Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 2002, volum 26, núm. 3
Segona època

Entitats patrocinadores:

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

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 2002, volum 26, núm. 3

SUMARI

EDITORIAL

La tercera època de *Qüestiió*: Statistics & Operations Research Transactions (SORT)

L'evolució de les revistes científiques reflecteix l'entorn científic, institucional i polític, algunes vegades estable i estimulant, i d'altres agitat i fins i tot opressiu, en el qual es desenvolupen. Algunes publicacions clàssiques de les grans acadèmies sembla que no han canviat al llarg d'uns segles però, si les examinem amb atenció, hi podrem veure un progrés intens dels plantejaments científics. En la història del nostre país, els canvis de les edicions acadèmiques són molt més ràpids i intensos que en d'altres, degut als seus esforços per consolidar-se, adaptar-se i competir.

Qüestiió és una continuació i potenciació dels «Cuadernos de Estadística Aplicada e Investigación Operativa», editats en un format humil en el període franquista que continuava la pobríssima i llarguíssima postguerra civil. El doctor Torrens Ibern havia portat del seu exil·li a França les noves tècniques del control estadístic de la qualitat i d'investigació operativa i les ensenyava a l'Escola d'Enginyers Industrials del carrer Urgell. Aleshores l'ensenyament universitari d'estadística es limitava als cursos de probabilitat i estadística matemàtica que impartia primer el doctor Ors i després el doctor Sales a la Secció de Matemàtiques de la Facultat de Ciències, i als cursos dedicats als actuaris d'assegurances.

Recollint la tradició del doctor Torrens Ibern, *Qüestiió* es va crear el 1977 al si de la Universitat Politècnica de Catalunya, i va comptar des del començament amb la col·laboració de professors d'altres universitats, especialment de l'àmbit de la Biologia. D'aquesta manera s'iniciava una etapa que reflectia l'extens creixement de les universitats catalanes, la introducció dels ensenyaments d'estadística en un gran nombre de titulacions i la intensificació de la recerca científica.

El reconeixement de la Generalitat de Catalunya, que va significar la recuperació d'alguns drets nacionals i de la seva institució històrica de govern, i el referèndum de l'Estatut d'Autonomia del desembre de 1979, van portar a una certa normalització del país. En aquest marc, a l'any 1989 es va crear l'Institut d'Estadística de Catalunya (Idescat) i el 1992 *Qüestiió* va iniciar una segona època amb l'Idescat com a editor i la fructífera col·laboració d'algunes de les universitats catalanes. Aquesta nova etapa (segona època) ha permès potenciar considerablement la revista, aprofundint en l'estadística aplicada, l'estadística oficial i la biometria, a més de la investigació operativa.

Després de la celebració dels seus vint-i-cinc anys, *Qüestiió* ha decidit emprendre una altra etapa, a fi d'aconseguir un major reconeixement internacional i assolir un nivell científic més alt. Aquest canvi es fonamenta en el desenvolupament natural que ha tingut la revista al llarg de l'etapa que ara tanquem. Per aquesta raó, el número present, que

és la culminació d'una llarga trajectòria, inclou un estudi detallat dels grups de recerca estadística d'universitats i institucions de Catalunya. El nombre i la qualitat d'aquests grups és alhora un testimoni del progrés científic assolit i una garantia dels projectes que estem iniciant.

A partir del número vinent, la revista tindrà el nom de «SORT» (Statistics and Operations Research Transactions), el qual vol fer una al·lusió a la seva història, i mantindrà la numeració amb el volum 27 (2003). SORT presentarà tots els articles en anglès (per bé que s'acompanyaran d'un breu resum dels mateixos en català) i preten satisfer les condicions que actualment la comunitat científica internacional exigeix per a la seva inclusió en els índexs d'impacte. D'aquesta manera, la tercera època de *Qüestiió* vol seguir sent un instrument cada cop més eficaç al servei dels investigadors.

Voldria aprofitar aquesta ocasió per felicitar a totes les persones que, d'una manera o d'una altra, han contribuït al desenvolupament de *Qüestiió* i pregar-los que, amb el mateix rigor i entusiasme, ajudin a assolir els objectius que es proposa SORT.

Eduard Bonet, director de *Qüestiió*

Comentari de la publicació

El tercer número del volum 26, corresponent a l'any 2002, recull l'edició d'un total de vuit articles, repartits en dues de les quatre seccions temàtiques de la revista i la secció reservada als treballs originals de caràcter docent, acompanyats de tres recensions a «Comentari de llibres» i de la ressenya de novetats editorials en matèria estadística. Com ja s'ha comentat, el número també inclou una panoràmica de la tasca investigadora a Catalunya que, sota el títol de *Ressenya sobre la recerca actual en estadística a Catalunya*, il·lustra els resultats assolits per 17 grups d'investigació catalans a l'entorn d'un nombre equivalent de línies de recerca en estadística. Al mateix temps, el present número publica la relació dels avaluadors i col·laboradors de *Qüestiió* que han participat en la revista en el període 1998-2002, completant d'aquesta manera el reconeixement dels autèntics promotors de la revista que ja es va fer l'any 1997 amb motiu dels vint anys de *Qüestiió*.

«Estadística» i «Investigació Operativa»

En aquest darrer número del volum 26 (2002) s'hi publiquen vuit articles, sis d'Estadística i dos d'Investigació Operativa, així com un article a la secció docent. Dins la secció «Estadística», el primer article titulat *Implementación del cálculo de polinomios zonales y aplicaciones en análisis multivariante*, de J. Rodríguez-Avi, A. J. Sáez-Castillo i A. Conde-Sánchez, conté un algorisme de càlcul de funcions especials lligades a les series hipergeomètriques d'argument matricial, que s'aplica a la distribució de valors propis d'una matriu Wishart. A continuació, l'article *Analysis on the individual efficiency prediction in the composed error frontier model. A Montecarlo study*, de R. Dios Palomares, A. Ramos-Millán i J. A. Roldán-Casas, estudia una funció de regressió separant el terme d'error en dues components, una normal i l'altra exponencial, half-normal, normal truncada o gamma. Les propietats del model en l'estimació de l'eficiència són analitzades i comprovades mitjançant simulació, arribant a conclusions interessants. En el tercer article, *Aplicación de redes neuronales artificiales a previsión de series temporales no estacionarias o no invertibles*, R. Pino, D. de la Fuente, J. Parreño i P. Priore calculen previsions de sèries temporals difícils de tractar aplicant models ARIMA segons la metodologia Box-Jenkins, mostrant el potencial de les xarxes neuronals artificials, especialment en el cas de sèries multivariants. El quart article, *Anàlisi de matrius quadrades no simètriques: un enfocament integral usant anàlisi de correspondències*, de J. Daunis, T. Aluja i S. Thió, aplica l'anàlisi de correspondències per estudiar matrius no simètriques motivats per l'estudi migratori de les comarques de Catalunya. Els autors proposen un nou mètode basat en reemplaçar les dades de la

diagonal per reconstitució i en descomposar la matriu en una part simètrica i una antisimètrica, però tenint en compte la diagonal de la taula i els moviments migratoris, aconseguint una visió més clarificadora del tractament d'aquestes matrius. Per últim, l'article *La distància de l'Eixample*, de M. Greenacre, defineix i estudia la distància entre dos punts de l'Eixample de Barcelona. És una mètrica no necessariàment euclidiana, més llarga que la Pitagoriana i que pot ser més curta o més llarga que la distància de Manhattan. L'autor proposa algunes aplicacions i una generalització multivariant.

Pel que fa a la secció «Investigació Operativa», el primer article inclòs es titula *Un algoritmo para el problema del biflujo máximo simétrico no dirigido*, d'A. Sedeño-Noda i C. González-Martín, on s'hi tracta el problema descrit en el mateix títol a partir de la introducció d'un canvi de variables que permet dividir el problema original en dos més simples de flux màxim, possibilitant així aplicar eines més conegudes. Cal ressenyar, a més, que els autors també calculen la capacitat computacional teòrica de l'algorisme. Seguidament, C. R. Delgado, S. Casado i J. F. Alegre descriuen, a *Problema de contratación de carretilleros para un almacén de productos manufacturados*, un problema de transport molt concret d'emmagatzement fins la distribució de productes manufacturats, coneixent la freqüència de sol·licitud dels clients. El problema del calendari de lliurament i l'assignació diària es resol mitjançant tres algorismes de cerca que els autors contrasten i comparen.

Altres seccions i apartats

La «Secció docent i problemes» publica l'article *Estadística, societat i veritat*, de C. M. Cuadras, en el qual es presenten les dades estadístiques des d'una perspectiva històrica i actual, comentant-se les diferències entre determinisme, indeterminisme i caos, així com el paper de l'estadística en l'estudi quantitatiu de l'atzar. Els descobriments, les curiositats, les paradoxes i la regularitat estadística són també comentades i il·lustrades, incloent-hi els comentaris de l'autor sobre el què hi ha de veritat en l'estadística i com aquesta disciplina pot ser útil a la societat. D'altra banda, aquesta secció inclou la presentació de nous enunciats, juntament amb la resolució dels problemes publicats en el número immediatament anterior. Atès que es tracta del darrer número de la segona època de *Qüestiió*, a partir de la qual deixaran de publicar-se enunciats i solucions de problemes en la versió impresa de la revista, les solucions dels problemes publicats en aquest número només es donaran a conèixer en la versió electrònica de la tercera etapa de la revista.

Seguidament, la secció «Comentari de llibres» acull tres ressenyes detallades de novetats editorials de Springer Verlag, a càrrec d'E. Bonet, C. Serrat i C. M. Cuadras, respectivament. En el primer cas, se'n destaca l'oportunitat i efectivitat de l'obra *Análi-*

sis estadístico de los datos (L. Lebart, A. Salem i M. Bécue) , sobretot pel que fa a la difusió de l'anàlisi multivariant aplicada a les «dades textuals», en la que els autors ofereixen una guia bàsica de conceptes, eines, procediments i aplicacions d'aquesta nova nova disciplina estadística. A continuació, el llibre *Statistical Consulting*, de J. Cabrera i A. McDougall, mereix una especial recomanació en l'extensa ressenya que fa C. Serrat, en la que valora sobretot l'aproximació via «cas studies» de l'entramat de regles i experiències que ha assolit l'estadística en els processos de consultoria. Per últim, C. M. Cuadras ressalta la quantitat i qualitat de continguts en *Análisis de datos multivariantes* (D. Peña), on el rigor i claredat expositiva es complementen amb l'accessibilitat a les aplicacions pràctiques que ofereix el llibre.

El darrer apartat, dedicat a «Ressenyes d'activitats institucionals», inclou, com ja és costum, una revisió actualitzada d'activitats de la Sociedad Española de Biometría i una recensió del «Training for European Statisticians Institute» que *Qüestiió* ha redifòs des del 1997, amb la relació dels cursos del *Core Programme 2003* que s'impartiran del gener al novembre del 2003, adreçats principalment als membres dels òrgans d'estadística oficial en l'àmbit comunitari. En tercer lloc, s'ofereix un anunci sobre el *27 Congreso Nacional de Estadística e Investigación Operativa* (Lleida, 8-11 abril 2003) que organitza la SEIO amb la Universitat de Lleida i la col·laboració, entre d'altres, de l'INE, l'Idescat i altres patrocinadors de *Qüestiió*. A continuació, es fa una breu presentació de la *International Conference on Correspondance Analysis and Related Methods* (CARME 2003, Barcelona, 29 juny-2 juliol 2003) que organitzarà la Universitat Pompeu Fabra, juntament amb altres entitats col·laboradores de la nostra revista. Per últim, es fa ressó del *Compositional Data Analysis Workshop* (CoDaWork'03, Girona, 15-17 octubre 2003) que la Universitat de Girona, amb altres grups de recerca catalans i internacionals, portarà a terme per primera vegada. El número conclou, com és habitual, amb les novetats editorials de la Generalitat de Catalunya que han estat publicades l'any 2002 en l'àmbit de l'estadística.

Per últim, com ja és habitual en el darrer número de cada volum, a continuació es presenta l'evolució d'articles en el decurs de l'interval gener-desembre del 2002, el qual ha enregistrat el moviment següent:

articles sotmesos: 11
articles en procés d'avaluació: 16
articles acceptats: 17 (tots ells publicats)
articles rebutjats: 13

També val la pena remarcar que, amb les dades de l'any 2002, el temps d'acceptació d'articles en els darrers anys se situa en una mitjana de 9,6 mesos, mentre que el percentatge de rebuig d'articles sotmesos a la revista augmenta fins a situar-se a l'entorn del 35 %.

Carles Cuadras, director executiu
Enric Ripoll, editor executiu

Ressenya sobre la recerca actual en estadística a Catalunya

PRESENTATION

The aim of this research report is to present the activities of the more representative groups in Catalonia doing research in statistics. Sixteen research groups (in alphabetical order of the coordinating author) presented a description of their research activity and a list of their publications.

The topics are quite varied, covering a wide spectrum in statistics: stochastic processes, data analysis, survival analysis, multivariate analysis, compositional data, structural equation modelling, statistical mathematics and finance, official statistics, statistics in behavioral research, survival and computation, structural models with latent variables, mathematics and statistics in Biosciences and regional quantitative analysis.

Most research groups show a high quality, publishing papers in very prestigious journals. In total, 16 groups published about 450 papers (including chapters of books). There are 32 authors who have five or more papers, and their distribution across numbers of papers is as follows:

Number of papers	*Frequency of authors*
> 30	1
21-30	3
11-20	15
5-10	13

About 25 authors published less than 5 papers.

DATA ANALYSIS AND DATA MINING GROUP

Tomas Aluja
Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

Michael Greenacre
Departament d'Estadística
Universitat Pompeu Fabra

For some time there has been a strong cooperation with the French group of «Analyse des Données», inspired by the seminal work of J.-P. Benzécri. This cooperation has led to continuous research exchanges centred in the Universitat Politècnia de Catalunya, Universitat Pompeu Fabra and Universitat de Girona, among others.

Data Analysis in this context refers to the conception of statistics based on data, without any probabilistic assumptions, building models that fit the data, instead of the classical approach of using the data to validate hypothesized models within a probabilistic data framework. When applied to large data sets, this approach is often referred to as Data Mining (Aluja, 2001).

The research has focused on methodological innovation as well as software implementations, some of them commercial, of the proposed methodologies.

Three international conferences have organized on this theme, jointly with the Central Archive for Empirical Social Research of the University of Cologne in Germany: in 1991, Correspondence Analysis in the Social Sciences; in 1995, Visualization of Categorical Data; and in 1999, Large Scale Data Analysis. Two of these conferences led to edited books (Greenacre and Blasius, 1994; Blasius and Greenacre, 1998). The next conference in this series is Correspondence Analysis and Related Methods, taking place in 2003 in Barcelona.

The following are topics of research by this group.

Detection of functional regions: This is a classical problem in regional statistics, a functional region being defined as a pole of attraction with its hinterland. It was applied to the case of the «comarcal» (county) delimitation of Catalonia. A new logical distance was used to measure the attraction between municipalities (Aluja, 1983).

Local principal components analysis: Very often, principal component analysis reveals a previously known global variability. In such cases it is interesting to go beyond the classical analysis, eliminating the known variability in the data. This has been proposed under different names and approaches. Following previous work by Ludovic Lebart, Aluja *et al.* (1985) and Aluja (1988) developed a non-parametric approach using a non-oriented graph on the individuals. This is a very flexible tool allowing to partial out

the variability «modelled» by the edges of the graph. It is possible to obtain a triplot of the local variables, the edges of the graph and the individuals after having eliminated the unwanted effect. This methodology was extended to the analysis of panel data by Aluja *et al.* (1993). Some equivalences between the different approaches are presented by Nonell, Thió and Aluja (2000).

Decision trees: The objective of decision trees methods is to automatically detect which variables serve to explain the behaviour of a given response variable, either quantitative or categorical. Our research focused on the problem of the stability of tree. The tree-growing process is highly dependent on the data, especially in actual data mining applications which have a large number of variables, where the construction of decision trees is somewhat arbitrary since we have to select between several splits possessing similar explanatory powers. Tackling this problem, Aluja and Nafria (1998*a*) proposed a new general geometric formulation of the impurity of a node, which allows us to compute the contribution of each individual to the impurity, and hence to detect influential individuals, and also defined an alternative criterion for the classification of trees based on a generalized Smirnov distance. In addition, Aluja and Nafria (1998*c*) studied the complexity of the CART methodology, proposing an algorithm with linear cost with the number of individuals. A complementary research line was to incorporate expert knowledge into the clustering process (Gibert, Aluja and Cortes, 1998).

Data fusion and grafting: Data fusion refers to the problem of merging information coming from independent sources. It is also known as statistical matching. The fusion consists of the transferring the specific y variables from the donor file to the receptor. There are two main methodologies for data fusion, one is modelling the relation of the y variables respect to the common x variables in the donor file and apply it to the receptor file. Thye other approach is the so-called «hot deck» methodology, which involves finding for each individual of the recipient file one or more similar individuals in the donor file and then transferring the y values of these individuals to the receptor. Our research is to assess the viability of the fusion process, to compare the accuracy of the different methodologies proposed, and to propose validation measures of the process (Aluja *et al.*, 1998*b*, Aluja and Thió, 2001) as well as implementing this in a complete system for data fusion. File grafting is a related methodology complementing the previous research line, developed to visually display information coming from independent sources in a common factorial subspace (Aluja, Morineau and Rius, 1999; Rius, Aluja and Nonell, 1999). Grafting is implemented within the SPAD software.

Textual data analysis: The usual techniques of multivariate data analysis can be applied to textual data, which form a homogeneous corpus full of redundancy. Correspondence analysis can be used to associate the lexical information with any external information about texts. The research has focused on the problem of definition of complex lexical units (repeated segments) to take into account the context in the analysis, the application of such methodology to massive corpus of textual data and the simul-

taneous analysis of different open questions (Bécue, 1993, 1997*a*, Bécue and Lebart, 1998, 2000) and the analysis of questions in different languages (Bécue, 1997*b*). A new methodology, Multiple Factorial Analysis, has been proposed to analyze three-way contingency tables, either textual or numerical (Bécue, 1998, Bécue and Pagès, 2001). The proposed methodologies have been implemented in the software system SPAD-textual.

Multivariate methods used in environmental research: This research line concerns the study of multivariate methods that are used in environmental studies. Important multivariate methods in the environmental sciences are principal component analysis (PCA), factor analysis, correspondence analysis (CA) and methods using linear constraints such as canonical correspondence analysis (CCA) and redundancy analysis (RDA). CCA and another approach based on multidimensional scaling are compared by Greenacre and Fieler (1995). Interesting applications of these methods have been obtained with a large database of marine biological samples. Quality statistics for the goodness of fit of all data matrices involved in CCA have been derived (Graffelman 2001*a*). Research has focused on the representation of supplementary information (cases and variables) in biplots obtained by these methods, with attention to the different types of scaling and the geometrical properties of the solution. Some theoretical results are given by Graffelman (1999, 2000, 2001) and Graffelman and van der Velden (1999). A different line of research undertaken involves various biometric analyses of the human sex ratio (Graffelman *et al.*, 1999), Graffelman and Hoekstra, 2000).

Correspondence analysis: Research here has focused on ways of interpreting multiple correspondence analysis and on a different algorithm for fitting multiway contingency tables called «joint correspondence analysis» (Greenacre, 1993*b*, 1994*a*, 1994*b*, 1998). In joint correspondence analysis the diagonal subtables on the super-diagonal of the so-called Burt matrix are not fitted by least-squares, only the extra-diagonal tables. This leads to correct measures of explained inertia in correspondence analysis. As a by-product, the usual solution in multiple correspondence analysis can be improved by a simple calculation which provides a lower bound, and usually a close lower bound, to the correct percentages of inertia (see, for example, Greenacre, 1993*a*).

Biplots and unfolding: Biplots arising from correspondence analysis and multiple correspondence analysis are also studied (Greenacre, 1993*c*) and a comparison with biplots for compositional data is given by Aitchison and Greenacre (2002). Gower and Greenacre (1996) showed how unfolding can be applied to a square symmetric matrix of distances, leading to a novel way of interpreting a multidimensional scaling display. Visualization of preference data in a marketing context is given by Torres and Greenacre (2002).

Analysis of square asymmetric tables and matched matrices: In this line of research the joint visualization of two or more tables is studied. The analysis of a square asymmetric table is studied, the two tables being the original table and its transpose (Gree-

nacre, 2000). In the case of two rectangular tables, matched by rows and columns, their joint visualization can be achieved by the singular value decomposition of a block matrix where each matrix appears twice (Greenacre, 2001). The case of two square asymmetric matrices, usually transition matrices, is given by Greenacre and Clavel (2002).

References

1. Aitchison, J. and Greenacre, M. J. (2002). «Biplots of Compositional Data». *Applied Statistics*, 51, 375-392.
2. Aluja, T. (1983). «Determinació dels centres d'atracció i llur zona d'influència en funció dels equipaments municipals».In *El Debat de la Divisió Territorial de Catalunya. Edició d'estudis propostes i documents (1939-1983)*, Oriol Nel·lo and Enric Lluch (eds.), Altafulla, Barcelona, 839-863.
3. Aluja, T. and Lebart, L. (1985). «Factorial analysis upon a graph». *Demandez le Programme*, 3, 3-34, CISIA, París.
4. Aluja, T. (1988). «Local and partial correspondence analysis. Application to the analysis of electoral data». *Computational Statistics Quarterly*, 4, 89-103.
5. Aluja, T. and Nonell, R. (1991). «Local principal components analysis». *Qüestiió*, 15, 267-278.
6. Aluja, T. and Nonell, R. (1993). «Multiple correspondence analysis on panel data». In *Multivariate Analysis: Future Directions 2*, C. M. Cuadras and C. R. Rao (eds.), North-Holland, Amsterdam, 233-244.
7. Aluja, T. (1994). «Formation of an index of economic capacity in Barcelona». *Qüestiió*, 18, 377-384.
8. Aluja, T. and Nafria, E. (1998*a*). «Generalized impurity measures and data diagnostics in decision trees». In *Visualisation of Categorical Data*, Jörg Blasius and Michael Greenacre (eds.), Academic Press, San Diego, 59-70.
9. Aluja, T., Nonell, R., Rius, R. and Martinez-Abarca, M. J. (1998*b*). «Data fusion and file grafting». *Analyses Multidimensionelles des Données*, 7-14.
10. Aluja, T. and Nafria, E. (1998*c*). «Robust impurity measures in decision trees». In *Data Science, Classification and Related Methods*, Springer Verlag, Heidelberg, 207-214.
11. Aluja, T., Morineau, A. and Rius, R. (1999). «La greffe de fichiers et ses conditions d'application. Méthode et exemple». In *Enquêtes et Sondages. Méthodes, Modèles, Applications, Nouvelles Approches*. Dunod, París, 94-102.
12. Aluja, T. and Thió, S. (2001). «Survey data fusion». *Bulletin de Méthodologie Sociologique*, 72, 20-36.
13. Aluja, T. (2001). «La minería de datos, entre la estadística y la inteligencia artificial». *Qüestiió*, 25, 479-498.
14. Álvarez, R., Bécue, M. and Lanero, J. J. (2000). «Le vocabulaire gouvernemental espagnol (1979-1996)». *Mots*, 62, 31-47.

15. Bécue, M. (1993). «Utilisation des logiciels de statistique: la question ouverte de l'enquête». *Modulad*, 11, 50-58.
16. Bécue, M. (1997*a*). «Visualization of open questions: a French study of pupils' attitudes to Mathematics».In *Visualization of Categorical Data*, Jörg Blasius and Michael Greenacre (eds.), Academic Press, San Diego, 151-158.
17. Bécue, M. (1997*b*). «Etude comparative de réponses ouvertes à différentes questions». In *Analyses Multidimensionnelles des Données*, Fernández Aguirre, K. and Morineau, A. (eds.), Cisia, París, 65-72.
18. Bécue, M. (1998). «Three-way textual data analysis». In *Advances in Data Science and Classification, Springer Series of Studies in Classification, Data Analysis and Knowledge Organization*, Rizzi, A., Vichi, M. and Bock, H. H. (eds.), Springer Verlag, Heidelberg, 457-464.
19. Bécue, M. and Lebart L. (1998). «Clustering of texts using semantic graphs. Application to open-ended questions». In *Surveys. Data Science, Classification and Related Methods, Springer Series of Studies in Classification, Data Analysis and Knowledge Organization*, Hayashi, C., Oshumi, N., Yajima, K., Tanaba, Y., Bock, H. H. and Baba, Y. (eds.), Springer Verlag, Tokyo, 480-487.
20. Bécue, M. and Lebart, L. (2000). «Analyse statistique des réponses ouvertes. Application à des enquêtes auprès de lycéens». In *Analyse des Correspondances et Techniques Connexes*, J. Moreau, P. A. Doudin and P. Cazes (eds.), 59-83.
21. Bécue. M. and Pagès, J. (2001). «Analyse simultanée de questions ouvertes et de questions fermées. Méthodologie, exemple». *Journal de la Société Statistique de France*, 142-4.
22. Blasius, J. and Greenacre, M. J. (1994). «Computation of correspondence analysis». In Greenacre, M. J. and Blasius, J. (eds.), *Correspondence Analysis in the Social Sciences*, Academic Press, London, 53-78.
23. Blasius, J. and Greenacre, M. J. (1998). *Visualization of Categorical Data*, Academic Press, New York.
24. Gibert, K., Aluja, T. and Cortés, U. (1998). «Knowledge discovery with clustering based on rules. Interpreting results». In *Principles of Data Mining and Knowledge Discovery*, Springer Verlag, Berlin, 83-93.
25. Gower, J. C. and Greenacre, M. J. (1996). «Unfolding a symmetric matrix». *Journal of Classification*, 13, 81-105.
26. Graffelman, J. (1998). «Solution 97.5.3: A fundamental matrix result on scaling in multivariate analysis. Solution 97.5.3». *Econometric Theory*, 14, 693-694.
27. Graffelman, J. (1999). «Solution 99.1.5: The justification of multidimensional scaling under Euclidean conditions. Solution». *Econometric Theory*, 15, 908-909.
28. Graffelman, J. (2000). «Solution 99.5.1: Use of the Moore-Penrose Inverse in Canonical Correspondence Analysis». *Econometric Theory*, 16, 792-793.
29. Graffelman, J. and van de Velden, M. (1999). «Problem 99.4.4: Upper bounds for the eigenvalues of the product of a symmetric idempotent and a non-negative definite matrix». *Econometric Theory*, 15, 631.

