Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 2000, volum 24, núm. 1
Segona època

Entitats patrocinadores:

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

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 2000, volum 24, núm. 1

SUMARI

EDITORIAL

En aquest primer número del volum 24 (2000) s'hi publiquen un total de vuit articles, repartits entre les quatre seccions temàtiques de la revista, als quals cal afegir-hi en aquesta ocasió dues ressenyes de l'apartat «Comentaris de llibres». D'altra banda, també s'incorpora un nou apartat al final de la revista que dóna compte de les novetats editorials en matèria estadística que publica l'Institut d'Estadística de Catalunya i la resta de la Generalitat de Catalunya. En conseqüència, el nombre d'articles i l'extensió d'altres continguts inclosos en aquest número permet suposar que la publicació final de l'actual volum superarà de nou els registres globals de la 2a època de *Qüestiió*, incloent-hi la producció mitjana dels darrers anys (que han esdevingut els més fructífers). També és interessant constatar el dinamisme de la seva versió electrònica: més de 18.300 consultes al web en el 1999, primer any complet que comptabilitza els accessos al domini de *Qüestiió*, que continuen creixent ininterrompudament fins a les 6.000 peticions enregistrades el mes de març d'enguany, situant el volum de consultes d'aquest primer trimestre per sobre de les 10.800.

Una altra novetat destacable, de caràcter més tècnic, és la imminent adopció de la nova classificació dels articles que es publiquin en el decurs del volum 24, basada en la «2000 Mathematical Subject Classification» (MSC2000) que l'*American Mathematic Society* ha desenvolupat a partir de l'actualització de l'anterior MSC 1991. En aquest sentit, *Qüestiió* posa a disposició dels autors i lectors que ho desitgin el sistema complet de la nova classificació, incloent-hi un enllaç electrònic per a l'adscripció dels descriptors classificatoris mitjançant la cerca de les paraules clau de l'article i la consulta de les correspondències amb l'antiga MSC 1991 que es facilita a través del web de l'Idescat.

Finalment, tal com s'avançava en el darrer número del 1999, el creixement sostingut de les despeses materials fa inajornable actualitzar l'import de la subscripció de *Qüestiió* per a l'any 2000 (volum 24). No obstant els esforços de les entitats patrocinadores per adequar les seves aportacions, l'augment dels ingressos en subscripcions i la inserció d'anuncis, no és possible mantenir el mateix preu de subscripció per cinquè any consecutiu. En qualsevol cas, s'espera que l'actualització del cost es compensi abastament per les millores qualitatives i quantitatives que la revista ha experimentat en els darrers anys.

C. Cuadras i E. Ripoll, editors executius

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

La secció «Estadística» conté dos originals. El primer, *Approximate maximum likelihood estimation for a spatial point pattern*, de J. Mateu i F. Montes, és un estudi comparatiu, mitjançant simulació, de cinc mètodes d'estimació màxim versemblant dels paràmetres d'un model espacial, amb conclusions diverses sobre la idoneïtat de cada mètode. En segon lloc, l'article *Spatial structure analysis using planar indices*, de J.M. Albert, J. Mateu i J.C. Pernías, és igualment un estudi comparatiu d'índexos d'agregació i de la seva distribució mitjançant tècniques de simulació, amb una aplicació a les coordenades geogràfiques i dades demogràfiques de les províncies de Barcelona i Madrid.

L'únic article publicat a la secció «Investigació Operativa», *Diseño de algoritmos para el problema del transporte escolar. Aplicación en la provincia de Burgos*, de J.A. Pacheco, A. Aragón i C. Delgado, és una aplicació al problema del transport d'alumnes que han de ser recollits i traslladats a centres escolars: els autors tracten de que el cost resultant sigui mínim, amb algunes restriccions, i que la solució sigui socialment racional en termes de comoditat, temps i millors rutes del transport.

Finalment, l'article inclòs a la secció «Biometria», *Comparación de dos tablas demográficas: aproximación a su significatividad estadística*, d'E.J. Veres, estudia la possible igualtat de dues taules demogràfiques en el mateix àmbit territorial, però referenciades en moments diferents del temps (contrast temporal), o bé amb la mateixa referència temporal, però situades en àmbits espacials diferents (contrast territorial): en aquest cas, l'autor utilitza el conegut test khi-quadrat d'homogeneïtat, amb algunes adaptacions a les especials característiques que ofereixen les dades demogràfiques.

Carles Cuadras, editor executiu

Comentari de la secció «Estadística Oficial» i altres apartats

Amb aquest número 1 del volum 24, la secció «Estadística Oficial» finalitza la publicació selectiva de vuit ponències presentades a les primeres jornades internacionals «Generació d'informació estadística: qualitat i limitacions» que va organitzar el 1998 la Xarxa temàtica «Enquestes i qualitat de la informació estadística». Els dos darrers treballs que s'editen corresponen a reflexions sobre el procés de producció estadística: d'una banda, sota el títol *Diversitat i complementarietat de les fonts estadístiques*, À. Costa presenta una novedosa caracterització de les fonts oficials d'àmbit preferentment regional, basada en la combinació dels paràmetres quantitat, puntualitat i fiabilitat

dels resultats amb la tipologia de processos i productes generats a l'Institut d'Estadística de Catalunya. Als efectes del planejament local, la ponència *Disseny i construcció d'una mostra estratificada a partir de dades censals*, de P. López, C. Lozares i M. Domínguez, il·lustra les possibilitats que ofereix la reutilització d'operacions mostrals, tan pel que fa a la fiabilitat dels estrats com a font d'informació per a l'anàlisi, mantenint la dimensió de la mostra general, com per facilitar criteris de validació de la mateixa enquesta. La secció es completa amb dos articles orientats a la producció d'estadístiques regionals més específiques, que poden extrapolar-se a nivell nacional. D'una banda, a *Comparative analysis of the alternative samplings plans used to created a farm accountancy data network for the agriculture of Navarra*, L. Júdez i C. Chaya analitzen l'impacte d'estratègies alternatives en el disseny mostral que suporta la xarxa comptable agrària de Navarra, avaluades segons la precisió dels estimadors i la introducció de les unitats geogràfiques oficials com a factors d'estratificació. Finalment, l'article *Avantatges i inconvenients de la metodologia de l'INE per elaborar indicadors de la producció industrial per a les regions espanyoles*, de M. Clar, R. Ramos i J. Suriñach, permet entreveure les prestacions desiguals que oferirien indicadors quantitatius de la producció industrial regional, basats en mètodes indirectes i/o sintètics, matitzant la generalització dels nivells de fiabilitat satisfactoris que, en canvi, s'assoleixen en el cas de Catalunya i el País Basc.

A continuació, la «Secció docent i problemes» incorpora de nou la presentació successiva de nous enunciats de problemes i la resolució dels publicats en el número anterior.

Seguidament, la secció «Comentaris de llibres» acull la recensió de tres obres recents de caràcter introductori. En primer lloc, E. Bonet comenta el llibre de text *Introducció a l'anàlisi i disseny d'algorismes*, de F.J. Ferri, J. Albert i G. Martín (Universitat de València), destacant-ne l'encert dels autors en la recreació d'una terminologia tècnica molt adient per a l'estudi de la programació computacional. D'altra banda, C.M. Cuadras recensiona dues publicacions força diferenciades: la primera correspon a la recent traducció al castellà de *The Cartoon Guide to Statistics*, de L. Gonick i W. Smith, on es manté el mateix to desenfadat amb el qual els autors il·lustren els fonaments de la probabilitat i l'estadística; més endavant, comenta l'actualitat i claredat de *An Introduction to Copulas*, de R.B. Nelsen, una publicació adient per familiaritzar-se amb les prestacions de les còpules en l'anàlisi de distribucions multivariades i nocions de dependència.

Per últim, l'apartat dedicat a «Ressenyes d'activitats institucionals» inclou, com ja és habitual, una referència actualitzada de la *Sociedad Española de Biometría*, amb les primeres informacions sobre la VIII Conferència Espanyola de Biometria (Pamplona, 28-30 de març 2001). En aquesta mateixa línia, es publica una nova recensió del «Training for European Statisticians Institute», on s'inclouen els cursos del *Core Programme 1999-2000* que s'impartiran fins el mes de juliol, adreçats preferentment als membres d'instituts d'estadística oficial d'àmbit comunitari. A continuació, s'anuncia una nova edició de l'«Applied Statistics Week» (Barcelona, 26-30 de juny 2000) que organitza per sisena vegada l'Institut d'Educació Contínua de la Universitat Pompeu

Fabra, amb la col·laboració sistemàtica de l'Idescat; el programa d'enguany inclou els tres cursos següents: *Statistical Information in Education, Probability and Statistics for Law i Statistics in Political and Social Opinion Polling*. Seguidament, també es reprodueix un darrer anunci més detallat del «Tercer Congrés Europeu de Matemàtiques» (Barcelona, 10-14 de juliol 2000) que organitza la Societat Catalana de Matemàtiques-IEC i que, sota els auspicis de la European Mathematical Society i en el marc de l'Any Internacional de la Matemàtica, hi col·laboren tres entitats patrocinadores de *Qüestiió* (Idescat, UB i UPC), entre d'altres organismes. Igualment, es reprodueix l'anunci de l'«International Workshop on Statistical Modelling» –15th IWSM– (Bilbao, 17-21 juliol 2000) que organitza la Universitat del País Basc. Finalment, tal com s'advertia a l'inici de l'editorial, les darreres pàgines es dediquen a les novetats editorials en l'àmbit de l'estadística per part de la Generalitat de Catalunya, informació que es renovarà puntualment en successius números de *Qüestiió*.

Enric Ripoll, editor executiu

Estadística

QÜESTIIÓ, vol. 24, 1, p. 3-25, 2000

APPROXIMATE MAXIMUM LIKELIHOOD ESTIMATION FOR A SPATIAL POINT PATTERN*

JORGE MATEU*
FRANCISCO MONTES**

Several authors have proposed stochastic and non-stochastic approximations to the maximum likelihood estimate for a spatial point pattern. This approximation is necessary because of the difficulty of evaluating the normalizing constant. However, it appears to be neither a general theory which provides grounds for preferring a particular method, nor any extensive empirical comparisons. In this paper, we review five general methods based on approximations to the maximum likelihood estimate which have been proposed in the literature. We also present the results of a comparative simulation study developed for the Strauss model.

Keywords: Gibbs distribution, maximum likelihood, Monte Carlo inference, stochastic approximation, Strauss model

AMS Classification: 62M05, 60G55

* This work has been partially supported by grant TIC 98-1019 of the Spanish Ministry of Education and Culture.
* Departament de Matemàtiques. Universitat Jaume I. Campus Riu Sec. 12071 Castelló. Spain.
** Departament de Estadística i I.O. Universitat de València. Dr. Moliner, 50. 46100 Burjassot. Spain.
- Received May 1998.
- Accepted October 1999.

1. INTRODUCTION

A spatial point pattern is a set of points

$$X = \{x_i \in A : i = 1, \ldots, n\}$$

for some planar region A. The x_i are called events to distinguish them from generic points $x \in A$. Very often, A is a sampling window within a much larger region and it is reasonable to regard X as a partial realization of a planar point process, the events consisting of all points of the process which lie within A.

Parameter estimation for two-dimensional point pattern data is difficult, because most of the available stochastic models have intractable likelihoods (see Ripley, 1977, 1988 and Diggle, 1983). An exception is the class of Gibbs or Markov point processes (Baddeley and Moller, 1989; Ripley, 1989), where the likelihood $l(X;\theta)$ typically forms an exponential family and is given explicitly up to a normalizing constant. However, the latter is not known analytically precluding the use of exact maximum likelihood, so parameter estimates must be based on approximations.

Gibbs point processes first appeared in the theory of statistical physics, where Gibbs distributions were applied to describe the equilibrium states of closed physical systems of interacting objects. In mathematical statistics Gibbs point processes are used as models of spatial point patterns. A preliminary paper introducing the Gibbs processing into the statistical literature is Ripley and Kelly (1977). Examples can be found in biology, plant ecology, forestry and economy.

The topic of this paper concerning Gibbs type processes has a general validity arising from two aspects: (i) It is a general way of proceeding in cases of exponential families with dependent samples, and (ii) it has theoretical value on its own. Examples of (i) are the applications of Markov random fields for lattice data (Besag, 1974; Geyer and Thompson, 1992), Markov random fields in image analysis (Geman and Geman, 1984), Gibbs point processes and germ-grain models in high level image analysis (Baddeley and van Lieshout, 1993), modelling of random graphs and general interaction models (Strauss, 1986). Gibbs processes are useful as prior distributions in image interpretation tasks, such as object recognition, edge detection and feature extraction (van Lieshout and Baddeley, 1995; Molina and Ripley, 1989). Maximum likelihood solutions tend to suffer from multiple response and the prior distribution serves to penalize scenes with too many almost identical objects, disconnected or crossing edges. Usually, the posterior distribution also possesses a Markov property, enabling sampling and optimization by iterative procedures that recursively update the scene by simple operations of addition or deletion.

In this paper, we consider generally applicable methods for estimating the parameter θ confining our attention to stochastic and non-stochastic approximations to the maxi-

mum likelihood estimate (MLE). We use a simple point process model, the Strauss process (Strauss, 1975), to illustrate and compare these methods which could be applied to more general and complex models. The Strauss process is a point process model which has been used in modelling (non-clustered) point patterns in some of the mentioned references and is a demanding member of the exponential family for a dependent sample.

The interest of the present paper relies on methods of estimation which can be used routinely in applications, and which do not place artificial restrictions on the parametric form of $l(X;\theta)$. The aim is to present a comparative study among the approximations to the MLE and to discuss the practical implications. We consider only homogeneous, i.e., stationary and isotropic processes. Throughout this paper, $N(A)$ stands for the number of events in A, $|A|$ denotes the area of A and $\lambda = E[N(A)]/|A|$ denotes the intensity of the process.

For a general introduction to statistical methodology for spatial point patterns, see for example Ripley (1981), Diggle (1983), Stoyan, Kendall and Mecke (1995) and Cressie (1993). Other parametric methods of estimation, not considered here, are maximum pseudo-likelihood and the Takacs-Fiksel method (Diggle et al., 1994; Takacs, 1986). In a different vein, Diggle, Gates and Stibbard (1987) develop a smooth, non-parametric estimator for the interaction function, to which a parametric family could be fitted by standard curve-fitting techniques such as non-linear least squares.

The plan of the paper is as follows. Section 2 describes the approximate MLE methods for a particular Gibbs process, the Strauss model. Section 3 shows the simulation study to compare the different methods. The paper ends with a section of final conclusions.

2. APPROXIMATE MLE FOR A GIBBS PROCESS

A class of stochastic models for patterns of n events in a bounded region A is the class of *pairwise interaction point processes*. The joint density for a pattern X, taken with respect to the Poisson measure μ, is given by

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

In (1), $\|.\|$ denotes Euclidean distance, $\Phi(.)$ is a *potential function* depending on a set of parameters θ, β is a parameter which determines the intensity of the process, and $C(\theta)$ is a normalizing constant. We call $U_n(X;\theta) = \sum_{i=1}^{n}\sum_{j>i}\Phi(\|x_i - x_j\|;\theta)$ the *total potential energy*. Often, (1) is written in terms of an *interaction function* $e(t) = \exp(-\Phi(t))$. Such class of point processes belongs to a more general kind of processes called *Gibbs processes* (Kelly and Ripley, 1976; Daley and Vere-Jones, 1988; Baddeley and Moller,

1989). Note that restrictions on the form of the potential $\Phi(.)$ are needed to ensure that the normalizing constant in (1) is finite.

A *Strauss process* (Strauss, 1975) is a pairwise interaction process in which the density depends only on the number of neighbour pairs defined by

$$s(X) = \sum_{i=1}^{n} \sum_{j>i} I(\|x_i - x_j\| \leq r).$$

Considering in (1) the Strauss potential function

$$\Phi(t) = \begin{cases} -\log(\theta), t \leq r \\ 0, t > r \end{cases}$$

the likelihood takes the form (Kelly and Ripley, 1976)

$$l(X;\theta) = \exp(-|A|)\alpha(\theta)^{-1}\beta^n\theta^{s(X)}$$

where the normalizing constant is $C(\theta) = \alpha(\theta)/\exp(-|A|)n!$ The case $\theta = 1$ corresponds to a Poisson process with intensity β. If $\theta = 0$, the result is a simple inhibition process that contains no events at a distance less than or equal to r. Values of $\theta < 1$ correspond to regularity of events, whilst for $\theta > 1$ the process should result in clustering (see Figures 1a, 1b and 1c). For a clustered pattern, as was pointed out by Kelly and Ripley (1976), the condition $\theta > 1$ violates the requirement of a finite normalizing constant $C(\theta)$ in (1). This problem can be removed by conditioning to the number of events, say $N = n$. This is not an artificial restriction because $n(X)$ usually provides little information about the interactions among the events. The effect on conditioning to the MLE for the Strauss family has been demonstrated by Geyer and Moller (1994). Furthermore, conditioning on n makes it easier to generate simulations by the discrete-time Markov chain method of Ripley (1979, 1987). The conditional likelihood function for the Strauss process is given by

$$l_n(X;\theta) = \theta^{s(X)}/C_n(\theta) \tag{2}$$

where the normalizing constant is given by

$$C_n(\theta) = \int_{A^n} \theta^{s(X)} dx_1 \cdots dx_n. \tag{3}$$

Maximum likelihood estimation of θ requires the evaluation of (3) which is not usually obtainable in closed form. We therefore try to maximize an approximation to the likelihood function. In the following, we develop approximations to the MLE for the Strauss conditional model.

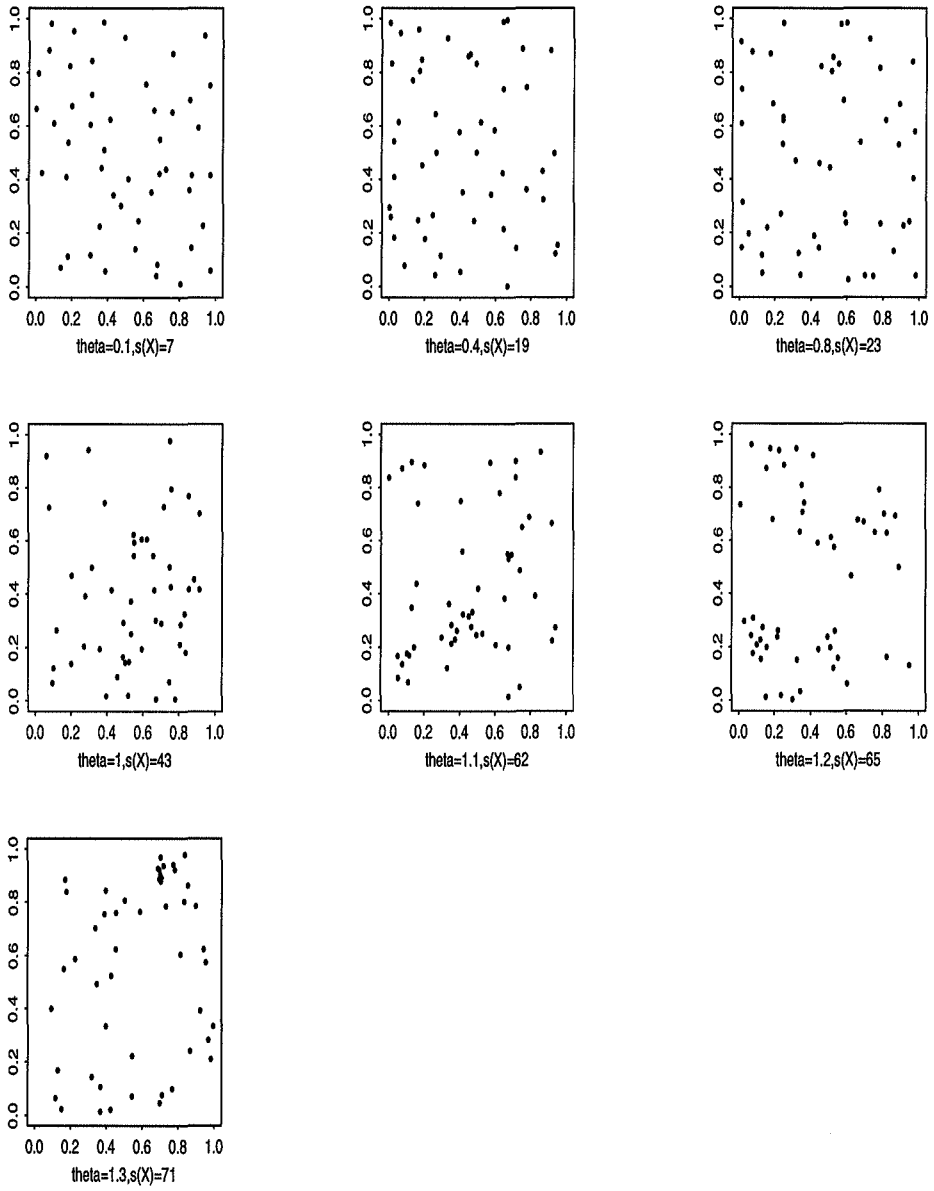


Figure 1a. Realizations of simulated patterns under the Strauss model for different values of parameter θ. In each pattern it is also included $s(X)$, the number of neighbour pairs. r=0.10

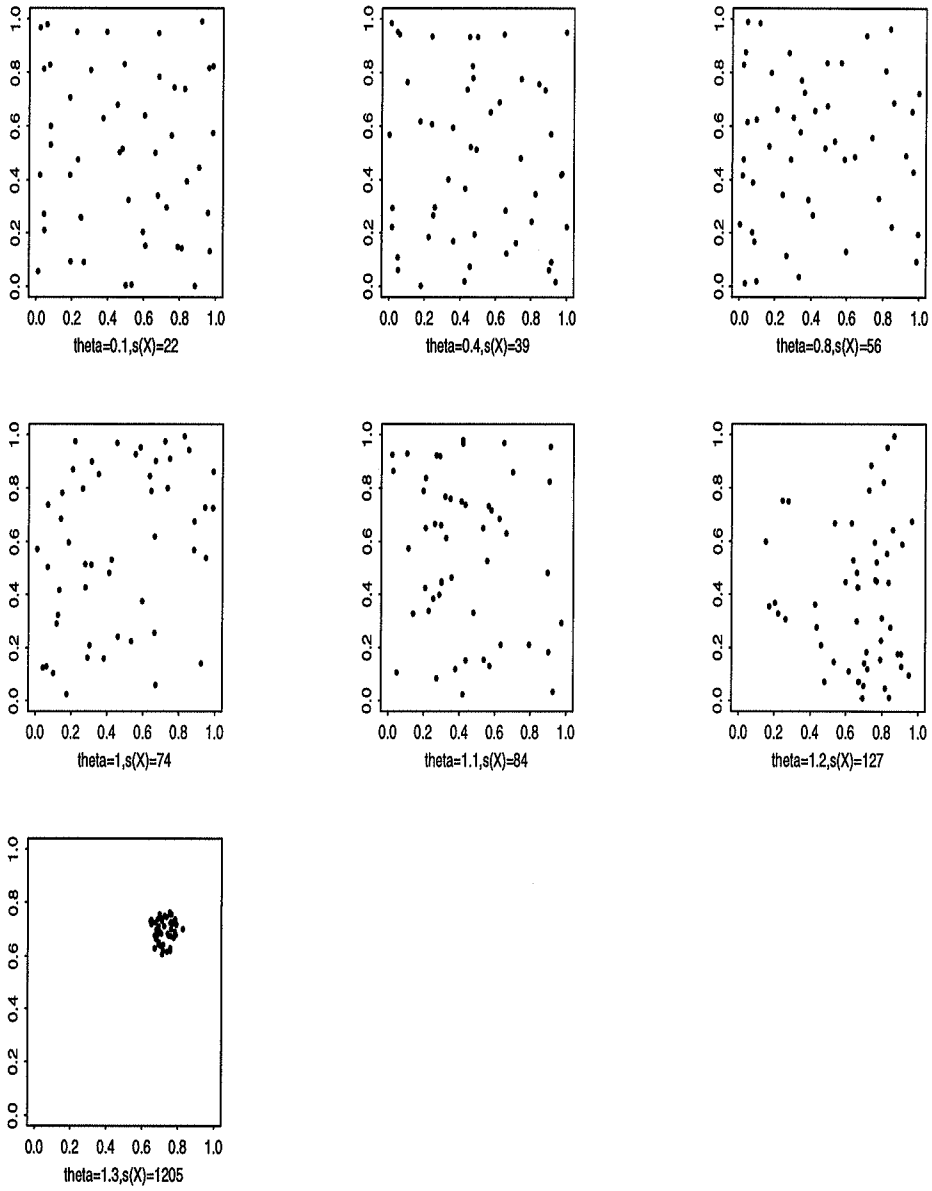


Figure 1b. Realizations of simulated patterns under the Strauss model for different values of parameter θ. In each pattern it is also included $s(X)$, the number of neighbour pairs. r=0.15

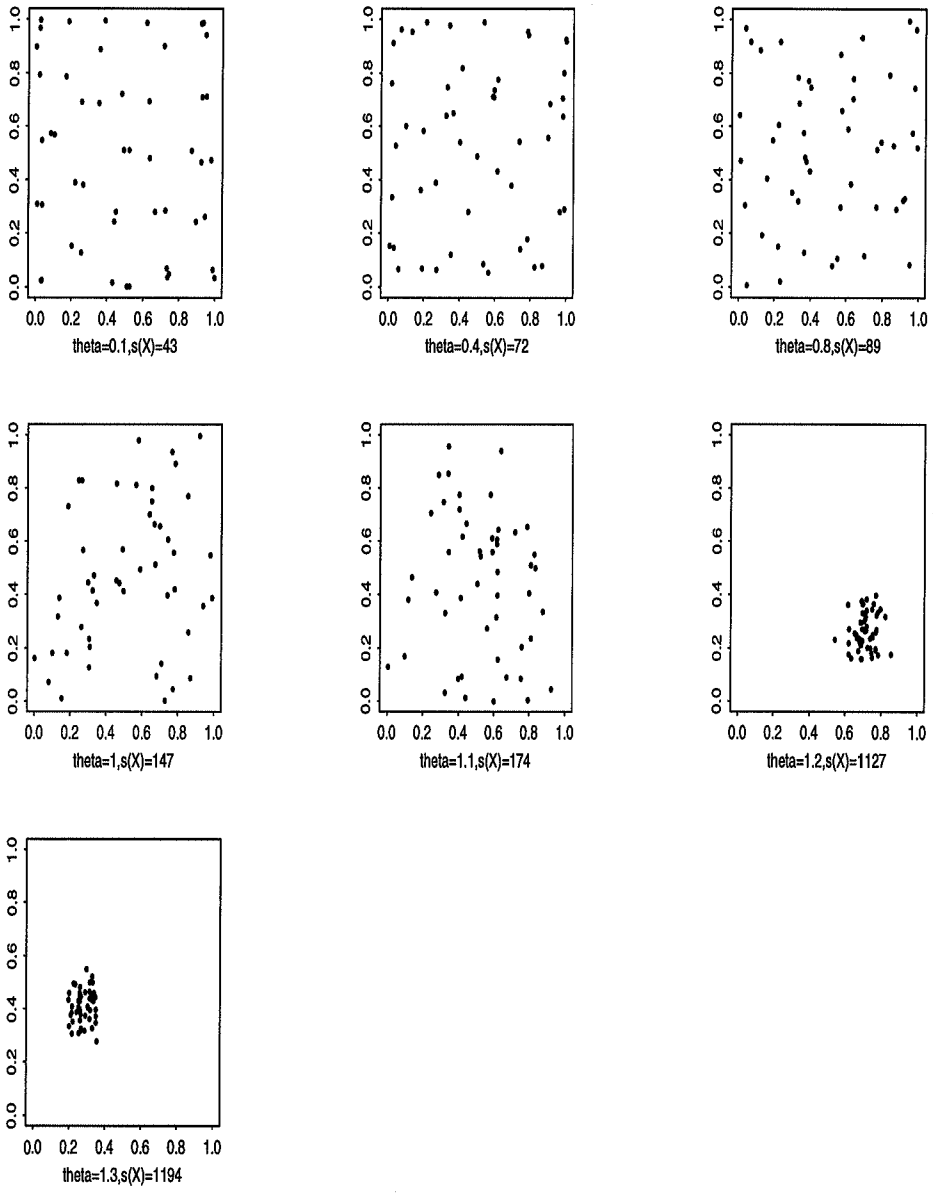


Figure 1c. Realizations of simulated patterns under the Strauss model for different values of parameter θ. In each pattern it is also included $s(X)$, the number of neighbour pairs. r=0.20

2.1. Method of Ogata-Tanemura

Ogata and Tanemura (1981) proposed to use a cluster-expansion method of statistical mechanics assuming that the events of the point process are sparsely distributed, so that third and higher-order cluster integrals are negligible. Then, using up to the second-order cluster integral, an approximation to the normalizing constant is given by

$$C_n(\theta) = |A|^n \left\{ 1 - \frac{b(\theta)}{|A|} \right\}^{n(n-1)/2}$$

where $b(\theta)$ is, for the Strauss model, $b(\theta) = \pi(1-\theta)r^2$. Then, the MLE is given by

$$\widehat{\theta} = \frac{s(X)\{2|A| - \pi r^2\}}{\pi r^2 (n(n-1)/2 - s(X))}. \tag{4}$$

2.2. Method of Penttinen

Penttinen (1984) proposed another sparse-data approximation to (3), which for the Strauss model takes the form

$$C_n(\theta) = \exp\{1/2n(n-1)\pi r^2(\theta - 1)\}$$

and the MLE is given by

$$\widehat{\theta} = \frac{s(X)}{1/2n(n-1)\pi r^2}. \tag{5}$$

2.3. Method of virial expansions

This method consists of the following approximation of (3),

(6)

$$n^{-1}\log(C_n) \approx (b_n/2)\int_{\Re^2} f_{12}dx_2 + (b_n^2/4)\int_{\Re^4} f_{12}f_{13}f_{23}dx_2dx_3 + (b_n^3/8)\int_{\Re^6}(f_{12}f_{13}f_{14}f_{23}f_{24}f_{34} + 6f_{12}f_{13}f_{14}f_{23}f_{24} + 3f_{12}f_{14}f_{23}f_{34})dx_2dx_3dx_4 + \cdots$$

where $b_n = n/|A|$ and $f_{ij} = \exp(-\Phi(\|x_i - x_j\|;\theta)) - 1$ (Ripley, 1988). To implement this method for the Strauss process we use the fourth order expansion obtained by calculating the integrals in (6):

$$\log(C_n(\theta)) = -\pi n(n-1)\Psi r^2/(2|A|) - 0.29325\pi^2 \tfrac{n!}{(n-3)!}\Psi^3 r^4/(6|A|^2) - \pi^3 \tfrac{n!}{(n-4)!}\{-0.27432\Psi^6 + 2.18542\Psi^5 - 1.37886\Psi^4\}r^6/(24|A|^3)$$

where $\Psi = 1 - \theta$.

Then solving

$$d\log(C_n(\theta))/d\theta = s(X)/\theta \tag{7}$$

we obtain $\widehat{\theta}$, the approximate MLE of θ.

2.4. Stochastic approximation based on a Newton-Raphson procedure

Penttinen (1984) suggested a Newton-Raphson type algorithm for solving the maximum likelihood estimating equation. Assume $\Phi(t;\theta)$ is twice differentiable with respect to θ. Differentiation of both sides of equation (3) yields

$$\frac{-\partial C_n(\theta)}{\partial\theta} = C_n(\theta)E_\theta[\partial U_n(x_1,\ldots,x_n;\theta)/\partial\theta]$$

where the *total potential energy*, for the Strauss process, is

$$U_n(x_1,\ldots,x_n;\theta) = -\log\theta^{s(X)}.$$

The MLE $\widehat{\theta}$ solves $\partial\log(l_n(X;\theta))/\partial\theta = 0$. If $\widehat{\theta}_0$ denotes an initial guess for $\widehat{\theta}$, then the Newton-Raphson algorithm consists of

$$\widehat{\theta}_{k+1} = \widehat{\theta}_k - \left[\overline{\Gamma}_T(\widehat{\theta}_k)\right]^{-1}\overline{\beta}_T(\widehat{\theta}_k) \quad k = 0,1,2,\ldots$$

where

$$\overline{\beta}_T(\widehat{\theta}_k) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\widehat{\theta}_k}[s(X) - s(\phi_n(t))] \tag{8}$$

and

$$\overline{\Gamma}_T(\widehat{\theta}_k) = \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\frac{1}{\widehat{\theta}_k^2}[s(\phi_n(t)) - s(X)]$$
$$-\left\{\tfrac{1}{\widehat{\theta}_k}[s(X) - s(\phi_n(t))] - \overline{\beta}_T(\widehat{\theta}_k)\right\}^2.$$

Note that $\phi_n(1),\ldots,\phi_n(T)$ are simulated according to a Strauss process with parameter $\widehat{\theta}_k$.

2.5. Stochastic approximation based on Robbins-Monro procedure

This stochastic approximation procedure was first introduced by Robbins and Monro (1951) and can be used to estimate the solution θ^* of an equation $F(\theta^*) = \varphi$ when there

is very little information about the function F but it is possible, for any given θ, to generate a random variable T_θ with expectation $E(T_\theta) = F(\theta)$.

For the Strauss model, the goal is to solve

$$M(\widehat{\theta}) = s(X) \tag{9}$$

for $\widehat{\theta}$, where X is the observed data and $M(\theta) = E_\theta[s(X)]$. Then we set $T_\theta = s(X_\theta)$, where X_θ is a simulated Strauss process with parameter θ and we obtain, recursively, a sequence of estimates of $\widehat{\theta}$ using

$$\theta_{k+1} = \theta_k + \frac{B}{k}\left\{s(X) - s(X_{\theta_k})\right\}.$$

Then $\theta_k \to \widehat{\theta}$ (a.s.) (Moyeed and Baddeley, 1991).

Defining $\mu = M'(\widehat{\theta})$ and $\sigma^2 = Var_\theta[s(X)]$, if $B > 1/(2\mu)$ then θ_k is asymptotically normally distributed with mean $\widehat{\theta}$ and variance $B^2\sigma^2/(2B\mu - 1)$.

The starting value θ_0 is arbitrary, but should be set to an initial approximation such as that holding in the sparse case

$$\theta_0 = \frac{2s(X)|A|}{n(n-1)\pi r^2}.$$

The optimum B, B_{opt}, could be estimated by

$$B_{opt} = \frac{1}{\mu} = \frac{1}{M'(\widehat{\theta})}$$

or

$$B_{opt} = \frac{2|A|}{n(n-1)\pi r^2}.$$

3. A SIMULATION STUDY

3.1. Edge-correction

Commonly, the region A is a sampled sub-region of a much larger region within which the phenomenon operates and some form of edge-correction is vital. When A is a rectangle, a possible strategy is to map A onto a torus by identifying opposite edges. This

periodic boundary is commonly used for computer experiments in statistical mechanics. However, for the analysis of real data, periodic boundaries can introduce undesirable artefacts: toroidal distances can be arbitrarily small even when the underlying process has a positive hard-core distance. In the present comparative simulation study, the points patterns were themselves generated using a periodic boundary, then this particular difficulty does not arise.

To compensate for the omission of contributions to the total potential from unobserved events outside A we replace summations of the form

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

by

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

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 is an adapted version of Ripley's correction (Ripley, 1977, 1988). The majority of available edge-corrections correct the bias using lengths or areas of parts of circles or discs, respectively.

In the simulation study, we also include results using the so-called free boundary conditions, in which no edge-correction at all is made.

3.2. Standard Errors

One possible way to obtain approximate standard errors is by using Monte Carlo methodology. For this approach, we simulate s realisations with $\theta = \hat{\theta}$, the point estimate under the chosen method for the original data. We then evaluate point estimates $\hat{\theta}_j, j = 1, \ldots, s$ from the simulated patterns and use the empirical distribution of the $\hat{\theta}_j$ as an approximation to the sampling distribution of $\hat{\theta}$. In particular, the sample mean and standard deviation of the $\hat{\theta}_j$ give useful indications of the bias and efficiency of estimation. This Monte Carlo approach is highly computer-intensive and it is usually known as parametric bootstrap.

3.3. Simulation method

The *spatial birth-and-death process* provides the framework under which Ripley (1977, 1979) 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 by Hastings (1970). Consider a set of particles interacting according

to a certain potential function on a set A with periodic boundary, i.e. A is identified with a torus. First, select n events from a uniform distribution on A; call this initial point pattern $\phi_n(0)$. At step $(t+1)$, delete systematically in turn one of the n events of $\phi_n(t) = \{x_1, \ldots, x_n\}$, say event x_i, and let $\phi_n(t) - \{x_i\}$ denote the point pattern formed by removing x_i from $\phi_n(t)$. Let

$$p(u; \phi_n(t) - \{x_i\}) = \frac{l_n(\phi_n(t) - \{x_i\}, u)}{l_{n-1}(\phi_n(t) - \{x_i\})}$$

denote the conditional intensity at $u \in A$ given $\phi_n(t) - \{x_i\}$. Define

$$M = \sup_{u \in A} p(u; \phi_n(t) - \{x_i\}).$$

Select an event u from a uniform distribution on A and set $\phi_n(t+1) = \{\phi_n(t) - \{x_i\}, u\}$ with probability $p(u; \phi_n(t) - \{x_i\})/M$; 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 $l_n(.)$ will occur.

Unfortunately, in the case of the Strauss model, for θ much larger than 1 the algorithm is very slow and may result in simulation difficulties (see Figures 1b and 1c when $\theta = 1.2$ and 1.3).

3.4. Design of the Simulation Study

For the simulation study we selected eight parameter values: $\theta = 0.1, 0.4$ and 0.8 corresponding to regular patterns; $\theta = 1$ for the random pattern (Poisson process) and $\theta = 1.1, 1.2$ and 1.3 for clustered ones (strongly interactive patterns). We also considered three different ranges of interaction: $r = 0.1, 0.15$ and 0.2. For each combination of parameter value and range of interaction we simulated 100 realizations, each one with $n = 50$ events on A the unit square. From the simulated realization we evaluated the estimate of θ using the five methods of approximation described in Section 2 and incorporating the edge-correction described in Section 3.1.

Each combination of parameter value, interaction range, method of estimation and edge-correction (no edge-correction, Ripley's and toroidal (periodic)) therefore yielded 100 estimates $\widehat{\theta}_j, j = 1, \ldots, 100$, which are summarised by the box-plots shown in Figures 2a, 2b and 2c.

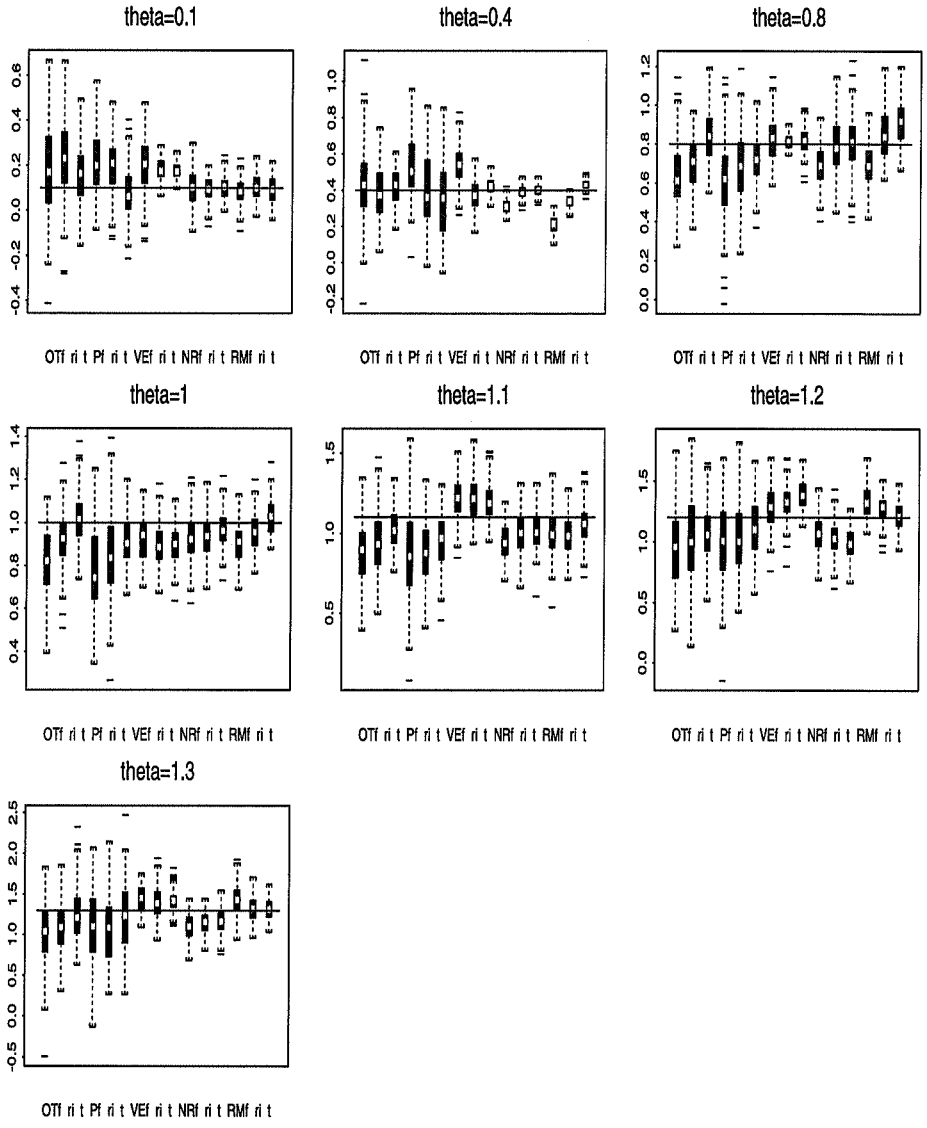


Figure 2a. Box-plots of simulated parameter estimates. The horizontal lines indicate the true value of θ. The upper case letter identifies the method of estimation (OT=Ogata-Tanemura, P=Penttinen, VE=Virial Expansions, NR=Newton-Raphson, RM=Robbins-Monro), the lower case letter identifies the boundary condition (f=free, ri=Ripley, t=toroidal). r=0.10

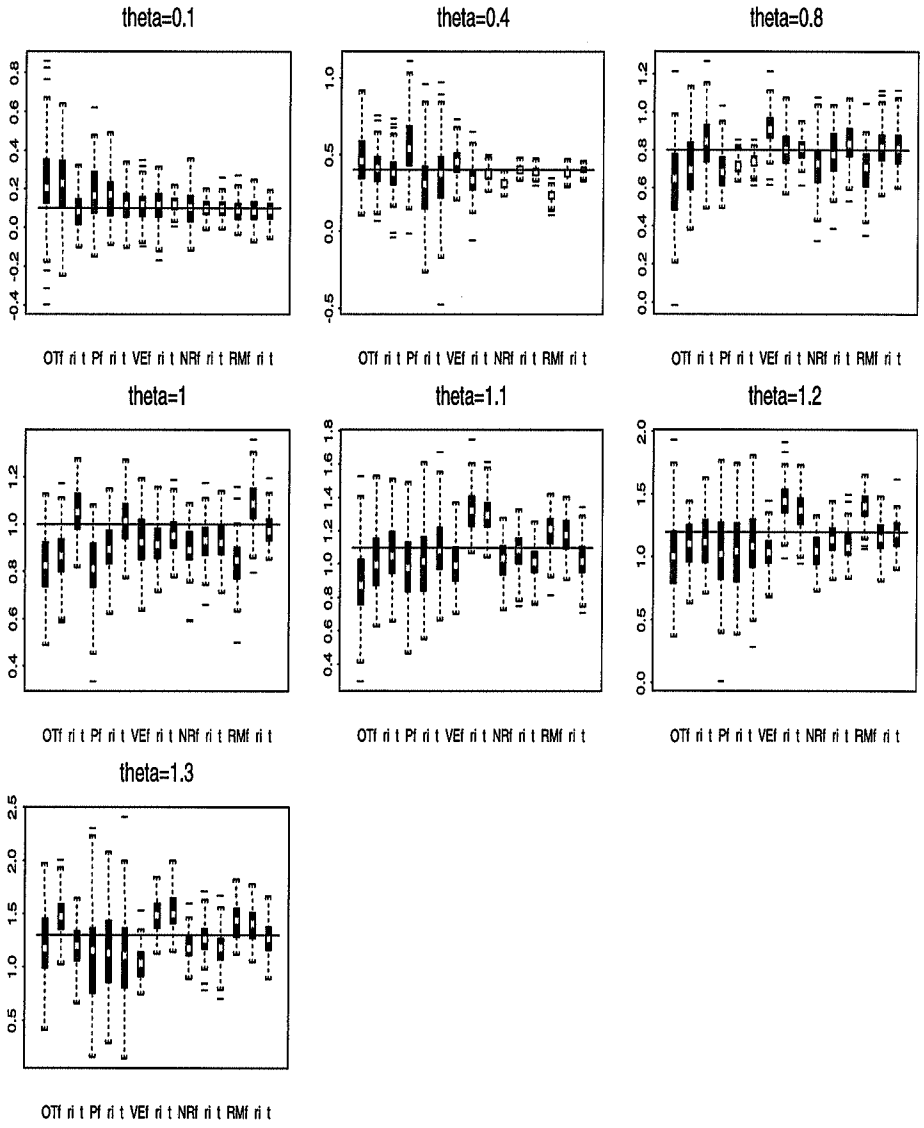


Figure 2b. Box-plots of simulated parameter estimates. See legend of Figure 2a. r=0.15

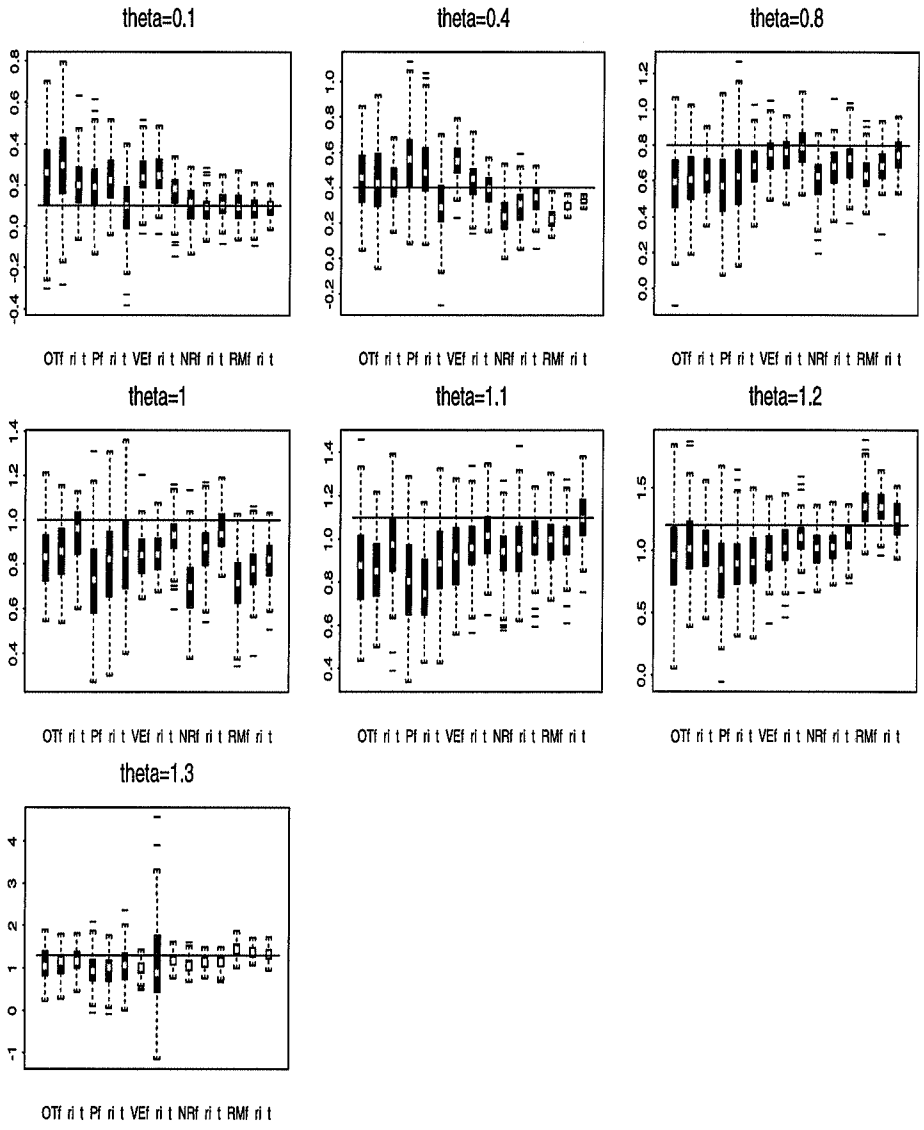


Figure 2c. Box-plots of simulated parameter estimates. See legend of Figure 2a. r=0.20

Table 1. Sample means and standard errors of parameter estimates when the interaction radius is $r = 0.10$. Each entry is based on 100 replicate simulations of n=50 events on the unit square. Lower case letters indicate the boundary condition: free, Ri=Ripley and toro=toroidal.

r=0.10	O-T method			Pent. method			V-E method			N-R method			R-M method		
	free	Ri	toro	free	Ri	toro	free	Ri	toro	free	Ri	toro	free	Ri	toro
θ	Sample Means														
0.1	0.247	0.231	0.091	0.193	0.176	0.087	0.112	0.125	0.122	0.095	0.097	0.098	0.093	0.091	0.096
0.4	0.434	0.427	0.407	0.541	0.308	0.345	0.456	0.319	0.376	0.305	0.403	0.393	0.231	0.378	0.404
0.8	0.645	0.714	0.835	0.703	0.715	0.732	0.913	0.809	0.805	0.704	0.791	0.807	0.710	0.803	0.824
1.0	0.831	0.873	1.041	0.810	0.912	1.013	0.919	0.920	0.946	0.910	0.935	0.931	0.847	1.079	0.979
1.1	0.914	0.979	1.093	0.973	1.007	1.081	1.013	1.315	1.291	1.004	1.073	0.993	1.215	1.183	1.032
1.2	0.997	1.101	1.103	1.031	1.046	1.097	1.035	1.416	1.335	1.053	1.143	1.103	1.392	1.194	1.197
1.3	1.124	1.445	1.213	1.093	1.106	1.148	1.056	1.496	1.531	1.197	1.292	1.197	1.431	1.393	1.245
θ	Standard Errors														
0.1	0.215	0.210	0.091	0.141	0.115	0.091	0.090	0.091	0.037	0.093	0.051	0.047	0.061	0.057	0.053
0.4	0.171	0.135	0.131	0.205	0.217	0.215	0.099	0.101	0.048	0.039	0.032	0.031	0.043	0.038	0.029
0.8	0.176	0.156	0.132	0.115	0.043	0.039	0.105	0.101	0.066	0.127	0.125	0.112	0.128	0.118	0.105
1.0	0.125	0.112	0.111	0.127	0.115	0.113	0.107	0.096	0.080	0.098	0.083	0.081	0.103	0.097	0.065
1.1	0.215	0.203	0.193	0.235	0.215	0.195	0.135	0.137	0.122	0.121	0.115	0.107	0.120	0.119	0.117
1.2	0.323	0.213	0.211	0.341	0.312	0.247	0.156	0.165	0.146	0.143	0.135	0.127	0.142	0.138	0.129
1.3	0.351	0.225	0.212	0.451	0.410	0.393	0.170	0.171	0.157	0.161	0.149	0.143	0.160	0.153	0.141

Table 2. Sample means and standard errors of parameter estimates when the interaction radius is $r = 0.15$.

r=0.15	O-T method			Pent. method			V-E method			N-R method			R-M method		
	free	Ri	toro	free	Ri	toro	free	Ri	toro	free	Ri	toro	free	Ri	toro
θ	Sample Means														
0.1	0.197	0.215	0.142	0.198	0.186	0.082	0.197	0.183	0.172	0.094	0.096	0.096	0.092	0.092	0.093
0.4	0.431	0.416	0.412	0.515	0.412	0.319	0.511	0.392	0.431	0.304	0.393	0.409	0.214	0.341	0.431
0.8	0.613	0.705	0.841	0.609	0.674	0.705	0.819	0.812	0.811	0.703	0.793	0.805	0.695	0.849	0.907
1.0	0.813	0.912	1.010	0.805	0.845	0.906	0.905	0.896	0.915	0.921	0.935	0.963	0.896	0.945	1.031
1.1	0.887	0.946	1.035	0.874	0.885	0.948	1.193	1.203	1.195	0.973	0.987	0.995	0.994	0.998	1.051
1.2	0.944	0.997	1.102	1.005	1.045	1.103	1.298	1.305	1.399	1.047	1.031	1.034	1.314	1.293	1.227
1.3	1.034	1.125	1.204	1.091	1.112	1.131	1.423	1.397	1.430	1.092	1.141	1.195	1.443	1.348	1.338
θ	Standard Errors														
0.1	0.212	0.205	0.146	0.150	0.141	0.115	0.102	0.052	0.036	0.091	0.050	0.045	0.061	0.060	0.055
0.4	0.202	0.165	0.108	0.210	0.231	0.212	0.103	0.096	0.047	0.039	0.037	0.031	0.048	0.035	0.028
0.8	0.176	0.135	0.126	0.212	0.195	0.118	0.105	0.037	0.071	0.131	0.129	0.125	0.127	0.121	0.113
1.0	0.146	0.131	0.127	0.210	0.196	0.121	0.104	0.096	0.091	0.103	0.095	0.091	0.113	0.093	0.091
1.1	0.215	0.201	0.153	0.235	0.221	0.198	0.135	0.131	0.118	0.125	0.121	0.119	0.125	0.122	0.119
1.2	0.345	0.303	0.246	0.319	0.312	0.251	0.158	0.162	0.135	0.148	0.137	0.132	0.157	0.138	0.118
1.3	0.431	0.397	0.353	0.425	0.418	0.401	0.177	0.182	0.152	0.159	0.152	0.151	0.191	0.153	0.139

Table 3. Sample means and standard errors of parameter estimates when the interaction radius is $r = 0.20$.

r=0.20	O-T method			Pent. method			V-E method			N-R method			R-M method		
	free	Ri	toro	free	Ri	toro	free	Ri	toro	free	Ri	toro	free	Ri	toro
θ	Sample Means														
0.1	0.240	0.273	0.184	0.210	0.221	0.082	0.253	0.231	0.165	0.092	0.093	0.093	0.090	0.091	0.093
0.4	0.450	0.443	0.415	0.539	0.495	0.324	0.551	0.453	0.397	0.241	0.295	0.335	0.221	0.298	0.321
0.8	0.595	0.593	0.625	0.605	0.614	0.693	0.719	0.771	0.789	0.615	0.693	0.705	0.630	0.710	0.747
1.0	0.841	0.855	0.931	0.741	0.793	0.845	0.810	0.839	0.921	0.710	0.845	0.947	0.708	0.793	0.810
1.1	0.872	0.879	0.947	0.793	0.815	0.897	0.915	0.986	1.023	0.941	0.943	0.998	0.983	0.995	1.108
1.2	0.936	0.979	0.998	0.847	0.895	0.913	1.005	1.039	1.103	1.009	1.015	1.074	1.334	1.321	1.253
1.3	1.013	1.115	1.197	0.915	0.945	1.041	1.027	1.093	1.154	1.051	1.123	1.147	1.451	1.382	1.324
θ	Standard Errors														
0.1	0.230	0.211	0.135	0.151	0.134	0.127	0.105	0.104	0.098	0.091	0.085	0.073	0.065	0.058	0.047
0.4	0.215	0.197	0.109	0.198	0.197	0.201	0.107	0.107	0.098	0.102	0.098	0.085	0.058	0.032	0.019
0.8	0.178	0.167	0.132	0.201	0.195	0.119	0.103	0.102	0.101	0.131	0.130	0.121	0.115	0.103	0.102
1.0	0.153	0.136	0.131	0.203	0.201	0.202	0.104	0.095	0.092	0.134	0.128	0.112	0.121	0.098	0.095
1.1	0.198	0.185	0.176	0.218	0.212	0.205	0.142	0.139	0.126	0.141	0.139	0.129	0.128	0.125	0.123
1.2	0.351	0.298	0.255	0.325	0.301	0.247	0.181	0.173	0.148	0.158	0.138	0.132	0.199	0.143	0.128
1.3	0.398	0.299	0.301	0.441	0.412	0.395	0.183	0.179	0.165	0.171	0.163	0.159	0.183	0.161	0.146

3.5. Results and discussion

Tables 1, 2 and 3 give the results of the simulation study, expressed in terms of the *sample means*, $\overline{\theta}$ and *standard errors*, s_θ. These two statistics characterize the sampling distribution of the parameter estimates as noted in section 3.2 above as neither the theoretical nor asymptotic approximations of the parameter distribution are not known. The bias and efficiency of the estimation can only be assessed by means of a Monte Carlo approach. However, Bayesian procedures could also be used to approach the theoretical parameter distribution as in Mateu and Montes (1995).

The table values indicate that both stochastic approximation methods, Newton-Raphson (N-R) and Robbins-Monro (R-M), exhibited better results, in terms of bias and standard errors, than Ogata-Tanemura (O-T) and Penttinen (P) methods for cases of strong regularity ($\theta \leq 0.4$) and clustering ($\theta \geq 1.1$). The approximate maximum likelihood method based on virial expansions (V-E) exhibited substantial bias, particularly when θ is large; however, this is qualitatively predictable on theoretical grounds, since the adequacy of the approximation to the likelihood deteriorates as θ increases. Implementation of this approximation for any potential is straightforward if only low-order virial coefficients are required. This method is not suited for estimation in cases of strong interaction.

The O-T and P approximate maximum likelihood methods provided substantial negative bias for medium-to-large values of θ giving relatively large standard deviations for these values. These two approximations are based on the sparseness assumption and are not reliable methods for clustered patterns for which higher-order interactions become important.

Inspecting the standard errors in conjunction with the range of θ, we observe that approximate maximum likelihood using O-T and P methods, provide large standard deviations for small ($\theta \leq 0.4$) and large ($\theta \geq 1.1$) values of parameter θ and, in any case, they are much larger than those obtained with the other three methods.

For different values of θ, the choice of boundary condition becomes important. Generally, for any method and parameter values, the periodic boundary condition produced better results than Ripley's, and in turn they are better than those obtained with the free boundary condition. The N-R and R-M approximate maximum likelihood methods provided unbiased and efficient estimates for all ranges of parameter values, under periodic and Ripley's boundary condition. However, they provided biased estimates under the free boundary condition.

Comparing the behaviour of the bias and the standard errors of estimates among the three interaction radii, we observe that, under the same parameter value, method of estimation and boundary condition, biases and standard errors increased with r providing worse estimates for $r = 0.2$ compared with $r = 0.1$. For example, for $r = 0.2$ and using the R-M procedure with Ripley's correction, we get significant bias compared with the

unbiased and efficient estimates obtained under the same conditions but with $r = 0.1$. Apart from this, all properties analysed above are also true for different interaction radii.

The Strauss process with $\theta > 1$ is not a good model for applications. It may result in simulation difficulties such as sensitivity to edge-conditions, poor mixing, etc (Gates and Westcott, 1986). Moreover, the spatial birth-and-death approach might not be the optimal choice. For the well-defined case $\theta < 1$, exact simulation of the Strauss process is possible using the Propp-Wilson algorithm (Moller, 1998; Kendall and Moller, 1999). Concerning edge-corrections, another possibility is to apply conditional simulation: simulate a point pattern into some guard area using the model conditional on the observed point pattern and then apply the guard area events in estimation.

4. CONCLUSIONS

The conclusions, taking into account the results of our simulation study, are the following:

1. Stochastic approximations generally provide better results, particularly for medium-to-large parameter values, than those based on the sparseness assumption which are not suited for estimation in cases of strong interaction. For small parameter values and small interaction radius ($r = 0.1$), the Ogata-Tanemura approximation exhibits very good results.

2. For small interaction radius and using stochastic approximations, Ripley's and periodic boundary condition provide unbiased and efficient estimates. This is not true when r increases.

3. Generally, periodic and Ripley's boundary condition exhibit better results than free boundary condition.

4. When r increases the biases and the standard errors increase for any method, particularly for the approximate maximum likelihood methods.

5. Finally, in cases of clustered processes we recommend to use stochastic approximations with Ripley's or toroidal boundary condition. In cases of strong regularity, we could also use approximations based on the sparseness assumption.

ACKNOWLEDGEMENTS

The referees are gratefully acknowledged for their helpful comments that have substantially improved an earlier version of the paper.

REFERENCES

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

Baddeley, A. and van Lieshout, M.N.M. (1993). «Stochastic geometry in high-level vision». In *Statistics and images*, K.V. Mardia and G.K. Kanji (Eds.), Vol. 1 of *Advances in Applied Statistics*, 231-256.

Besag, J.E. (1974). «Spatial interaction and the statistical analysis of lattice systems (with discussion)». *Journal of the Royal Statistical Society* B, 36, 192-236.

Cressie, N. (1993). *Statistics for Spatial Data.* New York: John Wiley and Sons, Revised edition.

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

Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns.* London: Academic press.

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

Diggle, P.J., Gates, D.J. and Stibbard, A. (1987). «A nonparametric estimator for pairwise interaction point processes». *Biometrika*, 74, 763-70.

Gates, D.J. and Westcott, M. (1986). «Clustering estimates for spatial point distributions with unstable potentials». *Annals of the Institute of Statistical Mathematics*, 38, 123-135.

Geman, S. and 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. and Moller, J. (1994). «Simulation and likelihood inference for spatial point processes». *Scandinavian Journal of Statistics*, 21, 359-373.

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

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

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

Kendall, W.S. and Moller, J. (1999). «Perfect Metropolis-Hastings simulation of locally stable point processes». Manuscript.

van Lieshout, M.N.M. and Baddeley, A. (1995). «Markov chain Monte Carlo methods for clustering of image features». In *Proceedings of the fifth international conference on image processing and its applications*, Vol. 410 of IEE Conference Publication, 241-245, London.

Mateu, J. and Montes, F. (1995). «Inferencia Bayesiana en procesos puntuales Markov». In *Proceedings of the V Spanish Conference of Biometry*, 65-68, Valencia.

Molina, R. and Ripley, B.D. (1989). «Using spatial models as priors in astronomical image analysis». *Journal of Applied Statistics*, 16, 193-206.

Moller, J. (1998). «Markov chain Monte Carlo and spatial point processes». In *Proceedings Seminaire Europeen de Statistique, "Stochastic geometry, likelihood and computation"*. Barndorff-Nielsen, Kendall and Van Lieshout (eds), Chapman and Hall.

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

Ogata, Y. and 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, 315-38.

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

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

Ripley, B.D. (1979). «Simulating spatial patterns: dependent samples from a multivariate density». *Applied Statistics*, 28, 109-12.

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

Ripley, B.D. (1987). *Stochastic Simulation*. New York: Wiley.

Ripley, B.D. (1988). *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University press.

Ripley, B.D. (1989). «Gibbsian interaction models». In *Spatial Statistics: past, present and future* (ed. D.A. Griffiths), New York: Image.

Ripley, B.D. and Kelly, F.P. (1977). «Markov point processes». *Journal of the London Mathematical Society*, 15, 188-192.

Robbins, H. and Monro, S. (1951). «A stochastic approximation method». *Annals of Mathematical Statistics*, 22, 400-407.

Stoyan, D., Kendall, W.S. and Mecke, J. (1995). *Stochastic Geometry and its Applications*. Berlin: Akademie-Verlag, 2nd Edition.

Strauss, D.J. (1975). «A model for clustering». *Biometrika*, 62, 467-75.

Strauss, D.J. (1986). «On a general class of models for interaction». *SIAM Review*, 28, 513-527.

Takacs, R. (1986). «Estimator for the pair-potential of a Gibbsian point process». *Math. Operationsf. Statist. Ser. Statist.*, 17, 429-33.

QÜESTIIÓ, vol. 24, 1, p. 27-51, 2000

SPATIAL STRUCTURE ANALYSIS USING PLANAR INDICES*

J.M. ALBERT*
J. MATEU**
J.C. PERNÍAS*
Universitat Jaume I

Spatial planar indices have become a useful tool to analyze patterns of points. Despite that, no simulation study has been reported in literature in order to analyze the behaviour of these quantities under different pattern structures. We present here an extensive Monte Carlo simulation study focused on two important indices: the Index of Dispersion and the Index of Cluster Size, usually used to detect lack of homogeneity in a spatial point model. Finally, an application is also presented.

Keywords: Index of cluster size, index of dispersion, point processes, spatial structure

AMS Classification: 62M30, 60G55

* The authors wish to thank financial support by the Spanish Ministry of Education (CICYT: SEC96-1435-C03-03).