30. Graffelman, J., Fugger, E. F., Keyvanfar, K. and Schulman, J. D. (1999). «Human live birth and sperm sex ratio compared». *Human Reproduction*, 14, 2917-2920.
31. Graffelman, J. and Hoekstra, R. F. (2000). «A statistical analysis of the effect of warfare on the human secondary sex ratio». *Human Biology*, 72, 433-445.
32. Graffelman, J. (2001*a*). «Quality statistics in canonical correspondence analysis». *Environmetrics*, 12, 485-497.
33. Graffelman, J. (2001*b*). «Factor Analysis». In El-Shaarawi, A. H. and Piegorsch, W. W. (eds), *Encyclopedia of Environmetrics. Volume 2*, John Wiley, Chichester, 763-767.
34. Greenacre, M. J. (1993*a*). *Correspondence Analysis in Practice*. Academic Press, London.
35. Greenacre, M. J. (1993*b*). «Different geometric approaches to correspondence analysis of multivariate data». In Opitz, O., Lausen, B. and Klar, R. (eds.), *Information and Classification*, Springer-Verlag, Berlin, 190-200.
36. Greenacre, M. J. (1993*c*). «Biplots in correspondence analysis». *Journal of Applied Statistics*, 20, 251-269.
37. Greenacre, M. J. (1994*a*). «Multiple and joint correspondence analysis». In Greenacre, M. J. and Blasius, J. (eds.), *Correspondence Analysis in the Social Sciences*, Academic Press, London, 141-161.
38. Greenacre, M. J. (1994*b*). «Correspondence analysis and its interpretation». In Greenacre, M. J. and Blasius, J. (eds.), *Correspondence Analysis in the Social Sciences*, Academic Press, London, 3-22.
39. Greenacre, M. J. (1995). «Multivariate generalisations of correspondence analysis». In Cuadras, C. M. and Rao, C. R. (eds.), *Multivariate Analysis: Future Directions 2*, North Holland, Amsterdam, 327-340.
40. Greenacre, M. J. (1998). «Diagnostics for joint displays in correspondence analysis». In Blasius, J. and Greenacre, M. J. (eds.), *Visualization of Categorical Data*, Academic Press, New York, 221-238.
41. Greenacre, M. J. (2000). «Correspondence Analysis of Square Asymmetric Matrices». *Applied Statistics*, 49, 297-310.
42. Greenacre, M. J. (2001). «Analysis of matched matrices». Working Report 539, Department of Economics and Business, Universitat Pompeu Fabra, submitted for publication.
43. Greenacre, M. J. and Blasius, J. (1994). *Correspondence Analysis in the Social Sciences*, Academic Press, London, 53-78.
44. Greenacre, M. J. and Clavel, J. G. (2002). «Simultaneous visualization of two transition tables». *Polish Journal of Statistics, Special Issue: Statistics in Transition*, in press.
45. Lebart, L., Morineau, A., Pleuvret, P. and Aluja, T. (1983). «Manuel SPAD. Système Portable pour l'Analyse des Données». CISIA, París.
46. Lebart, L., Morineau, A., Bécue, M. and Häusler, L. (1993). «Système pour l'Analyse de Données Textuelles, Manuel de l'Utilisateur». CISIA, París.

47. Nonell, R., Thió, S. and Aluja, T. (2000). «Some alternatives for conditional principal component analysis». *Applied Stochastic Models in Business and Industry*, 16, 189-200.
48. Rius, R., Aluja, T. and Nonell, R. (1999). «File grafting in market research». *Applied Stochastic Models in Business and Industry*, 15, 451-460.
49. Sieber, T. N., Petrini, O. and Greenacre, M. J. (1998). «Correspondence analysis as a tool in fungal taxonomy». *Systematic and Applied Microbiology*, 21, 433-41.
50. Torres, A. and Greenacre, M. J. (2002). «Dual scaling and correspondence analysis of preferences, ratings and paired comparisons». *International Journal for Research in Marketing*, to appear.

THE RESEARCH GROUP ON STATISTICAL ANALYSIS OF COMPOSITIONAL DATA

Carles Barceló
Departament d'Informàtica i Matemàtica Aplicada
Universitat de Girona

Compositional data analysis deals with positive, relative data, which express proportions of some whole. Typical examples are data recorded as percentages, ppm, ppb, g/kg, or in relative units, such as data in atomic or molecular proportions which are obtained as the ratio of weight percent to atomic or molecular weight. Compositional data are common in all fields of applied science, from natural to social science.

Historically, only the first type of data has been considered to be compositional, and the fact that percentages are naturally subject to a unit-sum constraint has been considered the special and intrinsic feature of this type of data. In applications, this unit-sum constraint has been widely ignored or wished away, and inappropriate 'standard' statistical methods, devised for and successfully applied to unconstrained data, have been used with disastrous consequences. The solution was given in the early 1980's by John Aitchison, who introduced a new methodology based on the additive logratio transformation from the d-dimensional simplex sample space to d-dimensional real space. This approach was extended to spatially dependent compositional data (Pawlowsky-Glahn, 1984, 1989; Pawlowsky-Glahn and Burger, 1992; Barceló-Vidal, and Pawlowsky-Glahn, 1995), to finite mixtures of compositions (Barceló-Vidal and Pawlowsky-Glahn, 1995; Barceló-Vidal, Pawlowsky-Glahn and Grunsky, 1995), and to the detection of outliers in compositional data (Barceló-Vidal, Pawlowsky-Glahn and Grunsky, 1996). Presently, a group of researchers from the Universitat de Girona, the Universitat Politècnica de Catalunya in Barcelona, and the Universitat de Barcelona –together with John Aitchison, professor emeritus of the University of Glasgow, and other researchers from the Free University of Berlin (Germany), the Universities of Jena (Germany), Firenze (Italy), and Kansas (USA), as well as the Canadian Geological Survey– are working on further developments of this methodology, emphasizing the geometric aspects of compositional data analysis and the applications to specific problems in geology and archaeometrics.

Geometric approach to statistical analysis on the simplex: According to Aitchison, perturbation and power transformation induce a real vector space structure in the simplex. Furthermore, an inner product, with associated norm and distance can be defined. The resulting Euclidean space structure of the simplex creates a framework within which all statistical methods devised for data in real space can be integrated. In particular, previous concepts specific to compositional data appear as the natural counterpart to equivalent concepts in real space, like measures of location or measures of variability (Pawlowsky-Glahn and Egozcue, 2001*a*, 2001*b*).

The natural sample space of compositional data: The recent developments of the vector space structure of the simplex have led to a more extensive study of the mathematical rationale underlying compositional data. This study has led to the definition of the natural sample space of compositional data as a quotient space, a datum thus being a class of equivalence, and reducing the simplex to one possible representative of this quotient space. This approach implies that equivalent representations are also compositional in nature, showing that the constant-sum constraint is not an inherent property to compositional data, while the relative nature certainly is. (Barceló-Vidal *et al.*, 2001).

Cluster analysis on the simplex: The compositional distance introduced by Aitchison and a new measure of difference for compositional data based on measures of divergence have been used to develop an appropriate clustering strategy for compositional data (Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn, 1997, 1998*a*, 1998*b*, 1999). Within this framework, a new multiplicative zero replacement algorithm has been proposed (Martín-Fernández Barceló-Vidal and Pawlowsky-Glahn, 2000).

Applications: Compositional data analysis is present in different scientific fields, for example in geology, archaeology, sociology and econometrics. The new methodology has been applied to solve specific problems of these fields giving rise to new concepts –coherent with the nature of the sample space– such as differential perturbation processes (Aitchison and Thomas, 1998; Buccianti, Pawlowsky-Glahn, Barceló-Vidal and Jarauta-Bragulat ,1999), compositional centering (Martín-Fernández, Bren, Barceló-Vidal and Pawlowsky-Glahn, 1999; Eynatten, Pawlowsky-Glahn and Egozcue, 2002), and others (Pawlowsky-Glahn and Barceló-Vidal, 1999; Martín-Fernández, Olea-Meneses and Pawlowsky-Glahn, 2001; Pawlowsky-Glahn and Buccianti, 2001). Sometimes the compositional methodology has been misunderstood by the specialists of these fields and therefore supplementary efforts have been devoted to convince the scientific community of the coherence and advantages of the approach (Barceló-Vidal, Martín-Fernández and Pawlowsky-Glahn, 1999; Aitchison, Barceló-Vidal, Martín-Fernández and Pawlowsky-Glahn, 2000, Aitchison, Barceló-Vidal and Pawlowsky-Glahn, 2001).

New probability distributions on the simplex: The multivariate skewnormal distribution introduced by Azzalini and Dalla Valle (1996), and Azzalini and Capitanio (1999) allowed to define a new parametric class of distributions on the simplex, the additive logistic skewnormal class of distributions (Mateu-Figueras, Barceló-Vidal and Pawlowsky-Glahn, 1998). This class includes the additive logistic normal distribution defined originally by Aitchison and expands the possibilities to capture the patterns of variability observed in practice on compositional data sets. The geometric approach to statistical analysis of compositional data mentioned before offers a new framework within which a systematic approach to distributions on the simplex can be undertaken.

References

1. Aitchison, J., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2001). «Some comments on compositional data analysis in archaeometry, in particular the fallacies in Tangri and Wright's dismissal of logratio analysis». *Archaeometry*, (in press).
2. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2000). «Logratio analysis and compositional distance». *Mathematical Geology*, 32, 271-275.
3. Aitchison, J. and Thomas, C. W. (1998). «Differential perturbation processes: a tool for the study of compositional processes». In A. Buccianti, G. Nardi and R. Potenza (eds.), *Proceedings of IAMG'98–The Fourth Annual Conference of the International Association for Mathematical Geology*. De Frede Editore, Napoli (I), 499-504.
4. Azzalini, A. and Dalla Valle, A. (1996). «The multivariate skew-normal distribution». *Biometrika*, 83, 715-726.
5. Azzalini, A. and Capitanio, A. (1999). «Statistical applications of the multivariate skew-normal distribution». *Journal of Royal Statistical Society, series B*, 61, 579-602.
6. Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1995). «Finite mixtures of compositional data». *Science de la Terre*, 32, 29-48.
7. Barceló-Vidal, C., Pawlowsky-Glahn, V. and Grunsky, E. (1995). «Classification problems of samples of finite mixtures of compositions». *Mathematical Geology*, 27, 129-148.
8. Barceló-Vidal, C., Pawlowsky-Glahn, V. and Grunsky, E. (1996). «Some aspects of transformations of compositional data and the identification of outliers». *Mathematical Geology*, 28, 501-518.
9. Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (1999). «Comment on "Singularity and Nonnormality in the Classification of Compositional Data" by Bohling, G. C. *et al.*», *Mathematical Geology*, 31, 581-586.
10. Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2001). «Mathematical foundations for compositional data analysis». *Proceedings of IAMG'01 –The Annual Conference of the International Association for Mathematical Geology*. CD-Rom. Cancún (México), 20 pp.
11. Buccianti, A., Pawlowsky-Glahn, V., Barceló-Vidal, C. and Jarauta-Bragulat, E. (1999). «Visualization and modeling of natural trends in ternary diagrams: a geochemical case study». In S. J. Lippard, A. Næss and E. Sinding-Larsen R. (eds.), *Proceedings of IAMG'99–The Fifth Annual Conference of the International Association for Mathematical Geology*. Tapir, Trondheim (N), 139-144.
12. Eynatten, H. V., Pawlowsky-Glahn, V. and Egozcue, J. J. (2002). «Understanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams». *Mathematical Geology*, (accepted for publication).

13. Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1997). «Different classifications of the Darss Sill data set based on mixture models for compositional data». In V. Pawlowsky-Glahn (ed.), *Proceedings of IAMG'97–The Third Annual Conference of the International Association for Mathematical Geology*. International Center for Numerical Methods in Engineering (CIMNE), Barcelona, 151-158.
14. Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998*a*). «Measures of difference for compositional data and hierarchical clustering methods». In A. Buccianti, G. Nardi and R. Potenza (eds.), *Proceedings of IAMG'98–The Fourth Annual Conference of the International Association for Mathematical Geology*. De Frede Editore, Napoli, 526-531.
15. Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998*b*). «A critical approach to non-parametric classification of compositional data». In A. Rizzi, M. Vichi and H. H. Bock (eds.), *Advances in Data Science and Classification. IFCS-98–Proceedings of the 6th Conference of the International Federation of Classification Societies*. Springer-Verlag, 49-56.
16. Martín-Fernández, J. A., Bren, M., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1999). «A measure of difference for compositional data based on measures of divergence». In S. J. Lippard, A. Næss and E. Sinding-Larsen R. (eds.), *Proceedings of IAMG'99–The Fifth Annual Conference of the International Association for Mathematical Geology*. Tapir, Trondheim, 211-215.
17. Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2000). «Zero Replacement in Compositional Data Sets». In: H. A. L. Kiers, J. P. Rasson, P. J. F. Groenen and M. Shader, *Proceedings of IFCS'2000, The 7th Conference of the International Federation of Classification Societies*, Namur, 155-160.
18. Martín-Fernández, J. A., Olea-Meneses, R. and Pawlowsky-Glahn, V. (2001). «Criteria to compare estimation methods of regionalized compositions». *Mathematical Geology*, (in press).
19. Mateu-Figueras, G., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998). «Modeling compositional data with multivariate skew-normal distributions». In A. Buccianti, G. Nardi and R. Potenza (eds.), *Proceedings of IAMG'98–The Fourth Annual Conference of the International Association for Mathematical Geology*. De Frede Editore, Napoli, 532-537.
20. Pawlowsky-Glahn, V. (1984). «On spurious spatial covariance between variables of constant sum». *Science de la Terre*, 21, 107-113.
21. Pawlowsky-Glahn, V. (1989). «Cokriging of regionalized compositions». *Mathematical Geology*, 21, 513-521.
22. Pawlowsky-Glahn, V. and Barceló-Vidal, C. (1999). «Confidence regions in ternary diagrams». *Terra Nostra*, 99, 37-47.
23. Pawlowsky-Glahn, V. and Buccianti, A. (2001). «Visualization and modeling of subpopulations of compositional data: statistical methods illustrated by means of

geochemical data from fumarolic fluids». *International Journal of Earth Sciences*, (in press).

24. Pawlowsky-Glahn, V. and Buger, H. (1992). «Spatial structure analysis of regionalized compositions». *Mathematical Geology*, 24, 675-691.
25. Pawlowsky-Glahn, V. and Egozcue, J. J. (2001*a*). «About BLU estimators and compositional data». *Mathematical Geology*, (in press).
26. Pawlowsky-Glahn, V. and Egozcue, J. J. (2001*b*). «Geometric approach to statistical analysis on the simplex». *Stochastic Environmental Research and Risk Assessment*, 15, 384-398.
27. Pawlowsky, V., Olea, R. A. and Davis, J. C. (1995). «Estimation of regionalized compositions: a comparison of three methods». *Mathematical Geology*, 27, 105-148.

STRUCTURAL EQUATION MODELLING

Joan M. Batista
Departament Mètodes Quantitatius de Gestió
Esade

Structural Equation Modeling (SEM) is a powerful technique that can combine complex path or simultaneous equation models with latent variables measured by a factor analysis model. This general approach includes Confirmatory Factor Analysis Models and Simultaneous Regression Models as special cases and can be extended to multiple populations, hierarchical data structures and non-metric variables. Thurstone or Maxwell and Lawley's Factor Analysis and Sewall Right's Path analysis can be considered as forerunners of these models. The general model was the result of a conference summoned by Goldberger in 1971 where for the first time measurement relationships between observable indicators and their underlying constructs –psychometrics– received the same attention as the substantive relationship between constructs –econometrics.

Since the late seventies, a group of researchers in Barcelona, all of them disciples and colleagues of Professor W. E. Saris (University of Amsterdam) has been working on this model mainly from its measurement perspective. Nowadays, research in this area revolves around Universitat de Barcelona (A. Maydeu-Olivares), Universitat de Girona (G. Coenders), Universitat Rovira i Virgili (P. J. Ferrando, U. Lorenzo-Seva), Universitat Pompeu Fabra (A. Satorra), and ESADE-URL (J. M. Batista-Foguet). Several research agendas are also under way. One of them is robust estimation and testing techniques (See Latent Variable Models by A. Satorra in this same issue). The remaining areas are outlined below, together with the major publications.

Multitrait-Multimethod Models: Special cases of SEM are not only used for reliability and validity assessment of measurement instruments in the social sciences, but also in the physical sciences. Their scope ranges from comparing different answer modes in a questionnaire to comparing different blood pressure measurement protocols. There have been contributions to generalise these models for different patterns of method effects, for bias evaluation instead of only reliability and validity, and for planned missing data designs (Batista-Foguet and Saris, 1988; Batista-Foguet, and Saris, 1992; Coenders, Batista-Foguet and Satorra, 1995; Coenders and Saris, 1998; Coenders and Saris, 2000*a*; Coenders and Saris, 2000*b*; Batista-Foguet, Coenders and Artés, 2001; Kogovšek *et al.*, in press; Corten *et al.*, in press).

Quasi simplex models: Special cases of SEM, they are used for reliability assessment in panel designs and quantification of change over time, specifically for evaluating stability of attitudes. There have been contributions in studying their statistical properties, their robustness and in generalising the analysis to mean structures instead of only analysing covariance structures (Batista-Foguet and Saris, 1997; Coenders *et al.*, 1999).

SEM for categorical ordered dependent variables: Since SEM has been established to deal with normally distributed variables, measures other than covariances are required in the ordinal case. A series of studies has been made concerning the robustness of alternative measures of association, comparing their properties with those of covariances, concerning the measurement implications of treating the observed variables as categorical or as continuous, and on using parametric or non-parametric approaches (Maydeu-Olivares, 1994; Coenders and Saris, 1995; Ferrando, 1996; Coenders, Satorra and Saris, 1997; Ferrando, 2000).

Development of new estimation methods and test statistics in SEM with non-metric and censored variables: Current estimation methods in SEM for non-metric dependent variables use a three-stage approach. A two-stage approach has been proposed and compared with the existing approach. In addition, new test statistics have been proposed for extremely sparse contingency tables and for dealing with censored data (Ferrando and Lorenzo-Seva, 1999; Maydeu-Olivares, 2001*a*; Maydeu-Olivares, in press).

Random utility modeling using a SEM approach: An overall model for fitting random utility models has been proposed that encompasses as special models, models for paired comparisons data and ranking (permutation) data. A SEM approach which relies on limited information estimation and testing procedures has been found to compare very favourably to full information approaches based on the EM algorithm or Gibbs sampling.

References

1. Batista-Foguet, J. M. and Saris, W. E. (1988). «Reduction in variation in response function for Social Science variables: Job-Satisfaction». In Willem E. Saris (ed.) *Variation in Response Function a Source of Measurement Error in Survey Research.* Amsterdam, Sociometric Research Foundation.
2. Batista-Foguet, J. M. and Saris, W. E. (1992). «A New Measurement Procedure for Attitudinal Research. Analysis of its Psychometric and Informational Properties». *Quality & Quantity*, 26, 127-146.
3. Batista-Foguet, J. M. and Saris, W. E. (1997). «Tests of Stability in Attitude Research». *Quality & Quantity*, 31, 269-285.
4. Batista-Foguet, J. M., Coenders, G. and Artés, M. (2001). «Using structural equation models to evaluate the magnitude of measurement error in blood pressure». *Statistics in Medicine*, 20, 2351-2368.
5. Coenders, G., Batista-Foguet, J. M. and Satorra, A. (1995). «Scale dependence of the true score MTMM model», in Saris, W. E. and Münnich, Á. (Eds.) *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments.* Eötvös University Press, Budapest, 71-87.

6. Coenders, G. and Saris, W. E. (1995). «Categorization and measurement quality. The choice between Pearson and polychoric correlations», in Saris, W. E. and Münnich, Á. (Eds.) *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments*. Eötvös University Press, Budapest, 125-144.
7. Coenders, G., Satorra, A. and Saris, W. E. (1997). «Alternative approaches to structural modeling of ordinal data. A Monte Carlo study», *Structural Equation Modeling*, 4, 261-282.
8. Coenders, G. and Saris, W. E. (1998). «Relationship between a restricted correlated uniqueness model and a direct product model for multitrait-multimethod data», in Ferligoj, A. (Ed.) *Advances in Methodology, Data Analysis and Statistics, Metodoloski Zvezki*, 14. FDV. Ljubljana, 151-172.
9. Coenders, G., Saris, W. E., Batista-Foguet, J. M. and Andreenkova, A. (1999). «Stability of three-wave simplex estimates of reliability», *Structural Equation Modeling*, 6, 135-157.
10. Coenders, G. and Saris, W. E. (2000*a*). «Systematic and random method effects. Estimating method bias and method variance», in Ferligoj, A. and Mrvar, A. (Eds.). *Developments in Survey Methodology, Metodoloski Zvezki*, 15. FDV, Ljubljana, 55-74.
11. Coenders, G. and Saris, W. E. (2000*b*). «Testing nested additive, multiplicative and general multitrait-multimethod models». *Structural Equation Modeling*, 7, 219-250.
12. Corten, I. W., Saris, W. E., Coenders, G., van der Veld, W., Aalberts, C. E. y Kornelis, C. (in press). «Fit of different models for multitrait-multimethod experiments». *Structural Equation Modeling*.
13. Ferrando, P. J. (1996). «Calibration of invariant item parameters in a continuous item response model using the extended Lisrel measurement submodel». *Multivariate Behavioral Research*, 31, 419-439.
14. Ferrando, P. J. (2000). «Testing the equivalence among different item response formats in personality measurement: A structural equation modeling approach». *Structural Equation Modeling*, 7, 271-286.
15. Ferrando, P. J, Lorenzo-Seva and U. (1999). «Implementing a test of underlying normality for censored variables». *Multivariate Behavioral Research*, 34, 421-439.
16. Kogovšek, T., Ferligoj, A. Coenders, G. and Saris, W. E. (in press). «Estimating reliability and validity of personal support measurements: Full information ML estimation with planned incomplete data». *Social Networks*.
17. Maydeu-Olivares, A. (1994). «Parametric vs. non-parametric approaches to individual differences scaling». *Psicothema*, 6, 297-310.
18. Maydeu-Olivares, A. (1999). «Thurstonian modeling of ranking data via mean and covariance structure analysis». *Psychometrika*, 64, 325-340.

19. Maydeu-Olivares, A. (2001). «Multidimensional item response theory modeling of binary data: Large sample properties of NOHARM estimates». *Journal of Educational and Behavioral Statistics*, 26, 49-69.

20. Maydeu-Olivares, A. (2001). «Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgement sampling». *Psychometrika*, 66, 209-228.

21. Maydeu-Olivares, A. (in press). «Limited information estimation and testing of Thurstonian models for preference data». *Mathematical Social Sciences.*

MATHEMATICAL STATISTICS AND FINANCE GROUP

Joan del Castillo
Departament de Matemàtiques
Facultat de Ciències
Universitat Autònoma de Barcelona

Since 1990 our group has been studying inference problems related to several application fields: reliability, survival analysis, log-linear models, goodness of fit tests and, more recently, mathematical finance. Our experience and research is mainly in the following topics:

Exponential models: Most of the classical models in Statistics are exponential models. These models have a unified theory that facilitates working with them. We have developed new exponential models related to different applied problems. For instance, we have developed new models in reliability and survival analysis (Castillo and Puig, 1999), new tests for Gamma and Rayleigh distributions (Castillo and Puig, 1997), and new models that generalize the Poisson distribution (Castillo and Pérez, 1998).

High order asymptotic methods: In applied work most of the statistics have moment generation functions. In these cases high order asymptotic methods improve greatly the classical asymptotic theory. We have used these methods in non-regular situations such as, for instance, the case where the parameters are in the boundary of the domain (Castillo and Puig, 1999). New results are now in progress.

Goodness of fit tests: Some goodness of fit tests have been developed in the context of exponential models in Castillo and Puig (1997, 1999). They are smooth tests in the sense that the distribution of interest is imbedded in a model with more parameters. The classical omnibus tests based on the EDF statistics have also been studied. These are especially convenient for location and scale models like the Laplace distribution (Puig and Stephens, 2000). In Puig and Stephens (2001) these tests are developed for the hyperbolic distribution that is usually employed for modelling log-returns in the stock market.

Levy processes and finance: The need for complex models to analyze financial data has produced a big interest in Levy processes. We have developed a Malliavin type calculus for such processes and have applied this to pricing and hedging an European option in a market with jumps (Leon, Solé, Utzet and Vives, 2002).

References

1. Castillo, J. (1994). «The singly truncated normal distribution a non-steep exponential family». *Annals of the Institute of Statistical Mathematics*, 46, 57-66.
2. Castillo, J. and Puig, P. (1997). «Testing departures from Gamma, Rayleigh and singly truncated normal distributions». *Annals of the Institute of Statistical Mathematics*, 49, 255-269.
3. Castillo, J. and Pérez, M. (1998). «Weighted Poisson distributions for overdispersion and underdispersion situations». *Annals of the Institute of Statistical Mathematics*, 50, 567-585.
4. Castillo, J. and Puig, P. (1999). «Invariant exponential models applied to reliability theory and survival analysis». *Journal of the American Statistical Association*, 94, 522-528.
5. Castillo, J. and Puig, P. (1999). «The best test of exponentiality against singly truncated normal alternatives». *Journal of the American Statistical Association*, 94, 529-532.
6. Leon, J., Solé, J. L., Utzet, F. and Vives, J. (2002). «On Levy processes, Malliavin Calculus and market models with jumps». *Finance and Stochastics*, to appear.
7. Puig, P. and Stephens, Michael A. (2000). «Tests of fit for the Laplace distribution, with applications». *Technometrics*, 42, 417-424.
8. Puig, P. and Stephens, Michael A. (2001). «Goodness-of-fit tests for the hyperbolic distribution». *Canadian Journal of Statistics*, 29, 309-320.

THE MULTIVARIATE ANALYSIS RESEARCH GROUP

Carles M Cuadras
Departament d'Estadística
Facultat de Biologia
Universitat de Barcelona

The set of statistical methods known as Multivariate Analysis covers a wide group of theoretical and applied methods, including factor analysis, classification, multivariate analysis of variance and multidimensional scaling. Since 1980, and especially since 1989, a group of researchers in Barcelona has been studying geometric aspects of statistics, with applications in statistical inference, regression and multivariate analysis.

Distance-based regression: In this approach a response variable is predicted using several explanatory variables on quantitative and qualitative measurement scales. The key idea is to define a distance between observations and to project the response variable on the principal dimensions obtained via multidimensional scaling. This distance-based (DB) model, which may be named «principal coordinate regression», generalizes and improves the multiple regression model, as well as the non-linear regression model by using suitable distance functions (Cuadras, 1989; Cuadras and Arenas, 1990; Cuadras, Arenas and Fortiana, 1996). Related results can be found in Cuadras (1993) and Fortiana and Cuadras (1997).

Distance-based discriminant analysis: The classic problem of allocating an observation to two or more known groups, can also be solved by using the DB approach. Using only distance matrices, one for each group, we can define a proximity function which reduces to the classic discriminant function for particular distances, such as Euclidean or Mahalanobis. This method allows us to handle nominal variables and missing data, and approach the problem of typicality in classification (Cuadras, 1989; Cuadras, 1992; Cuadras, Fortiana and Oliva, 1997; Cuadras, Atkinson and Fortiana, 1997; Cuadras and Fortiana, 2000; Villarroya, Rios and Oller, 1995).

Related metric scaling: A natural extension of this DB approach is to define a joint distance as a function of two given distances on the same set, which satisfies some specific rules (e.g., additivity in the case of independence) and preserves the redundancy between both distances. This approach allows us to relate distances and to represent multivariate data under two different kinds of information (Cuadras and Fortiana, 1995, 1996; Cuadras, 1998; Arenas *et al.*, 2000).

Principal components of a random variable: Just as we can obtain the principal components of a finite set of variables, we can define and obtain the principal directions of a Bernoulli process associated with a continuous random variable. This construction

is useful in goodness-of-fit assessment, in expanding a random variable in terms of its principal components, in constructing bivariate distributions with given marginals and in studying the asymptotic distribution of some statistics related to Rao's quadratic entropy (Cuadras and Fortiana, 1995; Cuadras and Lahlou, 2000).

Distributions with given marginals: The construction of joint distributions given the marginals and some dependence parameters, is of great interest in probability and statistics. Families of distributions have been obtained when an interdependence matrix between marginals is given and when the regression curve is given. A continuous extension of correspondence analysis has also been derived (Cuadras and Auge, 1981; Ruiz-Rivas and Cuadras, 1988; Cuadras, 1992; Cuadras, 1996; Cuadras and Fortiana, 1997; Cuadras, Fortiana and Greenacre, 2000; Cuadras, 2002).

Differential geometry in statistics: If we interpret a statistical model as a Riemannian manifold, and define a distance between parameters through geodesics by using the Fisher information matrix as the metric tensor, we obtain the Rao distance, a natural extension of the Mahalanobis distance. This allows us to study the geometry of any regular statistical model and approach some problems in statistical inference, such as intrinsic estimation, invariance, testing of hypotheses and representing parametric models. (Oller and Cuadras, 1985; Oller, 1989; Calvo and Oller, 1990; Oller and Corcuera, 1995; Rios, Villarroya and Oller, 1992; Villarroya and Oller, 1993; Villarroya, Rios and Oller, 1995; Garcia and Oller, 2001).

References

1. Arenas, C. and Cuadras, C. M. (2002). «Recent statistical methods based on distances». *Contributions to Science*, in press.
2. Arenas, C., Escudero, T., Mestres, F., Coll, M. D. and Cuadras, C. M. (2000). «Cacromos: a fortran program to reconstruct the position of human chromosomes». *Hereditas*, 132, 157-159.
3. Barata, C., Baird, D., Miñarro, A. and Soares, A. (2000). «Do genotype responses always converge from lethal to nonlethal toxicant exposure levels? Hypothesis tested using clones of Daphnia magna straus». *Environmental Toxicology and Chemistry*, 19, 2314-2322.
4. Burbea, J., Oller, J. M. and Reverter, F. (2002). «Some remarks on the information geometry of the gamma distribution». *Communications in Statistic: Theory and Methods*, 31, 1959-1975.
5. Calvo, M. and Oller, J. M. (1990). «A distance between multivariate normal distribution based in an embedding into the Siegel group». *Journal of Multivariate Analysis*, 35, 223-242.

6. Calvo, M. and Oller, J. M. (2002). «A distance between elliptical distributions based in an embedding into the Siegel group». *Journal of Computational and Applied Mathematics*, 145, 319-334.
7. Calvo, M., Villarroya, A. and Oller, J. M. (2002). «A biplot method for multivariate normal populations with unequal covariance matrix». *TEST*, 11, 143-165.
8. De la Cruz, X. and Calvo, M. (2001). «Use of surface area computations to describe atom-atom interactions». *Computer-Aided Molecular Design*, 15, 521-532.
9. Cuadras, C. M. (1989). «Distance analysis in discrimination and classification using both continuous and categorical variables», in Y. Dodge (ed.), *Statistical Data Analysis and Inference*, 459-473. Elsevier Science Publishers, Amsterdam.
10. Cuadras, C. M. (1992). «Probability distributions with given multivariate marginals and given dependence structure». *Journal of Multivariate Analysis*, 42, 51-66.
11. Cuadras, C. M. (1992). «Some examples of distance based discrimination». *Biometrical Letters*, 29, 1-18.
12. Cuadras, C. M. (1993). «Interpreting an inequality in multiple regression». *The American Statistician*, 47, 256-258.
13. Cuadras, C. M. (1996). «A distribution with given marginals and given regression curve», in: L. Rüschendorf, B. Schweizer and D. Taylor (eds.), *Distributions with Fixed Marginals and Related Topics*, 76-83. IMS-Lecture Notes-Monograph Series, Hayward, California.
14. Cuadras, C. M. (1998). «Multidimensional dependencies in classification and ordination», in: K. Fernández-Aguirre and A. Morineau (eds.), *Analyse Multidimensionnelles des Données*, 15-25. CISIA, Saint Mandé, France.
15. Cuadras, C. M. (1998). «Some cautionary notes on the use of principal components regression (Revisited)». *The American Statistician*, 52, 371.
16. Cuadras, C. M. (2000). «Distributional relationships arising from simple trigonometric formulas (Revisited)». *The American Statistician*, 54, 87.
17. Cuadras, C. M. (2000). «Discussion on "Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests" by E. del Barrio, J. A. Cuesta-Albertos and C. Matrán». *TEST*, 9, 1-96.
18. Cuadras, C. M. (2002). «On the covariance between functions». *Journal of Multivariate Analysis*, 81, 19-27.
19. Cuadras, C. M. (2002). «Correspondence analysis and diagonal expansions in terms of distribution functions». *Journal of Statistical Planning and Inference*, 103, 137-150.
20. Cuadras, C. M. (2002). «Diagonal distributions via diagonal expansions and tests of independence». In *Distributions with given marginals and statistical modelling*, (C. M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena, Eds.), Kluwer Academic Publishers Dordrecht, 35-42.
21. Cuadras, C. M. (2002). «Discussion on "Skewed multivariate models related to hidden truncation and/or selective reporting" by B. C. Arnold and R. J. Beaver». *TEST*, 11, 7-54.

22. Cuadras, C. M. (2002). «Geometrical understanding of the Cauchy distribution». *Qüestiió*, 26, 283-287.
23. Cuadras, C. M., Fortiana, J. and J. A. Rodriguez-Lallena (eds.) (2002). *Distributions with given marginals and statistical modelling*. Kluwer Academic Publishers, Boston/Dordrecht/London, 272 pp.
24. Cuadras, C. M. and Lahlou, Y. (2000). «Some orthogonal expansions for the logistic distribution». *Communications in Statistics: Theory and Methods*, 29, 2643-2663.
25. Cuadras, C. M. and Lahlou, Y. (2002). «Principal components of the Pareto distribution». In *Distributions with given marginals and statistical modelling* (C. M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena, Eds.), Kluwer Academic Publishers, Dordrecht, 43-50.
26. Cuadras, C. M. and Cuadras, D. (2001). «Principal directions for the normal random variable». *Statistical Review*, 2, 101-102.
27. Cuadras, C. M. and Cuadras, D. (2002). «Orthogonal expansions and distinction between logistic and normal». In *Goodness-of-fit Tests and Validity Models*. C. Huber-Carol, N. Balakrishnan, M. S. Nikulin, M. Mesbah, eds., 325-338, Birkhauser, Boston.
28. Cuadras, C. M. and Arenas, C. (1990). «A distance based regression model for prediction with mixed data». *Communications in Statistics: Theory and Methods*, 19, 2261-2279.
29. Cuadras, C. M., Arenas, C. and Fortiana, J. (1996). «Some computational aspects of a distance-based model for prediction». *Communications in Statistics: Simulation and Computation*, 25, 593-609.
30. Cuadras, C. M., Atkinson, R. A. and Fortiana, J. (1997). «Probability densities from distances and discriminant analysis». *Statistics and Probability Letters*, 33, 405-411.
31. Cuadras, C. M. and Augé, J. (1981). «A continuous general multivariate distribution and its properties». *Communications in Statistics: Theory and Methods*, A10, 339-353.
32. Cuadras, C. M. and Fortiana, J. (1995). «A continuous metric scaling solution for a random variable». *Journal of Multivariate Analysis*, 52, 1-14.
33. Cuadras, C. M. and Fortiana, J. (1996). «Weighted continuous metric scaling», in A. K. Gupta and V. L. Girko (eds.), *Multidimensional Statistical Analysis and Theory of Random Matrices*, 27-40. VSP, Zeist, The Netherlands.
34. Cuadras, C. M. and Fortiana, J. (1997). «Continuous scaling on a bivariate copula», in V. Benes and J. Stepan (eds.), *Distributions with Given Marginals and Moment Problems*, 137-142. Kluwer Academic Publishers, The Netherlands.
35. Cuadras, C. M. and Fortiana, J. (2000). «The importance of geometry in multivariate analysis and some applications», in C. R. Rao and G. Szekely, (eds.), *Statistics for the 21st Century*, 93-108. Marcel Dekker, N. York.

36. Cuadras, C. M., Fortiana, J. and Greenacre, M. J. (1999). «Continuous extensions of matrix formulations in correspondence analysis, with applications to the FGM family of distributions», in R. D. H. Heijmans, D. S. G. Pollock and A. Satorra, (eds.), *Innovations in Multivariate Statistical Analysis*, 101-116. Kluwer Academic Publishers, The Netherlands.
37. Cuadras, C. M., Fortiana, J. and Oliva, F. (1997). «The proximity of an individual to a population with applications in discriminant analysis». *Journal of Classification*, 14, 117-136.
38. Cuadras, C. M. and Fortiana, J. (1993). «Continuous metric scaling and prediction», in C. M. Cuadras and C. R. Rao (eds.), *Multivariate Analysis: Future Directions*, 2, Elsevier, Amsterdam, 47-66.
39. Cuadras, C. M. and Fortiana, J. (1994). «As certeining the underlying distribution of a data set», in *Selected Topics on Stochastic Modelling*, R. Gutierrez and M. J. Valderrama (eds.), World Scientific, Singapore, 223-230.
40. Cuadras, C. M. and Fortiana, J. (1998). «Visualizing categorical data with related metric scaling». In *Visualization of Categorical Data*, ch. 25, J. Blassius and M. Greenacre (eds.), Ac. Press, N. York, 365-376.
41. Cuadras, C. M. and Rao, C. R. (eds.) (1993). *Multivariate Analysis: Future Directions* 2, Elsevier, Amsterdam, 488 pp.
42. Cubedo, M. and Oller, J. M. (2002). «Hypothesis testing: a model selection approach». *Journal of Statistical Planning and Inference*, 108, 3-21.
43. Fortiana, J. and Grané, A. (2002). «A scale-free goodness-of-fit statistic for the exponential distribution based on maximum correlations». *Journal of Statistical Planning and Inference*, 108, 85-97.
44. Fortiana, J. and Grané, A. (2002). «Goodness-of-fit tests based on maximum correlations and their orthogonal decompositions». *Journal of Royal Statistical Society B*, in press.
45. García, G. and Oller, J. M. (2001). «Minimum Riemannian risk equivariant estimator for the univariate normal model». *Statistics & Probability Letters*, 52 109-113.
46. Grané, A. and Fortiana, J. (2002). «Maximum correlations and tests of goodness-of-fit». In *Distributions with given marginals and statistical modelling* (C. M. Cuadras, J. Fortiana and J. A. Rodriguez-Lallena, eds.), Kluwer Academic Publishers, Dordrecht, 113-122.
47. Ruiz-Rivas, C. and Cuadras, C. M. (1988). «Inference properties of a one-parameter curved exponential family of distributions with given marginals». *Journal of Multivariate Analysis*, 27, 447-456.
48. Fortiana, J. and Cuadras, C. M. (1997). «A family of matrices, the discretized Brownian Bridge and distance-based regression». *Linear Algebra and its Applications*, 263, 173-188.
49. Oller, J. M. (1989). «Some geometrical aspects of data analysis and statistics», in: Y. Dodge (ed.), *Statistical Data Analysis and Inference*, 41-58. Elsevier Science Publishers, Amsterdam.

50. Oller, J. M. and Corcuera, J. M. (1995). «Intrinsic analysis of statistical estimation». *Annals of Statistics*, 23, 1562-1581.
51. Ríos, M., Gracia, J. M., Sánchez, J. A. and Pérez, D. (2000). «A statistical analysis of the seasonality in Pulmonary Tuberculosis». *European Journal of Epidemiology*, 16, 483-488.
52. Oller, J. M. and Cuadras, C. M. (1985). «Rao's distance for negative multinomial distributions». *Sankhyã*, A 47, 75-83.
53. Ríos, M., Villarroya, A. and Oller, J. M. (1992). «Rao distance between multivariate linear models and their applications to the classification of response curves». *Computational Statistics and Data Analysis*, 13, 431-441.
54. Villarroya, A. and Oller, J. M. (1993). «Statistical tests for the inverse Gaussian distribution based on Rao distance». *Sankhyã*, A55, 80-103.
55. Villarroya, A., Ríos, M. and Oller, J. M. (1995). «Discriminant analysis algorithm based on a distance funtion and a Bayesian decision». *Biometrics*, 51, 908-919.

STATISTICAL DISCLOSURE CONTROL IN CATALONIA AND THE CRISES GROUP

Josep Domingo
Departament d'Enginyeria Química, Estadística i I.O.
ETSE
Universitat Rovira i Virgili

Statistical offices release two kinds of data through their statistical databases: tabular data and microdata sets (individual respondent records). In both cases, data dissemination should be performed in a way that does not lead to disclosure of individual information but preserves the informational content as much as possible. This is known as the statistical disclosure control (SDC) problem. While there is a long experience in table dissemination, microdata dissemination is a much more recent activity.

Data security, including statistical disclosure control, is the research topic of the CRISES group, based at the Universitat Rovira i Virgili, Tarragona, Catalonia (http://www.etse.urv.es/recerca/crises). In English CRISES stands for «CRyptography and Inference Surveillance in Electronic Systems»; in Catalan it stands for «CRiptografia i Secret EStadístic» (Cryptography and Statistical Data Confidentiality).

Although the activity of CRISES includes other data security topics beyond SDC (cryptography, secure e-commerce, etc.), the next sections focus on the contributions made by CRISES and other related Catalan research groups to the field of statistical disclosure control. Section 1 below describes Catalan activity on SDC during the period 1993-1999. Section 2 reports on U.S. government funded project OTTILIE (1999-2000) for assessment of microdata protection methods. Finally, Section 3 is an update on current research being carried by Catalan researchers under European 5th FP projects CASC and AMRADS.

Catalan activity in SDC (1993-1999): Probably the first Catalan paper published on SDC was by an IDESCAT researcher back in J. Turmo, 1993. Based on that paper and subsequent work, IDESCAT released a sample of the 1991 population census of Catalonia (IDESCAT, 1995). Sampling was the procedure used to control the risk of an individual being reidentified (Garín and Ripoll, 1999). In the period 1996-1999, CRISES was the only research group in Catalonia doing work on SDC, supported by CICYT project TIC95-0903-C02-02 and by two IDESCAT research contracts. Work done in this period encompasses both tabular data and microdata.

SDC applied to tabular data has the goal of avoiding exact disclosure of an individual attribute. For contingency or magnitude tables, this means that cells to which a small number of individuals contribute should be specially protected. Cell suppression is the most common approach to tabular data protection. In Domingo-Ferrer and Mateo-Sanz (1996), showed that cell suppression can fail to meet its security goals when the protec-

ted tables contain one or more quantitative factors. In Domingo-Ferrer and Mateo-Sanz (1999), a computationally efficient method for tabular data protection based on resampling proposed.

Regarding SDC for microdata, the CRISES group concentrated on microaggregation, which is a family of methods for «masking» (i.e. protecting) an original microdata set and turning it into a publishable protected microdata set. The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if the data records correspond to groups of k or more individuals, where no individual dominates (i.e. contributes too much to) the group and k is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with averages computed on small aggregates (microaggregates) prior to publication. This is the basic principle of microaggregation. The microaggregation problem consists of finding the k-partition (partition of the set of records into groups of size at least k) that minimizes the within group sum-of-squares: this can be seen as minimizing the information loss caused by the protection process, because the more homogeneous is a group, the less information loss is caused when replacing the group values by their average. In Mateo-Sanz and Domingo-Ferrer (1999), introduced data-oriented microaggregation, where the size of group can vary depending on data, with the only constraint that it be $\geqslant k$ (prior proposals required that all groups be of size k). Multivariate microaggregation was described in Mateo-Sanz and Domingo-Ferrer (1998). Optimal solutions to the microaggregation problem and heuristic solutions were the topic of Domingo-Ferrer and Mateo-Sanz (2002).

Another topic that was investigated at that time was delegation of statistical data. Delegation seeks to allow a data user to obtain exact statistical results without disclosing to him the exact value of data (see Domingo-Ferrer and Sánchez del Castillo, 1997 and the patent Domingo-Ferrer and Sánchez del Castillo, 1998).

A specially visible action by CRISES in this period was the organization of the *Statistical Data Protection'98* conference sponsored by Eurostat (Lisbon, Portugal, March 1998). See Domingo-Ferrer and Mateo-Sanz, 1998, Domingo-Ferrer (1999).

The OTTILIE project: In 1999 a one-year project called OTTILIE (*Optimizing the Tradeoff beTween Information Loss and dIsclosurE risk for microdata*) was awarded to the CRISES group and to a researcher at IIIA-CSIC (Institut d'Investigació en Intel·ligència Artificial) by the U. S. Bureau of the Census.

OTTILIE was a twelve-month project that sought to demonstrate a methodology to compare the existing methods for masking (SDC-protecting) microdata. In SDC, there is a tension between the information loss in by application of a particular masking method to an original dataset and the disclosure risk associated to the released masked dataset (probability that masked data may lead to disclosure of individual information). OTTILIE defined procedures to measure information loss and disclosure risk; then these measures were combined to construct an overall score for a masking method. Experi-

ments with a wide range of microdata masking methods were carried out and eventually a ranking of methods by score was produced (Domingo-Ferrer, Mateo-Sanz and Torra, 2001, Domingo-Ferrer and Torra, 2001). For continuous microdata, rank swapping and microaggregation were singled out as particularly well-performing masking methods (a method was defined to perform well if it scored low, i.e. if it could achieve low information loss and disclosure risk at the same time). For categorical microdata, none of the tried methods clearly outperformed the rest.

The European projects CASC and AMRADS: On January 1, 2001, the European projects CASC («Computational Aspects of Statistical Confidentiality», IST-2000-25069) and AMRADS («Accompanying Measure for Research and Development in Statistics», IST-2000-26125) were launched. Both projects span over three years and have connections with SDC.

The CASC project aims at perfecting the ARGUS SDC software (Argus, 1999) by enriching it with more methods for protecting tabular data and microdata. Four Catalan contractors belong to the CASC consortium: IDESCAT, UPC, IIIA-CSIC and URV (the CRISES group in the latter case). Their roles are as follows:

- IDESCAT will essentially take a user role and will provide testing for methods being developed.
- UPC is committed to work on tabular data protection, by exploring new heuristics for the secondary cell suppression problem based on network. See Castro (2001).
- IIIA-CSIC concentrates on microaggregation for categorical variables (Domingo-Ferrer and Torra, 2001) and also on new record linkage methods for empirical assessment of disclosure risk (Torra, 2000).
- CRISES-URV is working on microdata protection, specifically:
  - Implementing a library with all known microaggregation methods, including those proposed by Domingo-Ferrer and Mateo-Sanz (2002), in view of inclusion in the Argus software.
  - Characterizing the complexity of optimal microaggregation. An achievement has been to prove NP-hardness for optimal microaggregation (Oganian and Domingo-Ferrer, 2001).
  - Exploring microdata protection through synthetic data generation. Work is in progress to compare the best masking methods identified in Domingo-Ferrer and Torra, 2001) with new approaches based on synthetic data.
  - Providing mechanisms for multilevel access to microdata. Water-marking techniques are exploited to provide multilevel access to masked data: the higher the clearance of a user, the more masking she can remove (unprivileged users just see masked data, while top privileged users can retrieve the original data, see Domingo-Ferrer, Mateo-Sanz and Sebé, 2001).

Finally, the AMRADS projects intends to disseminate best practices in official statistics. The CRISES group is a subcontractor for this project in charge of organizing the AMRADS Workshop on Statistical Data Confidentiality (Luxembourg, December 13-14, 2001).

References

1. Hundepool, A. and Willenborg, L. (1999). «ARGUS: Software from the SDC project», *Joint ECE/Eurostat Work Session on Statistical Confidentiality*, Thessaloniki, Greece. http://www.unece.org/stats/docments/1999.03.confidentiality.htm
2. Castro, J. (2001). «Using modeling languages for the complementary suppression problem through network models», *2nd Joint ECE/Eurostat Work Session on Statistical Confidentiality*, Skopje, Macedonia. http://www.unece.org/stats/documents/2001/03/confidentiality/24.e.pdf
3. Domingo-Ferrer, J. and Mateo-Sanz, J. M. (1996). «On the security of cell suppression in contingency tables with quantitative factors», in *Proceedings of the 3rd International Seminar on Statistical Confidentiality*, Ljubljana: Eurostat-Statistical Office of the Republic of Slovenia, 208-217.
4. Domingo-Ferrer, J. and Sánchez del Castillo, R. X. (1997). «An implementable scheme for secure delegation of statistical data», a *Information Security-ICICS'97* (Lecture Notes in Computer Science 1334), eds. Y. Han, T. Okamoto and S. Qing, Berlin: Springer-Verlag, 445-451.
5. Mateo-Sanz, J. M. and Domingo-Ferrer, J. (1998). «A comparative study of microaggregation methods», *Qüestiió*, 22, 511-526.
6. Domingo-Ferrer, J. and Sánchez del Castillo, R. X. (2000). «A method for secure delegation of statistical data», Spanish patent P9800608, filed in the *Spanish Boletín Oficial de la Propiedad Industrial*, 16, 4559.
7. Domingo-Ferrer, J. and Mateo-Sanz, J. M. (1998). «Current directions in statistical data protection», *Research in Official Statistics*, 2, 105-112.
8. Domingo-Ferrer, J. and Mateo-Sanz, J. M. (1999). «On resampling for statistical confidentiality in contingency tables», *Computers & Mathematics with Applications*, 38, 13-32.
9. Domingo-Ferrer, J. (ed.) (1999). *Statistical Data Protection*, Luxembourg: Office for Official Publications of the European Communities.
10. Domingo-Ferrer, J., Mateo-Sanz, J. M. and Torra, V. (2001). «Comparing SDC methods for microdata on the basis of information loss and disclosure risk», *Proceedings of ETK-NTTS'2001*, 2, Luxembourg: Office for Official Publications of the European Communities.
11. Domingo-Ferrer, J. and Torra, V. «A quantitative comparison of disclosure control methods for microdata», in *Confidentiality, Disclosure and Data Access*. North-Holland, 113-135 (to appear).