* Department of Economics. Universitat Jaume I. Campus Riu Sec. 12071 Castelló. Spain.
** Department of Mathematics. Universitat Jaume I. Campus Riu Sec. 12071 Castelló. Spain (mateu@mat.uji.es)

- Received June 1999.
- Accepted December 1999.

1. INTRODUCTION

A spatial point pattern is a collection of data $\{(x_i, y_i) \;\; i = 1, \ldots, n\}$ consisting of n locations in an essentially planar region. Examples include the locations of cell nuclei in a microscopic tissue section, trees in a forest, or cases of disease in a geographical region. A fundamental assumption in the analysis of such data is that they can usefully be regarded as a partial realisation of a stochastic point process (Cox & Isham, 1980). Many systems of individuals can also be described by attaching to the locations measurable quantities like, say, diameter of a tree. This last case is an example of marked points patterns.

There are many contexts in which the use of spatial point patterns can be very interesting, for example, in the analysis of the human population's spatial distribution, since they can give us excellent information from both the demographic and the economic points of view.

The interest in a pattern of points is found in that, with an appropriate choice of scale, even huge objects may be best represented by a point. Given suitable scales, the actual physical sizes of objects that may be represented that way are unbounded. On one extreme, microscopes are required, on the other extreme it is telescopes that are needed. The range of disciplines (pathology, geology, marine biology, zoology, physical and human geography, astronomy, economy,...) dealing with similar phenomena reflects the applicability of these techniques.

The concept of *complete spatial randomness* (CSR) is fundamental to the quantitative description of a spatial pattern. A formal definition of CSR is that the events in the region of observation A constitute a partial realisation of a homogeneous, planar Poisson process (Diggle, 1983). This process incorporates a single parameter, λ, the intensity, or mean number of events per unit area. The actual number of events in A, n say, is an observation from a Poisson distribution with mean $\lambda|A|$, where $|A|$ denotes the area of the region A. If we consider n as fixed, we arrive at the following definition of CSR: (1) each of the n events is equally likely to occur at any point within A; (2) the n events are located independently of each other. Our interest in CSR is that it represents an idealized standard which, if strictly unattainable in practice, may nevertheless be tenable as a convenient first approximation. Most of the analysis begin with a test of CSR, and there are several good reasons for this. Firstly, a pattern for which CSR is not rejected scarcely deserves any further formal statistical analysis. Secondly, tests are used as a means of exploring a set of data, rather than because rejection of CSR has an intrinsic interest. Thirdly, CSR acts as a dividing hypothesis to distinguish between patterns which are broadly classifiable as «regular» or «aggregated» (Figure 1).

A question of immediate interest is the following: Is it reasonable to expect a pattern of real data events to display randomness? Naturally, the answer depends on what these

events represent. If, for example, the point pattern is defined as locations of cities, in order to evaluate the benchmark hypothesis of CSR, two contradictory forces are liable to have an effect on the observed cities locations. On one hand, the competition between neighbouring cities is likely to result in some thinning out of close neighbours, so that the pattern becomes rather more regular in appearance. On the other hand, variations in the local geography will result in some regions being more favourable for growth than are other regions. This will result in an apparent patchiness (or clustering) in the cities distribution. These two effects, and others, may well be sufficiently counterbalancing each other, so that the distribution of cities may yet retain the appearance of randomness.

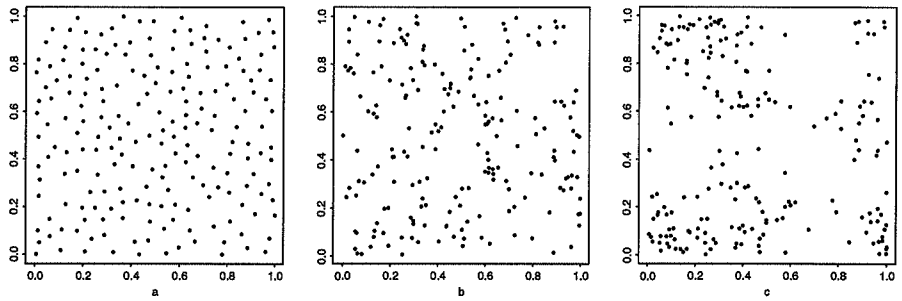


Figure 1. Examples of point patterns in the unit square showing three different spatial structures: *Left*, regular pattern; *Middle*, random pattern; *Right*, aggregated pattern.

In view of the remarkable diversity of mechanisms which may lead to an apparently random pattern, it must be stressed that a random distribution implies that the pattern has no discernible order and that its cause is undeterminable.

In literature, two major approaches have been suggested for the analysis of pattern. One involves measures of physical distances between points (Diggle 1983; Ripley 1981, 1988; Cressie 1993, Stoyan et al. 1995), the other involves analysis of the variation in the number of points in selected sub-areas of the region under study (Diggle, 1983; Upton & Fingleton, 1994). In this paper we concentrate on the second approach. All the counting methods rely on the use of quadrats. Several indices, like the *index of dispersion* and the *index of cluster size*, were proposed for scattered and contiguous quadrats. Generally, their behaviour under CSR is quite well known, though little is analyzed under departure of randomness. What is known is that the power of such indices depends in an unpredictable way on the size and shape of the individuals quadrats (Diggle, 1979, 1983; Perry & Mead, 1979; Stiteler & Patil, 1971).

This topic owes its generality as it is a general way of proceeding in detection of point pattern structures. Examples of this generality are the wide range of applications, the

majority refering to contiguous quadrats. We can find applications in environmental sciences such as pattern analysis of perennial shrubs (Gulmon & Mooney, 1977), of herbs in the savanna (Hopkins, 1965) or detection of pattern in Lansing Woods trees (Diggle, 1983); in economics such as analysis of the distribution of houses (Moellering & Tobler, 1972), etc.

The goal of this paper is to analyze exhaustively the behaviour of quadrat-based indices to detect spatial structures. Particularly, we present an extensive simulation study to compare the performance of two of these indices under clear departures of CSR. In addition we present our own developed software, S.P.P.A. (1997), which can be used in this spatial context.

The plan of the paper is as follows. Section 2 presents the spatial indices. Section 3 is devoted to the simulation study presenting the whole set of results. Finally, Section 4 develops the analysis of a real application.

2. SPATIAL INDICES

Throughout this section, given an observed pattern, we attempt to deduce the nature of the process that gave rise to that pattern. Quadrat sampling involves collecting counts of events in subsets of the study region. Traditionally, these subsets are rectangular (hence the name of quadrats), although any shape is possible. Quadrats may be placed either randomly or layed out contiguously in the region.

This is very easy to implement as it is only needed to position quadrats randomly in the study region (Figure 2) and count the numbers of events that fall in each quadrat. It is true that scattered quadrats counts provide some limited information about the nature of a point pattern and that is our present concern. Counts from scattered quadrats are, at best, a crude indicator of pattern because they take into account neither the relative positions of the points within the quadrats nor the relative positions of the quadrats themselves.

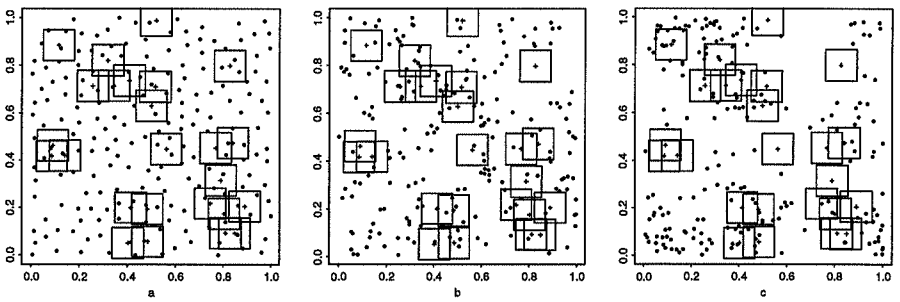


Figure 2. Equi-sized scattered quadrats superimposed over the point patterns.

The counts from equi-sized scattered quadrats (Figure 2) over a Poisson pattern will be observations from a Poisson distribution with parameter equal to the product of the intensity of the events per unit area and the area of the quadrats. A useful characteristic of the Poisson distribution is that its parameter is equal both to the mean and the variance of the distribution. If the pattern is more regular than a Poisson one, then the quadrat counts will be more uniform in size and will therefore have a relatively small variance (when compared with the size of the mean). On the other hand, if there are clusters, then some quadrats will have large counts so that the variance of the counts will be relatively large.

The analysis of grids of contiguous quadrats takes advantage of information on quadrat locations. A grid of contiguous quadrats is a spatial lattice (Figure 3). The advantage of such a grid is that neighbouring quadrats can be combined so that we may obtain information about quadrats of more than one size.

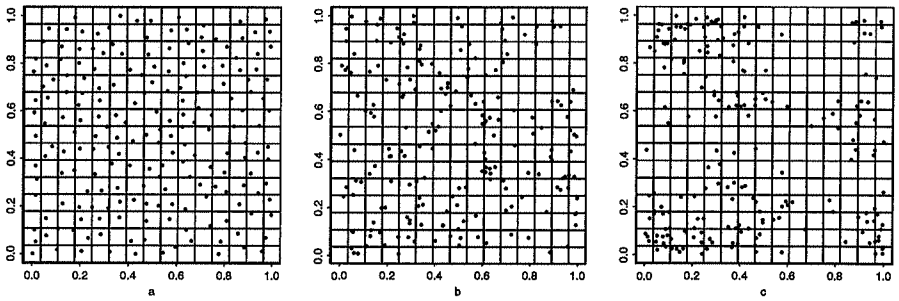


Figure 3. Spatial lattice of 15×15 contiguous quadrats superimposed over the point patterns.

A natural test of a Poisson distribution is therefore provided by examining the value of the ratio sample variance/sample mean or *index of dispersion* (ID). If we denote the observed counts in nq quadrats by $x_1, x_2, \ldots, x_{nq}$, then these counts have mean $\bar{x} = \sum x_i / nq$ and variance $s^2 = \dfrac{\sum (x_i - \bar{x})^2}{(nq-1)}$, where the summations are over the values of i from 1 to nq. Hoel (1943) showed that

$$ID = \frac{(nq-1)s^2}{\bar{x}} \tag{1}$$

has an approximate χ^2_{nq-1} distribution under CSR. This approximation is reasonable provided that $nq > 6$ and $\bar{x} > 1$. Values of $ID > \chi^2_{nq-1}; (1-\alpha)$ are indicative of clustering and regularity is given by values of $ID < \chi^2_{nq-1}; \alpha$, where α stands for the usual significance level.

Perry & Mead (1979) examined the behaviour of the index of dispersion test and concluded both, that it is remarkably sensitive at detecting a lack of homogeneity within a

point pattern, and that the behaviour of the test was principally dependent upon the size of the mean of the quadrat counts: the larger the mean the more likely was the ID test to recognize any heterogeneity presented in the pattern.

A number of other indices have been suggested, principally for situations in which clusters may be present. All are sensitive to changes in quadrat sizes. David and Moore (1954) suggested that the quantity

$$ICS = \left(\frac{s^2}{\bar{x}}\right) - 1 \tag{2}$$

would provide an approximate index of clumping or contagiousness. This test is called *index of cluster size.* For a Poisson pattern, ICS has mean 0 and is independent of the quadrat size. An interpretation of a positive value for ICS is as the number of other events intimately associated with a randomly chosen event. A negative value for ICS indicates some regularity in the positioning of the events.

Theoretically, if we have the observed iid counts $x_1, x_2, \ldots, x_{nq}$ in nq quadrats following any distribution, then ICS satisfies (Serfling, 1980)

$$ICS \sim N\left(\frac{\sigma^2}{\mu} - 1, \frac{1}{n}\left(\frac{\sigma^6}{\mu^4} + \frac{\mu_4 - \sigma^4}{\mu^2} - \frac{2\mu_3\sigma^2}{\mu^3}\right)\right) \tag{3}$$

where $E(x_i) = \mu$, $Var(x_i) = \sigma^2$, $E(\{x_i - \mu\}^3) = \mu_3$ and $E(\{x_i - \mu\}^4) = \mu_4$.

Particularly, if the counts come from a Poisson distribution, i.e., if we have a CSR point pattern, then for large nq we have that

$$ICS \sim N(0, 2/n) \tag{4}$$

However, for small nq a better approximation can be found for ICS which consists of

$$ICS \sim \frac{\chi^2_{nq-1}}{nq-1} - 1 \tag{5}$$

Note that for large values of nq, we can approximate $\chi^2_{nq-1} \sim N(nq-1, 2(nq-1))$.

A number of other indices have been suggested, principally for situations where it is thought that clusters may be present. All are sensitive to changes in quadrat sizes. However, for quadrat count data, ID appears to have no serious rivals as a test statistic (Diggle, 1983). Cormack (1979) notes that alternative indices proposed by Morisita (1959) and Lloyd (1967) need to be converted to ID in order to test CSR.

It was first Greig-Smith (1952) who proposed contiguous quadrats to analyze data presented as counts by means of the index of dispersion. He suggested that a 16×16

grid of quadrats should be used. Then, Diggle (1983) used the same index of dispersion within a regular $k \times k$ grid of contiguous square sub-regions of equal area to test CSR. Particularly, significantly small values of ID indicate a tendency towards a regular spatial distribution of events, whereas significantly large values indicate aggregation.

3. SIMULATION STUDY

We present here an extensive simulation study of the two selected count-based indices, ID and ICS based on contiguous quadrats superimposed over the selected region. To consider any possible alternatives to the random (CSR) pattern, we also consider regular and aggregated point patterns. CSR patterns are generated according to a homogeneous Poisson process following the definition given in the introduction.

Regular or inhibitory point patterns are defined through a hard-core distance, δ, using a sequential inhibition process based on the following facts: a) x_1 is uniformly distributed in the region A; b) Given $\{x_j, j = 1,\ldots,i-1\}$, x_j is uniformly distributed on the intersection of A with $\{y : d(y,x_j) \geq \delta,\ j = 1,\ldots,i-1\}$.

Aggregated patterns are defined in terms of Poisson clustered processes as defined by Neyman & Scott (1958). These processes incorporate an explicit form of spatial clustering, and as such provide a more satisfactory basis for the modelling of aggregated spatial point patterns. They are defined based on the following three postulates: a) Parent events form a Poisson process with a fixed intensity; b) Each parent produces a random number of offsprings, realized independently and identically for each parent; c) The positions of the offsprings relative to their parents are independently and identically distributed according to a bivariate normal density function.

The S.P.P.A. computer software has been developed to generate planar coordinates of points in a fixed region with a particular spatial structure and then, among other things, calculate quadrat-based counts for a given quadrat size.

3.1. Design of the simulation study

The process of simulations has been carried out for three qualitatively different spatial structures: randomness, clustering and regularity. We have used several total number of points per pattern: for random and regular structures, $n = 400, 1000$ and for clustered structures, $n = 400, 1000, 2500$. Clustered pattern simulation is done with several number of fathers, ranging from 1 to 4. Two hard-core radius for inhibitory patterns have been used, $ir = 0.04295$ (with $n = 400$), and $ir = 0.02688$ (with $n = 1000$). For each combination of selected quantities, we simulated $r = 2000$ realizations in the unit square, $(0,1) \times (0,1)$.

For each pattern, and to apply the contiguous quadrat system, we have used several grid sizes shown in Table 1. Each line of Table 1 gives us information about the type of simulated pattern (*Pattern*), the number of points in each pattern (*Points*), the number of replications for each experiment (*Repl.*), the inhibition value used in the regular pattern simulations (*Inhib.*), and grid order values. All grid orders are magnitudes to the square, i.e., $(10 \times 10, 15 \times 15, 20 \times 20$, etc), but they are shown simplified (10, 15, 20, etc).

Table 1. Design of the simulation study.

Pattern	**Points**	**Repl.**	**Inhib.**	**Grid order**
Random	400	2000		2,3, ... ,6,8,10,11,13, ... ,19,20,21,23,25,30, ... ,100
Random	1000	2000		2,3, ... ,6,8,10,11,13, ... ,19,20,21,23,25,30,40, ... ,100
Regular	400	2000	.04295	5,6,8,10,11,13, ... ,19,20,21,23,25
Regular	1000	2000	.02688	5,6,8,10,11,13, ... ,19,20,21,23,25,30,35,40
1 Cluster	400	2000		4,5,6,8,10,11,13, ... ,19,20,21,23,25,30,35,40
1 Cluster	1000	2000		6,8,10,11,13, ... ,19, ... ,30,35,40,50, ... ,100,120, ... ,200
1 Cluster	2500	2000		10,20,30, ... ,100,120,140, ... ,200
2 Cluster	1000	2000		10,20,30, ... ,100,120,140, ... ,200
2 Cluster	2500	2000		10,20,30, ... ,100,120,140, ... ,200
3 Cluster	1000	2000		10,20,30, ... ,100,120,140, ... ,200
3 Cluster	2500	2000		10,20,30, ... ,100,120,140, ... ,200
4 Cluster	1000	2000		10,20,30, ... ,100,120,140, ... ,200
4 Cluster	2500	2000		10,20,30, ... ,100,120,140, ... ,200

Each combination of pattern, replicate, grid size, inhibition radius and number of clusters yielded 2000 estimates of both indices which are summarised by box-plots and tables.

Particularly interesting is to show the goodness of fit of the approximation given in (5) for CSR patterns with a small number of nq quadrats. As we are interested in analyzing values belonging to the tails of the chi-square distribution to safely contrast overdispersion or underdispersion, we focus upon the following percentiles, $\alpha = 1\%, 5\%, 10\%$ and $\alpha = 99\%, 95\%, 90\%$. For each percentile the *relative error* is calculated as

$$re = \frac{100|\alpha - p|}{\alpha} \tag{6}$$

where p is the observed proportion of simulated values of ICS under the theoretical value obtained through (5) for $\alpha = 0.01, 0.05$ and 0.10 and over the theoretical value from (5) for $\alpha = 0.99, 0.95$ and 0.90.

3.2. Analysis of ID index

Table 2 shows the summarized results of the simulations with the ID index. Each entry represents the mean value of 2000 simulations. The following results are observed.

Table 2. Means of ID index. Each column entry indicates grid size and pattern structure (Ran=random, Reg=regular, c=cluster) followed by the number of points in the unit square. Blank boxes indicate that the corresponding simulation has not been analyzed.

Grid	Ran400	Ran1000	Reg400	Reg1000	1c400	1c1000	1c2500	2c1000	2c2500	3c1000	3c2500	4c1000	4c2500
2	3	3											
3	8	8											
4	15	15			257								
5	24	24	4	3	274								
6	35	35	6	5	290	677							
8	62	63	12	10	323	716							
10	99	99	21	16	363	759	1752	1390	3349	1534	3699	1031	
11	119	120	27	21	384	782							
13	167	168	44	32	433	832							
15	223	224	65	46	490	890							
17	287	288	86	62	553	955							
19	359	359	109	85	626	1028							
20	398	399	121	100	665	1067	2068	1768	3833	1981	4369	1457	3065
21	438	440	136	116	707	1107							
23	528	527	178	150	796	1197							
25	622	624	246	183	891	1293							
30	898	899		270	1167	1569	2576	2281	4369	2508	4935	1988	3628
35	1224	1224		394	1490	1894							
40	1598	1597		648	1867	2271	3275	2987	5078	3220	5658	2689	4349
45	2023												
50	2497	2499				3173	4180	3890	5985	4122	6567	3594	5258
55	3022												
60	3597	3602				4273	5273	4990	7087	5234	7685	4697	6366
65	4221												
70	4898	4902				5572	6579	6297	8391	6530	8985	5999	7664
75	5626												
80	6397	6398				7067	8077	7791	9889	8034	10481	7497	9174
85	7224												
90	8097	8102				8771	9779	9496	11593	9728	12186	9202	10875
95	9022												
100	9996	10005				10675	11682	11392	13497	11629	14096	11107	12769
120						15071	16077	15797	17895	16035	18491	15509	17177
140						20278	21281	20998	23095	21234	23704	20711	22377
160						26269	27287	26989	29095	27235	29685	26708	28375
180						33072	34079	33806	35891	34033	36502	33503	35173
200						40663	41696	41392	43501	41630	44067	41121	42772

3.2.1. CSR point pattern distribution

- In patterns with 400 points, ID follows a perfect χ^2_{nq-1} distribution (pval= 0.98) for grid sizes of 2×2 up to 35×35. For bigger grid sizes, ID still follows a chi-square distribution (pval= 0.39) but with a slightly smaller mean (Figure 4). Then, theory is generally satisfied except for very large grid sizes.

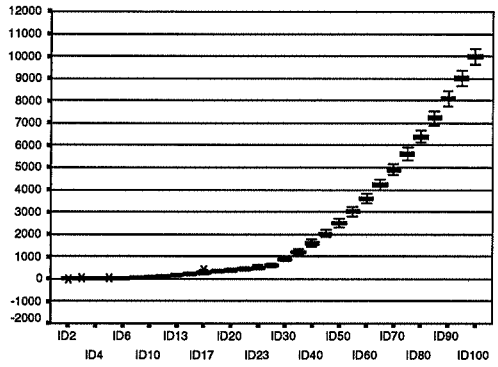


Figure 4. Boxplots of ID values under CSR pattern structure with 400 points.

- Again, in patterns with 1000 points, ID follows a perfect χ^2_{nq-1} distribution (pval= 0.97) for grid sizes from 2×2 up to 70×70. This distribution is still observed for bigger grid sizes (pval= 0.57) but with some little changes in the mean value (Figure 5). Generally, there seems to be no clear distinction in the ID performance between CSR patterns with different number of points.

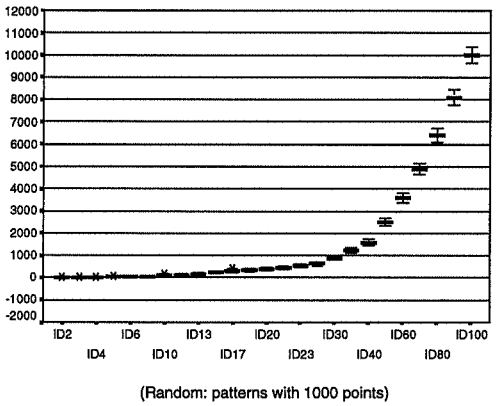


Figure 5. Boxplots of ID values under CSR pattern structure with 1000 points.

3.2.2. Regular point pattern distribution

- For patterns with 400 points, we observe in any case that $ID < \chi^2_{nq-1}; \alpha = 0.01$ (Figure 6). The magnitude of the ID index is increased with the grid size, though they are significantly smaller (pval= 0.01) compared to the corresponding ID value under CSR structure. The chi-square distribution is no longer satisfied.

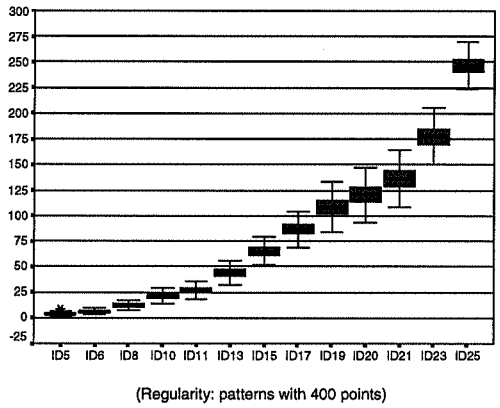


Figure 6. Boxplots of ID values under regular pattern structure with 400 points and inhibition radius $ir = 0.04295$

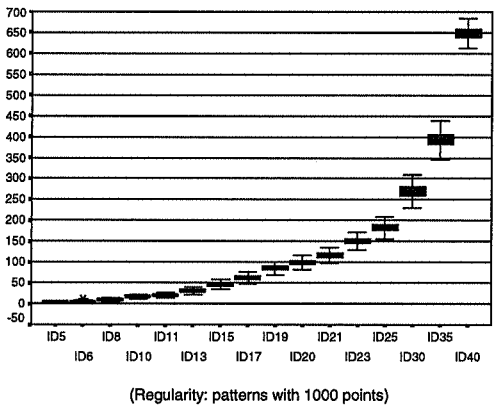


Figure 7. Boxplots of ID values under regular pattern structure with 1000 points and inhibition radius $ir = 0.02688$.

- For patterns with 1000 points, again it is generally observed that $ID < \chi^2_{nq-1}; \alpha = 0.01$ (Figure 7). ID values increase with the grid size though still significantly

smaller (pval= 0.01) than the corresponding values under CSR condition. It is also observed that for equal grid size, ID values with 1000 points are smaller than ID values for patterns with 400 points. This is clearly due to the fact that we are using different values for the inhibition radius and this is detected by the ID index.

3.2.3. Cluster point pattern distribution

In this section we comment the results for the four different cluster pattern structures, though, for shortness, we only show the corresponding boxplots for 1 cluster patterns (see Figures 8, 9 and 10).

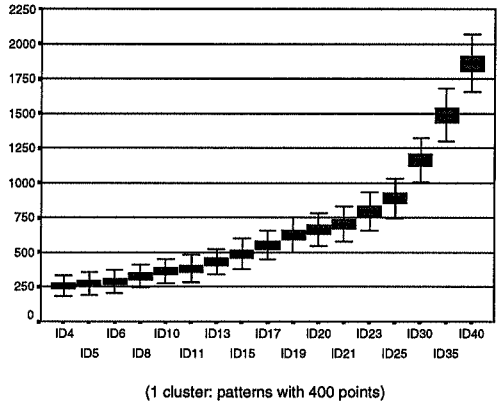


Figure 8. Boxplots of ID values under 1 cluster pattern structure with 400 points.

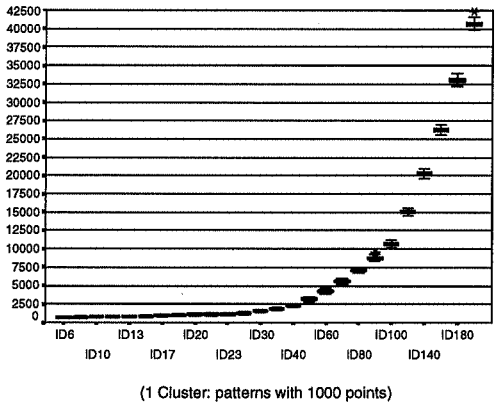


Figure 9. Boxplots of ID values under 1 cluster pattern structure with 1000 points.

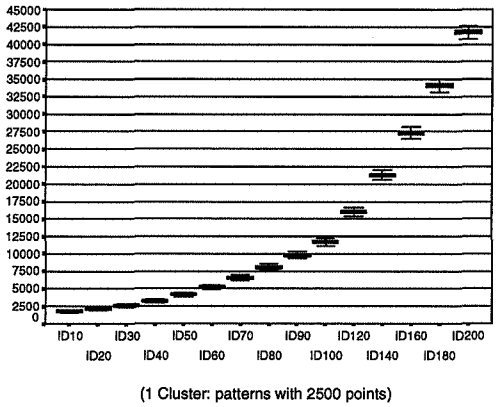


Figure 10. Boxplots of ID values under 1 cluster pattern structure with 2500 points.

- In any case, we observe that $ID > \chi^2_{nq-1}; (1-\alpha = 0.99)$ and its value increases with the grid size. ID index is a very sensitive index to detect clustering in that not only the ratio standard deviation/mean decreases as the grid size increases, but also ID increases with the number of clusters, except for the four cluster case. In fact, the pattern with four clusters may appear to be less aggregated that those patterns with a smaller number of clusters. Then, ID index is good in detecting scales of aggregation in point patterns.
- Within the same scale of aggregation, ID values show less variability in those patterns with a larger number of points.
- The statistical distribution of ID under clustering depends on the grid size and the number of points. The chi-squared distribution is always rejected though ID values seem to follow a gaussian distribution in several grid sizes. This behaviour is independent of the scale of aggregation.
- Note that grid size depends on the number of points in the pattern. Using a large grid size so that most of our cells or quadrats have zero counts causes the results to be biased. This comment is also clearly true for any other kind of pattern structure.

3.3. Analysis of ICS index

Table 3 presents the summarized results of the corresponding simulations with the ICS index. Each entry in the table represents the mean value of 2000 simulations. The following results are obtained.

Table 3. Means of ICS index. Each column entry indicates grid size and pattern structure (Ran=random, Reg=regular, c=cluster) followed by the number of points in the unit square. Blank boxes indicate that the corresponding simulation has not been analyzed.

Grid	Ran400	Ran1000	Reg400	Reg1000	1c400	1c1000	1c2500	2c1000	2c2500	3c1000	3c2500	4c1000	4c2500
2	-0,01509	0,0250											
3	-0,02333	-0,0062											
4	0,00239	-0,0003			16,11								
5	-0,00448	0,0005	-0,843	-0,867	10,44								
6	-0,00608	-0,0073	-0,832	-0,862	7,30	18,34							
8	-0,01696	0,0012	-0,807	-0,848	4,13	10,37							
10	0,00039	0,0030	-0,788	-0,835	2,66	6,67	16,697	13,0392	32,8244	14,493	36,361	9,4166	
11	-0,00718	0,0005	-0,778	-0,827	2,20	5,52							
13	-0,00426	0,0021	-0,739	-0,812	1,58	3,95							
15	-0,00281	0,0002	-0,709	-0,796	1,19	2,97							
17	-0,00192	-0,0011	-0,700	-0,784	0,92	2,32							
19	-0,00201	-0,0026	-0,698	-0,763	0,74	1,86							
20	-0,00217	0,0004	-0,696	-0,749	0,67	1,67	4,1821	3,4315	8,60691	3,9654	9,9505	2,6509	6,68246
21	-0,00439	-0,0002	-0,690	-0,736	0,61	1,52							
23	-0,00092	-0,0011	-0,664	-0,717	0,51	1,27							
25	-0,00273	0,0001	-0,605	-0,707	0,43	1,07							
30	-0,00069	-0,0002		-0,700	0,30	0,75	1,8657	1,5372	3,85966	1,7893	4,4896	1,2118	3,03538
35	-0,00038	-0,0001		-0,678	0,22	0,55							
40	-0,00078	-0,0010		-0,595	0,17	0,42	1,0482	0,8679	2,17584	1,0139	2,5382	0,6819	1,71952
45	-0,00033												
50	-0,0009	0,0001				0,270	0,6725	0,5567	1,39478	0,6494	1,6277	0,4381	1,10421
55	-0,00052												
60	-0,00045	0,0009				0,187	0,4652	0,3865	0,96915	0,4542	1,1353	0,3051	0,76895
65	-0,0006												
70	-0,00011	0,0005				0,137	0,343	0,2854	0,71275	0,3329	0,834	0,2245	0,56445
75	0,00028												
80	-0,00028	-0,0001				0,105	0,2623	0,2176	0,54534	0,2556	0,6379	0,1717	0,43366
85	-8,9E-06												
90	-0,00026	0,0003				0,083	0,2074	0,1725	0,43137	0,2012	0,5047	0,1362	0,34277
95	-0,00018												
100	-0,00032	0,0006				0,068	0,1683	0,1394	0,34981	0,163	0,4097	0,1108	0,27705
120						0,0467	0,1165	0,09711	0,2428	0,1136	0,2842	0,0771	0,19292
140						0,0346	0,0858	0,07139	0,17839	0,0834	0,2095	0,0567	0,14175
160						0,0262	0,0659	0,0543	0,13656	0,0639	0,1596	0,0433	0,10846
180						0,0208	0,0519	0,04342	0,10779	0,0504	0,1267	0,0341	0,08562
200						0,0166	0,0424	0,03483	0,08755	0,0408	0,1017	0,0281	0,06933

3.3.1. CSR point pattern distribution

- For CSR patterns with 400 points, ICS values oscillate around zero. However, for small to medium grid sizes, there exist quite large standard deviations and also some outliers are present (Figure 11). Standard deviations rapidly decrease when grid size increases, which is natural in terms of the theoretical expression of the variance in formula (4) for large samples. The results show that for small samples, the chi-square distribution given by (5) is satisfied and also ICS values follow the gaussian distribution (4) when large samples are considered.

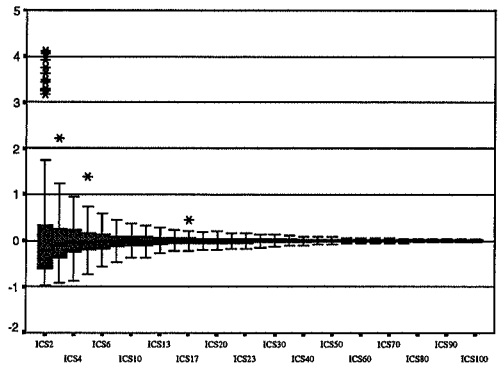


Figure 11. Boxplots of ICS values under CSR pattern structure with 400 points.

- Generally, the same results can be found for CSR patterns with 1000 points. The difference is that the larger the number of points, we need smaller grid sizes to get the same or smaller standard deviations (Figure 12).

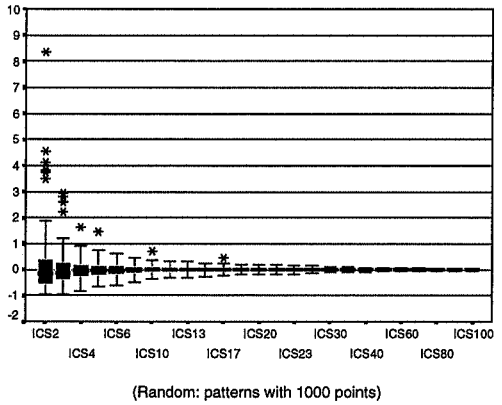


Figure 12. Boxplots of ICS values under CSR pattern structure with 1000 points.

- Concerning the evaluation of the goodness of fit of the approximation given in (5) for CSR patterns with a small number of nq quadrats in terms of the relative error, we calculated the relative errors following (6) for six percentiles, $\alpha = 0.01, 0.05, 0.10, 0.90, 0.95, 0.99$, and for several grid sizes up to $nq = 100$ (10×10). The results are shown in Table 4. In general, and independently of the number of points per pattern, the relative errors are below 20%, which can be considered as small enough to trust on the results. However, we find relative errors bigger that 20% at the very end of the tails, i.e., for $\alpha = 0.01$ and 0.99,

which means that under very severe conditions the results of a CSR contrast might be misleading.

Table 4. Relative errors of the ICS index under CSR pattern structure for several percentiles and for patterns with $n = 400$ points and with $n = 1000$ points (in parenthesis) in the unit square.

		grid size						
		2 × 2	3 × 3	4 × 4	5 × 5	6 × 6	8 × 8	10 × 10
α	0.99	5 (30)	40 (10)	20 (20)	25 (5)	25 (10)	15 (15)	20 (0)
	0.95	5 (3)	12 (3)	8 (1)	4 (6)	4 (19)	14 (0)	7 (12)
	0.90	1.5 (3)	8.5 (3)	3 (3)	5.5 (0)	10 (16)	19 (1.5)	5 (4.5)
	0.10	1.5 (15)	16 (1.5)	4.5 (3.5)	5.5 (9)	9.5 (2)	16.5 (4.5)	5.5 (15)
	0.05	6 (30)	22 (1)	12 (1)	1 (9)	16 (10)	1 (4)	15 (10)
	0.01	40 (95)	5 (40)	30 (15)	25 (0)	0 (45)	10 (55)	25 (15)

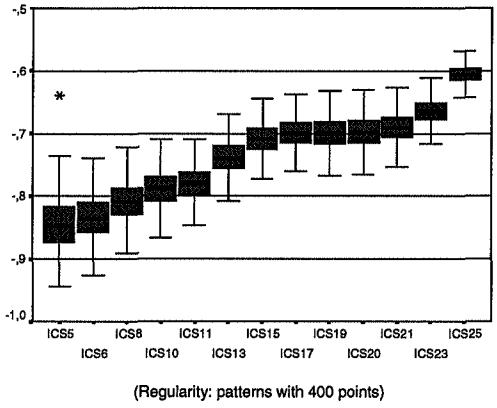


Figure 13. Boxplots of ICS values under regular pattern structure with 400 points and inhibition radius $ir = 0.04295$.

3.3.2. Regular point pattern distribution

- We find that ICS values for regular patterns are significantly smaller than those under CSR (pval= 0.01), showing always clearly negative values. However, there seems to be an increasing tendency towards zero as the grid size increases (Figures 13 and 14). This shows that there should be an optimum grid size, for example

around 20×20 in patterns with 400 points or 30×30 when the number of points is 1000. If we surpass them, then ICS values are biased. The chi-squared distribution is no longer satisfied in favour of the Gaussian distribution which is only observed for those grid sizes near the optimum (Figures 13 and 14).

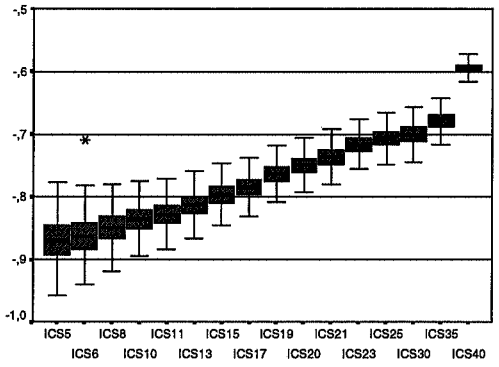


Figure 14. Boxplots of ICS values under regular pattern structure with 1000 points and inhibition radius $ir = 0.02688$.

3.3.3. Cluster point pattern distribution

Again we comment the results for all cluster structures though, for shortness, we only show those boxplots corresponding to one cluster with 400, 1000 and 2500 points (Figures 15, 16 and 17).

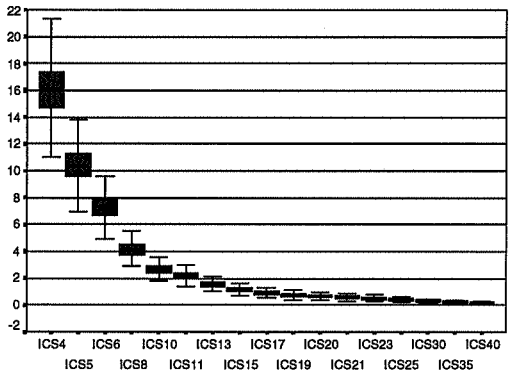


Figure 15. Boxplots of ICS values under 1 cluster pattern structure with 400 points.

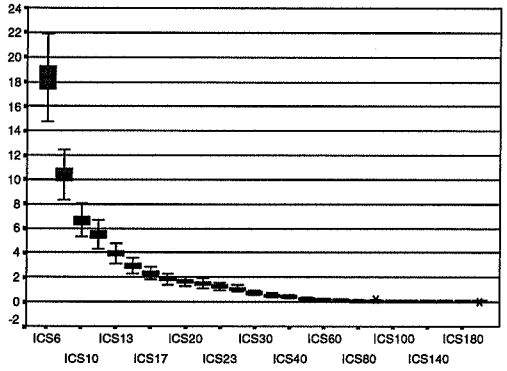


Figure 16. Boxplots of ICS values under 1 cluster pattern structure with 1000 points.

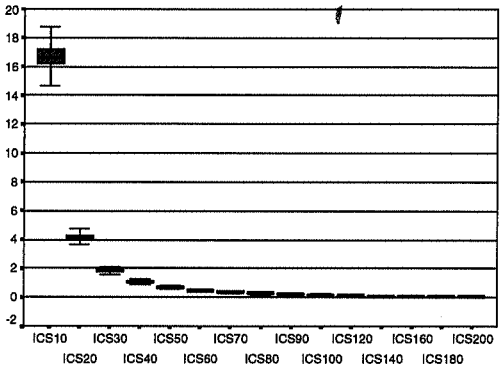


Figure 17. Boxplots of ICS values under 1 cluster pattern structure with 2500 points.

Generally speaking, under cluster spatial structures, ICS takes positive values, significantly different from zero (pval= 0.01, which is characteristic of ICS under CSR patterns). However, ICS tends to decrease to zero as the grid size increases. The relative standard deviation of this indicator (standard deviation/mean) is increased when the grid size increases.

We have also found that ICS index is very sensitive to both, the scale of aggregation (given by the number of clusters) and the number of points. In fact, the mean values of the ICS index for patterns with 1000 points can be obtained by multiplying those obtained with 400 points times 2.5. This is equally true if we compare patterns with 400 points with those patterns with 2500 points. In this case we have to multiply times 6.25 the corresponding values of ICS.

- The distribution of ICS under clustering is somewhat complicated and depends on the number of points and the degree of aggregation. For example, when we consider patterns with one cluster, gaussianity is satisfied only within a few grid sizes (from 4×4 to 13×13) for patterns with 400 points. The number of grid sizes in which the gaussian assumption is satisfied, increases when the number of points increases (from 6×6 to 60×60, for n=1000, and in most of grid sizes for n=2500).
- For those patterns with 2 clusters, values of ICS are more indicative of spatial clustering compared to patterns with only one cluster. The gaussian behaviour of ICS is similar as commented above.
- In general, ICS values clearly indicate the degree of clustering by taking larger values. The gaussian distribution is comfortable reached when the number of points is large.

3.4. General conclusions

After analyzing step by step the results obtained in the simulation study, the following general results can be outlined.

- The results confirm that the *index of dispersion* follows an approximate χ^2_{nq-1} distribution under CSR when $nq > 6$ and $\bar{x} > 1$. Moreover, we can enlarged this condition to $nq < 6$ (grid 2×2) and $\bar{x} < 1$. Equally, our results show that values of $ID > \chi^2_{nq-1};(1-\alpha)$ are indicative of clustering and regularity is given by values of $ID < \chi^2_{nq-1};\alpha$ (see Figures 18 and 19). Therefore, we confirm by simulation Hoel's (Hoel, 1943) and Diggle's results (Diggle, 1983).

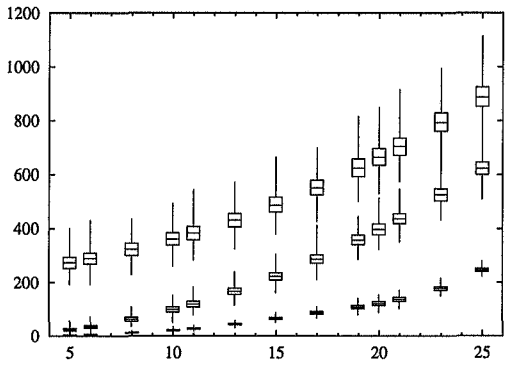


Figure 18. Comparison of boxplots for the ID index under the three pattern structures (1 cluster, CSR and regularity) for patterns with n=400 points in the unit square.

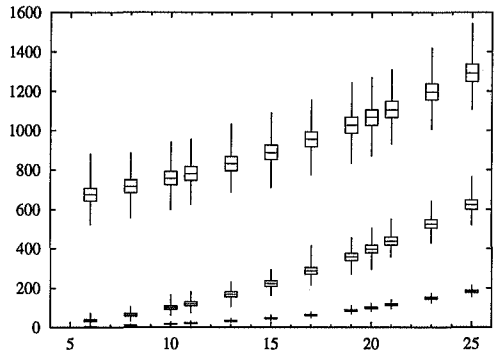


Figure 19. Comparison of boxplots for the ID index under the three pattern structures (1 cluster, CSR and regularity) for patterns with n=1000 points in the unit square.

- Similarly, we have also confirmed Douglas theory adapted for contiguous quadrats (Douglas, 1975) for the *index of cluster size*: for a Poisson pattern, ICS has mean 0 and is independent of the quadrat size; ICS has a positive mean for clustered patterns and a negative mean for regular patterns (see Figures 20 and 21).
- We have checked the distribution of both indices under alternative spatial patterns. ID index is no longer chi-squared distributed if CSR is rejected in favour to regular or aggregated patterns. However, under these alternatives, it seems that ID follows a gaussian distribution, particularly when the number of points is large enough (at least 1000 points). On the other hand, ICS index follows a chi-squared distribution under small grid sizes and a gaussian distribution for bigger grid sizes as theory states.

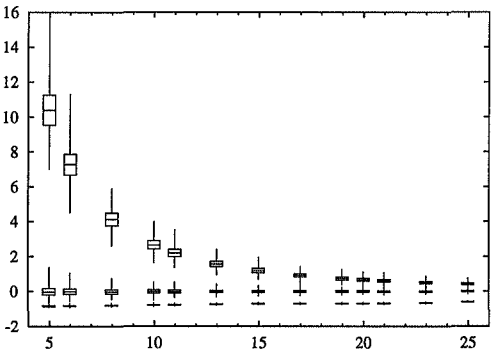


Figure 20. Comparison of boxplots for the ICS index under the three pattern structures (1 cluster, CSR and regularity) for patterns with n=400 points in the unit square.

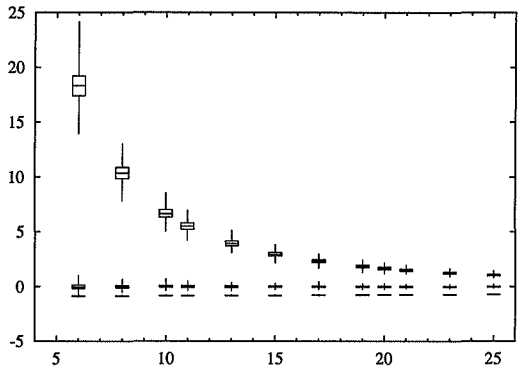


Figure 21. Comparison of boxplots for the ICS index under the three pattern structures (1 cluster, CSR and regularity) for patterns with n=1000 points in the unit square.

- Both indices are very sensitive at detecting not only the degree of clustering but also the number of points. The latter element is crucial to get lower standard deviation in the index values.
- To detect randomness, the larger the grid size, the better indication of CSR pattern (in terms of bias and standard deviation) we get from both indices. This is not generally true when detecting regularity or aggregation. There seems to be an optimum grid size, from which onwards the results are clearly biased, even confusing. The optimum grid size depends on the number of points. Generally speaking, and according to our simulations, the number of contiguous quadrats should not exceed the total number of points. A low grid order means that we may have insufficient information. But also a too large grid order means that we are introducing irrelevant information to our data structure, introducing bias to the estimates. This was not previously found in literature.

4. APPLICATION

It is known that information about the spatial distribution of the population may give us insights of interesting economic phenomena (Richardson, 1986; Hudson & Fowler, 1966; Lösch, 1954). This is due to the fact that human settlements have a history conditioned by economy which has been developed in specific geographical areas. Consequently, we present here a practical use of both, the ID and ICS indices, to describe qualitative and quantitatively the spatial structure of two important spanish peninsular provinces, Madrid and Barcelona as this analysis will provide economists and geographers with relevant information.

4.1. Data and results

The data set represents coordinates, longitude (eastings) and latitude (northings) of points, where each point defines the location of a 20,000 inhabitants crowd. The data set was transformed adequately to have planar coordinates. The origin of the coordinate axis is set up as the point of the meridian at 9^o-west longitude and the parallel at 36^o-north latitude. Let each measurement unit be equal to 1500 metres (approximately). Then, we assign p_i points to $i-th$ city as follows

$$p_i = \frac{h_i}{20000}, \tag{7}$$

where h_i denotes the number of inhabitants in each city, obtained from INE (1994), and the coordinates from Dirección General I.G.N. (1994). The spatial locations of Madrid and Barcelona are shown in Figure 22.

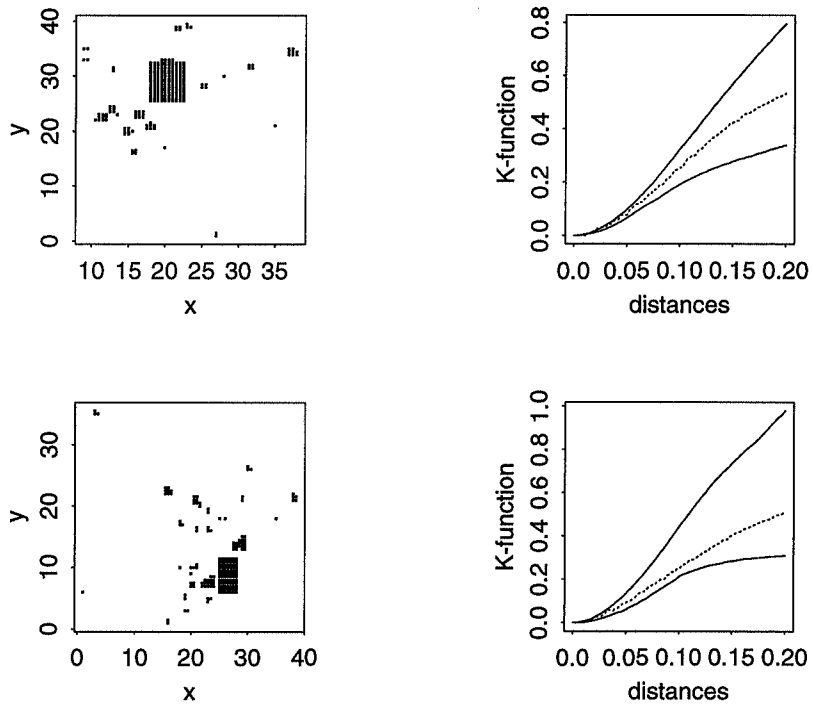


Figure 22. *Upper row*: Spatial locations of Madrid and empirical (dotted line) and confidence intervals (solid lines) of the K-function under a Neyman-Scott process with two clusters. *Lower row*: Spatial locations of Barcelona and empirical (dotted line) and confidence intervals (solid lines) of the K-function under a Neyman-Scott process with two clusters.

In Tables 4 and 5 the results are shown for the two selected provinces and compared to the total spanish peninsular territory. The column entries give us the number of points included in each territorial pattern, and the ID and ICS values for the following grid orders 10×10, 20×20 and 30×30.

Table 4. ID and ICS values for two spanish provinces

	Points	$\mathbf{ID_{10}}$	$\mathbf{ID_{20}}$	$\mathbf{ID_{30}}$	$\mathbf{ICS_{10}}$	$\mathbf{ICS_{20}}$	$\mathbf{ICS_{30}}$
Spain	1183	8393	27491	59763	83,78	67,90	65,48
Barcelona	185	1961	3020	3367	18,81	6,57	2,74
Madrid	235	2587	3843	4974	25,14	8,63	4,53

Table 5. Quantitative comparison of ID and ICS values for two spanish provinces

	Points	$\mathbf{ICS_{10}}$	$\mathbf{ICS_{20}}$
Barcelona	185	18,81	6,57
Madrid	$\frac{235}{1,2702703} = 185$	19,79	6,79

The indices values indicate that we have demographic structures characterized by cluster spatial processes (pval= 0.01 for both indices and grid order when testing CSR). But if we want to use more quantitatively the indices information, let us limit ourselves to compare the provinces of Barcelona and Madrid, as they are set on surfaces of (approximately) the same magnitude. Keeping in mind the total number of points in each pattern structure (185 and 235), we concentrate on grid orders of 10×10 and 20×20, as 30×30 will add irrelevant information as commented previously on the paper. Also, though both indices are sensitive at detecting the spatial structure depending on the number of points and degree of clustering, if there exists, the ICS values present the interesting characteristic of picking up the proportion among the number of points of different patterns within the same cluster degree. Due to this characteristic we have considered more appropriate to use the ICS index in order to make the comparisons in quantitative terms.

In order to remove from Madrid ICS index the component due to the increment in the number of points with regard to Barcelona, the values of the ICS index are divided by 1,27 (note that 235/185=1,2702703). As a result (see Table 5) we can conclude that the demographic spatial structure of Madrid province presents bigger intensity and bigger cluster degree than that of Barcelona.

Knowing that both, Madrid and Barcelona spatial patterns are clustered spatial structures, we tried to fit Neyman-Scott processes (as defined in section 3) to both patterns. As

a result, Madrid corresponds to a clustered pattern with two parents (pval= 0.89), one with 155 offsprings and a dispersion from the parent of 0.06 and other parent with 80 offsprings dispersed 0.10 from the parent. On the other hand, Barcelona corresponds to a clustered pattern with again two parents (pval= 0.83), one with 90 offsprings and a dispersion parameter of 0.06 and other parent with 100 offsprings and a dispersion parameter of 0.068.

The goodness of fit of both processes (see Figure 22) has been measured by means of the K-function, a second-order property defined as (Diggle, 1983)

$$K(t) = \lambda^{-1} E(NFE(t))) \tag{8}$$

where λ stands for the first-order intensity function and NFE(t) represents the number of further events within distance t of an arbitrary event (Diggle, 1983).

Then we have showed an application of the use of spatial indices in detecting a pattern structure and also in making quantitative comparisons between point patterns.

ACKNOWLEDGEMENTS

The referees are gratefully acknowledged for their helpful comments that have substantially improved an earlier version of the paper.

REFERENCES

Cormack, [1979). «Spatial aspects of competition between individuals». In *Spatial and Temporal Analysis in Ecology* (Cormack, R.M. and Ord, J.K., eds.), Internatiuonal Co-operative Publishing House, Maryland, 151-212.

Cox, D.R. & Isham, V. [1980). *Point processes*, Chapman & Hall, London.

Cressie, N. [1993). *Statistics for Spatial Data.* John Wiley & Sons, 2nd Edition, New York.

David, F.N. & Moore, P.G. [1954). «Notes on contagious distributions in plant populations». *Annals of Botany of London*, 18, 47-53.

Diggle, P.J. [1979). «Statistical methods for spatial point patterns in ecology». In *Spatial and Temporal Analysis in Ecology* (Cormack, R.M. and Ord, J.K., eds.), International Co-operative Publishing House, Maryland, 95-150.

Diggle, P.J. [1983). *Statistical Analysis of Spatial Point Patterns*, Academic Press, London.

Dirección General del Instituto Geográfico Nacional [1994). *Atlas Nacional de España*, MOPTMA.

Greig-Smith, P. [1952). «The use of random and contiguous quadrats in the study of the structure of plant communities». *Annals of Botany*, 16, 293-316.

Gulmon, S.L. & Mooney, H.A. [1977). «Spatial and temporal relationships between two desert shrubs *Atriplex hymnelytra* and *Tidestromia oblongifolia* in Death Valley, California». *Journal of Ecology*, 65, 831-838.

Hoel, P.G. [1943). «On indices of dispersion». *Annals of Mathematical Statistics*, 14, 155-162.

Hopkins, B. [1965). «Observations on savanna burning in the Olokemeji Forest Reserve». *Journal of Applied Ecology*, 2, 367-381.

Hudson, J.C. & Fowler, P.M. [1966). *The concept of Pattern in Geography*. Department of Geography, University of Iowa, Discussion paper: Series 1.

INE, [1994). *Anuario Estadístico de España de 1993, 1994.*

Lloyd, M. [1967). «Mean crowding». *Journal of Animal Ecology*, 36, 1-30.

Lösch, A. [1954). *The economics of location.* Yale University Press, New Haven.

Moellering, H. & Tobler, W.R. [1972). «Geographical variances». *Geographical Analysis*, 4, 34-50.

Morisita, M. [1959). «Measuring of the dispersion and analysis of distribution patterns. *Memoires of the Faculty of Science, Kyushu University», series E. Biology*, 2(4), 215-235.

Neyman,J. & Scott, E.L. [1958). «Statistical approach to problems of cosmology (with discussion)». *Journal of the Royal Statistical Society*, B 20, 1-43.

Perry, J.N. & Mead, R. [1979). «On the power of the index of dispersion test to detect spatial pattern». *Biometrics*, 35, 613-622.

Richardson, H.W. [1986). *Economía regional y urbana.* Alianza Editorial, Madrid.

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

Ripley, B.D. [1988). *Statistical Inference for Spatial Processes.* Cambridge University Press, Cambridge.

S.P.P.A. [1997). *Spatial Point Pattern Analysis.* Computer Software developed by Albert, Albert, Mateu & Pernias, Universitat Jaume I ,Castellón.

Serfling, R.J. [1980). *Approximation Theorems of Mathematical Statistics.* Wiley & Sons, New York.

Stiteler, W.M. & Patil, G.P. [1971). «Variance to mean ratio and Morisita's index as measures of spatial pattern in ecological populations». In *Statistical Ecology*, Vol. 1, (Patil, Pielou and Waters, eds.), University Park: Pennsylvania State University Press, 423-459.

Stoyan, D., Kendall, W.S. & Mecke, J. [1995). *Stochastic Geometry and its Applications.* Akademie-Verlag, 2nd Edition, Berlin.

Upton, G.J.G. & Fingleton, B. [1994). *Spatial Data Analysis by Example.* **Vol. 1**, John Wiley & Sons, New York.

Investigació Operativa

QÜESTIIÓ, vol. 24, 1, p. 55-82, 2000

DISEÑO DE ALGORITMOS PARA EL PROBLEMA DEL TRANSPORTE ESCOLAR. APLICACIÓN EN LA PROVINCIA DE BURGOS

J.A. PACHECO
A. ARAGÓN
C. DELGADO
Universidad de Burgos*

La problemática del transporte escolar es en Burgos especialmente significativa al ser una provincia extensa con muchos núcleos de población muy dispersos y poco poblados. En este trabajo se describen las aportaciones realizadas por los autores para dar solución a dicho problema, a través de técnicas que den soluciones lo más racionales posibles. En este sentido, hay que indicar que el término de racionalidad no sólo hace referencia a la minimización del coste total del transporte, sino también al cuidado de determinados aspectos como el tiempo que permanecen los alumnos en el vehículo, la elección de carreteras cómodas, etc., en especial a partir de determinados acontecimientos recientes. Asimismo, se muestran los resultados obtenidos con los datos del actual curso.

Algorithm design for school transportation. Application to Burgos province

Palabras clave: Transporte escolar, VRP, VRPTW, búsqueda tabú

Clasificación AMS: 90B20

*Departamento de Economía Aplicada. Universidad de Burgos. Parralillos, s/n. 09001 Burgos.

–Recibido en junio de 1999.
–Aceptado en enero de 2000.

1. INTRODUCCIÓN

Considérese el problema del transporte de un conjunto de alumnos que han de ser recogidos en una serie de localizaciones distribuidas geográficamente y llevados a un centro de enseñanza. Se denota por 1 el punto correspondiente al centro de enseñanza, y por $2,\ldots,n$ los puntos correspondientes a las localizaciones donde se recogen los alumnos. Sea $q(i)$ el número de alumnos que se recogen en cada punto $i, i = 2,\ldots,n$. Para cumplir estos requerimientos la legislación autoriza diferentes tipos de vehículos; cada tipo de vehículo con un número de plazas diferente. Para cada alumno el tiempo que trascurre desde que sube al autobús y entra en clase no debe sobrepasar de un determinado tiempo máximo, que denotamos por *tmax*, y la ruta debe finalizar antes del inicio de las clases en el instante *tinicio*. Las distancias d_{ij} y tiempos t_{ij} entre cada par de puntos $i, j \in \{1,2,\ldots,n\}$ son conocidas. El número de tipos de vehículos autorizados se denota por *ntipos* y las capacidades por *capactipo(i)*, $i = 1,\ldots,$*ntipos*.

Se ha de diseñar un conjunto de rutas de coste mínimo, verificando que se respeten los horarios y tiempos de conducción, y que el número de alumnos transportados en cada ruta sea inferior a la capacidad del vehículo asignado a dicha ruta.

Además, se va a imponer la restricción de que los alumnos de una determinada localización sean transportados por un solo vehículo. (En el caso en que en una localización el número de alumnos de ésta superen la máxima capacidad de todos los tipos de vehículo, se trataría dicha localización como dos diferentes: una con un número de alumnos igual a dicha capacidad, y otra con el resto. De todas formas, este caso no se ha dado en los datos reales analizados.)

El coste de transporte de cada ruta viene dado básicamente por el número de kilómetros recorridos, aunque también interviene el número de alumnos transportados y el número de paradas. En concreto, recientemente (13 de diciembre de 1997), el Ministerio de Educación y Cultura, a través de la Secretaría General de Educación y Formación Profesional, sugirió una fórmula que podría servir de referencia para el cálculo de las cantidades necesarias para las contrataciones de las rutas del transporte escolar. Estas cantidades (en pesetas) vendrían descompuestas en los 3 apartados antes mencionados (kilómetros, alumnos, paradas) de la siguiente forma:

- Coste por kilómetros $= \begin{cases} Pr * k & \text{si } k \leqslant 35 \\ Pr * 35 + (k-35) * (1'33) * Pr & \text{si } k > 35 \end{cases}$

  donde Pr es el precio de referencia por Km. que oscila entre 125 y 163 (en función de la calidad del vehículo), y k el nº de Kms. La cantidad por este concepto no podrá exceder de 14.250.

- Coste por alumnos $= \begin{cases} 100 * (M-33) & \text{si } M > 33 \\ 0 & \text{si } M \leqslant 33 \end{cases}$

  donde M es el nº de alumnos.

– Coste por paradas $= 75 * (L1 - 6)$, si $L1 > 6$
0 si $L1 \leqslant 6$

siendo $L1 = \min\{L, M/3\}$ y L el número de paradas.

Sin embargo, en otros casos los responsables en cada provincia han de negociar las cantidades por diferentes sistemas de tarifas. Es decir, se paga por kilómetro recorrido, y dicho precio por kilómetro depende del número de alumnos transportados en la ruta. En cualquier caso, la fórmula anteriormente diseñada supone una buena aproximación y es la que utilizaremos en este trabajo.

Finalmente, en cuanto al coste se ha de mencionar que la distancia recorrida comienza a contar desde el primer punto de recogida, y no desde el punto en el que sale el autobús[1], y en cualquier caso no hay un coste fijo en cada ruta por el tipo de vehículo utilizado. Esto es importante en el diseño de las estrategias de solución, puesto que de esta forma no va a ser necesario minimizar el número de vehículos a utilizar, es más, se ha observado que, en muchos casos, aumentando el número de rutas se puede llegar a soluciones menos costosas.

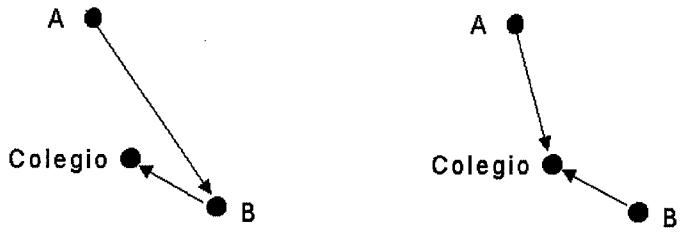


Figura 1. Al no haber coste fijo por vehículo, y al comenzar a contar desde el primer punto de recogida (origen y destino no coinciden), es fácil encontrar soluciones que con más rutas son más baratas que otras con menos. (La solución con las rutas A-Colegio y B-Colegio, recorre menos distancias que la solución con la ruta A - B - Colegio.)

Así considerado, este modelo es un caso particular del conocido *Problema de Rutas de Vehículos* o VRP (Vehicle Routing Problem) o, para ser más precisos, es un *Problema de Rutas de Vehículos con Restricciones de carga y tiempo* (ver Laporte y otros (1984) y (1985)).

Existen muchos algoritmos de solución para el VRP (y/o de variantes, principalmente del VRPTW) en la literatura. Se pueden encontrar recopilaciones de los principales

[1]Obviamente a partir de ahora $d_{1i} = t_{1i} = 0$, para $i = 2, \ldots, n$.

en trabajos como los de Bodin y Golden (1981), Desrochers y otros (1988), Haouri y otros (1990), Laporte (1992) y Laporte y Osman, (1995). En los últimos años han tomado importancia el desarrollo de algoritmos basados en procesos denominados Metaheurísticos como *Algoritmos Genéticos, Temple Simulado, Búsqueda Tabú, GRASP, Búsqueda Local Guiada (GLS), Colonias de Hormigas, etc.*, especialmente a partir de los trabajos de Gendreau y otros (1991), (1994) en versión posterior, y de Osman (1993); y más recientemente en los de Potvin y otros (1993) y (1994), Thangiah y otros (1993) y (1994), Campos y Mota (1995), Kantoravdis (1995), Rochat (1995), Kilby y otros (1997), Backer y otros (1997), Bullnheimer y otros (1997), o Rego (1998).

Al ser éste un problema NP-Hard (complejidad no Polinomial, ver Lenstra et al. (1981)), los algoritmos exactos (es decir, aquellos que garantizan la solución óptima) requieren un tiempo de computación que crece de forma exponencial con el número de elementos que intervienen en el problema. Por tanto, el uso de estos algoritmos puede requerir un tiempo de computación excesivo, incluso en problemas no muy grandes (menos de 100 puntos) cuando, como es el caso, se resuelven en ordenadores personales.

Por tanto, se va a optar por el diseño de una técnica heurística, es decir, que no garantiza la obtención del óptimo, pero sí una buena solución en un tiempo de computación más razonable. Esta estrategia está basada en dos partes: la primera consiste básicamente en una serie de procesos de Búsqueda Local, que dan soluciones rápidamente; la segunda (opcional según el tiempo de cálculo disponible), está basada en un proceso de Búsqueda Tabú que lleva incorporado un novedoso método de Intensificación. Por tanto, el algoritmo propuesto es una técnica compuesta por varias partes o subalgoritmos, basados en su mayor parte en movimientos vecinales.

En las dos siguientes secciones se describen las diferentes definiciones de vecindarios y un método para calcularlos. En la cuarta sección se describe la estructura del algoritmo general. En la quinta se describe un algoritmo basado en un proceso de Búsqueda Tabú que complementa el algoritmo anterior y que puede mejorar en muchas ocasiones la solución obtenida. Finalmente, en la sexta se describen los resultados obtenidos por los diferentes algoritmos y subalgoritmos, así como los tiempos de computación usados para una serie de experiencias en problemas reales con datos obtenidos en la provincia de Burgos.

A partir de ahora se denotará por S el conjunto de soluciones factibles del problema, y f la función de costes a minimizar definida en S.

2. CONSTRUCCIÓN DE SOLUCIONES VECINAS

Se van a considerar dos tipos de soluciones vecinas, una basada en la idea propuesta por Or (1976), para el Problema del Viajante o TSP, y otra más reciente que extiende el usado en los trabajos de Gendreau et al (1991), Osman (1993), Campos y Mota

(1995), etc., que se explicará en la siguiente sección. (En los trabajos de Taillard y otros (1995) y (1997), se definen interesantes estructuras vecinales parecidas aunque algo más complejas.)

2.1. Vecindario tipo Or

En este subapartado se van a definir como soluciones vecinas las obtenidas por el método de intercambio propuesto por Or (1976) que sean factibles. El método de intercambio Or es una variante de los conocidos intercambios r-óptimos desarrollados por Lin, (1965) y Lin & Kernighan (1973) para el TSP simétrico. Como se verá mas adelante, el método de Or se puede utilizar en problemas asimétricos. La eficacia de este método para el TSP ha sido contrastada en trabajos como el de Nurmi (1991).

Obsérvese que una solución puede expresarse de forma sencilla, como una única secuencia de puntos; por ejemplo, la solución formada por las dos rutas siguientes:

$$\textit{ruta 1: } 1-3-5-1; \quad \text{y} \quad \textit{ruta 2: } 1-4-2-1$$

puede expresarse como:

$$1-3-5-1-4-2-1$$

los '1' representan la vuelta de un vehículo al origen y la salida del siguiente (obviamente el último y el primer elemento siempre serán '1'). De esta forma, como se ilustra a continuación, se puede aplicar los Or-intercambios para obtener soluciones vecinas de la actual, considerando a ésta como una única secuencia.

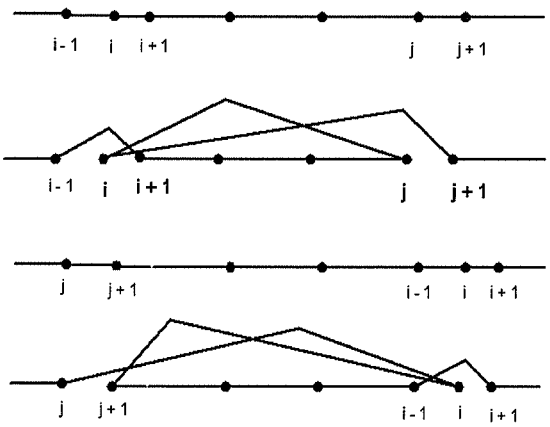


Figura 2. Posible recolocación del elemento i hacia adelante y hacia atrás entre j y $j+1$.

Or propone restringir la búsqueda de intercambios a los *3-intercambios* en los que cadenas[2] de uno, dos o tres puntos consecutivos son recolocadas entre otras dos. Nótese que con estos intercambios no se cambia el sentido de los diferentes tramos.

En nuestro caso seguiremos la misma idea, pero sólo consideraremos recolocaciones hacia adelante (se ha observado que considerando también las recolocaciones hacia atrás apenas hay diferencias en los resultados finales y el tiempo de computación es el doble); además, se debe chequear la factibilidad de cada posible recolocación respecto a las restricciones del problema. A continuación se ilustra la recolocación de una cadena de k elementos comenzando en i entre j y $j+1$.

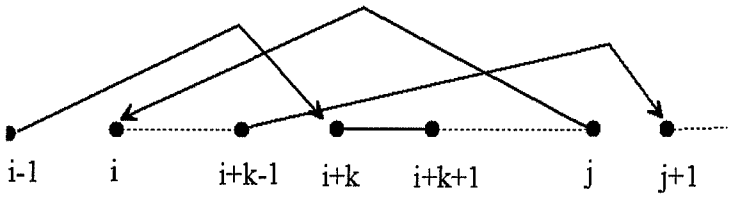


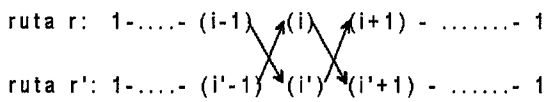
Figura 3. Recolocación de una cadena de k elementos.

Para cada $s \in S$ se va a denotar por $N_1^k(s)$ al conjunto de soluciones factibles obtenidas por recolocaciones hacia adelante de cadenas de a lo sumo k elementos en s. Se denota por $N_1^\infty(s)$ el conjunto de soluciones factibles obtenidas por todas las recolocaciones.

2.2. Vecindarios tipo Gendreau-Clarke

El segundo tipo de vecindarios no considera las soluciones como una única secuencia de puntos, sino que sólo considera 3 tipos de intercambios entre 2 rutas diferentes:

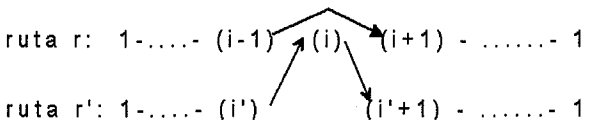
- Tipo I: Intercambio del elemento i de la ruta r y con el elemento i' de la ruta r':
  - Eliminación de los arcos $(i-1,i),(i,i+1),(i'-1,i'),(i',i'+1)$
  - Incorporación de los arcos $(i-1,i'),(i',i+1),(i'-1,i),(i,i'+1)$



[2]En este trabajo se denomina cadena a toda secuencia de puntos consecutivos en la solución actual.

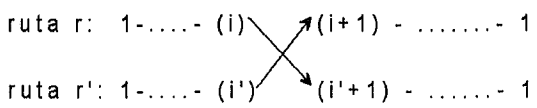
– Tipo II: Inserción del elemento i de la ruta r entre los elementos i' e $i'+1$ de la ruta r':

  - Eliminación de los arcos $(i-1,i),(i,i+1),(i',i'+1)$
  - Incorporación de los arcos $(i',i),(i,i'+1),(i-1,i+1)$



– Tipo III: Cruce de las rutas r y r' por los elementos i e i' según la figura:

  - Eliminación de los arcos $(i,i+1),(i'i'+1)$
  - Incorporación de los arcos $(i',i+1),(i,i'+1)$.



Los dos primeros tipos aparecen en el trabajo de Gendreau et al. (1991) y otros mencionados anteriormente; el tercero, sin embargo, no aparece en dichos trabajos: se basa en algunas de las ideas propuestas por Clarke and Wright (1964).

Obsérvese que el tipo I es en realidad un 4-intercambio, el tipo II un 3-intercambio y el tipo III un 2-intercambio. Para cada $s \in S$ se denotará a los conjuntos de soluciones factibles generadas por cada uno de estos 3 tipos de intercambio como $N_2^1(s), N_2^2(s)$ y $N_2^3(s)$ respectivamente; asimismo, se denota por $N_2(s)$ la unión de estos 3 conjuntos.

Una vez descritos los tipos de vecindarios que se van a utilizar, se deben hacer las siguientes puntualizaciones:

– En los subapartados 2.1. y 2.2. anteriores se han considerado rutas como ciclos que comienzan y acaban en 1, pero en la descripción del problema las rutas son caminos abiertos que comienzan en alguna población donde se recogen a los primeros niños y finaliza en el colegio 1. Sin embargo, como se indica en la descripción del problema, se va a considerar o redefinir $d_{1i} = t_{1i} = 0$, para $i = 2,\ldots,n$; lo que nos permite considerar a las rutas como ciclos (aunque en realidad sean caminos abiertos) y esto, al menos para los autores, resulta más cómodo para la programación de los algoritmos.

– A las rutas que componen una solución s se añade una ruta ficticia vacía $(1-1)$. El objeto de esto es permitir obtener soluciones vecinas con más rutas reales (no vacías) añadiendo en la ruta ficticia elementos de las otras rutas. Recuérdese, como se dijo en la introducción, que a veces se consiguen mejores soluciones con más rutas.

El chequeo de la factibilidad y la valoración de cada intercambio (tanto de tipo Or como de tipo I, II o III) puede suponer un tiempo de computación excesivo. En los trabajos de Pacheco y Delgado (1996) y (1997) se propone el uso de variables globales, con lo que el número de operaciones para chequear y evaluar cada intercambio es constante, es decir, independiente del tamaño del problema. En la siguiente sección se intenta resumir las ideas básicas de estos trabajos.

3. FACTIBILIDAD Y VALORACIÓN DE CADA INTERCAMBIO

El chequeo de la factibilidad y la valoración de cada intercambio (tanto de tipo Or como de tipo I, II o III) puede suponer un tiempo de computación excesivo. Valorar un cambio en el tipo de problema que se está estudiando, no sólo consiste en calcular la diferencia entre los costes de los arcos que se añaden y los que se quitan: hay que determinar también y tener en cuenta los nuevos tipos de vehículos requeridos. En los trabajos de Pacheco y Delgado (1996) y (1997) se propone el uso de variables globales, con las que el número de operaciones para chequear y evaluar cada intercambio es constante, es decir, independiente del tamaño del problema.

3.1. Valoración y factibilidad respecto a la carga

Sea una ruta de nr puntos $r(1) - r(2) - \cdots - r(nr-1) - r(nr)$, (obviamente $r(nr) = r(1) = 1$). Para calcular los espacios requeridos tras cada intercambio se define:

- *esp_ocup(s)* como la carga en el vehículo después de visitar el punto $r(s)$, para $s = 2, \ldots, nr$;

asimismo sea:

- *cadena* $= r(p) - r(p+1) - \cdots - r(q)$, cualquier cadena de la solución actual,

se define

- *maximo_esp_ocup(p,q)* como la máxima carga en el vehículo durante la visita a los puntos de cadena;

obviamente

- *maximo_esp_ocup(p,q)* $= \max_{s=p,\ldots,q}$ *esp_ocup(s)* y
- *maximo_esp_ocup(p,q)* $= \min\{$*maximo_esp_ocup*$(p,q-1),$ *esp_ocup(q)*$\}$, con
- *maximo_esp_ocup(p,p)* $=$ *esp_ocup(p)* como valor inicial.

4.2. Factibilidad respecto a las ventanas de tiempo

Obsérvese que los valores *tmax* y *tinicio*, definidos en la introducción, implícitamente definen a su vez unos intervalos de visita en cada punto $i, [e,l]$, donde e = *tinicio-tmax*, y l = *tinicio*, para $i = 1,\ldots,nr$.

Para todo $s = 2,\ldots,nr$ se van a definir las siguientes variables, inspiradas en el trabajo de Savelsberg (1985):

- $A(s)$: Tiempo de llegada del vehículo a $r(s)$,
- $D(s)$: Tiempo de salida de $r(s)$; se tiene que $D(s) = \max\{A(s), e\}$;
- $M(s)$: Margen de tiempo en la llegada a $r(s)$, es decir, $M(s) = l - A(s)$;
- $E(s)$: Tiempo de espera del vehículo en $r(s)$, es decir, $E(s) = \max\{e - A(s), 0\}$.

A continuación se muestra un gráfico con un ejemplo que ilustra estas definiciones.

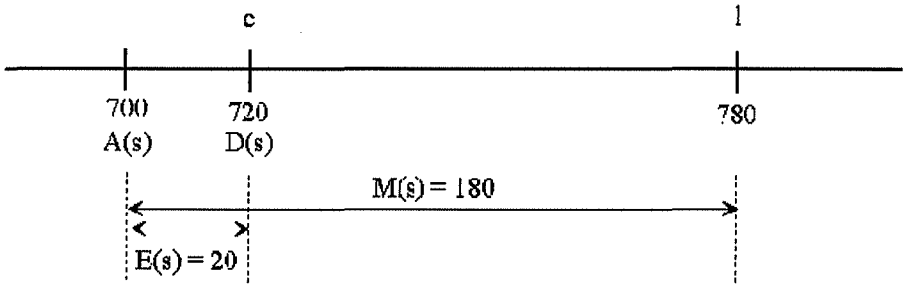


Figura 6. Ejemplo de la llegada del vehículo a un punto $r(s)$, con er(s)=720 y lr(s)=780

Si la llegada $A(s)$ se produce en 700, entonces el tiempo de salida es $D(s)$ = 720, el tiempo de espera $E(s)$, es 20, y el margen, $M(s)$, es 80.

Proposición 1. *Para cualquier $p = 1,\ldots,nr$, considérese un retardo en la llegada a $r(p)$ de x unidades de tiempo; sea A^* y D^* los nuevos tiempos de llegada y salida que este incremento produce en los puntos siguientes de la solución, se tiene que:*

$$A^*(p+h) = A(p+h) + \max\left\{x - \sum_{s=p,\ldots,p+h-1} E(s), 0\right\}, \quad \text{para } h = 0,\ldots,nr-p.$$

Demostración. Por inducción en los valores de h:

- para $h = 0$ es trivial: $A^*(p) = A(p) + x$.
- sea cierto para $h-1$, entonces,

$$A^*(p+h-1) = A(p+h-1) + \max\left\{x - \sum_{s=p,\ldots,p+h-2} E(s), 0\right\}.$$

■

Observando la figura 6 anterior y teniendo presente las relaciones entre A, D y E, es fácil de comprobar que un retardo en la llegada a un punto, repercute en el tiempo de salida en dicho punto, si este retardo es mayor o igual que el tiempo de espera; en este caso el aumento en el tiempo de salida será igual a la diferencia entre el retardo y el tiempo de espera; por tanto:

$$D^*(p+h-1) = D(p+h-1) + \max\left\{\max\left\{x - \sum_{s=p,\ldots,p+h-2} E(s), 0\right\} - E(p+h-1), 0\right\} =$$

$$= D(p+h-1) + \max\left\{x - \sum_{s=p,\ldots,p+h-1} E(s), 0\right\}$$

luego:

$$A^*(p+h) = D^*(p+h-1) + tr(p+h-1)r(p+h) =$$

$$= A(p+h) + \max\left\{x - \sum_{s=p,\ldots,p+h-1} E(s), 0\right\}.$$

Corolario 1. Sea $H(p,q)$, siendo $p = 1,\ldots,nr$, y $q = p,\ldots,nr$, el máximo retardo en la llegada a $r(p)$, es decir, incremento en $A(p)$ que no produce violación de las ventanas de tiempo en $r(q)$ (se mantiene $A^*(q) \leqslant 1$); se tiene que:

$$H(p,q) = \sum_{s=p,\ldots,q-1} E(s) + M(q)$$

Sea:

$$cadena = r(p) - r(p+1) - \cdots - r(q),$$

se define

- *maxretraso(p,q)* = Máximo retardo en la llegada a $r(p)$, que no produce violaciones de las ventanas de tiempo en los puntos de cadena;
- *tespera(p,q)* = Tiempo total de espera acumulado en cadena;

obviamente

$$\textit{tespera}(p,q) = \sum_{s=p,\ldots,p+h-1} E(s) \quad \text{y}$$

$$\textit{tespera}(p,q) = \textit{tespera}(p,q-1) + E(q);$$

por comodidad si $q < p$ (si cadena no contiene elementos), se define *tespera(p,q) = 0.*

Proposición 2. *Se tiene que los valores de maxretraso se pueden obtener de la forma siguiente:*

$$\textit{maxretraso}(p,q) = \min_{s=p,\ldots,q} H(p,s) = \min_{s=p,\ldots,q} \{M(s) + \textit{tespera}(p,s-1)\}$$

Demostración. Obvio a partir de Proposición 1 y Corolario 1 y la definición de maxretraso.

■

Un método recursivo para obtener el valor de maxretraso viene dado por la siguiente fórmula:

$$\textit{maxretraso}(p,q) = \min\{\textit{maxretraso}(p,q-1), M(q) + \textit{tespera}(p,q-1)\}$$

con

$$\textit{maxretraso}(p,p) = M(p) \quad \text{como valor inicial.}$$

Asimismo se define:

- *maxadelanto(p,q)* = Máximo adelanto en la llegada a $r(p)$, es decir disminución de $A(p)$, que no produce tiempo de espera en dicha cadena (es decir, se mantiene $A(s) \geqslant e$ para $s = p,\ldots,q$); obviamente si en algún punto de cadena ya existe espera maxadelanto(cadena) es 0.

Proposición 3. *Los valores de maxadelanto pueden obtenerse de la siguiente forma:*

$$\textit{maxadelanto}(p,q) = \min_{s=p\ldots q} \{\max[0, A(s) - e]\}$$

A partir de aquí se puede obtener un método recursivo para su cálculo:

$$\textit{maxadelanto}(p,q) = \min\{\textit{maxadelanto}(p,q-1), \max[0, A(q) - e]\}$$

con

$$maxadelanto(p,p) = \max[0, A(p) - e] \quad \text{como valor inicial.}$$

Estas variables son importantes a la hora de determinar cómo afecta la alteración en el tiempo de llegada al primer punto de una cadena, al tiempo de salida del último punto de esa cadena y a la posible violación o no de las ventanas de tiempo en los puntos de la cadena. Concretamente, sea para una determinada cadena $r(p) - r(p+1) - \cdots - r(q)$, adelanto_inicial la cantidad de tiempo en que se adelanta la llegada a $r(p)$, y adelanto_final la cantidad de tiempo en que se adelanta la salida de $r(q)$, se tiene que:

$$adelanto\_final = \min\{adelanto\_inicial, maxadelanto\}$$

análogamente, sean *retraso_inicial* y *retraso_final* respectivamente los atrasos en los tiempos de llegada a $r(p)$ y salida de $r(q)$, se tiene que:

$$retraso\_final = \max\left\{ retraso\_inicial - \sum_{s=p,\ldots,q} E(s), 0 \right\}.$$

Las variables globales definidas en este apartado pueden ser calculadas con orden $\theta(n^2)$ de operaciones, facilitando posteriormente el chequeo de la factibilidad y valoración de los intercambios.

4. DISEÑO DEL ALGORITMO INICIAL

Como se ha mencionado en la introducción, el algoritmo propuesto se divide en varios subalgoritmos que, en la mayoría de los casos, se basan en movimientos vecinales. Sea $s^* \in S$ la solución actual en cada momento, y $k \in IN$ un número entero prefijado, inicialmente se van a definir los siguientes procedimientos:

▸ Procedimiento Búsqueda_Local_Or(k, sf)

Repetir

Determinar $s' \in N_1^k(sf)$ *verificando* $f(s') = \min\{f(s)/s \in N_1^k(sf)\}$

Si $f(s') < f(sf)$ *entonces hacer* $sf = s'$

hasta que $f(s) \geqslant f(sf)$, $\forall s \in N_1^k(sf)$.

▸ Procedimiento Búsqueda_Local_Ge(sf)

Repetir

Determinar $s' \in N_2(sf)$ *verificando* $f(s') = \min\{f(s)/s \in N_2(sf)\}$

Ejecutar Búsqueda_Local_Or$(3, r')$ y Búsqueda_Local_Or$(3, r'')$ *donde* r' *y* r'' *son las dos rutas de* sf *modificadas para dar lugar a* s'

Si $f(s') < f(sf)$ *hacer* $sf = s'$
hasta que $f(s) \geqslant f(sf)$, $\forall s \in N_2(sf)$.

Estos procedimientos de Búsqueda Local son análogos, aunque obsérvese como en el segundo caso, cuando se realiza un intercambio (tipo Gendreau-Clarke) a continuación se ejecuta el primero para mejorar cada una de las dos rutas implicadas en este intercambio.

Una vez descritos estos procedimientos el Algoritmo Inicial queda de la siguiente forma:

▶ Procedimiento Algoritmo_Inicial(Output $s^* \in S$)
Paso 1: Obtención de una Solución Inicial s^* por un método constructivo
Paso 2: Ejecutar Búsqueda_Local_Ge(s^*)
Paso 3: A cada ruta de s^* aplicar Búsqueda_Local_Or con $k = \infty$
Paso 4: Ejecutar Búsqueda_Local_Or(∞, s^*)

Para la obtención de una solución inicial se usa una adaptación del algoritmo de Fisher & Jaikumar (1981) para este modelo. Aunque en este caso también se obtienen soluciones aceptables con una adaptación para este modelo del algoritmo de *inserción más cercana*. (Una colección exhaustiva de estos métodos constructivos para el TSP se puede encontrar en el trabajo de Golden y otros, (1980).)

Tras el paso 2, cada una de las rutas de la solución s^* obtenida es óptimo local con respecto a N_1^3, es decir, no son mejorables con recolocaciones de tipo Or de cadenas de hasta 3 elementos. Se ejecuta el paso 3 por si pueden ser mejoradas con recolocaciones de cadenas de mayor tamaño (obsérvese que en este paso sólo se consideran recolocaciones de elementos dentro de una misma ruta).

Finalmente, en el paso 4 se vuelve a ejecutar el mismo procedimiento *Búsqueda_Local_Or* a la solución obtenida, pero considerada como una sola cadena, con el objeto de analizar las posibles recolocaciones que afecten a rutas diferentes.

5. MEJORA CON UN PROCEDIMIENTO DE BÚSQUEDA TABÚ

El algoritmo descrito en el apartado anterior da resultados satisfactorios mejorando sensiblemente las soluciones usadas actualmente en los Centros Escolares analizados, con un tiempo de cálculo razonable (ver apartado siguiente).

Sin embargo, las soluciones obtenidas aún podrían ser ligeramente mejoradas, en algunos casos, añadiendo un paso posterior en el que se aplique alguna técnica Me-

taheurística desarrollada en los últimos años. Más concretamente, se propone un procedimiento basado en un proceso de búsqueda tabú que se describe a continuación.

La *búsqueda tabú* es un procedimiento o estrategia dado a conocer en los trabajos de Glover (1989) y (1990), y que está teniendo grandes éxitos y mucha aceptación en los últimos años. Según su creador, es un procedimiento que «*explora el espacio de soluciones más allá del óptimo local*» (Glover y Laguna (1993)). Se permiten cambios *hacia arriba* o que empeoran la solución, una vez que se llega a un óptimo local. Simultáneamente, los últimos movimientos se califican como *tabúes* durante las siguientes iteraciones para evitar que se vuelva a soluciones anteriores y el algoritmo cicle. El termino tabú hace referencia a «*un tipo de inhibición a algo debido a connotaciones culturales o históricas y que puede ser superada en determinadas condiciones...*». (Glover (1996)). Recientes y amplios tutoriales sobre búsqueda tabú, que incluyen todo tipo de aplicaciones, pueden encontrarse en Glover y Laguna (1997) y (1999).

5.1. El algoritmo básico

El algoritmo que se propone básicamente actúa de la forma siguiente:

▸ Procedimiento Búsqueda_Tabú_Básico
  Leer como solución inicial s^* (obtenida por el Algoritmo Inicial descrito en la sección 4)
  Hacer $s_0 = s^*; T = \emptyset, niter = 0, kiter = 0$
  Repetir
    niter:=niter+1;
    Seleccionar $s \in N_2(s_0)/s \notin T$ *o s verifica criterio de 'aspiración' con* $f(s)$ *mínimo*
    Hacer $s_0 = s$
    Ejecutar Búsqueda_Local_Or(3,r') *y* Búsqueda_Local_Or(3,r') *donde* r' *y* r'' *son las dos rutas de* s_0 *modificadas*
    Si $f(s_0) < f(s^*)$ *entonces: hacer* $s^* = s_0$ *y kiter = niter*
    Actualizar T
  hasta niter $-$ *kiter* $\geqslant$ *maxiter*

Se denota por s^* a la solución óptima en cada momento. T es el conjunto de movimientos tabúes y se obtiene determinando qué conjunto de soluciones tienen ciertos *atributos tabús activos*. Por tanto, se han de definir estos atributos tabús y durante cuántas iteraciones van a permanecer activos (y, por tanto, las soluciones que les contienen).

El objeto de aplicar el *criterio de aspiración* es determinar en qué condiciones un movimiento tabú puede ser admisible. Habitualmente se considera que una solución s cumple el *criterio de aspiración* si $f(s) < f(s^*)$. Este movimiento facilita una nueva dirección de búsqueda y garantiza que no se produzcan ciclos.

Por otra parte, cualquier movimiento vecinal de este tipo supone la incorporación de un conjunto de arcos y la eliminación de otros (según se ilustró en el apartado 3). Para que en las iteraciones siguientes no se vuelva a la solución anterior, se van a impedir los movimientos que supongan la incorporación de algunos de estos arcos en la solución actual. En otras palabras, los *atributos tabús* van a ser los arcos que componen cada solución, y un movimiento es tabú si supone la incorporación de algún arco eliminado en iteraciones recientes (*atributo tabú activo*).

Para identificar qué atributos tabús (arcos) van a estar activos se define arco_tabú una matriz $n \times n$ de la siguiente forma:

$$arco\_tabú(r,l) = \text{n}^{\text{o}} \text{ de la última iteración en la que fue eliminado el arco}(r,l).$$

Un determinado arco (r,l) será un atributo tabú activo si:

$$niter - arco\_tabú(r,s) < maxiter\_tabú$$

siendo *maxiter_tabú* el número de iteraciones que permanece activo como atributo tabú desde que es eliminado de la ruta actual.

Inicialmente se define $arco\_tabú(r,l) = 0$, para cada arco (r,l) en la solución inicial, $arco\_tabú(r,l) = -maxiter\_tabú$ (o un valor más negativo) para el resto. De esta forma se impide que en las primeras iteraciones sean declarados tabús activos arcos que no formen parte de la solución actual. Así, inicialmente se asegura que $T = \emptyset$.

- Se va a dar al parámetro maxiter_tabú el valor de $maxiter\_tabú = 2^* n^{1/2}$.
- Para el criterio de parada se toma maxiter = $10 \cdot n$.

En esta sección se ha descrito un algoritmo de búsqueda tabú básico, pero que puede ser complementado y ampliado con procedimientos basados en lo que se denomina habitualmente memoria a *largo* y *medio plazo* como *diversificación* e *intensificación*; también se puede enriquecer con la posibilidad de visitar de forma controlada soluciones infactibles, por medio de funciones de penalización (*oscilación estratégica*).

5.2. Fase de intensificación

Como señala el nombre, en esta fase se intensifica la exploración de las regiones y los vecindarios donde se hallan las mejores soluciones encontradas en fase inicial (algoritmo básico) con la esperanza de encontrar aún ligeras mejoras. Esta fase puede ser

diseñada de diferentes formas. En este caso, está inspirada en los trabajos de Rossing (1997), Rossing y otros (1997) y (1998) sobre *Concentración Heurística*, una estrategia para encontrar soluciones en dos fases.

En la primera se ejecuta varias veces un procedimiento de búsqueda local, registrándose los mejores óptimos locales obtenidos; en la segunda se construye un *conjunto de concentración*, CS, con los elementos de las mejores soluciones obtenidas en la primera, y se ejecuta un algoritmo exacto o heurístico pero considerando sólo los elementos de CS.

En este trabajo se va a tomar la idea de construir un *conjunto de concentración* con los elementos de las mejores soluciones obtenidas en el algoritmo básico. Más concretamente, para este modelo se va a considerar como elementos de las soluciones a los arcos que la componen; por ejemplo la solución compuesta por las dos rutas:

$$\textit{ruta 1:}1-3-5-1; \quad \text{y} \quad \textit{ruta 2:}1-4-2-1$$

que puede expresarse como la secuencia $1-3-5-1-4-2-1$ (según se vio en el apartado 2), estará formada por los arcos (1, 3), (3, 5), (5, 1), (1, 4), (4, 2) y (2, 1).

Para formar el conjunto de concentración utilizamos el siguiente procedimiento:

▶ Procedimiento Construcción_Conjunto_de_Concentracion
Ordenar las soluciones obtenidas en una lista según el valor de f (comenzando con la mejor)
Hacer $i = 0$ *y* $CS = \emptyset$
Repetir
hacer $i = i + 1$
Añadir los elementos de la i-ésima solución de la lista a CS
hasta Cardinal(CS) $\geqslant$ *num_arcos*

Obsérvese que a diferencia de la propuesta por Rossing no se fija un número predeterminado de soluciones cuyos elementos se introducen, sino que se fija un número mínimo de elementos a introducir *num_arcos*.

A continuación, se va a diseñar un procedimiento que *concentre* la búsqueda de elementos en CS. Inicialmente, se define la siguiente matriz de distancias auxiliar $d1$ de la siguiente forma:

$$d1(i,j) = d(i,j) \qquad \text{si } (i,j) \in \text{CS}$$
$$d1(i,j) = d(i,j) + 10^{*}max\_d \qquad \text{si } (i,j) \notin \text{CS}$$

donde $max\_d = \max\{d(i,j)/i,j = 1,\ldots,n\}$; además se define $f1(s)$ como el valor de

la función objetivo considerando el coste de los arcos el dado por $d1$. Se propone el siguiente procedimiento de búsqueda que tiene en cuenta ambas matrices d y $d1$:

▶ Procedimiento Búsqueda_Local_Guiada (sf)
Hacer $s_0 = sf$
Repetir
Hacer coste_anterior = f(sf)
Repetir
Buscar $s' \in N_2(s_0)/f1(s') = \min\{f1(s)/s \in N_2(s_0)\}$
En s' *Ejecutar* Búsqueda_Local_Or(3, r') *y* Búsqueda_Local_Or(3, r'')
(considerando f1 en vez de f) donde r' *y* r'' *son las dos rutas de* s_0 *modificadas para dar lugar a* s'
Si $f1(s') < f1(s_0)$ *entonces* $s_0 = s'$
Buscar $s'' \in N_2(s_0)/f(s'') = \min\{f(s)/s \in N_2(s_0)\}$
En s'' *Ejecutar* Búsqueda_Local_Or(3, r') *y* Búsqueda_Local_Or(3, r'') *donde* r' *y* r'' *son las dos rutas de* s_0 *modificadas para dar lugar a* s''
Si $f(s'') < f(sf)$ *entonces* $sf = s''$
hasta $f1(s') \geqslant f1(s_0)$
Hacer $s_0 = sf$
hasta f(sf) = coste_anterior

Se trata de un procedimiento de búsqueda local anidado: en cada paso la solución actual s_0 se sustituye por otra mejor según $f1$, es decir según $d1$ y buscando por tanto soluciones que contengan elementos de CS; cuando no hay mejora en $f1$, se sustituye s_0 por sf, la mejor solución según f observada en los vecindarios explorados y se reinicia la búsqueda local. El proceso acaba cuando no hay mejora en $f(sf)$.

En definitiva, es un procedimiento de búsqueda 'guiado' por $f1$, i.e. por $d1$, hacia soluciones que contengan el mayor número de elementos de CS posibles. En cierta manera, podría ser considerado como una forma de reencadenamiento de trayectorias. Obsérvese que $d1$ 'penaliza' a los arcos no pertenecientes a CS, pero no 'impide' su elección en cada paso, ya que esto podría hacer excesivamente reducido el número de soluciones a considerar y 'encajonar' el proceso. Obviamente, si se usara un algoritmo exacto la estrategia debería ser impedir en vez de penalizar. Este procedimiento se inserta en la fase de intensificación que queda de la siguiente forma:

▶ Procedimiento intensificación
Ejecutar Construcción_Conjunto_de_Concentracion;
Desde i:=1 hasta num_soluciones hacer

Tomar s_i *la i-esima solución de la lista ordenada obtenida en*
Búsqueda_Tabú_Básico
Ejecutar Ejecutar Búsqueda_Local_Ge(s_i)
A cada ruta de s_i *aplicar* Búsqueda_Local_Or *con* $k = \infty$,
Ejecutar Búsqueda_Local_Guiada(s_i);
Si $f(s_i) < f(s^*)$ *hacer* $s^* = s_i$

Se toma $f1$ en vez de f al generar las soluciones iniciales y se ejecuta el procedimiento de *Búsqueda_Local_Guiada* dirigido por $f1$, en vez de la búsqueda local habitual. En definitiva, se 'concentra' o intensifica la búsqueda de soluciones en las regiones con elementos de CS.

Para este trabajo se ha tomado *num_soluciones* = 50. En cuanto al valor de este parámetro *num_arcos*, en el trabajo de Pacheco y Delgado (1999), se describen los resultados de diferentes experiencias que aconsejan tomar *num_arcos* como el 10% del total de arcos (i.e. $num\_arcos = 0.1^* n^* (n-1)$).

Al añadir la fase de intensificación el algoritmo búsqueda tabú queda de la siguiente forma:

▶ Procedimiento Búsqueda_Tabú_Principal;
Ejecutar Búsqueda_Tabú_Básico
Ejecutar Intensificación

Se han de tener en cuenta las siguientes consideraciones: durante la ejecución de la fase básica es necesario crear y actualizar una lista con las mejores soluciones para ser usada en la fase de intensificación; para formar parte de esta lista se considera en cada iteración la mejor de todas las soluciones del vecindario analizado independientemente de contener o no elementos tabú o de cumplir el criterio de aspiración. Por otra parte se podría añadir al procedimiento Búsqueda_Tabú_Principal una fase de diversificación y volver a ejecutarlo una o varias veces. Sin embargo, en este trabajo no va a ser así para evitar tiempos de computación excesivos.

6. RESULTADOS COMPUTACIONALES

En esta sección se van a analizar los diferentes problemas o instancias del transporte escolar de secundaria en la provincia de Burgos. Cada problema viene definido por un centro escolar, por las localizaciones donde se recogen los alumnos que van a ese centro y por el número de alumnos a recoger en cada una de esas localizaciones. En

todos los casos el tiempo máximo de un alumno en ruta, *tmax*, es de 60′. Cada una de estas instancias se ha resuelto con los algoritmos anteriormente descritos.

Para el cálculo de la matriz de distancias y del camino entre cada par de puntos del problema se ha ponderado la distancia de cada tramo según el tipo de carretera, de forma que favorezca la elección de carreteras nacionales antes que autonómicas, éstas antes que carreteras sin revestimiento, etc., de forma que en muchas ocasiones los caminos obtenidos no son los más cortos, pero si los más cómodos y rápidos.

A continuación, para los problemas correspondientes al transporte de centros de secundaria se muestra la población donde está cada centro, el número de localidades donde se recogen los niños que van a esos centros y el número de niños, los resultados con los costes y los tiempos de computación de los algoritmos anteriormente descritos, así como de la solución que se usa en la realidad (siempre teniendo en cuenta la función de costes descrita en la sección 1).

Pr.	Centro (Población)	n. loc. alumn.	S. Actu.	Algoritmo inicial				B. Tabú	
				Paso 1	Paso 2	Paso 3	Paso 4	Básico	Intensif.
1°	ARANDA DE DUERO	57 429	61.177,5 12	67.123,6 10 8,72	57.973,4 13 0,61	57.973,4 13 0,03	57.973,4 13 0,40	57.973,4 13 26,87	56.196,1 14 72,78
2°	BELORADO	24 175	21.045,0 5	22.960,5 4 0,29	16.962,5 8 0,22	16.962,5 8 0,02	16.962,5 8 0,12	16.962,5 8 7,08	16.962,5 8 20,74
3°	BRIVIESCA	24 101	25.051,3 6	24.055,0 5 0,75	23.333,8 6 0,07	23.333,8 6 0,01	23.333,8 6 0,06	23.197,4 7 7,56	23.197,4 7 7,02
4°	L. Mendoza BURGOS	19 127	18.608,8 3	18.104,4 4 0,26	17.542,4 4 0,03	17.542,4 4 0,00	17.542,4 4 0,03	16.880,3 5 5,30	16.863,6 5 6,26
5°	Diego Marín BURGOS	23 59	15.902,5 4	14.773,3 3 0,35	14.614,9 3 0,02	14.614,9 3 0,00	14.614,9 3 0,03	14.614,9 3 5,25	14.614,9 3 4,43
6°	Diego Siloé BURGOS	32 141	30.720,0 4	34.831,6 5 1,81	29.742,0 6 0,18	29.742,0 6 0,01	29.742,0 6 0,08	27.410,8 6 11,65	27.410,8 6 12,29
7°	S.de Colonia BURGOS	36 158	34.835,0 6	28.147,5 5 1,98	22.820,9 7 0,31	22.820,9 7 0,01	22.820,9 7 0,17	22.539,1 7 27,71	22.539,1 7 19,12
8°	LERMA	53 302	51.112,5 9	52.730,5 9 6,80	45.152,3 11 0,46	45.152,3 11 0,02	43.152,9 10 1,40	41.006,9 10 53,77	40.542,0 11 42,49
9°	MEDINA DE POMAR	39 182	31.075,0 5	35.005,8 6 2,87	30.790,4 7 0,19	30.790,4 7 0,01	30.790,4 7 0,15	30.189,1 7 16,55	30.189,1 7 15,81
10°	MELGAR	22 71	19.402,5 6	19.758,9 4 0,44	18.887,0 5 0,02	18.887,0 5 0,01	18.887,0 5 0,05	18.847,9 6 5,27	18.847,9 6 5,16
11°	MIRANDA	13 41	16.622,5 4	19.038,1 4 0,17	18.847,9 5 0,03	18.847,9 5 0,01	18.847,9 5 0,01	18.847,9 5 2,27	18.847,9 5 3,60
12°	QUINTANAR	4 105	8.025,0 2	6.925,0 2 0,01	3.312,5 4 0,02	3.312,5 4 0,00	3.312,5 4 0,00	3.312,5 4 0,66	3.312,5 4 1,79
13°	ROA	28 207	22.975,0 6	28.072,4 5 0,51	20.237,5 7 0,26	20.237,5 7 0,01	20.200,0 7 0,17	20.200,7 7 8,83	19.437,5 7 21,85
14°	SALAS	23 81	21.488,8 5	21.371,1 5 0,55	18.586,8 4 0,04	18.586,8 4 0,00	18.586,4 4 0,03	18.512,5 4 5,26	18.512,5 4 3,27
15°	VILLADIEGO	31 128	29.088,8 7	29.604,8 6 1,75	25.212,5 8 0,10	25.212,5 8 0,01	25.212,5 8 0,12	25.000,0 7 12,22	24.981,2 8 8,71
16°	VILLARCAYO	9 23	8.252,5 2	13.205,5 2 0,05	10.847,0 2 0,01	10.847,0 2 0,01	10.847,0 2 0,00	10.847,0 2 1,61	10.847,0 2 1,75
T.	**TOTAL PROVINCIA**	**437 2.330**	**415.382'7 86**	**435.708'0 79 27,31**	**374.863'8 100 2,57**	**374.863'8 100 0,16**	**372.826'5 99 2,82**	**366.342'9 101 197,86**	**363.302'0 104 247,07**

En cada celda se muestra el coste, el número de vehículos y el tiempo de computación en segundos

A continuación se muestran dos gráficos correspondientes a los costes y tiempos de computación de cada algoritmo:

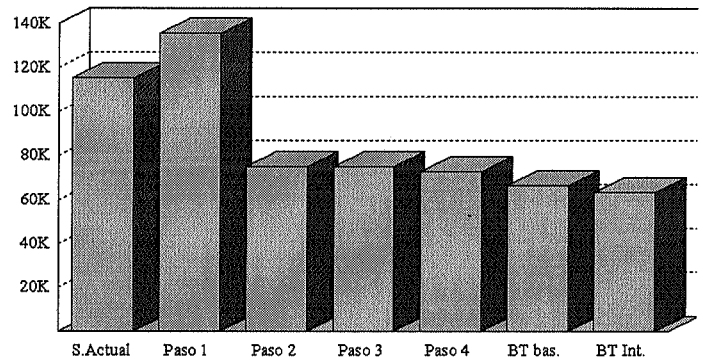


Figura 4. Costes de las soluciones obtenidas. El paso 2 supone la disminución más significativa. La búsqueda tabú también aporta reducciones importantes.

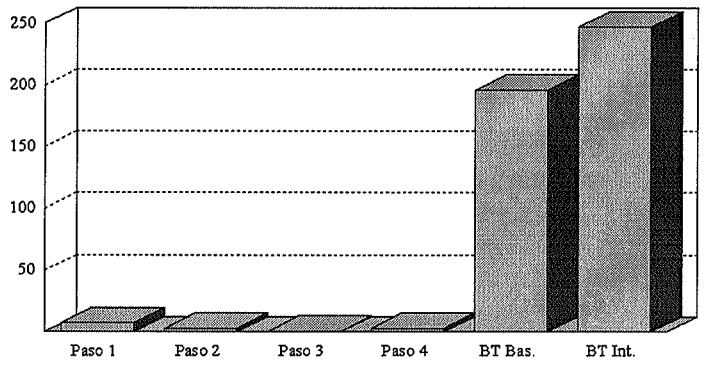


Figura 5. Tiempos de computación (en segundos) empleados. Insignificantes tiempos de las partes del algoritmo inicial, frente al empleado por la búsqueda tabú

Los algoritmos diseñados en este trabajo se han programado usando el compilador Borland Pascal 7.0 y Borland Delphi 3.0. El equipo informático donde se han programado y se han realizado las pruebas es un ordenador personal *Pentium MMX-200* Mhz.

Como se ve en la tabla de resultados, el algoritmo inicial aporta soluciones rápidamente (apenas poco más de 11 segundos para todos los problemas de secundaria), que mejoran las soluciones actuales en más de un 10% en el coste. Si se dispone de más tiempo de computación, la búsqueda tabú aporta soluciones que mejoran algo más de un 2% las del algoritmo inicial, y un 13% las soluciones actuales. Resultados similares se han registrado para primaria.

7. CONCLUSIONES Y REFLEXIONES

Es claro que los algoritmos propuestos (A.Inicial y A.Inicial+B.Tabú), globalmente dan soluciones muy interesantes en cuanto a ahorro de costes. Obsérvese, sin embargo, que existen instancias, como se refleja en la tabla anterior, para las que esta mejora es escasa. Incluso en algún caso, los algoritmos descritos aparentemente aportan soluciones peores que las utilizadas actualmente (Miranda y Villarcayo). Esto se debe a las siguientes causas:

- Para el cálculo de los caminos, distancias y tiempos se ha usado la red de carreteras obtenida de la cartografía digitalizada suministrada por el CNIG (Centro Nacional de Información Geográfica). Esta cartografía es muy completa y de mucha calidad. Sin embargo, en nuestro caso se ha detectado alguna imprecisión (falta de algún tramo o cruces no detectados) que puede dar lugar a que los caminos hallados, en algún caso concreto, sean más largos y costosos que los que se podrían usar realmente.
- Muchas de las rutas de las soluciones actuales, a la hora de la verdad tienen una duración mayor que el máximo programado de 60′. (En algún caso hasta de 90′, según responsables de la propia Dirección Provincial de Educación, con el consiguiente esfuerzo de intentar 'desdoblar' estas rutas.) Esto se debe a que en la planificación de las rutas se supuso unas velocidades medias mayores de las que se pueden conseguir en la realidad.
- Sin embargo, en nuestro caso se han supuesto unas velocidades bastante moderadas para cada tipo de carretera; concretamente 80 km/h para autopista y autovía, 70 km/h para nacional, 60 para autonómica de 1^{er} orden, 55 para autonómicas de 2^{o} orden, 50 para autonómicas de 3^{er} orden, 40 para carreteras sin revestimiento, 30 para carreteras de enlace y 20 para travesías. Obviamente, estas velocidades se pueden conseguir fácilmente en la realidad e incluso mayores. Lo importante es que las soluciones planificadas teóricamente puedan ser llevadas a cabo en la realidad con cierto margen. (Al contrario de lo que pasa en algunas soluciones usadas actualmente.)

Lo señalado en el párrafo anterior hace reflexionar en el siguiente punto: la reducción de costes es algo bueno, pero esta reducción se ha de conseguir manteniendo, e incluso mejorando, la comodidad de los trayectos (menores distancias y tiempos; mejores carreteras, etc.). En otras palabras, se ha de buscar la *racionalidad* en el sentido más amplio de la palabra, que es algo socialmente muy valioso tratándose del problema del transporte escolar.

RECONOCIMIENTO

Nuestro más sincero agradecimiento a los responsables de transporte de la Delegación Provincial del Ministerio de Educación y Ciencia de Burgos, por los datos suministrados y por las facilidades dadas en general.

REFERENCIAS

Backer (de), B. Furnon, V. Kilby, P., Prosser, P. and Shaw, P. (1997). «Solving Vehicle Routing Problems using Constraint Programming and Metaheuristics». *Journal of Heuristics*, 1 - 16.

Bodin, L.D. and Golden, B.L. (1981). «Classification in Vehicle Routing and Scheduling». *Networks*, 11, 2, 97-108.

Bullheimer, B., Harti, R.F. and Strauss, C. (1997). *Applying the Ant System for the Vehicle Routing Problem.* 2nd Metaheuristics International Conference (MIC-97), Sophie-Antipolis, France, July 1997.

Campos, V. y Mota, E. (1995). «Metaheurísticos para el CVRP». *XXII Congreso Nacional de Estadística e Investigación Operativa.* Sevilla, Noviembre 1995.

Clarke, G. and Wright, J.W. (1964). «Scheduling of Vehicles from a Central Depot to a Number of Delevery Points». *Oper. Res.*, 12, (1964), 568-581.

Desrochers, M., Lenstra, J.K., Savelsbergh, M.W.P. and Soumis, F. (1988). «Vehicle Routing with Time Windows: Optimization and Approximation». In *Vehicle Routing: Methods and Studies*, (Studies in Management Sciences and Systems, vol. 16), eds: Golden, B.L. and Assad, A.A., Nort-Holland, 65-84.

Fisher, M.L. y Jaikumar, R. (1981). «A Generalized Assignment Heuristic for Vehicle Routing». *Networks*, 11, 2, 109-124.

Gendreau, M., Hertz, A. and Laporte, G. (1991). «A Tabu Search Heuristic for Vehicle Routing Problem». *Report CRT-777.* Centre de Recherche sur les Transports. Univ. Montréal.

Gendreau, M., Hertz, A. and Laporte, G. (1994). «A Tabu Search Heuristic for Vehicle Routing Problem». *Management Sci.*, 40 (10), 1276-1290.

Glover, F. (1989). «Tabú Search: Part I». *ORSA Journal on Computing*, 1, 190-206.

Glover, F. (1990). «Tabú Search: Part II». *ORSA Journal on Computing*, 2, 4-32.

Glover, F. (1996). *Búsqueda Tabú en Optimización Heurística y Redes Neuronales.* Adenso Díaz (coordinador). Paraninfo. Madrid. pp. 105-143.

Glover, F. y Laguna, M. (1993). *Tabu Search in Modern Heuristic Techniques for Combinatorial Problems.* C. Reeves, ed., Blackwell Scientific Publishing, pp. 70-141.

Glover, F. y Laguna, M. (1997). *Tabu Search.* Kluver Academic Publishers, Boston.

Glover, F. y Laguna, M. (1999). *Tabu Search, aparecerá en Handbook of Applied Optimization*, P.M. Paradalos and M.G.S. Resende (eds). Oxford Academic Press.

Haouari, M., Dejax, P. et Desrochers, M. (1990). «Les Problems de Tournées avec Contraintes des Fenênetres de Temps: L'Etat de l'Art». *Recherche Operationnelle/Operations Research*, 24, 3, 217-244.

Kilby, P., Prosser, P. and Shaw, P. (1997). *Guided Local Search for the Vehicle Routing Problem.* 2nd Metaheuristics International Conference (MIC-97), Sophie-Antipolis, France, July 1997.

Kontoravdis, G. and Bard, J.F. (1995). «A Grasp for the Vehicle Routimg Problem with Time Windows». *ORSA Journal on Computing*, 7, 10-23.

Laporte, G. (1992). «The Vehicle Routing Problem: An overview of exact and approximate algorithms». *European Journal of Operations Research*, 59, 345-358.

Laporte, G., Desrochers, M. and Nobert, Y. (1984). «Two Exact Algorithms for the Distance-Constrained Vehicle Routing Problem». *Networks*, 14, 161-172.

Laporte, G., Nobert, Y. and Desrochers, M. (1985). «Optimal routing under capacity and distance restrictions». *Operations Research*, 33, 1075-1073.

Laporte, G. and Osman, I.H. (1995). «Routing Problems: A Bibliography». *Ann. Oper. Res.*, 61, 227-262.

Lenstra, J.K. and Rinnoy Kan, A.H.G. (1981). «Complexity of Vehicle Routing and Scheduling Problems». *Networks*, 11, 2, 221-228.

Lin, S. (1965). «Computer Solutions to the Traveling Salesman Problem». *Bell Syst. Tech. Jou.*, 44, 2245-2269.

Lin, S. y Kernighan, B.W. (1973). «An Effective Heuristic Algorithm for the Traveling Salesman Problem». *Operations Researsch*, 20, 498-516.

Nurmi, K. (1991). «Traveling Salesman Problem Tools for Microcomputers». *Computers & Ops. Res.*, 18, 8, 741-749.

Or, I. (1976). *Traveling Salesman Type Combinatorial Problems y their Relations to the Logistics of Blood Banking.* Ph. Thesis, Dpt. of Industrial Enginnering y Management Sciences, Northwestern Univ.

Osman, I.H. (1993). «Metastrategy Simulated Annealing and Tabu Search Algorithms for the Vehicle Routing Problem». *Annals of Operations Research*, 41, 421-451.

Pacheco, J. y Delgado, C. (1996). *Adaptación del Algoritmo de Or al VRPTW con Carga y Descarga simultánea.* X Reunión Asepelt-España, Albacete, Junio 1996.

Pacheco, J. y Delgado, C. (1997). «Problemas de Rutas con Ventanas de tiempo y carga y Descarga simultánea: Diseño de Filtros para algoritmos de intercambio (caso de un sólo vehículo)». *Estudios de Economía Aplicada*, 7 , 79-100.

Pacheco, J. y Delgado, C. (1999). «Diseño de Metaheurísticos híbridos para Problemas de Rutas con Flota Heterogénea: Concentración Heurística». Aceptado para su publicación en *Estudios de Economía Aplicada.*

Potvin, J.Y. and Bengio, S. (1994). «A Genetic Approach to the Vehicle Routing Problem with Time Windows». *Technical Report CRT-953*, Centre de Recherche sur les Transports. Univ. Montréal.

Potvin, J.Y., Kervahut, T., García, B.L. and Rousseau, J.M. (1993). «A Tabu Search Heuristic for Vehicle Routing Problem with Time Windows». *Report CRT-777. Management Sci.*, 40 (10), 1276-1290.

Rego, C. (1998). «A Subpath Ejection Method for the Vehicle Routing Problem». *Management Science*, 44, 10, 1447-1459.

Rochat, Y. and Taillard, E.D. (1995). «Probabilistic Diversification and Intensification in Local Search for Vehicle Routing». *Journal of Heuristics*, 1 (1), 147-167.

Rosing, K.E. (1997). *Heuristic Concentration: An Introduction with Examples. The Tenth Meeting of the European Chapter on Combinatorial Optimization.* Tenerife. Spain. May, 1997.

Rosing, K.E. and Revelle, C.S. (1997). «Heuristic Concentration: Two Stage solution Construction». *European Journal of Operational Research*, 97, 75-86.

Rosing, K.E., Revelle, C.S., Rolland, E., Schilling, D.A. and Current, J.R. (1998). «Heuristic Concentration and Tabu Search: A head to head comparison». *European Journal of Operational Research*, 104, 93-99.

Solomon, M.M. (1987). «Algorithms for the Vehicle Routing and Scheduling Problem with Time Windows Constraints». *Operations Research* 35, 254-265.

Taillard, E., Badeau, P., Gendreau, M., Guertain, F. and Potvin, J.Y. (1995). *A new Neighbourhood structure for the Vehicle Routing Problem with Time Windows.* Technical Repport CRT-95-66, Centre de Recherche sur les Transports. Univ. Montréal.

Taillard, E., Badeau, P., Gendreau, M., Guertain, F. and Potvin, J.Y. (1997). «A Tabu Search heuristic for the Vehicle Routing Problem with Time Windows». *Transportation Science*, 31, 170-186.

Thangiah, S.R., Osman, I.H. and Sun, T. (1994). «Hybrid Genetic Algorithm, Simulated Annealing, and Tabu Search methods for the Vehicle Routing Problem with Time Windows». *Working paper UKC/OR94/4*, Institute of Mathematics and Statistics, University of Kent, Canterbury.

Thangiah, S.R., Vinayagamoorty, R. and Sun, T. (1993). «Vehicle Routing Problem with Time Deadlines using Genetic and Local Algorithms». In *5th International Conference on Genetic Algorithms.*

ENGLISH SUMMARY

ALGORITHM DESIGN FOR SCHOOL TRANSPORTATION. APPLICATION TO BURGOS PROVINCE

J.A. PACHECO
A. ARAGÓN
C. DELGADO
Universidad de Burgos*

The issue of school transportation in Burgos is particularly significant since it is a large province with many inhabited areas widely dispersed and only scantly populated. In this study, the authors = attempts to provide a solution to this problem in the most reasonable way possible are presented. It should be noted that, in this context, the term 'reasonable' does not refer solely to the minimization of total transport costs, but also to the consideration of certain aspects such as the amount of time students are in the vehicle, the selection of good highways, etc., particularly in light of certain recent incidents. Algorithms based principally on Local Search are proposed and described and then a Tabu Search algorithm is also suggested as an option depending on the computation time. Likewise, the results obtained from the data from the current year are also shown.

Keywords: Routes, school transportation, local search, tabu search, intensification, heuristic concentration

AMS Classification: 90B20

---

* Departamento de Economía Aplicada. Universidad de Burgos. Parralillos, s/n. 09001 Burgos.

-Received June 1999.
-Accepted January 2000.

Consider the transportation problem of a group of students that need to be picked up from a series of geographically distributed places and taken to an educational center. The number 1 can indicate the educational center, 2, ... ,n points corresponding to the places where the children are picked up, and $q(i)$ the number of students that are picked up at each point $i, i = 2, \ldots, n$. In order to meet these demands, the legislation authorizes different types of vehicles, each one with a different number of seats. Each student should not be in route more than a determined maximum amount of time, which is indicated by *tmax*, and the route should be completed before the start of classes in the moment *tinicio*. The distances d_{ij} and the times t_{ij} between each pair of points i, j are known. The number of types of authorized vehicles is indicated by *ntipos* and the capacities by *capactipo(i)*, $i = 1, \ldots,$ *ntipos*.

A group of routes needs to be designed with a minimum total cost making sure that: the driving hours and times and respected, and that the number of students to be transported on each route does not exceed the total capacity of the vehicle assigned to that route. Considered in this way, this model is a specific case of the already established Vehicle Routing Problem (VRP), or to be more precise of the Vehicle Routing Problem with Time Windows (VRPTW), since there are time restrictions.

To design an algorithm for this problem, we have chosen to use a heuristic technique which provides a good solution in a more reasonable computation time. This strategy is based on two parts: the first consists basically in a series of Local Search processes which give quick solutions; the second (optional depending on the available calculation time) is based on a process of Tabu Search that includes a new Intensification method. Therefore, the proposed algorithm is a technique made up of various parts or subalgorithms based, for the most part, on neighborhood movement.

In the following two sections, the different definitions used here of neighborhood are presented. In the four section, he general algorithm structure is described. In the fifth section an algorithm based on a Tabu Search process is described which complements the preceding algorithm and which on many occasions improves the results. Finally, in the sixth section, the results obtained by the different algorithms and subalgorithms are described, as are the computation times used for a series of experiments on real problems with data from Burgos Province.

Estadística Oficial

QÜESTIIÓ, vol. 24, 1, p. 85-110, 2000

DIVERSITAT I COMPLEMENTARIETAT DE LES FONTS ESTADÍSTIQUES*

ÀLEX COSTA
Institut d'Estadística de Catalunya (Idescat)

Es presenta la diversitat de les fonts generals d'estadística oficial, posant de relleu la importància de censos, enquestes, aprofitament de registres administratius i estimacions comptables. Es justifica aquesta diversitat per les diferents prestacions de cada forma de producció, des del punt de vista de la puntualitat i de la quantitat d'informació i també des del punt de vista de la fiabilitat dels resultats. Tot seguit es passa a l'àmbit de les fonts regionals on, de nou, es destaca la diversitat i la complementarietat de les mateixes. En aquest punt destaca la complementarietat en els processos de producció d'informació estadística i en els resultats finals. Pel que fa als processos, es presenta una tipologia de formes de col·laboració entre l'administració central (INE) i l'autonòmica (Idescat). En la complementarietat dels resultats es tornen a fer servir els mateixos conceptes emprats per a les fonts generals: puntualitat, quantitat d'informació i fiabilitat.

Diverseness and complementariness of statistical sources

Paraules clau: Fonts estadístiques, diversitat de fonts, complementarietat de fonts, correspondència, coherència, territori/temps/conceptes, col·laboració institucional

Classificació AMS: 62P25

---

*Aquest treball és el resultat d'una ponència presentada a les «Jornades internacionals sobre generació d'informació estadística: qualitat i limitacions», organitzades a Barcelona el novembre de 1998 per la xarxa temàtica *Enquestes i qualitat de la informació estadística*. El text és una ampliació de l'original feta amb ocasió d'una conferència sobre el mateix tema a la *Escuela de Estadística Gonzalo Arnáiz*, en un seminari sobre estadística regional celebrat a la Universitat Internacional Menéndez Pelayo (Santander, juliol de 1999).

– Rebut el setembre de 1999.
– Acceptat el novembre de 1999.

INTRODUCCIÓ

La major part dels llibres de text d'estadística, quan tracten el tema de la producció de les dades estadístiques, fan referència a una única forma de producció. El típic procés d'obtenció de dades, l'enquesta per mostreig, queda presentat en molts textos com una sèrie de fases de la investigació estadística. Aquestes són: 1) identificació del tema objecte d'estudi, 2) disseny del qüestionari, 3) determinació del directori, 4) disseny de la mostra, 5) treball de camp, 6) depuració, imputació i validació de les microdades, 7) elevació de les dades de la mostra, 8) tabulació dels resultats i 9) anàlisi i difusió.

Aquesta visió sobre la forma de generar informació estadística s'adapta bé a la investigació aplicada i planteja pocs dubtes en l'entorn de treball universitari. En canvi, després d'uns anys a l'Institut d'Estadística de Catalunya, crec que aquesta concepció més aviat uniformista pot ser matisada d'una manera significativa. En aquests moments, si s'observa el panorama de l'estadística oficial rellevant per a una societat, resulta clar que existeixen altres importants formes de producció, tant en l'àmbit de les fonts generals com en les fonts regionals.

En aquesta exposició es presenta, en la primera part, la diversitat i la complementarietat de les formes de producció i de les fonts estadístiques generals. S'apreciarà que la diversitat identificada no és inútil, sinó que respon a unes distintes prestacions i limitacions de cada font. En aquest sentit, es pot dir que la diversitat de fonts està justificada per la complementarietat entre les diverses formes de producció. Aquesta complementarietat de les fonts estadístiques generals pot ser explicada des de dues perspectives, que seran objecte d'atenció en aquests treball. En primer lloc, la quantitat i la puntualitat de la informació estadística i, en segon lloc, la fiabilitat de la informació estadística.

A la segona part del treball es consideren les fonts regionals. Novament es torna a presentar la diversitat i la complementarietat de fonts. En aquest punt resulta determinant la convergència d'interessos i activitats estadístiques de diversos organismes estadístics, de l'administració central i de les comunitats autònomes. La complementarietat s'analitza des del punt de vista dels processos i dels productes generats. Es podrà observar que la quantitat i puntualitat de la informació i la fiabilitat tornen a jugar un paper rellevant en el diagnòstic global del sistema d'estadística regional.

Aquestes reflexions, si bé tenen en comú el fet de la diversitat de fonts i la seva justificació, se situen en àmbits clarament diferenciats i presenten argumentacions autònomes. Les relacionades amb les fonts generals són idees més abstractes, fins i tot filosòfiques en tractar el tema de la fiabilitat. En canvi, en la segona part, relativa a les fonts regionals, es tracta un tema d'organització del sistema estadístic i està formulat en bona part des d'una òptica més institucional. En tot cas, les apreciacions del text es fan a títol personal, i no representen necessàriament una posició oficial de l'Institut d'Estadística de Catalunya.

LA DIVERSITAT DE LES FONTS ESTADÍSTIQUES

A l'estadística oficial coexisteixen diverses formes de producció d'informació: les operacions censals, les investigacions per mostreig, l'aprofitament de registres administratius i les estimacions de síntesi comptable. La realitat actual de la producció estadística mostra que tots aquests diferents sistemes de producció són importants, responen a uns diferents processos de treball i, en bona part per aquesta mateixa raó, tenen limitacions i prestacions clarament diferenciades. Veurem aquesta diversitat d'una forma molt senzilla i empírica, fent una estadística sobre les estadístiques que en l'actualitat conté el catàleg de publicacions de l'Institut d'Estadística de Catalunya.

En el catàleg de l'Institut d'Estadística de Catalunya, a més dels llibres de recopilació com l'Anuari Estadístic de Catalunya, i també del que nosaltres coneixem com estadística instrumental (publicacions sobre nomenclatures o metodològiques), trobem uns àmbits temàtics de producció estadística pròpiament dita: estadística demogràfica, estadística social, estadística econòmica estructural i estadística de conjuntura econòmica. Per tal de saber de quina manera es presenta la diversitat de processos esmentada no hi ha res millor que fer una estadística d'estadístiques.

Taula 1. Estadística sobre les formes de producció estadística. Relació de publicacions de l'Idescat

Estadística demogràfica		**Forma de producció**
1)	Moviment natural de la població	Registre adm.
2)	Moviments migratoris	Registre adm.
3)	Estimacions de població	Síntesis
4)	Estadística de població 96 (5 vol)	Cens
5)	Cens de Població 91 (17 vol)	Cens
6)	Padró d'habitants 86 (3 vol)	Cens
Estadística social		**Forma de producció**
1)	Estadística de biblioteques	Enquesta
2)	Estadística de finançament i despeses de l'ensenyament privat	Enquesta
3)	Cens d'habitatges 1991	Cens
4)	Cens d'edificis 1990	Cens
5)	Els comptes de la protecció social a Catalunya	Síntesi
6)	Eleccions municipals a Catalunya	Registre adm.

Taula 1. (Cont.)

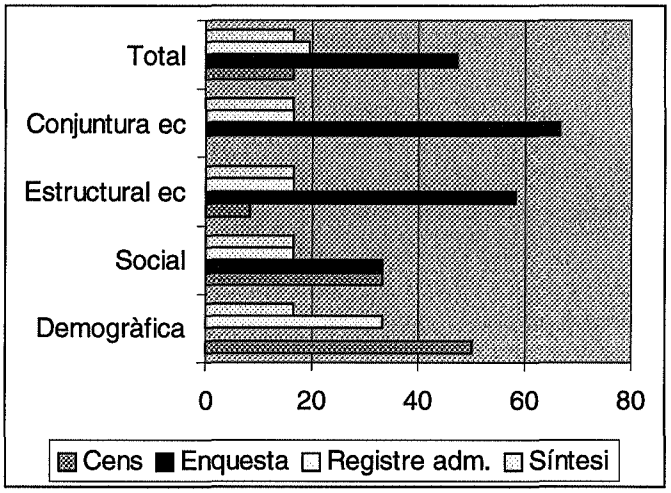
Estadística econòmica estructural		**Forma de producció**
1)	Evolució de les principals macromagnituds de l'economia catalana	Síntesi
2)	Macromagnituds bàsiques de les economies locals i comarcals	Síntesi
3)	Comptes de les administracions públiques	Enquesta
4)	Estadística de les explotacions agràries	Enquesta
5)	Cens agrari (4 vol)	Cens
6)	Estadística de producció i comptes de la indústria	Enquesta
7)	Estadístiques i comptes del comerç interior	Enquesta
8)	Estadístiques i comptes de l'hoteleria	Enquesta
9)	Estadística i comptes del transport públic de mercaderies	Enquesta
10)	Comerç amb l'estranger (import/export)	Registre adm.
11)	Mercat de treball (ampliació EPA)	Enquesta
12)	Localització de l'activitat econòmica (aprofitament IAE)	Registre adm.
Conjuntura econòmica		**Forma de producció**
1)	Producte interior brut de Catalunya trimestral	Síntesi
2)	Índex de producció industrial (IPI)	Enquesta
3)	Índex de Preus Industrials (IPRI)	Enquesta
4)	Indicadors d'Activitat de la Construcció	Enquesta
5)	Indicadors d'Activitat Hotelera	Enquesta
6)	Viatges dels catalans	Enquesta
7)	Viatges dels espanyols d'altres comunitats a Catalunya	Enquesta
8)	Índex de vendes en grans superfícies (IVGS)	Enquesta
9)	Comerç amb l'estranger (import/export)	Registre adm.
10)	Contingut tecnològic del comerç amb l'estranger	Registre adm.
11)	Indicadors de posició competitiva	Síntesi
12)	Clima exportador	Enquesta

De forma més sintètica, aquesta informació condueix als resultats següents:

Taula 2. Estadística sobre les formes de producció estadística. Resultats en percentatge. Publicacions Idescat.