12. Domingo-Ferrer, J., Mateo-Sanz, J. M. and Sebé, F. (2001). «Watermarking for multilevel access to statistical databases», in *IEEE International Conference on Information Technology: Coding and Computing-ITCC'2001*, Piscataway NJ: IEEE Computer Society, 243-247.
13. Domingo-Ferrer, J. and Torra, V. «Aggregation techniques for statistical confidentiality», in *Aggregation Operators: New Trends and Applications*, (eds. R. Mesiar and T. Calvo), Physica Verlag (to appear).
14. Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). «Practical data-oriented microaggregation for statistical disclosure control», *IEEE Transactions on Knowledge and Data Engineering* (to appear). Preprint available at http://www.computer.org/tkde/STATISTICAL DATABASE.htm.
15. IDESCAT-Statistics Catalonia. (1996). *Sample of 1991 Population Census of Catalonia.*
16. Garín, A. and Ripoll, E. (1999). «Performance of μ-Argus in disclosure control of uniqueness in populations», *Joint ECE/Eurostat Work Session on Statistical Confidentiality*, Thessaloniki, Greece. http://www.unece.org/stats/docments/1999.03.confidentiality.htm
17. Mateo-Sanz, J. M. and Domingo-Ferrer, J. (1999). «A method for data- oriented multivariate microaggregation», in *Statistical Data Protection*, ed. J. Domingo-Ferrer, Luxembourg: Office for Official Publications of the European Communities, 89-99.
18. Oganian, A. and Domingo-Ferrer, J. «On the complexity of optimal microaggregation for statistical disclosure control», *UNECE Statistical Journal* (to appear).
19. Torra, V. (2000). «Towards the re-identification of individuals in data files with non-common variables», in *European Conference on Artificial Intelligence*, Amsterdam: IOS Press.
20. Turmo, J. (1993). «Pertorbacions aleatòries amb compensació: una tècnica per a la protecció d'informació estadística confidencial», *Qüestiió*, 17, 413-435.

APPLICATION OF DATA ANALYSIS TECHNIQUES IN THE BEHAVIOURAL SCIENCES

Montserrat Freixa Blanxart
Departament Metodologia de les Ciències del Comportament
Facultat de Psicologia
Universitat de Barcelona

A large number of different definitions are grouped under the umbrella of the so-called behavioural sciences. Since the end of the 1980s a stable working group has formed itself within the Department of Methodology of the Behavioural Sciences at the University of Barcelona. The group is primarily concerned with the study, analysis, dissemination and application of various techniques of statistical analysis in psychology, a field of enquiry understood in the broadest of terms. In general, a range of research interests and techniques have been adopted which, in brief, have centred on both single factor and multivariate exploration and estimates.

Exploratory data analysis: This is primarily concerned with the characterisation of the various possibilities afforded by exploratory techniques. Specifically, it deals with techniques based in the study of object rankings which are unaffected by anomalies in the distribution (Freixa *et al.*, 1992). The impact of «robust» techniques has been used in various situations, although the most interesting are those linked with multivariate phenomena

Analytical techniques in epidemiological studies: Here we are concerned with the application of the techniques of epidemiology (Freixa *et al.*, 1998) to psychological phenomena, in particular those of a clinical nature. The difficulties encountered have been analyzed in a number of studies. These have examined the difficulties of diagnosis, the prediction of results; specific statistical indicators or the estimation of bias (Peró and Guàrdia, 2000; Freixa *et al.*, 2000; Peró and Guàrdia, 2002*a*; Peró and Guàrdia, 2002*b*).

Multivariate techniques and their application in psychology: The application of multivariate techniques is an ever present matter in psychology today. A multivariate approach to psychological phenomena remains essential given their complexity. Within a strictly methodological line of investigation, various possibilities have been analysed linked, for example, with research design (Guàrdia and Arnau, 1990; Solanas, Salafranca and Guàrdia, 1992); the associated covariance structure (Hernández, San Luis and Guàrdia, 1995; 1996) and with the metrics of the variables. In the applied field, various phenomena have been modelled by applying this perspective. For example, in the study of Environmental Psychology (Valera *et al.*, 1998; Valera, Guàrdia and Pol, 1999; Guàrdia and Pol, 2002; Valera and Guàrdia, 2002); the application of models of confirmatory factor analysis (Adan and Guàrdia, 1993, 1997; Guàrdia and Adan, 1997;

Marcet *et al.*, 2000); in the field of cognitive impact on neuropsychological disorders (Peña, Guàrdia and Jarne, 1991; Martinez, Guàrdia and Peña, 1996); in road safety, in diagnostic interviews and in the study of posttraumatic stress syndrome (Jarne and Guàrdia, 2000).

Statistical analysis of functional magnetic resonance images: In recent times the analysis of the undertaking of simple tasks has been predominantly carried out using functional magnetic resonance imaging. This approach requires images of the brain and the associated study of the activation zones and their statistical significance. This analytical procedure is somewhat controversial given the considerable difficulties inherent to the specificity of the signal and its distortion in relation to the system's noise (Mañeru *et als.*, 2001).

References

1. Adan, A. and Guàrdia, J. (1993). «Circadian variations of self-reported activation: A multidimensional approach». *Chronobiologia*, 20, 233-244.
2. Adan, A. and Guàrdia, J. (1997). «Efectos de la hora del día y la personalidad en la activación autoevaluada». *Psicothema*, 9, 133-143.
3. Aznar, J. A., Amador, J. A., Freixa, M. and Turbany, J. (2000). «Consumo atencional en la estimación de la profundidad retrovisual». *Psicothema*, 12, 71-78.
4. Freixa, M., Guàrdia, J., Honrubia, M. L. and Peró, M. (1998). *Estudis Epidemiològics*. Barcelona: Edicions Universitat de Barcelona.
5. Freixa, M., Salafranca, Ll., Guàrdia, J., Ferrer, R. and Turbany, J. (1992). *Nuevas técnicas estadísticas: Análisis Exploratorio de datos*. Barcelona: PPU.
6. Freixa, M., Guàrdia, J., Honrubia, M. L. and Peró, M. (2000). «Estimación de la prevalencia a partir de los métodos de captura-recaptura». *Psicothema*, 12, 231-235.
7. Guàrdia, J. and Adan, A. (1997). «Confirmatory factor analysis applied to Matthews adjective checklist of self-reported activation: effect of time of day». *Quality & Quantity*, 31, 95-106.
8. Guàrdia, J. and Arnau, J. (1990). «Diseños longitudinales en panel: alternativa de análisis de datos mediante los sistemas de ecuaciones estructurales». *Psicothema*, 2, 57-71.
9. Guàrdia, J. and Pol, E. (2002). «A Critical study of theoretical models of Sustainability using Structural Equation Systems». *Enviromental and Behavior*, in press.
10. Hernández, J. A., San Luis, C. and Guàrdia, J. (1995). «Acerca de la robustez de los estimadores multinormales y elípticos bajo ciertas condiciones de asimetría, tamaño muestral y complejidad de los modelos de estructura de covarianza». *Anales de Psicología*, 11, 203-217.
11. Hernández, J. A., San Luis, C. and Guàrdia, J. (1996). «Aplicación de los modelos de ecuaciones estructurales en los diseños en panel: Introducción». In Arnau, J.

(coord.), *Métodos y técnicas avanzadas de análisis de datos en ciencias del comportamiento*, 207-223. Barcelona: Edicions Universitat de Barcelona.

12. Jarne, A. and Guàrdia, J. (2000). «The relationship of the degree of exposure to a technological disaster and emotional response: A structural model approach». *Quality & Quantity*, 35, 161-171.
13. Mañeru, C., Junqué, C., Bargalló, N., Olondo, M., Botet-Mussons, F., Tallada, M., Guàrdia, J. and Mercader, J. M. (2001). «H-MR Spectroscopy is sensitive to subtle effects of Perinatal Asphyxia». *Neurology*, 57, 1115-1118.
14. Marcet, C., Guàrdia, J., Almirall, H. and Lorenzo, U. (2000). «Factorial Structure of the Spanish version of the Dots-R questionnaire». *European Journal of Psychological Assessment*, 16, 59-65.
15. Martínez, J. A., Guàrdia, J. and Peña, J. (1996). «Validación de las subpruebas del Test Barcelona relacionadas con subtest de la escala de inteligencia de Wechsler para adultos». *Neuropsychologya Latina*, 2, 10-14.
16. Peña, J., Jarne, A. and Guàrdia, J. (1991). «Programa Integrado de Exploración Neuropsicológica-Test Barcelona: Validez de Contenidos». *Revista de Logopedia, Fonologia y Audiologia*, 11, 96-107.
17. Peró, M. and Guàrdia, J. (2000). «Efecto del sesgo de desclasificación no diferencial e independiente sobre el trastorno, la exposición y una tercera variable». *Psicothema*, 12, 447-452.
18. Peró, M. and Guàrdia, J. (2002*a*). «Indicadores epidemiológicos descriptivos para el estudio de trastornos que cursan con recaidas». *Psicothema*, 14, in press.
19. Peró, M. and Guàrdia, J. (2002*b*). «The rate ratio for the study of disorders with multiple episodes». *Quality & Quantity*, in press.
20. Solanas, A., Salafranca, L. and Guàrdia, J. (1992). «Análisis estadístico de diseños conductuales: Estadístico β_n». *Psicothema*, 4, 253-259.
21. Valera, S., Guàrdia, J. and Pol, E. (1999). «A study of the symbolic aspects of space using non-quantitative techniques of analysis». *Quality & Quantity*, 32, 367-381.
22. Valera, S. and Guàrdia, J. (2002). «Social Identity and Sustainability: Barcelona's Olympic Village». *Environmental and Behavior*, in press.
23. Valera, S., Guàrdia, J., Cruells, E., Paricio, A., Pol, E., Reixach, N., Schilman, N. and Vallés, N. (1998). «Environment and Urban Social Identity. A case study: BCN's Olympic Village». In J. Tekkenburg, J. Van Andel, J. Smeets & A. Seidel (eds.). *Shifting Balances. Changing Roles in Policy, Research and Design*, 222-223. Eindhoven: EIRASS.

THE GRASS GROUP (Grup de Recerca en Anàlisi eStadística de la Supervivència)

Guadalupe Gómez
Departament d'Estadística i Investigació Operativa
Facultat de Matemàtiques
Universitat Politècnica de Catalunya

This group was formally born during a SEIO meeting in Sevilla in 1995. Its aim is to bring together researchers from Catalunya and abroad to collaborate in the theoretical developments, as well as in the applied methodologies, of the analysis of survival data. Even before its conception, the group had been actively collaborating with Professor Stephen W. Lagakos from the Biostatistics Department of the Harvard School of Public Health. During these years, our applied work, as well as some of the methods developed, have been based in problems proposed by clinicians and epidemiologists from the Institut Municipal de la Salut (IMS), from the Institut Municipal d'Investigacions Mèdiques (IMIM), from the Hospital Germans Trias i Pujol (Can Ruti), as well as from the AIDS Clinical Trial Group (ACTG) from Harvard University.

The main topics in our methodological research are the analysis of interval-censored data and the study of missing data problems. The first has been approached using frequentist and Bayesian methods, while the second has benefit from parametric as well as semiparametric developments. Among others, and as a consequence of the collaboration with the above mentioned Catalan groups, several papers have been published where survival analysis techniques have been applied to several complex scenarios such as the modelling of breast cancer survival as a function of the elapsed time between symptom and treatment (Gómez *et al.*, 1996), the study of the impact of the functional capacity into survival in the elderly population of Barcelona (Lamarca *et al.*, 1998) and the short term survival of individuals with tuberculosis who are infected with HIV (Falqués *et al.*, 1999).

Analysis in survival/sacrifice experiments: The analysis of the distribution of the time-to-tumor in experiments with rodents where several sacrifice times are scheduled has been investigated. The use of a four-variate counting process imbedded in their corresponding martingale framework is used to derive a new estimator and to study its asymptotic properties (Gómez and Julià, 1990; Gómez and Van Ryzin, 1992).

Properties for left-censored data: A theoretical study is undertaken for the asymptotic properties of the left-censored Kaplan-Meier estimator derived as the solution of a backward Dóleans equation (Gómez, Julià and Utzet, 1992; Gómez, Julià and Utzet, 1994).

Self-consistency approaches for interval-censored data: The self-consistency concept was first developed by Efron (1967) and it represented a different way of deriving the Kaplan-Meier estimator for the survival function. Turnbull used this same idea to estimate the distribution function from a sample containing interval-censored observations. Based on that, an algorithm is derived which is appropriate for other censoring mechanisms. We have developed this idea when interest lies in the elapsed latency time from an interval-censored origin to a right-censored end-point (double-censored data). The method has been applied to a cohort of hemophiliacs that became infected with HIV in the early 80's and subsequently developed AIDS (Gómez, 1992; Gómez and Lagakos, 1994; Gómez and Calle, 1999). These ideas are further developed and extended to estimate the parameters in a linear regression model when the covariate is interval-censored. The methodology is applied to the analysis of the viral load baseline history of an HIV+ group of patients that were subsequently randomized to six different therapies (Gómez, Espinal and Lagakos, 2001).

Nonparametric Bayesian analysis: We approach the estimation of the survival function based on interval-censored data from a nonparametric Bayesian point of view. We propose a methodology that accommodates the theory for right-censored data based on Dirichlet processes to an interval-censoring scheme by using Markov Chain Monte Carlo methods. We apply this methodology to analyze the risk of HIV infection in a cohort of injecting drug users in Barcelona. The nonparametric Bayesian approach allows, not only the incorporation of prior believes about the survival function, but also, the analysis of the risk of seroconversion without assuming restrictive parametric models. Furthermore, the estimator for the distribution function is smooth and thus differences between groups can be easily interpreted. (Gómez *et al.*, 2000; Calle and Gómez, 2001*a*). We extend these ideas to the analysis of regression models when one of the covariates is interval-censored. We encountered this situation in an AIDS clinical trial where the goal was to assess the association between duration of viral load suppression on a failed regimen and subsequent viral load. A completely parametric approach to this problem is, in general, not appropriate because it requires a model for the interval-censored covariate which cannot be validated. We propose the use of a mixture of Dirichlet processes. This mixture enables us to specify parametrically every component in the model except the distribution of the interval-censored covariate, which is treated nonparametrically. We develop in detail the proposed methodology for the linear regression model (Calle and Gómez, 2001*b*).

Survival analysis with missing covariates: In this topic the goal is to estimate the survival when part of the covariates of interest are missing. The main statistical problem is, on one hand, that any complete-case based inference is potentially biased and, on the other, in general it is not possible to assess the ignorability of the non-response mechanism, nor to test the assumptions on the distributions. Under a missing at random non-response pattern, we have developed some imputation techniques in order to complete the unobserved subsample and to apply a standard methodology. In the most

general framework, we have adapted maximum likelihood based strategies to evaluate the impact of the model assumptions on the resulting estimates (Serrat *et al.*, 1998; Gómez and Serrat, 1999).

Semiparametric methods for missing covariates: We develop semiparametric strategies to deal with missing covariates in the context of a survival analysis study. We define a grouped Kaplan-Meier estimator as a discrete time version of the usual Kaplan-Meier estimator, when the covariates of interest are categorical and completely observed. We also developed inverse probability weighted generalized estimating equations. We use them to estimate the proportion of individuals at risk/censored/events in each one of the categories, in presence of missing data. The resulting estimator for the survival function is asymptotically unbiased and normal distributed. Its properties for finite samples have been also illustrated by simulation. To summarize the resulting inferences we propose a sensitivity analysis perspective on a rank of plausible values for the non-ignorability parameters in the non-response pattern. We have applied this methodology to the analysis of a cohort of pulmonary tuberculosis HIV-infected patients with a large proportion of missingness in the main predictors –CD4+ lymphocytes count and tuberculin skin test– (Serrat and Gómez, 2001*a*, 2001*b*).

References

1. Berger, A., Gómez, G. and Wallenstein, S. (1988). «A Homogeneity Test for Follow-Up Studies». *IMA Journal of Mathematics Applied in Medicine and Biology*, 5, 101-112.
2. Calle, M. L. and Gómez, G. (2001*a*). «Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods». *Journal of Statistical Planning and Inference*, 98, 73-87.
3. Calle, M. L. and Gómez, G. (2001*b*). «Semiparametric Bayesian analysis of regression models with an interval-censored covariate». *Technical Report, 2001/04, Department of Statistics and Operations Research.* Universitat Politècnica de Catalunya.
4. Falqués, M., Langohr, K., Gómez, G., Garcia de Olalla, P., Jansà, J. M. and Caylà, J. (1999). «Supervivencia en pacientes con tuberculosis infectados con VIH. Estudio de los fallecimientos en los primeros nueve meses de tratamiento». *Revista Española de Salud Pública*, 73, 549-562.
5. Gómez, G. (1992). «Estimation of Induction Distribution with Doubly Censored Data and Application to AIDS». *Teoría de la Probabilidad y sus Aplicaciones*, 37, 36-45 (published version is in russian).
6. Gómez, G. and Calle, M. L. (1999). «Nonparametric Estimation with Doubly Censored Data». *Journal of Applied Statistics*, 26, 45-58.
7. Gómez, G., Calle, M. L., Egea, J. M. and Muga, R. (2000). «Estimation of the risk of HIV infection as a function of the length of intravenous drug use. A nonparametric Bayesian approach». *Statistics in Medicine*, 19, 2641-2656.

8. Gómez, G., Espinal, A. and Lagakos, S. (2001). «Inference for a linear regression model with an interval-censored covariate». *Technical Report, 2000/18, Department of Statistics and Operations Research.* Universitat Politècnica de Catalunya.
9. Gómez, G. and Julià, O. (1990). «Estimation and Asymptotic Properties of the Distribution of Time-to-Tumor in Carcinogenesis Experiments». *IMA Journal of Mathematics Applied in Medicine and Biology*, 7, 109-123.
10. Gómez, G., Julià, O. and Utzet, F. (1992). «Survival Analysis for Left-Censored Data in Survival Analysis: State of the Art (editors: J. P. Klein and P. Goel): 269-288, Kluwer Academic Publishers, Dordrecht, Holanda.
11. Gómez, G., Julià, O. and Utzet, F. (1994). «Asymptotic Properties of the Left Kaplan-Meier Estimator». *Communications in Statistics: Theory and Methods*, 23, 123-135.
12. Gómez, G. and Lagakos, S. (1994). «Estimation of the Infection Time and Latency Distribution of AIDS with Doubly Censored Data». *Biometrics*, 50, 204-212.
13. Gómez., G., Porta, M., Griful, E., Maguire, A., Calle, M. L., Malats, N., Fernández, E., Piñol, J. L. and Gallén, M. (1996). «Modelling breast cancer survival and the symptom-to-treatment interval». *Journal of Epidemiology and Biostatistics*, 1, 175-182.
14. Gómez, G. and Serrat, C. (1999). «Estudios de supervivencia con datos no observados. Dificultades inherentes al enfoque paramétrico». *Qüestiió*, 23, 365-392.
15. Gómez, G. and Van Ryzin, J. (1992). «Estimation of the Subsurvival Function for Time-to-Tumor in Survival/Sacrifice Experiments». *Statistics and Probability Letters*, 13, 5-13.
16. Lamarca, R., Alonso, J., Gómez, G. and Muñoz, A. (1998). «Left-Truncated Data with Age as Time Scale: An Alternative for Survival Analysis in the Elderly». *Journal of Gerontology: Medical Sciences*, 53A, M337-M343.
17. Serrat, C. and Gómez, G. (2001*a*). «Estimating the stratified survival with missing covariates: I. A semiparametric approach». *Technical Report, 2001/12, Department of Statistics and Operations Research.* Universitat Politècnica de Catalunya.
18. Serrat, C. and Gómez, G. (2001*b*). «Estimating the stratified survival with missing covariates: II. A simulation study». *Technical Report, 2001/13, Department of Statistics and Operations Research.* Universitat Politècnica de Catalunya.
19. Serrat, C., Gómez, G., García, P. and Caylà, J. (1998). «CD4+ lymphocytes and tuberculin skin test as survival predictors in pulmonary tuberculosis HIV-infected patients». *International Journal of Epidemiology*, 27, 703-712.

MEASUREMENT INVARIANCE RESEARCH GROUP

Juana Gómez Benito
Departament Metodologia de les Ciències del Comportament
Facultat de Psicologia
Universitat de Barcelona

One of the largest applied areas of statistical research within the behavioural sciences is undoubtedly that of psychometry, in which instruments are developed for measuring psychological traits. It is here that one finds the research group «Studies of measurement instrument invariance», whose perspective is that the appropriateness of psychological diagnoses depends directly upon the quality of the measures provided by such instruments and their fairness with respect to the different groups with which they are used.

Psychometric techniques: The measurement instrument must be a reliable and valid indicator of the trait being evaluated; the requirements which have to be met for this to be so are analysed at a theoretical level in Gómez (1989, 1996), particular emphasis being placed on the contribution which covariance structure models can make in terms of a more complex analysis of test reliability and validity. On the basis of this, new measurement instruments (Castro *et al.*, 1997; Puyuelo *et al.*, 1998) and computer programs (Martínez and Renom, 1993; Puyuelo *et al.*, 2002) are developed (Castro *et al.*, 1997), the quality of measurement provided by various existing scales is analysed (Forns and Gómez, 1990, 1994; Gómez and Forns, 1993*a*, 1993*b*; Stock, Okun and Gómez, 1994) and the equivalence of measures with respect to original versions is considered (Balluerka and Gómez, 2000; Balluerka *et al.*, 2000; Maydeu *et al.*, 2000).

Analysing the effectiveness of techniques for detecting DIF: The possible lack of test fairness with respect to certain variables (demographic, ethnic, cultural, etc.) would undermine validity. Differential item functioning (DIF) analysis aims to detect those items which may function differently for different groups, favouring some and being to the detriment of others. In order for such analysis to be accurate it is essential to use statistical techniques that offer higher rates of correct identifications and fewer classification errors. Simulation studies, in which various conditions, such as the amount of DIF, type of DIF, percentage of items with DIF, distribution of group ability, and sample size, are manipulated, can be used to compare the efficacy of the most widely used techniques for detecting DIF (Hidalgo and Gómez, 2000; Gómez and Navas, 2000) and the degree to which this efficacy may be optimized by applying iterative purification procedures to the trait measure (Navas and Gómez, 2001).

Bias of items and/or tests: Research in this area aims to elucidate if the differences found between groups reflect different levels of the trait being measured (impact)

or whether they are caused by systematic sources of variation unrelated to the trait (bias). In order to achieve this, the subgroups of items which are invariant with respect to possible bias variables must be identified across various instruments, as it is only these items which enable the measure to have the same meaning for all subjects. To this end, invariance studies are carried out at different levels: developmental (Gómez and Forns, 1996), linguistic (Ferreres, González and Gómez, 2000), gender (Gómez and Navas, 1998) and questionnaire translation (Tomás, González and Gómez, 2000). In terms of techniques the following are applied: logistic regression, loglinear models, Mantel-Haenszel statistic, item response theory and confirmatory factor analysis. All these form part of what are known as conditional invariance methods, which match trait levels within groups so that they are comparable, thus enabling impact to be distinguished from bias.

References

1. Balluerka, N. and Gómez, J. (2000). «Comparación entre los resultados obtenidos en la escala TDA-H (Trastorno por Déficit de Atención con Hiperactividad) en una muestra americana y en una muestra española de adultos». *Psicothema*, 12, 64-68.
2. Balluerka, N., Gómez, J., Stock, W. and Caterino, L. (2000). «Características psicométricas de las versiones americana y española de la escala TDA-H (Trastorno por Déficit de Atención con Hiperactividad): un estudio comparativo». *Psicothema*, 12, 629-634.
3. Castro, J., de Pablo, J., Gómez, J., Arrindell, W. A. and Toro, J. (1997). «Assessing rearing behaviour from the perspective of the parents: a new form of the EMBU». *Social Psychiatry and Psychiatric Epidemiology*, 32, 230-235.
4. Ferreres Traver, D., González Romá, V. and Gómez Benito, J. (2000). «Comparación del estadístico Mantel-Haenszel y la regresión logística en el funcionamiento diferencial de los ítems en dos pruebas de aptitud intelectual en un contexto bilingüe». *Psicothema*, 12, 214-219.
5. Forns Santacana, M. and Gómez Benito, J. (1990). «Factor Structure of the McCarthy Scales». *Psychology in the Schools*, 27, 111-115.
6. Forns-Santacana, M. and Gómez-Benito, J. (1994). «The cognitive, linguistic and adaptative development and academic achievement of pre-school children within the catalan immersion programme». In C. Laurén (ed.) *Evaluating Immersion Programs*. Vaasa: Tutkimuksia.
7. Gómez Benito, J. (1989). «Estimates of the internal consistency of a factorially complex composite». *Educational and Psychological Measurement*, 49, 571-578.
8. Gómez Benito, J. (1996). «Aportaciones de los modelos de estructuras de covariancia al análisis psicométrico». In J. Muñiz (ed.) *Psicometría*. Madrid: Universitas.