	Cens	Enquesta	Registre adm.	Síntesi	Total
Demogràfica	50,0	–	33,3	16,6	100
Social	33,3	33,3	16,6	16,6	100
Econòmica estructural	8,3	58,3	16,6	16,6	100
Conjuntura econòmica	–	66,6	16,6	16,6	100
Total	16,6	47,2	19,4	16,6	100

En aquests resultats, tant a la taula com al gràfic, es pot veure que malgrat la gran significació de les enquestes per mostreig, que són indubtablement la forma de producció predominant (pràcticament suposen el 50% de les publicacions), els altres sistemes no són marginals, sinó que tenen una presència rellevant. Per tant, aquesta informació sobre les formes de producció estadística, sense negar el protagonisme de les enquestes per mostreig, trenca amb la visió uniforme i avala la idea de la diversitat que s'ha anunciat al principi del treball. A partir d'aquest punt la qüestió quedarà centrada en saber si aquesta diversitat és positiva i està justificada o és inútil i redundant.

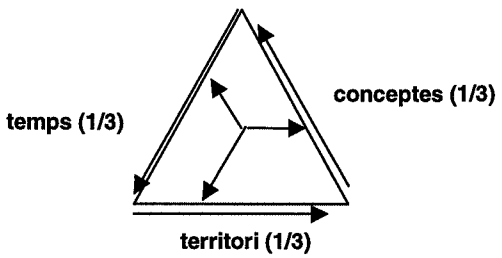


Gràfic 1. Formes de producció per àrees temàtiques

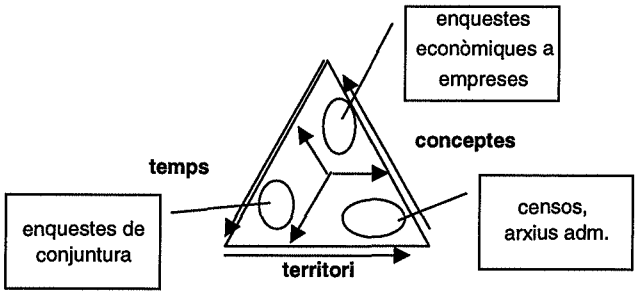
LA COMPLEMENTARIETAT DE LES FONTS ESTADÍSTIQUES (I): QUANTITAT I PUNTUALITAT DE LA INFORMACIÓ

La màxima quantitat i la puntualitat de la informació estadística són objectius naturals i legítims. La quantitat de la informació pot ser expressada en termes d'un òptim detall conceptual i d'una forta desagregació territorial. L'experiència quotidiana ens mostra que aquestes tres dimensions, puntualitat, detall conceptual i desagregació territorial, rarament es troben en nivells molt satisfactoris en una única estadística. Normalment, una enquesta amb un gran detall conceptual (per exemple, una enquesta a empreses), tindrà un cert desfassament temporal i difícilment tindrà una gran aproximació al territori. D'altra banda, la puntualitat, tan apreciada en el seguiment de la conjuntura, es paga de vegades amb una pobre aproximació territorial i amb un detall conceptual limitat. Finalment, hi ha algunes operacions estadístiques, com els censos de població, que tenen el seu gran valor en la seva capacitat de desagregació territorial, i no són tan satisfactoris en el detall conceptual o en la seva puntualitat.

Això no vol dir que no s'hagi de procurar que les enquestes a les empreses siguin puntuals, o que no es puguin explotar territorialment. Tampoc es pretén que les estadístiques de conjuntura tinguin que ser necessàriament molt deficients des d'un punt de vista de detall conceptual. El que s'ha volgut expressar és que, a la pràctica, existeix un cert *trade-off* entre aquestes tres virtuds d'una estadística. Aquesta situació pot ser visualitzada mitjançant un gràfic en forma de triangle amb projeccions en cada un dels costats. Una estadística perfectament centrada en l'espai triangular tindria una puntuació de 1/3 per a cada costat, amb uns eixos que mesuren el seu valor en puntualitat, detall conceptual i desagregació territorial. Aquesta idea és mostrada en el gràfic 2. La regla que fixa aquest triangle pot expressar-se de la següent forma: per a una quantitat fixa de recursos i tecnologia estadística (metodologia i organització), millorar significativament un aspecte dels tres es pagarà amb una pèrdua en els altres. Altra forma d'enunciar aquesta mateixa idea: cada operació estadística té una ubicació en aquest espai, prioritzant un o altre aspecte. Per exemple, tal com mostra el gràfic 3, l'enquesta industrial a les empreses, el cens de població i l'índex de vendes en grans superfícies, tenen els seus punts forts en cada cas en un dels costats del triangle.



Gràfic 2. Tres prestacions d'una estadística: la regla del triangle



Gràfic 3. Tres operacions primàries i tres prestacions: mostra de la regla del triangle

Aquesta complementarietat de prestacions en la quantitat d'informació i la puntualitat és la primera de les justificacions de la diversitat de les fonts. Pel que fa a les modalitats de producció, podem veure a la Taula 3 la complementarietat de les formes de producció primària.

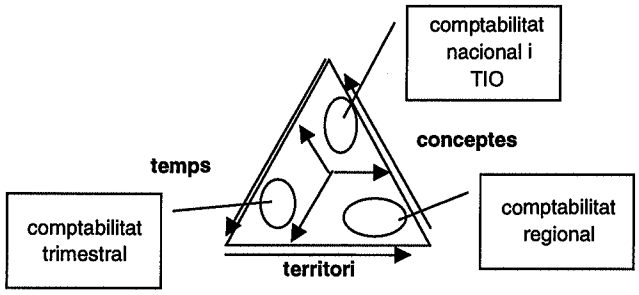
Taula 3. Prestacions de les formes de producció primària d'estadístiques

	Puntualitat temporal	Desagregació territorial	Detall conceptual
Censos	–	++	+-
Enquestes estructurals	+-	–	++
Enquestes conjunturals	++	–	+-
Registre adm (i)	+-	++	–
Registre adm (ii)	++	+-	–

La regla del triangle no és una regla fixa i sense excepcions. Les dades duaneres de comerç internacional extracomunitari derivades del DUA són molt puntuals i també tenen un fort detall en la seva nomenclatura. L'enquesta de població activa és francament bona tant en detall conceptual com en puntualitat, encara que en desagregació territorial sí que està afectada per la regla del triangle. Malgrat la casuística que pugui detectar-se, la regla del triangle és suficientment vàlida amb caràcter general, i a més resulta molt didàctica per a l'usuari d'estadístiques, especialment si aquest té algunes prioritats ben marcades, com és el cas dels usuaris d'estadístiques regionals.

En l'àmbit de la síntesi comptable també es dóna aquesta regla de contraposició entre la puntualitat, el detall conceptual i la desagregació territorial, com és normal tenint en compte que la síntesi es nodreix de la informació. En el gràfic 4 es pot veure aquesta situació de les estadístiques de comptabilitat.

La regla del triangle és la primera expressió que la diversitat de les fonts i dels processos de producció en estadística, i en la mesura en què es tradueix en una complementarietat informativa, és una diversitat positiva. Aquesta diversitat permet tenir estratègies d'aproximació a la realitat adaptades a les òptiques preferents de la nostra investigació, quan una informació puntual i a escala conceptual i territorial 1:1 resulta inassolible.



Gràfic 4. Tres operacions de comptabilitat i tres prestacions: mostra de la regla del triangle.

LA COMPLEMENTARIETAT DE LES FONTS ESTADÍSTIQUES (II): FIABILITAT DE LA INFORMACIÓ

Existeix un segon aspecte, possiblement tan rellevant com el primer, encara que més complex, que també justifica la diversitat de fonts en termes de complementarietat. En aquest cas la diversitat considerada no se situa en l'àmbit dels diferents sistemes de producció primària o de síntesi considerades per separat, sinó que precisament es refereix a la coexistència d'informació primària i d'informació de síntesi comptable. L'àmbit en què es detecta la complementarietat no és el de la quantitat o puntualitat de la informació, sinó el de la fiabilitat.

A l'estadístic els usuaris sovint li pregunten per la veritat de la informació estadística. Aquesta questió sempre causa una certa incomoditat. No resulta fàcil contestar preguntes com si és veritat que la taxa d'atur, tal com és estimada per l'EPA per al quart trimestre de 1998, va ser del 13,6% a Catalunya, o si és veritat que el creixement del PIB de l'economia catalana és d'un 4,1% entre 1997 i 1998. Es pot contestar des d'un punt de vista professional, dient que aquests són els resultats d'aplicar unes metodologies i tècniques actualment vigents en l'activitat estadística o, en termes més científics, fent referència a que el resultat en realitat hauria de ser un interval de confiança. Aquestes respostes, però, presenten encara algun problema.

La suma de metodologies i tècniques hauria de fer convergents els resultats de les estadístiques. La percepció de l'usuari freqüentment no és aquesta. Per a l'any 1996 (any de l'estadística de població a Catalunya), per a l'economia catalana disposem de quatre mesures d'ocupació industrial: la comptable, la de l'EPA (mitjana anual), la de l'estadística de població (cens) i la dels registres administratius de la Seguretat Social. Tots aquests resultats, o com a mínim els tres primers, han estat obtinguts mitjançant la metodologia i les tècniques de la professió estadística. Aleshores, ¿perquè difereixen més enllà del que és acceptable des del punt de vista estadístic? Aquesta pregunta pot ser contestada fent referència a diferències de concepte i diferències de procés: errors de mostreig en l'enquesta, no exhaustivitat del cens, aspectes conceptuals en la comptabilitat. Ara bé, un cop justificades les diferències i la diversitat d'aproximacions es pot insistir: quin és el resultat més fiable? Quin s'aproxima més a la veritat?

Per tal de resoldre aquestes preguntes resulta natural preguntar-se pel significat dels termes que en elles apareixen i, de forma principal, d'un concepte venerable i que, en principi, pot semblar molt lluny de la terminologia professional o científica de l'estadístic modern. Em refereixo al concepte de «veritat». Pot ser instructiu per als estadístics (i per als usuaris d'estadística) conèixer esquemàticament la trajectòria del concepte de «veritat» al llarg de la història del pensament.

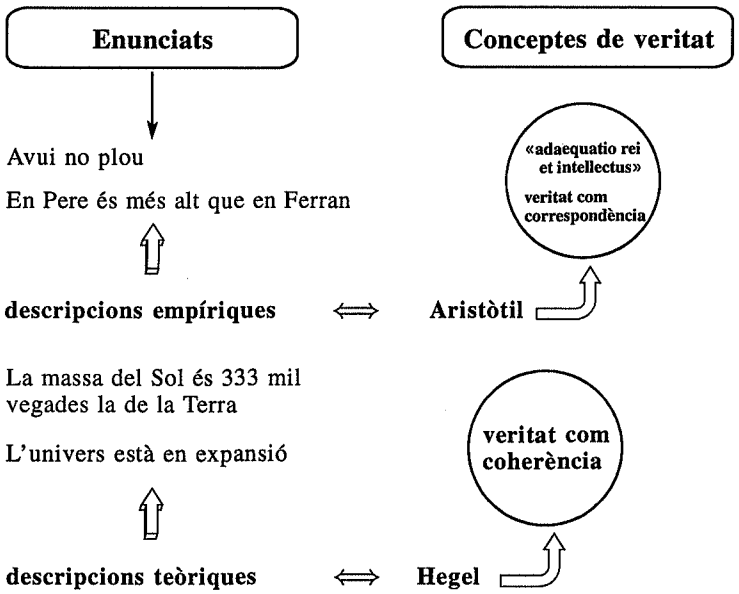
El filòsof Josep Ferrater Mora, en el seu diccionari enciclopèdic, ens ensenya que a partir d'Aristòtil, el concepte de «veritat» ja té un sentit modern, com a atribut d'un enunciat o d'un pensament. L'afirmació aristotèlica és: «dir del que és que no és, o del que no és que és, és fals; dir que el que és que és, i del que no és que no és, és la veritat». La sentència aristotèlica dóna cobertura a diferents concepcions de veritat, però la més immediata i predominant és la idea que la veritat és la *correspondència* entre el pensament i la realitat. Aquesta definició és recollida per l'aforisme escolàstic «*adaequatio rei et intellectus*». Aquest plantejament s'adapta molt bé a enunciats empírics senzills, com «avui no plou» o «En Pere és més alt que en Ferran». No s'adapta tant bé, en canvi, als enunciats formals, com «2+3=5» o més teòrics, com «la massa del sol és 333 mil vegades la de la Terra» o «l'univers està en expansió».

Precisament sobre la base d'enunciats més abstractes, el filòsof alemany Hegel trenca amb la definició clàssica de la veritat com a correspondència i formula una concepció de la veritat no com atribut d'un enunciat singular, sinó com a atribut d'un conjunt de coneixements coherents. La teoria mecànica de Newton és veritat, mentre que la física de Ptolomeo és falsa. La veritat no és parcial o local, sinó que es refereix a concepcions globalitzadores de la realitat. Té més a veure amb la *coherència* d'un conjunt de coneixements que amb la correspondència d'una afirmació i uns fets específics.

Aquestes dues concepcions de la veritat s'han mantingut vigents al llarg de la història fins arribar al segle XX. La idea dels enunciats «falsables» de Popper queda més propera a la concepció de veritat com a correspondència, encara que ara la veritat és un

estat provisional d'un enunciat (veritat equival a «encara no falsat»). Per la seva part, conceptes com el de «paradigma» de Kuhn encaixen millor amb la idea de veritat com a concepció global coherent de la realitat.

Aquesta història del concepte de «veritat», presentada aquí de forma telegràfica, mostra una concepció dual del concepte: la veritat com a «correspondència» davant la veritat com a «coherència». Resulta clar que aquests dos conceptes de veritat són complementaris i els dos són útils: en funció del tipus d'enunciat considerat, pot resultar més natural o aplicable un o altre concepte. S'ha intentat mostrar aquesta complementarietat en el gràfic adjunt.

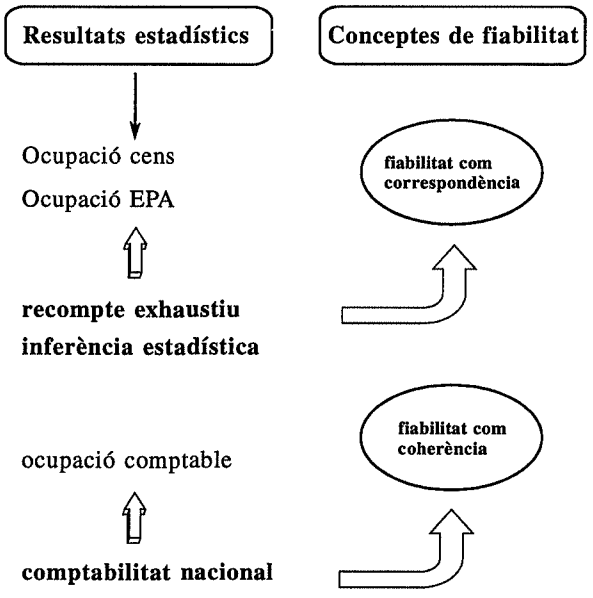


Gràfic 5. Concepció dual del concepte de «veritat»: correspondència *versus* coherència

Aquesta dualitat del concepte de veritat sembla afectar el camp de la informació estadística a través de la idea de fiabilitat. Algunes informacions estadístiques estan molt vinculades a realitats específiques. Aquest és el cas d'un cens o d'un registre administratiu. També una enquesta per mostreig té una estreta relació o correspondència amb la realitat, encara que potser no tant com la d'una investigació exhaustiva (ja que requereix la teoria de la inferència estadística). En canvi, la comptabilitat és un sistema molt més teòric, en el qual la relació amb la informació empírica no és sempre tan directa i transparent però, a canvi, té al seu favor precisament la coherència de la informació

que produeix. La seva virtud és que fa compatibles informacions diverses, presentant la realitat econòmica d'una forma coherent, on ocupació, consum, producció i rendes, per exemple, formen part d'una mateixa aproximació a la realitat econòmica. A la comptabilitat s'afegeix una nova estimació de la població ocupada que no acaba de coincidir amb cap d'altra, però que les aprofita totes. Per a molts estadístics aquest resultat, que té la fiabilitat de la coherència, potser és el més vàlid.

Resultaria un error contraposar una forma de fiabilitat a l'altra, ja que res no val la coherència si no té vincles amb la informació empírica. Precisament per aquesta raó, trobem aquí, novament, una diversitat d'aproximacions estadístiques que troba la seva justificació en la complementarietat. La informació estadística primària i la informació comptable són complementàries en virtut de la seva fiabilitat. La informació primària gaudeix d'una fiabilitat derivada de la seva correspondència directa amb una base empírica concreta. És una fiabilitat que pot vincular-se al concepte aristotèlic de veritat. En canvi, la informació comptable perd aquesta relació directa, però ens ofereix una informació coherent, raó per la qual podem dir que presenta una fiabilitat pròxima al concepte hegelià de veritat, més adaptat a les descripcions teòriques. En el gràfic 6 s'intenta esquematitzar aquesta situació.



Gràfic 6. Complementarietat a la fiabilitat estadística: informació primària *versus* comptabilitat

LA DIVERSITAT DE LES FONTS ESTADÍSTIQUES REGIONALS

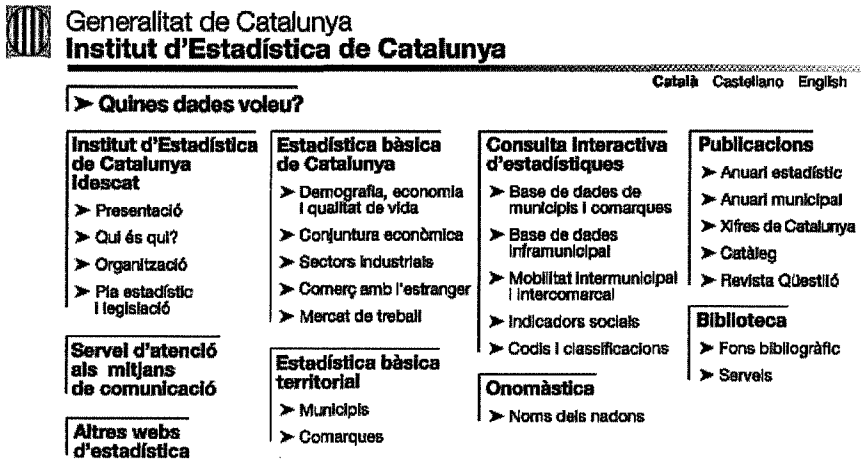
Tot el que s'ha dit fins aquí sobre la diversitat de les fonts estadístiques generals i sobre la seva complementarietat pot considerar-se universal, generalitzable a qualsevol sistema estadístic. Ara bé, en un àmbit estadístic com el nostre, on l'estadística regional és objecte d'interès i competència legítima tant de l'organisme central d'estadística, l'INE, com dels organismes regionals de les comunitats autònomes, s'afegeix un nou eix de diversitat i de complementarietat, relacionat amb la pluralitat dels organismes productors. En aquest punt cal plantejar-se quina és la naturalesa de la diversitat de fonts estadístiques regionals produïda per aquesta convergència d'administracions estadístiques i haurem també de plantejar-nos la seva complementarietat.

Un primer element bàsic d'aquest tema és el marc jurídic. Aquest marc fixa la competència exclusiva de les estadístiques d'interès estatal a l'administració central i la competència exclusiva de les estadístiques d'interès autonòmic a les comunitats autònomes. Ara bé, l'estadística oficial no és un àmbit d'investigació lliure, sinó que està determinat per un consens internacional (en el nostre cas bàsicament europeu) que defineix un catàleg bastant tancat sobre quines són les estadístiques interessants per al seguiment d'una realitat econòmica o demogràfica. Això fa que el solapament d'interessos entre l'administració central i les regionals sigui pràcticament total, amb algunes diferències d'èmfasi o d'enfocament. Davant aquesta situació i a la vista d'una estadística central ja en marxa en el moment de desenvolupar l'estadística oficial de les comunitats autònomes, s'identifiquen tres estratègies generals: 1) minimització d'activitats de producció, a favor de la difusió, 2) duplicació d'operacions, 3) operacions complementàries.

La gran major part de les comunitats autònomes han optat bàsicament per la primera o la tercera via i només en casos molt comptats, s'ha plantejat la segona via de la duplicació. Aquest segon camí és possible, encara que costós i complex, i des d'un punt de vista de cost/benefici social resulta difícil de justificar. En el cas de l'Idescat, s'ha donat prioritat a l'estratègia primera i tercera.

Pel que fa a la rellevància de la difusió cal considerar en primer lloc el servei de documentació de la biblioteca, amb més d'11 mil consultes ateses al llarg de 1998, de les quals un 20% són per correu/fax, un 25% per visita personal i un 55% per telèfon. Aquest servei es complementa amb les peticions a mida, de les quals es van atendre l'any passat 722 usuaris i que són sol·licituds de major complexitat, resoltes per unitats pròpiament estadístiques. Aquí es troben des de tabulacions relativament senzilles fins a cessió de microdades amb tractament de preservació del secret estadístic. Una tercera via de difusió és el servei de premsa, amb 504 comandes ateses i 350 impactes en els mitjans de comunicació l'any 1998. Juntament amb aquests tres sistemes, en l'actualitat el medi més potent és sens dubte internet. La pàgina web de l'Idescat conté una informació prou rica des de diferents òptiques, tal com mostra la seva pàgina principal (gràfic 7). Es pot veure que hi ha informació de conjuntura, territorial, de

sectors industrials, de mercat de treball, de comerç exterior i d'altres. Aquesta web ha rebut unes 150 mil peticions http mensuals al llarg d'aquest any, això és, unes 5.000 consultes diàries. Pel que fa al nombre de visites, la mesura és difícil, però pel nombre de peticions a la primera pàgina, se suposa que ens visiten diàriament entre 250 i 300 persones.



Nota: totes les estadístiques de l'Idescat poden consultar-se de forma completa o parcial al web de l'Institut d'Estadística de Catalunya (Idescat).

Gràfic 7. Pàgina principal de la Web de l'Idescat (www.idescat.es)

Però el nostre interès primordial està centrat en les fonts i, per tant, en l'activitat de producció de l'oficina regional d'estadística que, com s'ha esmentat anteriorment, està definida estratègicament com de millora i complementació de la informació disponible.

L'estratègia no ha estat fixada arbitràriament per part del nivell tècnic de l'Idescat, sinó que és el resultat de les directrius polítiques i jurídiques del sistema estadístic de Catalunya. En efecte, la llei del Pla Estadístic de Catalunya, aprovada pel Parlament de Catalunya, estableix en el capítol 2, article 8, el següent objectiu:

«*L'objectiu central del Pla estadístic de Catalunya 1997-2000 és desenvolupar i consolidar el sistema estadístic de Catalunya i aconseguir un conjunt coherent, ordenat, fiable i actualitzat de dades estadístiques, comparable amb els dels sistemes estadístics de l'entorn que, amb el mínim cost possible i aprofitant al màxim les fonts existents, permeti el coneixement de la realitat econòmica, demogràfica i social de Catalunya, sigui útil per a la presa de decisions de les institucions públiques i els agents socials i minimitzi les molèsties als ciutadans i els en garanteixi el secret estadístic*».

Resulta evident que per a l'organisme estadístic de la Generalitat, l'Idescat, les idees de la coherència i la comparabilitat amb els sistemes estadístics de l'entorn, juntament amb l'objectiu de la minimització dels costos i l'aprofitament màxim de les fonts, ja indica ben clarament el camí a seguir.

S'ha de tenir present que aquesta complementarietat no implica, però, que l'actuació de l'Idescat no comporti una notable i enriquidora diversificació de fonts sobre la realitat catalana, que és el nostre objectiu en termes d'informació. Aquesta diversificació de fonts complementàries serà presentada seguidament, diferenciant entre la complementarietat de les activitats o processos i la complementarietat dels productes finals.

LA COMPLEMENTARIETAT DE LES FONTS ESTADÍSTIQUES REGIONALS (I): COL·LABORACIÓ I APROFITAMENT EN L'ACTIVITAT

Una primera línia de complementarietat ve donada en l'àmbit de l'activitat, de la producció de grans operacions estadístiques que, amb caràcter censal o d'enquesta, són dutes a terme per part de l'estadística de l'administració central, l'INE. Resulta evident que aquestes operacions són de màxim interès per a l'estadística regional. Per exemple: els censos de població, els censos agraris, les enquestes anuals a les empreses industrials i, més recentment, les enquestes a les empreses de serveis. El plantejament de l'Idescat va ser, des del primer moment, col·laborar i donar suport a les operacions de l'INE i, alhora, aprofitar-les al màxim. Aquesta activitat de col·laboració i aprofitament genera una diversitat important de fórmules d'actuació. A la taula 4 es recullen distintes possibilitats d'intervenció i aprofitament de les dades. Aquestes diverses possibilitats dónen lloc a una sèrie d'opcions que queden recollides esquemàticament a les cinc taules següents (taules 5 a 9). Cada taula representa una fórmula de col·laboració o aprofitament amb elements diferencials.

Taula 4. Línies d'activitat de l'Idescat en relació amb processos de l'INE

Qüestionari	Edició de qüestionaris bilíngües amb anagrama INE/Idescat	Addició d'un mòdul amb preguntes d'interès de l'Idescat
Directori	Verificació exhausitivitat del directori	Actualització contínua de directoris
Mostra	Sense participació (dins d'uns límits de fiabilitat general, podria definir-se una afixació d'interès per a Idescat)	
Treball de camp	Presentació de l'enquesta amb anagrama i signatures INE/Idescat	Execució del treball de camp de l'Idescat
Depuració, imputació i validació de microdades	Execució de la depuració, imputació i validació de l'Idescat	
Elevació de dades mostrals	Elevació de l'Idescat complementària a la de l'INE	Elevació de l'Idescat autònoma de l'elevació INE
Tabulació a partir de microdades	Tabulació de l'Idescat complementària a la de l'INE	Tabulació de l'Idescat autònoma de l'elevació INE

Taula 5. Línies d'activitat de l'Idescat en relació amb processos de l'INE

Col·laboració en la producció (i): Suport institucional i actuació operativa forta

Exemples: censos de població, cens agrari, enquesta de biblioteques, enquesta econòmica de l'ensenyament privat		
Qüestionari	Edició de qüestionaris bilingües amb anagrama INE/Idescat	Addició d'un mòdul amb preguntes d'interès de l'Idescat
Directori	Verificació exhaustivitat del directori	Actualització contínua de directoris
Mostra	Sense participació (dins d'uns límits de fiabilitat general, podria definir-se una afixació d'interès per a Idescat)	
Treball de camp	Presentació de l'enquesta amb anagrama i signatures INE/Idescat	Execució del treball de camp de l'Idescat
Depuració, imputació i validació de microdades	Execució de la depuració, imputació i validació de l'Idescat	
Elevació de dades mostrals	Elevació de l'Idescat complementària a la de l'INE	Elevació de l'Idescat autònoma de l'elevació INE
Tabulació a partir de microdades	Tabulació de l'Idescat complementària a la de l'INE	Tabulació de l'Idescat autònoma de l'elevació INE

Taula 6. Línies d'activitat de l'Idescat en relació amb processos de l'INE

Col·laboració en la producció (ii): Suport institucional i actuació operativa dèbil

Exemples: enquesta industrial d'empreses, enquesta industrial de productes, enquesta a empreses de serveis		
Qüestionari	Edició de qüestionaris bilingües amb anagrama INE/Idescat	Addició d'un mòdul amb preguntes d'interès de l'Idescat
Directori	Verificació exhaustivitat del directori	Actualització contínua de directoris
Mostra	Sense participació (dins d'uns límits de fiabilitat general, podria definir-se una afixació d'interès per a Idescat)	
Treball de camp	Presentació de l'enquesta amb anagrama i signatures INE/Idescat	Execució del treball de camp de l'Idescat
Depuració, imputació i validació de microdades	Execució de la depuració, imputació i validació de l'Idescat.	
Elevació de dades mostrals	Elevació de l'Idescat complementària a la de l'INE	Elevació de l'Idescat autònoma de l'elevació INE
Tabulació a partir de microdades	Tabulació de l'Idescat complementària a la de l'INE	Tabulació de l'Idescat autònoma de l'elevació INE

Taula 7. Línies d'activitat de l'Idescat en relació amb processos de l'INE
Col·laboració en la difussió (amb microdades)

Exemples: enquesta de població activa, enquesta d'ocupació hotelera		
Qüestionari	Edició de qüestionaris bilíngües amb anagrama INE/Idescat	Addició d'un mòdul amb preguntes d'interès de l'Idescat
Directori	Verificació exhaustivitat del directori	Actualització contínua de directoris
Mostra	Sense participació (dins d'uns límits de fiabilitat general, podria definir-se una afixació d'interès per a Idescat)	
Treball de camp	Presentació de l'enquesta amb anagrama i signatures INE/Idescat	Execució del treball de camp de l'Idescat
Depuració, imputació i validació de microdades	Execució del la depuració, imputació i validació de l'Idescat.	
Elevació de dades mostrals	Elevació de l'Idescat complementària a la de l'INE	Elevació de l'Idescat autònoma de l'elevació INE
Tabulació a partir de microdades	Tabulació de l'Idescat complementària a la de l'INE	Tabulació de l'Idescat autònoma de l'elevació INE

Taula 8. Línies d'activitat de l'Idescat en relació amb processos de l'INE
Aprofitament per a producció complementària

Exemples: índex de producció industrial		
Qüestionari	Edició de qüestionaris bilíngües amb anagrama INE/Idescat	Addició d'un mòdul amb preguntes d'interès de l'Idescat
Directori	Verificació exhaustivitat del directori	Actualització contínua de directoris
Mostra	Sense participació (dins d'uns límits de fiabilitat general, podria definir-se una afixació d'interès per a Idescat)	
Treball de camp	Presentació de l'enquesta amb anagrama i signatures INE/Idescat	Execució del treball de camp de l'Idescat
Depuració, imputació i validació de microdades	Execució del la depuració, imputació i validació de l'Idescat.	
Elevació de dades mostrals	Elevació de l'Idescat complementària a la de l'INE	Elevació de l'Idescat autònoma de l'elevació INE
Tabulació a partir de microdades	Tabulació de l'Idescat complementària a la de l'INE	Tabulació de l'Idescat autònoma de l'elevació INE

Taula 9. Línies d'activitat de l'Idescat en relació amb processos de l'INE

Aprofitament per a producció pròpia

Exemples: índex de preus industrials		
Qüestionari	Edició de qüestionaris bilíngües amb anagrama INE/Idescat	Addició d'un mòdul amb preguntes d'interès de l'Idescat
Directori	Verificació exhaustivitat del directori	Actualització contínua de directoris
Mostra	Sense participació (dins d'uns límits de fiabilitat general, podria definir-se una afixació d'interès per a Idescat)	
Treball de camp	Presentació de l'enquesta amb anagrama i signatures INE/Idescat	Execució del treball de camp de l'Idescat
Depuració, imputació i validació de microdades	Execució del la depuració, imputació i validació de l'Idescat.	
Elevació de dades mostrals	Elevació de l'Idescat complementària a la de l'INE	Elevació de l'Idescat autònoma de l'elevació INE
Tabulació a partir de microdades	Tabulació de l'Idescat complementària a la de l'INE	Tabulació de l'Idescat autònoma de l'elevació INE

A les anteriors taules s'han identificat fins a cinc activitats de l'Idescat associades a processos de l'INE, algunes que podrien ser considerades com a coproduccions amb col·laboració en el procés (en un sentit més fort o dèbil), com a codifusió i com a producció autónoma amb aprofitament de microdades de l'INE.

Totes aquestes fórmules de relació amb l'INE queden situades en el camp dels censos i de les enquestes. Concretament, les cinc fórmules es troben presents en les enquestes mentre que en els censos de fet només en tenim dues: la col·laboració en producció (forta) i la codifusió.

Per la seva naturalesa, els registres administratius són una activitat que es realitza en col·laboració, ja que es tracta d'un aprofitament d'un organisme extern (duanes, registre mercantil central, etc.). En tot cas, pot diferenciar-se entre processos molt senzills i directes, o processos més complexos. Per la seva part, la producció secundària, de síntesi comptable, es realitza actualment sense col·laboració en el procés, encara que sí que hi ha relació en l'assessorament.

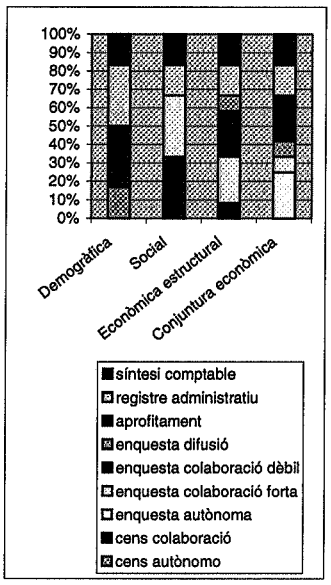
Per tot l'exposat anteriorment, si volem avaluar la diversitat de formes de producció (i també la difusió a partir de microdades), tindrem que a les anteriors quatre formes, hauríem d'afegir les noves detectades.

Per aquest motiu, l'estadística presentada a les taules 1 i 2 del principi d'aquest treball s'hauria de completar amb la següent estadística:

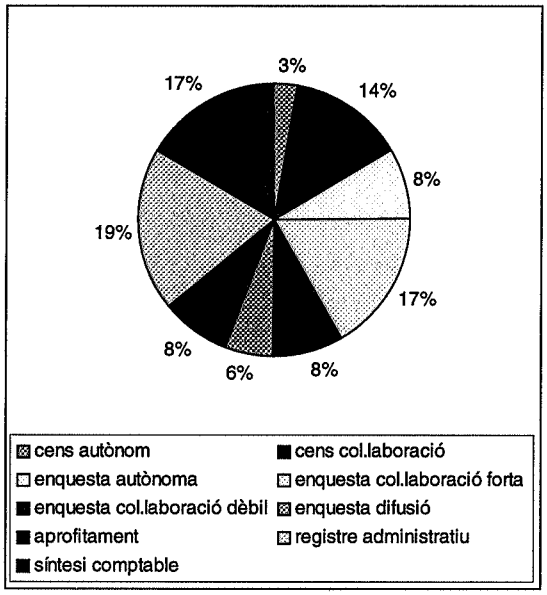
Taula 10. Estadística sobre les formes de producció estadística (Idescat).
Censos i enquestes segons relació amb activitats Adm. Central. (INE i Ministeris)

	Demogràfica	Social	Econòmica estructural	Conjuntura econòmica	Total
Cens autònom (amb Padró)	*Cens 96				1
Cens amb forta col·laboració	*Cens 91 *Padró 86	*Cens d'habitatge	*Cens agrari		4
Cens amb col·laboració difusió		*Cens d'Edificis			1
Enquesta autònoma				*Viatges cat. *Viatges esp. *IVGS	3
Enquesta amb col·laboració forta		*Biblioteques *Ensenyament	*Expl. Agràries *Transport *Ad. Públiques	*Construc.	6
Enquesta amb col·laboració dèbil			*Industrial *Comerç *Hotels		3
Enquesta amb col·laboració difusió			*EPA Ampliació (tabulació anual)	*Clima exportador *Activitat hotelera	3
Aprofitament producció complementària				*IPI	1
Aprofitament producció pròpia				*IPRI	1
Total	3	4	8	8	23

El recompte anterior ens permet renovar l'estadística del gràfic 1 sobre formes de producció en el cas de l'Idescat, com a organisme d'estadística regional. En els gràfics 6 i 7 es pot veure com les diferents formes de col·laboració i aprofitament de l'estadística oficial regional en relació amb l'administració central, lluny de generar un panorama empobridor o uniforme encara produeix una major diversitat de formes d'actuació, de manera que la diversitat de les fonts generals queda multiplicada per les fonts regionals a partir de l'estratègia de la complementarietat en el procés de producció.



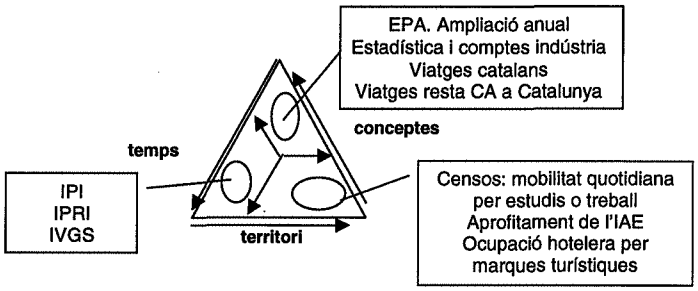
Gràfic 8. Formes de producció estadística (Idescat): per àrees temàtiques i col·laboracions



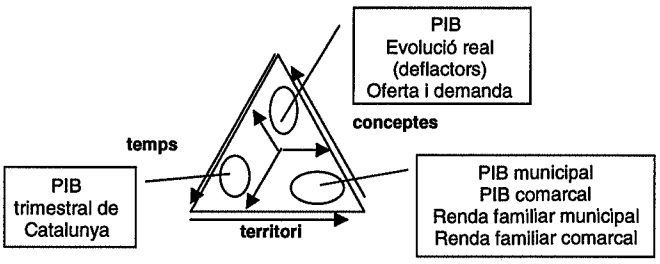
Gràfic 9. Diversitat de les formes de producció estadística (Idescat): totes les àrees temàtiques

LA COMPLEMENTARIETAT DE LES FONTS ESTADÍSTIQUES REGIONALS (II): QUANTITAT, PUNTUALITAT I COMPARABILITAT DELS RESULTATS

La segona línia de complementarietat queda ubicada en l'àmbit dels resultats. Aquesta complementarietat es deriva de l'acceptació de la informació regional proporcionada per l'estadística de l'administració central, bàsicament de l'INE. D'aquesta forma, l'activitat més operativa i socialment més justificable passa per complementar aquest cos de dades en els punts en els que aquest resulti poc informatiu del nostre àmbit d'interès. Sobre aquesta base, el treball consisteix en fer una anàlisi dels productes de caràcter general que tenen informació regional, detectar carències i plantejar-se la seva superació. Des d'aquest punt de vista, i tornant a les tres òptiques presentades en la primera part d'aquest treball, poden identificar-se una sèrie de resultats estadístics que completen la informació sobre Catalunya i que fan èmfasi en un més gran detall conceptual, un millor coneixement del territori o uns productes estadístics més puntuals i útils per al seguiment de la conjuntura. En l'àmbit de la producció primària (amb tota la diversitat de formes de producció que s'acaba de mostrar) i en el de la síntesi comptable, exemples d'aquests diferents resultats estadístics són mostrats en els gràfics 10 i 11.



Gràfic 10. Productes estadístics primaris de l'Idescat



Gràfic 11. Productes estadístics comptables de l'Idescat

Cadascun dels productes d'aquests dos gràfics té una justificació clara ja que superen limitacions de la informació oferida per l'INE.

Aquestes estadístiques es produeixen mitjançant distintes estratègies. De vegades amb enquestes autònomes, com en el cas de l'IVGS, dels viatges turístics dels catalans o dels viatges dels residents en altres comunitats de la resta de l'estat espanyol a Catalunya. En d'altres casos, resulta necessari un mòdul específic d'interès de la Generalitat de Catalunya, com en el cas de la mobilitat obligada per raó d'estudi o treball. Sovint resulta molt recomanable fer un aprofitament d'arxiu de l'INE, com en el cas de l'IPRI i del IPI. Finalment, en operacions tan significatives com l'EPA, l'enquesta industrial o el cens agrari, es tracta de realitzar una tabulació complementària. Aquesta tabulació pot estar acompanyada (o no) per una col·laboració en el procés de producció, operativament dèbil o forta. Aquests són els casos, respectivament, de l'enquesta industrial o del cens agrari.

El plantejament d'assumir els resultats de l'adminitració central i de complementar-los és especialment clar en el cas de les macromagnituds. Des del primer moment, l'Idescat es va plantejar la idoneïtat d'assumir els resultats definitius de la Comptabilitat regional d'Espanya (CRE) de l'INE. Cal assenyalar que un cop s'assumiren aquests resultats es va fer manifesta la necessitat d'ampliar-los conceptualment oferint resultats anuals del PIB no només pel cantó de l'oferta, sinó també pel cantó de la demanda, així com la necessitat de deflactar totes les macromagnituds per a conèixer l'evolució anual de l'economia catalana en termes reals.

En l'àmbit territorial, es detectava la necessitat d'oferir dades del compte de renda de les llars i del PIB a un nivell comarcal i per a municipis d'una certa dimensió (més de 10.000 habitants). Finalment, pel que fa al temps, resultava un objectiu essencial arribar a elaborar la Comptabilitat Trimestral de Catalunya, realizada de forma paral·lela i comparable a la Comptabilitat Trimestral d'Espanya. En l'actualitat aquests resultats, tots ells compatibles amb les estimacions definitives de Valor Afegir Brut i de renda de les famílies de l'INE per a Catalunya, són força apreciats pels nostres usuaris que, d'aquesta manera, troben en l'activitat de l'Idescat una font d'enriquiment i complementarietat amb la informació de l'INE, i no un problema d'alternativa d'unes dades enfront les altres.

La clau d'aquest plantejament consisteix, com s'ha esmentat, en assumir els resultats de l'INE com a interessants i susceptibles de ser ampliats i complementats. Per això, com ja s'ha vist, hi ha raons de tipus jurídic i institucional. També hi ha raons d'eficàcia, quan s'enten que l'estadística oficial és, sobre tot, un servei públic.

Però addicionalment, de forma també significativa, existeixen raons de naturalesa més abstracta i teòrica. Hem vist anteriorment que la veritat, en una de les seves accepcions, significa coherència (i no només una correspondència local amb la realitat). En l'àmbit de l'estadística regional coherència vol dir comparabilitat i assumpció de resultats gene-

rals. Como ja s'ha apuntat amb anterioritat, en aquesta coherència, molt especialment en l'àrea de les macromagnituds, es troba bona part de la fiabilitat de les dades.

Arribats a aquest punt, val la pena destacar algunes connotacions que, al nostre entendre, té la idea d'assumir resultats i de treballar en l'àmbit de la complementarietat. Assumir resultats no vol dir tenir una actitud acrítica, de la mateixa forma que buscar la complementarietat no implica adoptar una posició pasiva de resignada supeditació i de menor importància o ambició.

Pel que fa a la crítica, tot el que s'ha dit no s'oposa a la possibilitat de crítica enfront els resultats que puguin considerar-se poc versemblants per part de l'INE. Al contrari, aquesta crítica és absolutament vàlida i imprescindible, entesa com una sèrie d'aportacions que pretenen introduir millores en els processos de producció dels organismes responsables de l'estadística oficial. Precisament en el camí de la col·laboració i l'aprofitament és on es troben amb més facilitat les oportunitats per a exercir amb prou coneixement de causa aquesta activitat crítica.

Pel que fa a la supeditació o dependència i a la importància de la nostra activitat com a organisme d'estadística regional, la nostra valoració és la més natural en un economista. És evident que l'actitud d'assumir i complementar dades d'un altre organisme genera una certa dependència, però aquesta dependència no és negativa en sí mateixa. Si aquesta dependència permet l'aprofitament d'informacions útils, la reducció de costos i, en definitiva, la possibilitat d'oferir un millor servei per la via de la divisió del treball, aquesta dependència és positiva. En tot cas, del que es tracta és de controlar els riscos, garantizant mitjançant acords institucionals el marc d'unes relacions en les que guanyen les dues parts.

Pel que fa a la importància, la discusió sobre la producció primària clàssica versus sistemes de producció en col·laboració o d'aprofitament fa recordar la disputa entre els economistes fisiòcrates sobre quines activitats creaven riquesa. Durant força temps s'assegurava que només l'activitat agrària creava riquesa, i no la transformació industrial o els serveis. Més endavant es plantejà que l'agricultura i la indústria podien generar riquesa, però no els serveis. Actualment els economistes tendeixen a pensar que qualsevol activitat que sigui valorada pel mercat té valor. En la producció d'informació estadística la situació és equivalent. La producció en col·laboració o producció lleugera no té perquè ser menys important que la producció pesant. Ambdues són reclamades per la societat i ambdues es necessiten mútuament.

D'altra banda, la divisió del treball ens permet ser tan ambiciosos com volguem en la nostra activitat, en producció estadística o en qualitat del servei que s'ofereix. Mostra d'aquesta ambició poden ser, en l'Idescat, estadístiques com la Comptabilitat Trimestral de Catalunya pel cantó de l'oferta i de la demanda, el sistema d'imputació de la despesa declarada dels turistes que s'està desenvolupant en col·laboració amb la Universitat de Girona, o el treball en l'estimació de petites àrees que es fa amb la Universitat Pompeu

Fabra. En l'àmbit de la difusió, es pot esmentar el programa de qualitat dels nostres serveis de biblioteca i atenció a peticions a mida, que tenim en marxa amb enquestes de satisfacció a usuaris realitzades en contacte amb les Universitats Politècnica de Catalunya i Autònoma de Barcelona.

Les reflexions presentades potser serviran per il·luminar quelcom del sentit de la diversitat de les fonts estadístiques i el valor d'aquesta diversitat. Poden servir també per entendre l'aportació que poden realitzar els instituts d'estadística regional. Una aportació en la qual un perill és caure en un mimetisme mecànic amb el treball que es du a terme per part de l'administració de l'Estat. És essencial per a definir les nostres pròpies estratègies com a estadístics advertir que el nostre camí no ha de ser necessàriament el mateix que el de l'INE. Precisament en aquesta diferència pot residir la possibilitat que, com a professionals de l'estadística regional, puguem ser creatius en la nostra activitat i elaborar productes realment útils per a la societat a la qual servim.

ENGLISH SUMMARY

DIVERSENESS AND COMPLEMENTARINESS OF STATISTICAL SOURCES

ÀLEX COSTA
Institut d'Estadística de Catalunya (Idescat)

In this article we show the diversity in official statistical sources. These sources can be produced by more than one system of production: census, sample surveys, administrative registers or accounting estimations. In our opinion, these systems are positive, because every different system of production has different performances. Census are very good for territorial data, surveys are very good in conceptual information and administrative registers are very good in time reference or in territorial information.

We can also find another kind of complementary sources. From a reliability point of view, primary sources (census, surveys, administrative registers) are very close to the reality, but sometimes have contradictory results. On the other hand, accounting estimations have not a so clear relationship with the reality, but accounting estimations are coherent, and this coherence is very positive for the reliability.

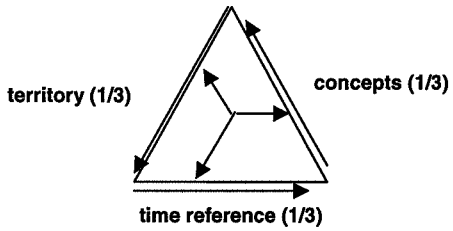
Finally, in this article we can find a presentation of the activities by Idescat (Regional Institute) in relationship with INE's activities (Central Institute). Idescat's activities are a complement to INE's statistics. Idescat can help to INE in statistical operations in Catalunya and also statistics from Idescat complement statistics from INE results when it is necessary.

Keywords: Statistical sources, diverseness, complementariness, correspondence, coherence, territory, time reference, concepts, institutional relationships

AMS Classification: 62P25

–Received September 1999.
–Accepted November 1999.

The statisticians would like only one very good source for all points of view (territorial information, conceptual information and time reference) but that perfect source does not exist. If we look at graphic 1, we can understand the reason why. Statistical sources are placed in a triangle. Each dimension is territory, concepts and time. If one statistical source is very good in time, it will not be very good in concepts or in territorial information. This is the case of short-term surveys. If one statistical source is very good in territorial data (case of the census), it will not be very good in time reference and it will have not a lot of details in concepts. Finally, if one source is very good in concepts, may be it will not be very interesting in its territorial data or in its time reference.



Graphic 1.

There is another kind of complementary in statistical sources, not so easy to understand. In this case among primary sources and accounting estimation. The main idea is this: primary statistical sources give us information very close to the facts, but sometimes this information is contradictory. This problem is solved thanks to the accounting estimation, because this system works with all the primary sources to elaborate consistent statistical results. In this sense, primary sources and the accounting estimation are complementary. The first is very near to the facts, the second give us consistency.

The second part of the report talks about regional statistical sources. We present two kinds of co-operation. The first is a co-operation in progress. In Catalunya our institute (Idescat) and the Spanish institute (INE) co-operate to do surveys and census together. This co-operation has two ways:

- Idescat gives institutional help to the survey (with a letter to the people that will be asked) and, also, Idescat makes the questionnaire in two languages (Catalan and Spanish)
- In addition to the considered aspects, sometimes Idescat also does the field work and codify the information.

In both cases Idescat can use all the information for its publications and databases.

The second line of co-operation is placed in the results. We accept the results of INE, we analyze these results and when we detect some limitation, then we try to solve it with a complementary work. For instance: regional accounting give us the evolution of the economic activity, but in currency, not in real rates. Then, we correct the currency quantities, because we want to know the real rate of increase of the economy of Catalunya.

In the regional cooperation we find the ideas we used in the first part of this article. Frequently, our work with INE improve our information with better data in one of these points of view: territory, time or concepts.

We will finish with an example of this complementary work from Idescat in the three dimensions of the triangle: reference of time, concepts and territory.

Time References:

The INE estimations of GNP for economy of Catalunya are very good, but not in time. For instance, in 1999 INE shows the data for 1996. Idescat complements this data with an advance of the GNP. We have our estimation with a gap of nine months. For this reason, our advance of GNP complements the GNP of INE.

Territory:

INE elaborates an estimation of household income in Catalunya and also for the four departments in Catalunya (Barcelona, Tarragona, Lleida and Girona). This information is very interesting for us, but we would like more details about our territory. The public administration and the economic agents of Catalunya, would like to know the income in small territorial areas, like counties («comarca» in Catalan) and cities and villages. This is an objective for Idescat. We accept the general estimation from INE and we estimate the income for these small areas.

Concepts:

INE and Idescat have an agreement to do an industrial companies survey yearly. From this survey, INE published a lot of information for Spain economy but not a lot regional information. For instance, for Catalunya INE only shows results for nine variables and twelve industrial sectors. Idescat is interested in more information, and for this reason we do a additional tabulation: in our statistics we show information for forty-one industrial sectors and more than twenty-five concepts.

QÜESTIIÓ, vol. 24, 1, p. 111-136, 2000

DISSENY I CONSTRUCCIÓ D'UNA MOSTRA ESTRATIFICADA A PARTIR DE DADES CENSALS*

P. LÓPEZ*
C. LOZARES*
Universitat Autònoma de Barcelona
M. DOMÍNGUEZ**
Universitat de Barcelona

L'objectiu de l'article és presentar les principals característiques del disseny i del procés de construcció de la mostra estratificada de l'«Enquesta metropolitana de Barcelona. Condicions de vida i hàbits de la població» de l'edició de l'any 1995. A partir de la informació que prové del cens de població s'agrupen a les persones en seccions censals i es procedeix a la construcció dels estrats d'acord amb un procediment on s'impliquen tècniques d'anàlisi multivariable: d'anàlisi factorial de components principals i de classificació automàtica. Finalment, s'afixa una grandària de mostra segons el criteri òptim de Neyman. El procediment seguit i els criteris i decisions que s'utilitzen van més enllà dels resultats habituals esperats a tota mostra, doncs proporciona conclusions d'interès per a l'anàlisi sociològica i la planificació del territori, al mateix temps que serveix de criteri de validació de les conclusions posteriors a l'anàlisi de l'enquesta.

Desing and construction of a stratified sample from census data

Paraules clau: Enquesta per mostreig, anàlisi de dades, anàlisi de components principals, anàlisi de classificació

Classificació AMS: 62D05, 62-07, 62H25, 62H30

⋆ Els tres autors són membres del *Grup d'Estudis Sociològics sobre la Vida Quotidiana i el Treball* (QUIT) del Departament de Sociologia de la Universitat Autònoma de Barcelona.

* Pedro López Roldán (Pedro.Lopez.Roldan@uab.es); Carlos Lozares Colina (Carlos.Lozares@uab.es). Departament de Sociologia. Universitat Autònoma de Barcelona.

** Màrius Domínguez Amorós (marius@riscd2.eco.ub.es). Departament de Sociologia. Universitat de Barcelona.

– Rebut el setembre de 1999.
– Acceptat el novembre de 1999.

1. INTRODUCCIÓ

L'objectiu d'aquest article és presentar les característiques de major rellevància relatives al disseny i al procés de construcció de la mostra estratificada que serveix de base per a la recollida (producció) d'informació de l'*Enquesta Metropolitana de Barcelona. Condicions de Vida i Hàbits de la Població* (EM) de l'edició de l'any 1995, elaborada al sí de l'*Institut d'Estudis Metropolitans de Barcelona.*[1]

L'EM constitueix una investigació permanent que va ser projectada l'any 1984 per a analitzar les activitats, condicions i formes de vida de la població de l'Àrea Metropolitana de Barcelona davant l'absència de dades estadístiques sistemàtiques i objectives, de caràcter social i sobre aquest contingut. L'EM s'ha convertit en un instrument periòdic de recollida d'informació que ens ajuda a precisar l'evolució i els canvis de les tendències més estructurals del conjunt de fenòmens socials que s'analitzen a l'enquesta[2].

Encarregada inicialment per l'antiga Corporació Metropolitana de Barcelona (CMB), la primera edició (1985) considerà els 27 municipis que constituïen la CMB. L'edició de 1990, encarregada per la Mancomunitat de Municipis de la Regió Metropolitana de Barcelona i la Diputació de Barcelona, va estendre l'univers poblacional a l'anomenada Regió Metropolitana de Barcelona (RMB) reunint les 5 comarques que conformen la Regió I dins de la divisió territorial de Catalunya. Finalment, la darrera edició de 1995 va afegir dins la RMB les comarques de l'Alt Penedès i del Garraf[3].

L'EM recull informació a partir del disseny d'una mostra estratificada que s'ha mantingut al llarg de les tres edicions, si bé l'àmbit territorial de referència s'ha anat modificant. El punt de partida és la informació que prové del cens de població. A partir de l'agrupació de les persones en seccions censals es disposa d'informació en termes de la proporció de persones que posseeixen determinada característica sobre el total de la secció censal, segons un conjunt d'indicadors seleccionats. A partir d'aquesta

[1] Aquest article va ser presentat pels autors en forma de comunicació a les «Jornades internacionals sobre generació d'informació estadística: qualitat i limitacions», organitzades a Barcelona el novembre de 1998 per la xarxa temàtica *Enquestes i qualitat de la informació estadística.* L'exposició detallada del disseny i del procés de construcció de la mostra, per a l'edició de l'Enquesta Metropolitana de 1990, es pot consultar a Lozares i López (1990). Una versió reduïda d'aquest document va ser publicat a Lozares i López (1991b). Posteriorment, prenent com a base aquesta mostra es va considerar el tractament de subpoblacions per a la construcció de zones socials (Lozares i Domínguez, 1993, 1996). A partir d'aquests antecedents fonamentem el contingut d'aquest article, però considerant el material inèdit del procés de construcció de la mostra de l'any 1995. L'Enquesta Metropolitana de l'any 1995 ha estat dirigida per Marina Subirats.

[2] Són nombroses les publicacions que s'han generat amb les anàlisis de les dades de l'EM, l'edició de les quals ha anat a càrrec de l'Institut d'Estudis Metropolitans de Barcelona.

[3] En el projecte de la nova edició de l'any 2000 s'està considerant tota la província de Barcelona. Al llarg d'aquests anys, i davant la necessitat de mantenir la sèrie temporal, sempre s'ha considerat el manteniment de submostres representatives que facilitessin la continuïtat i la comparació entre les diferents edicions.

informació es procedeix a la construcció dels estrats d'acord amb un procediment en el que s'impliquen tècniques d'anàlisi multivariable: d'anàlisi factorial de components principals i de classificació automàtica.

Amb els estrats prèviament analitzats i construïts es procedeix a l'afixació d'una grandària de mostra donada segons el criteri òptim de Neyman. Després de la distribució per estrats es realitza l'assignació de quotes de mostra a cada secció censal de cada estrat segons la seva població. Les unitats finalment s'extreuen de forma aleatòria a partir del marc de la mostra que s'obté del registre del cens electoral.

El procediment seguit i els criteris i decisions que s'utilitzen van més enllà dels resultats habituals esperats a tota mostra, doncs, com destacarem, proporciona conclusions d'interès per a l'anàlisi sociològica, al mateix temps que serveix de criteri de validació de les conclusions posteriors en l'anàlisi de l'enquesta.

D'altra banda, i com a resultat d'una anàlisi específica realitzada amb posterioritat a la recollida d'informació, s'ha reutilitzat la mostra amb l'objectiu de construir el que hem anomenat com a zones socials. Cada estrat de la mostra estratificada pot constituir una base mostral per a dur a terme anàlisis amb totes les garanties de precisió, com a mínim calculables i acceptables, sense augmentar la grandària de la mostra per a l'enquesta general.

Tots aquests aspectes ressenyats són els que articulen el contingut de l'article als apartats següents.

2. LA CONSTRUCCIÓ DE LA MOSTRA DE L'ENQUESTA METROPOLITANA

La construcció de la mostra de l'EM per a l'any 1995 manté les mateixes característiques de procediment que en les dues anteriors edicions. Tot seguit especificarem els trets més significatius del disseny mostral i del procés de construcció que s'organitza a partir de la distinció de cinc moments:

1. Definició de l'univers poblacional i selecció del camp de les variables/criteri d'estratificació.
2. Anàlisi de dimensionalització per components principals amb l'objectiu de reduir i sintetitzat els factors principals de variabilitat de la informació original.
3. Anàlisi de classificació automàtica que, en funció dels factors principals, agrupi les seccions censals en grups homogenis, en els estrats de la mostra.
4. Determinació de la grandària i de l'error mostrals amb la corresponent afixació de la mostra entre els estrats.

5. Ponderació a posteriori de la mostra per a restituir l'equiprobabilitat de que un individu sigui elegit a l'atzar.

2.1. Disseny de la mostra: univers poblacional i variables criteri

En el disseny de la mostra es va establir com a objectiu l'extracció d'una mostra aleatòria estratificada representativa de la població de la Regió Metropolitana de Barcelona. L'elecció del procediment de mostreig estratificat es justifica bàsicament per criteris de precisió front altres mètodes i per l'heterogeneïtat social que caracteritza la població objecte d'estudi. Així, la construcció dels estrats ens permet disposar d'una variable d'estratificació que constituirà una variable densa de tipificació social i que ens garantirà la presència a la mostra de tipus socials homogenis característics de la població, tipus o grups socials que estan en la base, es correlacionen, amb els objectius d'estudi de l'EM.

L'univers poblacional es defineix com el conjunt de persones majors de 18 anys de la RMB. A l'edició de 1995 la RMB comprèn les 7 comarques catalanes següents: Barcelonès, Baix Llobregat, Maresme, Vallès Occidental, Vallès Oriental, Alt Penedès i Garraf. Es tracta d'un territori que comprèn 162 municipis i una població total de 3.275.458 persones, segons les dades del Cens de Població de 1991.

La condició indispensable de poder disposar de les dades poblacionals del cens ens ofereix la possibilitat de l'estratificació. De la informació disponible del qüestionari del cens hem seleccionat una sèrie d'indicadors que són els que actuen com a variables/criteri d'estratificació i que caracteritzaran, per l'agregació de les persones, a les seccions censals de la RMB. Així doncs, la unitat elemental a estratificar que hem considerat no són, en primera instància, les persones censades, sinó les seccions censals on aquestes resideixen. Procedir d'aquesta manera es justifica per un doble motiu: per les dificultats que es deriven de tractar una matriu d'individus d'aquesta magnitud; i perquè tan sols l'agregació en seccions censals ens permet un tractament percentual mètric de les variables seleccionades. Tornarem a l'individu en el moment d'aplicar la fórmula per a calcular la grandària de la mostra. Llavors establirem el supòsit, per a fer el salt de les seccions als individus, segons el qual els individus d'una mateixa secció censal posseeixen la mateixa condició socioeconòmica.

La selecció final de les variables/criteri que considerarem obeeix en primer terme a la seva disponibilitat com informació recollida al cens de l'any 1991, també respon a criteris de pertinença conceptual d'acord amb els objectius de l'estudi de l'EM, i finalment a criteris de tipus estadístic que tenen a veure amb els resultats del procés d'anàlisi que presentarem a continuació i que ens van conduir a l'eliminació d'alguna de les variables inicialment considerades (per manifesta combinació lineal o per l'escàs valor o dispersió de les variables).

En conseqüència, tenim una matriu de dades amb 3586 seccions censals de la RMB amb la selecció dels 16 d'indicadors o variables socioeconòmiques que apareixen a Taula 1. Es tracta de variables de caracterització demogràfica, cultural-educativa, d'activitat laboral i professional, de mobilitat i de grandària poblacional. Les variables a la matriu original expressen el percentatge de la població de la secció censal que posseeix una determinada característica sobre el total de la població de la secció censal corresponent. Les mitjanes i les desviacions són estadístics computats sobre valors de percentatges de cada secció sobre total de seccions.

Taula 1. Mitjana, desviació i descripció de les variables/criteri de la mostra

Variable	Mitjana	Desviació	Descripció
P1	14,23	5,02	Joves de menys de 15 anys
P2	16,24	6,76	Vells majors de 65 anys
P4	33,59	1,40	Immigració fora Catalunya
P7	2,18	0,38	Analfabets majors de 10 anys
P8	27,63	5,38	Titulats mitjans-superiors majors 20 anys
P9	57,96	3,95	Escolarització 14-24 anys
P11	12,16	4,16	Aturats abans ocupats
P12	3,20	1,93	Atur busca primera feina
P14	38,69	5,79	Dones actives majors 15 anys
P15	17,82	12,82	Professions altes
P16	37,28	17,91	Professions baixes
P17	19,07	4,49	Terciari mitjà/comerç/hosteleria
P18	4,28	3,09	Terciari alt/finances
P19	0,91	0,60	Agropecuari
P20	43,09	13,80	Vehicle privat treball
P23	4,54	4,77	Població Secció/Municipi

La construcció dels estrats homogenis de població és l'objectiu buscat. Es tracta de garantir que a la mostra hi siguin representades una sèrie de característiques de la població –i els fenòmens que en depenen–. El procés de construcció dels estrats es va fer mitjançant la utilització de dues tècniques d'anàlisi multivariable independents i complementàries: l'anàlisi factorial de components principals i l'anàlisi de classificació automàtica.

2.2. L'anàlisi de components principals

Amb l'anàlisi de components principals (ACP) es pretén reduir la informació original per tal d'obtenir un subespai vectorial de menys dimensions o factors, on aquests són

base i per tant linealment independents, i que, ordenats de manera jeràrquica, conserven la major part de la variança total. Obtenim així les dimensions fonamentals de diferenciació de la població de la RMB que estructuren inicialment la realitat social segons la informació introduïda. L'ACP es concep així com una etapa prèvia i complement necessari de la categorització de les unitats, d'obtenció dels estrats.

En el procés de l'ACP distingim cinc moments bàsics que passem a comentar: càlcul dels eixos factorials o components; càlcul dels valors propis i nombre d'eixos a retenir; correlació de les components amb les variables originals, comunalitats i recomposició de la matriu de correlacions; interpretació dels factors amb la rotació dels eixos; i recomposició de la matriu d'unitats en els nous eixos retinguts.

El càlcul dels eixos parteix de les 16 variables originals que apareixen a la Taula 1, la matriu de correlacions de les quals es presenta a la Taula A1 de l'annex[4]. L'examen de les correlacions de la matriu i altres mesures ens confirmen l'adequació de la informació considerada: la significativitat de totes les correlacions, el determinant no nul i pròxim a zero, el comportament de la matriu antiimatge de correlacions, la significativitat del test de Barlett i l'obtenció d'un índex d'adequació mostral de Kaiser-Meyer-Olkin de 0,83.

Essent la matriu de correlacions no singular, el càlcul dels valors propis o variança incorporada a cada eix ens porta a l'obtenció de 16 vectors propis i valors propis associats que es presenten a la Taula 2. Si apliquem els criteris habitualment emprats a l'hora de decidir el nombre d'eixos a retenir, l'espai vectorial original pot reduir-se a només 4 dimensions o variables factorials, acumulant el 76,6% de la variança total.

La relació entre les variables primitives i les components obtingudes ens permetrà simultàniament recompondre les variables originals en els nous eixos, el que ens mostrarà l'estructura el primer espai d'atributs, i donar identitat a les components. A la Taula A2 de l'annex es presenta la matriu de saturacions que mostra aquesta relació, en aquest cas després d'haver aplicat una rotació varimax. Als Gràfics 1 i 2 es representen igualment aquestes dades que complementaran la interpretació de la identitat dels factors o components.