9. Gómez Benito, J. and Forns Santacana, M. (1993*a*). «Concurrent validity between the Columbia Mental Maturity Scales». *Perceptual and Motors Skills*, 76, 1177-1178.
10. Gómez Benito, J. and Forns Santacana, M. (1993*b*). «Confirmatory Factor Analysis of the McCarthy Scales». In R. Steyer, K. F. Wender and K. F. Widaman (eds.). *Psychometric Methodology*. Stuttgart: Gustav Fisher Verlag.
11. Gómez-Benito, J. and Forns-Santacana, M. (1996). «Factor structure of the McCarthy Scales in 7 years old Spanish Children». *Psychology in the Schools*, 33, 231-238.
12. Gómez, J. and Navas, M. J. (1998). «Impacto y funcionamiento diferencial de los ítems respecto al género en una prueba de aptitud numérica». *Psicothema*, 10, 717-728.
13. Gómez, J. and Navas, M. J. (2000). «A comparison of chi-square, RFA and IRT based procedures in the detection of DIF». *Quality & Quantity*, 34, 17-31.
14. Hidalgo Montesinos, M.ª. D. and Gómez Benito, J. (2000). «Comparación de la eficacia de regresión logística politómica y análisis discriminante logístico en la detección del DIF no uniforme». *Psicothema*, 12, 298-300.
15. Martínez, N. and Renom, J. (1993). *DEMOTAC: programa de ordenador para tests adaptativos computerizados*. Barcelona: PPU.
16. Maydeu-Olivares, A., Rodríguez-Fornells, A., Gómez-Benito, J. and D'Zurilla, T. J. (2000). «Psychometric properties of the Spanish adaptation of the Social Problem-Solving Inventory-Revised (SPSI-R)». *Personality and Individual Differences*, 29, 699-708.
17. Navas Ara, M. J. and Gómez Benito, J. (2001). «Effects of ability scale purification on identification of DIF». *European Journal of Psychological Assessment*.
18. Puyuelo, M., Renom, J., Solanas, A. and Wiig, E. (2002). *Evaluación del lenguaje: BLOC Screening. Manual de usuario*. Barcelona: Masson, S.A.
19. Puyuelo, M., Wiig, E., Renom, J. and Solanas, A. (1998). *Batería del lenguaje objetiva y criterial: Manual de usuario*. Barcelona: Masson, S.A.
20. Stock, W. A., Okun, M. A. and Gómez Benito, J. (1994). «Subjective well-being measures: reliability and validity among spanish elders». *International Journal of Aging and Human Development*, 38, 221-235.
21. Tomás Marco, I., González-Romá, V. and Gómez Benito, J. (2000). «Teoría de respuesta al ítem y análisis factorial confirmatorio: dos métodos para analizar la equivalencia psicométrica en la traducción de cuestionarios». *Psicothema*, 12, 540-544.

THE RISK IN FINANCE AND INSURANCE RESEARCH GROUP

Montserrat Guillén
Departament d'Econometria, Estadística i Economia Espanyola
Facultat de Ciències Econòmiques i Empresarials
Universitat de Barcelona

The activity of the research group «Risk in Finance and Insurance» is concentrated on studying measurement, control and evaluation of institutional risk and insurance risk. The group studies the value and risk of an investment portfolio, and all the other hazards derived from the choice of customers. It was established during the 90s and nowadays it is formed by ten researchers at the University of Barcelona. It is a multidisciplinary group, working in the field of Actuarial Statistics, Insurance Econometrics, Economic Theory and Financial Mathematics.

The econometric methodology is used to evaluate the risk undertaken by a company. We study methods to improve pricing systems and interest rate for credit products (Dionne, Artís and Guillén, 1996, Guillén and Soldevilla, 1996 and Soldevilla and Guillén, 1997). We also develop modeling techniques to detect adverse selection, to control fraud and, finally, to deal with the presence of the asymmetric information in the insurance contract (Artís, Ayuso, and Guillén, 1999, Artís, Ayuso and Guillén, 2002, Ayuso and Guillén, 1999 and Pinquet, Guillén and Bolance, 2002). In this context, the group also focuses on the study of risk in life insurance products and pension schemes (Betzuen, Felipe and Guillén, 1997 and Borras, Guillén, Sánchez, Junca and Vicente, 1999).

Another research area is the choice of a globally optimal portfolio (García Mínguez, 1988 and Oliva Furés and García Mínguez, 1995). We analyze the temporary structure of interest rate and establish how to design the strategies for immunization (Alegre and Fontanals, 1993).

Currently the group is involved in two research projects, SEC2001-3672 «Long Term Care Insurance in Spain: Longevity and Demand for Benefits» and SEC2001-2581-C02-02 «Observatory of Judicial Culture». The first project analyses how the rising longevity of the Spanish population and the changes in social patterns may affect the future demand of long-term health care for the elderly population. Due to disability or a prolonged illness, long-term care is the assistance given when a person is unable to provide for himself or equivalently, when the insured is unable to perform (ADL) activities of daily living. The analysis of longevity together with the study of disabilities and impairments of the elderly population is an outstanding issue in developed countries (Felipe and Guillén, 1998, Felipe, Guillén and Artís, 1998 and Felipe, Guillén and Nielsen, 2001). Statistical data currently available in Spain allow for the prediction and perspectives of both longevity and the demand for health care. In addition the group is

concerned with the pricing of private insurance covering this risk (Séculi, Fusté, Brugulat, Juncá, Rué and Guillén, 2001).

The second research project in which the group is participating aims to study the decision making process of judges in court decisions and to shed some light on the legal and professional problems of novel magistrates. The economic costs of judicial verdicts in insurance claims compensations are specially considered.

The main focus of the research group is mostly on the application of statistical tools. Nevertheless, the adaptation of existing approaches to the kind of problem and data under study often require a great deal of innovation in the methodology and a deep expertise in the use of statistical techniques. Our main field of application usually requires generalised linear models, survival analysis, nonparametric methods and general econometrics (Bolancé, Guillén and Nielsen, 2000, García Minguez and Sáez, 1992, García Minguez and Sáez, 1993, Guillén and Martin, 1996, Guillén, Junca, Rue and Aragay, 2000, Lin and Guillén, 1998, Salsas, Guillén and Alemany, 1999 and Upton and Guillén, 1995).

References

1. Alegre, A. and Fontanals, H. (1993). «Análisis financiero de un empréstito». *Cuadernos de Economía*, 21, 165-188.
2. Artís, M., Ayuso, M. and Guillén, M. (1999). «Modelling different types of automobile insurance fraud behaviour in the Spanish market». *Insurance: Mathematics & Economics*, 24, 67-81.
3. Artís, M., Ayuso, M. and Guillén, M. (2002). «Detection of automobile insurance fraud with discrete choice models and misclassified claims». *Journal of Risk and Insurance, 69*, 325-340.
4. Ayuso, M. and Guillén, M. (1999). «Modelos de deteccion de fraude en el seguro de automovil». *Cuadernos Actuariales*, 8, 135-149.
5. Betzuen, A., Felipe, A. and Guillén, M. (1997). «Modelos de tablas de mortalidad en España y situación actual». *Anales del Instituto de Actuarios Españoles*, 3, 79-104.
6. Bolancé, C., Guillén, M. and Nielsen, J. P. (2000). «Kernel density estimation of actuarial loss functions». Working paper series of the University of Aarhus Business School (WP-D-00-4). Forthcoming in *Insurance: Mathematics & Economics*.
7. Borras, J. M., Guillén, M., Sánchez, V., Junca, S. and Vicente, R. (1999). «Educational level, voluntary private health insurance and oppportunistic cancer screening among women in Catalonia (Spain)», *European Journal of Cancer Prevention*, 8, 427-434.
8. Dionne, G., Artís, M. and Guillén, M. (1996). «Count data models for a credit scoring system». *Journal of Empirical Finance*, 3, 303-325.

9. Felipe, A. and Guillén, M. (1998). *Evolución y predicción de tablas de mortalidad para la población española*, Fundación Mapfre Estudios. Madrid.
10. Felipe, A., Guillén, M. and Artís, M. (1998). «Recent Mortality Patterns in the Spanish Population». *Transactions of the 26th International Congress of Actuaries*, 9, 250-269. Brimingham, Reino Unido.
11. Felipe, A., Guillén, M. and Nielsen, J. P. (2001). «Longevity studies based on kernel hazard estimation». *Insurance: Mathematics & Economics*, 28, 2, 191-204.
12. García Mínguez, P. (1988). «Shifting Shares: Un Comentario». *Cuadernos de Economía*, 16, 315-322.
13. García Minguez, P. and Sáez, M. (1992). «Advances in integration tests: the PPP hypothesis in the Spanish case». *Estudios económicos*, 178-189.
14. García Minguez, P. and Sáez, M. (1993). «Nuevos contrastes de integrabilidad y la hipótesis del PPP». *Estudios de Economía Aplicada*, 63-76.
15. Guillén, M. and Martin, X. (1996). «Estimació de la variancia mostral a l'enquesta de població activa». *Qüestiió*, 20, 259-272.
16. Guillén, M. and Soldevilla, C. (1996). «On the performance of backpropagation networks in econometric analysis». *INFORMATICA*, 20, 435-441.
17. Guillén, M., Junca, S., Rue, M. and Aragay, J. M. (2000). «Efecto del diseño muestral en el analisis de encuestas de diseño complejo. Aplicación a la Encuesta de Salud de Cataluña». *Gaceta Sanitaria*, 14, 5, 399-402.
18. Lin, T. and Guillén, M. (1998). «The rising hazards of party incumbency: a discrete renewal analysis». *Political Analysis*, 7, 31-59.
19. Oliva Furés, M. and García Mínguez, P. (1995). «Una nota sobre 'Valoración de los derechos de suscripción utilizando la teoría de opciones'». *Revista Española de Economía*, 12, 399-402.
20. Pinquet, J., Guillén, M. and Bolance, C. (2002). «Long-range contagion in automobile insurance data: estimation and implications for experience rating». *Astin Bulletin*, 2, 337-348.
21. Salsas, P., Guillén, M. and Alemany, R. (1999). «Perfect value and outlier detection in logistic binary choice models». *Communications in Statistics. Theory and Methods*, 28, 1447-1460.
22. Séculi, E., Fusté, J., Brugulat, P., Juncá, S., Rué, M. and Guillén, M. (2001). «Percepción del estado de salud en hombres y mujeres en las últimas etapas de la vida». *Gaceta Sanitaria*, 15, 217-223.
23. Soldevilla, C. and Guillén, M. (1997). «Consumer Credit Scoring Using Artificial Neural Networks», in *Intelligent Technologies in Accounting and Finance*, Boson E. (ed.) 117-128.
24. Upton, G. and Guillén, M. (1995). «Perfect cells, direct models and contingency table outliers». *Communications in Statistics: Theory and Methods*, 24, 1843-1862.

STATISTICS IN HEALTH SCIENCES AT THE SCHOOLS OF MEDICINE OF BARCELONA

Víctor Moreno
Dept. de Pediatria, Obstetrícia i Ginecologia i de Medicina Preventiva
Àrea de Medicina Preventiva i de Salut Pública
Universitat Autònoma de Barcelona

The Units of Biostatistics of the schools of medicine of Barcelona have diverse fields of interest that include epidemiology of different diseases, clinical pharmacology and drug development and public health information systems. The most interesting areas of the activity of these groups will be summarized here.

Design and analysis of clinical trials: An important area of work is methodological support for clinical trials, most of them carried out by the pharmaceutical industry. There is a need to keep high standards in the statistical analyses used, since regulatory authorities are severe in these aspects. Factorial designs (Roca-Cusachs *et al.*, 2001) have been applied to optimal dose finding for drug combinations. Many of the trials carried out nowadays have the aim to prove bioequivalence of generic drugs. Methods that account for measurement error such as structural regression models have been applied to the design and analysis of these trials which focus on «proving the null hypothesis» (Carrasco and Jover, 2001, 2002). In this area, two measures of agreement, intraclass correlation coefficient and Lin's concordance correlation, were compared and their potential usefulness was assessed for bioequivalence studies (Carrasco and Jover, 2001). Replicated designs that improve efficiency have been studied and applied to compare the bioavailability of different formulations of a drug (Calvo *et al.*, 1999).

Another area of interest is the application of sequential methods to the design and analysis of clinical trials. The methods based on continuous boundaries such as the double truncated sequential probability ratio test have been applied to a large trial on the efficacy of antiplatellet drugs to prevent mortality after a myocardial infarction (Cruz-Fernandez *et al.*, 2000). For a simple review of sequential methods in clinical trials see Moreno (1995). Similar sequential methods have been compared with inverse sampling in another context. Here the interest was to efficiently design a study to compare the proportion of chromosomal abnormalities estimated from laboratory assays in a sample of individuals with a known reference population proportion. In this study the required sample size of a classical design was compared with that of a modified inverse sampling scheme and a formal sequential method based on triangular continuous boundaries (Moreno *et al.*, 2002).

Another area which involves interesting statistical methods is meta-analysis of published clinical trials. Dose-response models to estimate the effect of morphine in drug addicts based on logistic regression with random effects were applied in Llamas *et al.*,

1994. More recently, generalized linear mixed models have also been applied to study the relative efficacy of different drugs used for detoxification of morphine drug addicts (Farré *et al.*, 2002).

Epidemiology: In the field of epidemiology, methods to estimate risk from the combined or pooled analysis of several cases-control studies have been developed for the situation where one wishes to combine studies with matched and unmatched designs (Moreno *et al.*, 1996). These methods were applied to the study of progression factors of cervical cancer (Moreno *et al.*, 1995). More recently, efforts have been devoted to develop methods to study gene-environment interactions. In a classical case-control study to estimate risk of colorectal cancer related to diet, tumors of the cases were studied to detect mutations in the K-ras oncogene. Models based on polytomous logistic regression were used to simultaneously estimate the risk for mutated and wild-type tumors compared to the same group of general population healthy controls (Bautista *et al.*, 1997).

Much work has also been done in the field of descriptive epidemiology. Analysis of temporal trends of cancer incidence has been studied in detail for cervical cancer using age-period-cohort models (Vizcaino *et al.*, 1998; 2000). A recent review of these methods has been published (González *et al.*, 2002). Models with joinpoint regression have been used to show the recent decline in trends for cancer mortality in Catalonia (Fernández *et al.*, 2001). Life time cumulative risk of cancer incidence and mortality has been estimated (Moreno *et al.*, 1998). Also extrapolation of cancer incidence to the complete Spanish territory from areas with cancer registries has been done using generalized linear mixed models to account for spatial heterogeneity. In these models the log-ratio of cancer incidence to mortality was modelled as a linear function of age (using smoothing splines), period and province of residence. Bayesian methods of estimation were used (Moreno *et al.*, 2001). A related subject of interest is the use of geostatistics to study spatial distribution of disease. In this area much research is ongoing using kriging methods involved in generalized linear mixed models with spatial correlation. These methods are being applied to model the distribution of respiratory diseases (Ascaso and Abellana, 2001*a*), malaria (Ascaso and Abellana, 2001*b*) and cancer incidence around a chemical factory (González *et al.*, 2001). Recurrent events have been studied with generalized linear models using the negative binomial distribution to account for the overdispersion observed in the distribution of events (Navarro *et al.*, 2001).

Clinical epidemiology: This is a heterogeneous area that covers clinical research using epidemiological methods. Multivariate models for prediction of prognosis after cardiac surgery have been developed using stepwise strategies to select the best subset of covariates. Models were validated with the subsample method and compared to other published models regarding calibration and predictive ability (Pons *et al.*, 1997). In the field of diagnosis, a predictive model was developed to diagnose organic dyspepsia using clinical symptoms. Bootstrap validation methods were used in this case (Barenys *et al.*,

2000). Two techniques of computer tomography scans were compared with respect to their ability to diagnose liver metastasis. Comparison of ROC curves was done with jackknife methods (Valls *et al.*, 1998).

Design and analysis of laboratory assays: Experiments in the laboratory often use complex designs that require advanced statistical methods for proper analysis. Generalized linear mixed models have been applied to compare the evolution in time of tumor markers in relation to treatments (Martín-Henao *et al.*, 2000). For the diagnosis of microsatellite instability in tumors, models with a mixture of two or three binomial distributions with covariates were used (González-García *et al.*, 2000).

References

1. Ascaso, C. and Abellana, R. (2001). «Evaluacion geografica de la incidencia de infecciones respiratorias en Manhiça (Mozambique)». *Gaceta Sanitaria 2001*, 15 (supp. 2), 91.
2. Ascaso, C. and Abellana, R. (2001). «Modelizacion espacial del riesgo de malaria mediante aproximaciones bayesianas». *VIII Conferencia Española de Biometria 2001*, 247.
3. Barenys, M., Abad, A., Pons, J. M., Moreno, V., Rota, R., Granados, A., Admetlla, M. and Piqué, J. M. (2000). «Scoring system has better discriminative value than Helicobacter pylori testing in patients with dyspepsia in a setting with high prevalence of infection». *European Journal of Gastroenterology and Hepatology*, 12, 1275-82.
4. Bautista, D., Obrador, A., Moreno, V., Cabeza, E., Canet, R., Benito, E., Bosch, X. and Costa, J. (1997). «KI-RAS mutation modifies the protective effect of dietary monounsaturated fat and calcium on sporadic colorectal cancer». *Cancer Epidemiology Biomarkers and Prevention*, 6, 57-61.
5. Calvo, G., Llinares, J., Antonijoan, R. M., Morros, R., Zsolt, I., Delgadillo, J., Horas, M., Zurita, N. and Jané, F. (1999): «A replicated steady-state design to assess the comparative bioavailability of two formulations of loratadine». *Methods and Findings in Experimental and Clinical Pharmacology*, 21 (suppl B), 64.
6. Carrasco, J. L. and Jover, L. (2001). «Assessing individual bioequivalence using the structural regression model. Statistical Methods in Biopharmacy. Integrating Issues of Efficacy, Safety and Cost-Effectiveness». 4th International Meeting September 24-25, París (France).
7. Carrasco, J. L. and Jover, L. (2002). «The error-in-equation model as approach to assess individual bioequivalence». *XXIth International Biometric Conference 2002*.
8. Carrasco, J. L. and Jover, L. (2002). «Comparison of two measures of agreement: intraclass correlation coefficient and Lin's concordance correlation». *XXIth International Biometric Conference 2002*.

9. Cruz-Fernández, J. M., López-Bescos, L., García-Dorado, D., López García-Aranda, V., Cabades, A., Martin-Jadraque, L., Velasco, J. A., Castro-Beiras, A., Torres, F., Marfil, F. and Navarro, E. (2000). «Randomized comparative trial of triflusal and aspirin following acute myocardial infarction». *European Heart Journal*, 21, 457-6.
10. Farré, M., Mas, A., Torrens, M., Moreno, V. and Cami, J. (2002). «Retention rate and illicit opioid use during methadone maintenance interventions: a meta-analysis». *Drug and Alcohol Dependence*, 65, 283-290.
11. Fernández, E., González, J. R., Borras, J. M., Moreno, V., Sánchez, V. and Peris, M. (2001). «Recent decline in cancer mortality in Catalonia (Spain). A joinpoint regression analysis». *European Journal of Cancer*, 37, 2222-8.
12. González, J. R., Llorca, F. J. and Moreno, V. (2002). Algunos aspectos metodológicos sobre los modelos edad-período-cohorte. Aplicación a las tendencias de mortalidad por cáncer». *Gac. Sanit.*, in press.
13. González, J. R. and Moreno, V. (2001). «Modeling cancer risk around a chemical factory». *22nd Annual Conference The International Society for Clinical Biostatistics.*
14. González-García, I., Moreno, V., Navarro, M., Martí-Ragué, J., Marcuello, E., Benasco, C., Campos, O., Capellà, G. and Peinado, M. A. (2000). «Standardized Approach for Microsatellite Instability Detection in Colorectal Carcinomas». *Journal of the National Cancer Institute*, 92, 544-549.
15. Lamas, X., Farré, M., Moreno, V. and Camí, J. (1994). «Effects of morphine in postaddict humans: a meta-analysis». *Drug and Alcohol Dependence*, 36, 147-152.
16. Martín-Henao, G. A., Quiroga, R., Sureda, A., González, J. R., Moreno, V. and García, J. L. (2000). «Selectin expression is low on CD34 + cells from patients with chronic myeloid leukemia and interferon-α up-regulates this expression». *Hæmatologica*, 85, 139-146.
17. Moreno, V. (1995). «El método secuencial en el ensayo clínico». *Investigación Clínica y Bioética*, 13, 1-4.
18. Moreno, V., Martín, I, Torres, F., Horas, M., Ríos, J. and González, J. R. (2002). «Inverse sampling and triangular sequential designs to compare a small proportion with a reference value». *Qüestiió*, 26, 259-271.
19. Moreno, V., Martín, M., Bosch, F. X., de Sanjosé, S., Torres, F. and Muñoz, N. (1996). «Combined analysis of matched and unmatched case-control studies. Comparison of risk estimates from different studies». *American Journal of Epidemiology*, 143, 293-300.
20. Moreno, V., Muñoz, N., Bosch, F. X., de Sanjosé, S., González, L. C., Tafur, L., Gili, M., Izarzugaza, I., Navarro, C., Vergara, A., Viladiu, P., Ascunce, N. and Shah, K. (1995). «Risk factors for progression of CIN III to invasive cervical cancer». *Cancer Epidemiology Biomarkers & Prevention*, 4, 459-467.

21. Moreno, V., Sánchez, V., Galcerán, J., Borràs, J. M., Borràs, J. and Bosch, F. X. (1998). «Riesgo de enfermar y morir por cáncer en Cataluña». *Medicina Clínica (Barcelona)*, 110, 86-93.
22. Moreno, V., González, J. R., Soler, M., Bosch, F. X., Kogevinas, M. and Borràs, J. M. (2001). «Estimación de la incidencia del cáncer en España: período 1993-1996». *Gaceta Sanitaria*, 15, 380-88.
23. Navaro, A., Utzet, F., Puig, P., Caminal, J. and Martín, M. (2001). «La distribución binomial negativa frente a la de Poisson en el análisis de fenómenos recurrentes». *Gaceta Sanitaria*, 15, 447-52.
24. Pons, J. M. V., Granados, A., Espinas, J. A., Borràs, J. M., Martín, I. and Moreno, V. (1997). «Assessing open heart surgery mortality in catalonia (spain) through a predictive risk model». *European Journal of Cardio-thoracic Surgery*, 11, 415-423
25. Roca-Cusachs, A., Torres, F., Horas, M., Ríos, J., Calvo, G., Delgadillo, J. and Teran, M. (2001). «Spanish Nitrendipine/Enalapril Collaborative Study Group. Nitrendipine and enalapril combination therapy in mild to moderate hypertension: assessment of dose-response relationship by a clinical trial of factorial design». *Journal of Cardiovascular Pharmacology*, 38, 840-9.
26. Valls, C., López, E., Guma, A., Gil, M., Sánchez, A., Andía, E., Serra, J., Moreno, V. and Figueras, J. (1998). «Helical CT *versus* CT arterial portography in the detection of hepatic metastasis of colorectal carcinoma». *American Journal of Roentgenology*, 170, 1341-1347.
27. Vizcaíno, A. P., Moreno, V., Bosch, F. X., Muñox, N., Barros-Dios, X. M. and Parkin, D. M. (1998). «International trends in the incidence of cervical cancer: I. Adenocarcinoma and adenosquamous cell carcinomas». *International Journal of Cancer*, 75, 536-545.
28. Vizcaíno, A. P., Moreno, V., Bosch, F. X., Muñoz, N., Barros-Dios, X. M., Borràs, J. and Parkin, D. M. (2000). «International trends in incidence of cervical cancer: II. Squamous-cell carcinoma». *International Journal of Cancer*, 86, 429-35.

THE SIMULATION AND COMPUTATIONAL STATISTICS RESEARCH GROUP

Jordi Ocaña
Departament d'Estadística
Facultat de Biologia
Universitat de Barcelona

Since 1989, a group of researchers in the department of Statistics of the University de Barcelona has been working in resampling methods and in simulation, especially in applications of simulation techniques in statistics and statistical methodologies based on simulation, and in the application of these techniques in diverse application fields, mainly biology and health sciences.

Simulation methodology in statistics: The main goal of this research is to improve the quality of simulation studies in statistics by means of variance reduction techniques (VRT), better random variate generators, etc. Ocaña and Ruiz-Rivas (1990) propose an efficient random variate generator for the Cuadras-Augé system of bivariate distributions. Vegas and Ocaña (1992) introduce a VRT especially devoted to improve the precision of simulation studies with dichotomic response variables (i.e. studies of the power of a test or the coverage of a confidence interval). This VRT is developed in Ocaña and Vegas (1995). Vegas (1997) and Vegas and Ocaña (2000) illustrate its application a concrete simulation studies in statistics. Vegas, del Castillo and Ocaña (2000) analyze the properties of this VRT from the perspective of exponential models and differential geometry.

Resampling methods: This research line is devoted to the study of some inferential problems posed by the use of some distance and diversity indices in diverse application fields. These inferential problems are approximated by means of resampling techniques (especially the bootstrap) and involve theoretical and simulation studies. Ribó, Ocaña and Prevosti (1989) and Ocaña, Ruiz de Villa and Ribó (1991) are studies on the Levene's Z index of sexual selection. Sánchez, Ocaña and Utzet (1995) and Sánchez, Ocaña, Utzet and Serra (2001) are studies on Prevosti's distance index, widely used in genetics. Pardo, Morales, Salicrú and Menéndez (1997), Salicrú, Vives and Ocaña (2000) and Vives, Salicrú and Ocaña (2001) are studies on entropy and diversity measures.

Statistical and simulation software: Statisticians are frequently also software developers. For example, a new statistical method should be implemented in order to make it usable, or its properties may be studied by simulation. Object orientation is a way to improve the quality and reusability of (statistical) software. Sánchez, Ocaña and Ruiz de Villa (1992) present an object-oriented software oriented to statistical simulation.

Ocaña and Sánchez (1996) is a reflection on the use of object-oriented methodologies in statistical and simulation software.

References

1. Ocaña, J., Ruiz de Villa, M. C. and Ribó, G. (1991). «Bias and efficiency in estimating sexual selection. Levene's Z index and some resampling alternatives». *Heredity*, 67, 95-102.
2. Ocaña, J. and Ruiz-Rivas, C. (1990). «Computer generation and estimation in a one-parameter system of bivariate distributions with specified marginals». *Communications in Statistics. Simulation and Computation*, 19, 37-55.
3. Ocaña, J. and Sánchez, A. (1996). «Object-Oriented Analysis and Design in Statistical Simulation». *Proceedings of the II Workshop on Simulation*. S. Ermakov and V. B. Melas, eds., Pub. House of St. Petersburg University, 15-20.
4. Ocaña, J. and Vegas, E. (1995). «Variance reduction for Bernoulli response variables in simulation». *Computational Statistics and Data Analysis*, 19, 631-640.
5. Pardo, L., Morales, D., Salicrú, M. and Menéndez, M. L. (1997). «Large sample behaviour of entropy measures when the parameters are estimated». *Communications in Statistics. (Theory and Methods)*, 26, 483-501.
6. Ribó, G., Ocaña, J. and Prevosti, A. (1989). «Effect of larval crowding on adult mating behavior in *Drosophila melanogaster*». *Heredity*, 63, 195-202.
7. Salicrú, M., Vives, S. and Ocaña, J. (2000). «Power of some asymptotic tests for maximum entropy». In *Advances in Stochastic Simulation Methods*, N. Balakrishnan, V. B. Melas, S. Ermakov, eds., Birkhauser, Chap. 15, 337-354.
8. Sánchez, A. and Ocaña, J. (1998). «Comparison of L^1 Distances Between Multinomials. Theory and Software Tools». *Proceedings of the III Workshop on Simulation*, S. Ermakov, Y. N. Kashtanov and V. B. Melas, eds., St. Petersburg University Press, 283-287.
9. Sánchez, A., Ocaña, J. and Ruiz de Villa, F. (1992). «An environment for Monte Carlo simulation studies (EMSS)». In *Computational Statistics*. Y. Dodge, J. Whittaker, eds. Physica-Verlag, vol. 2, 195-199.
10. Sánchez, A., Ocaña, J. and Utzet, F. (1995). «Sampling Theory, Estimation and Significance Testing for Prevosti's Estimate of Genetic Distance». *Biometrics*, 51, 1216-1235.
11. Sánchez, A., Ocaña, J., Utzet, F. and Serra, Ll. (2002). «Comparison of Prevosti genetic distances». *Journal of Statistical Planning and Inference*, in press.
12. Vegas, E. (1997). «El problema de Behrens-Fisher en la investigación biomédica. Análisis crítico de un estudio clínico mediante simulación». *Qüestiió*, 21, 293-316.
13. Vegas, E. and Ocaña, J. (1992). «Variance reduction for Bernouilli response variables». In *Computational Statistics*, Y. Dodge, J. Whittaker, eds. Physica-Verlag, vol. 2, 103-107.

14. Vegas, E. and Ocaña, J. (2000). «Variance reduction in the study of a test concerning the Behrens-Fisher problem». *Communications in Statistics-Simulation and Computation*, 29, 463-470.
15. Vegas, E., del Castillo, J. and Ocaña, J. (2000). «Efficiency and exponential models in a variance-reduction technique for dichotomous response variables». *Journal of Statistical Planning and Inference*, 85, 61-74.
16. Vives, S., Salicrú, M. and Ocaña, J. (2001). «Asymptotic confidence regions in (h, ϕ)-entropies». *Proceedings of the IV Workshop on Simulation*, V. B. Melas, S. Ermakov, eds., 492-499.

RESEARCH GROUP ON STOCHASTIC PROCESSES

Marta Sanz
Departament d'Estadística
Facultat de Matemàtiques
Universitat de Barcelona

The research group on Stochastic processes started his activities by the end of the seventies at the Universitat de Barcelona. During the past twenty years, the group has developed into several teams in three universities in Catalonia –Universitat de Barcelona, Universitat Autònoma de Barcelona and Universitat Pompeu Fabra. These teams are being funded by the main research programmes from the Spanish government, the European Union and the Catalan autonomous government. The American Mathematical Society database MathSciNet includes 259 publications authored by members of this group since 1976.

The research interests of the group are in Stochastic Analysis; that is, the study of stochastic processes and their related analytical tools with the purpose of mathematical random modelling of nonsmooth evolution dynamics. Since 1990, the range of topics studied by the group goes from Malliavin calculus and anticipating stochastic calculus, with applications to a broad variety of problems, to stochastic differential and stochastic partial differential equations. For the basis of the first two topics and an overwiew of some of their applications, we refer to the text books (Nualart, 1995 and Nualart, 1998).

Malliavin calculus is an infinite dimensional differential calculus on the Wiener space. It has been introduced by P. Malliavin in a seminal work in 1976 to better understand the interplay between probabilistic and deterministic problems in analysis and differential geometry. The first application of Malliavin calculus has been to derive a criterion to decide whether the law of a random vector is absolutely continuous with respect to Lebesgue measure and if the density is a smooth function. By applying this result to diffusion processes one can ensure hypoellipticity for a class of differential operators.

The paper by Florit and Nualart (1995) provides a new version of the criterium and it is applied to an unsmooth functional of the Wiener process. The papers by Caballero, Fernández and Nualart (1995), Rovira and Sanz-Solé (1996), Rovira and Sanz-Solé (1997), Millet and Sanz-Solé (1999) and Márquez-Carreras, Mellouk and Sarrà (2001) show different examples of new applications of the criterion to anticipative stochastic differential equations and stochastic partial differential equations.

The techniques of Malliavin calculus have led to many research directions. We first mention some applications of one of its main ingredients –the integration by parts formula– to the study of local time in different frameworks (see for instance Nualart

and Wschebor (1991), Imkeller and Nualart (1993) and Imkeller, and Nualart (1994)) or as the startpoint of the development of Poisson-based Malliavin Calculus (see Nualart and Vives (1995)).

In the mid eighties Malliavin calculus has been at the origin of the development of the anticipating stochastic calculus through the work of many researchers –Nualart, Pardoux, Zakai, among others. An extension of this calculus to multiparameter stochastic processes is given in Jolis, and Sanz-Solé (1990), Solé and Utzet (1991), Delgado, and Sanz-Solé (1995) (see also Delgado and Sanz-Solé (1992), Jolis and Sanz-Solé (1992) for additional contributions to the essentials).

Besides the mathematical interest of getting rid of adaptedness of the integrands in the Itô stochastic calculus, anticipating stochastic calculus has other solid motivations, like the analysis of stochastic differential systems with an anticipating initial condition or with boundary conditions. Both situations lead in a natural way to anticipating schemes which can be formulated by means of different types of anticipating stochastic integrals, basically the Skorohod and the Stratonovich integral. The papers Nualart, and Pardoux (1992), Buckdahn and Nualart (1994), provide a small sample of such examples. Numerical approximations schemes for these type of equations are presented in Ahn, and Kohatsu-Higa (1995) and Ferrante, Kohatsu-Higa and Sanz-Solé (1996).

An important question in connection with the properties of the solution of boundary value stochastic problems is the validity of the Markov field property. Some techniques of Malliavin calculus have been successfully applied to analyze this issue. A representative sample of tools and results can be found in Nualart and Pardoux (1991), Alabert, Ferrante, and Nualart (1995), Ferrante and Nualart (1995) and Alabert and Marmolejo (2001).

Anticipating stochastic calculus is also an appropiate tool to give a rigourous mild type formulation to stochastic evolution equations defined through differential operators with random coefficients. This approach has been undertaken by León and Nualart (1998) and Alòs, León and Nualart (1999).

Malliavin calculus, initially developped for the Wiener process, is also valid for an arbitrary Gaussian process. In particular one can consider Gaussian processes given by integration of either singular or regular deterministic kernels with respect to the Wiener process. This idea has been implemented by Alòs, Mazet and Nualart (2001) to develop a stochastic calculus with respect to some classes of Gaussian processes including in particular the fractional Brownian motion (see also Alòs, Mazet and Nualart (2000)). This is a challenging new field with applications to mathematical financial models.

Perhaps the most essential tool of Itô stochastic calculus is the change of variable formula, usually known as the Itô formula. In its classical version this formula gives the

martingale decomposition of a twice differentiable function of a semimartingale. Using Malliavin calculus it is possible to extend the formula to particular cases of semimartingales, such as diffusion processes, by relaxing the requirements on the function. Some achievements in this direction have been published in Bardina and Jolis (1997), Moret and Nualart (2000) and Bardina and Jolis (2002).

Probabilists have been led to study stochastic partial differential equations motivated by problems coming from physics, engineering, biology, chemistry, etc. This is quite a young subject; problems like existence and uniqueness of solution, the search for the appropriate functional analytic framework, numerical approximations, are still not closed. Some important contributions on these issues can be found in Nualart and Rozovskii (1997), Gyöngy and Rovira (1999), Gyöngy, and Nualart (1999), Gyöngy and Rovira (2000), Nualart and Viens (2000), Alabert and Gyöngy (2001), Gyöngy and Nualart (1995) (see also Nualart and Pardoux (1992) and Nualart and Tindel (1995)).

The law of the solution of stochastic equations gives relevant quantitative information. Besides its absolutely continuity an important issue is to know a characterization of the topological support of the law. Since the solution of a stochastic partial differential equation is a random stochastic process with values in an infinite-dimensional space, the question is not easy to handle. An approach based on approximations of the equation by smoothing the noise has been used by Millet and Sanz-Solé (1994), Bally, Millet and Sanz-Solé (1995), Gyöngy, Nualart and Sanz-Solé (1995) and Millet and Sanz-Solé (2000) to study hyperbolic and parabolic type equations.

The analysis of the difference between the stochastic and the deterministic model can be approached by perturbing the driving noise of the stochastic equation by a deterministic parameter and then studying different limit problems when this parameter tends to zero. Problems of this type include large deviations estimates, logarithmic estimates and Taylor expansions of the density. Malliavin calculus also enter here as a natural tool. Our contributions go from rather general settings, such as in Márquez-Carreras and Sanz-Solé (1999), to the more concrete models of anticipating type studied in Millet, Nualart and Sanz-Solé (1992), delay equations in Ferrante, Rovira and Sanz-Solé (2000) and the heat equation in Kohatsu-Higa, Márquez-Carreras and Sanz-Solé (2001), Rovira and Tindel (2001).

The preceding overwiew aims to give a sketch of the main currents of our research along the last decade, based on published work; it is by no means complete. The day-by-day activity is being posted at our URL site http://www.orfeu.mat.ub.es/~gaesto.

Stochastic Analysis has also lead to a rapid development of the area of Mathematical Finance. The group of Stochastic Analysis in Barcelona has also been interested in these developments and has studied problems related to the study of stock markets generated by Lévy processes (see León, Solé, Utzet and Vives (2002), Nualart and Schoutens (2001)) and fractional Brownian motion (see Alòs, Mazet and Nualart (2000), Alòs,

Mazet and Nualart (2001)), as well as the study of asymmetric information problems (see León, Navarro and Nualart (2002), Corcuera, Imkeller, Kohatsu-Higa and Nualart (2002)).

On the area of numerical simulations the applications of Malliavin Calculus have also lead to a rapid development of new numerical methods to the approximations of sensitivity quantities in Finance (see Fournié, Lasry, Lebuchoux, Lions and Touzi (1999), Bermin, Kohatsu-Higa and Montero (2003) and Kohatsu-Higa and Pettersson (2002)).

References

1. Jolis, M. and Sanz-Solé, M. (1990). «On generalized multiple stochastic integrals and multiparameter anticipative calculus». *Stochastic analysis and related topics*, II, 141-182. Lecture notes in Mathematics, 1444, Springer, Berlin.
2. Nualart, D. and Üstünel, A. S. (1991). «Geometric analysis of conditional independence on Wiener space». *Probability Theory Related Fields*, 89, 407-422.
3. Nualart, D. and Wschebor, M. (1991). «Intégration par parties dans l'espace de Wiener et approximation du temps local». *Probability Theory Related Fields*, 90, 83-109.
4. Nualart, D. and Pardoux, E. (1991). «Boundary value problems for stochastic differential equations». *Annals of Probability*, 19, 1118-1144.
5. Solé, J. LL. and Utzet, F. (1991). «Skorohod and Stratonovich line integrals in the plane». *Stochastic Processes and their Applictions*, 39, 239-262.
6. Delgado, R. and Sanz-Solé, M. (1992). «The Hu-Meyer formula for nondeterministic kernels». *Stochastics and Stochastics Reports*, 38, 149-158.
7. Jolis, M. and Sanz-Solé, M. (1992). «Integrator properties of the Skorohod integral». *Stochastics and Stochastics Reports*, 41, 163-176.
8. Millet, A., Nualart, D. and Sanz-Solé, M. (1992). «Large deviations for a class of anticipating stochastic diferential equations». *Annals of Probability*, 20, 1902-1931.
9. Nualart, D. and Pardoux, E. (1992). «White noise driven quasilinear spdes with reflection». *Probability Theory Related Fields*, 93, 77-89.
10. Nualart, D. (1992). «Randomized stopping points and optimal stopping on the plane». *Annals of Probabilty*, 20, 883-900.
11. Imkeller, P. and Nualart, D. (1993). «Continuity of the ocupation density for anticipating stochastic integral processes». *Potential Analysis*, 2, 137-155.
12. Malliavin, P. and Nualart, D. (1993). «Quasi-sure analysis and Strtonovich anticipative stochastic differential equations». *Probality Theory Related Fields*, 96, 45-55.
13. Buckdahn, R. and Nualart, D. (1994). «Linear stochastic differential equations and Wick products». *Probability Theory Related Fields*, 99, 501-526.

14. Imkeller, P. and Nualart, D. (1994). «Integration by parts on Wiener space and the existence of occupation densities». *Annals of Probability*, 22, 469-493.
15. Millet, A. and Sanz-Solé, M. (1994). «The support of the solution to a hyperbolic spde». *Probability Theory Related Fields*, 98, 361-387.
16. Alabert, A., Ferrante, M. and Nualart, D. (1995). «Markov field property of stochastic differential equations». *Annals of Probability*, 23, 1262-1288.
17. Ahn, H. and Kohatsu-Higa, A. (1995). «The Euler scheme for anticipating stochastic differential equations». *Stochastics and Stochastics Reports*, 54, 247-269.
18. Bally, V., Millet, A. and Sanz-Solé, M. (1995). «Approximation and support theorem in Hölder norm for parabolic stochastic partial differential equations». *Annals of Probability*, 23, 178-222.
19. Caballero, M. E., Fernández, B. and Nualart, D. (1995). «Smoothness of distributions for solutions of anticipating stochastic differential equations». *Stochastics and Stochastics Reports*, 52, 303-322.
20. Delgado, R. and Sanz-Solé, M. (1995). «Green formulas in anticipating stochastic calculus». *Stochastic Processes and their Applications*, 57, 113-148.
21. Ferrante, M. and Nualart, D. (1995). «Markov field property for stochastic differential equations with boundary conditions». *Stochastics and Stochastics Reports*, 55, 55-69.
22. Florit, C. and Nualart, D. (1995). «A local criterion for smoothness of densities and application to the supremum of th Brownian sheet». *Statistics Probability Letters*, 22, 25-31.
23. Gyöngy, I. and Nualart, D. (1995). «Implicit scheme for quasi-linear parabolic partial differential equations perturbed by space-time white noise». *Stochastic Processes and their Applications*, 58, 57-72.
24. Gyöngy, I., Nualart, D. and Sanz-Solé, M. (1995). «Approximation and support theorems in modulus spaces». *Probability Theory Related Fields*, 101, 495-509.
25. Nualart, D. (1995). *The Malliavin calculus and related topics. Probability and its Applications*. Springer-Verlag, New York.
26. Nualart, D. and Tindel, S. (1995). «Quasilinear stochastic elliptic equations with reflection». *Stochastic Processes and their Applications*, 57, 73-82.
27. Nualart, D. and Vives, J. (1995). «A duality formula on the Poisson space and some applications». Seminar on Stochastic Analysis, Random Fiels and Applications (Ascona, 1993), 205-213, *Progress in Probability*, 36, Birkhäuser, Basel.
28. Rovira, C. and Sanz-Solé, M. (1996). «The law of the solution to a nonlinear hyperbolic spde». *Journal of Theoretical Probability*, 9, 863-901.
29. Ferrante, M., Kohatsu-Higa, A. and Sanz-Solé, M. (1996). «Strong approximations for stochastic differential equations with boundary conditions». *Stochastic Processes and their Applications*, 61, 323-337.
30. Bardina, X. and Jolis, M. (1997). «An extension of Ito's formula for elliptic diffusion processes». *Stochastic Processes and their Applications*, 69, 83-109.

31. Nualart, D. and Rozovskii, B. (1997). «Weighted stochastic Sobolev spaces and bilinear spdes driven by space-time white noise». *Journal of Functional Analysis*, 149, 200-225.
32. Rovira, C. and Sanz-Solé, M. (1997). «Anticipating stochastic differential equations: regularity of the law». *Journal of Functional Analysis*, 143, 157-179.
33. León, J. and Nualart, D. (1998). «Stochastic evolution equations with random generators». *Annals of Probability*, 26, 149-186.
34. Nualart, D. (1998). «Analysis on Wiener space and anticipating stochastic calculus». *Lectures on probability theory and statistics* (Saint-Flour, 1995), 123-227, *Lecture notes in Mathematics*, 1690, Springer, Berlin.
35. Alòs, E., León, J. and Nualart, D. (1999). «Stochastic heat equation with random coefficients». *Probability Theory Related Fields*, 115, 41-94.
36. Gyöngy, I. and Rovira, C. (1999). «On stochastic partial differential equations with polynomial nonlinearities». *Stochastics and Stochastics Reports*, 67, 123-146.
37. Gyöngy, I. and Nualart, D. (1999). «On the stochastic Burgers' equation in the real line». *Annals of Probability*, 27, 782-802.
38. Márquez-Carreras, D. and Sanz-Solé, M. (1999). «Expansion of the density: a Wiener-chaos approach». *Bernoulli*, 5, 257-274.
39. Millet, A. and Sanz-Solé, M. (1999). «A stochastic wave equation in two space dimension: smoothness of the law». *Annals of Probability*, 27, 803-844.
40. Alòs, E., Mazet, O. and Nualart, D. (2000). «Stochastic calculus with respect to fractional Brownian motion with Hurst parameter lesser than 1/2». *Stochastic Processes and their Applications*, 86, 121-139.
41. Bardina, X. and Jolis, M. (2000). «Weak approximation of the Brownian sheet from a Poisson process in the plane». *Bernoulli*, 6, 653-665.
42. Ferrante, M., Rovira, C. and Sanz-Solé, M. (2000). «Stochastic delay equations with hereditary drift: estimates of the density». *Journal of Functional Analysis*, 177, 138-177.
43. Gyöngy, I. and Rovira, C. (2000). «On Lp solutions of semilinear stochastic partial differential equations». *Stochastic Processes and their Applications*, 90, 83-108.
44. Millet, A. and Sanz-Solé, M. (2000). «Approximation and support theorem for a wave equation in two space dimensions». *Bernoulli*, 6, 887-915.
45. Moret, S. and Nualart, D. (2000). «Quadratic covariation and Itô's formula for smooth nondegenerate martingales». *Journal of Theoretical Probability*, 13, 193-224.
46. Nualart, D. and Viens, F. (2000). «Evolution equation of a stochastic semigroup with white-noise drift». *Annals of Probability*, 28, 36-73.
47. Alabert, A. and Marmolejo, M. A. (2001). «Differential equations with boundary conditions perturbed by a Poisson noise». *Stochastic Processes of their Applications*, 91, 255-276.

48. Alabert, A. and Gyöngy, I. (2001). «On stochastic reaction-diffusion equations with singular force term». *Bernoulli*, 7, 145-164.
49. Alòs, E., Mazet, O. and Nualart, D. (2001). «Stochastic calculus with respect to Gaussian processes». *Annals of Probability*, 29, 766-801.
50. Kohatsu-Higa, A., Márquez-Carreras, D. and Sanz-Solé, M. «Asymptotic behavior of the density in a parabolic spde». *Journal of Theoretocal Probability*, 14, 427-462.
51. Kohatsu-Higa, A. (2001). «Weak approximations. A Malliavin calculus approach». *Math. Comp.*, 70, 135-172.
52. Márquez-Carreras, D., Mellouk, M. and Sarrà, M. (2001). «On stochastic partial differential equations with spatially correlated noise: smoothness of the law». *Stochastic Processes and their Applications*, 93, 269-284.
53. Rovira, C. and Tindel, S. (2001). «Sharp large deviation estimates for the stochastic heat equation». *Potential Analysis*, 14, 409-435.
54. Bardina, X. and Jolis, M. (2002). «Estimation of the density of hypoelliptic diffusion processes with application to an extended Itô's formula». *Journal of Theoretical Probability*, 15, 223-247.
55. León, J. A., Solé, J. L., Utzet, F. and Vives, J. (2002). «On Lévy processes, Malliavin calculus and market models with jumps». *Finance and Stochastics*, 6, 197-225.
56. León, J. A., Navarro, R. and Nualart, D. (2002). *An anticipating calculus approach to the utility maximization of an insider*. Preprint.
57. Corcuera, J. M., Imkeller, P., Kohatsu-Higa, A. and Nualart, D. (2002). *Additional utility of insiders with imperfect dynamical information*. Preprint.
58. Fournié, E., Lasry, J.-M., Lebuchoux, J., Lions, P.-L. and Touzi, N. (1999). «Applications of Malliavin calculus to Monte Carlo methods in finance». *Finance and Stochastics*, 3, 391-412.
59. Bermin, H.-P., Kohatsu-Higa, A. and Montero, M. (2003). «Local Vega index and variance reduction methods». To appear in *Mathematical Finance*.
60. Kohatsu-Higa, A. and Pettersson, R. (2002). «Variance reduction methods for simulation on Wiener space». *SIAM Journal of Numerical Analysis*, 40, 431-450.
61. Alòs, E., Mazet, O. and Nualart, D. (2000). «Stochastic calculus with respect to fractional Brownian motion with Hurst parameter less that 1/2». *Stochastic Processes and Their Applications*,86, 121-139.
62. Alòs, E., Mazet, O. and Nualart, D. (2001). «Stochastic calculus with respect to Gaussian processes». *Annals of Probability*, 29, 766-801.
63. Nualart, D. and Schoutens, W. (2001). «BSDE's and Feynman-Kac formula for Lévy processes with applications in finance». *Bernoulli*, 7, 761-776.

STRUCTURAL EQUATION MODELS WITH LATENT VARIABLES

Albert Satorra
Departament d'Economia i Empresa
Universitat Pompeu Fabra

Structural Equation Modeling (SEM) is widely used in behavioural, social and economic studies to analyse structural relations between variables, some of which may be latent (i.e., unobservable). At present, SEM encompasses a wide variety of models and methods for multivariate analysis, such as multiple and multivariate regression, errors-in-variable models, ordered-probit regression, multiple-indicator models, factor analysis, simultaneous equation models, models for panel data, growth-curve models, and so forth. Furthermore, the observed variables can be normally distributed, but also continuous non-normal, or just ordinal; in addition, the data can have a single- or multiple-group, or multilevel structure, or can be mixture-distribution.

Nowadays, there exists a variety of commercial computer programs to carry out the practice of SEM, under general conditions on the model and the type of data. The most typical programs are LISREL, of Jöreskog and Sörbom; EQS, of Bentler; Mplus, of Muthén; the procedure CALIS of SAS, among others. For a thorough historical perspective of SEM, and its present state of the art, the reader should consult the recent book of Cudeck, Du Toit & Sörbom (2001).

Catalunya has various research groups with recognized international status in the area of SEM. There is Batista and Coenders's group (Universitat Ramon Llull and Universitat de Girona) that has contributed applications of SEM in research in social and behavioural sciences, specially on the methodology of measurement. The group of Maydeu-Olivares (Universitat de Barcelona) that has contributed with fundamental work on SEM analysis of rank-ordered and general discrete-type of data. The group of Ferrando and Lorenzo-Seva (Universitat Rovira i Virgili) working on SEM analysis of discrete data and issues of rotation in Factor Analysis. Since some of these groups are also contributing to the present issue, below I will just concentrate on SEM at UPF.

Power of the test and model search: Power analysis is required when we want to assess the substantive significance of coefficients of the model, and/or the validity of the model. A procedure for computing the power of the test in SEM is developed in Satorra and Saris (1985). Goodness of fit testing and power issues are also worked out in Satorra (1989) and, more recently, in Satorra (2001). A procedure for model search and model modification is worked out in Saris, Satorra and Sörbom (1987). The so-called "expected parameter change", which was developed in this paper, is nowadays implemented in most of the software for SEM.

Asymptotic robustness for multiple-group data: Asymptotic robustness (AR) concerns the validity of inferences derived under the NT assumption when the data deviates from normality. Since multivariate data often deviates from the normality assumption, assessing the validity of an analysis based on the normality assumption is of crucial importance in practice. AR issues for statistics arising in weighted-least-square analysis of SEM are investigated in Satorra and Bentler (1990). AR for mean-and-covariance structures is investigated in Satorra (1992) and Satorra and Neudecker (1994). AR issues for multiple-group mean-covariance structures is investigated in Satorra (1993) and Satorra (2002).

Scaled and adjusted test statistics: In many applications, chi-square goodness-of-fit tests are used when data are non-normal and there is non-compliance of assumptions required for asymptotic robustness. In such situations, corrected test statistics may be useful. Scaled and adjusted test statistics (for chi-square goodness-of-fit test, difference, score and Wald type test statistics) appropriate for non-normal data are developed in Satorra and Bentler (1994), Satorra (2000) and Satorra and Bentler (2001) (the scaled statistic of Satorra and Bentler (1994) is now available in most of the software for SEM).