El primer eix o component, que acumula el 41% de la variança total[5], té un pes cabdal, sent per tant una dimensió determinant de l'estructura de relacions de les variables originals. L'eix o la dimensió es defineix a partir de la polaritat que oposa, d'una banda, la presència de professions baixes, d'immigració de fora de Catalunya, de l'analfabetisme

[4]Tots els resultats de les anàlisis presentades per a la construcció de la mostra s'han realitzat mitjançant el programari estadístic SPSS.

[5]Considerem el percentatge de varianza explicat pels factors inicialment, sense considerar a l'explicació la redistribució que implica la rotació dels eixos.

i de l'atur; de l'altra banda, la dimensió recull en aquest pol el fet de tenir una professió alta i una ocupació en el terciari alt/finances, de posseir un nivell de titulació mitjà o superior, amb altes taxes d'escolarització entre 14 i 24 anys. Es tracta doncs d'una dimensió de *categoria social entesa com un compost de categoria professional, nivell educatiu, inserció laboral i origen immigrant.*

El segon eix, amb el 17,7% del total de la variança, reflecteix una dimensió que oscil·la entre la presència d'altes proporcions de població jove de menys de 15 anys, on hi ha una major taxa de dones actives i s'empra el vehicle privat per anar a treballar, en un extrem, i el predomini de població vella de més de 65 anys, juntament amb un certa presència del terciari mitjà/comerç/hosteleria, en l'altre extrem. És una dimensió *de cicle vital o d'edat*, que oposa a les seccions censals de població més jove, resident a la perifèria metropolitana i dels municipis front al nuclis urbans més cèntrics on la proporció de població vella és més important.

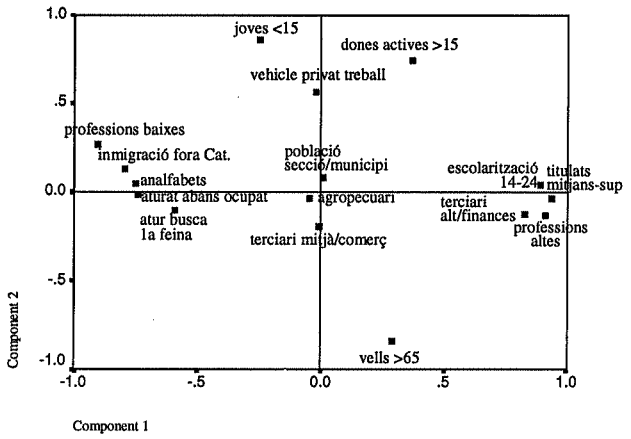
Taula 2. Valors propis (autovalors) obtinguts a l'ACP

Component	Autovalors inicials			Sumes de les saturacions al quadrat de l'extracció		
	Total	% de la variança	% acumulat	Total	% de la variança	% acumulat
1	6,561	41,0	41,0	6,561	41,0	41,0
2	2,828	17,7	58,7	2,828	17,7	58,7
3	1,891	11,8	70,5	1,891	11,8	70,5
4	,978	6,1	76,6	,978	6,1	76,6
5	,822	5,1	81,8			
6	,586	3,7	85,4			
7	,476	3,0	88,4			
9	,324	2,0	92,9			
10	,299	1,9	94,8			
11	,255	1,6	96,4			
12	,234	1,5	97,8			
13	,152	,9	98,8			
14	,130	,8	99,6			
15	,038	,2	99,8			
16	,028	,2	100,0			

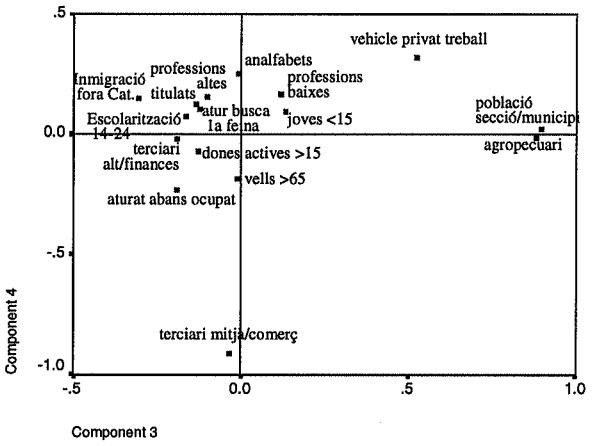
El tercer eix, que té l'11,8% del total de la variança, es defineix con una dimensió d'*identitat metropolitana*, d'oposició rural-urbà o grau de «metropolització», doncs ens mostra la polaritat entre una alta proporció de població ocupada al sector agropecuari i valors alts en el quocient poblacional secció/municipi, i pràcticament la resta de variables.

El quart eix, finalment acumula el 6,1% de la variança, essent una dimensió definida netament per l'oposició entre la variable terciari mitjà/comerç/hosteleria i la resta; és per tant una dimensió de *terciarització* que diferencia les seccions censals on hi ha una important presència d'activitat de comercial i hostelera de la resta.

Amb aquesta caracterització dels eixos factorials, finalment, la recomposició de les unitats seccions censals en el subespai vectorial ens porta a obtenir les puntuacions factorials en l'espai reduït de quatre dimensions que seran reutilitzades com a punt de partida de l'anàlisi de classificació automàtica.



Gràfic 1. Components 1 i 2 en l'espai rotat



Gràfic 2. Components 3 i 4 en l'espai rotat

2.3. L'anàlisi de classificació automàtica

L'estratificació de la mostra descansa sobre el principi de que la població en estudi careix d'homogeneïtat a efectes estadístics. L'homogeneïtat estadística fa referència a la relació que s'estableix entre les distintes característiques d'una mateixa població, quant més correlacionades estiguin aquestes característiques més homogènia serà la població considerada. L'objectiu de l'estratificació serà doncs la classificació de les seccions censals en estrats que, a efectes del mostreig, de guany en la precisió, seran l'expressió de conjunts de seccions el més homogènies possible a dintre de cada estrat, i el més heterogènies entre elles, segons les variables/criteri factorials derivades de l'anàlisi anterior.

Aquesta classificació es realitza sense tenir en compte cap restricció de contigüitat territorial, per la qual cosa s'obtindrà com a resultat un mapa de seccions de la RMB de diferent caracterització social segons la seva pertinença als estrats, la presència dels quals tindrà per tant una distribució diferenciada al territori.

A partir de la matriu reduïda que es deriva de l'ACP, de 4 variables de puntuacions factorials que identifiquen les 3586 seccions censals, hem considerat afegir, per la importància desigual de la grandària de les seccions en relació al municipi i els seus efectes sobre la mostra, una cinquena variable que manifestés explícitament aquesta relació.

En el procés de classificació automàtica, l'aplicació de l'anàlisi factorial previ ens proporciona dues avantatges d'interès: l'obtenció de quatre components factorials que són les variables reduïdes que més discriminen a les seccions censals i la incorrelació d'aquestes, dues característiques essencials en la construcció dels grups o estrats. El mètode de classificació emprat, a partir de considerar com a mesura de proximitat la distància quadràtica euclidiana, es desglossa en dues etapes: primer s'aplica una classificació jeràrquica ascendent mitjançant el procediment ward (o de mínima pèrdua d'inèrcia), en un segon moment, amb el nombre de grups determinat i els centres inicials definits, s'opera una classificació no jeràrquica per l'agregació al voltant de centres mòbils amb l'objectiu d'optimitzar l'assignació als estrats. D'aquesta forma finalment s'obtenen 8 estrats, la distribució de freqüències i els centres finals dels quals es presenten a la Taula 3.[6]

[6]El procés de classificació comporta a més diversos procediments de validació: una anàlisi sistemàtica de les classificacions entre 15 i 5 estrats, la comparació amb el resultats obtinguts en edicions anteriors de l'enquesta, la comparació del procediment de classificació *ward* amb altres jeràrquics ascendents (distàncies màximes, mínimes, distàncies promig, mediana i centroide), la replicació de la classificació a partir de submostres així com la utilització de criteris teòrics-interpretatius.

Taula 3. Distribució final de les seccions censals per estrat i centres finals

Estrat	Seccions finals		Centres finals				
			FSC1	FSC2	FSC3	FSC4	ZP23
1	159	4,5%	0,1163	0,5210	1,8965	0,2306	1,6172
2	661	18,5%	-0,6742	0,8633	-0,0267	0,2349	-0,0464
3	539	15,1%	0,5725	1,0295	-0,2933	-0,5932	-0,2002
4	71	2,0%	0,2549	-0,3544	5,5014	-0,0575	6,1028
5	642	18,0%	-1,1779	-0,3439	-0,2648	0,7265	-0,2211
6	449	12,6%	-0,2770	-0,8112	-0,1391	-1,5785	-0,2509
7	647	18,1%	0,5030	-0,8780	-0,2312	0,0141	-0,2832
8	400	11,2%	1,6390	-0,0754	-0,3357	0,9127	-0,2850
Total	3568	100,0%	0,0000	0,0000	0,0000	0,0000	0,0000

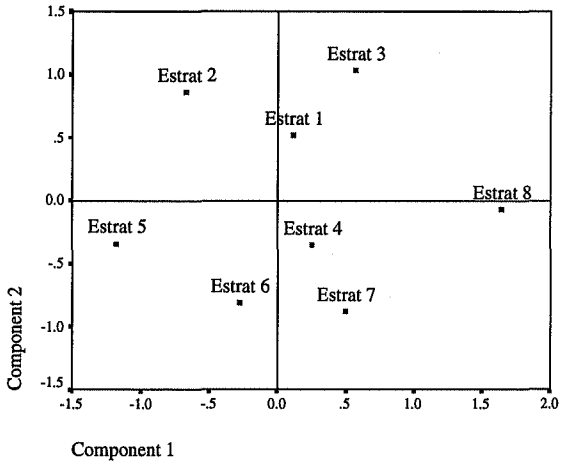
La caracterització sociològica dels vuit estrats és la que presentem breument tot seguit, de manera descriptiva i esquemàtica[7]. Les dades de referència es poden observar a la Taula A3 de l'annex, amb les mitjanes i desviacions de les diferents variables, i visualitzar en els Gràfics 3 i 4, d'ubicació dels estrats en els eixos factorials. La caracterització dels estrats es presenta a continuació en ordre decreixent d'acord amb la gradació de trets que es configuren amb la categoria socioprofessional, el nivell d'estudis, l'ocupació i l'origen de naixement.

L'estrat 8 identifica a la població de seccions censals on predominen els grups o les posicions socials de classe alta, són persones formades i ocupades, nascudes a Catalunya, és una població madura i urbana, i amb una taxa d'activitat femenina alta.

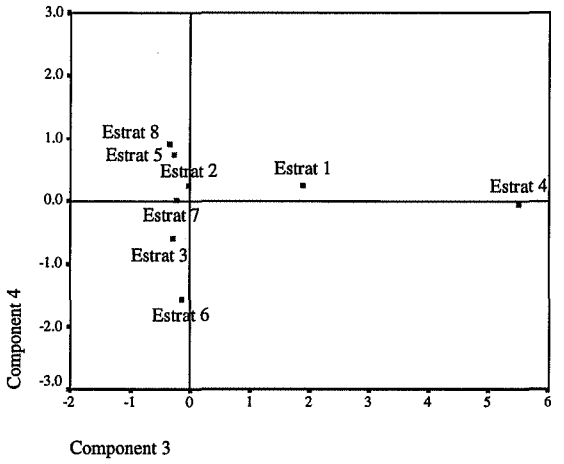
L'estrat 7 es correspon amb una posició social de classe mitjana, mitjana-alta, de persones formades, d'origen català, major índex d'envelliment, de medi urbà i amb menor activitat femenina.

L'estrat 3 també es caracteritza per una posició social de classe mitjana, mitjana-alta, formada i d'origen català, si bé es tracta d'una població més jove i on es dona la major taxa d'ocupació femenina; es correspon a una situació mitjana segons el grau d'urbanització.

[7]Insistim en que es tracta d'una identificació bàsicament descriptiva que no incorpora consideracions de naturalesa teòrica ni es vincula amb dades externes a les aquí tractades. De la mateixa forma no s'ha realitzat una anàlisi comparativa de desigualtats i disparitats entre els estrats si bé s'apunta en certa mesura en la pròpia descripció.



Gràfic 3. Representació dels estrats a les components 1 i 2



Gràfic 4. Representació dels estrats a les components 3 i 4

L'estrat 1 recull seccions censals amb posicions socials de classe mitjana-baixa i amb un nivell educatiu mitjà-baix. És població d'origen català, d'edats intermèdies o joves, localitzada als municipis més petits amb una presència important de l'activitat agropecuària.

L'estrat 4 coincideix amb l'anterior en el predomini de classe mitjana-baixa i nivells educatius mitjans-baixos, així com per la localització en petits nuclis de població. És

una població ocupada, d'edats intermèdies sobre tot, on l'origen català és més accentuat i on també s'intensifica l'activitat al sector agropecuari.

L'estrat 6 presenta nivells professionals i educatius intermedis, en canvi l'atur és dels més alts. L'origen immigrant es combina amb el català, essent un dels grups amb una població més vella, ubicada en nuclis urbans.

L'estrat 2 agrupa a la població de categoria professional baixa i nivells formatius baixos, amb xifres d'atur per sobre de la mitjana. Es tracta de les seccions amb major proporció de població jove, amb alts nivells de persones d'origen immigrant, ubicades a municipis de diferents grandàries.

L'estrat 5 finalment reprodueix, accentuant-la, la imatge de l'estrat anterior: major proporció de categories professionals baixes, mínims nivells educatius, màximes taxes d'atur i amb major presència de persones d'origen immigrant. En aquest cas es tracta de població menys jove i que resideix a municipis de més grandària.

A continuació l'exposició seguirà amb les dues etapes finals del disseny de mostra proposat: el càlcul de la grandària de la mostra i l'afixació, i la ponderació a posteriori de la mostra.

2.4. Grandària mostral, afixació de la mostra i ponderació a posteriori

Una vegada determinats els estrats amb la caracterització socioeconòmica obtinguda i com a expressió de conjunts homogenis de seccions censals, correspon determinar la distribució dels individus de la mostra en aquests estrats a partir de la determinació d'un nombre total. El càlcul de la grandària mostral s'inscriu en la fixació dels paràmetres bàsics: grandària de la població, estimacions de la mitjana i de la variabilitat, nivell de significació i error mostral.

Com a mesura de variabilitat es considera la distància quadràtica euclidiana de cada secció al centre global del núvol de punts que forma l'espai vectorial dimensionalitzat i reduït, i com a paràmetre la mitjana d'aquesta distància. D'aquesta manera s'aconsegueix estimar i reflectir a la mostra, no tan sols una característica determinada d'interès de l'estudi, sinó tot un conjunt, atès que es pren un punt mitjà i una desviació d'aquest conjunt de característiques dimensionalitzades a partir de les dades censals.

Si considerem un nivell de significació de 2σ i un error mostral relatiu de l'1,76% per a dades globals de la RMB, la grandària mostral n surt d'aplicar la fórmula següent:

$$n = \frac{z^2 \cdot \sigma_y^2}{e^2 \cdot \overline{Y}^2}$$

on: z és el nombre de sigmes de nivell de significació.

σ_y^2 és la variança de la distància quadràtica euclidiana de les seccions censals al centre de la totalitat del núvol (valor obtingut de 0,2794).

e és l'error mostral relatiu.

$\overline{Y}$ és la mitjana de les distàncies quadràtiques euclidianes (valor obtingut de 1,0654).

i on s'obté un total de 5200 individus.

Aquest nombre d'individus va ser distribuït entre els estrats amb el criteri d'afixació òptima de Neyman. Amb aquest criteri s'opera l'efecte de l'estratificació de la mostra segons el qual quan més gran i variable és un estrat major proporció de mostra se l'assigna. Per tant no es tracta d'una distribució estrictament proporcional a la població de cada estrat, sinó que a efectes d'optimització, de guany en la precisió de les estimacions, s'adopta aquest doble criteri que s'expressa a la fórmula:

$$n_h = \frac{N_h \cdot \sigma_h}{\sum_{h=1}^{K} N_h \cdot \sigma_h} \cdot n$$

on: n_h és la grandària mostral de l'estrat $h (h = 1 \ldots 8)$

N_h és la població major de 15 anys de l'estrat h

σ_h és la desviació de la distància quadràtica euclidiana de les seccions censals de l'estrat h al centre del seu estrat.

Els valors que s'obtenen de l'afixació es presenten a la Taula 4.

Amb el nombre mostral de cada estrat es procedeix a l'assignació proporcional de quotes de mostra en termes d'individus corresponents a cada secció censal de l'estrat. Amb aquest repartiment es garanteix l'acompliment de l'aleatorietat de la mostra en l'elecció d'un individu que pertany a una secció determinada. L'assignació de quotes es fa segons la fórmula:

$$n_{sh} = \frac{N_{sh}}{N_h} \cdot n_h$$

on: n_{sh} és la quota de mostra de la secció s de l'estrat h

N_{sh} és la població major de 15 anys de la secció s l'estrat h

N_h és la població major de 15 anys de l'estrat h

n_h és la grandària mostral de l'estrat h

Taula 4. Valors d'afixació mostral òptima de Neyman per a cada estrat

Estrat	N_h	σ_h	$N_h.\sigma_h$	Coeficient d'afixació	n_h
1	223.835	0,6025869	134.880,0	0,0820091	426,45
2	834.715	0,3801628	317.327,6	0,1929400	1.003,29
3	588.235	0,4386580	258.034,0	0,1568886	815,82
4	64.298	1,1315886	727.58,88	0,0442385	230,04
5	627.659	0,4814569	302.190,7	0,1837366	955,43
6	323.600	0,5777149	186.948,5	0,1136676	591,07
7	519.470	0,4496771	233.593,8	0,1420286	738,55
8	331.165	0,4196152	138.961,9	0,0844909	439,35
Total	3.512.977	0,5285464	1.644.695,4	1,0000000	5.200,00

L'assignació que s'obté dona quotes de mostra no enteres per a cada secció censal, per la qual cosa, a efectes d'elecció dels individus, es computa l'arrodoniment generant una grandària mostral final de 5263 enquestats/des que varen ser extrets aleatòriament del marc mostral que es deriva del Cens Electoral corresponent a l'any 1994.

Cal assenyalar per tant que aquest disseny mostral estratificat suposa l'afixació no proporcional de les quotes de mostra de cada estrat. Segons la grandària poblacional de l'estrat i la variabilitat de les característiques socioeconòmiques pròpies d'aquest, hi haurà individus majors de 18 anys que tindran una probabilitat major de ser elegits a partir de la quota que s'assigna a l'estrat on s'ubica la secció censal a la qual pertany, és a dir, no es garanteix el criteri d'equiprobabilitat quan un individu és elegit a l'atzar. Aquest criteri té un sentit instrumental doncs ens assegura la presència a la mostra d'aquelles característiques menys freqüents a la població, però, alhora, sobredimensiona la presència dels individus que les posseeixen. En conseqüència, una vegada obtinguda la mostra, es procedeix a la restitució del valor real de les freqüències ponderant el seu pes en el conjunt i així garantir una mostra estrictament aleatòria.

Aquesta ponderació és una magnitud que transforma la probabilitat real que un individu hagi estat escollit en la probabilitat teòrica sota hipòtesi d'estricta aleatorietat, i que es pot expressar amb la relació següent:

$$PES = \frac{\text{Probabilitat Teòrica}}{\text{Probabilitat Real}} = \frac{N_h/N}{n_h/n}$$

És a dir, s'atorga un menor pes a aquells individus que tenen una probabilitat major de ser elegits, i un major pes a aquells altres amb una probabilitat menor.

3. COMENTARIS FINALS

El procediment seguit per a la construcció de la mostra estratificada de l'Enquesta Metropolitana de Barcelona i els criteris i decisions que s'utilitzen van més enllà dels resultats habituals esperats del disseny mostral, ja que el procés de construcció proporciona conclusions d'interès per a l'anàlisi sociològica o pel coneixement social del territori, el que pot ser també de gran ajut en la tasca de planificació de l'Administració. Al mateix temps l'anàlisi realitzada serveix de criteri de control i validació de les conclusions posteriors a la pròpia anàlisi de les dades obtingudes a l'enquesta.

En aquest sentit i com a resultat d'una anàlisi específica que es pot efectuar amb posterioritat a la recollida d'informació, és possible reutilitzar la mostra amb l'objectiu de construir el que hem anomenat com a *zones socials* (Lozares i Domínguez, 1993, 1996), i considerar les necessitats presents habitualment en els estudis que plantegen el tractament de submostres per àmbits territorials més restringits amb errors mostrals que siguin acceptables donades les limitacions econòmiques. El que aquí proposem és que cada estrat de la mostra estratificada pot constituir una base mostral per a dur a terme anàlisis amb totes les garanties de precisió, com a mínim calculables i acceptables, sense augmentar la grandària de la mostra per a l'enquesta general.

El procediment consisteix en considerar els estrats homogenis generats en la construcció de la mostra com a zones socials, geogràficament localitzables i sense restricció de contigüitat territorial, als quals se'ls hi apliquen les dades obtingudes a l'enquesta. Per tant, les zones socials són caracteritzades separadament per la informació de l'enquesta amb un error mostral acceptable que en molts casos garanteix la representativitat. D'altra banda en aquestes zones socials es poden realitzar les anàlisis que es considerin adients, en particular de dimensionalització i estructuració de grups socials a l'interior de les zones on es poden observar noves diversitats i perfils. Posteriorment, la comparació de les diferents zones genera models generals de composició de grups i la idea d'una certa dinàmica social.

BIBLIOGRAFIA

Cochran, W.G. (1971). *Técnicas de muestreo*. México: CECSA.

Grosbas, J.M. (1987). *Méthodes statistiques des sondages*. Paris: Economica.

Lébart, L.; Morineau, A.; Fénelon, J.P. (1985). *Tratamiento estadístico de datos*. Barcelona: Marcombo.

López Roldán, P. (1996). «La construcción de tipologías: metodología de análisis». *Papers. Revista de Sociologia*, 48, 9-29.

Lozares, C.; López, P. (1990). *Enquesta Metropolitana de la Regió Metropolitana de*

Barcelona. Construcció de la mostra estratificada. Sèrie Documents de Treball 90/1. Bellaterra: Institut d'Estudis Metropolitans de Barcelona.

Lozares, C.; López, P. (1991). «El análisis de componentes principales. Aplicación al análisis de datos secundarios». *Papers. Revista de Sociologia*, 37, 31-63.

Lozares, C.; López, P. (1991). «El muestreo estratificado por análisis multivariado». En: *El pluralismo metodológico en la investigación social: ensayos típicos*, editado por M. Latiesa. Granada: Universidad de Granada, 107-160.

Lozares, C.; Domínguez, M. (1993). *Enquesta de la Regió Metropolitana de Barcelona 1990. Territori i realitat social: les zones sòcio-demogràfiques de la Regió Metropolitana de Barcelona.* Barcelona: Mancomunitat de Municipis de l'Àrea Metropolitana de Barcelona y Diputació de Barcelona.

Lozares, C.; Domínguez, M. (1996). «Tratamiento multivariado de subpoblaciones en una gran encuesta social: la construcción de zonas sociales». *Papers. Revista de Sociologia*, 48, 71-87.

Nel·lo, O.; Recio, A.; Solsona, M.; Subirats, M. (1998). *Enquesta Metropolitana de Barcelona. Condicions de Vida i Hàbits de la població. La transformació de la societat metropolitana.* Barcelona: Mancomunitat de Municipis de l'Àrea Metropolitana de Barcelona i Diputació de Barcelona.

Sánchez Carrión, Juan Javier (Ed.) (1984). *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales.* Madrid: Centro de Investigaciones Sociológicas.

ANNEX

Taula A1. Matriu de correlacions de les variables emprades a l'ACP

	P1	P2	P4	P7	P8	P9	P11	P12	P14	P15	P16	P17	P18	P19	P20	P23
P1	1,000	-,791	,248	,250	-,276	-,178	,102	,108	,430	-,316	,472	-,241	-,323	,109	,553	,165
P2	-,791	1,000	-,417	-,271	,258	,161	-,092	-,145	-,405	,371	-,517	,290	,336	-,030	-,495	-,091
P4	,248	-,417	1,000	,571	-,705	-,597	,520	,373	-,206	-,719	,756	-,128	-,622	-,211	-,074	-,222
P7	,250	-,271	,571	1,000	-,607	-,660	,530	,506	-,183	-,557	,665	-,133	-,577	,056	,161	-,023
P8	-,276	,258	-,705	-,607	1,000	,845	-,647	-,427	,335	,955	-,884	-,062	,767	-,151	-,034	-,107
P9	-,178	,161	-,597	-,660	,845	1,000	-,672	-,520	,281	,801	-,781	-,045	,724	-,170	-,078	-,138
P11	,102	-,092	,520	,530	-,647	-,672	1,000	,441	-,155	-,626	,566	,137	-,526	-,107	-,091	-,166
P12	,108	-,145	,373	,506	-,427	-,520	,441	1,000	-,234	-,406	,413	,024	-,401	-,040	-,035	-,101
P14	,430	-,405	-,206	-,183	,335	,281	-,155	-,234	1,000	,267	-,189	-,158	,245	-,096	,269	-,016
P15	-,316	,371	-,719	-,557	,955	,801	-,626	-,406	,267	1,000	-,888	-,073	,739	-,125	-,045	-,091
P16	,472	-,517	,756	,665	-,884	-,781	,566	,413	-,189	-,888	1,000	-,214	-,818	,098	,280	,122
P17	-,241	,290	-,128	-,133	-,062	-,045	,137	,024	-,158	-,073	-,214	1,000	,029	-,039	-,327	-,085
P18	-,323	,336	-,622	-,577	,767	,724	-,526	-,401	,245	,739	-,818	,029	1,000	-,169	-,204	-,146
P19	,109	-,030	-,211	,056	-,151	-,170	-,107	-,040	-,096	-,125	,098	-,039	-,169	1,000	,341	,667
P20	,553	-,495	-,074	,161	-,034	-,078	-,091	-,035	,269	-,045	,280	-,327	-,204	,341	1,000	,458
P23	,165	-,091	-,222	-,023	-,107	-,138	-,166	-,101	-,016	-,091	,122	-,085	-,146	,667	,458	1,000

Determinant de la Matriu de Correlacions = 0,0000004

Mesura d'Adequació Mostral de Kaiser-Meyer-Olkin = 0,83302

Test d'Esfericitat de Bartlett = 52574,674 (Significació = 0,00000)

Taula A2. Matriu factorial o de saturacions de l'ACP

	Component			
	1	2	3	4
P8 Titulats mitjans-superiors majors 20 anys	,935	-,032	-,134	,121
P15 Professions altes	,911	-,128	-,101	,155
P16 Professions baixes	-,904	,268	,117	,164
P9 Escolarització 14-24 anys	,891	,044	-,166	,072
P18 Terciari alt/finances	,828	-,125	-,190	-,019
P4 Inmigració fora Catalunya	-,798	,133	-,306	,148
P7 Analfabets majors de 10 anys	-,749	,045	-,010	,252
P11 Aturats abans ocupats	-,742	-,011	-,192	-,230
P12 Atur busca primera feina	-,592	-,101	-,126	,103
P1 Joves de menys de 15 anys	-,248	,863	,133	,092
P2 Vells majors de 65 anys	,293	-,844	-,013	-,186
P14 Dones actives majors 15 anys	,371	,747	-,127	-,073
P20 Vehicle privat treball	-,023	,569	,525	,320
P23 Població Secció/Municipi	,008	,081	,896	,020
P19 Agropecuari	-,047	-,033	,880	-,015
P17 Terciari mitjà/comerç/hosteleria	-,009	-,192	-,036	-,915

Mètode de rotació: Normalització Varimax amb Kaiser

Taula A3. Descripció dels estrats

	Mitjanes de cada estrat								Total
	Estrat 1	Estrat 2	Estrat 3	Estrat 4	Estrat 5	Estrat 6	Estrat 7	Estrat 8	RMB
Joves de menys de 15 anys	19,61	21,18	19,71	17,22	16,02	12,31	11,56	14,23	16,24
Vells majors de 65 anys	11,70	8,25	10,25	14,72	12,03	20,57	20,69	16,35	14,23
Index d'envelliment > 65/ < 15	,63	,41	,57	,91	,84	1,87	1,99	1,31	1,10
Inmigració fora Catalunya	25,66	41,84	30,29	16,34	46,53	32,17	28,17	20,18	33,59
Estrangers	1,62	1,31	1,95	1,37	1,12	2,53	1,92	3,60	1,91
Analfabets majors de 10 anys	1,87	3,11	,95	1,41	5,16	1,71	,91	,37	2,18
Titulats mitjans-superiors majors 20 anys	23,57	17,19	33,61	21,74	14,31	22,01	33,40	57,84	27,63
Escolarització 14-24 anys	53,91	50,66	65,79	49,94	44,93	52,15	64,62	79,15	57,96
Aturats abans ocupats	9,51	13,85	10,92	8,03	15,26	14,58	10,74	7,43	12,16
Atur busca primera feina	2,28	3,63	2,22	2,18	4,94	3,45	2,63	2,17	3,20
Atur total	11,80	17,48	13,14	10,21	20,20	18,03	13,37	9,60	15,36
Dones actives majors 15 anys	39,29	40,47	45,10	35,72	34,53	35,19	36,34	41,80	38,69
Professions altes	15,44	8,81	20,75	13,85	6,97	13,80	23,77	42,68	17,82
Professions baixes	43,77	54,58	29,94	41,06	55,36	32,94	25,30	10,59	37,28
Terciari mitjà/comerç/hosteleria	17,41	17,23	20,34	17,77	16,91	26,55	19,33	15,93	19,07
Terciari alt/finances	3,21	1,91	5,91	2,67	1,61	3,79	6,19	8,43	4,28
Agropecuari	4,72	,80	,34	12,59	,60	,57	,28	,15	,91
Vehicle privat treball	65,36	51,76	44,37	68,98	41,61	30,35	34,33	44,44	43,09
Vehicle privat estudi	6,70	2,21	2,91	9,47	1,15	1,12	1,78	4,49	2,51
Vehicle privat treball + estudi	31,91	21,04	20,70	36,18	15,19	11,69	14,47	22,29	18,49
Població Secció/Municipi	28,43	3,85	1,58	94,68	1,28	,83	,36	,33	4,54

ENGLISH SUMMARY

DESING AND CONSTRUCTION OF A STRATIFIED SAMPLE FROM CENSUS DATA*

P. LÓPEZ*
C. LOZARES*
Universitat Autònoma de Barcelona
M. DOMÍNGUEZ**
Universitat de Barcelona

The purpose of the article is to show the most important features of the design and construction process of the stratified sample from the «Enquesta metropolitana de Barcelona. Condicions de vida i hàbits de la població» *(Metropolitan survey of Barcelona. Life conditions and habits of the population) 1995 edition. Persons are gathered in census sections according to the information from the population census and strata are constructed by means of a procedure in which techniques of multivariate analysis are implied: principal components factor analysis and cluster analysis. Finally, a sample size is determined and allocated according to Neyman's optimum criterium. The procedure followed and criteria and decisions used go beyond the usual results expected in any sample since it provides interesting conclusions for sociological analysis and territorial planning as well as it may be used as validation criteria for later conclusions in the survey's analysis.*

Keywords: Sample surveys, data analysis, factor analysis and principal components, cluster analysis

AMS Classification: 62D05, 62-07, 62H25, 62H30

⋆ The three authors are members of the *Grup d'Estudis Sociològics sobre la Vida Quotidiana i el Treball* (QUIT) of the Departament de Sociologia of the Universitat Autònoma de Barcelona.

* Pedro López Roldán (Pedro.Lopez.Roldan@uab.es); Carlos Lozares Colina (Carlos.Lozares@uab.es). Departament de Sociologia. Universitat Autònoma de Barcelona.

** Màrius Domínguez Amorós (marius@riscd2.eco.ub.es). Departament de Sociologia. Universitat de Barcelona.

– Received September 1999.
– Accepted November 1999.

1. INTRODUCTION

The purpose of this article is to show the most relevant features regarding the design and construction process of the stratified sample which is the basis for information collection of the «Enquesta Metropolitana de Barcelona. Condicions de Vida i Hàbits de la Població» (EM, Metropolitan Survey of Barcelona. Life Conditions and Habits of the Population) 1995 edition, carried out by the «Institut d'Estudis Metropolitans de Barcelona» (Institute for Metropolitan Studies of Barcelona).

The EM is a permanent research which was designed in 1984 in order to analyse activities, life conditions and lifestyles of population in the Metropolitan Area of Barcelona, facing the lack of systematic and objective statistic data, of social character and about these contents. The EM has become a periodical tool for collecting information which helps us to specify the evolution and most structural tendency changes of all social phenomena which are analysed in the survey. Up to now, there have been three editions (1985, 1990, 1995) and the organisation of a fourth edition is being planned for 2000.

The EM collects information from the design of a stratified sample which has been kept through the three editions, however, the specific territorial scope has been modified. Next, the most relevant features of the sampling design and construction procedure are specified.

2. CONSTRUCTION OF THE SAMPLE FROM «METROPOLITAN SURVEY»

2.1. Sample design: population universe and criteria variables

Extracting a stratified random sample which was representative of the population in the Metropolitan Area of Barcelona was the objective established in the design of the sample. Choosing the stratified sample procedure is basically justified by precision criteria, compared to other methods, and by social heterogeneity which is a feature of the target population. Thus, strata construction enables us to have a stratification variable which will be a dense variable of social typification and which will guarantee the presence in the sample of homogeneous social types, which are characteristic of the population, types or social groups defining a basic variable which correlates with other targets in the EM.

The population universe is defined as the collection of persons above 18 in the RMB (Metropolitan Area of Barcelona), a territory including 162 municipalities and a total population of 3,275,458 people according to data from the Population Census in 1991. From the information available in the census survey, we have selected some indicators

which are the ones acting as variables/stratification criteria, and which will define, as a result of people addition, the census sections in the RMB. Thus, we have not considered individuals in the census as basic units to be stratified at first stage, but census sections where these individuals live. This procedure is justified by a double reason: because of difficulties deriving from dealing with such a big matrix of individuals; and because only addition in census sections enables us to use a metric percentage treatment of the selected variables. Consequently, we have a data matrix with 3,586 census sections in the RMB with the selection of the 16 indicators-variables of demographic characterising, cultural-educative, of labour and professional activity, of mobility and of population extension.

2.2. Strata construction

From this information we proceed to construct strata in a procedure in which techniques of multivariate analysis are involved: principal components factor analysis and cluster analysis. The stratification of the sample is based on the principle of considering the studied population without homogeneity as far as statistics is concerned. Statistic homogeneity refers to the relation established amongst different features in the same population, the more correlated these features are, the more homogeneous the population will be considered. The purpose of stratification will then be the classification of census sections in strata which, with the object to sampling, and improving in precision, will be the expression of as much as possible homogeneous groups of sections inside every stratum, and as much as possible heterogeneous amongst them according to variables/factor criteria deriving from the previous analysis.

This classification is made without considering any restriction in territorial continuity, so the result will be a map of RMB sections divided into a number of strata with different socio-economic features.

With the principal components factor analysis (ACP), we attempt to reduce original information in order to obtain a vector subspace where factors are bases and therefore lineally independent, and which, in a hierarchical order, keep most of the total variation. By this means, we obtain the fundamental dimensions of population difference in the RMB, which initially structure social reality according to the information introduced. So the ACP is thought as a previous stage and necessary complement of categorisation of units, of strata obtaining. In the analysis presented, we finally retain 4 components.

In the reduced matrix deriving from the ACP, from 4 factor scoring variables identifying the 3,586 census sections, we proceed to cluster analysis. The classification process must be divided into two different stages: in the first one, we apply a forward hierarchical clustering (ward's method), in which a first classification in 8 classes or strata is established; in a second stage, with the number of groups determined and initial centres

defined, a non-hierarchical classification operates out of addition around mobile centres in order to optimise strata assignments.[8]

2.3. Sample size, allocation and sample weighting

Once the strata have been defined as the expression of uniform sets of censal sections, the individuals of the sample are distributed among the strata on the basis of the predetermined sample size. The sample size is calculated by determining the basic parameters: the size of the population, the estimations of mean and its variability, the significance level, and the sampling error.

We compute the Euclidean distance of each section to the global centroid of the cluster of points in space of the dimensionalised factors as a measure of variability, and the mean of this distance as the parameter. In this way, it was possible to estimate and reflect in the sample not only a specific characteristic of interest of study, but also a whole set of characteristics, given that a midpoint was taken as well as a deviation of this set of dimensionalised poblation characteristics.

If we take a significance level of 2 sigma and a relative sampling error of 1,76%, the calculation of the sample size was carried out using the following formula:

$$n = \frac{z^2 \cdot \sigma_y^2}{e^2 \cdot \overline{Y}^2}$$

where: z Number of sigma of the significance level.

σ_y^2 Variance of the squared Euclidean distance from the censal sections to the centroid of the cluster as a whole (the resulting value was 0,2794).

e Relative sampling error.

$\overline{Y}$ Mean of the squared Euclidean distances (the resulting value was 1,0654).

This resulted in a total of 5,200 individuals. Using this figure we distributed the individuals among the different strata. The homogeneity of these strata facilitate the proper allocation of the individuals who belong to the censal sections of a specific stratum, using Neyman's optimum allocation criterion. It was the use of this optimum allocation process that allowed us to obtain the real effect of the stratification of the sample. According to Neyman's optimum allocation criterion, the distribution of the 5,200 individuals is not carried out in strict proportion of the population of each strata, but with a

[8] Classification process also means different validation procedures: a systematic analysis of classification between 15 and 5 strata, comparison between results obtained in previous editions of the survey, comparison between ward classification procedure and other upward hierarchical methods (average linkage between groups, average linkage within groups, complete linkage, centroid clustering and median clustering), repetition of the cluster analysis from sub-samples as well as use of theoretic-interpretative criteria.

view to optimising the result, that is, in order to increase the accuracy of the estimations. To this end, a two-fold strategy is adopted in which the larger and the more variable the strata, the larger the proportion of the sample that is allocated to it. This is expressed in the formula:

$$n_h = \frac{N_h \cdot \sigma_h}{\sum_{h=1}^{K} N_h \cdot \sigma_h} \cdot n$$

where: n_h Sample size of the strata $h(h = 1 \ldots 8)$

N_h Population over 15 in strata h

σ_h Standard deviation of the squared Euclidean distances from the censal sections of strata h to the centroid of its strata.

n Resulting sample size.

Once the sample size for each strata was obtained, the sample quotas were allocated to the different censal sections that make up each strata. This procedure ensures the randomness of the sample in the selection of an individual belonging to any given selection. Quotas were allocated according to the following formula:

$$n_{sh} = \frac{N_{sh}}{N_h} \cdot n_h$$

where: n_{sh} Sample quota of section s of the strata h.

N_{sh} Population over 15 in section s of strata h.

N_h Population over 15 strata h.

n_h Sample size of strata h.

The resulting allocation yielded sample quotas for each population section that were not integers, and consequently, for the purposes of selecting individuals, these figures were rounded off to the nearest integer. After arounded-off, the sample size was ultimately established at 5,263 individuals, who were selected at random from the 1994 Electoral List.

It should be noted that this sample design involved non-proportional allocation of the sample quotas of each strata. Depending on the population size of the strata, and the variability of the socio-economic characteristics of the strata, there will be individuals of over 18 who have a greater probability of being selected from the quota allocated to the strata containing the censal section to which that individual belongs, i. e. equal probability is not guaranteed when an individual is selected at random. This criteria make sense in this context because it ensure the presence in the sample of characteristics that are less common in the population, since these are the more variable phenomena. At the same time, however, this criteria gives rise to overrepresentation of individuals with such characteristics. Consequently, once we obtained the sample, we restored the

real values of the frequencies by weighting them according to the proportion of the overall sample they represented in order to guarantee a truly random sample.

This weighting is a figure that transforms the real probability that an individual has been selected into the theoretical probability governed by the hypothesis of strict randomness, and can be expressed as follows:

$$WEIGHT = \frac{Theoretical\ probability}{Real\ probability} = \frac{N_h/N}{n_h/n}$$

where: N_h Population over 18 in censal section h.
N_h Population over 18 in the Region.
n_h Sample quota of section h.
n Total sample.

3. FINAL COMMENTS

The procedure followed to construct the sample and criteria and decisions used go beyond the usual expected results in the sample design since it provides interesting conclusions for the sociological analysis or of social knowledge of the territory which might be of a great help in the planning task of the Administration. At the same time, it might be used as control and validation criteria for later conclusions in the survey's analysis.

On the other hand, and as a result of a specific analysis carried out after information collection, the sample has been used again with the purpose of constructing what we have called «social areas». Every stratum in the stratified sample may constitute a sampling base in order to carry out analysis with all precision guarantees, at least calculable and acceptable, without enlarging the extension of the sample for the general survey. Analysis of these areas with the survey's data also enables to obtain a social profile, a structural analysis of dimensionalisation inside the homogeneous stratum without corresponding with an administration unit.

QÜESTIIÓ, vol. 24, 1, p. 137-150, 2000

COMPARATIVE ANALYSIS OF ALTERNATIVE SAMPLING PLANS TO CREATE A FARM ACCOUNTANCY DATA NETWORK FOR THE AGRICULTURAL SECTOR OF NAVARRA

LUCINIO JÚDEZ
CAROLINA CHAYA
Escuela Técnica Superior de Ingenieros Agrónomos de Madrid*

This study presents the method that was followed and the results of the analysis for establishing a sampling plan for the Farm Accountancy Data Network (FADN) of the commercial agricultural sector of the Spanish Autonomous Community of Navarra. The first part of the study presents the categories considered of the different stratification criteria for the commercial farms of Navarra: geographical units (subregions), types of farming (TF) and economic size in ESU (European Size Unit). Then the authors define sampling plans with different objectives to be compared. These objectives are: i) maximum accuracy in the estimation of the Standard Gross Margin (SGM) for the whole of the commercial agricultural sector, ii) the same accuracy in the various types of farming (TF) and iii) the same accuracy in the individual strata. Special attention is given to the effects of introducing geographical units as a stratification criterion. Given the sample size and the characteristics of the population of commercial farms in Navarra the plan without geographical stratification that gives approximately the same accuracy in each TF seems to be the most suitable sampling plan for the FADN of this region. Even though this study is limited to Navarra, it may be of help when considering sampling plans for FADN not only in other Autonomous Communities but also on a national scale.

Keywords: Farm accountancy data network (FADN), stratified sampling, sample surveys, Navarra

AMS Classification: 62D05

* Unidad de Estadística. Departamento de Economía y Ciencias Sociales Agrarias. Escuela Técnica Superior de Ingenieros Agrónomos de Madrid. Ciudad Universitaria s/n. 28040 Madrid (Spain).

– Received November 1998.
– Accepted November 1999.

1. INTRODUCTION

In spite of the increasing use of data from farm accountancy networks in European countries to analyse various aspects of the agricultural sector, the accuracy of the information used is not usually studied. Publications on this subject are few and, in general, of restricted diffusion.

This may be due to the complexity of obtaining samples for the networks with which the accuracy of the estimates, relevant to the numerous variables collected from the farms, can be determined.

And this is perhaps the reason why each country in the European Union adopts a different method of obtaining the sample of its farm accountancy data network (FADN), as shown in the Table of Appendix 1, taken from Commission of the European Communities (1989).

So it is not surprising that the statistical sample design of the FADN is one of the aspects suggested for improvement by a group of experts of the concerted action Pacioli[1] (Beers et al., 1995, p. 58).

The aim of this study is to present the method followed to obtain the sample of the FADN of the Autonomous Community of Navarra, a territorial Unit for Statistics of the European Union[2].

The field of observation is the commercial agricultural sector of Navarra that includes farms of an economic size larger than or equal to 4 European Size Units (ESU)[3], which were 11388 in 1989, the date of the last agricultural census in Spain.

There is wide consensus that the sample of an FADN should be obtained after stratification of the field of observation (population), as well as of the stratification criteria to be taken into account: types of farming (TF), economic size (measured in ESU) and geographic area (subregions). These are the criteria suggested by the Commission of the EU and accepted by member states[4].

[1]The concerted action Pacioli (Panel in Accounting of Innovation Offering a Lead-up to the use of Information modelling) recently submitted to the Management Committee of the FADN of the EU a series of reflection papers with suggestions for improvement (Poppe and Beers, 1995 and 1996a; Poppe et al., 1996b). These papers proceed from workshops attended by international experts; the discussions and summaries have been published (see references of the publications in the above-mentioned reflection papers).

[2]The name given by Eurostat to the geografical units that are the base of the organisation of the FADN of the European Union.

[3]A European Size Unit (ESU) is a number of ECUs (1200 in 1997) of standard gross margin. For more information on the definition of this and any other term used in the Farm Accountancy Data Network refer to Commission of the European Communities (1989).

[4]It should be noted, however, that the categories of criteria used differ from one country to another (see

In our opinion, a sample plan for building a Farm Accountancy Data Network for a region must allow for the analysis of different results in the sector (by subregion or TF, for example) without losing sight of the fact that they are integrated in the whole of the regional agricultural sector. The network must also allow a study of the evolution of the sector.

On this basis, the objectives we set will be less ambitions the smaller the size of the sample, i.e. from a small sample size we cannot expect to obtain very accurate estimates in farms of a particular size, belonging to a TF and a particular subregion. In this case, good estimates for the whole TF, the whole region and/or the whole size class, will have to suffice, and it is not possible to obtain accurate estimates in a single stratum: size class * subregion * TF, unless a large part of the sample was allocated to this stratum of the population, thus losing the aforementioned global perspective.

In the case of Navarra, budget limitations restrict the sample size to 400 farms per year, so the problem is to obtain the distribution of the fixed sample size among the different strata of the population.

The first part of this study describes the methodology used, in particular, the classes or groups of different stratification criteria considered by the FADN of Navarra before this study and how they have been modified to obtain the sample. The evaluation procedure and the allocation method of the sample are stated later, as well as the different sampling plan objectives to be compared.

The second part gives the results of comparing the various sampling plans, each one corresponding to a different objective, and the influence of the introduction of subregions as a stratification criterion.

The information for this study was provided by the Department of Agriculture of the Government of Navarra.

As already stated, although the key objective of the analysis of the different plans is to obtain a sampling plan for the FADN of Navarra, we are not aware of studies of farm accountancy networks in which sampling plans with different objectives are compared, nor do we know of studies which analyze the effect of considering or not considering the stratification by geographic units. We believe that these aspects, although referring here only to Navarra, can help in drawing up sampling plans in other Autonomous Communities, as well as on a national scale. Given the diversity of plans used in the EU the problem is still unsolved.

Commission of the European Communities, 1989) and that some countries add other stratification criteria in their national networks, such as age of the farmer in Holland (Boers and al. 1994) or the farm area and the farm system of work, full or part-time, in Denmark (Institute of Agricultural Economics, 1994).

2. METHODOLOGY

2.1. Starting Point

The number of agricultural subregions called «comarcas» usually considered in the analysis of the agricultural sector of Navarra is 7[5], and the number of TFs in this Community is 51. The FADN for Navarra considered 8 classes of economic size, in ESU, defined by the boundaries[6]:

4-6, 6-8, 8-12, 12-16, 16-40, 40-60, 60-100 and >100

Consequently, the field of observation should be divided in a first approximation into:

7 SUBREGIONS * 51 TFs * 8 SIZE CLASSES = 2.856 STRATA

Although many of these strata are empty, the number with at least one farm, which is 1036, is much higher than the sample size. So if we want the farms represented by the sample are to be close to the field of observation[7], it is necessary to consider fewer categories in one or more of the three stratification criteria; those taken into account to analyze the distinct sampling plans are presented next.

2.2. Categories of the stratification criteria

The seven «comarcas» have been considered in the plans when including the geographical stratification criterion.

The decision on the TFs, or more precisely on the aggregations of TFs, to be considered was one of the most difficult aspects in establishing a sampling plan, since it is impossible to avoid a certain amount of subjectivity in its handling. Table 1 shows the TF groups finally selected, along with the codes adopted in the study (Roman Numerals I to XI), their composition and relative importance in relation to the standard gross margin (SGM) and the total number of farms in the commercial agricultural sector of Navarra.

[5]Nord-occidental, Pirineos, Cuenca de Pamplona, Tierra de Estella, Navarra media, Ribera alta and Ribera baja.

[6]Muñoz Segura, J.C. and Beperet Aizkorbe, M. (1993, p.19).

[7]All or nearly all the strata have to be sampled.

Table 1. Aggregation of the TFs for the fadn of Navarra. Importance of the groups of TFs selected

Notation of the group in this document	Composition of the group (*)	% of the total SGM of the population	% of the total number of farms of the population
TF I	111+112+113	20.0	18.9
TF II	123	9.6	12.3
TF III	1244	8.1	11.0
TF IV	311	3.2	3.7
TF V	411+412	8.2	7.9
TF VI	441	12.8	8.6
TF VII	421+422+431+432+442+443+444	7.2	8.9
TF VIII	5011+5012+5013+5021+5022+5023+ 5031+5032	7.9	3.3
TF IX	121+122+1241+601+602+603+604+605+6061+6062	12.1	14.8
TF X	711+712+721+722+723+811+813+814+821+822+8232	7.6	6.5
TF XI	2011+2012+2013+3211+3213+340	3.4	4.2

(*) The codes of TF and their meaning correspond to those in the «Commission Decision 85/377/EEC, of 7 June 1985», O.J. n° L220, 17-8-85.

To obtain these groups, the relative importance of each of the 51 initial TFs in relation to the standard gross margin of the field of observation in 1989 is used as a starting point. The basic criteria for the creation of the 11 groups were the following: i) not to disregard any TF. This allows estimates for all the field of observation and the analysis of the commercial agricultural sector of Navarra, as well as an indication of its evolution from a global perspective. ii) not to aggregate the most important TFs, so that they can be studied separately. iii) to aggregate the less important TFs with a certain degree of similarity[8].

Finally, five farm size classes were adopted, in ESU, which in EU terminology[9] correspond to: 1. Small (4-8 ESU), 2. Medium-low (8-16 ESU), 3. Medium-high (16-40 ESU), 4. Large (40-100 ESU) and 5. Very Large ((100 ESU).

This division of farm sizes, the one suggested by the Commission of the EU, also corresponds to the optimum aggregation into five classes, if all commercial farms of Navarra are initially divided into the eight classes mentioned earlier. This partition is optimum in relation to the criterion of the accumulated distribution of the square root of the number of farms, proposed by Dalenius and Hodge (1959).

[8]For further details of this or of other aspects of the methodology see Júdez and Chaya (1994).
[9]See Commission of the European Communities (1989, p. 4)

2.3. Evaluation Procedure

The essential characteristic used to evaluate the sample of each plan is the coefficient of variation of the estimator of the total of the SGM, which is the same as that of the mean of the SGM[10]. This coefficient is obtained for: i) all the commercial farms of the agricultural sector of Navarra, ii) the farms belonging to the various TFs, iii) the farms of the different «comarcas», when a geographical stratification is used, and iv) the farms of each of the individual strata.

2.4. Allocation of the sample to the strata

The two most usual methods of allocating sample numbers to different strata, the proportional method and Neyman method, are compared. Table 2 shows the coefficients of variation of the estimators of the total SGM for the whole of commercial agricultural sector of Navarra and for each type of farming, when the sample is distributed among the strata TF * SIZE CLASS using the proportional method (representative sample) and using Neyman method (optimum allocation).

Table 2. Coefficients of variation, as percentages, of the estimators of the total SGM with the neyman allocation and with the proportional allocation for all the commercial farms of Navarra and for different TFs

	Allocation		**(Proportional / / Neyman)*100**
	Neyman	**Proportional**	
Navarra	1.14	4.23	372
TF I	2.66	3.87	145
TF II	3.75	5.59	149
TF III	4.04	4.15	103
TF IV	7.10	8.20	116
TF V	3.98	6.48	163
TF VI	3.05	26.66	874
TF VII	4.34	5.20	120
TF VIII	3.36	22.44	668
TF IX	3.19	9.12	286
TF X	4.18	7.88	188
TF XI	6.29	7.27	115

[10]In the future when we talk of the accuracy of the estimator of the SGM, we refer indistinctly to the estimator of its mean and of its total, since they have the same coefficient of variation.

The coefficient of variation for the whole of Navarra is 272% higher when proportional allocation is used. This percentage varies between 3% and 774% in the estimates of the different TFs. The considerable advantage of Neyman method over the proportional allocation in our case led us to compare sampling plans which we now analyze, using only the optimum allocation.

2.5. Sampling plans to compare

Once the categories of different criteria to stratify the field of observation have been defined, the sampling plans corresponding to the following objectives are evaluated: i) maximum accuracy in the estimates of the whole of the commercial agricultural sector of Navarra, ii) the same accuracy in the estimates of each TF, iii) the same accuracy in each of the strata.

These plans are first evaluated without introducing the «comarcas» as a stratification criterion; the consequences of its introduction studied later[11].

3. RESULTS

3.1. Sampling plans without geographical stratification

Table 3 shows the coefficients of variation of the estimators of the SGM of the sampling plans associated with each of the three objectives mentioned earlier, and Table 4 presents the characteristics (maximum value, average and coefficient of variation) of the coefficient of variation of the estimators of the SGM in the distinct TF * SIZE CLASS strata[12] for these plans. We make the following comments on these Tables:

1. The best accuracy in the estimates of the SGM for the whole of the field of observation is obtained, logically, in the plan whose objective is to maximise this accuracy, which corresponds to a coefficient of variation of 1.14%. This plan presents a strong heterogeneity in the accuracy of the estimates of the SGM in each TF (coefficients of variation varying between 2.66% and 7.10%) and in each stratum, where the average of the coefficient of variation is approximately 10%.
2. In the plan designed to obtain the same accuracy for the different TFs the average of its coefficient of variation, 3.74%, is the lowest of the three plans studied.

[11] The determination of the SGM, its mean and variance in the new strata when the categories of some stratification criteria were aggregated or divided the allocation of the sample to the strata and the evaluation of different sampling plans were carried out using Fortran programs elaborated by the authors.

[12] This concerns the characteristics of the non zero coefficients of variation.

3. The homogeneity of the coefficients of variation of the estimators of the SGM of the strata, for the plan designed to obtain the same accuracy in the estimates of the SGM of each stratum, is accompanied by a very high coefficient of variation for these estimators (about 8 % on average).

4. Having to disregard good accuracy at an individual stratum level (the average of the coefficients of variation of the estimators of the SGM in the strata go from 8% to 10% in the plans studied), the choice of a sample plan must be made by comparing the accuracy of the estimators of the total SGM for all the commercial farms of Navarra and for each of the TFs.

In view of these results, the plan aimed at achieving accuracy among the TFs seems to be the most interesting; it leads, in general, to the most accurate estimates in each TF, and the accuracy of the estimator of the SGM for all the commercial farms of Navarra is only slightly inferior to that of the plan aimed at optimising the accuracy of this estimator. The distribution of the sample among the strata is shown in Table 5.

Table 3. Coefficients of variation (CV), as percentage, of the estimators of the total SGM for all the commercial farms of Navarra and for the TF, according to different objetives

	Objectives		
	Maximum accuracy Navarra estimate	**Same accuracy TF estimates**	**Same accuracy strata estimates**
Navarra	1.14	1.26	1.35
TF I	2.66	3.76	4.21
TF II	3.75	3.74	3.92
TF III	4.04	3.71	4.16
TF IV	7.10	3.76	4.13
TF V	3.98	3.71	3.18
TF VI	3.05	3.74	3.62
TF VII	4.34	3.72	4.10
TF VIII	3.36	3.77	4.22
TF IX	3.19	3.76	3.95
TF X	4.18	3.75	3.93
TF XI	6.29	3.73	3.82
Average of TF coefficients	4.18	3.74	3.99
CV (%) of TF coefficient	31.07	0.50	4.48

Table 4. Maximum value, average and coefficient of variation (CV) of the coefficients of variation (%),of the estimators of the total SGM of the strata according to different objectives

	Objectives		
	Maximum accuracy Navarra estimate	**Same accuracy TF estimates**	**Same accuracy strata estimates**
Maximum	22.845	21.315	9.401
Average	10.004	9.031	8.057
CV (%)	37.035	32.235	4.378

Table 5. Sample to obtain the same accuracy of the estimators of the total SGM of each TF

TFs	**SIZE CLASS**					**Total**
	1	**2**	**3**	**4**	**5**	
TF I	5	8	18	8	1	40
TF II	9	8	10	4	6	37
TF III	9	11	10	2	3	35
TF IV	6	8	15	7	1	37
TF V	5	9	11	5	6	36
TF VI	2	6	12	8	4	32
TF VII	8	11	12	5	1	37
TF VIII	1	2	7	8	23	41
TF IX	8	10	9	3	7	37
TF X	3	7	13	7	8	38
TF XI	8	6	9	5	2	30
TOTAL	64	86	126	62	62	400

3.2. Influence of the introduction of the geographical stratification

The introduction of the geographical stratification, which, as seen above, adds a criterium with seven categories, increases the number of strata (11 TFs * 5 SIZE CLASS) to 385. Of these strata, 90 contain no farms.

If the aim is to obtain unbiased estimates, it is necessary to sample at least one farm from each stratum with farms. This means that given the small sample size, in many

of the strata only one farm can be selected, and that the handling of this sampling plan will be more difficult than if the subregions were not used as a stratification criterion[13].

Table 6 allows us to compare the coefficients of variation of the estimators of the SGM, in the whole commercial agriculture sector of Navarra and in the different TFs, when the allocation of the sample is made with and without geographical stratification[14]

Table 6. Coefficients of variation (%) of the estimators of the total SGM for the allocations with and without geographical stratification for all the commercial farms of Navarra and for the different TF

	Without geographical stratification	**With geographical stratification**
Navarra	1.14	1.34
TF I	2.66	3.20
TF II	3.75	4.34
TF III	4.04	4.66
TF IV	7.10	8.18
TF V	3.98	4.75
TF VI	3.05	3.72
TF VII	4.34	4.78
TF VIII	3.36	4.18
TF IX	3.19	3.77
TF X	4.18	4.81
TF XI	6.29	6.78

The Table shows that the accuracy of the estimators of the SGM for all the commercial farms of Navarra and for each TF worsens when the subregion is introduced as a stratification criterion.

This disadvantage of the stratified sample by «comarcas», illustrated in Table 6 in the context of a sampling plan to obtain the maximum accuracy in the estimate of the SGM for the whole of the field of observation, is also found when the objective is to reach

[13]Difficulties will be found in obtaining farms for all the 293 strata, especially given the small number of farms in many of them (33 strata contain only one farm).

[14]When geographical stratification is used, the coefficients of variation of the estimator of the SGM in each «comarca» are: 3.32 in the Nord-occidental, 4.54 in Pirineos, 3.98 in Cuenca de Pamplona, 3.57 in Tierra de Estella, 3.88 in Navarra media, 2.96 in Ribera alta and 3.16 in Ribera baja.

the same accuracy in each TF. The lower level of accuracy is unexpected. In our case, it is due to a combination of two facts: the large number of strata with a small number of farms, and the small size of the sample compared to the number of strata. A detailed analysis of this result can be found in Júdez and Chaya (1999).

4. DISCUSSION AND CONCLUSIONS

We have seen that in our case, the optimum allocation of the sample gives estimates that are considerably more precise than those using proportional allocation. However, this type of allocation calls for a high proportion of holdings of the largest size classes in the sample[15], due to the high variance of the SGM in these classes. This can create a problem if it is difficult to rely on the collaboration of the large farms with the network. In this case, although one or more of the farms that should appear in the sample is not surveyed and the objective sample cannot be reached, it is foreseen that the estimates obtained from this «possible» sample are more accurate than those that would be reached with a proportional allocation. One can also analyze the larger strata in detail and detect the farms which contribute most to the variance of the strata, in order to include all of them in the sample. Then, the rest of the farms in the strata can show a decrease in the variance, and in consequence the number of farms to be included in the sample of these strata will also decrease.

For estimates of means and totals, a weighting is needed for each stratum. This makes the estimates with the sample in which the «comarca» is not considered as a stratification criterion, simpler and less subject to error than those when this criterion is adopted. The problem of weighting is worsened if the same weights are maintained for ten years (time between two censuses). Besides, as seem above, the estimates are less accurate in our case, and the sampling plan more difficult to handle, when the geographic stratification is considered.

If it were essential to obtain estimates at a subregional level, the handling of the sampling plans dealt with in this study can be improved in two ways: one, by reducing the number of categories of some of the stratification criteria (the criterion chosen could be the economic size) and secondly, by not sampling the less important strata. The first option when reducing the number of economic size classes increase the variance of the estimates[16], and the second, used by some EU countries lead to biased estimates. An analysis of the consequences of this later procedure in our case will be made in a future study.

[15] In some TFs, the optimum allocation contains all the farms of the size class (100 ESU.

[16] For a detailed analysis of the effects on the SMG estimates of reducing the categories of the stratification criteria considered here, see the above mentioned work of Júdez and Chaya (1999).

Considering the disadvantages of the sampling plans with geographical stratification if estimates by subregion are not essential, as in the case of Navarra[17], a sampling plan in which these are aggregated seems to be more suitable. To try to obtain good estimates in the individual strata does not make much sense in our case given the size of the sample, and these estimates are not very accurate in any of the plans studied.

When the estimates of great interest are those related to the different TF, the sampling plan without subregional stratification, aimed at of obtaining comparable accuracy in the estimates of the distinct TFs, is the best of the plans studied. In fact, this plan, not only provides the greatest homogeneity in the accuracy of the estimates of the SGM of the TF, but it also gives the smaller coefficient of variation of the estimator in each TF. The accuracy that can be expected in the estimate of the whole commercial agricultural sector of Navarra with this sampling plan is very close to that of the plan aimed at maximising the accuracy of this estimate.

ACKNOWLEDGEMENTS

This study is part of a large research project spread over several years, for modelling the agricultural sector using FADN data, funded by the Government of Navarra and the CICYT and now funded by the Commission of the European Union.

The authors would like to thank Mr. Robson and Mr. Vard of the Directorate General of Agriculture (Unit VI/A-3) of the EU Commission, and Mr. Teurlay, responsible for the French network, for the information they kindly provided on the farm accountancy data networks.

The authors also thank FADN delegates of several EU countries for the information provided about the sampling plan and the procedures for estimation used in their countries, in particular Mr. Price (United Kingdom), Miss Abitabile (Italy), Mr. Poppe (Holland), Mr. Mollenberg (Denmark) and Mr. Goffinet (Institute of Agricultural Economics of Belgium).

The authors consider themselves responsible for any errors in this article.

[17] In Navarra, the estimates of the agricultural sector of greatest interest at a subregional level, are actually already obtained by other procedures.

REFERENCES

Beers, G., Poppe, K.J., Spiering, D.F. and Pruis, H.C. (1995). *Pacioli 1; Farm Accoun tancy Data Networks and Information Analysis.* Workshop Report. The Hague: Agricultural Economics Research Institute (LEI-DLO). Mededeling 532.

Boers, A., Dijk, J., van Dijk, J.P.M., Poppe, K.J. and Welten, J.P.P.J. (1994). *Report on Farm Selection 1993 and Selection Plan 1994.* The Hague, Agricultural Economics Research Institute (LEI-DLO). Periodeke Rapportage 4-93.

Commission of the european communities (1989). *Farm Accountancy Data Network: An A to Z of methodology.* Luxembourg: Office for Official Publications of the European Communities.

Cochran, W.G. (1977). *Sampling Techniques.* New York: John Wiley and Sons. (Traducción española en C.E.C.S.A., Mexico, 1980).

Dalenius, T. and Hodges, J.L., Jr. (1959). «Minimum variance stratification», *Jour. Amer. Stat. Assoc.*, 54, 88-101.

Diario oficial de las comunidades europeas n° L220/1 de 17.8.85: «Decisión de la Comisión 85/377/CEE de 7 de junio de 1985 por la que se establece una tipología comunitaria de las explotaciones agrícolas».

Goffinet, R. (1986). «Actualisation du plan d'échantillonage pour le réseau de comptabilités agricoles de l'Institut Economique Agricole (IEA)». *Documento n°3 de l'IEA.* Ministère d'Agriculture: IEA, Bruselas.

Institute of agricultural economics (1994). *Agricultural Accounts Statistics 1993/94.* Serie A nr. 78. Dinamarca, Valby (Copenhagen).

Júdez, L. and Chaya, C. (1994). *Elementos para el establecimiento de un plan de muestreo para la red de información contable agrícola de Navarra.* Documento de trabajo. Unidad de Estadística de la E.T.S. Ingenieros Agrónomos de Madrid.

Júdez, L. and Chaya, C. (1999). «Effects of Geographical Stratification in a Farm Accountacy Data Network on the Accuracy of the Estimates». *Journal of Agricultural Economics*, 50, 388-399.