SEM with categorical (ordinal) data, multilevel, and complex samples: Muthén and Satorra (1989) and Muthén and Satorra (1995) investigate SEM issues for multilevel and complex-sample structure of the data. Theory for SEM for continuous and categorical (ordinal) data can be found in Muthén and Satorra (1995).

Combining data sets from different sources: Multiple-group SEM has been shown to play a central role on the problem of combining data sets from different sources (Satorra, 1997). This is an issue that is recently attracting a lot of attention from different areas of applications.

Matrix algebra aspects of SEM: Matrix algebra has become essential for multivariate analysis. Advanced matrix algebra methods are needed to implement SEM in practice (specially when using high-level computer programming such as Matlab, or R). A general matrix formulation for testing equality of moments can be found in Satorra and Neudecker (1995). A matrix equality useful for goodness of fit testing in SEM is developed in Satorra and Neudecker (2003). Expressions for best linear prediction (as in estimation of factor scores) under general conditions are obtained by Neudecker and Satorra (2003).

Applications: On the side of applications, Ventura and Satorra (2001) use SEM on Spanish ECPF survey data to investigate variation of expenditure-patterns across various family groups, while Rivera and Satorra (2002) use multiple group SEM on ISSP-1993 data (a survey on twenty-two countries) to investigate the variation across countries of attitudes toward environmental issues.

Finally, we would like to mention other research areas in which statisticians at the UPF are involved, areas that relate with statistical modeling but that go beyond the scope of SEM. Work on the **description and visualization of multivariate data** is represented elsewhere under «Data Analysis and Data Mining» [Greenacre, at UPF]. Research on **information theory and non-parametric methods** is represented by Delicado (2000) [Delicado, now at UPC], Devroye, Lugosi and Udina (2001) Lugosi and Nobel (1999), and Udina (1999) [Lugosi and Udina, at UPF]. **Multilevel analysis** applied to data in education is represented by Cuxart-Jardí and Longford (2000) [Cuxart-Jardí, at UPF]. **Small are estimation** applied to regional data is represented by Costa, Satorra and Ventura (2002) [Ventura, at UPF]. **Analysis of preference data** by Van de Velden and Neudecker (2000) [Van de Velden, at UPF]. **Modeling human birth data and environmental statistics** is represented elsewhere under «Data Analysis and Data Mining» [Graffelman, now at UPC]. **Survival analysis with measurement error on covariates** is represented by Espinal-Berenguer and Satorra (1996) [Espinal, now at UAB]. Research on **subsampling and improved estimation** is represented by Romano and Wolf (2001) and Ledoit and Wolf (2002) [Wolf, at UPF]. Work on **stochastic processes** is represented elsewhere under «Research group on Stochastic Processes» [Kohatsu and Alós, at UPF]. There is also a vast amount of research on Operations Research and in Econometrics published by other colleagues at the UPF.

References

1. Costa, A., Satorra, A. and Ventura, E. (2002). «Estimadores Compuestos en Estadística Regional: una Aplicación a la Estimación de la Tasa de Variación de la Ocupación en la Industria». *Qüestiió*, 26, 213-243.
2. Cuxart-Jardí, A. and Longford, N. J. (1998). «Monitoring the University Admission Process in Spain». *Higher Education in Europe*, 23.
3. Cudeck, R., Du Toit, S. and Sörbom, D. (eds.). (2001). *Structural Equation Modeling: Present and Future*, SSI Scientific Software International, Lincolnwood, IL.
4. Devroye, L., Lugosi, G. and Udina, F. (2000). «Inequalities for a new data-based method for selecting nonparametric density estimates», in *Asymptotics in Statistics and Probability. Papers in Honor of George Gregory Roussas*, M. L. Puri (ed.), 133-154, International Science Publishers, Amsterdam.
5. Espinal-Berenguer, A. and Satorra, A. (1996). «Survival Analysis with Measurement Error on Covariates», in *Proceedings in Computational Statistics*, A. Prat, (ed.), 253-263, Springer Verlag, Heidelberg.
6. Ledoit, O. and Wolf, M. (2002). «Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size». *Annals of Statistics*, 30, 1081-1102.

7. Lugosi, G. and Nobel, A. B. (1999). «Adaptive model selection using empirical complexities». *The Annals of Statistics*, 27, 1830-1864.

8. Muthén, B. and Satorra, A. (1989). «Multilevel Aspects of Varying Parameters in Structural Models», in R. D. Bock (ed.), *Multilevel Analysis of Educational Data*, 87-99, New York: Academic Press.

9. Muthén, B. and Satorra, A. (1995*a*). «Complex Sample Data in Structural Equation Modeling». *Sociological Methodology*, 25, 267-316.

10. Muthén, B. and Satorra, A. (1995*b*). «Technical aspects of Muthen's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model». *Psychometrika*, 60, 489-503.

11. Neudecker, H. and Satorra, A. (1991). «Linear structural relations: Gradient and Hessian of the fitting function». *Statistics & Probability Letters*, 11, 57-61.

12. Rivera, P. and Satorra, A. (2001). «Analysing Group Differences: A Comparison of SEM Approaches», in *Latent Variable and Latent Structure Models* (G. Marcoulides and I. Moustaki, eds.) Lawrence Erlbaum Associates, Inc, Publishers, NJ.

13. Romano, J. P. and Wolf, M. (2001). «Subsampling intervals in autoregressive models with linear time trend». *Econometrica*, 69, 1283-1314.

14. Saris, W. E., Satorra, A. and Sörbom, D. (1987). «Detection and Correction of Specification on Errors in Structural Equation Models». *Sociological Methodology*, 17, 105-129.

15. Satorra, A. (1997). «Fusion of Data Sets in Multivaraite Linear Regression with Errors-in-Variables». *Classification and Knowledge Organization*, 195-207, R. Klar and O. Opitz (eds.), Springer Verlag: London.

16. Satorra, A. (1989). «Alternative Test Criteria in Covariance Structure Analysis: A Unified Approach». *Psychometrika*, 54, 131-151.

17. Satorra A. (1992). «Asymptotic robust inferences in the analysis of mean and covariance structures». *Sociological Methodology*, 22, 249-278.

18. Satorra, A. (1993). «Asymptotic Robust Inferences in Multi-Sample Analysis of Augmented-Moment Matrices», in *Multivariate Analysis: Future Directions*, 2, 211-229, (C. R. Rao, C. M. Cuadras, eds.), North-Holland, New York.

19. Satorra, A. (2000). «Scaled and Adjusted Restricted Tests in Multi-Sample Analysis of Moment Structures», in *Innovations in Multivariate Statistical Analysis: A Festschrift for Heinz Neudecker*. Heijmans, D. D. H., Pollock, D. S. G. and Satorra, A. (eds.), 233-247, Kluwer Academic Publishers, Dordrecht.

20. Satorra, A. (2001). «Goodness of fit testing of structural equation models with multiple group data and nonnormality», in *Structural Equation Modeling: Present and Future*, R. Cudeck, S. Du Toit and D. Sörbom (eds.), SSI Scientific Software International, Lincolnwood, IL.

21. Satorra, A. (2002). «Asymptotic robustness in multiple group linear-latent variable models». *Econometric Theory*, 18, 297-312.

22. Satorra, A. and Bentler, P. M. (1990). «Model Conditions for Asymptotic Robustness in the Analysis of Linear Relations». *Computational Statistics & Data Analysis*, 10, 235-249.

23. Satorra, A. and Bentler, P. M. (1994). «Correctios to test statistics and standard errors in covariance structure analysis», in *Latent variable Analysis in Developmental Research*, 285-305 (A. van Eye and C. C. Clogg, eds.), SAGE Publications, Inc., Thousand Oaks, CA.

24. Satorra, A. and Bentler, P. M. (2001). «A scaled difference chi-square test statistic for moment structure analysis». *Psychometrika*, 66, 507-514.

25. Satorra A. and Saris, W. E. (1985). «Power of the likelihood ratio test in covariance structure analysis». *Psychometrika*, 50, 83-89.

26. Satorra, A. and Neudecker, H. (1994). «On the asymptotic optimality of alternative minimum-distance estimators in linear latent-variable models». *Econometric Theory*, 10, 867-883.

27. Satorra, A. and Neudecker, H. (1997). «Compact Matrix Expressions for Generalized Wald Tests of Equality of Moment Vectors». *Journal of Multivariate Analysis*, 63, 259-276.

28. Neudecker, H. and Satorra, A. (2003). «On Best Affine Prediction». *Statistical Papers*, (forthcoming).

29. Satorra, A. and Neudecker, H. (2003). «A matrix equality useful in goodness-of-fit testing of structural equation models». *Journal of Statistical Planning and Inference*, (forthcoming).

30. Ventura, E. and Satorra, A. (2001). «Product expenditure patterns n the ECPF survey: An analysis using multiple group latent-variables models». *Qüestiió*, 25, 93-117.

31. Udina, F. (1999). «Implementing interactive computing in an object-oriented environment». *Journal of Statistical Software*, 5, 1-20.

32. Van de Velden, M. and Neudecker, H. (2000). «On an eigenvalue property relevant in correspondence analysis and related methods». *Linear Algebra and Its Applications*, 324, 347-364.

THE COMPUTER TECHNOLOGY IN THE BEHAVIOURAL SCIENCE RESEARCH GROUP

Antonio Solanas Pérez
Departament Metodologia de les Ciències del Comportament
Facultat de Psicologia
Universitat de Barcelona

The computer technology in the behavioural science research group develops mathematical and statistical models for describing behavioural and psychological processes. Since the eighties a group of researchers in Barcelona has been studying behaviour sequences analysis, statistical analysis of $N = 1$ designs, analytical techniques for simulating and describing social behaviour, measurement of typicality in scene categorization processes, and statistical analysis of bioelectric signals.

Behaviour sequences analysis: A basic property of behaviour is that it unfolds in time. In order to be able to describe and explain certain behavioural phenomena, it is necessary to observe, code, and analyze sequences of behaviour. Such analysis is helpful for uncovering patterns of behaviour and for depicting their temporal dynamics. Since the early eighties the behavioural science group has developed software for the statistical analysis of behavioural sequences (Bakeman and Quera, 1992, 1995*a*, 2000), and has proposed some improvements to the techniques used in that field (Bakeman, McArthur and Quera, 1996; Bakeman and Quera, 1995*b*; Bakeman, Quera, McArthur and Robinson, 1997; Bakeman, Robinson and Quera, 1996; Quera, 1990; Quera and Bakeman, 2000).

Statistical analysis of $N = 1$ designs: It is well known that problems arise when standard statistical techniques (such as the t-test and ANOVA) are used for analyzing data from $N = 1$ designs. Time series analysis is generally useless in this case because of the small number of data points in these designs. Although several analysis techniques had been proposed that are claimed to control for the autocorrelation present in series of data in $N = 1$ designs, those techniques do not provide satisfactory solutions. Our research confirms and extends to other statistical tests the results obtained by previous researchers (Sierra, Quera and Solanas, 2000; Solanas, Salafranca and Guardia, 1992; Solanas and Sierra, 1995).

Analytical techniques for simulating and describing social behaviour: Group behaviour and processes are viewed as phenomena that emerge from a simple and small set of interaction rules among agents. Agents are abstract entities that have basic psychological properties and adapt to the environment and to the behaviour of other agents. Specific models are run in an artificial world (Quera, Solanas *et al.*, 2000; Zibetti *et al.*, 2001), and developed until a reasonable fit with empirical data is reached (Quera,

Beltran *et al.*, 2000). Simulation makes it possible to analyze some situations difficult to deal with the real world, and also to uncover unknown phenomena that can be subsequently subjected to empirical test. This approach allows us to understand psychological and sociological mechanisms of social behaviour of human and non-human organisms.

Measuring typicality in scene categorization processes: Typicality is a basic property of both the internal organization of the lexical categories, and scenes. The typicality of a given exemplar is usually measured from its frequency, or from judgments in a Likert scale. Those procedures and analysis were not appropriate for scene categorization. We proposed a procedure to categorize objects present in environmental scenes, bassed on psychometric IRT model. (Beltran, 1990; Beltran and Herrando, 1995; Beltran, Herrando and Pelegrina, 1992; Beltran, Herrando and Salavert, 1998).

Statistical analysis of bioelectric signals: We have investigated spatio-temporal mechanisms underlying lingual activity in speech production (Recasens *et al.*, 1993) and the effect of thalamic lesions on sleep spindles activity (Santamaría *et al.*, 2000). This approach allows us to identify indicators of psychological and physiological processes.

References

1. Bakeman, R., McArthur, D. and Quera, V. (1996). «Detecting group differences in sequential association using sampled permutations: Log odds, kappa, and phi compared». *Behavior Research Methods, Instruments and Computers*, 28, 446-457.
2. Bakeman, R. and Quera, V. (1992). «SDIS: A sequential data interchange standard». *Behavior Research Methods, Instruments and Computers*, 24, 554-559.
3. Bakeman, R. and Quera, V. (1995*a*). *Analyzing interaction: Sequential analysis with SDIS and GSEQ*. New York: Cambridge University Press.
4. Bakeman, R. and Quera, V. (1995*b*). «Log-linear approaches to lag-sequential analysis when consecutive codes may and cannot repeat». *Psychological Bulletin*, 118, 272-284.
5. Bakeman, R. and Quera, V. (2000). «OTS: A program for converting Noldus Observer data files to SDIS files». *Behavior Research Methods, Instruments and Computers*, 32, 207-212.
6. Bakeman, R., Quera, V., McArthur, D. and Robinson, B. F. (1997). «Detecting sequential patterns and determining their reliability with fallible observers». *Psychological Methods*, 2, 357-370.
7. Bakeman, R., Robinson, B. F. and Quera, V. (1996). «Testing sequential association: Estimating exact P values using sampled permutations». *Psychological Methods*, 1, 4-15.
8. Beltran, F. S. (1990). «Typicality effects when processing objects into environmental scenes». *Current Psychology: Research & Reviews*, 9, 69-74.

9. Beltan, F. S. and Herrando, S. (1995). «Measuring the typicality of objects included in environmental scenes: A logistic model for atypicality». *Perceptual and Motor Skills*, 80, 1343-1352.
10. Beltran, F. S., Herrando, S. and Pelegrina, M. (1992). «Measuring the typicality of objects included in environmental scenes: A first scale». *Perceptual and Motor Skills*, 74, 827-831.
11. Beltran, F. S., Herrando, S. and Salavert, F. (1998). «Typicality of objects in environmental scenes: Effects of stimuli». *Perceptual and Motor Skills*, 87, 928-930.
12. Quera, V. (1990). «A generalized technique to estimate frequency and duration in time sampling». *Behavioral Assessment*, 12, 409-424.
13. Quera, V. and Bakeman, R. (2000). «Quantification strategies in behavioral observation research». In T. Thompson, D. Felce and F. Symons (eds.), *Behavioral observation: Technology and applications in developmental disabilities*, 297-315. Baltimore: Brookes Publishing.
14. Quera, V., Beltran, F. S., Solanas, A., Salafranca, Ll. and Herrando, S. (2000). «A dynamic model for inter-agent distances». In J. A. Meyer, A. Berthoz, D. Floreano, H. L. Roitblat and S. W. Wilson (eds.), *From Animals to Animats 6, SAB2000 Proceedings Supplement*, 304-313. Honolulu, Hawaii.
15. Quera, V., Solanas, A., Salafranca, Ll., Beltran, F. S. and Herrando, S. (2000). «P-SPACE: A program for simulating spatial behavior in small groups». *Behavior Research Methods, Instruments, and Computers*, 32, 191-196.
16. Recasens, D., Fontdevila, J., Pallarés, M. D. and Solanas, A. (1993). «An electropalatographic study of stop consonant clusters». *Speech Communication*, 12, 335-355.
17. Santamaría, J., Pujol, M., Orteu, N., Solanas, A., Cardenal, C., Santacruz, P., Chimeno, E. and Moon, P. (2000). «Unilateral thalamic stroke does not decrease ipsilateral sleep spindles». *Sleep*, 3, 333-339.
18. Sierra, V., Quera, V. and Solanas, A. (2000). «Autocorrelation effect on type I error rate of Revusky's Rn test: A Monte Carlo study». *Psicológica*, 32, 91-114.
19. Solanas, A., Salafranca, Ll. and Guàrdia, J. (1992). «Análisis estadístico de diseños conductuales». *Psicothema*, 1, 253-259.
20. Solanas, A. and Sierra, V. (1995). «Análisis de incrementos-decrementos». *Psicothema*, 1, 187-199.
21. Zibetti, E., Quera, V., Beltran, F. S. and Tijus, Ch. (2001). «Contextual categorization: A mechanism linking perception and knowledge in modeling and simulating perceived events as actions». In V. Akman, P. Bouquet, R. Thomason and R. A. (eds.), *Modeling and using context. Third International and Interdisciplinary Conference Context 2001 Proceedings*, 395-408. Springer, Lecture Notes in Artificial Intelligence.

MATHEMATICAL AND STATISTICAL MODELS IN THE BIOSCIENCES RESEARCH GROUP

Albert Sorribas
Departament de Ciències Mèdiques Bàsiques
(Bioestadística)
Universitat de Lleida

A complex system, such as a metabolic pathway, a physiological network or a population, requires developing specific tools for analyzing its properties. Mathematical models deal with such complexity by building-up approximate descriptions that can be used to critically inspect system's properties. Among other possibilities, ordinary differential equations models (ODE) provide an appropriate framework from elementary chemical reactions to population dynamics. A critical step in definning appropriate ODE models is to choose a suitable formalism for representing the underlying processes. If the goal is to analyze the whole system, a mathematical description based on a too detailed representation of individual processes may be inappropriate and may lead to unpractical models. Our group, established since 1993 in the Department of Basic Medical Sciences of the University of Lleida, has specialized in a specific approach known as *power-law formalism*. This formalism provides an approximate representation for non-linear functions and leads to mathematical models that can be used for characterizing large-system's properties. Basically, this formalism is based on a Taylor series approximation in logarithmic coordinates, which provides a power-law approximation in Cartesian coordinates. As a result, the models based in this formalism provide a good representation of the non-linear properties of the underlying processes and yet they are easily analyzable both algebraically and numerically. Modeling issues that have been addressed using this formalism encompases from regulatory networks to modeling statistical distributions. An up-to-date information on published papers and last results can be found at the web page: www.udl.es/usuaris/q3695988/WebPL/main.htm.

Development of the power-law formalism for system modeling and analysis: One of the main concerns in using mathematical models is parameter estimation. In metabolic systems experimental data come from different sources and design of optimal experiments for parameter estimation is seldom possible. In such cases, alternative strategies must be used, including the possibility of mixing both experimental and qualitative data based on published observations and on expertise. Issues concerning parameter estimation either from steady-state measurements or from dynamic observations and a method for identifying the regulatory estructure of a pathway can be found in: Sorribas *et al*, 1993; Cascante, Sorribas and Canela, 1994; Sorribas and Cascante, 1994. Formally, the power-law formalism can be related to other existing formalisms such as Generalized Lotka Volterra models. Within metabolic studies, a common alternati-

ve is to use the Metabolic Control Analysis approach. Comparison between different formalisms can be found in: Curto, Sorribas and Cascante, 1995; Soribas, Curto and Cascante, 1995; Cascante, Curto and Sorribas, 1995. More recently, we have turned to develop a new approach based on least-squares criteria as a definition for the power-law formalism. This new approach may be more suitable for model definition from steady-state measurements (Hernández-Bermejo, Fairén and Sorribas, 1999, 2000; Sorribas and Hernández-Bermejo, 2002).

Modeling and analyzing large systems: One of the critical steps in using a mathematical model is validation. This is specially important in models that are build using different sources of information. The power-law formalism provides tools for a systematic validation in large models. These tools are related to sensitivity analysis and a sound knowledge of the biological framework (Curto *et al*, 1997, 1998). Tools for evaluating the effect of parameter uncertainties on large models can be found in: de Atauri, Sorribas and Cascante, 2000. Current work on modeling and analyzing large systems through the use of these techniques in our group includes: (*i*) A model of cell cycle in yeast, with inclusion of stochastic effects related to cellular compartimentation, (*ii*) Bridging the gap between Genomics and Physiology through the development of a model based on yeast response to oxidative stress and data from DNA-microarrays, and (*iii*) A model of the spread of HIV infection in our community.

S-distributions: The power-law formalism can be used to define approximate models in many instances. One of the most interesting applications is the modeling of statistical distributions by means of a family of distributions known as S-distribution (Voit, 1992). We have studied some important characteristics of this family, including its potential use as general method for generating a random sample of a given distribution (Hernández-Bermejo and Sorribas, 2001) and the representation of dynamic changes in distribution associated to a growth process (Voit and Sorribas, 2000). From a practical point of view, we have developed a method for estimating conditional S-distributions (Sorribas, March and Voit, 2000; March *et al.*, 2002) and for Receiver Operating Characteristic curves in medical diagnostic problems (Sorribas, March and Trujillano, 2002). A numerical method for maximum-likelihood estimation for S-distributions has also been defined (March *et al.*, 2002). Our current research aims to generalize the S-distribution family, to further develop numerical tools for random number generation based on this family, and to define hypothesis tests based on S-distributions.

References

1. Cascante, M., Curto, R. and Sorribas, A. (1995). «Comparative characterization pathway of Saccharomyces cerevisiae using Biochemical Systems Theory and Metabolic Control Analysis: Steady-state analysis». *Mathematical Bioscience*, 130, 51-69.

2. Cascante, M., Sorribas, A. and Canela, E. I. (1994). «Enzyme-enzyme interactions and metabolite channelling: alternative mechanisms and their evolutive significance». *Biochemical Journal*, 298, 313-320.
3. Curto, R., Sorribas, A. and Cascante, M. (1995). «Comparative characterization pathway of Saccharomyces cerevisiae using Biochemical Systems Theory and Metabolic Control Analysis: Model definition and nomenclature». *Mathematical Biosciences*, 130, 25-50.
4. Curto, R., Voit, E. O., Sorribas, A. and Cascante, M. (1997). «Validation and steady-state analysis of a power-law model of purine metabolism». *Biochemical Journal*, 324, 761-775.
5. Curto, R., Voit, E. O., Sorribas, A. and Cascante, M. (1998). «A power-law model of purine metabolism: comparison of alternative modelling strategies». *Mathematical Biosciences*, 151, 1-49.
6. de Atauri, P., Sorribas, A. and Cascante, M. (2000). «Evaluation of the effect of boundary model assumptions». *Biotechnology Bioengineering*, 68, 18-30.
7. Hernández-Bermejo, B. and Sorribas, A. (2001). «Analytical quantile solution for the S-distribution, random number generation and statistical data modelling». *Biomedical Journal*, 43, 1017-1025.
8. Hernández-Bermejo, B., Fairén, V. and Sorribas, A. (1999). «Power-law modelling based on least-squares minimization criteria». *Mathematical Biosciences*, 161, 83-97.
9. Hernández-Bermejo, B., Fairén, V. and Sorribas, A. (2000). «Power-law modelling based on least-squares criteria: consequences for systems analysis and simulation». *Mathematical Biosciences*, 167, 87-107.
10. March, J., Trujillano, J., Tort, M. and Sorribas, A. (2002). «Estimating conditional distributions using a method based on S-distributions: Reference percentile curves for body mass index in Spanish Children». *Growth Development and Aging* (submitted for publication).
11. Sorribas, A. and Cascante, M. (1994). «Structure identifiability in metabolic pathways: parameter estimation in models based on the power-law formalism». *Biochemical Journal*, 298, 303-311.
12. Sorribas, A. and Hernández-Bermejo, B. (2002). «Modeling a metabolic pathway from steady-state measurements on the intact system: on the usefulness of least-squares derived power-law models». *Mathematical Biosciences* (submitted for publication).
13. Sorribas, A., Curto, R. and Cascante, M. (1995). «Comparative characterization pathway of *Saccharomyces cerevisiæ* using Biochemical Systems Theory and Metabolic Control Analysis: Model validation and dynamic behaviour». *Mathematical Biosciences*, 130, 71-84.
14. Sorribas, A., March, J. and Trujillano, J. (2002). «A new parametric mtehod based on S-distributions for computing Receiver Operating Characteristic curves for continuous diagnostic tests». *Statistics in Medecine* (in press).

15. Sorribas, A., March, J. and Voit, E. O. (2000). «Data modelling using S-distributions: Utility in estimating age-related trends in cross-sectional studies». *Statistics in Medecine*, 19, 697-713.
16. Sorribas, A., Samitier, S., Canela, E. I. and Cascante, M. (1993). «Metabolic pathway characterization from transient response data obtained in situ: parameter estimation in S-system models». *Journal of Theoretical Biology*, 162, 81-102.
17. Voit, E. O. (1992). «The S-distribution: A tool for approximation and classification of univariate, unimodal probability distributions». *Biomedical Journal*, 7, 855-878.
18. Voit, E. O. and Sorribas, A. (2000). «Computer modeling of dynamically changing distributions of random variables». *Mathematical and Computer Modelling*, 31, 217-225.

THE REGIONAL QUANTITATIVE ANALYSIS GROUP

Jordi Suriñach
Departament d'Econometria, Estadística i Economia Espanyola
Facultat de Ciències Econòmiques i Empresarials
Universitat de Barcelona

The «Grup d'Anàlisi Quantitativa Regional» (AQR) (Regional Quantitative Analysis Group) co-directed by Dr. Manuel Artís and Dr. Jordi Suriñach, is a research unit of the University of Barcelona, and is linked to the Department of Econometrics, Statistics and Spanish Economics of the same University. With more than thirty researchers and twenty years of experience, it has been recognized by the *Generalitat de Catalunya* (the regional government) as a Quality Research Group. Since 1987, the AQR Group has been working for many public institutions and private companies.