Muñoz Segura, J.C. y Beperet Aizkorbe, M. (1993). *Red Contable Agraria. Navarra 1992.* Gobierno de Navarra. Departamento de Agricultura, Ganadería y Montes. Secretaría Técnica. Serie Agraria, n° 14.

Poppe, K.J. and Beers, G. (1995). *Pacioli 1; On data management in farm accountancy data networks; Reflection Paper.* The Hague: Agricultural Economics Research Institute (LEI-DLO). Mededeling 533.

Poppe, K.J. and Beers, G. (1996a). *Pacioli 2; On innovation management in farm accountancy data networks; Reflection Paper.* The Hague: Agricultural Economics Research Institute (LEI-DLO). Mededeling 535.

Poppe, K.J., Beers, G. and Pruis, H.C. (1996b). *Pacioli 3; RICA: Reform Issues Change the Agenda; Reflection Paper.* The Hague: Agricultural Economics Research Institute (LEI-DLO). Mededeling 537.

Appendix 1. Sampling procedures in different countries of the european union

	Determination of sample size			**Method of selection**
	A fixed number of farms from each cell in the field of observation	A fixed proportion of farms from each cell in the field of observation	A variable proportion-taking account of variability in the field of observation	
Belgique			Yes	Non-random
Danmark			Yes	Random
Deutschland			Yes (Neymann-Tschuprow)	Random
Ellas	Yes			Non-random
España		1 % (except for large cells where sample is increased)		Non-random
France			Yes	Random
Ireland	Yes			Random
Italia			Yes (Neymann-Pearson)	Non Random
Luxembourg		Yes		Random
Nederland			Yes	Non-random
Portugal		Yes		Non-random
United Kingdom			Yes	Random

Source: Commission of the european communities (1989, p. 21).

QÜESTIIÓ, vol. 24, 1, p. 151-186, 2000

AVANTATGES I INCONVENIENTS DE LA METODOLOGIA DE L'IDESCAT/INE PER ELABORAR INDICADORS DE LA PRODUCCIÓ INDUSTRIAL PER A LES REGIONS ESPANYOLES*

MIQUEL CLAR
RAÚL RAMOS
JORDI SURIÑACH
Universitat de Barcelona*

Conèixer l'evolució conjuntural del sector industrial, tant a nivell nacional com regional, és de gran importància. En aquest sentit, el retard en la publicació de les xifres de les Comptabilitats Nacionals/Regionals, fa necessària l'elaboració d'indicadors que permetin dur a terme un seguiment a curt termini de l'activitat industrial. Així, l'INE elabora un IPI mensual obtingut pel mètode directe pel conjunt de l'Estat. D'altra banda, al llarg dels darrers anys, a algunes comunitats autònomes espanyoles, s'han engegat projectes centrats en l'elaboració d'indicadors de l'activitat industrial regional, tot i que a partir de metodologies no homogènies. Per corregir aquesta situació, d'un temps ençà, a diferents fòrums s'ha proposat emprar la metodologia emprada per l'Idescat per elaborar l'indicador de la comunitat catalana com a alternativa per construir indicadors de l'activitat industrial regional, atès el seu bon comportament per a Catalunya. Així, l'INE recentment ha publicat uns IPIs per a les CA espanyoles d'acord amb dita metodologia. A aquest treball s'estudia la idoneïtat d'estendre l'esmentada metodologia a totes les regions espanyoles. Per això, es duu a terme una anàlisi comparativa centrada en tres de les (quatre) regions que disposen d'un IPI elaborat pel mètode directe: Andalusia, Astúries i el País Basc.

Advantages and disadvantages of Idescat/INE's methodology to elaborate industrial production indicators for the Spanish regions

Paraules clau: Activitat industrial, índex de producció industrial, indicadors regionals, conjuntura

Classificació AMS: 62P20, 90A19

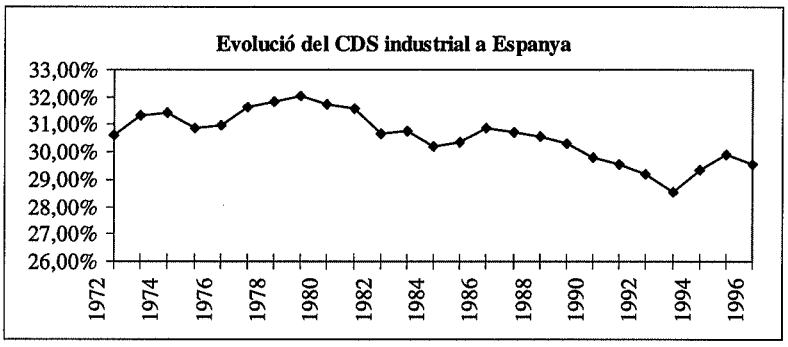
*Els autors agraeixen el suport rebut de la DGICYT projecte SEC99-0700 i del Plan Nacional de I+D projecte 2FD97-1004-C03-01.

* Miquel Clar (mclar@eco.ub.es); Raúl Ramos (rrlobo@eco.ub.es); Jordi Suriñach (surinach@eco.ub.es). Grup de recerca *Anàlisi Quantitativa Regional*. Universitat de Barcelona. Av. Diagonal, 690. 08034 Barcelona.

– Rebut el juny de 1999.
– Acceptat el desembre de 1999.

1. INTRODUCCIÓ

Tot i el procés de terciarització que han experimentat les economies occidentals al llarg de les darreres dècades, l'activitat industrial té (segueix tenint) un pes important. En aquest sentit, el nostre país no és una excepció: la participació del sector industrial en termes de Valor Afegit Brut (VAB) respecte al total espanyol s'ha mantingut al voltant del 30% als darrers vint-i-cinc anys (vegi's gràfic 1.1)[1], la qual cosa vol dir que la creixent importància relativa del sector serveis a l'economia espanyola s'ha produït principalment per una pèrdua de pes de l'agricultura i de la construcció (vegi's gràfics 1.2 a 1.6). A més a més, cal assenyalar que una part important del creixement (en termes de VAB) experimentat pel sector terciari és degut al desenvolupament d'activitats relacionades amb la indústria, en concret el creixement que al llarg dels darrers anys ha viscut el subsector dels serveis destinats a les empreses.

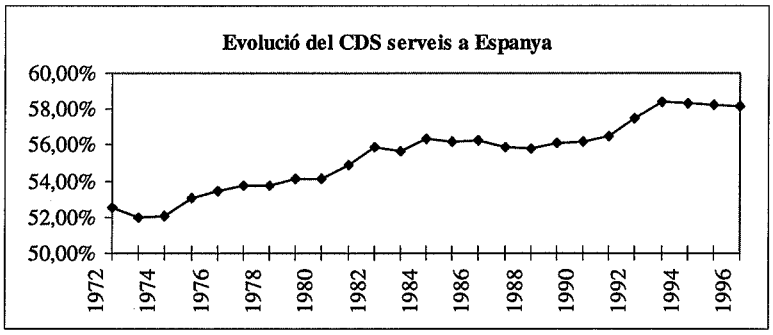


Gràfic 1.1.

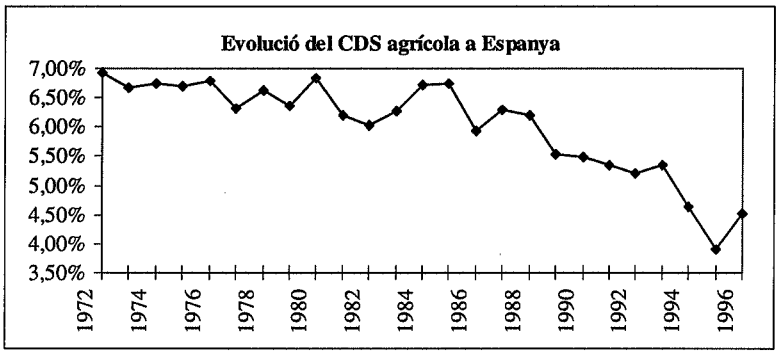
D'altra banda, cal tenir en compte l'efecte d'arrossegament que exerceix el sector industrial sobre la resta i l'elevat pes del comerç exterior de productes industrials en la demanda agregada. D'acord amb tot l'anterior, doncs, conèixer el comportament de l'activitat industrial és (segueix sent) clau a l'hora de caracteritzar l'evolució tant a curt com a llarg termini de les economies i, en conseqüència, té (segueix tenint) interès analitzar l'evolució del sector industrial.

[1] CDS simbolitza el coeficient de distribució (o participació) sectorial, estadístic que recull el pes relatiu de les diferents branques d'activitat en l'economia d'un determinat àmbit territorial. Per tant, ofereix informació sobre la composició sectorial d'una variable. Per a una magnitud X (el VAB en aquest treball) i un sector j es calcula com segueix:

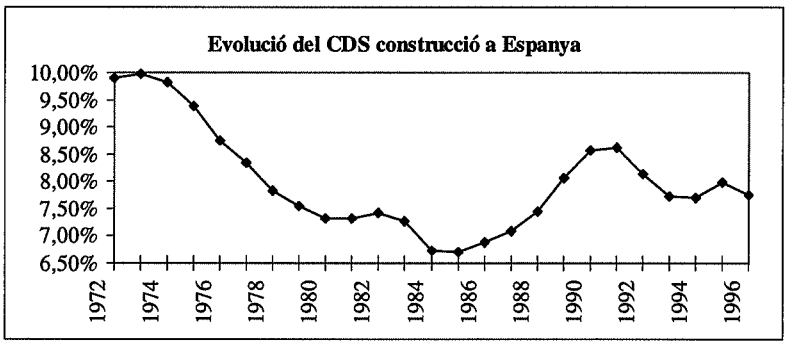
$$CDS_j = \frac{X_j}{\sum_{j=1}^{J} X_j}$$



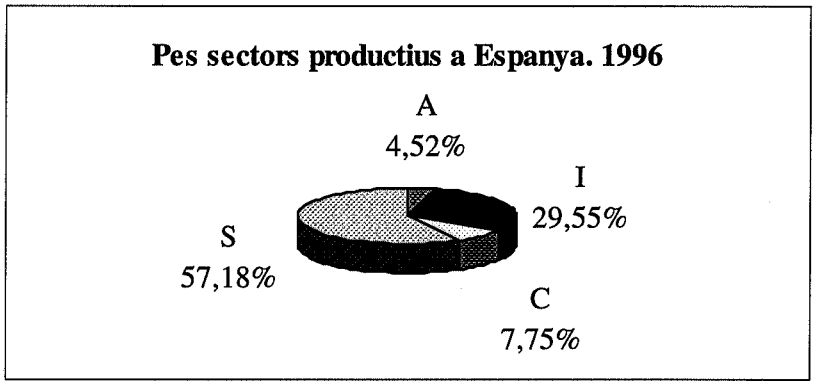
Gràfic 1.2.



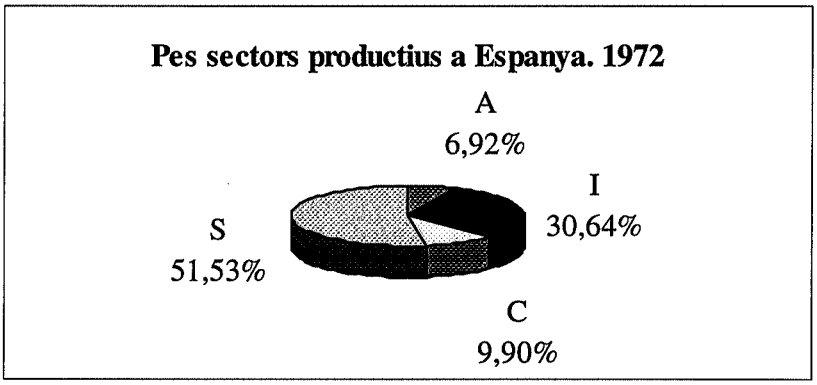
Gràfic 1.3.



Gràfic 1.4.



Gràfic 1.5.



Gràfic 1.6.

Per dur a terme aquest seguiment s'acostuma a emprar les dades corresponents al VAB i/o al Producte Interior Brut (PIB) a preus de mercat i a cost de factors (segons si es consideren o no els impostos indirectes i les subvencions) que ofereixen les Comptabilitats Nacionals/Regionals. En concret, una anàlisi sobre l'evolució de l'activitat industrial s'hauria de centrar entorn el PIB generat pel sector industrial (PIBIN) en sentit estricte, és a dir, sense considerar la construcció. Al nostre país, així com a altres països, el principal problema que presenta emprar aquesta informació per dur a terme una anàlisi conjuntural es deriva del fet que no es disposa d'ella amb la rapidesa que seria desitjable[2], la qual cosa dificulta en gran mesura (per no dir que impedeix) la possibilitat

[2]Sobre aquest punt vegi's el treball de Muñoz *et al.* (1996) referent al retard del *Instituto Nacional de Estadística* (INE) en la publicació de les xifres oficials (definitives) de la Comptabilitat Nacional. Així

A més a més, també juga un paper rellevant en l'elaboració de Comptabilitats Trimestrals per mètodes indirectes.

Al nostre país, l'INE elabora un indicador mensual quantitatiu pel seguiment de l'activitat industrial del conjunt de l'Estat, el *Índice de Producción Industrial* (IPI), a partir de la informació obtinguda d'una enquesta adreçada a una mostra representativa d'unitats productives de tots els sectors d'activitat (mètode directe)[7]. D'aquesta manera, doncs, per l'àmbit nacional resta solucionat el problema de la manca d'informació estadística per a dur a terme una anàlisi conjuntural quantitativa industrial[8].

Però, en l'àmbit regional (fins no fa gaire temps) hi havia certes dificultats a l'hora de realitzar un seguiment quantitatiu a curt termini de l'activitat industrial, donat que existien algunes deficiències quant a la disponibilitat d'informació estadística d'aquestes característiques[9]. Davant d'aquest marc, al llarg dels darrers anys es varen encetar a algunes regions espanyoles[10] diverses iniciatives, públiques i/o privades, l'objectiu de les quals era superar aquestes mancances. Però tot i l'important esforç realitzat, la situació era que no totes les comunitats espanyoles disposaven d'un indicador quantitatiu de l'activitat industrial i, a més a més, els indicadors regionals existents no eren directament comparables atès que s'empraven metodologies no homogènies per elaborar-los[11].

Així doncs, a diferents fòrums es va encetar un debat centrat en estudiar quina era la metodologia més adient per elaborar indicadors de la producció industrial regional

[7]Per a un detall sobre el procés seguit per l'INE en l'elaboració de l'IPI pel conjunt de la indústria espanyola vegi's INE (1982) i Eurostat (1978). D'altra banda, a Clar (1998) pot trobar-se una anàlisi comparativa entre l'actual IPI (base 1990) i l'anterior (base 1972) elaborats per l'INE.

[8]A més a més, cal tenir en compte que el *Ministerio de Industria y Energía* (MINER) elabora, d'acord amb la metodologia establerta en el sí de la *Dirección General de Asuntos Económicos y Sociales*, un índex mensual qualitatiu pel conjunt de la indústria espanyola anomenat *Indicador de Clima Industrial* (ICI) a partir dels saldos de tres de les variables (tendència de la producció, nivell de la cartera de comandes i nivell d'estocs) sobre les que la *Encuesta de Opiniones Empresariales*, realitzada pel propi MINER, recull informació. Per a un detall sobre el procés d'elaboració d'aquest índex pot consultar-se Comisión de la CE (1991) o Cordero *et al.* (1996), entre d'altres.

[9]De tota manera, cal assenyalar que el MINER elabora uns ICIs de periodicitat mensual per a les regions espanyoles seguint fonamentalment les directrius fixades per la *Dirección General de Asuntos Económicos y Sociales*. A més a més, a algunes CA existeixen diverses iniciatives l'objectiu de les quals és elaborar indicadors qualitatius pel seguiment de l'activitat industrial de les seves regions. Així, per exemple, a Catalunya la Cambra Oficial de Comerç, Indústria i Navegació de Barcelona (COCINB) elabora un indicador qualitatiu de periodicitat bimensual que es publica a *Perspectiva Econòmica de Catalunya*.

[10]Andalusia, Astúries, Balears, Canàries, Catalunya, Extremadura, Madrid, Navarra, País Basc i La Rioja.

[11]Per a una anàlisi sobre les metodologies emprades per l'*Instituto de Estadística de Andalucía* (IEA), la *Sociedad Asturiana de Estudios Económicos e Industriales* (Sadei), els Governs de Balears i de Canàries, l'Institut d'Estadística de Catalunya (Idescat), la *Dirección General de Planificación y Presupuestos de la Consejería de Economía, Industria y Hacienda de Extremadura*, l'*Instituto de Estadística de la Comunidad de Madrid* (IEM), el Govern de Navarra, l'*Instituto de Estadística del País Vasco* (Eustat) i el Govern de la Rioja, per a elaborar els indicadors quantitatius pel seguiment de l'activitat industrial per a les regions d'Andalusia, Astúries, Balears, Canàries, Catalunya, Extremadura, Madrid, Navarra, el País Basc i La Rioja respectivament, vegi's Clar (1998).

amb un alt grau de fiabilitat i que, a l'hora, tingués associat un cost baix[12]. El resultat d'aquest debat ha estat que l'INE, recentment, ha publicat uns indicadors de la producció industrials per les regions espanyoles seguint un mètode indirecte molt semblant a la metodologia que l'Idescat empra des de fa anys per elaborar l'indicador de la comunitat catalana[13]. En concret, dites sèries comencen a l'octubre del 1991 i fan referència l'índex general, però no es facilita informació desagregada ni per branques d'activitat ni per destinació econòmica dels béns[14]. D'aquesta manera, doncs, s'han superat (parcialment) les mancances existents en aquest àmbit fins fa poc temps.

Davant d'aquest marc, l'objectiu del present treball és analitzar la fiabilitat dels indicadors regionals obtinguts amb la metodologia que empra l'INE. Per a assolir aquest objectiu en primer lloc es presenta dita metodologia; a continuació, es duu a terme una anàlisi comparativa entre els índexs publicats per l'INE per Andalusia, Astúries i el País Basc i els IPIs elaborats per l'IEA, el Sadei i l'Eustat[15]. En tercer lloc es construeixen

[12]A l'hora d'elaborar un indicador quantitatiu per a aproximar l'evolució de la producció industrial de qualsevol economia existeixen dues vies clarament diferenciades des del punt de vista metodològic segons el mètode emprat (directe o indirecte) per a elaborar-lo. Els indicadors quantitatius *directes* s'elaboren prenent com a font d'informació dades corresponents a la producció industrial realitzada en l'economia investigada. Aquesta informació prové d'una enquesta que acostuma a ésser especialment dissenyada per a aquest fi. En aquest cas, el procés de recopilació de dades implica necessàriament dissenyar un qüestionari apropiat i definir una mostra d'unitats productives i productes que representi correctament la composició sectorial i geogràfica de la producció industrial. Sense cap tipus de dubte, doncs, aquest mètode permet obtenir els millors índexs quantitatius per a efectuar un seguiment de l'evolució de la producció industrial però presenta l'inconvenient de tenir associat un (molt) elevat cost.

D'altra banda, els indicadors quantitatius *indirectes* es caracteritzen per aproximar la producció industrial a partir d'informació preexistent. En conseqüència, l'aproximació no és (generalment) tan exacta com la que s'assoleix amb els indicadors directes, però té l'avantatge que els costos que s'han de suportar són molt més reduïts. Per aquest motiu, han estat (i estan) sent molt emprats en un gran nombre d'economies, principalment d'àmbit regional que acostumen a enfrontar-se a majors restriccions pressupostàries per a dedicar a la informació estadística.

[13]En qualsevol cas, però, cal assenyalar que l'INE no ha publicat, almenys fins avui (desembre del 1998), cap nota metodològica on es presenti la metodologia que empra per a elaborar els indicadors regionals. Únicament es coneix que «*el índice general por comunidades autónomas se obtiene calculando la estructura de ponderaciones en cada comunidad y aplicando este sistema de pesos, diferente en cada territorio, a los índices de las distintas actividades industriales según la CNAE. Para calcular las ponderaciones en cada comunidad, se han utilizado los valores añadidos de las actividades industriales en el año base del índice, facilitados por la Encuesta Industrial. El procedimiento de regionalización asegura que el índice obtenido como suma ponderada de los índices de las 17 comunidades autónomas es idéntico al índice general*» (http://www.ine.es/htdocs/daco/daco43/notaipi.htm).

[14]Aquests índexs poden consultar-se a la base de dades Tempus de l'INE (en el moment d'escriure aquest article, desembre del 1998, la darrera actualització era maig del 1998, http://www.ine.es/tempus).

[15]El fet de centrar l'anàlisi en les tres comunitats esmentades és degut a que són tres de les quatre úniques regions espanyoles que disposen d'un indicador de l'activitat industrial elaborat pel mètode directe. L'anàlisi no es realitza per a Extremadura (que és l'altra comunitat on s'elabora l'indicador pel mètode directe) donat que la *Dirección General de Planificación y Presupuestos de la Consejería de Economía, Industria y Hacienda* del Govern Extremeny (que és l'entitat elaboradora de l'índex) va començar a elaborar i publicar (a *Coyuntura Económica de Extremadura*, revista semestral editada per la *Junta de Extremadura*) l'IPI (amb

uns Índexs de Producció de Productes Industrials (IPPI)[16] per a les tres regions esmentades seguint la metodologia de l'INE i es comparen amb els IPPIs estimats a partir de la informació publicada pels índexs sectorials per les entitats regionals elaboradores. Finalment, es presenten les conclusions de l'estudi.

2. METODOLOGIA DE L'IDESCAT/INE PER A ELABORAR INDICADORS DE L'ACTIVITAT INDUSTRIAL REGIONAL

D'acord amb el comentat anteriorment, donat que es desconeix exactament la metodologia que empra l'INE per elaborar els indicadors de la producció regionals i, en canvi, sí que es disposa d'informació sobre el procediment seguit per l'Idescat per construir el de la comunitat catalana; en aquest apartat es presenta (a mode d'exemple) aquesta darrera. En aquest sentit, cal assenyalar que de la informació publicada per l'INE es desprèn que la metodologia de l'Idescat és molt semblant a l'aplicada a la resta de CA espanyoles. Així doncs, es pren com a punt de partida de l'estudi la metodologia de l'Idescat que es la publicada i, a continuació, s'analitza sota quins supòsits aquesta metodologia es podria estendre a altres regions.

2.1. Metodologia

L'indicador elaborat per l'Idescat és un indicador quantitatiu indirecte on la informació de base prové d'informació preexistent. En concret, l'Idescat pren com a punt de partida les sèries dels IPIs al màxim nivell de desagregació sectorial (quatre dígits de la CNAE-74, subgrup) elaborades per l'INE per a la indústria del conjunt de l'Estat.

Tot seguit, en una primera etapa es duu a terme un procés de censura de les sèries d'IPIs corresponents a les branques d'activitat que no són representatives en la indústria investigada (catalana) per a garantir, d'una banda que la informació de base emprada és representativa de l'estructura productiva de l'economia de la comunitat investigada (catalana) i, d'altra banda, que no s'introdueix informació sobre altres CA en l'indicador de la regió considerada (en l'indicador català). A més a més, en partir de la màxima desagregació que permet la CNAE-74 (més de dues-centes cinquanta sèries) és possible ajustar la informació de base prou bé a l'estructura industrial investigada (catalana)[17].

---

periodicitat trimestral) a partir del primer trimestre del 1996, estant disponibles en el moment de fer aquest treball les dades fins el quart trimestre del 1997. Així doncs, únicament es disposa de vuit dades la qual cosa fa que sigui una sèrie massa curta per a l'anàlisi que es realitza en aquest treball.

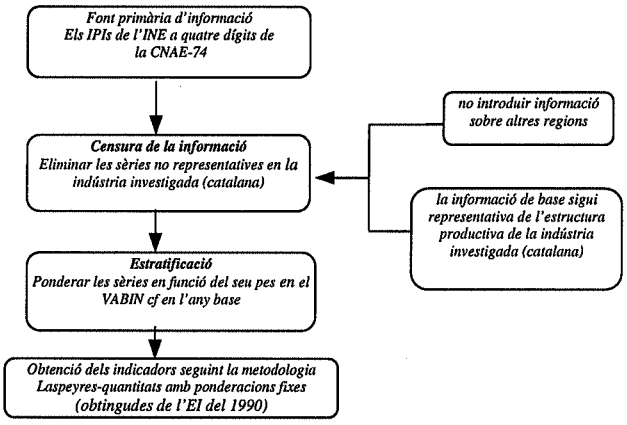
[16]El nom d'IPPI reflexa que són indicadors que aporten informació sobre la producció de productes industrials i no sobre la producció industrial. És a dir, no consideren la producció dels sectors integrants de la divisió d'Energia i aigua (divisió 1) de la *Clasificación Nacional de Actividades Económicas* del 1974 (CNAE-74).

[17]En concret, l'Idescat en l'elaboració de l'indicador català, elimina les sèries corresponents als subsectors 21 (extracció i preparació de minerals metàl·lics) i 37 (construcció naval, reparació i manteniment de vaixells)

En una següent etapa s'estratifiquen (ponderen) les sèries seleccionades a l'etapa anterior d'acord amb la importància de cada sector en el total de la producció de la indústria investigada (catalana) en termes de VAB a cost dels factors a l'any triat com a base[18]. Un cop realitzades aquestes dues etapes, el procediment consisteix en obtenir les sèries des del nivell de desagregació sectorial menor fins l'indicador general. Per això es construeixen índexs compostos (de tipus Laspeyres-quantitats) dels índexs del nivell d'agregació immediatament anterior (al quadre 2.1 es sintetitza aquest procés)[19]. Es tracta doncs, d'aprofitar la informació derivada de l'enquesta realitzada per l'INE.

Així doncs, aquesta metodologia front al mètode directe presenta l'avantatge que el seu cost és molt més reduït donat que el punt de partida per elaborar els indicadors de producció industrial regionals és el mateix per a totes les CA: els IPIs elaborats per l'INE per a la indústria del conjunt de l'Estat per branques d'activitat al màxim nivell de desagregació sectorial.

Quadre 2.1. Metodologia de l'Idescat per a elaborar l'indicador d'activitat industrial regional (català)



Font: Elaboració pròpia a partir de Costa i Galter (1994).

de la CNAE-74. A més a més, tampoc considera la producció dels subsectors energètics (divisió 1 de la CNAE-74) atès que, com assenyalen Costa i Galter (1994), es va comprovar que l'excessiva variabilitat d'aquests subsectors empitjoraven l'indicador enlloc de millorar-lo. D'aquesta manera, doncs, el nombre de sèries d'IPIs nacionals a partir de les quals l'Idescat elabora l'indicador de la indústria catalana és de cent cinquanta-tres.

[18]En concret, en el cas de l'indicador català, les cent cinquanta-tres ponderacions s'obtenen de la *Encuesta Industrial* (EI) que, al territori català, realitzen en col·laboració l'INE i l'Idescat i són fixes per l'any 1990. Aquestes ponderacions es poden trobar a Costa i Galter (1994, taula 3) o a Suriñach i Royuela (1995, annex 2).

[19]Per a un major detall sobre la metodologia emprada per l'Idescat per a elaborar l'indicador de l'activitat industrial catalana, vegi's Costa i Galter (1994).

2.2. Nota metodològica sobre la possibilitat d'estendre la metodologia de l'Idescat/INE a altres regions

Com s'ha assenyalat, la metodologia analitzada pren com a punt de partida per elaborar l'indicador regional la informació corresponent als IPIs sectorials nacionals. Es tracta doncs, d'analitzar sota quins supòsits els IPIs nacionals poden oferir una bona aproximació als indicadors regionals. Com s'ha dit anteriorment, d'acord amb el procés d'elaboració dels índexs nacionals, l'IPI general pel conjunt de l'Estat pot expressar-se com segueix:

$$IPI = \sum_{s=1}^{N} IPI_s \alpha_s, \tag{2.1}$$

on α_s representa el pes, en termes de VAB, de cada sector s en el total de la producció del conjunt de l'Estat, això és, $\alpha_s = \frac{VAB_{cfs}}{VAB_{cf}}$. Les variables IPI_s són els índexs de producció industrial nacionals corresponents a cadascun dels N sectors considerats. Aplicant aquest mateix procediment, pot obtenir-se una expressió anàloga a (2.1) per una regió j:

$$IR_j = \sum_{s=1}^{N} IR_{js} \alpha_{js}, \tag{2.2}$$

on IR_j és l'indicador de producció industrial per a la comunitat j, IR_{js} són els indicadors sectorials d'aquesta regió, i α_{js} representa el pes de cadascun dels s sectors en el total de la producció de la comunitat: $\alpha_{js} = \frac{VAB_{cf_{js}}}{VAB_{cf_j}}$. La major dificultat a l'hora d'aplicar (2.2) consisteix en obtenir les estimacions dels diferents indicadors sectorials en l'àmbit regional, IR_{js}. En qualsevol cas, però, si totes les regions disposessin d'un IPI propi, l'IPI pel conjunt de l'Estat podria obtenir-se a partir de la següent expressió:

$$IPI_s = \sum_{j=1}^{17} IR_{js} \mu_{js}, \tag{2.3}$$

on μ_{js} recull la importància que en el total de l'Estat té la comunitat j en el sector s, és a dir, $\mu_{js} = \frac{VAB_{cf_{js}}}{VAB_{cf_s}}$. A partir de (2.3) i, donat que per definició $\sum_{j=1}^{17} \mu_{js} = 1 \; \forall s$, si μ_{js} val u per la CA j, valdrà zero per la resta de comunitats, la qual cosa vol dir que la regió j és l'única regió productora dels productes del sector s. Així doncs, sempre que μ_{js} s'aproximi a la unitat per una comunitat serà pràcticament zero per la resta, de manera que serà possible obtenir una bona aproximació als índexs sectorials en l'àmbit regional a partir dels seus homòlegs nacionals. En conseqüència, si existeix un alt grau de concentració territorial de la producció (2.2) podria aproximar-se per:

(2.2.bis) $$IPI_s \approx IR_{js} \Rightarrow IR_j = \sum_{s=1}^{N} IPI_s \alpha_{js}.$$

És evident que l'expressió (2.2.bis) només es compleix estrictament quan la producció de tots i cadascun dels sectors industrials es realitza en la seva totalitat en una regió, és a dir, quan el grau de concentració territorial és del 100%. L'incompliment d'aquesta condició implicaria introduir informació d'altres regions en l'elaboració de l'indicador de la comunitat *j*. Aquest problema, però, serà menor quant major sigui el nivell de desagregació sectorial que es prengui com a punt de partida donat que, degut a l'especialització productiva de les diferents CA, augmentarà el grau de concentració territorial de la producció.

3. COMPARACIÓ ENTRE ELS INDICADORS REGIONALS DIRECTES I ELS ELABORATS PER L'INE

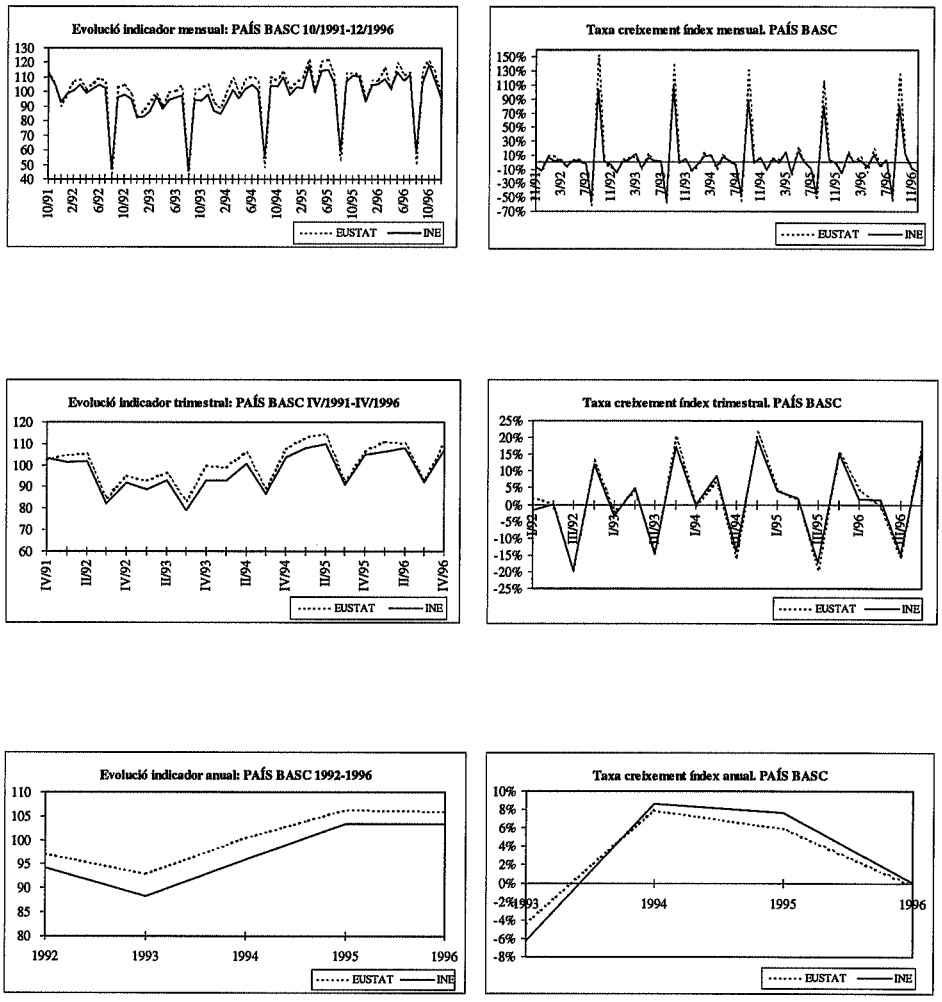
A aquest apartat es duu a terme una anàlisi comparativa centrada en les comunitats del País Basc, Astúries i Andalusia entre els indicadors indirectes elaborats per l'INE i els índexs directes elaborats respectivament per l'Eustat, el Sadei i l'IEA[20] pel període comprés entre octubre del 1991 i desembre del 1996[21]. Així, en primer lloc s'han comparat gràficament l'evolució d'ambdós indicadors en termes mensuals, trimestrals i anuals tant en nivells com en taxes de creixement (gràfics 3.1 a 3.3).

Els resultats obtinguts mostren que l'aproximació obtinguda a l'índex directe del País Basc i d'Andalusia és prou satisfactòria per a tot el període analitzat (tret del cas d'Andalusia pel període comprés entre octubre del 1992 i desembre del 1993[22]), però per a Astúries s'observen més discrepàncies.

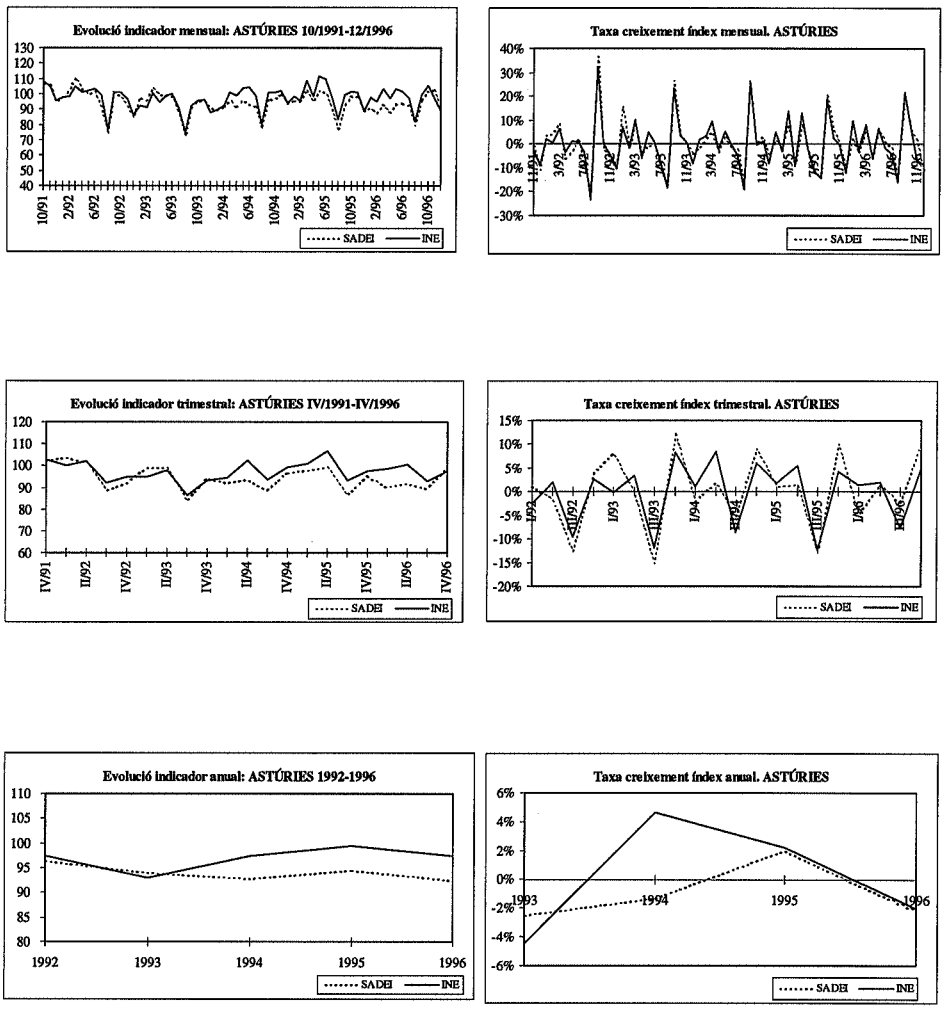
[20]S'han triat aquestes tres regions per a efectuar l'anàlisi atès que, com s'ha dit anteriorment, són (tres de les quatre) comunitats que disposen (d'unes sèries prou llargues) d'un indicador quantitatiu obtingut a partir del mètode directe.

[21]Cal assenyalar, però, que donat que l'INE no ha publicat cap nota metodològica referent al procés seguit per a elaborar els indicadors de producció regionals hi ha tot un seguit de qüestions que no es poden respondre, com ara, Quines són les ponderacions emprades per cada regió? Es duu a terme un procés de censura particularitzat per a cada CA? En aquest cas, quines són les sèries d'IPIs nacionals eliminades a l'hora d'elaborar l'indicador de cada regió? Per què l'inici de les sèries dels indicadors regionals és octubre del 1991 si la informació de base per a elaborar-los està disponible des de gener del 1975? Per què no s'ha publicat més que l'indicador general sense cap tipus de desagregació sectorial per branques d'activitat i/o per destinació econòmica dels béns quan aquesta és una de les avantatges d'aquesta metodologia front a altres metodologies indirectes com ara emprar com a *proxy* de la producció industrial el consum d'energia elèctrica per a usos industrials?

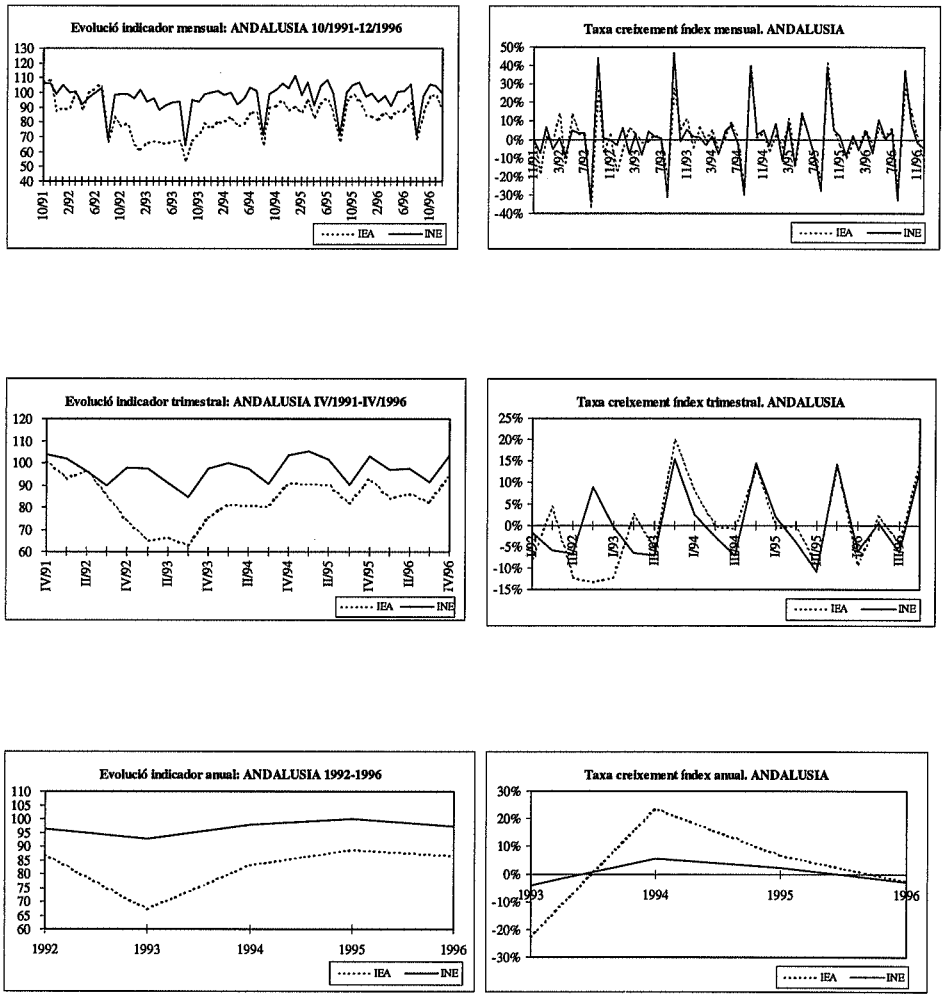
[22]De tota manera, però, l'anàlisi gràfica mostra que per l'any 1994 també es produeixen certes discrepàncies entre l'índex de l'IEA i el de l'INE que afecten als nivells però no (als signes de) les taxes de creixement.



Gràfic 3.1. Comparació entre l'indicador elaborat per l'INE i l'índex elaborat per l'Eustat pel País Basc



Gràfic 3.2. Comparació entre l'indicador elaborat per l'INE i l'índex elaborat pel Sadei per a Astúries



Gràfic 3.3. Comparació entre l'indicador elaborat per l'INE i l'índex elaborat per l'IEA per a Andalusia

Per confirmar les conclusions que es deriven de l'anàlisi gràfica realitzada s'ha calculat, d'una banda, l'error percentual absolut mig (EPAM)[23] en termes mensuals, trimestrals i anuals entre ambdues sèries (quadre 3.1) i, d'altra, el percentatge d'errors en el signe de les taxes de creixement d'ambdós índexs (quadre 3.2) per a les tres comunitats considerades.

Quadre 3.1. EPAMs entre els índexs publicats per l'IEA, el Sadei i l'Eustat i els de l'INE

	Mensual	Trimestral	Anual
Andalusia	18,10%	17,73%	18,10%
Astúries	4,42%	4,08%	3,51%
País Basc	4,83%	3,33%	3,53%

Quadre 3.2. Percentatge d'errors en el signe de les taxes de creixement dels índexs publicats per l'IEA, el Sadei i l'Eustat i els de l'INE

	Mensual	Trimestral	Anual
Andalusia	20,10%	20,00%	0,00%
Astúries	14,52%	30,00%	25,00%
País Basc	4,84%	10,00%	25,00%

* Els resultats a nivell anual s'han d'interpretar amb prudència donat que fan referència a un curt període de temps (1993 a 1996).

Els resultats obtinguts confirmen, en línies generals, les conclusions de l'anàlisi gràfica. En particular cal destacar dos fets:

a) els EPAMs obtinguts per a Andalusia prenen valors molt elevats (al voltant del 18% en tots els casos). Això és degut a que el comportament de l'indicador elaborat per l'INE i el de l'índex de l'IEA divergeixen molt significativament pel període

[23] L'Error Percentual Absolut Mig es calcula com segueix:

$$EPAM = \frac{\sum_{t=1}^{T} \frac{|Y_t - y_t|}{Y_t}}{T} 100,$$

on Y_t són els valors dels índexs de l'IEA, el Sadei i l'Eustat i y_t els dels indicadors elaborats per l'INE pel període t.

10/1992-12/1993. De fet, la major part dels errors en els signes de les taxes de creixement d'ambdues sèries es donen en l'esmentat període: en termes mensuals, es cometen set errors (d'un total de tretze) i en termes trimestrals dos (d'un total de quatre). Descomptant aquest efecte, a nivell mensual es cometen sis errors (que suposa un 9,68%), i a nivell trimestral dos (un 10%); i,

b) tot i que per Astúries en termes d'EPAM els resultats són relativament satisfactoris, al valorar els errors comesos en els signes de les taxes de creixement de l'índex del Sadei i de l'indicador de l'INE s'observa que, de les tres regions analitzades, és en la que es produeix un percentatge d'errors més elevat (si es descompta l'efecte del període 10/1992-12/1993 a Andalusia).

L'anterior porta a la conclusió que la metodologia emprada per l'INE per elaborar indicadors de producció regionals permet obtenir uns indicadors la fiabilitat dels quals no es pot garantir plenament per a totes les regions. Cal esbrinar doncs quin són els factors que determinen la fiabilitat dels indicadors obtinguts seguint dita metodologia.

4. SENSIBILITAT DE LA METODOLOGIA DE L'IDESCAT/INE A LA DISPONIBILITAT I CENSURA DE LA INFORMACIÓ BASE EMPRADA I EL PERÍODE MOSTRAL CONSIDERAT

A aquest apartat, tenint en compte l'assenyalat a la nota a peu número 21, s'estimen uns indicadors per a les tres comunitats considerades seguint la metodologia emprada per l'INE tot i que amb petites variacions[24] per tal de disposar d'unes sèries més llargues que permetin esbrinar els factors que fan que la metodologia de l'Idescat/INE sigui adequada aplicar-la a totes les CA (atès que la informació oficial de l'INE es disposa per a un període curt de temps, no presenta una desagregació sectorial). La comparació entre els indicadors obtinguts, tenint en compte aquestes variacions, i els directes permetran avançar en la identificació dels factors determinants de l'adequació de la metodologia analitzada.

4.1. Anàlisi de la disponibilitat i censura de la informació estadística de base

En primer lloc cal assenyalar que les ponderacions s'han obtingut de l'EI del 1990, elaborada per l'INE en termes de la producció bruta (donat que és la font que presenta

[24]Motivades, com es veurà més endavant, per la impossibilitat d'accedir a la informació de base al mateix nivell de desagregració que el que disposa l'INE com a conseqüència de l'aplicació de la *Ley de la Función Pública Estadística.*

informació amb un major nivell de desagregació sectorial). Així, es disposa, de ponderacions per un nivell de desagregació sectorial de vuitanta-nou sectors industrials.

Quant a la informació de base, els IPIs sectorials nacionals (base 1990), pel període comprés entre gener del 1975 i setembre del 1991 únicament s'ha tingut accés a les sèries mensuals per un nivell de desagregació sectorial de dos dígits de la CNAE-74 tot i que no per tots els sectors. D'una banda, no s'ha disposat d'informació relativa a la majoria dels sectors integrants de la divisió d'Energia i aigua per la qual cosa s'ha optat per no considerar la producció energètica en la construcció dels indicadors regionals[25]. D'altra banda, la informació corresponent al sector 49 de la CNAE-74 (altres indústries manufactureres) tampoc està disponible. Per tant, el nombre de sectors que s'han pogut considerar per elaborar els indicadors regionals pel període assenyalat ha estat de vint-i-un. En canvi, a partir d'octubre del 1991 i fins desembre del 1996, es disposa de la major part dels IPIs sectorials a un nivell de desagregació sectorial de tres/quatre dígits de la CNAE-74[26]. De tota manera, per mantenir l'homogeneïtat respecte al primer subperíode, en aquest segon tampoc s'han considerat els sectors de la divisió 1 de la CNAE-74 (que es corresponen amb els sectors 1 a 8 de l'EI). Així doncs, el nombre de sectors considerats ha estat, en aquest segon subperíode, de setanta-vuit[27].

D'acord amb l'anterior i donada la desagregació sectorial de la font a partir de la qual s'han obtingut les ponderacions, pel primer subperíode ha estat necessari agrupar els vuitanta-nou sectors de l'EI en els vint-i-un disponibles i, pel segon, s'ha hagut d'agrupar alguns dels índexs nacionals en funció del seu pes (en termes de VAB a cost dels factors) respecte al conjunt de l'economia. En concret, s'han estimat uns IPIs per a la indústria espanyola amb el mateix nivell de desagregació sectorial que el de l'EI. Per això s'han estimat les ponderacions emprades per l'INE (i no publicades a aquest nivell de detall) per a l'àmbit nacional mitjançant regressions entre els IPIs corresponents al nivell d'agregació superior amb els de l'immediatament inferior.

Pel que fa a la informació regional, a Andalusia, la informació disponible sobre l'IPI base 1994 elaborat per l'IEA comença al gener del 1984, mentre que a Astúries, l'índex elaborat pel Sadei comença al gener del 1990, prenent com a any base el 1989. En canvi, per al País Basc, la informació sobre l'IPI elaborat per l'Eustat comença al gener del 1986 sent l'any base 1990. Així doncs, en el cas d'Andalusia i d'Astúries ha estat necessari estimar prèviament un índex que pogués ésser comparable amb els indicadors elaborats, que tenen com any base 1990.

[25] Així doncs, els indicadors elaborats en aquest apartat són uns IPPI i, per tant, aporten informació sobre la producció de productes industrials.

[26] Els sectors sobre els quals no es disposa d'informació són el 425 (indústria vinícola), el 454 (confecció a mida de roba i complements del vestit) i el 495 (indústries manufactureres diverses) de la CNAE-74 (que es corresponen amb els sectors 60, 73 i 89 de l'EI).

[27] Per a un detall sobre les equivalències entre els sectors de l'EI i els de la CNAE-74, vegi's Clar *et al.* (1998).

D'altra banda, en cap dels tres casos, aquests índexs no són directament comparables amb els elaborats en aquest treball, atès que inclouen informació sobre els sectors energètics. Així doncs, ha estat necessari estimar prèviament un IPPI per cadascuna de les tres regions considerades. Les ponderacions que s'han emprat (quadre 4.1) s'han obtingut a partir de la informació publicada per les entitats regionals elaboradores referent als índexs sectorials de cada regió.

Quadre 4.1. Ponderacions IPI (publicades) i IPPI

	Andalusia		Astúries		País Basc	
Divisió	IPI	IPPI	IPI	IPPI	IPI	IPPI
1	12,94%	–	38,15%	–	11,24%	–
2	11,81%	13,56%	31,49%	50,93%	21,33%	24,03%
3	36,04%	41,39%	15,21%	24,58%	43,10%	48,56%
4	39,21%	45,05%	15,155	24,49%	24,33%	27,41%
Total	100%	100%	100%	100%	100%	100%

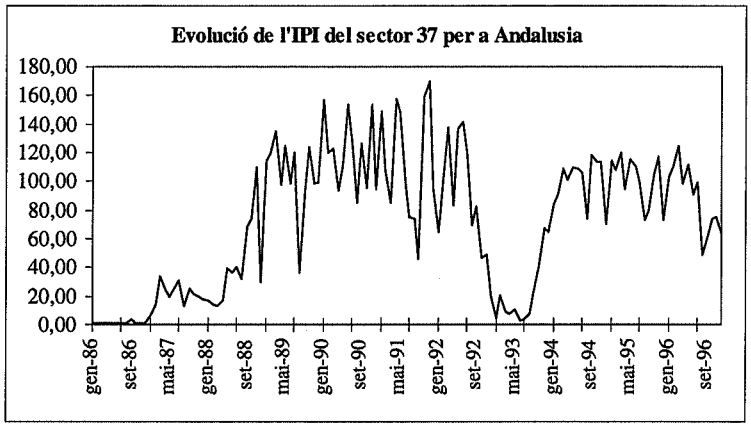
Font: IEA, Sadei, Eustat i elaboració pròpia.

En el cas d'Andalusia, s'han realitzat, a més a més, certes modificacions addicionals. Si bé sembla lògic emprar com a indicador de referència per procedir a la validació de la metodologia analitzada l'IPPI d'Andalusia, la producció dels sectors corresponents a les agrupacions 37 (construcció naval) i 42 (que engloba indústries alimentàries diverses) de la CNAE-74 mostren uns comportaments atípics.

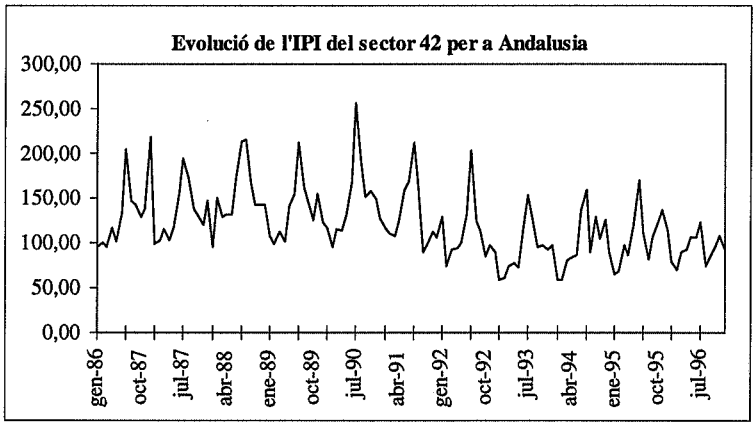
Pel que fa a l'agrupació 37 s'observen (vegi's gràfic 4.1) importants oscil·lacions al llarg del temps i donat que el seu pes és prou elevat[28] augmenta la variabilitat de l'IPPI reduint així la seva capacitat com a indicador de conjuntura[29].

D'altra banda, en el cas de l'agrupació 42 es produeix un trencament en el comportament de la sèrie de l'índex (directe) elaborat per l'IEA a partir del 1993 (vegi's gràfic 4.2) que no fa sinó empitjorar els resultats en considerar-la (cal tenir en compte addicionalment que el pes d'aquesta agrupació és del 9,43%). Així doncs, s'ha procedit a estimar un IPPI per Andalusia que no incorporés la producció d'aquestes dues agrupacions a partir de les dades dels indicadors a dos dígits facilitades per l'IEA, repartint el pes d'ambdós sectors dins de les respectives divisions.

[28]En concret, el pes de l'agrupació 37 és del 22,35%.
[29]Vegi's Morales *et al.* (1997) i Predyco (1994).



Gràfic 4.1.



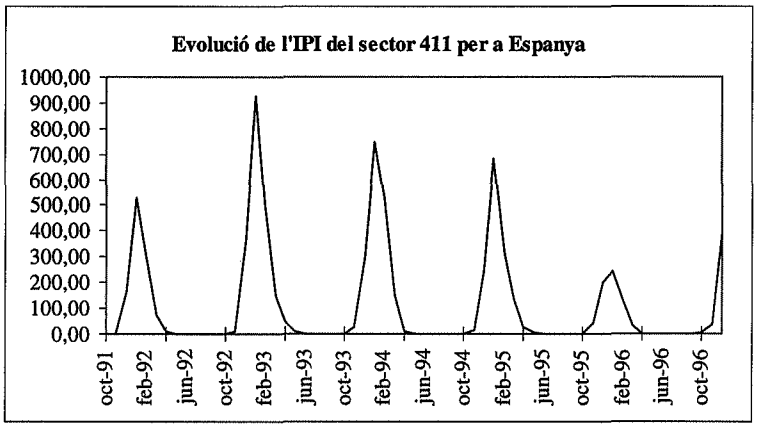
Gràfic 4.2.

4.2. Obtenció i comparació dels indicadors pel País Basc, Andalusia i Astúries

Per elaborar els indicadors regionals per a les regions considerades seguint la metodologia de l'INE, únicament cal ponderar (fent servir un o altre nivell d'agregació depenent del període) els índexs nacionals d'acord amb la seva importància relativa a l'estructura productiva de cadascuna de les regions a partir de l'expressió (2.2.bis)[30]. Cal assenya-

[30]Les ponderacions utilitzades, determinades a partir del pes de la producció bruta de cada sector en la

lar, però, que l'índex nacional corresponent al sector 411 de la CNAE-74 (fabricació d'oli d'oliva) no s'ha considerat en l'elaboració de l'indicador per a Andalusia donat que des d'octubre del 1991 (període a partir del qual es disposa d'informació estadística per a aquest sector) presenta una estacionalitat molt important així com una certa erraticitat en el seu comportament (vegi's gràfic 4.3). A més a més, cal tenir en compte que aquest sector va experimentar una profunda crisi a Andalusia en aquest període per la qual cosa l'índex nacional (en aquest període) reflecteix la producció realitzada a altres regions espanyoles. Així doncs, la consideració de l'índex nacional d'aquest sector en l'elaboració de l'indicador d'Andalusia únicament introduiria un biaix respecte a l'índex directe de l'IEA[31].



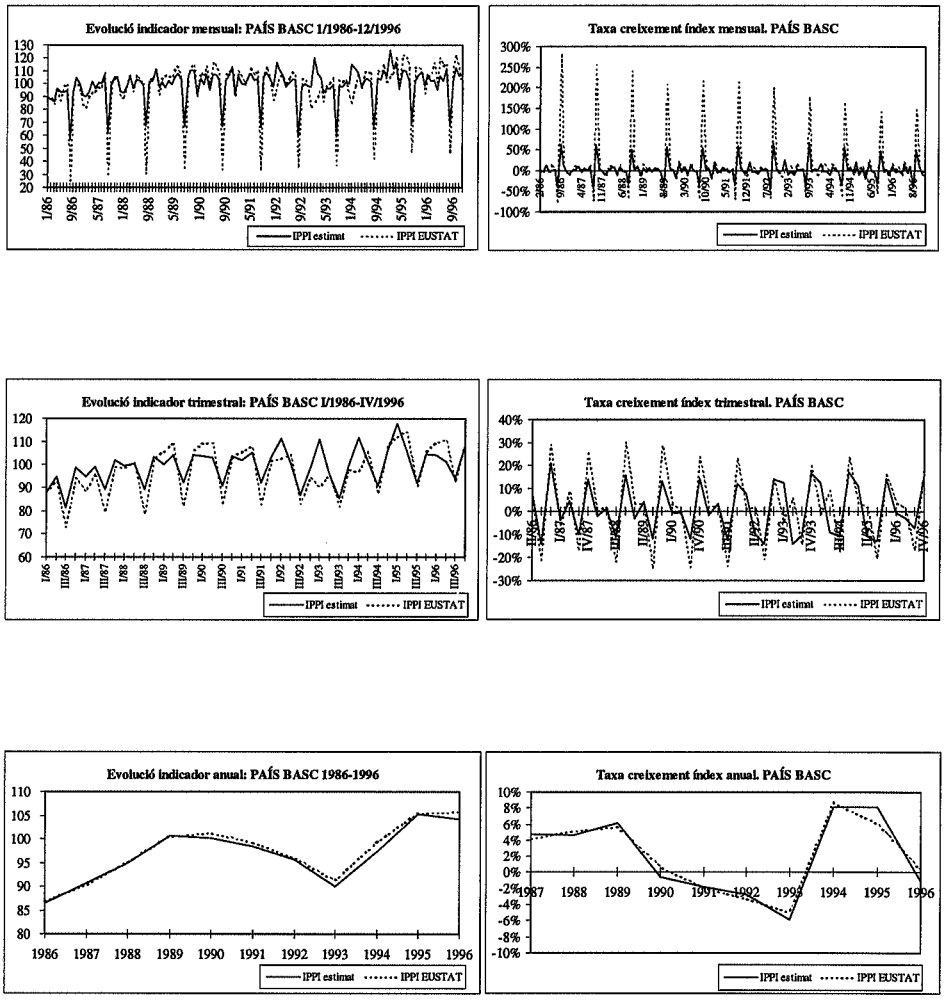
Gràfic 4.3.

Els gràfics 4.4 a 4.6 mostren el comportament dels indicadors directes i indirectes a nivell mensual, trimestral i anual. Els resultats són prou satisfactoris, especialment des de l'octubre del 1991, atès que es consideren un major nombre de sectors en la informació de base[32]. Tot i així, els resultats no són del tot satisfactoris a nivell mensual donat que l'estacionalitat dels índexs no és del tot recollida pels indicadors elaborats. Per tant, tot i que aquests indicadors ofereixen una bona aproximació a l'evolució a curt termini de la producció industrial, recullen en major mesura la seva tendència.

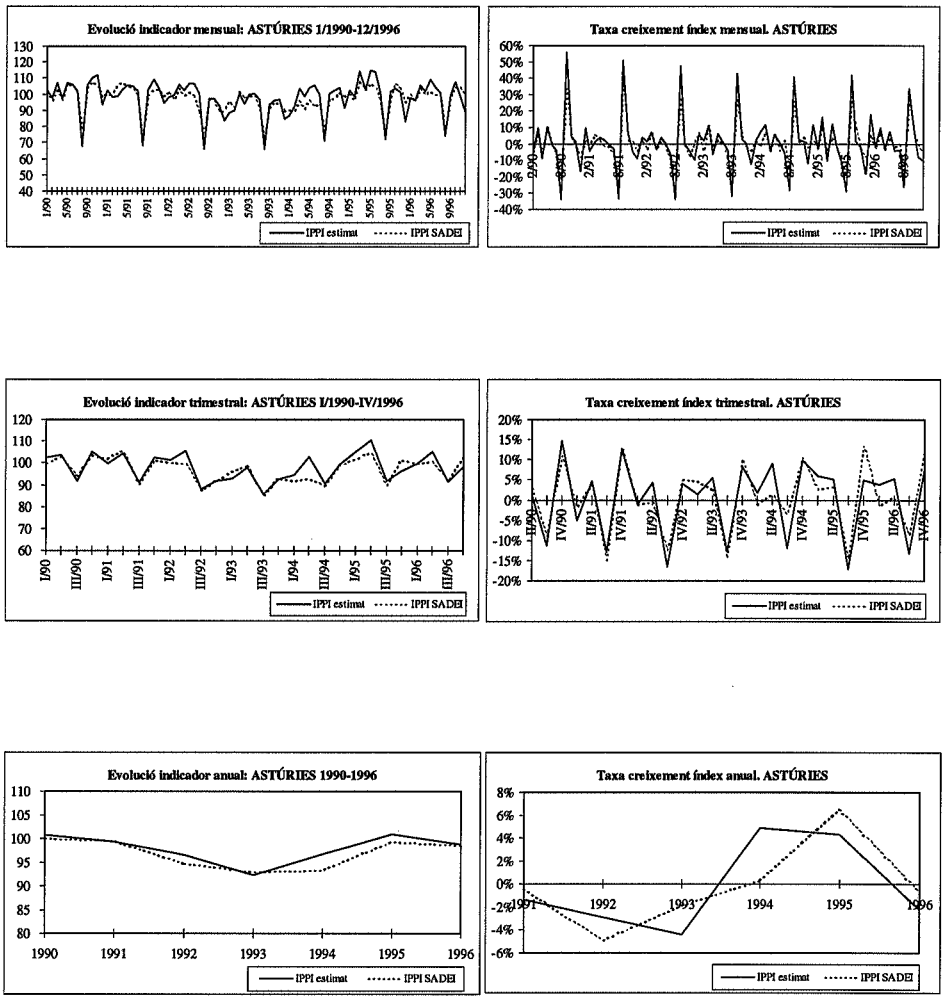
producció bruta total obtinguda de l'EI del 1990, es poden trobar a Clar *et al.* (1998).

[31]De fet, una de les limitacions de la metodologia analitzada és que els indicadors obtinguts seran tant més bons quant major sigui la informació *a priori* disponible, la qual cosa *ex-post* no suposa cap problema però, lògicament, si ho és *ex-ante*.

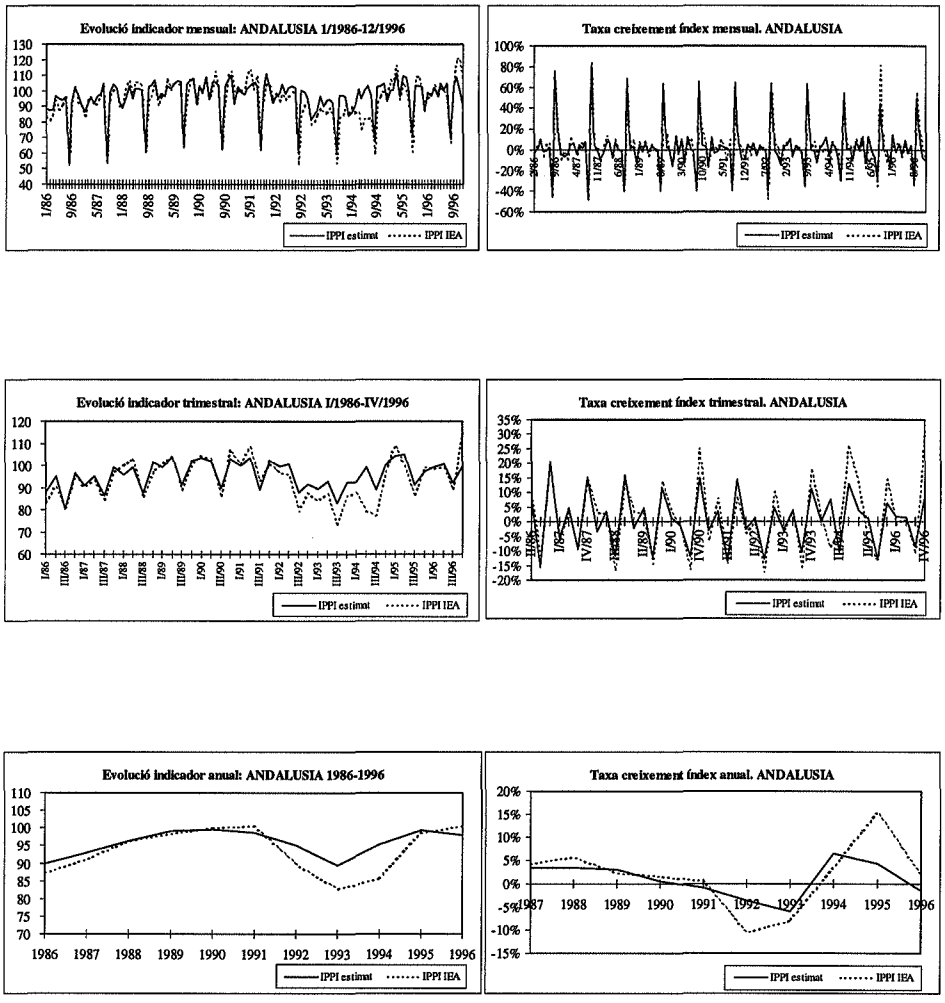
[32]De tota manera, en tots tres casos s'observa que a l'any 1993 s'obté un pitjor ajust. Això és degut a que l'economia espanyola va viure en aquest any el final d'una etapa de recessió i l'inici d'una recuperació. Aquest fet fa que l'índex nacional no reculli prou bé el comportament de l'activitat industrial regional.



Gràfic 4.4. Comparació entre l'IPPI elaborat seguint la metodologia de l'Idescat/INE i el de l'Eustat



Gràfic 4.5. Comparació entre l'IPPI elaborat seguint la metodologia de l'Idescat/INE i el del Sadei



Gràfic 4.6. Comparació entre l'IPPI elaborat seguint la metodologia de l'Idescat/INE i el de l'IEA

Com element addicional de validació s'ha calculat l'EPAM entre els índexs construïts a partir de les dades de l'IEA, del Sadei i de l'Eustat i els indicadors indirectes elaborats. La comparació s'ha efectuat a nivell mensual, trimestral i anual. Els resultats obtinguts (vegi's quadre 4.2) pels indicadors trimestrals i anuals són inferiors al 3%, fet que reflecteix el bon comportament dels indicadors obtinguts tret del cas d'Andalusia[33]. Tot i així, en termes mensuals l'indicador no presenta un comportament tan acceptable en cap dels tres casos.

Addicionalment s'ha calculat l'EPAM a partir de la component tendència-cicle de les sèries. Els resultats obtinguts milloren respecte als anteriors confirmant el fet que els indicadors sintètics elaborats seguint la metodologia de l'INE recullen en major mesura el comportament tendencial mensual de l'activitat industrial, però no les erraticitats de les sèries (normalment més associades a factors regionals propis que són recollits per dita metodologia).

Quadre 4.2. EPAMs entre els IPPIs elaborats a partir dels índexs publicats per l'IEA, el Sadei i l'Eustat i els elaborats seguint la metodologia de l'INE

ANDALUSIA			
Indicador	**Període**	**EPAM**	**EPAM***
Mensual	01/1986 - 12/1996	5,67%	3,55%
Trimestral	I/1986 - IV/1996	4,63%	3,52%
Anual	1986 - 1996	3,36%	–
ASTÚRIES			
Indicador	**Període**	**EPAM**	**EPAM***
Mensual	01/1990 - 12/1996	4,32%	1,72%
Trimestral	I/1990 - IV/1996	2,33%	1,72%
Anual	1990 - 1996	1,29%	–
PAÍS BASC			
Indicador	**Període**	**EPAM**	**EPAM***
Mensual	01/1986 - 12/1996	6,83%	1,18%
Trimestral	I/1986 - IV/1996	2,54%	1,10%
Anual	1986 - 1996	0,67%	–

* Calculat a partir de la component tendència-cicle de les sèries obtinguda mitjançant el (nou) filtre de Línies Aèries Modificat (LAM) de l'INE.

[33]Recordi's, però, que com s'ha pogut observar a l'apartat anterior, pel període comprés entre octubre del 1992 i desembre del 1993 l'índex directe elaborat per l'IEA presenta un comportament (molt) atípic, la qual cosa fa que els resultats estiguin esbiaixats a l'alça.

Per últim, per garantir la validesa de les conclusions que es deriven de la comparació entre els indicadors indirectes obtinguts en aquest apartat (adaptant la metodologia de l'Idescat) amb els directes pel cas dels publicats per l'INE i, alhora confirmar que la metodologia emprada a aquest apartat per elaborar els IPPIs és molt semblant a la que empra l'INE per elaborar els IPIs regionals, s'han calculat els coeficients de correlació entre les taxes de creixement de les sèries dels IPIs de l'INE i dels IPPIs elaborats en termes mensuals, trimestrals i anuals per les tres comunitats considerades (vegi's quadre 4.3).

Quadre 4.3. Coeficients de correlació entre les taxes de creixement de les sèries d'IPIs de l'INE i les d'IPPIs elaborades. Període 11/1991-12/1996

	Mensual	Trimestral	Anual
País Basc	0,95%	0,71%	0,84%
Astúries	0,98%	0,97%	0,98%
Andalusia	0,96%	0,78%	0,97%

Els resultats obtinguts mostren que la metodologia seguida en aquest apartat per elaborar els IPPIs regionals és consistent amb la que empra l'INE en l'elaboració dels IPIs regionals. Així doncs, l'elaboració de les sèries d'IPPIs permet prendre a aquestes com a sèries de referència més llargues que les de l'INE, la qual cosa permet avançar en la identificació dels factors determinants de l'adequació de la metodologia de l'INE.

5. FACTORS DETERMINANTS DE LA FIABILITAT DELS INDICADORS REGIONALS ELABORATS SEGUINT LA METODOLOGIA DE L'INE

Els resultats obtinguts als dos apartats anteriors permeten afirmar que la fiabilitat dels indicadors elaborats seguint la metodologia emprada per l'INE per una regió depèn de cinc factors:

a) *Del grau de concentració geogràfica de la producció industrial.* D'acord amb l'anàlisi realitzada al segon apartat, només si tota la producció de cada sector considerat es produeix en una única regió, la metodologia analitzada és completament fiable, atès que en aquest cas l'indicador sectorial nacional és igual al regional. Tot i així, pot obtenir-se una bona aproximació a l'evolució de la producció industrial regional si el grau de concentració geogràfica de la producció és elevat. Així doncs, un test inicial per a validar aquesta metodologia consisteix en calcular els coeficients de concentració geogràfica de Gini a partir de la producció bruta de cada sector a l'any base. En calcular l'esmentat coeficient pel nivell de desagregació de dos dígits

de la CNAE-74, dotze dels vint-i-un sectors considerats (el 57,14%) tenen un coeficient de Gini superior a 0,7. D'altra banda, pel nivell de desagregació sectorial utilitzat a l'EI, el grau de concentració geogràfica és, com era d'esperar, superior: cinquanta-un dels setanta-vuit sectors considerats (això és, el 65,39%) presenten un coeficient de Gini superior o igual a 0,7. De tota manera, però, aquests resultats no semblen prou alts com per poder afirmar que el nivell de concentració geogràfica de la producció és elevat i suficient per a aplicar la metodologia analitzada, introduint, en conseqüència, errors en els indicadors proposats que poden ésser relativament importants.

b) *Del nivell de desagregació de la informació de base.* Sembla clar a partir dels resultats obtinguts pels dos nivells de desagregació considerats a l'apartat anterior, que quant més gran sigui el nombre d'indicadors sectorials utilitzats, millor s'ajustarà el comportament de l'indicador elaborat amb la metodologia analitzada a la producció industrial. De fet, aquest segon factor està (molt) relacionat amb l'anterior, donat que quant més gran sigui el nivell de desagregació considerat major és el grau de concentració. A més a més, en treballar amb un major nivell de desagregació a l'hora de censurar la informació de base es perd menys informació.

c) *Del pes de la producció industrial de la regió sobre la del conjunt de l'Estat.* La metodologia analitzada proporciona millors resultats per a aquelles regions on la producció industrial té un pes important en la producció del total nacional. Aquesta és una de les raons per la qual la metodologia analitzada funciona millor per regions com Catalunya i el País Basc que per a altres comunitats (vegi's quadre 5.1).

Quadre 5.1. Ordenació de les CA en funció del pes de la producció industrial del conjunt de l'Estat

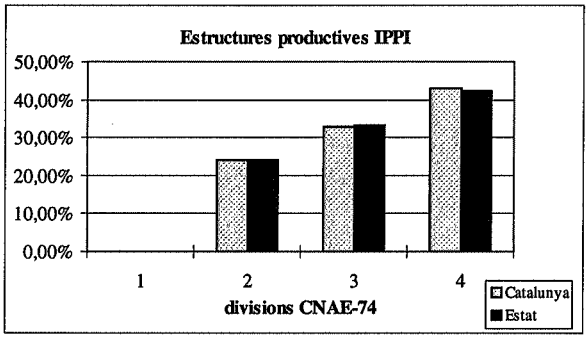
CA	Pes relatiu de la producció regional sobre el total espanyol (posició d'ordre respecte a les disset CA)*
Catalunya	26,30% (1)
País Basc	9,63% (4)
Andalusia	9,21% (5)
Astúries	2,39% (11)

* En termes de la producció bruta industrial del 1990.

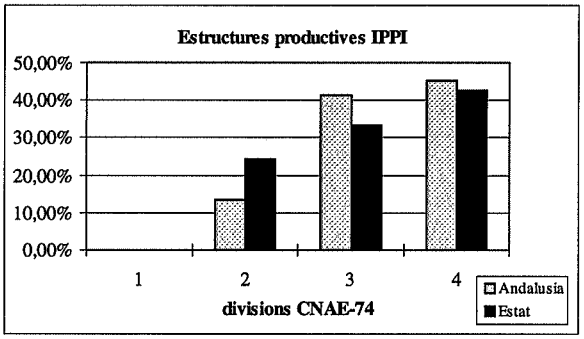
d) *De la semblança de l'estructura productiva regional amb la nacional.* Quant més similar sigui l'estructura productiva de la regió i del conjunt de l'Estat, més representativa serà la mostra emprada en l'enquesta nacional de l'estructura productiva regional i, donat que aquesta metodologia empra els índexs nacionals com informació de base per a obtenir els indicadors regionals, els resultats seran millors.

Com pot veure's al gràfic 5.1 les estructures productives de l'economia catalana i espanyola són pràcticament coincidents, la qual cosa garanteix que la utilització de la informació emprada per l'INE per a elaborar l'índex per la indústria del conjunt de l'Estat és representativa per a Catalunya.

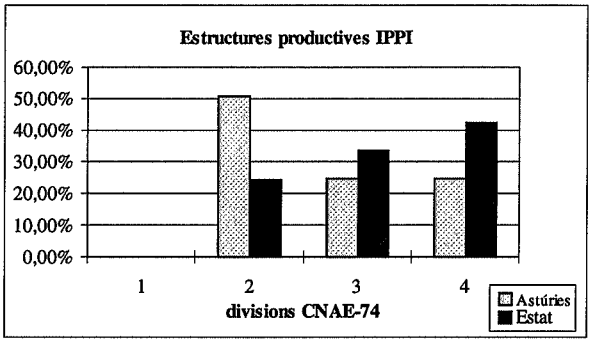
En canvi, per a les comunitats d'Andalusia, Astúries i el País Basc hi ha més diferències (vegi's els gràfics 5.2 a 5.4). En concret, el País Basc és una regió on el pes de la divisió 3 de la CNAE-74 (Indústries transformadores dels metalls. Mecànica de precisió) és molt més gran que en el conjunt de l'Estat, mentre que el de la divisió 4 (Altres indústries manufactureres) és molt més petit. Per la seva banda, a Astúries les principals diferències es centren en les divisions 2 (Extracció i transformació de minerals no energètics i productes derivats. Indústria química) i 4: la divisió 2 té molta més importància a Astúries que a la resta de l'Estat i, pel contrari, la divisió 4 a Astúries té un pes molt menor. Pel que fa a Andalusia, les principals diferències es centren en les divisions 2 i 3: la divisió 2 té menys pes a Andalusia que en el conjunt de l'Estat i la 3 més.



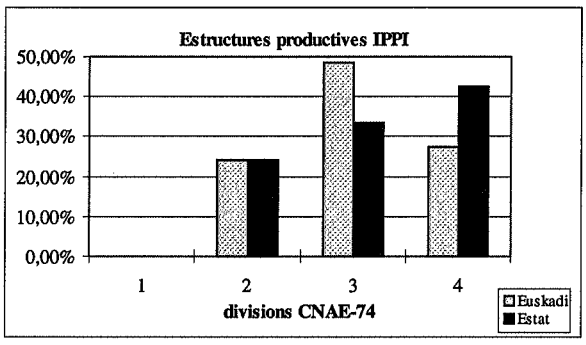
Gràfic 5.1.



Gràfic 5.2.



Gràfic 5.3.



Gràfic 5.4.

e) De la disponibilitat d'informació a priori. Tal i com s'ha posat de manifest anteriorment en el cas del sector de fabricació d'oli d'oliva (sector 411 de la CNAE-74) pel cas d'Andalusia, els indicadors obtinguts seran tant més bons quant major sigui la informació *a priori* disponible atès que el procés de censura de la informació de base pot dur-se a terme de forma més eficient. En qualsevol cas, però, aquest problema sempre es presenta *ex-ante* la qual cosa suposa un element d'incertesa addicional a l'hora de predir el comportament de l'indicador sintètic.

6. CONCLUSIONS

La pràctica inexistència d'indicadors (quantitatius) de la producció industrial en l'àmbit regional al nostre país va fer que, recentment, l'INE elaborés uns indicadors per a les

CA espanyoles a partir d'una metodologia homogènia. No hi ha cap dubte que la millor opció seria elaborar uns indicadors a partir del mètode directe, però els elevats costos associats a aquest mètode, juntament amb les restriccions pressupostàries, no la fan viable. Per aquest motiu, la millor opció a la pràctica consisteix en emprar informació preexistent. Així, l'INE ha adoptat el mètode que empra l'Idescat per elaborar l'indicador de la comunitat catalana atès el bon funcionament d'aquesta metodologia per dita comunitat.

A aquest treball s'ha estudiat la idoneïtat d'estendre l'esmentat mètode indirecte per a construir indicadors de la producció industrial per a totes les regions espanyoles. Els resultats obtinguts mostren que els indicadors elaborats d'acord amb aquesta metodologia ofereixen una bona aproximació a nivell trimestral i anual pel conjunt de la regió. Tot i això, però, els indicadors que s'obtenen no recullen de manera totalment correcta l'evolució de la producció industrial donat que la fiabilitat de la metodologia emprada depèn de tot un seguit d'hipòtesis i supòsits que no es compleixen a moltes de les regions espanyoles.

D'acord amb tot l'anterior, cal dir que la metodologia analitzada està plenament justificada per a determinades regions (com ara Catalunya i el País Basc), però la fiabilitat dels indicadors sintètics obtinguts per a altres regions no pot garantir-se a nivell mensual. En conseqüència, proposem un sistema mixt de càlcul dels IPIs regionals, en el que la informació dels IPIs nacionals al màxim nivell de desagregació sectorial es complementi amb altre tipus d'informació pròpia de la regió que permeti obtenir un bon indicador conjuntural de l'activitat industrial.

BIBLIOGRAFIA

Artís, M., J. Pons, M.A. Sierra i J. Suriñach (1994). «Elaboració d'un Sistema d'Indicadors d'Activitat per a l'Economia Catalana», *Perspectiva Econòmica de Catalunya*, 176, 83-102.

Artís, M., J. Pons, M.A. Sierra i J. Suriñach (1997a). «Estimación de la Actividad Económica a Corto Plazo Mediante Indicadores de Coyuntura», *Revista de Economía Aplicada*, 13, 129-147.

Artís, M., E. Pons, J. Pons i J. Suriñach (1997b). *Evolución Cíclica de las Comunidades Autónomas y Análisis Cíclico*, Escuela de Economía Regional, Universidad Internacional Menéndez Pelayo, Santander.

Artís, M., E. Pons, J. Pons i J. Suriñach (1997c). «Comptabilitat Econòmica de Catalunya i Mètodes de Trimestralització. Components de la Demanda», *Document de Treball 97R02*, Grup d'Anàlisi Quantitativa Regional, Universitat de Barcelona.

Clar, M. (1998). *Una Anàlisi Metodològica pel Seguiment Conjuntural de l'Activitat Industrial de les Regions Espanyoles*, Tesi Doctoral, Universitat de Barcelona.

Clar, M., R. Ramos i J. Suriñach (1998). «Algunes Reflexions sobre la Construcció d'Indicadors Indirectes pel Seguiment de l'Activitat Industrial Regional», *Document de Treball E98/40*, Divisió de Ciències Jurídiques i Socials, Universitat de Barcelona.

Comisión de la CE (1991). «Business and Consumer Survey», *European Economy*, Supplement B, Edició especial, Direcció General d'Assumptes Econòmics i Socials.

Cordero, G., A. Gayoso, A. Pavón i E. Rodríguez (1996). «Los Indicadores de Clima Industrial Regionales como Instrumento para el Análisis Espacial del Ciclo en la Industria: Metodología y Resultados», *Document de Treball SGPR-96002*, Direcció General de Planificació, Secretaria d'Estat d'Hisenda, Ministeri d'Economia i Hisenda.

Costa, A. i J. Galter (1994). «L'IPPI, un Indicador molt Valuós per Mesurar l'Activitat Industrial Catalana», *Revista d'Indústria*, 3, 2ona etapa, 3er trimestre 1994, 6-15, Generalitat de Catalunya, Departament d'Indústria i Energia.

Eurostat (1978). *L'Indice de la Production Industrielle de la Communaute Europeenne*, Suplement Metodològic 1/78.

González, M., P. Revilla i P. Rey (1992). «Los Nuevos Índices de Producción y Precios Industriales», *Situación*, 3-4, 109-117, BBV.

Instituto de Estadística de Andalucía (1997). *Índice de Producción Industrial. Metodología del Cambio de Base y Presentación de Resultados.*

Institut d'Estadística de Catalunya: *L'Índex de Producció de Productes Industrials*, informe mensual.

Instituto Nacional de Estadística (1982). *Números Índice de la Producción Industrial. Base, 100 en 1972*, Monografia Tècnica, Ministeri d'Economia i Comerç, Madrid.

Instituto Nacional de Estadística (1993). *Contabilidad Nacional Trimestral de España, Metodología y Series*, Madrid.

Instituto Nacional de Estadística (1994). *Encuesta Industrial 1988-1991*, Madrid.

Instituto Nacional de Estadística: *Boletín Mensual de Estadística*, varis números.

Instituto Vasco de Estadística: *Índice de Producción Industrial*, varis números.

Junta de Andalucía (1988). *Memoria Técnica y Metodología del Índice de Producción Industrial de Andalucía*, Conselleria de Foment i Treball.

Kmietowicz, Z.W. (1995). «Accuracy of Indices of Industrial Production in Developing Countries», *The Statistician*, 44, 3, 295-307.

Morales, E., R. Mínguez i L. Dávila (1997). *Comparación de Métodos de Estimación de Tendencias y Análisis de Coyuntura. Aplicación al Caso del Índice de Producción Industrial de Andalucía*, Comunicació presentada a la XXIII Reunión de Estudios Regionales, València, 18-21 de novembre.

Muñoz, J., E. Pons i J. Pons (1996). «Les Revisions de les Estimacions de la Comptabilitat Nacional», *Questiió*, 20, 1, 293-324.

Prado, C. (1988). «Elaboración de un Índice de Producción y Precios Industriales», *Ekonomiaz*, 11, 297-313.

Predyco (1994). *Realización de un Indicador Sintético para Estimar el Crecimiento del Producto Interior Bruto no Agrario de Andalucía*, mimeo.

Revilla, P. (1997). «El IPI como Principal Indicador Económico de Oferta», *Fuentes Estadísticas*, 30, 11.

Rey, P., M. González i P. Revilla (1993). «Principales Características de los Nuevos Índices de Producción y Precios Industriales», *Boletín Trimestral de Coyuntura*, 47, 60-84.

Sadei (1993). *Índice de Producción Industrial de Asturias*. Año 1991, Principat d'Astúries, Conselleria d'Hisenda, Economia i Planificació.

Smith, P. (1993). «The Timeliness of Quarterly Income and Expenditure Accounts: An International Comparison», *Australian Economic Indicators*.

Suriñach, J. i V. Royuela (1995). «L'Índex de Producció de Productes Industrials per Catalunya. Extensió de la Sèrie fins l'Any 1975», *Document de Treball 95R03*, Grup d'Anàlisi Quantitativa Regional, Universitat de Barcelona.

Suriñach, J., E. Pons i J. Pons (1996). *Comptabilitat Econòmica de Catalunya i Mètodes de Trimestralització*, Institut d'Estadística de Catalunya, Barcelona. Una versió reduïda pot trobar-se a *Revista Econòmica de Catalunya*, 30, 38-56.

ENGLISH SUMMARY

ADVANTAGES AND DISADVANTAGES OF IDESCAT/INE'S METHODOLOGY TO ELABORATE INDUSTRIAL PRODUCTION INDICATORS FOR THE SPANISH REGIONS*

MIQUEL CLAR
RAÚL RAMOS
JORDI SURIÑACH
Universitat de Barcelona*

The analysis of the conjunctural evolution of the industrial sector, both at a national and regional level, is relevant. In this sense, the delay in the publication of National/Regional Accounts data makes necessary the elaboration of indicators that permit to analyse the short-term evolution of industrial activity. To correct this deficit at the national level, the INE elaborates a monthly IPI from specific data survey. At a regional level, during the last few years, different projects have focused on the elaboration of indicators of industrial activity using non-homogeneous indirect methods. In this sense, one of the most widely accepted methodologies has been the one applied by the Idescat to elaborate the indicator for Catalonia. Thus, the INE has recently published IPIs for the Spanish regions following this methodology. In this paper, we analyse the reliability of extending this indirect methodology to all the Spanish regions comparing the INE's indirect indicators with the direct ones elaborated by other institutions in three of the four regions which have it: Andalucía, Asturias and País Vasco.

Keywords: Industrial activity, industrial production index, regional indicators, conjuncture

AMS Classification: 62P20, 90A19

⋆Financial support is gratefully acknowledged from DGICYT SEC99-0700 project and Plan Nacional de I+D 2FD97-1004-C03-01 project.

*Miquel Clar (mclar@eco.ub.es); Raúl Ramos (rrlobo@eco.ub.es); Jordi Suriñach (surinach@eco.ub.es). Grup de recerca *Anàlisi Quantitativa Regional*. Universitat de Barcelona. Av. Diagonal, 690. 08034 Barcelona.

–Received June 1999.
–Accepted December 1999.

1. INTRODUCTION

The most commonly used measure to analyse the evolution of the manufacturing sector is the Gross Added Value (GAV) or the Gross Domestic Product (GDP), in strict sense, this is, without including data on the construction sector. However, in Spain, as well as in other countries, the main problem to use this information to analyse the short-term evolution of manufacturing is related to the fact that these data are not available as soon as it would be desirable. This fact makes very difficult to evaluate the short-term behaviour of industrial activity. It is necessary, then, to obtain indicators that permit to analyse the conjunctural evolution of industrial GDP overcoming the previously mentioned limitations.

In Spain, the National Institute of Statistics (INE) elaborates a monthly quantitative index to monitor the national industrial activity, called *Índice de Producción Industrial* (Industrial Production Index –IPI–), using data from surveys addressed to a representative sample of productive units from all sectors of activity (direct method). So, at a national level, the problem of the lack of statistical information to carry out a complete industrial quantitative conjunctural analysis is partially solved.

However, at a regional level (until very recently) there were big difficulties to analyse the short-term industrial activity evolution as there were great deficiencies regarding the availability of statistical information of these characteristics. In front of this situation, during the last years, in some Spanish regions several public and private initiatives were initiated to overcome these deficiencies. Although an important effort was carried out, the real situation was that not every Spanish region had a quantitative indicator of the industrial activity evolution and, moreover, the available regional indicators were not directly comparable as non-homogenous methodologies were used to elaborate them. In relation to this topic, in different forums a debate was initiated about which was the most appropriate methodology to elaborate regional industrial production indicators with a high level of reliability and, at the same time, a low cost. The result was that, at last, the INE recently published regional industrial production indicators following an indirect method, which is very similar to the Idescat methodology for the regional indicator for *Cataluña*. In particular, the published series begin in October 1991 and refer only to the general index, and no information is provided for the different activity branches or for the economic destination of the goods. In this sense, some of the existing deficiencies have been partially overcome.

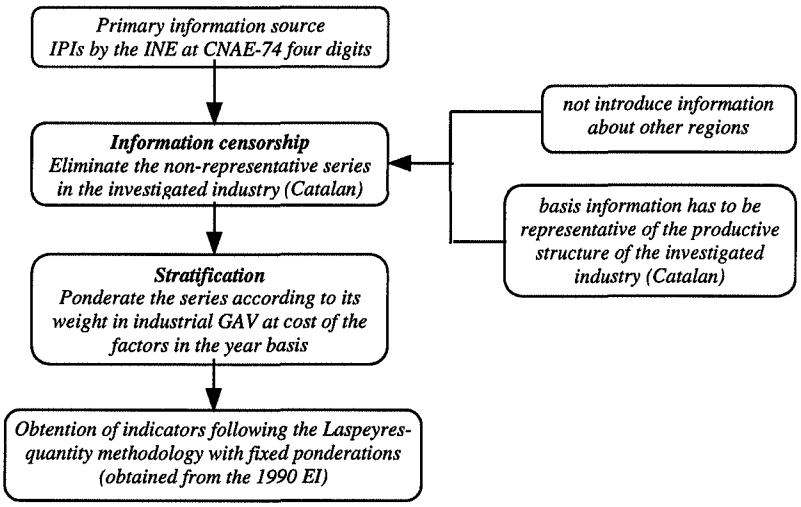
In front of this situation, the objective of this paper is to analyse the reliability of the regional indicators obtained with the methodology used by the INE. The structure of the paper is as follows: first, this methodology is presented; second, a comparative analysis between the indexes published by the INE for *Andalucía, Asturias* and *País Vasco* and the IPIs elaborated by the IEA, the Sadei and the Eustat using direct methods is done; next, the Idescat methodology is applied for these three regions to obtain longer series

of Industrial Products Production Indexes (IPPI) which are compared with the regional direct indexes; and, last, the main conclusions are presented.

2. THE IDESCAT/INE'S METHODOLOGY TO ELABORATE REGIONAL INDUSTRIAL ACTIVITY INDICATORS

The indicator elaborated by the Idescat/INE is an indirect quantitative indicator, so basic information comes from pre-existent available information. This methodology is summarised in figure 1.

The main result of the analysis in this section is that the considered methodology provides a good approximation to the evolution of the regional manufacturing production but it will provide a «perfect» indicator when the whole production of each industrial branch is done in one region (when the geographical concentration level is 100%). The failure to fulfil this condition implies to introduce information from other regions in the elaboration of the indicator of the considered region. This problem loses relevance when more disaggregated sectorial national information is used because the level of geographical concentration of the different regions, increases when a higher number of branches is considered.



Source: Own elaboration from Costa and Galter (1994).

Figure 1

3. COMPARISON BETWEEN DIRECT REGIONAL INDICATORS AND THE ONES ELABORATED BY THE INE

In this section a comparative analysis is carried out among the indirect indicators elaborated by the INE and the direct indexes elaborates respectively by the Eustat, the Sadei and the IEA for *País Vasco, Asturias* and *Andalucía* for the period within October 1991 and December 1996. The obtained results show that the adjustment for *País Vasco* and *Andalucía* direct indexes' is satisfactory for the whole considered period, but for *Asturias* more disagreements can be observed. These results lead to the conclusion that the methodology used by the INE to obtain regional indicators is not completely reliable for all regions at a monthly frequency.

4. SENSITIVENESS OF THE INE METHODOLOGY TO THE AVAILABILITY AND CENSORSHIP OF THE BASIS INFORMATION AND TO THE CONSIDERED PERIOD

As the INE has not published any methodological note referring to the process followed to elaborate regional production indicators, there are some questions that cannot be answered, as for example: Which are the weights employed in each region? Is a particular censorship process carried out for each region? In this case, which are the national IPI's series that are not included in the regional indicators? For this reason, In this section, we have estimated indicators for the three considered regions following the Idescat (INE's) methodology with some little variations due to data availability in order to obtain evidence on the determinants of the methodology adequacy for all the Spanish regions and to identify the determining factors of its reliability.

5. DETERMINING FACTORS IN REGIONAL INDICATORS' RELIABILITY ELABORATED FOLLOWING INE'S METHODOLOGY

The results obtained at the two previous sections allow to affirm that the reliability of the obtained indicators for a particular region depends on five factors: on the geographical concentration degree of industrial production; on the aggregation level of basis information; on the regional industrial production share in the total national; on the similarity of regional productive structure with the national one; and, on the availability of *a priori* information.

6. CONCLUSIONS

The main conclusion is that the considered methodology is fully justified for some regions (such as *Cataluña* and *País Vasco*), but the reliability for other regions cannot be guaranteed at monthly frequency. As a consequence, we propose to obtain regional indicators of industrial activity using information about national IPIs at the maximum level of detail but complementing it with region-specific information, which will permit to obtain a good conjunctural indicator of industrial activity.

Biometria

QÜESTIIÓ, vol. 24, 1, p. 189-204, 2000

COMPARACIÓN DE DOS TABLAS DEMOGRÁFICAS: APROXIMACIÓN A SU SIGNIFICACIÓN ESTADÍSTICA

ERNESTO J. VERES FERRER
Universidad de Valencia*

En este trabajo se analiza la significación estadística de la posible igualdad de dos tablas demográficas. Concretamente, se presenta la aplicabilidad de un clásico contraste de los métodos estadísticos —el de homogeneidad de dos distribuciones— para contrastar la hipótesis nula «H_0 = las dos tablas demográficas son iguales, esto es, responden a una misma estructura del fenómeno demográfico estudiado», frente a la alternativa que niega la anterior. Ambas tablas se refieren a un único fenómeno demográfico para el mismo ámbito territorial y dos momentos diferentes de tiempo (contraste temporal), o para la misma referencia temporal en dos poblaciones de ámbito territorial distinto (contraste territorial). Se aplica la metodología descrita en dos situaciones: para la comparación de los niveles de mortalidad de dos provincias, y sobre dos tablas-tipo correspondientes a sendos niveles de mortalidad.

Comparison of two demographic tables: approximation to their statistical significance

Palabras clave: Calendario, contraste de hipótesis, intensidad, tabla demográfica, tablas-tipo de mortalidad, test de la χ^2

Clasificación AMS: 62F03, 62G10, 62P05, 62P25, 92 H20

*Departamento de Economía Aplicada. Facultad de Ciencias Económicas y Empresariales. Universidad de Valencia. Campus de los Naranjos. 46022 Valencia. E-mail: Ernesto.Veres@uv.es.

–Recibido en marzo de 1999.
–Aceptado en noviembre de 1999.

1. INTRODUCCIÓN

El modelo conocido como *tabla demográfica* permite analizar un fenómeno demográfico F a través de cierto suceso característico A, que supondremos irrepetible. La información para el análisis es proporcionada por la observación de la incidencia del suceso A sobre una cohorte, estudiándose la frecuencia con la que ese suceso característico va apareciendo desde una edad inicial —que denotamos por x_0—, hasta una edad final en la que el suceso deja de hacer su aparición —denotada por x_ω—.

En nuestro desarrollo, supondremos —salvo indicación al contrario— que la tabla demográfica es *completa*, esto es, que las edades o duraciones consideradas están tomadas unidad a unidad (por edades o duraciones simples), y que éstas tienen la consideración de *duraciones o edades exactas*. También la supondremos definida a partir de las tres series biométricas fundamentales siguientes:

- serie de individuos no alcanzados por A antes de la edad x, denotada por $\{L_x\}_{x=x_0}^{x_\omega}$ (serie de supervivientes);
- serie del flujo relativo del suceso A entre dos edades consecutivas x y $x+1$, denotada por $\{D(x,x+1)\}_{x=x_0}^{x_\omega-1}$ (serie de flujo de sucesos); y,
- serie de probabilidades de que un individuo, en el momento de llegar a la edad x, sea alcanzado por el suceso A antes de llegar a $x+1$, denotada por $\{q_x\}_{x=x_0}^{x_\omega-1}$ (serie de probabilidades de ocurrencia del suceso A).

Suponiendo una cohorte ficticia con L_0 efectivos iniciales (generalmente, 10^4 ó 10^5 individuos), las relaciones entre las tres series anteriores son las siguientes:

$$D(x,x+1) = L_x - L_{x+1}$$

$$q_x = \frac{D(x,x+1)}{L_x}$$

por lo que, conocida una cualquiera de las series, son conocidas las otras dos.

La *intensidad* y el *calendario* son dos índices analíticos básicos que se deducen fácilmente de una tabla demográfica. La intensidad I —expresada en términos absolutos— representa el número de individuos que acaban por ser alcanzados por el suceso A a lo largo de la vigencia del fenómeno estudiado:

$$I = L_{x_0} - L_{x_\omega} = \sum_{x=x_0}^{x_\omega-1} D(x,x+1)$$

mientras que, en términos relativos, esa intensidad puede definirse como el porcentaje de individuos alcanzados por A sobre el total de efectivos iniciales L_{x_0}.

El calendario $d(x,x+1)$ representa la distribución por edades de la intensidad anterior. Se trata de una distribución de probabilidad condicional:

$$d(x,x+1) = \frac{D(x,x+1)}{I}$$

Como síntesis analítico del calendario puede utilizarse cualquier medida estadística, cuya interpretación sea la propia de la Estadística Descriptiva. En particular, la media aritmética:

$$\overline{d} = \sum_{x=x_0}^{x_\omega - 1} \left(\left(x + \frac{1}{2} \right) \times d(x,x+1) \right) =$$

$$= \frac{1}{2} + \frac{1}{I} \left(x_0 L_{x_0} + \sum_{x=x_0+1}^{x_\omega - 1} L_x - (x_\omega - 1) L_{x_\omega} \right)$$

donde se supone la distribución uniforme del suceso A dentro de cada intervalo de edades x y $x+1$.

Con los elementos anteriores, pretendemos determinar, con significación estadística, si dos tablas demográficas —correspondientes a dos momentos diferentes, o a dos territorios distintos— son iguales. Para ello se propone el contraste de homogeneidad basado en el test de la χ^2, aplicado sobre las respectivas series $D(x,x+1)$.

Cuando las diferencias entre las dos tablas demográficas son grandes, resulta evidente a simple vista su significación. En efecto, el orden de magnitud de las diferencias entre sus respectivas series biométricas confirman la variación —en uno u otro sentido— entre los niveles alcanzados por el fenómeno F expresados por ambas tablas. Resulta irrelevante, pues, efectuar cualquier otro tipo de análisis. Sin embargo, cuando las diferencias alcanzadas en dichas series biométricas y en los indicadores clásicos deducidos de ellas no son lo suficientemente grandes como para valorar a simple vista su significación, podemos preguntarnos por la *significación estadística* de esas pequeñas diferencias. Comprendemos, pues, la plena aplicación de las técnicas y modelos estadísticos para intentar explicarla.

2. CONTRASTE DE HOMOGENEIDAD

La serie del flujo relativo del suceso A entre dos edades consecutivas x y $x+1$, $\{D(x, x+1)\}_{x=x_0}^{x_\omega - 1}$, expresa sobre una generación ficticia de L_0 elementos (la potencia de la tabla, que generalmente suele tomar los valores 10.000 ó 100.000 personas) la frecuencia de aparición del fenómeno demográfico estudiado F, para el nivel expresado a través de la serie de probabilidades $\{q_x\}_{x=x_0}^{x_\omega - 1}$ de la correspondiente tabla.

Consideremos, pues, que los flujos recogidos en sendas series de las dos tablas demográficas consideradas corresponden a las observaciones de dos universos —las derivadas de la incidencia del fenómeno F en dos momentos de tiempo, o en dos territorios, a comparación—, y que se han clasificado atendiendo al mismo criterio «edad de la persona al incidir sobre él el suceso característico A». En una tabla completa, el criterio puede llegar, por ejemplo y en el caso de la mortalidad, a 101 alternativas excluyentes (de «0 años» a «100 o más años», edad por edad simple), mientras que en una abreviada, las alternativas excluyentes pueden ser, por ejemplo, de 21 (de «0 años», de «1 a 4 años», de «5 a 9 años», de «10 a 14» años, ..., de «95 o más años», esto es, agrupando quinquenalmente la edad y distinguiendo las muertes de 0 años). Para otros fenómenos, la extensión de las clasificaciones son más reducidas: por ejemplo y para la fecundidad, 35 categorías de clasificación, si es medida con edades simples, y 7 en el caso de clasificación por grupos quinquenales de edad. Denotemos por $L_0^{t_1}$ y $L_0^{t_2}$ las respectivas potencias de ambas tablas demográficas, que corresponden a los tiempos o territorios diferentes t_1 y t_2.

La población asociada a cada una de estas tablas puede representarse a través de una variable aleatoria ξ_x, que adopta valores sobre las categorías de clasificación contenidas en aquéllas. La formulación de las hipótesis a contrastar se establecen en los siguientes términos muy conocidos:

H_0: *los niveles del fenómeno F en* t_1 *y* t_2 *son homogéneos, o, lo que es lo mismo, las distribuciones que expresan la ocurrencia por edades de F son iguales*

frente a la alternativa

H_1: *los niveles del fenómeno F en* t_1 *y* t_2 *no son homogéneos, o, lo que es lo mismo, las distribuciones que expresan la ocurrencia por edades de F son diferentes.*

Observemos que el contraste propuesto atiende solamente a la estructura de incidencia por edades del fenómeno estudiado F, esto es, a su calendario. De ahí que las hipótesis a contrastar puedan reformularse así:

H_0: *los calendarios del fenómeno F son iguales*

frente a la alternativa

H_1: *los calendarios del fenómeno F son diferentes.*

Planteadas las hipótesis de esta manera, el contraste considerará que el fenómeno F actúa de forma diferente siempre y cuando los respectivos calendarios sean distintos, aún pudiendo ser igual la intensidad de las dos tablas a comparar.

Este contraste de homogeneidad entre poblaciones conduce a considerar los flujos $D(x, x+1)$ como manifestaciones de la variable aleatoria ξ_x estudiada sobre el colectivo teórico de las L_0 personas. La tabla de presentación de datos, obtenida a partir de sendas tablas demográficas, tendría, pues, la siguiente estructura:

Tabla 1.

Año \ \ Territorio	Edad					Total
	x_0	x_1	.....	$x_\omega - 1$	x_ω	
t_1	$D_0^{t_1}$	$D_1^{t_1}$	.....	$D_{\omega-1}^{t_1}$	$L_\omega^{t_1}$	$L_0^{t_1}$
t_2	$D_0^{t_2}$	$D_1^{t_2}$	.....	$D_{\omega-1}^{t_2}$	$L_\omega^{t_2}$	$L_0^{t_2}$
Total	T_0	T_1	.....	$T_{\omega-1}$	T_ω	$L_0^{t_1} + L_0^{t_2}$

y en donde: $L_0^{t_i}$ es la potencia respectiva de la tabla t_i; para simplificar la notación, se establece que $D(x_k, x_{k+1}) = D_k$ y $L_{x_\omega}^{t_i} = L_\omega^{t_i}$; ω expresa la última categoría de clasificación utilizada y que corresponde a la última edad en la que ya no incide sobre la población el fenómeno demográfico F estudiado; y, finalmente, $D_k^{t_1} + D_k^{t_2} = T_k$, $\forall\, x_k \neq x_\omega$, y $L_\omega^{t_1} + L_\omega^{t_2} = T_\omega$.

En la tabla anterior se incluye como última categoría de clasificación la de los supervivientes finales no afectados nunca por el fenómeno demográfico F. En efecto, la correcta aplicación de la χ^2 en un test de homogeneidad obliga a la consideración de sucesos excluyentes como criterio de clasificación para las poblaciones tratadas (en nuestro caso, las dos tablas demográficas), sucesos que además constituyan una partición de las poblaciones de ambas.

Sobre la tabla anterior se calcula el conocido estadístico:

$$\chi^2 = \sum_{i=t_1,t_2} \sum_{k=0}^{\omega-1} \frac{\left(D_k^i - \frac{L_0^i \times T_k}{L_0^{t_1} + L_0^{t_2}}\right)^2}{\frac{L_0^i \times T_k}{L_0^{t_1} + L_0^{t_2}} +} + \sum_{i=t_1,t_2} \frac{\left(L_\omega^{t_i} - \frac{L_0^i \times T_\omega}{L_0^{t_1} + L_0^{t_2}}\right)}{\frac{L_0^i \times T_\omega}{L_0^{t_1} + L_0^{t_2}}}$$

que se distribuye aproximadamente según una χ^2 con ω grados de libertad. La resolución del contraste, previa fijación de la región crítica, es inmediata: un alto valor para el estadístico χ^2 está indicando que las diferencias entre los valores observados y los esperados son lo suficientemente significativas para determinar comportamientos diferentes del fenómeno F en el tiempo o en el espacio.

Conviene hacer notar que en la expresión anterior de χ^2 el valor absoluto tanto del numerador como el denominador se ven afectados por el orden de magnitud de los

datos utilizados. Esto es, dicho valor se ve afectado por el orden de magnitud del tamaño de la muestra. En efecto, si todos los datos de la tabla que define al estadístico χ^2 se multiplican por una constante, el estimador también se ve afectado por esa misma constante. Este es el motivo por el que, a la hora de trabajar con datos concretos, deben prepararse previamente para que el contraste, por la alta magnitud de las frecuencias a comparar, no rechace sistemáticamente la hipótesis nula: valores inflados del tamaño muestral total invalidan la prueba. Runyon & Haber (1967) advierten de este efecto no deseado hablando del error de la N inflada, y Cochran (1952) demuestra que la potencia del contraste tiende a la unidad cuando el tamaño muestral es grande.

Dado que cada tabla demográfica parte de un efectivo inicial $L_0^{t_i}$ que es arbitrario, la fijación de este valor incial afecta a la aplicación posterior del contraste. De ahí que corresponda fijar la potencia de cada tabla (el efectivo de la generación ficticia inicial). Para una tabla por edades simples es de preveer, aproximadamente, una tabla resultante para el estadístico χ^2 de $2 \times (\omega + 1)$ casillas. Por tanto, el número mínimo de muestra, esto es, de población a considerar deberá ser mayor de $5 \times 2 \times (\omega + 1)$ personas, a fin de que en todas las casillas pudiera aparecer el número mínimo exigido en la metodología general del contraste. En caso contrario, ese número total podría disminuir previa agrupación de categorías de clasificación. También disminuiría considerablemente este número en el caso de tablas con edades agrupadas. En cuanto al valor máximo, su fijación debe atender al efectivo total real del que se han obtenido los datos con los que se calcularon las tablas demográficas a comparar. La correcta aplicación del test de la χ^2 exige que los valores de todas las celdillas de la tabla de datos reflejen la intensidad real del fenómeno demográfico. Se hace necesario, pues, introducir el factor de elevación

$$F^{t_i} = \frac{\widehat{I}_0^{t_i}}{I_0^{t_i}} = \frac{\widehat{I}_0^{t_i}}{\sum_{x=0}^{\omega-1} D_x^{t_i}}$$

en donde $\widehat{I}_0^{t_i}$ es el *número total de sucesos reales A ocurridos en la población de referencia* t_i, factor a partir del cual vuelven a calcularse la potencia y el calendario de cada una de las tablas a utilizar en el contraste.

Así pues, las nuevas potencias respectivas para ambas tablas, $\widehat{L}_0^{t_i}$, serán:

$$\widehat{L}_0^{t_i} = L_0^{t_i} \times \frac{\widehat{I}_0^{t_i}}{I_0^{t_i}} = L_0^{t_i} \times \frac{\widehat{I}_0^{t_i}}{\sum_{x=0}^{\omega-1} D_x^{t_i}} = L_0^{t_i} \times F^{t_i}$$

En el caso de que la tabla se hubiera calculado con los datos de más de un año, el total $\widehat{I}_0^{t_i}$ utilizado en el factor de elevación sería la correspondiente media de los totales de los años considerados. A partir de la potencia $\widehat{L}_0^{t_i}$, los flujos utilizados en la definitiva

tabla de cálculo del estadístico χ^2 se recalculan según:

$$\hat{D}_x^{t_i}=\hat{L}_x^{t_i}\times q_x^{t_i} \qquad x=0,1,\ldots,\omega-1$$

siendo

$$\hat{L}_x^{t_i}=L_x^{t_i}\times F^{t_i} \qquad x=0,1,\ldots,\omega-1,\omega$$

los correspondientes valores de la serie de supervivientes de la tabla demográfica acomodada a la nueva potencia $\hat{L}_0^{t_i}$, por lo que el estadístico χ^2 toma ahora la expresión:

$$(1)\quad \chi^2=\sum_{i=t_1,t_2}\sum_{k=0}^{\omega-1}\frac{\left(\hat{D}_k^i-\dfrac{\hat{L}_0^i\times(\hat{D}_k^{t_1}+\hat{D}_k^{t_2})}{\hat{L}_0^{t_1}+\hat{L}_0^{t_2}}\right)^2}{\dfrac{\hat{L}_0^i\times(\hat{D}_k^{t_1}+\hat{D}_k^{t_2})}{\hat{L}_0^{t_1}+\hat{L}_0^{t_2}}}+\sum_{i=t_1,t_2}\frac{\left(\hat{L}_\omega^{t_i}-\dfrac{\hat{L}_0^i\times(L_\omega^{t_1}+L_\omega^{t_2})}{\hat{L}_0^{t_1}+\hat{L}_0^{t_2}}\right)}{\dfrac{\hat{L}_0^i\times(L_\omega^{t_1}+L_\omega^{t_2})}{\hat{L}_0^{t_1}+\hat{L}_0^{t_2}}}$$

El proceso descrito es equivalente a considerar en la expresión de χ^2 un valor modificado para $D_x^{t_i}$ según la expresión:

$$\hat{D}_x^{t_i}=D_x^{t_i}\times\frac{\hat{L}_0^{t_i}}{L_0^{t_i}}=D_x^{t_i}\times F^{t_i}$$

manteniéndose por tanto la misma estructura de proporcionalidad entre los flujos $\hat{D}_x^{t_i}$ que la existente entre los flujos $D_x^{t_i}$ de la tabla original:

$$\frac{\hat{D}_x^{t_i}}{\hat{D}_{x+1}^{t_i}}=\frac{D_x^{t_i}}{D_{x+1}^{t_i}}$$

siendo además iguales los calendarios de ambas tablas (la original, y la transformada con la nueva potencia) $\hat{d}\,(x,x+1)=d(x,x+1)$, y las respectivas medias de los calendarios $\hat{\bar{d}}=\bar{d}$, al verificarse:

$$\hat{d}(x,x+1)=\frac{\hat{D}_x}{\hat{I}}=\frac{D_x\times F}{I\times F}=d(x,x+1)$$

$$\hat{\bar{d}}=\frac{1}{2}+\frac{1}{\hat{I}}\left(x_0\hat{L}_{x_0}+\sum_{x=x_0+1}^{x_\omega-1}\hat{L}_x-(x_\omega-1)\hat{L}_{x_\omega}\right)=$$

$$=\frac{1}{2}+\frac{1}{\hat{I}}\times\frac{\hat{I}}{I}\left(x_0L_{x_0}+\sum_{x=x_0+1}^{x_\omega-1}L_x-(x_\omega-1)L_{x_\omega}\right)=\bar{d}$$

En definitiva, la correcta aplicación del contraste obliga a homogeneizar la escala de las series de flujo de sucesos de ambas tablas demográficas a utilizar, para que así la generación ficticia inicial de las mismas sea de $\hat{L}_0^{t_i}$ efectivos reales, agrupando posteriormente

las edades necesarias para conseguir frecuencias superiores a 5, según se establece en la teoría general de estos contrastes. Finalmente, la generalización a la comparación simultánea de más de dos tablas demográficas es inmediata.

3. APLICACIÓN

Planteamos dos aplicaciones distintas del contraste de homogeneidad anterior. En la primera de ellas, se estudia la significación estadística de la diferencia entre los calendarios de la mortalidad para la población de dos provincias (comparación territorial); en la segunda, la metodología anterior se aplica para determinar la significación de dos modelos de tablas-tipo, también de mortalidad. Como el fenómeno demográfico considerado en ambos ejemplos —la mortalidad— tiene como intensidad relativa la unidad, la aplicación de la metodología anterior se simplifica notablemente, al resultar ser

$$L_{\omega}^{t_i} = 0$$

y, por lo tanto

$$L_0^{t_i} = l_0^{t_i} \tag{2}$$

con la consiguiente desaparición de la última categoría de clasificación en la Tabla 1.

3.1. Aplicación primera

Para las provincias de Valencia y Alicante, y para la población total (sin distinguir sexo), son conocidas sus tablas completas de mortalidad, centradas en el período 1989-1992, expresadas ambas para unos efectivos iniciales de 10^5 personas. Dados los bajos niveles de mortalidad alcanzados por las poblaciones de ambas provincias —que puede apreciarse comparando, por ejemplo, sus esperanzas de vida como indicador sintético más utilizado—, tiene pleno sentido plantearse por su significación estadística, toda vez que, a simple vista, no llega a apreciarse el sentido y la fuerza de las inevitables diferencias entre ellas.

Las defunciones teóricas para ambos ámbitos, una vez reducida la escala de sus tablas de mortalidad correspondientes a la de una generación ficticia de $\widehat{L}_0^{t_1} = \widehat{l}^{t_1} = 10.401$ personas para Alicante y $\widehat{L}_0^{t_2} = \widehat{l}^{t_2} = 19.073$ para Valencia (media respectiva de las defunciones ocurridas en ambas provincias en el cuatrienio 1989-1992) y agrupadas las edades para conseguir frecuencias esperadas superiores a 5, se recogen en la tabla siguiente:

Tabla 2. Defunciones teóricas. Población total. Tablas de mortalidad 1989-1992. Provincias de Alicante y Valencia

Edad	Alicante	Valencia	Edad	Alicante	Valencia	Edad	Alicante	Valencia
0	63	134	39	14	30	71	201	394
1	8	14	40	15	31	72	221	425
2	5	8	41	17	33	73	236	460
3 y 4	6	13	42	19	35	74	254	490
5 y 6	6	9	43	20	36	75	270	522
7 y 8	5	7	44	21	39	76	287	556
9 a 11	6	10	45	24	44	77	307	600
12 y 13	5	9	46	25	49	78	332	627
14 y 15	9	16	47	28	54	79	358	664
16	7	12	48	28	57	80	377	698
17	8	14	49	32	65	81	402	726
18	9	16	50	34	73	82	417	736
19	9	17	51	38	78	83	420	755
20	9	18	52	40	84	84	418	751
21	9	18	53	46	97	85	415	725
22	10	19	54	48	100	86	401	693
23	10	21	55	53	105	87	376	655
24	11	22	56	57	116	88	352	602
25	11	23	57	64	125	89	328	551
26	11	25	58	69	135	90	299	495
27	11	26	59	75	146	91	251	412
28	12	26	60	82	160	92	217	353
29	12	26	61	89	173	93	188	305
30	11	26	62	96	185	94	158	258
31	11	26	63	104	203	95	128	209
32	12	25	64	114	220	96	96	159
33	11	25	65	127	240	97	67	111
34	11	27	66	136	259	98	43	70
35	11	27	67	148	282	99	24	40
36	12	27	68	162	308	100	12	20
37	12	28	69	173	332	101 ó +	7	13
38	13	29	70	185	361			

A partir de esas defunciones teóricas, el estadístico χ^2 definido en (1) toma el valor:

$$\chi^2 = 35,31$$

lo que, para una significación del 5% en una χ^2_{94}, no permite rechazar la hipótesis nula.

Conclusión: las estructuras por edades de la mortalidad de las provincias de Alicante y Valencia, para el período 1989-1992, son significativamente iguales.

3.2. Aplicación segunda

La segunda aplicación se realiza sobre un caso muy particular de tablas demográficas: las *tablas-tipo* o *tablas-modelo* de mortalidad. Estas tablas suministran las relaciones

empíricas entre ciertos datos sobre la mortalidad conocidos, pero incompletos, y las series biométricas de una tabla de mortalidad abreviada. Su uso está muy extendido, y es de plena aplicabilidad en los procesos proyectivos. La notación desarrollada con anterioridad se simplifica notablemente ahora, ya que el fenómeno demográfico mortalidad tiene como intensidad la unidad y las potencias de ambas tablas son iguales.

El objetivo de esta segunda aplicación, pues, consiste en determinar la significación de dos modelos de tablas-tipo, correspondientes a las mujeres para los niveles 26 y 27 de la zona occidental, tomadas de Coale y Guo (1991). Y, concretamente, determinar a partir de qué potencia esas tablas no pueden considerarse representativas de mortalidades diferentes, para cierta significación estadística previamente fijada.

La Tabla 3 recoge la serie de defunciones teóricas de dichas tablas. La potencia utilizada en ellas es de 10^6 efectivos:

Tabla 3. Tablas-tipo de mortalidad

	Nivel 27		Nivel 26	
Edad	**Supervivientes**	**Defunciones**	**Supervivientes**	**Defunciones**
0	1,000,000	2,831	1,000,000	4,033
1	997,169	1,194	995,967	1,454
5	995,975	145	994,513	459
10	995,830	128	994,054	383
15	995,702	571	993,671	1,003
20	995,131	1,126	992,668	1,433
25	994,005	1,495	991,235	1,825
30	992,510	1,673	989,410	2,170
35	990,837	2,205	987,240	3,027
40	988,632	3,458	984,213	4,711
45	985,174	5,306	979,502	7,309
50	979,868	9,276	972,193	11,523
55	970,592	11,549	960,670	16,333
60	959,043	17,190	944,337	24,888
65	941,853	24,787	919,449	37,934
70	917,066	43,813	881,515	64,897
75	873,253	89,801	816,618	115,903
80	783,452	167,203	700,715	185,228
85	616,249	237,742	515,487	225,531
90	378,507	226,049	289,956	184,253
95	152,458	120,321	105,703	85,478
100	32,137	32,137	20,225	20,225
Total		1,000,000		1,000,000

La expresión del estadístico χ^2, en este caso de coincidencia de potencias e intensidad unidad, es más simple (ver 2):

$$\chi^2 = \sum_{i=niveles26,27} \sum_{k=0}^{\omega-1} \frac{\left(D_k^i - \frac{I^i \times T_k}{I^{26}+I^{27}}\right)^2}{\frac{I^i \times T_k}{I^{26}+I^{27}}} = \sum_{i=niveles26,27} \sum_{k=0}^{\omega-1} \frac{\left(D_k^i - \frac{T_k}{2}\right)^2}{\frac{T_k}{2}} =$$

$$= \sum_{k=0}^{\omega-1} \frac{\left(D_k^{26} - \frac{D_k^{26}+D_k^{27}}{2}\right)^2}{\frac{D_k^{26}+D_k^{27}}{2}} + \sum_{k=0}^{\omega-1} \frac{\left(D_k^{27} - \frac{D_k^{26}+D_k^{27}}{2}\right)^2}{\frac{D_k^{26}+D_k^{27}}{2}} =$$

$$= \frac{1}{2} \cdot \sum_{k=0}^{\omega-1} \frac{\left(D_k^{26}-D_k^{27}\right)^2}{\left(D_k^{26}+D_k^{27}\right)} + \frac{1}{2} \cdot \sum_{k=0}^{\omega-1} \frac{\left(D_k^{27}-D_k^{26}\right)^2}{\left(D_k^{26}+D_k^{27}\right)} = \sum_{k=0}^{\omega-1} \frac{\left(D_k^{26}-D_k^{27}\right)^2}{D_k^{26}+D_k^{27}}$$

Una primera aproximación a la situación planteada queda recogida en la Tabla 4, en la que se aplica el contraste desarrollado en el apartado anterior para diferentes potencias. Dado el objetivo previsto, en esta aplicación no es necesario modificar las potencias de las tablas-tipo:

Tabla 4. Nivel de significación: 5%

Potencia de la tabla	χ^2	Grados libertad	Significación
1.000.000	28157.7	21	Si
10.000	281.2	19	Si
1.000	28.0	13	Si
500	14.0	11	No

De la Tabla 4 se deduce que la potencia de las tablas-tipo que hace no rechazable la hipótesis de su igualdad debe estar situada entre los valores 500 y 10^3. Por lo tanto, y tras la aplicación sucesiva del contraste anterior para un nivel de significación del 5%, se concluye que la potencia límite a partir de la cual existe significación es de 752 individuos ($\chi^2 = 21.02$ y 12 grados de libertad).

Conclusión: hay que descender a una población extraordinariamente pequeña —con un número total anual máximo de defunciones de 752 mujeres— para poder afirmar que las tablas correspondientes a los niveles 26 y 27 para las mujeres de la zona occidental, tomadas de Coale y Guo (1991), responden a una estructura de mortalidad semejante.

BIBLIOGRAFÍA

Coale, A. & Guo, G. (1991). «Utilización de nuevas tablas modelo de mortalidad para tasas de mortalidad muy bajas en proyecciones demográficas», en *Boletín de Población de las Naciones Unidas*, 30. Naciones Unidas. Nueva York.

Cochran, W.G. (1952). «The χ^2 test of goodness of fit». *Ann. Math. Statist.*, 23, 315-345.

Cochran, W.G. (1954). «Some methods for strengthening the common χ^2 test». *Biometrics*, 10, 417-451.

Leguina, J. (1981). *Fundamentos de Demografía.* Siglo XXI Editores. Madrid.

Runyon, R. & Haber, A. (1967). *Fundamentals of Behavioral Statistics.* Addison-Wesley. Massachusetts.

Suchindran, C.M. & Namboodiri, K. (1987). *Life Table Techniques and their Application.* Orlando, Academic Press.

ENGLISH SUMMARY

COMPARISON OF TWO DEMOGRAPHIC TABLES: APPROXIMATION TO THEIR STATISTICAL SIGNIFICANCE

ERNESTO J. VERES FERRER
Universidad de Valencia*

In this work is analyzed the statistical significance of the possible equality of two demographic tables. Concretely, it is presented the applicability of a classic tests of the statistical methods —that Chi-Square test— to contrast the null hypothesis «H_0 = two demographic tables are equal, this is, answer to a same structure of the studied demographic phenomenon», against the alternative that denies it. Both tables are referred to an only demographic phenomenon for the same territorial area and two different moments of time (temporal test), or for the same temporal reference in two populations of different territorial area (territorial test). The methodology applied is described for two situations: for the comparison of the mortality levels of two provinces, and on two tables-type corresponding at each mortality levels.

Keywords: Calendar, Chi-Square test, demographic table, hypothesis testing, intensity, life tables-type

AMS Classification: 62F03, 62G10, 62P05, 62P25, 92 H20

---

* Departamento de Economía Aplicada. Facultad de Ciencias Económicas y Empresariales. Universidad de Valencia. Campus de los Naranjos. 46022 Valencia. E-mail: Ernesto.Veres@uv.es.

- Received March 1999.
- Accepted November 1999.

The model known as *demographic table* allows one to analyse a single demographic phenomenon F by means of a certain characteristic event A, which we suppose is unrepeatable. The information for the analysis is provided by observing the incidence of event A on a cohort, studying the frequency with which this characteristic event begins to appear from an initial age −which we denote by x_0- , until a final age when the event no longer appears −denoted by $x_\omega-$.

The demographic table includes three fundamental biometric series. Once any of the series is known, the other two are also known: *series of survivors*, $\left\{L_x\right\}_{x=x_0}^{x_\omega}$; *event flow series*, $\left\{D(x,x+1)\right\}_{x=x_0}^{x_\omega-1}$; and *series of probabilities of an occurrence of event A, or hazard series*, $\left\{q_x\right\}_{x=x_0}^{x_\omega-1}$.

Intensity I and *Calendar* $d(x,x+1)$, are two basic analytical indices which are easily deduced from a demographic table.

With the previous elements, in this paper we intend to determine, with statistical significance, if two demographic tables −corresponding to two different moments, or to two distinct territories− are equal. For this we propose the homogeneity test based on the χ^2 Chi-Square test, applied to the respective series $D(x,x+1)$.

When there are large differences between the two demograghic tables, their significance is plainly evident. In fact, the order of magnitude of the differences between their respective biometric series confirms the variation −in one direction or another− between the levels reached by phenomenon F expressed by both tables. It is irrelevant, therefore, to carry out any other type of analysis. However, when the differences are reached by said biometric series and the classic indicators deduced from these are not sufficiently large to evaluate their significance at a simple glance, we can wonder about the *statistical significance* of these small differences. We can therefore understand the full application of the statistical techniques and models to attempt to explain it (this approach involves incorporating technical demographic tables into the study, which are Statistical Methods) and, in particular, the hypothesis testing.

We therefore consider that the flows of events A gathered in each series of the two demographic tables considered correspond to the observations of the two populations −those derived from the incidence of phenomenon F in two moments in time, or in two territories, to be compared−, and that they have been classified according to the same criteria «age of the person when characteristic event A takes place». We denote by $L_0^{t_1}$ and $L_0^{t_2}$ the respective powers of both demographic tables, which correspond to the different times or territories t_1 and t_2.

The population associated with each of these tables may be represented by means of a random variable ξ_x, which adopts values on the categories of classification contained in

these. The formulation of the hypothesis to be contrasted is established in the following well-known terms:

H_0: levels of phenomenon F in t_1 and t_2 are homogeneous,
or, equally, distributions which express the occurrence of ages of F are equal
as opposed to the alternative
H_1: the levels of phenomenon F in t_1 and t_2 are not homogenous, or,
equally ,distributions which express the occurrence by ages of F are different

We observed that the proposed contrast only looks at the structure of incidence by ages of phenomenon F studied, that is, to its calendar. Hence the hypothesis to be contrasted may be reformulated in the following way:

H_0: the calendars of phenomenon F are equal
as opposed to the alternative
H_1: the calendars of phenomenon F are different

When the hypotheses are described in this way, the contrast will consider that phenomenon F behaves in a different way provided that the respective calendars are different, even though the intensity of the two tables to be compared are equal.

This homogenous contrast between populations leads one to consider flows $D(x,x+1)$ as manifestations of the random variable ξ_x studied on the collective theory of L_0 people. The data double entry table, obtained from these democratic tables, would therefore have the following structure:

Year \ \ Territory	**Age**					**Total**
	x_0	x_1	.....	$x_\omega - 1$	x_ω	
t_1	$D_0^{t_1}$	$D_1^{t_1}$	.....	$D_{\omega-1}^{t_1}$	$L_\omega^{t_1}$	$L_0^{t_1}$
t_2	$D_0^{t_2}$	$D_1^{t_2}$	.....	$D_{\omega-1}^{t_2}$	$L_\omega^{t_2}$	$L_0^{t_2}$
Total	T_0	T_1	.....	$T_{\omega-1}$	T_ω	$L_0^{t_1}+L_0^{t_2}$

in which: $L_0^{t_i}$ is the respective power of table t_i; to simplify the notation one establishes that $D(x_k,x_{k+1}) = D_k$ and $L_{x_\omega}^{t_i} = L_\omega^{t_i}$; ω expresses the last classification category used which corresponds to the last age when demographic phenomenon F studied has no

more incidence on the population; and, finally, $D_k^{t_1} + D_k^{t_2} = T_k, \forall x_k \neq x_\omega$, and $L_\omega^{t_1} + L_\omega^{t_2} = T_\omega$.

The previous table includes as a last category of classification those final survivors who were never affected by demographic phenomenon F, in order for the correct application of the χ^2 in a homogeneity test, which forces one to consider the excluded events as a criteria of classification for the populations dealt with (in our case, the two demographic tables), events which also constitute a partition of both populations.

One should note that the homogeneity contrast is sensitive to the order of magnitude of the data used. This is why, when working with concrete data, these must be previously prepared so that the contrast, by the high magnitude of frequencies to compare, does not systematically contradict the null hypothesis: inflated samples of total sample size invalidate the test. The preparation of the original information is carried out taking into account the effective totals of the real flow of events from those which have been obtained from the data with which the demographic tables were calculated to be compared.

Finally, in this paper two different applications of the previous homogeneity contrast are proposed. In the first of these, we study the statistical significance of the difference between the mortality calendars for the population of the two provinces (territorial comparison); and in the second, the previous methodology is applied in order to determine the significance of the two life table-type models.

Secció Docent i Problemes

SECCIÓ DOCENT I PROBLEMES

La «Secció docent i problemes» té l'objectiu de publicar articles de caire docent, difícilment publicables en revistes de recerca. A cada número de *Qüestiió* s'inclouen d'un a tres problemes i les solucions es donen en el número següent.

Els lectors poden proposar problemes amb les solucions pertinents i enviar-los a *Qüestiió,* que farà una selecció i en publicarà els més adequats, fent la corresponent referència a l'autor.

També seran ben rebudes solucions alternatives a les propostes fetes per l'autor dels problemes. L'editorial es reservarà, però, el dret a publicar-les.

SOLUCIONS ALS PROBLEMES PROPOSATS AL VOLUM 23 N. 3

El problema n. 82, amb un nou enunciat i la seva solució, es publicaran en el proper número de *Qüestiió*.

PROBLEMA N. 83

Los estimadores de regresión lineal multivariante (que incluyen al estimador de regresión lineal univariante tradicional) verifican que para todo $\varepsilon > 0$, existe $\varepsilon' = \varepsilon/(k+1) > 0$, tal que

$$1 \geqslant P\left\{\left|\hat{\bar{Y}} - \bar{Y}\right| < \varepsilon = (k+1)\varepsilon'\right\} =$$

$$= P\left\{\left|\bar{y} - \bar{Y} + \sum_{j=1}^{k} b_j\left(\bar{X}_j - \bar{x}_j\right)\right| < (k+1)\varepsilon'\right\} \geqslant$$

$$\geqslant P\left\{\left|\bar{y} - \bar{Y}\right| < \varepsilon', |b_1|\left|\bar{x}_1 - \bar{X}_1\right| < \varepsilon', \ldots, |b_k|\left|\bar{x}_k - \bar{X}_k\right| < \varepsilon'\right\} =$$

$$= P\left(\bigcap_{j=0}^{k}\left\{\left|\bar{x}_j - \bar{X}_j\right| |b_j| < \varepsilon'\right\}\right) \geqslant$$

$$\geqslant P\left(\bigcap_{j=0}^{k}\left\{\left|\bar{x}_j - \bar{X}_j\right| < \varepsilon'/b\right\}\right) \longrightarrow 1$$

cuando $n \longrightarrow \infty$, donde $b_0 = 1, \bar{x}_0 = \bar{y}$ es consistente para $\bar{X}_0 = \bar{Y}$, y además $\bar{x}_j$ es consistente para $\bar{X}_j$ $(j = 1,2,\ldots,k)$. También hemos denotado por b a la cota

$$b = \sup\{|b_i| : i = 0,1,2,\ldots,k\} < \infty.$$

La convergencia a 1 (cuando $n \longrightarrow \infty$) puede demostrarse por inducción en $k = 1,2,\ldots$, lo que concluye la demostración.

M. Ruiz Espejo
UNED

Housila P. Singh
Vikram University

PROBLEMES PROPOSATS

PROBLEMA N. 84

Sea X una variable aleatoria con distribución absolutamente continua, función de distribución $F(x)$ y función de densidad $f(x)$. Supongamos $P(X > 0) = 1$. Sea X_β la variable condicionada a $X \leqslant \beta$, siendo $\beta > 0$ un valor constante. Se pide:

1) Hallar la función de densidad de X_β.

2) Sea $\sigma(\beta)$ la desviación típica de X_β, que suponemos existe y es finita. Probar la desigualdad

$$\beta^2 \inf_{0 \leqslant x \leqslant \beta} \{f(x)\} \leqslant \sqrt{12} F(\beta) \sigma(\beta).$$

C.M. Cuadras
Universitat de Barcelona

PROBLEMA N. 85

Sea $(X_1, Y_1), \ldots, (X_n, Y_n)$ una muestra aleatoria simple de (X, Y) con distribución bivariante continua. Consideremos la hipótesis nula

$$H_0: \quad (X, Y) \text{ tiene la misma distribución que } (Y, X).$$

Sea $Z = X - Y$. Se pide:

1) Probar que bajo H_0 la distribución de Z es simétrica respecto del origen.

2) Probar que aceptar

$$H_1: \quad \text{la mediana de } Z \text{ es positiva,}$$

implica rechazar H_0.

3) Proponer un test no paramétrico para contrastar H_0 frente H_1.

C.M. Cuadras
Universitat de Barcelona

Comentaris de llibres

Introducció a l'anàlisi i disseny d'algorismes

Francesc J. Ferri, Jesús V. Albert, Gregorio Martín

Publicacions de la Universitat de València (1998), 307 pp.

Sempre he cregut que els primers cursos d'universitat han de ser impartits per professors molt bons, carregats de seny i d'experiència. Els llibres de text corresponents han de ser exposicions clares i profundes que només poden donar-se en la conjunció entre una comprensió aguda de la matèria i una preocupació didàctica per exposar-la. Al meu entendre el seu valor formatiu, científic i didàctic rau en una formulació rigorosa, en una ordenació lògica impecable i en estil de llenguatge fluent i seductor. En altres paraules, un llibre de text ha de combinar les exigències lògiques amb les suavitats retòriques que estan implicades en tota construcció científica i en tota comunicació.

El llibre «Introducció a l'Anàlisi i Disseny d'Algorismes» és un llibre de text que, al meu parer, intenta assolir les qualitats ideals esmentades i ho aconsegueix en gran mesura. El text s'adreça a estudiants d'informàtica i ofereix un estudi dels algorismes, com un tema central de la programació computacional i la resolució de problemes, que recull l'estructura lògica dels programes que després es poden desenvolupar en llenguatges i ordinadors determinats.

Per tant, el llibre és una introducció a la programació computacional que explica els conceptes bàsics i s'estén en les diverses classes d'algorismes i ens ofereix els instruments lògics i matemàtics que necessitem per treballar-los. El text conté molts exemples de problemes i d'algorismes i aconsegueix un bon equilibri entre la presentació conceptual i el desplegament de casos.

La preparació del llibre va comptar amb una beca del Servei de Normalització Lingüística de la Universitat de València per a la redacció de manuals en català. En aquest aspecte, he llegit el llibre reconeixent-lo escrit en aquesta llengua comuna de molta gent, de moltes regions i de diversos antics regnes, en les harmòniques variacions valencianes. La seva aportació en el llenguatge científic te dues qualitats remarcables. La primera és l'adaptació i creació d'una terminologia tècnica molt encertada, que té arrels profundes en la nostra llengua. Al capdavall, la vida d'una llengua viva es manifesta en la seva creativitat per adaptar els seus termes i expressions a les noves situacions, objectes i conceptes, incloent-hi totes les ciències i tècniques. La segona és l'estil fluid de llenguatge, que ens incita a la lectura i que ens aplana les dificultats de comprensió. Aquestes virtuts desgraciadament no són una característica de molts dels textos científics actuals.

Eduard Bonet

La estadística en cómic

Larry Gonick y Woollcott Smith

Editorial Zendrera Zariquey, Barcelona, 1999, 232 pp.

Traducción: Laura Monero

Revisado por Erik Cobo, Guadalupe Gómez y Pilar Muñoz

(Traducción de la Edición Original *The Cartoon Guide to Statistics*, Haper-Collins Publishers, Inc., 1993)

Antes de redactar lo que sigue me he hecho la siguiente pregunta: ¿debería un estadístico serio como yo escribir este comentario? Pero he llegado a la conclusión de que soy más estadístico que serio, así que me he decidido a escribirlo para los lectores de Qüestiió.

Desde los albores de la moderna estadística, ésta se ha caracterizado por la presentación de sus resultados y conclusiones mediante gráficos y figuras. Aunque unos pocos valores numéricos pueden compendiar una larga tabla de datos, el usuario agradece un gráfico que le permita visualizar el material estadístico. *La Estadística en Cómic* lleva esta filosofía a su grado máximo. Como el título nos indica, explica e ilustra la estadística mediante viñetas de tebeo, en el más puro estilo del cómic. La fórmula es siempre la misma: una explicación técnica, un breve comentario y un chiste. Con esta estructura tan simple como divertida y eficaz, los autores ilustran, a lo largo de doce capítulos, lo esencial de la probabilidad y la estadística.

Los capítulos 1 y 2 se dedican a introducir la estadística e ilustrar sus aspectos descriptivos. En general, las viñetas siguen un guión (en la introducción, las dudas estadísticas de una pareja eligiendo menú en un restaurante). El hilo argumental, en su parte más «seria», lo lleva un señor E (el del restaurante), calvo, con gafas y cara de ingenuo. Los capítulos 3 al 5 se dedican a la probabilidad, variables aleatorias y distribuciones, con especial atención a la binomial y normal y al teorema central del límite (ilustración que acompaña este comentario). Aparecen muchos dados, histogramas, curvas, tablas y la complicidad de algunos personajes históricos (el caballero De Meré, empedernido jugador y mujeriego, Pascal que se cartea con Fermat, Bernoulli, De Moivre). Aparece también un tal señor X, personaje al principio bastante normal, pero que se va volviendo loco a medida que se va dando cuenta de que para entender la estadística hay que saber probabilidad.

El capítulo 6 trata del muestreo en sus diferentes formas, todas bien ilustradas, y se introduce la t de Student. Al llegar aquí, el señor X ya necesita una camisa de fuerza. El capítulo 7 trata de la estimación y los intervalos de confianza, se explica el razonamiento inductivo, con la inestimable ayuda de Sherlock Holmes y su ayudante Dr. Watson. Es

quizás el capítulo 8, sobre contraste de hipótesis, donde mejor se aprecia la agudeza de los chistes y comentarios, tratando por separado los casos de muestras grandes y pequeñas, así como los diferentes pasos en la toma de decisiones. Aquí acaba el primer curso de estadística clásica, con el señor E imponiendo la birreta a una señora que se entera bastante, aunque no lo ve muy claro (el señor X sigue loco).

En efecto, llegan las poblaciones y el diseño experimental, contenido de los capítulos 9 y 10, en los cuales las ideas se concretan en situaciones más cotidianas, con aspirinas, muchos coches y precios de la gasolina. La regresión se ilustra en el capítulo 11, motivando al lector con intrigantes preguntas (¿leer libros de estadística te convierte en mejor persona?) sobre relación y causalidad. Los principales temas (modelo lineal, múltiple, no lineal, intervalos de confianza, etc.) se tratan del mismo modo: un guión, chistes, ejemplos, el señor E que aclara las cosas y el señor X que empeora.

El capítulo 12 y último es un breve compendio de métodos avanzados, una viñeta para cada uno. Empieza (no podían faltar en un libro así) comentando las caras de Chernoff y diferentes métodos modernos: análisis de conglomerados, discriminante, factorial,

paseos aleatorios, series temporales, tratamiento de imágenes y remuestreo. La viñeta que ilustra el análisis factorial (un psicoanalista y una señora en un diván en lugar del señor X) demuestra que el señor X, o ya se ha curado o es un caso perdido (quizás, cuando ya empezaba a entender la estadística, alguien le ha dicho que ahora hace falta dominar la teoría de la medida). Es posible que los autores, en mensaje oculto, pretendan identificar a la mayoría de estadísticos como sanos señores E o desconcertados señores X. Hay una referencia a R.A. Fisher y a la necesidad de cuidar la calidad de los datos. No se pierdan el gráfico estadístico final, que prueba lo teóricos (por no decir inútiles) que nos hemos vuelto muchos estadísticos desde 1900 hasta la actualidad.

La bibliografía, comentada también en estilo cómic, orienta al lector a buscar referencias según los temas: estudiante, estudiante apasionado, juegos y apuestas, legislación y sociedad, gráficos, historia y software estadístico.

Estamos ante un libro inédito en su presentación, muy original y útil para hacer llevadero y divertido el aprendizaje de la estadística, recomendable para estudiantes universitarios, profesionales y usuarios de la estadística, como complemento de otros libros de texto convencionales.

C.M. Cuadras

An Introduction to Copulas

Roger B. Nelsen

Lecture Notes in Statistics, 139.

Editorial: Springer, New York, 1999, 216 pp.

Tanto el interés de esta obra sobre modelos bivariantes de probabilidad como la proximidad del congreso *Distributions with Given Marginals and Statistical Modelling,* Barcelona, 17-20 de julio de 2000, justifican este comentario.

Una «cópula» es una función de distribución bivariante $C(u,v)$, con $u,v \in [0,1]$, tal que $C(1,v) = v, C(u,1) = u$, es decir, las distribuciones marginales son uniformes en el intervalo $[0,1]$. Cualquier función de distribución bivariante $H(x,y)$ define una cópula (que es única si H es continua), y recíprocamente, mediante una cópula podemos generar una distribución bivariante (Teorema de Sklar).

Las cópulas más importantes son las cotas de Fréchet $\min\{u,v\}, \max\{u+v-1,0\}$ y la cópula de independencia uv. Otras cópulas se construyen mediante expresiones que dependen de uno o más parámetros, algunas sencillas y otras complicadas. Todo esto lo explica el autor en los capítulos 1 y 2, sobre propiedades generales, incluyendo el caso multivariante, las cópulas asociadas y de supervivencia, las relaciones entre cópulas y la generación de datos siguiendo una cópula. Las cotas, en el caso general de marginales F, G, son $\min\{F(x), G(y)\}, \max\{F(x)+G(y)-1,0\}$, que proporcionan las distribuciones con mínima y máxima correlación entre las variables, un hecho descubierto por Hoeffding, por lo que tales acotaciones para toda distribución $H(x,y)$ con marginales F, G, reciben el nombre de cotas de Fréchet-Hoeffding.

El capítulo 3 trata de la construcción de cópulas, con una exposición sistemática de las cópulas y distribuciones bivariantes más conocidas (Farlie-Gumbel-Morgenstern, Marshall-Olkin, Cuadras-Augé, Mardia, Gumbel, Pareto, Kimeldorf-Sampson, etc.), cópulas con secciones y la construcción llamada «suffle of Min», que resulta de repartir al azar secciones de la cópula $\min\{u,v\}$ y que tiene la curiosa propiedad de que se puede aproximar tanto como se quiera a la cópula de independencia uv, a pesar de que para la «suffle of Min» existe relación biunívoca entre las variables aleatorias. Las figuras son muy instructivas, demostrando buen dominio en la explicación mediante gráficos. (Roger B. Nelsen es también autor de *Proofs Without Words: Exercises in Visual Thinking.)*

El capítulo 4 trata de las cópulas Arquimedianas, que son las que pueden expresarse en la forma $C(u,v) = \phi^{-1}(\phi(u)+\phi(v))$, y por tanto sólo dependen de una función ϕ. Las cópulas de Frank, Clayton-Oakes, Ali-Mikhail-Haq, son Arquimedianas. Un amplio cuadro central presenta tales cópulas, incluyendo la función generadora ϕ, el rango del

parámetro y las distribuciones límites. El interés de tales cópulas reside en que tienen un tratamiento estadístico más simple, por ejemplo, a través de la τ de Kendall, que es un funcional en ϕ.

El capítulo 5 estudia la dependencia entre variables aleatorias a través de las cópulas. Sólo necesitamos conocer la cópula $C(u,v)$ de una distribución (lo que Kimeldorf y Sampson llaman la representación uniforme) para definir y estudiar ciertos conceptos de dependencia estocástica. Por ejemplo, la correlación ρ de Spearman y el coeficiente de dependencia τ de Kendall se pueden expresar como

$$12\int(C(u,v)-uv)dudv, \quad 4\int(C(u,v)-uv)dC(u,v),$$

respectivamente. El autor proporciona fórmulas explícitas para familias particulares, así como un estudio de las relaciones entre ρ y τ. Se explican muchos otros conceptos de dependencia, como el de «positivamente cuadrante dependiente»: $C(u,v) \geq uv$, que implica que correlación y otras medidas de dependencia son positivas. Se incluye una sistemática relación entre tales conceptos, cuyo origen y justificación a menudo procede del análisis de supervivencia, que son invariantes por transformaciones monótonas de las variables, y que tienen interés en estadística no paramétrica.

El capítulo 6 sobre tópicos adicionales, trata de la minimización de distancias entre distribuciones, el uso de cópulas y de las llamadas quasi-cópulas, para relacionar operaciones sobre funciones de distribución y sobre las variables aleatorias correspondientes. También estudia una operación $C_1 * C_2$ entre cópulas, que goza de interesantes propiedades (es asociativa, uv es elemento nulo, $\min\{u,v\}$ es elemento unidad, etc.) y que tiene interés para construir procesos de Markov, pues proporciona una condición equivalente a las ecuaciones de Chapman-Kolmogorov. Sin embargo, no trata de temas como las llamadas expansiones diagonales, o la modelización de datos bivariantes, pero que se pueden encontrar en otras obras de contenido similar.

Con alrededor de 100 ejemplos y 150 ejercicios seleccionados, *An Introduction to Copulas* es una excelente introducción al tema, de gran utilidad para estadísticos y probabilistas que deseen tener una visión clara y actualizada sobre las distribuciones bivariantes y multivariantes, y sobre los diferentes conceptos de dependencia.

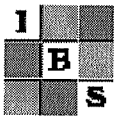
C.M. Cuadras

Ressenyes d'activitats institucionals

Sociedad Española de Biometría

Región Española
de la Sociedad Internacional de Biometría

Spanish Region
of the International Biometric Society



http://www.iata.csic.es/ibsresp

La **Sociedad Española de Biometría/Región Española de la Sociedad Internacional de Biometría** (abreviadamente **SEB** o ***REsp***) tiene como objetivos promover, impulsar y difundir el desarrollo y la aplicación de los métodos matemáticos y estadísticos a la biología, medicina, psicología, farmacología, agricultura y otras ciencias afines (ciencias relacionadas con los seres vivos). Cualquier profesional o alumno de estas disciplinas puede ser miembro de la SEB.

Consejo Directivo

Presidenta:	Guadalupe Gómez Melis (Biología)
Vicepresidente:	María Jesús Bayarri García (Medicina)
Secretario y Tesorero:	Fernando López Santoveña (Agronomía)
Vocal en calidad de Miembro del Consejo de la IBS:	Emilio A. Carbonell Guevara (Agronomía)
Vocales:	Juan Luis Chorro Gascó (Psicología)
	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

La VIII CONFERENCIA ESPAÑOLA DE BIOMETRÍA

se celebrará en Pamplona

los días 28, 29 y 30 de marzo 2001.

La información preliminar puede ser consultada en

http://www.unavarra.es/directo/congresos/apoyo/biometria.htm

The TES Institute
Training of European Statisticians

GENERAL INTRODUCTION

Seeing the need of harmonising statistics at the European level, Eurostat decided in the early nineties to set up the TES Project. This programme was in charge of offering truly European vocational training and staff development opportunities at post-graduate level through annual training programmes for target groups ranging from young statisticians to executives of National Statistical Institutes.

In November 1996. the TES Project became the TES Institute, a non-profit association created by ten Member States of the European union and the four Member States of the European Free Trade Association. At present it counts among its members the representatives of seventeen European National Statistical Institute and of the Centre Universitaire de Luxembourg.

The training programmes offered by the TES Institute provide both theoretical and practical background but the courses have all a very strong applied character.

These programmes also offer participants the opportunity to meet colleagues form all over Europe and other countries since the TES Institute has extended its activities to the Central European, Mediterranean Basin and TACIS countries.

The above characteristics represent the basic conditions to acquire sharper competence in their work environment and highlight the European dimension of their activity.

After ten years of existence, the programme became entire part of the statistical world. For the time being, around 500 participants coming from more than 30 countries are trained every academic year. Such an interest is mainly due to the large number of courses on offer. Indeed, the TES portfolio comprises more than 80 courses of short duration all at post-graduate level.

After a few years of co-operation with the Central European countries, the TES Institute has recently extended the co-operation to the MEDSTAT and TACIS region. Such an internationalisation is the direct result of the growing importance of training as a part of the current technological and intellectual development. Therefore, as far as statistics and economics are concerned, it is of the utmost importance to extend the best national practices to an international level.

It is obvious that the TES programmes should be considered as a complement and not a substitute to the training provided at national level.

In brief, one may say that by offering training opportunities which are complementary to the ones provided at national level, the TES Institute is offering a new approach of the subsidiarity concept.

TRAINING

The 1999-2000 Programme offering 28 courses has started from September 1999 and will last until July 2000.

You will find an updated overview of the remaining courses at the end of this paper. Please note that the following four courses have been postponed compared to the initial planning:

- The Revised System of Accounts (ESA95) - Sector Accounts
- The Revised System of Accounts (ESA95) - Quarterly Accounts
- The European Statistical System
- Confidentiality and Protection of Privacy

Beside the execution of the subsidised annual vocational training programme –which is only one of our many activities– 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 the training of trainers, the setting up of training centres, reform of training policy and curricula. For example, the TES Institute has also been involved in consulting activities with Ukraine and the Russian Federation for the set-up of training centres and for drafting recommendations for the improvement of the vocational training system for the Russian Statistical Agency.

RESEARCH

In view of constantly upgrade and update its services, the TES Institute has also recently been involved in various *Research* projects such as:

- A project on distance education for developing a course on *European Economic Statistics* in co-operation with the University of Ljubljana The outputs of the project were a CD-Rom, a manual of the course as well as an online site.
- Several distance education projects (Virtual Library - Computer Assisted Training in Statistics) within the scope of the 5th Framework Programme of the EU.
- A project concerning the development of a «Thematic Network» of technology transfer in co-operation with the Joint Research Centre of Ispra and the Universitat Politecnica de Catalunia within the 5th Framework Programme of the EU.

OTHER ACTIVITIES

The TES institute has been associated with the Maastricht School of Management as training and advice provider in the framework of their MBA in Decision Support Systems.

You can directly contact the TES Institute for any further information on this MBA.

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.

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 Indices for bilateral and *multilateral Comparison of Prices, Quantities and Values.*
- Further manuals will cover topics of *Sampling Techniques, Seasonal Adjustment Methods* and *Social Statistics.*

TES NEWSLETTER

Beginning of February, the TES Institute has disseminated the first issue of its Newsletter named «Facts & Visions». 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.

We hope that our partners will use «Facts & Visions» to report on their own activities. So, may we encourage you to use your keyboards for European information purposes.

Copies of «Facts & Visions» are available at the TES Institute as well as guidelines for the submission of articles.

GENERAL INFORMATION

For further details on any of the above mentioned topics, please contact directly Ms Valérie Vandewalle (co-ordinator for *Evaluation and Information* matters):

by phone: (352) 29.85.85.34
fax: (352) 29.85.29
or e-mail: vvandewalle@tes-institute.lu

OVERVIEW OF COURSES IN CHRONOLOGICAL ORDER				
Code	**Title**	**Course Leader**	**Location**	**Date**
DAT-102-E	Dealing with non-Response	LYNN	London	3-7 Apr 2000
SUP-201-E	Seasonal Adjustment Methods	MARAVALL	Luxembourg	10-14 Apr 2000
ECO-203-E	The Revised European System of Accounts (ESA 95) Financial Accounts	COIN	Luxembourg	26-28 Apr 2000
DAT-001-E	Notions of Sampling and Survey for Managers	DROESBEKE	Rome	8-10 May 2000
ECO-202-E	The Revised European System of Accounts (ESA 95) Sector Accounts	NEWSON	Luxembourg	15-17 May 2000
ECO-205-E	The Revised European System of Accounts (ESA 95) Quarterly Accounts	MAZZI	Luxembourg	22-25 May 2000
SOC-106-E	Demographic Data and Their Analysis	ANDERSEN	Copenhagen	22-26 May 2000
EU-SOC-221-E	Summer School on Social Statistics - Cohesion, Integration and Policy Analysis	EVERAERS, GARONNA, TEEKENS, ...	Siena	29 May - 3 June 2000
SUP-104-E	Comparative Analysis of Statistical Packages and Data Bases for Statistics	COLE & CAMPBELL	Manchester	5-7 Jun 2000
PDU-105-E	Marketing and Sales of Statistical Products and Services	MUNCH HAAGENSEN	Copenhagen	19-21 Jun 2000
COM-001-E	The European Statistical System	To be announced	Luxembourg	19-21 Jun 2000
ECO-001-F	National Accounts Statistics in Practice	LEQUILLER	Paris	19-30 Jun 2000
DAT-003-E	Sampling Techniques and Practice	SMITH	Southampton	19-30 Jun 2000
COM-003-E	Confidentiality and Protection of Privacy	NANOPOULOS	Luxembourg	5-7 July 2000

Institut d'Educació Contínua
Universitat Pompeu Fabra

Applied Statistics Week

(6th edition) 2000

Short Courses

Barcelona, from 26 to 30 June 2000

Aims

The sixth APPLIED STATISTICS WEEK is being organized by the Pompeu Fabra University (UPF) from 26 to 30 June 2000, 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) and «Statistics in Marketing Research» (1999). As this is the Millennium year, we have decided to enlarge the scope of the courses to include three important issues of today's society: education, law and politics. This year's theme is «Statistics in Society».

The first course, presented by Harvey Goldstein of the Institute of Education, Universiy of London, deals with statistical information in education in the context of two important problems: the impact of class size on pupil achievement, and the comparison of schools performance. The second course, by David Kaye of Arizona State University's College of Law, aimed at the legal community, explains the use of probability and statistics in the process of law and litigation. The third course, by Nick Moon of the research company NOP in London, is a comprehensive course on public opinion polling in social and political surveys.

After attending these courses, participants will have an insight into the important role played by statistics in assisting with problem-solving and decision-making in these three crucial areas of our society.

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 a general audience interested in the themes to be discussed in education, law and political and social surveys. No previous statistical knowledge will be assumed. The course on education is aimed at educationalists, teachers, policy-makers, and educational researchers. The course on the law is aimed at the legal community in general, lawyers, advocates, judges and law academics. The course on opinion polling is aimed at sociologists, journalists, politicians, market researchers and political analysts.

Place: classes and enrolment

Classes will be held at IDEC (Continuing Education Institute, Pompeu Fabra University) - Balmes, 132. 08008 Barcelona.
Telephone: (34) 93 542 18 00
Fax (34) 93 542 18 08
http://www.upf.es/idec
e-mail: idec@upf.es

Language

All the courses will be taught in English.

Supporting Institution

As in all five previous editions, the sixth APPLIED STATISTICS WEEK is being organized also in cooperation with the **Institut d'Estadística de Catalunya (Idescat).**

Programme

Course 1.

STATISTICAL INFORMATION IN EDUCATION

Harvey Goldstein
Institute of Education, University of London

The course will look at ways in which statistical information is used in education. It will develop ideas using two areas of topical interest: first, research on the effects of class size on attainment; and second, the use of comparative performances of schools on tests and examinations to produce rankings («league tables»). In research on class size we deal with issues of study design, especially how to reach conclusions from both randomised and non-randomised studies. The results of some recent data analyses will be presented and discussed. One of the methodological issues to be addressed is the use of multilevel models for data which have a hierarchical structure, and an introduction to these models will be presented.

In the case of school performance indicators, there are also issues of modelling hierarchically structured data, but especially there are issues related to ways in which such rankings are presented. Because these issues have become politically important in many countries, this raises interesting points about the interface between statistical information and policy making. These questions will be explored using databases on longitudinal achievement from England.

Participants will be encouraged to raise problems from their own experience, which can form the basis of group discussions.

Harvey Goldstein is Professor of Statistics at the Institute of Education, University of London. He is the director of the project on Multilevel Modelling, which conducts research and develops software in this area, especially applied to educational data (see web page: http://www.ioe.ac.uk/multilevel).
He has published a book Multilevel Statistical Modelling, an electronic version of which is available through his personal web page. (http://www.ioe.ac.uk/hgpersonal/index.html).

Date: 26-27 June, from 9.30 to 13.30, and from 15.00 to 18.00 hours.

Course 2.

PROBABILITY AND STATISTICS FOR LAW

David H. Kaye
College of Law, Arizona State University

Modern courts, lawyers, and legislators confront an enormous range of statistical issues. In cases involving criminal prosecutions, product liability, environmental law, antitrust enforcement,

voting rights, and discrimination, for example, crucial evidence now comes from economists, psychologists, social scientists, epidemiologists, geneticists, and other scientists who make use of statistical tools.

This beginning course surveys the concepts of probability theory and statistics as they apply to the proof of facts in courts of law and in the administrative process. It offers a basic introduction to probability, sampling theory, descriptive statistics, regression analysis, and statistical and causal inference. These principles are explained in connection with such topics as the burden of persuasion in criminal and civil cases, the definition of relevant evidence, trademark infringement, discrimination in employment, toxicological and epidemiological proof of causation, parentage testing, and DNA profiling. The emphasis is on ideas rather than computation. No previous knowledge of probability or statistics is required. The course is based on the book Prove It with Figures: Empirical Methods in Law and Litigation (1997), by David Kaye and the late Hans Zeisel, and on the Reference Guide on Statistics, prepared by David Kaye and David Freedman for the U.S. Federal Judicial Center's Reference Manual on Scientific Evidence (2000).

David Kaye is Regents' Professor, Arizona State University College of Law. He teaches and conducts research into the law of evidence, particularly scientific and statistical evidence. His publications include 7 books and 85 articles, reviews, or letters in journals of law, philosophy, medicine, genetics, and statistics. He has taught or delivered invited lectures at many universities, including Cornell, Duke, and Oxford, and he has presented short courses on statistics to judges through programs of the Federal Judicial Center and the National Judicial College of the United States. More details on David's web page: http://www.law.asu.edu/kaye.

Date: 28 June, from 9.30 to 13.30, and from 15.00 to 19.00 hours.

Course 3.

STATISTICS IN POLITICAL AND SOCIAL OPINION POLLING

Nick Moon
NOP Social and Political, London

The course will briefly introduce the history of opinion polling, before moving on to discuss the two main areas of methodological consideration: sampling, and questionnaire design.

The sampling part of the course will discuss the relative merits of random and quota sampling, and the ways in which they can be applied to opinion poll design. Some examples will be given of sample designs in practice. This discussion will cover face-to-face and telephone interviewing, and also self-completion surveys such as postal or Web-based surveys.

The discussion of questionaire design will highlight the dangers of the questions asked influencing the answers given. The course will also provide a guide to assist the lay reader in interpreting polls, including examples of common pitfalls in understanding data. While predominantly about pre-election polling, exit polls will also be covered.

Nick Moon joined NOP as a graduate trainee in 1977, having gained a degree in History from Cambridge, and has worked in social research for practically his whole career. He is now manager of a team of eleven, and although responsible for the overall direction of the Social and Political group, works closely on surveys for much of his time. He is responsible for NOP's considerable

body of work in the field of political opinion polling, and has a very wide range of experience of conducting social research. For more details about NOP consult: http://www.nop.co.uk.

Date: 29 June, from 9.30 to 13.30, and from 15.00 to 18.00, and 30 June from 9.30 to 12.30 hours.

Course fees

The course fees include course notes, a Diploma of Institut d'Educació Contínua, meals and refreshments during the courses.

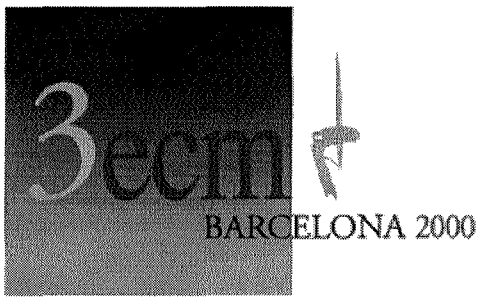
- **Course 1:** 300€ (49.916 ptas. if paid before the 20th of May 2000). (450€; 74.874 ptas.)
- **Course 2:** 180€ (29.949 ptas. if paid before the 20th of May 2000). (260€; 43.260 ptas.)
- **Course 3:** 225€ (37.437 ptas. if paid before the 20th of May 2000). (335€; 55.739 ptas.)

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: 225€ (37.437 ptas.)
Course 2: 135€ (22.462 ptas.)
Course 3: 170€ (28.286 ptas.)

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.



Tercer Congrés Europeu de Matemàtiques

Barcelona, del 10 al 14 de juliol del 2000

Palau de Congressos de Barcelona

El Comitè Organitzador es complau a anunciar que el **Tercer Congrés Europeu de Matemàtiques (3ecm)** tindrà lloc a Barcelona del 10 al 14 de juliol de l'any 2000. L'organitza la Societat Catalana de Matemàtiques (SCM), sota els auspicis de la Societat Matemàtica Europea (EMS).

Programa científic

El programa del congrés inclou nou conferències plenàries, trenta conferències invitades en sessions paral·leles, minisimposis, taules rodones i sessions de pòsters. També s'organitzaran demostracions de programari matemàtic, vídeo i material multimèdia. Els minisimposis són una de les novetats d'aquest congrés; el comitè científic ha escollit una llista de temes molt actuals i amb vinculacions importants fora de les matemàtiques.

Tal com es va fer en els congressos europeus anteriors, s'atorgarà un cert nombre de premis a investigadors/res joves en matemàtiques, de menys de trenta-dos anys d'edat.

Conferències plenàries

- **Robbert Dijkgraaf** (Universitat d'Amsterdam, Holanda)
- **Hans Föllmer** (Universitat Humboldt de Berlín, Alemanya)
- **Hendrik W. Lenstra, Jr.** (Universitat de Califòrnia a Berkeley, Estats Units, i Universitat de Leiden, Holanda)
- **Yuri I. Manin** (Institut Max Planck de Matemàtiques, Bonn, Alemanya)
- **Yves Meyer** (Escola Normal Superior de Cachan, França)
- Carles Simó (Universitat de Barcelona)
- **Marie-France Vignéras** (Universitat de París 7, França)
- **Oleg Viro** (Universitat d'Uppsala, Suècia, i POMI de Sant Petersburg, Rússia)
- **Andrew J. Wiles** (Universitat de Princeton, Estats Units)

Conferències paral·leles

- **Rudolf Ahlswede** (Bielefeld)
- **François Baccelli** (París)
- **Volker Bach** (Mainz)
- **Viviane Baladi** (París)
- **Joaquim Bruna** (Barcelona)
- **Xavier Cabré** (Barcelona)
- **Peter J. Cameron** (Londres)
- **Ciro Ciliberto** (Roma)
- **Zoé Chatzidakis** (París)
- **Gianni Dal Maso** (Trieste)
- **Jan Denef** (Lovaina)
- **Barbara Fantechi** (Udine)
- **Alexander B. Givental** (Berkeley)
- **Alexander Goncharov** (Providence)
- **Alexander Grigor'yan** (Londres)
- **Michael Harris** (París)
- **Kurt Johansson** (Estocolm)
- **Konstantin M. Khanin** (Edimburg, Cambridge i Moscou)
- **Pekka Koskela** (Jyväskylä)
- **Steffen L. Lauritzen** (Aalborg)
- **Gilles Lebeau** (Palaiseau)
- **Nicholas S. Manton** (Cambridge)
- **Ieke Moerdijk** (Utrecht)
- **Eric M. Opdam** (Leiden)
- **Thomas Peternell** (Bayreuth)
- **Alexander Reznikov** (Durham)
- **Henrik Schlichtkrull** (Copenhaguen)
- **Bernhard Schmidt** (Augsburg)
- **Klaus Schmidt** (Viena)
- **Bálint Tóth** (Budapest)

Minisimposis

Computer Algebra. Wolfram Decker (coordinador), Manuel Bronstein, Gaston H. Gonnet, Gert-Martin Greuel, Erich Kaltofen, Hendrik W. Lenstra Jr., Tomás Recio.

Curves over Finite Fields and Codes. Gerard van der Geer (coordinador), Noam D. Elkies, Andrew Kresch, Christian Maire, Henning Stichtenoth, Chaoping Xing.

Free Boundary Problems. José Francisco Rodrigues (coordinador), Giovanni Bellettini, Klaus Deckelnick, Irina V. Denisova, Harald Garcke, Josephus Hulshof, Régis Monneau, Henrik Shahgholian, José Miguel Urbano.

Mathematical Finance: Theory and Practice. Hélyette Geman (coordinadora), Tomas Björk, M.A.H. Dempster, Ernst Eberlein, Jean Jacod, Dilip Madan, Marek Musiela, Stanley R. Pliska, Ton Vorst.

Mathematics in Modern Genetics. Peter Donnelly (coordinador), David Balding, Alison M. Etheridge, Warren J. Ewens, Augustine Kong, Simon Tavaré.

Quantum Chaology. Sir Michael Berry (coordinador), Eugene Bogomolny, Monique Combescure, Alex Eskin, Christopher Howls, Jonathan Keating, Jens Marklof, Zeév Rudnick, André Voros.

Quantum Computing. Sandu Popescu (coordinador), Richard Cleve, Artur Ekert, Rolf Tarrach, Umesh Vazirani.

String Theory and M-Theory. Michael Douglas (coordinador), Duiliu-Emanuel Diaconescu, Jaume Gomis, Chris M. Hull, Albrecht Klemm, J.M.F. Labastida, Marcos Mariño, Nikita Nekrassov, Christoph Schweigert, Angel M. Uranga.

Symplectic and Contact Geometry and Hamiltonian Dynamics. Mikhail B. Sevryuk (coordinador), Paul Biran, Yu.V. Chekanov, Hansjörg Geiges, Viktor L. Ginzburg, Alberto Ibort, Àngel Jorba, Dietmar Salamon, Vladimir M. Zakalyukin.

Wavelet Applications in Signal Processing. Andrew T. Walden (coordinador), Richard G. Baraniuk, Peter Craigmile, Patrick Flandrin, Emma McCoy, Vasily Strela.

Taules rodones

Building Networks of Cooperation in Mathematics. *Moderador:* Friedrich Hirzebruch (Institut Max Planck, Bonn).

How to Increase Public Awareness of Mathematics. *Moderador:* Felipe Mellizo (Radio Nacional de España).

Mathematics Teaching at the Tertiary Level. *Moderador:* Vladimir Tikhomirov (Universitat Estatal de Moscou).

Shaping the 21st Century. *Moderador:* Miguel de Guzmán (Universitat Complutense de Madrid).

The Impact of Mathematical Research on Industry and Viceversa. *Moderador:* Irene Fonseca (Universitat Carnegie Mellon, Pittsburgh).

The Impact of New Technologies on Mathematical Research. *Moderador:* Rafael de la Llave (Universitat d'Austin).

What is Mathematics Today? *Moderador:* Zbigniew Semadeni (Universitat de Varsòvia).

Inscripció i allotjament

Les inscripcions al congrés es poden fer a través de la web **www.iec.es/3ecm**, seguint les indicacions del servidor. També es pot imprimir el formulari d'inscripció, omplir-lo i enviar-lo a *Viajes El Corte Inglés* per fax o per correu a l'adreça següent: Gran Via, 613, 08003 Barcelona, telèfon 93 317 02 02, fax 93 317 58 59. Si no teniu accés a Internet, podeu demanar el full d'inscripció a la Societat Catalana de Matemàtiques. Es pot optar per inscriure's i reservar allotjament al mateix temps, o bé fer-ho separada-ment. Les inscripcions es consideraran vàlides quan es rebi el pagament de la quota.

Quota d'inscripció

	Abans de l'1 d'abril	*Després de l'1 d'abril*
Membres de l'EMS o la SCM	23.000 PTA (138,23 Eu)	33.000 PTA (198,33 Eu)
Altres participants	29.000 PTA (174,29 Eu)	41.000 PTA (246,41 Eu)
Acompanyants	12.000 PTA (72,12 Eu)	18.000 PTA (108,18 Eu)

La quota d'inscripció dels participants inclou un llibre amb el programa del congrés i un CD-Rom amb una versió preliminar de les actes. A més, la inscripció dóna dret als participants i als seus acompanyants a transport públic gratuït durant el congrés i a assistir a tots els actes socials que s'organitzin en el marc del congrés.

Comitès

Comitè científic

Sir Michael F. Atiyah (president), Vladimir Arnold, Robert Azencott, Fabrizio Catanese, Ildefonso Díaz, Antti Kupiainen, Jack van Lint, Colette Moeglin, Johannes Sjöstrand, A.F.M. Smith, Domokos Szász, Stanislaw L. Woronowicz, Don Zagier.

Comitè de premis

Jacques-Louis Lions (president), Noga Alon, Werner Ballmann, David Crighton, Jan Derezinski, Maxim Kontsevich, Eduard Looijenga, Angus Macintyre, José María Montesinos, David Nualart, A.N. Parshin, Ragni Piene, Itamar Procaccia, Mario Pulvirenti, Rolf Rannacher, Caroline Series, Vladimir Sverák, Dan Voiculescu.

Comitè de taules rodones

Miguel de Guzmán (president), Andrey Bolibrukh, Heinz W. Engl, Juan José Manfredi, Carles Perelló, Tomás Recio, Zbigniew Semadeni, Vinicio Villani.

Comitè organitzador

Sebastià Xambó (president), Lluís Alsedà, Jaume Amorós, Carles Broto, María J. Carro, Carles Casacuberta (informació i comunicacions), Teresa Crespo, Julià Cufí (finances), Josep M. Font, Gábor Lugosi, Rosa M. Miró (programació i activitats), Jaume Moncasi, Antoni Montes, Joaquín M. Ortega, August Palanques-Mestre†, Antoni Ras, Jordi Saludes, Marta Sanz (secretaria d'organització), Oriol Serra, Frederic Utzet, Marta València (infraestructura), Joan Verdera, Santiago Zarzuela.

Comitè d'honor

President: S.M. el Rey de España
Molt Honorable President de la Generalitat de Catalunya
Excelentísimo Ministro de Educación y Cultura
Excel·lentíssim Alcalde de Barcelona
Comissionat per a Universitats i Recerca
Director General de la UNESCO
Excel·lentíssim President de l'Institut d'Estudis Catalans
Excel·lentíssim i Magnífic Rector de la Universitat de Barcelona
Excel·lentíssim i Magnífic Rector de la Universitat Autònoma de Barcelona
Excel·lentíssim i Magnífic Rector de la Universitat Politècnica de Catalunya

Patrocinadors

Generalitat de Catalunya, Comissionat per a Universitats i Recerca
Generalitat de Catalunya, Departament d'Ensenyament
Ministeri d'Educació i Cultura, S.E.U.I.D.
Comissió Europea
Fundació Catalana per a la Recerca
Ajuntament de Barcelona
Institut d'Estudis Catalans
Universitat de Barcelona
Universitat Autònoma de Barcelona
Universitat Politècnica de Catalunya
Universitat Pompeu Fabra
Institut d'Estadística de Catalunya
Unió Matemàtica Internacional
Real Sociedad Matemática Española
Sociedad Española de Matemática Aplicada
Fundación Retevisión
Borsa de Barcelona
Port de Barcelona
Fundació Caixa Catalunya
Fundació Banc Sabadell
Fundació Caixa de Sabadell
Fundació Caixa de Manresa
Logic Control
COMSOL AB
Nokia
Compaq
Codorníu
Springer-Verlag

Adreces de contacte

Correu electrònic: **3ecm@iec.es**

Web: **http://www.iec.es/3ecm/** o també **http://www.si.upc.es/3ecm/**

Correu ordinari: Societat Catalana de Matemàtiques
Institut d'Estudis Catalans
Carrer del Carme, 47
08001 Barcelona

Telèfon: +34 93 270 16 20

Fax: +34 93 270 11 80

INTERNATIONAL WORKSHOP ON STATISTICAL MODELLING

Bilbao, Spain: Monday 17 to Friday 21 July, 2000

15th IWSM

New Trends in Statistical Modelling

First Announcement and Call for Papers

The International Workshop on Statistical Modelling concentrates on the various aspects of statistical modelling, including theoretical developments, applications and computational methods. Papers motivated by real practical problems are desirable, but theoretical contributions addressing problems of practical importance or related to software developments are also welcome.

The scientific programme is characterized by having invited lectures & tutorials, contributed papers, posters and software demonstrations. Contributed papers should be suitable for a 20 to 30 minutes oral presentation (including discussion) and focus on motivation, statement of key results and conclusions, and emphasize examples, wherever possible.

Invited speakers:

Christopher Bishop (Cambridge, UK), Joel L. Horowitz (Iowa City, Iowa, USA), Johannes Ledolter (Vienna, Austria), Winfried Stute (Giessen, Germany), Mark Steel (Edinburgh, UK), James Zidek (Vancouver, British Columbia, Canada), Dale L. Zimmerman (Iowa City, Iowa, USA).

A Tutorial on Goodness of Fit Tests for Regression Models will be given by W. González-Manteiga (Santiago de Compostela, Spain).

Students:

Professors should encourage students to attend the workshop. The programme is designed to allow for discussions and interchange between junior and senior scientists. A special session will be devoted for students contributions, an award for the best presentation will be given.

Scientific programme committee:

Ludwig Fahrmeir (Munich, Germany), Eva Ferreira (Bilbao, Spain, Co-chair), John Hinde (Exeter, U.K., Secretary), Michel Mouchart (Louvain-La-Neuve, Belgium), Vicente Núñez-Antón (Bilbao, Spain, Chair), Jean D. Opsomer (Ames, Iowa, U.S.A.), Juan Romo (Madrid, Spain), Esther Ruíz-Ortega (Madrid, Spain), Bill Venables (Australia).

Local organizing committee:

Ma. Victoria Esteban-González, Eva Ferreira, Petr Mariel, Vicente Núñez-Antón, Jesús Orbe-Lizundia, Susan Orbe-Mandaluniz, Marta Regúlez-Castillo, Juan M. Rodríguez-Póo, Gonzalo Rubio-Irigoyen, Fernando Tusell-Palmer.

Further information:

Details about registration for the workshop, instructions for authors and further information is available from the workshop homepage

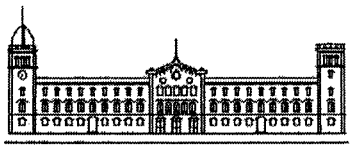
http://iwsm.bs.ehu.es

Deadlines:

Jan 31: Submission of abstracts
Mar 13: Notification of acceptance
Apr 17: Submission of final manuscripts.

For additional information please contact:

Vicente Núñez-Antón
Departamento de Econometría y Estadística
Facultad de Ciencias Económicas y Empresariales
Universidad del País Vasco
Avda. Lehendakari Aguirre, 83
48015 Bilbao, Spain
Phone: +34 94 601 37 49
Fax: +34 94 601 37 54
E-mail: vn@alcib.bs.ehu.es



UNIVERSITAT DE BARCELONA
Departament d'Estadística

UNIVERSIDAD DE ALMERÍA
Departamento de Estadística y
Matemática Aplicada

DISTRIBUTIONS WITH GIVEN MARGINALS AND STATISTICAL MODELLING

To continue the Rome (1990), Seattle (1993) and Prague (1996) Conferences

1	14	14	4
11	7	6	9
8	10	10	5
13	2	3	15

Cryptogram on the west front in Gaudí's Sagrada Familia

July 17-20, 2000
Barcelona (Spain)

Organizing Committee	Scientific Committee
• C.M. Cuadras • J. Fortiana • F. Oliva • J.M. Oller • J.A. Rodríguez-Lallena	• C. Alsina • C.M. Cuadras • J.A. Cuesta • C. Genest • R. Nelsen • I. Olkin • J. Quesada-Molina • B. Schweizer • C. Sempi

For information: http://www.bio.ub.es/estad/personal/cuadras/promar.htm

Informació per als autors i lectors

NORMES PER A LA PRESENTACIÓ D'ARTICLES A *QÜESTIIÓ*

La revista accepta, per a la seva publicació, articles originals no sotmesos a consideració en cap altra revista dins els àmbits de l'Estadística, la Investigació Operativa, l'Estadística Oficial i la Biometria. Els articles poden ser teòrics o aplicats, incloent aspectes computacionals i/o de caire docent, i poden presentar-se en anglès, francès, català o qualsevol altra llengua oficial a l'Estat espanyol.

Tots els originals destinats a les esmentades seccions temàtiques de *Qüestiió*, incloent-hi els articles per a la «Secció docent i problemes», seran sotmesos sistemàticament a una avaluació prèvia a càrrec d'especialistes independents i/o membres del Consell Editorial, llevat dels articles convidats per la revista i les reimpressions d'articles. El resultat de l'avaluació serà comunicat a l'autor principal als efectes d'eventuals correccions formals o dels seus continguts.

Per a totes les trameses d'originals, la revista emetrà un acusament de recepció la data del qual figurarà com a «data de rebuda» en la publicació de l'article. Per la seva banda, la «data d'acceptació» de l'article serà la data de recepció de la versió definitiva.

Per a la presentació d'articles, l'autor trametrà a la Secretaria de *Qüestiió* (Institut d'Estadística de Catalunya) dues còpies del treball mecanografiat en DIN A4, a una sola cara, a doble espai i amb marges amplis. Cada article ha d'incloure el títol, el nom de l'autor o autors, la seva afiliació i l'adreça completa, així com un resum de 75-100 paraules al principi de l'article, seguit de les principals paraules clau (en l'idioma original) i la seva adscripció a la classificació MSC2000 de l'American Mathematical Society. Abans de sotmetre els articles a la revista, s'aconsella els autors que revisin la correcció lingüística de textos d'acord amb l'idioma original i les eventuals traduccions a l'anglès.

Les referències bibliogràfiques es faran indicant el cognom de l'autor seguit de l'any de la publicació entre parèntesi [i.e.: Mahalanobis (1936), Rao (1982b)] i seran llistades alfabèticament al final de l'article; les referències múltiples d'un mateix autor s'ordenaran cronològicament. Les notes explicatives es numeraran correlativament i han d'aparèixer al peu de la pàgina corresponent. Les taules i figures també es numeraran correlativament en el text i seran reproduïdes directament dels originals tramesos en cas que no sigui possible la seva autoedició.

Una vegada avaluat satisfactòriament l'article cal que, a més de la versió impresa, l'autor el trameti en disquet de 3.5 polsades i en format MS-DOS, on han de constar de forma clara els noms dels autors i el títol de l'article. Aquesta versió final s'ha de trametre preferiblement en el processador de textos LaTeX2ε [subsidiàriament, es poden trametre els textos i les taules en Word Perfect —versió 6.0A o anterior— o ASCII]; en el cas de figures, diagrames o gràfics es recomanen els formats adients per als programes-editors PS, EPS o PCX. Els autors han de garantir la correspondència exacta entre la versió impresa i la còpia electrònica. D'altra banda, si l'article no està escrit en llengua anglesa s'haurà d'adjuntar la traducció del títol original, de l'abstract i de les paraules clau, així com un ampli resum en anglès (amb una extensió d'entre 2 i 5 pàgines i amb la mateixa estructura de l'article original).

La Secretaria de *Qüestiió* posa a disposició dels autors que ho sol·licitin plantilles en format LaTeX2ε per a la seva edició i les referències adients de la classificació de l'AMS.

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE SUBMISSION OF ARTICLES FOR *QÜESTIIÓ*

The journal well comes submission of articles and contributions that are not being considered for publication in any other journal in the fields of Statistics, Operational Research, Official Statistics or Biometrics. Articles may be theoretical or applied, including teaching aspects and applications, and will be accepted in English, French, Catalan or any of the other official languages in Spain.

All originals assigned to the thematic sections of Qüestiió, including articles for the «Teaching section and problems» will be systematically reviewed by independent referees and/or members of the Editorial Board, who will send a report to the main author of the article in order to correct, if necessary, any formal or content aspects. The articles invited by the journal and articles reprinted will be excluded from this evaluation process.

For all submissions, the journal will issue a receipt corresponding to the submission date, which will appear as «date received» in the final publication of the article. The «acceptance date» of the article, which will appear in its final publication, will be the date of sending the final version to the journal.

For the presentation of original articles, the author should send, to the Secretary of Qüestiió (Institut d'Estadística de Catalunya), two copies of the paper typed on A4 sheets, one side of the paper only, double spaced and with wide margins. Each article should include the title, the name of the author or authors, their affiliation, full address and also an abstract of the paper (75-100 words) at the beginning of the article, followed by the main keywords (in the original language) and its assignation in the MSC2000 classification of American Mathematical Society. Before submitting their papers, authors are advised to seek assistance in the writing of their articles for the correct use of English and/or of original language.

Bibliographical references should state the author's name followed by the year of publication in brackets [e.g.: Mahalanobis (1936), Rao (1982b)] and they should be listed at the end of the article in alphabetical order; multiple references to the same author should be given in chronological order. Footnotes should be numbered in the article and appear at the foot of the corresponding page. Figures and tables are to be numbered in consecutive order in the text using Arabic numerals and will be directly reproduced from the originals submitted if it is not impossible to print them electronically.

Once the evaluation has been passed, the author is required to provide the article on a diskette (a 3.5-inch disk in MS-DOS format) together with its paper copy; it must be a new diskette and must bear very clearly the names of the authors and the title of the article. This final version should be processed by LaTeX2ε, preferably, or, failing that, by Word Perfect (6.0A or earlier) or ASCII for text and tables; for figures, diagrams or graphs, the appropriate formats of PS, EPS or PCX software tools are strongly recommended. Authors must ensure that the version of the electronic copy is exactly the same as the paper copy which accompanies it. Furthermore, if the article is not written in English, the translation of its original title, short abstract and keywords should be enclosed, as well as a full summary of the article in English (that is, 2-5 pages with the same structure as the original).

The Secretary of Qüestiió can send, by request of the authors, the LaTeX2ε style for manuscript preparation and the appropriate classification references of AMS.

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS INSTITUCIONALS A *QÜESTIIÓ*

Qüestiió convida les entitats patrocinadores, les institucions col·laboradores, els organismes públics i privats, i tota la comunitat científica vinculada a l'estadística o la investigació operativa, a la publicació d'anuncis institucionals sobre cursos, seminaris, congressos i activitats similars que, preferentment, tinguin lloc en el nostre país. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les entitats interessades, de manera que *Qüestiió* no fa una cerca sistemàtica d'esdeveniments d'aquesta naturalesa, ni té cap ànim d'exhaustivitat en les ressenyes d'activitats finalment publicades.

Una vegada aprovada la inclusió dels anuncis sol·licitats es procedirà a la seva publicació, i es reproduirà directament dels originals tramesos amb les mides adequades i la màxima qualitat tipogràfica possible; en aquest cas, *Qüestiió* no procedeix a cap mena de procés d'autoedició de la versió impresa que l'anunciant hagi tramès. Si els originals es trameten en els mateixos termes electrònics exigits per als articles (vegeu «Normes per a la presentació d'articles a *Qüestiió*»), la revista procedirà a la seva autoedició. Si es desitja una qualitat superior a la reproducció simple o l'autoedició, o bé la seva publicació en color, els sol·licitants hauran de posar-se en contacte amb la Secretaria de *Qüestiió* per tal de trametre els fotolits dels textos originals corresponents.

La disposició dels textos i les figures adjuntes dels anuncis han de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de *Qüestiió* es reserva la decisió final pel que fa a la seva publicació.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la inserció en els números/volums de *Qüestiió* que prèviament s'hagi establert. La revista no es fa responsable dels retards, per part de l'anunciant, que impedeixin la publicació de l'anunci en els termes previstos.

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR INSTITUTIONAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió invites all sponsor entities, collaborating institutions, other public and private bodies and the entire scientific community related to Statistics or Operations Research to submit institutional advertisements on courses, seminars, congress and similar activities that will be held, preferably in our country. These will be accepted in English, French, Catalan or any of the other official languages in Spain. The initiative should always come from the entities interested in advertising them so that *Qüestiió*'s aim is not to do a systematic search of these events and therefore does not publish a comprehensive list of such activities.

Once their insertion is approved the advertisements will be reproduced from the most accurate photocopy of the originals sent by the advertiser to *Qüestiió* in paper copy, with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editing process to the printed version that the advertiser has sent. If the original advertisements are sent in the same electronic format requested by the articles (please see «Guidelines for the submission of articles for *Qüestiió*») the journal will print it directly from the file. If a better quality than the simple reproduction or automatic printing or a colour version of the adverts is desired, the authors should contact the Secretary of *Qüestiió* in order to negotiate this.

The typesetting of texts and figures in the advertisement should have maximum intelligibility and clearness, neither compressing the information too much nor using formats or letter fonts that are too small. Furthermore, the information has to be reliable, without errors and respectful of the people and institutions. The management of *Qüestiió* has the right to a final decision concerning the insertion of the advertisement.

Advertisers commit themselves to give the text/materials on request in order to insert them in the issues of *Qüestiió* that have been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published on the agreed terms.

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS PRIVATS O AMB FINALITAT COMERCIAL A *QÜESTIIÓ*

Qüestiió accepta la publicació d'anuncis privats o amb finalitat comercial sobre productes, serveis o altres eines promocionals a l'entorn de l'estadística o la investigació operativa. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les organitzacions que hi estiguin interessades, de manera que *Qüestiió* no fa una cerca sistemàtica de novetats o productes d'aquesta naturalesa ni té cap ànim d'exhaustivitat en els anuncis finalment publicats.

Els anuncis en **blanc i negre** s'elaboren a partir de la fotocòpia més acurada possible dels originals que trameti l'anunciant en versió impresa, amb les mides adequades i la màxima qualitat tipogràfica. Per tant, en aquest cas la revista no efectua cap procés d'edició ulterior respecte de la versió impresa que l'anunciant hagi tramès. Alternativament, si els anuncis originals es trameten en els mateixos termes formals exigits per als articles (vegeu «Normes per a la presentació d'articles a *Qüestiió*»), la revista procedirà a la seva autoedició. Igualment, si es desitja una qualitat superior a la reproducció simple, els sol·licitants hauran de trametre els fotolits dels originals corresponents o encarregar-los a *Qüestiió,* que els facturarà separadament.

Els anuncis **en color** requereixen els fotolits dels textos originals, que poden ser subministrats directament per l'anunciant o bé encarregats per la revista a compte de l'anunciant; en el segon cas, l'anunciant ha de trametre a la revista els originals impresos en color amb la màxima qualitat, per tal de filmar-los amb les millors garanties i condicions. El cost dels fotolits realitzats per *Qüestiió* serà sempre a càrrec de l'anunciant, a qui se li repercutirà l'import i l'IVA d'aquests, juntament amb les tarifes que corresponen a la modalitat d'anunci per la qual hagi optat.

La disposició dels textos i figures adjuntes dels anuncis ha de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de *Qüestiió* es reserva la decisió final de la seva inclusió.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la seva inserció en el(s) número(s)/volum(s) de *Qüestiió* que prèviament s'hagi establert. La revista no es fa responsable dels retards per part de l'anunciant que impedeixin la publicació de l'anunci en els termes previstos.

Imports:

1 pàgina en color (un número aïllat):	125.000 PTA + IVA
1 pàgina en color (tres números consecutius):	200.000 PTA + IVA
1 pàgina en blanc i negre (un número aïllat):	30.000 PTA + IVA
1 pàgina en blanc i negre (tres números consecutius):	50.000 PTA + IVA
1/2 pàgina en blanc i negre (un número aïllat):	20.000 PTA + IVA
1/2 pàgina en blanc i negre (tres números consecutius):	35.000 PTA + IVA

Mides opcionals dels anuncis:

1 pàgina sencera (espai intern):	19.0 cm. × 12.3 cm.
1 pàgina sencera (espai extern):	23.8 cm. × 17.0 cm.
1/2 pàgina (espai intern):	9.5 cm. × 12.3 cm.
1/2 pàgina (espai extern):	11.9 cm. × 17.0 cm.

Forma de Pagament:

- Transferència bancària al compte: 2013-0100-53-0200698577
- Xec bancari nominatiu a l'Institut d'Estadística de Catalunya
- Pagament amb targeta de crèdit

El pagament serà per l'import total de la factura corresponent, on hi figurarà el cost dels fotolits en el cas que l'edició de l'anunci hagi estat a càrrec de l'Institut. En el cas que l'anunciant necessiti una factura proforma, només cal que ho faci saber amb l'antelació suficient.

Correspondència:

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE PRIVATE OR COMMERCIAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió accepts for their publication both private and commercial advertisements on products, services or other promotional tools related to statistics or operational research and will be accepted in English, French, Catalan or any of the official languages in Spain. The initiatives should always come from entities interested in advertising them so that *Qüestiió*'s aim is not to do a systematic search of news and therefore does not publish a comprehensive list of such private or profit activities.

The **black and white** advertisements are made out from the most accurate photocopy of the originals sent by the advertiser to *Qüestiió* in paper copy with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editorial process to the printed version that the advertiser has sent. Alternatively, if the original advertisements are sent in the same formal terms required by the articles (please see «Guidelines for the submission of articles for *Qüestiió*»), the journal will proceed to its autoedition. In the same way, if a better quality than the simple reproduction is wanted, the authors should send the photolits of the corresponding original texts or, on the other hand, order to *Qüestiió* their fulfilment, which will be invoiced separately from the rates charged as advertisements.

The advertisements **in colour** need the photolits of the original texts, which can be provided directly by the advertiser or requested by *Qüestiió* to the advertiser charge; in the second case, the advertiser must sent to the journal the originals printed in colour with the best possible quality, so that they can be filmed at the best conditions and guarantees. The cost of the photolits made by *Qüestiió* will always be charged to the advertiser together with the VAT derived from it, plus the prices corresponding to the type of the advertisement that has been chosen.

The set up of texts and figures of the advertisement should provide the maximum intelligibility and clearness, neither squeezing together the information nor using set ups or letter types that are too small. On the other hand the publicity has to be reliable, without fraud and respectful to the persons and institutions. The direction of *Qüestiió* has the right of the last decision concerning the insertion of the advertisement.

The advertiser committs himself to give the texts/materials on request, in order to insert them in the issue(s) of *Qüestiió* that had been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published in the agreed terms.

Rates:

1 colour page (only one issue):	125.000 PTA + VAT
1 colour page (three consecutive issues):	200.000 PTA + VAT
1 black and white page (only one issue):	30.000 PTA + VAT
1 black and white page (three consecutive issues):	50.000 PTA + VAT
1/2 black and white page (only one issue):	20.000 PTA + VAT
1/2 black and white page (three consecutive issues):	35.000 PTA + VAT

Advertisement sizes (optional):

1 full page (internal space):	19.0 cm. × 12.3 cm.
1 full page (external space):	23.8 cm. × 17.0 cm.
1/2 page (internal space):	9.5 cm. × 12.3 cm.
1/2 page (external space):	11.9 cm. × 17.0 cm.

Payment:

- A bank transfer to account number: 2013-0100-53-0200698577
- A bank cheque to Institut d'Estadística de Catalunya
- Charge on a credit card

The payment should be for the amount shown at the invoice, where it will be shouwn the total cost of the photolits, in case that *Qüestiió* would be in charge of the filmation of the advertisement. If advertiser need a pro-forma invoice, he should let us know some time in advance so that *Qüestiió* could send it to the proper address.

Mail address:

Mònica M. Jaime
Secretaria de *Qüestiió*
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

Butlleta de subscripció a la revista *Qüestiió*

Nom i cognoms _____

_____

Empresa/Institució _____

_____

Adreça _____

_____

Codi postal _____ Ciutat _____

Tel. _____ Fax _____ NIF_____

Data _____

Signatura

Desitjo subscriure'm a ***Qüestiió*** per a l'any 2000

Preu de subscripció vigent:

– Estat espanyol: 3.600 Pta/any (21,64 €) (IVA inclòs)
– Estranger: 4.000 Pta/any (24,04 €) (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: 1.500 Pta/exemplar (9,02 €) (IVA inclòs)
– Estranger: 1.700 Pta/exemplar (10,22 €) (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-abril 2000

- **Anuari Estadístic de Catalunya, 1992-1999 CD-ROM**
  1999, 3.000 PTA (18,03 €)
  ISBN: 84-393-4914-9
- **Xifres de Catalunya 1999**
  versions en català, castellà, francès, anglès i alemany. Gratuït
- **Revista Qüestiió Any 1999 Vol. 23 núm. 3**
  desembre 99, 3.000 PTA (18,03 €),
  ISSN 0210-8054
- **Estadística, producció i comptes de la indústria 1998**
  gener 2000, 1.300 PTA (7,81 €), 371 pp.,
  ISBN: 84-393-5038-4
- **Projeccions de població de Catalunya 2010. Comarques i àmbits del Pla territorial**
  gener 2000, 1.600 PTA (9,62 €), 395 pp,
  ISBN: 84-393-5009-0
- **Comptes de les administracions públiques de Catalunya 1996**
  maig 2000, 1.200 PTA (7,21 €),138 pp.,
  ISBN: 84-393-5075-9
- **Estadística de població 1996 Vol. 8 Relació de la població amb l'activitat econòmica. Dades comarcals i municipals**
  maig 2000, 1.250 PTA (7,51 €),192 pp.,
  ISBN: 84-393-5049-X

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

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