The group's main research activities are focused on urban and regional economic analysis and the realization of macroeconomic forecasts. Currently, the main research areas are as follows:

- **Regional and urban economic analysis:** The concept of functional areas is linked to the interrelations between neighbouring municipalities, due to job or study commuting, or due to the fact these municipalities share services. These aggregations are very important when urban studies are carried out, because they continually take into account the rest of the municipalities which are part of their urban system. Commuting is a result of many municipal necessities, such as public transport, traffic, infrastructures and common policies for towns in the same urban system.
- **Economic growth factor and regional convergence:** This line of research is focused on the analysis of the process of regional convergence in European regions, industry location factors and the effect of production factors, such as public and private capital, technology or infrastructure on the economic growth. In addition, much work has been done on externalities across economies using spatial econometrics, and macroeconomic implications of EMU at regional level.
- **Statistical and econometric methodology:** In this mostly theoretical area, we try to solve statistical and econometric problems related to the stationary, temporary aggregation and crossing of economic variables, including research on the real cycles, dynamic modelling, filters, seasonality, long memory, spatial econometrics and cyclical tendencies.
- **Drawing up, updating and treatment of series:** To complete historical series about consumption, labour market, prices, industrial production and other related themes, we need a strict levelling down process of database and comprehensive research on existing information. In addition, the treatment of the data series in-

cludes enrolment, interpolations, treatment of outliers, stationary analysis, cyclical tendency, connection of the series and rationalisation of the variables.

- **Design, execution and exploitation of surveys:** The complete process is carried out, according to the technique used, to reach the main aim of the survey. This has allowed us to carry out important surveys such as one related to the degree of preparation of firms in the years 1999, 2000 and 2001 before the introduction of the Euro, in sectors such as tourism, industry and finance.
- **Macroeconomic forecasts and estimations:** The expert's capacity to make economic forecasts and estimations is very helpful to the agents to take good decisions about economic measures and drawing up budgets and, moreover, the information given could help analysts, businessmen etc. In this field there is a great deal of work done with the collaboration of several institutions.
- **Cost-benefit valuation of economic impact:** The cost-benefit valuation has been applied in three major areas: infrastructures, TAV and health.
- **Synthetic indicators of living standards and economic activity:** These synthetic indicators have great advantages: they concentrate a lot of information in only one variable, and they also permit us to carry out both cross-sectional and longitudinal comparisons. Quality of life indices at the municipal level and the urban system level have been calculated, based on a large number of statistics and local data from different areas, such as car parking, per capita income, cultural infrastructures, health and education, financial situation of the town council, housing, and unemployment. All these indicators are usually published in statistical yearbooks or similar publications, to show a global view of the progress of the regional, municipal, or local economy.
- **Strategic plans:** Strategic plans are a complete and multidisciplinary view about the actual situation of the town or the city in study, focusing on its weaknesses, its strong points and its potentialities. They consider especially the social and economic part: economic activity, real state market, unemployment, firms, sector activity distribution, supply and demand of land, housing quality, housing and the land prices, and population-housing adaptation.
- **Economics reports:** Several reports about punctual subjects related with economy have been carried out. This research is possible thanks to a number of EU Projects and some private contracts, such as:
- «Small and Medium Enterprises, Economic Development and Regional Convergence in Europe». **Cost Action 17 Social Sciences**. European Commission (2000-2004).
- «European Forecast Network». **European Commission** (2001-2002).

- «Estudi de l'evolució del trànsit a la ciutat de Barcelona» (Study of transit evolution in Barcelona city). **Ajuntament de Barcelona** (1997-2003).
- «Estimació de a curt i a mig termini de l'activitat econòmica a Catalunya» (Medium and short term estimation of catalan economic activities). **Cambra de Comerç de Barcelona** (COCINB) (1992-2001).
- «Enquesta sobre la Introducció de l'Euro a les empreses catalanes» (Survey about Euro introduction to catalan firms). **Departament d'Economia i Finances de la Generalitat de Catalunya i el CIDEM** (1999-2001).
- «Pla de Desenvolupament econòmic i social. Pla Estratègic"(Economic and social development plan. Strategic plan). **Ajuntament de Figueres** (2000).
- «Anàlisi de l'evolució macroeconòmica de les Illes Balears i per Illes» (Macroeconomic analysis of the Balearic Islands and of each Island). **Govern Balear** (1999-2002).
- «Estimació de magnituds econòmiques als països de la zona Euro» (Estimation of economic magnitudes for the Euro zone countries). **La Caixa d'Estalvis i Pensions de Barcelona "La Caixa"** (1999).
- «Construcció d'un indicator sintètic de Qualitat de vida i d'Activitat Econòmica» (Drawing up a synthetic indicators of life quality and economic activity). **Diputació de Barcelona** (1999).

Finally some selected publications from the last five years are listed, which reflect the importance and quality of these research lines:

Economic and policy analysis at the European level

Moreno, R., López-Bazo, E. and Artís, M. (2002). «Public infraestructure and the performance of manufactures: short and long run effects», *Regional Science and Urban Economics*, 32, 97-121.

Ramos, R., Clar, M. and Suriñach, J. (2001). «Comercio y variabilidad del tipo de cambio: Evidencia para los países de la Unión Europea», *ICE. Revista de Economía. Consejo Superior de Investigaciones Científicas*, 794, 77-89.

Ramos, R. (2000). «Macroeconomic implications of EMU: The Catalonian case». This work was awarded with the first prize of the universitary contest «L'adaptació de l'economia catalana a la moneda única europea', organized by the CIDEM» (Generalitat de Catalunya) in co-operation with the «Quatre Moteurs pour l'Europe».

Ramos, R., Clar, M. and Suriñach, J. (1999). «Specialisation in Europe and asymmetric shocks: potential risks of EMU», en M. M. Fischer y P. Nijkamp (eds.), *Spatial Dynamics of European Integration. Political and Regional Issues at the Turn of the Millenium*, Springer-Verlag, Berlin, 63-93.

López-Bazo, E., Vayà, E., Mora, A. and Suriñach, J. (1999). «Regional economic dynamics and convergence in the European Union», *Annals of Regional Science*, 33, 343-370.

Sanromá, E. and Ramos, R. (1999). «Interprovincial Wage Differences in Spain. A microdata Analysis», *Jahrbuch fuer Regionalwissenschaft-Review of Regional Science Research*, 19, 35-54.

Moreno, R., Artís, M., López-Bazo, E. and Suriñach, J. (1999). «Evidence on the complex link between infrastructures and regional growth», *International Journal of Development Planning Literature*, 12, 81-108.

Sanromá, E. and Ramos, R. (1998). «El mercado de trabajo español en la Unión Monetaria. Flexibilidad de salarios y política laboral», in J. C. Jimenez (ed.): *La economía española ante el Euro*, Ed. Civitas, Madrid, 133-176. (The Spanish labour market and the Monetary Union. Wage flexibility and labour market policies).

Forecasting the economic evolution of European countries

European Forecasting Network (2002). «Report on the Euro-Area Outlook. Autum 2002», research report for the *European Commission*, EFN project.

European Forecasting Network (2002). «Report on the Euro-Area Outlook. Spring 2002», research report for the *European Commission*, EFN project.

Artís, M., Clar, M., Ramos, R., Sansó, A. and Suriñach, J. (2000). «Previsions econòmiques per als països de la zona euro», Research report for *"La Caixa d'Estalvis i Pensions de Barcelona"*. (Economic forecasts for the Euro Zone countries).

Forecasting at the regional level and conjunctural analysis

Clar, M. and Ramos, R. (several years). «Situación actual y perspectivas de las regiones de España-Cataluña», *Jornadas HISPALINK*, Instituto L. R. Klein y Consejo Superior de Cámaras de Comercio, Industria y Navegación de España, Madrid. (Present situation and economic perspectives of Spanish regions).

Clar, M. and Ramos, R. (2001). «Un modelo econométrico para predecir el VAB subsectorial de la economía catalana», in *Diez Años de análisis regional: el proyecto Hispalink*, Madrid, 63-75. Depósito legal M-14950-2001. (An Econometric model to forecast the subsectoral GAV of Catalan Economy).

Clar, M., Ramos, R. and Suriñach, J. (2001). «A state-space approach for measuring regional manufacturing production indices», *Regional Science*, 80, 357-369.

Clar, M., Ramos, R. and Suriñach, J. (2000). «Avantatges i inconvenients de la metodologia de l'INE per a elaborar indicadors de la producció industrial per a les regions espanyoles», *Qüestiió*, 24, 151-186. (Advantages and inconvenients of INE's methodology to elaborate industrial production indices for the Spanish regions).

Forecasting: methodological issues

Ramos, R., Clar, M. and Suriñach, J. (2000). «Comparación de la capacidad predictiva de los modelos de coeficientes fijos frente a variables en los modelos econométricos regionales: un análisis para Cataluña», *Estudios de Economía Aplicada*, 15, 125-162. (Comparison of the predictive accuracy of fixed coefficients models versus variable coefficients models in regional econometric models: an analysis for Catalonia).

Artís, M., Clar, M., Ramos, R., Sansó, A. and Suriñach, J. (1999). «Metodologia per a l'anàlisi de les previsions econòmiques dels països de la zona euro», Research report for *"La Caixa d'Estalvis i Pensions de Barcelona"*. (Methodology for the analysis of economic forecasts in the Euro zone countries).

Theoretical articles

Del Barrio, T., Pons, E. and Suriñach, J. (2002). «The effects of working with seasonally adjusted data when testing for unit root». *Economics Letters*, 75, 249-256.

Barrio, T., Sur, A. and Suriñach, J. (2001). «Comportamiento de los contrastes ADF, PP y KPSS al trabajar con series ajustadas de estacionalidad», *Qüestiió*, 25, 19-46.

Carrión, J. Ll., Sansó, A. and Artís, M. (2001). «Unit root and stationarity tests' wedding». *Economic Letters*, 70, 1-8.

Sansó, A., Artís, M. and Suriñach, J. (1999). «Comportamiento en muestra finita de los contrastes de integrabilidad estacional para datos mensuales: Un ejercicio de simulación». *Revista Estadística Española*, 39, 141-184.

Carrión, J. Ll., Sansó, A. and Artís, M. (1999). «Response surfaces estimates for the Dickey-Fuller unit root test with structural breaks». *Economic Letters*, 63, 279-283.

Sansó, A., Artís, M. and Suriñach, J. (1998). «Una nota sobre el contraste de relaciones de cointegración entre índices de precios». *Qüestiió*, 22, 83-102.

Other articles

Del Barrio, T., López-Bazo, E. and Serrano, G. (2002). «New evidence on international R&D spillovers, human capital and productivity in the OECD». *Economics Letters*, 77, 41-45.

Artís, M., Carrión, J. L., Costa, A. and Suriñach, J. (2002). «Smoothing the Catalan tourism micro-data time series». *Qüestiió*, 26, 197-211.

López-Bazo, E., del Barrio, T. and Artís, M. (2002). «The regional distribution of Spanish unemployment. A spatial analysis». *Papers in Regional Science*, 81, 365-389.

López-Bazo, E., del Barrio, T. and Artís, M. (2002). «La distribución provincial del desempleo en España». *Papeles de Economía*, 93, 195-208.

Artís, M., Romaní, J. and Suriñach, J. (2000). «Determinants of individual commuting in Catalonia, 1986-1991». *Urban Studies*, 37, 1431-1450.

Referència d'avaluadors i col·laboradors de *Qüestiió*: 1998-2002

Referència d'avaluadors de *Qüestiió*: 1998-2002

Qüestiió desitja agrair a les persones següents la seva participació en qualitat d'avaluadors d'articles per a la revista des del 1998 fins al 2002, tenint en compte que en el volum 21 (1997) ja es va publicar la relació d'avaluadors que van participar en el període 1987-97. L'asterisc que precedeix a cada nom indica la seva col·laboració en més d'una ocasió.

A. Alabert
* J. M. Alonso Meijide
* T. Aluja
L. Aguilar Zuil
J. M. Angulo
* A. Arcas
A. Arcos Cebrián
R. Ardanuy
C. Armero
J. Arnau
B. C. Arnold
T. Baiget
C. Barceló
J. M. Bas
J. Basulto
E. Beamonte
* E. Boj del Val
* E. Bonet
N. Bozzo
C. Bravo Llatas
M. L. Calle
* B. Campos
* V. Campos
M. A. Canela
* J. Capellades
* E. Carbonell
F. Carmona
F. Carreras
* J. A. Casco
J. del Castillo
J. M. Cisquella
E. Cobo
* G. Coenders
* À. Costa
J. Costa Bouzas
* C. Cuadras
U. Dafni
* P. Delicado
A. del Toro
J. de Uña
I. Domínguez
J. J. Egozcue
* M. D. Esteban
M. Falguera
* M. Farré
M. Febrero Bande
E. Fernández
* F. R. Fernández García
* A. Fernández Militino
J. Ferrándiz
* J. Fortiana
* I. García Jurado
C. García Olaverri
A. García Pérez
P. García Soidán
C. Gomà
* M. Greenacre
* L. Gómez
X. Heredia
J. de la Horra
A. Igelmo
J. M. Jiménez
* O. Julià
M. J. Lombardía
J. López
M. A. López Cerdá

* A. Luceño
J. Luna
* M. Martí Recober
M. Martín
A. Martín Andrés
J. M. Martínez
* M. Masats
* J. A. Mayor Gallego
V. Meléndez
J. L. Moreno Rebollo
* N. Nabona
C. Nieto
* R. Nonell
D. Nualart
* F. Oliva
* J. Oliveres
J. M. Oller
B. O. Oluyede
J. Orbe
R. Pérez Ocón
L. Pozuela
J. M. Prada
* A. Prat Bartés
* J. Puerto
* P. Puig
V. Quesada
A. Quintela del Río
A. Ramírez Granados
* M. del Río
* E. Ripoll
R. Romero Villafranca
C. Rovira
J. M. Ruiz
* M. Ruiz Espejo
J. J. Salazar
M. Salicrú
* I. Sánchez
J. A. Sánchez
* J. M. Sarabia
* A. Satorra
C. Serrat
C. Trilla
* L. Ugarte
F. Utzet
* M. J. Valderrama
L. Varona
* E. Vegas
J. F. Vera
J. A. Vilar Fernández

Altres col·laboradors de *Qüestiió*

Qüestiió desitja agrair igualment el suport tècnic i administratiu en l'edició material de la revista per part de les persones següents: C. Coll (fins el 1990) i M. C. Pallejà (des del 1991) com a responsables de la secretaria i gestió administrativa a càrrec de les universitats patrocinadores (UPC i UB, respectivament); R. M. Capo (fins el 1993), M. M. Jaime (des del 1994) i T. Patau (des del 2002) com a responsables de la secretaria i de la gestió administrativa a càrrec de l'Institut d'Estadística de Catalunya; F. Nogueras, responsable de la reprografia de l'Institut d'Estadística de Catalunya i, finalment, M. Aicart, responsable de la composició de cada número de la revista.

D'altra banda, també s'agraeix la col·laboració d'E. Aznar, C. Gómez, N. Ruíz i G. Quesada, de l'Institut d'Estadística de Catalunya, que, amb la seva col·laboració especialitzada, han ajudat a l'acurada edició d'aquesta revista.

Estadística

QÜESTIIÓ, vol. 26, 3, p. 429-441, 2002

IMPLEMENTACIÓN DEL CÁLCULO DE POLINOMIOS ZONALES Y APLICACIONES EN ANÁLISIS MULTIVARIANTE

J. RODRÍGUEZ-AVI
A. J. SÁEZ-CASTILLO
A. CONDE-SÁNCHEZ
Universidad de Jaén*

En este trabajo se describe la implementación de un algoritmo para el cálculo de polinomios zonales, así como dos aplicaciones explícitas de éstos en el ámbito del análisis multivariante. Concretamente, esta implementación permite obtener resultados de sumación aproximados para funciones hipergeométricas de argumento matricial que, a su vez, pueden utilizarse en la génesis de distribuciones multivariantes discretas con frecuencias simétricas. De igual forma, se pone en práctica un conocido resultado teórico que caracteriza la distribución de la menor raíz característica de una matriz aleatoria con distribución de Wishart.

Implementing calculus of zonal polynomials and applications in Multivariate Analysis

Palabras clave: Polinomios zonales, distribuciones discretas, distribución de Wishart, funciones hipergeométricas de argumento matricial

Clasificación AMS (MSC 2000): 60E05, 62E15

* Dep. de Estadística e I.O. Despacho 147, Ed. D-3, Paraje Las Lagunillas, Universidad de Jaén, 23071 Jaén, España. Telf: + 34 953 012.207; fax: + 34 953 012.222; email: jravi@ujaen.es.

- Recibido en diciembre de 2000.
- Aceptado en abril de 2002.

1. INTRODUCCIÓN

Los polinomios zonales son una extensión multivariante de las funciones potenciales. La familia de polinomios simétricos y homogéneos que constituyen ha sido ampliamente utilizada dentro de la Estadística Matemática clásica en la expresión de densidades y distribuciones de formas cuadráticas en poblaciones normales, para extender resultados conocidos de la estadística univariante a un ambiente multivariante (Muirhead, 1982). No obstante, no se conocen fórmulas generales para estos polinomios, ya que su definición, como vamos a ver a continuación, los caracteriza como autofunciones de un cierto operador diferencial.

En concreto, dada una matriz simétrica $X_{m\times m}$ con raíces características $x_1,\ldots,x_m$ y $\kappa=(k_1,\ldots,k_m)$ una partición de k en no más de m partes, el polinomio zonal de X correspondiente a κ, denotado por $C_\kappa(X)$, es un polinomio simétrico homogéneo de grado k en las raíces características $x_1,\ldots,x_m$ que debe verificar las siguientes tres condiciones:

1. El término de mayor peso es, salvo constante, $x_1^{k_1}\cdots x_m^{k_m}$.
2. Verifica la ecuación diferencial

   (1) $$\Delta_X C_\kappa(X)=\alpha_\kappa C_\kappa(X)$$

   donde
   $$\Delta_X=\sum_{i=1}^{m}x_i^2\frac{\partial^2}{\partial x_i^2}+\sum_{i=1}^{m}\sum_{1=j\neq i}^{m}\frac{x_i^2}{x_i-x_j}\frac{\partial}{\partial x_i}$$
   y
   $$\alpha_\kappa=\sum_{i=1}^{m}k_i(k_i-i)+k(m-1).$$
3. $(trX)^k=(x_1+\cdots+x_m)^k=\sum_\kappa C_\kappa(X)$, donde por $\sum_\kappa$ denotamos la suma en todas las particiones del grado k.

Alternativamente a esta última condición, que se utiliza para la normalización de los polinomios zonales, puede considerarse la evaluación de estos polinomios en un punto concreto. En este sentido

(2) $$C_\kappa(I_m)=2^{2k}k!\left(\frac{1}{2}m\right)_\kappa\frac{\prod_{i<j}^{p}(2k_i-2k_j-i+j)}{\prod_{i<j}^{p}(2k_i+p-i)!}$$

donde p es el número de partes no nulas de κ y, en general,

$$(a)_\kappa = (a)_{k_1}\left(a-\frac{1}{2}\right)_{k_2}\cdots\cdots\left(a-\frac{1}{2}(m-1)\right)_{k_m}, \tag{3}$$

con $(a)_k = a(a+1)\cdots(a+k-1), (a)_0 = 1$.

2. ALGORITMO DE CÁLCULO

Como se ha comentado, de la definición no puede deducirse una fórmula explícita, aunque sí un algoritmo de cálculo (James, 1968), que describimos seguidamente.

Sea $\kappa = (k_1,\ldots,k_m)$ una partición del entero k. Entonces, el monomio simétrico de una matriz $X_{m\times m}$ con raíces características $x_1,\ldots,x_m$ correspondiente a κ se define como

$$M_\kappa(X) = x_1^{k_1}\cdots x_m^{k_m} + \text{términos simétricos.}$$

El algoritmo de James establece que los polinomios zonales son combinaciones lineales de monomios simétricos. Concretamente:

$$C_\kappa(X) = \sum_{\lambda\le\kappa} c_{\kappa,\lambda} M_\lambda(X), \tag{4}$$

donde la sumatoria se hace sobre todas las particiones λ de k tales que $\lambda \le \kappa$ y

$$c_{\kappa,\lambda} = \sum_{\lambda<\mu\le\kappa} \frac{[(l_i+t)-(l_j-t)]}{\rho_\kappa - \rho_\lambda} c_{\kappa,\mu}. \tag{5}$$

En esta expresión,

$$\rho_\kappa = \sum_{i=1}^{m} k_i(k_i-i); \tag{6}$$

además, $\lambda = (l_1,\ldots,l_m)$ y $\mu = (l_1,\ldots,l_i+t,\ldots,l_j-t,\ldots,l_m)$, para $t = 1,\ldots,l_j$, son tales que cuando las partes de la partición μ se organizan en orden descendente, μ es mayor que λ y menor o igual que κ en orden lexicográfico. La sumatoria en (5) es sobre todas las posibles μ, y una suma vacía se toma como cero.

Mediante este algoritmo se determinan todos los coeficientes $c_{\kappa,\lambda}$ salvo el de mayor peso; es decir, se determina el polinomio zonal salvo normalización. Desde el punto de vista computacional, el algoritmo puede inicializarse considerando $c_{\kappa,\kappa} = 1$ y calculando los coeficientes $c_{\kappa,\lambda}$ salvo constante multiplicativa, que se obtiene mediante (2).

La implementación del algoritmo se ha realizado empleando el programa MATLAB. Los correspondientes archivos fuente y las instrucciones para su ejecución no aparecen aquí por brevedad, aunque pueden encontrarse en la dirección

http://www.ujaen.es/dep/estinv/pdi/ajsaez/software.html.

3. APLICACIONES

En este apartado se presentan sendas aplicaciones en el ámbito de la estadística multivariante que pueden llevarse a cabo gracias al conocimiento explícito de un buen número de polinomios zonales calculados mediante la metodología expuesta en la sección anterior.

3.1. Aproximación de funciones hipergeométricas de argumento matricial y génesis de distribuciones multivariantes

Se define la función hipergeométrica de argumento matricial de parámetros $a_1, \ldots, a_p$ y $b_1, \ldots, b_q$ asociada a la matriz simétrica $X_{m\times m}$ como

$$(7) \qquad {}_pF_q(a_1,\ldots,a_p;b_1,\ldots,b_q;X) = \sum_{k=0}^{\infty}\sum_{\kappa}\frac{(a_1)_\kappa\cdots(a_p)_\kappa}{(b_1)_\kappa\cdots(b_q)_\kappa}\frac{C_\kappa(X)}{k!},$$

donde $\sum_\kappa$ denota la suma sobre todas las particiones $\kappa = (k_1, \ldots, k_m)$ de k y $C_\kappa(X)$ es el polinomio zonal de X correspondiente a κ. Ningún parámetro b_j del denominador puede ser cero o un medio de un entero menor o igual que $\frac{1}{2}(m-1)$, ya que, en caso contrario, alguno de los denominadores de la serie se anularía. De igual forma, si algún parámetro del numerador es un entero negativo, esto es, $a_i = -n$, entonces la función es un polinomio de grado mn.

La serie converge para todo X si $p \leq q$; converge para $\|X\| < 1$ si $p = q+1$, donde por $\|X\|$ notamos el máximo de los valores absolutos de las raíces características de X; y salvo que la serie tenga un número finito de términos no nulos, diverge para todo $X \neq 0$ si $p > q+1$.

En el caso $m = 1$, la serie (7) se reduce la función hipergeométrica generalizada clásica

$${}_pF_q(a_1,\ldots,a_p;b_1,\ldots,b_q;x) = \sum_{k=0}^{\infty}\frac{(a_1)_k\cdot\cdots\cdot(a_p)_k}{(b_1)_k\cdot\cdots\cdot(b_q)_k}\frac{x^k}{k!}.$$

Podemos subrayar dos casos especiales de (7): los dados por la función ${}_0F_0(X) = etr(X)$, generalización de la serie exponencial, y la función ${}_1F_0(a;X) = \det(I_m - X)^{-a}$,

generalización de la serie binomial $(1-x)^{-a}$. Más allá de estos dos casos no existen expresiones generales de las funciones hipergeométricas de argumento matricial. Es por ello que para obtener resultados de sumación se suele acudir a aproximaciones numéricas mediante técnicas de análisis que proporcionan el valor de las funciones en algunos casos concretos.

Nosotros proponemos desarrollar estas funciones como series de monomios simétricos, utilizando la implementación del algoritmo de James. De esta forma, pueden obtenerse resultados aproximados de sumación de dichas funciones para un amplio rango de valores de los parámetros. Asimismo, este desarrollo en serie es el adecuado para utilizarlas como generatrices de distribuciones multivariantes discretas de probabilidad, distribuciones que cuentan con una interesante propiedad de simetría en sus frecuencias.

Tabla 1. Desarrollo truncado de la función ${}_2F_1(a,b;c;I_2)$.

a	b	c	${}_2F_1(a,b;c;I_2)$	$\sum_{k=0}^{20}\sum_{\rho} f_\rho M_\rho(I_2)$	Diferencia
2	2	10	3.0569	3.0547	0.0022
3	3	10	20.0909	19.0750	1.0159
5	5	25	13.8359	13.8299	0.006
4	5	20	15.7342	15.7046	0.0296
8	10	50	55.1096	55.0856	0.024
10	10	50	172.7901	172.2222	0.5679
12	12	65	260.1934	259.5012	0.6922
15	15	85	699.2147	695.7047	3.51

Concretamente, si en el desarrollo de la serie hipergeométrica dada en (7), expresamos los polinomios zonales en términos de los monomios simétricos, se tiene que

$$
\begin{aligned}
(8)\qquad {}_pF_q(a_1,\ldots,a_p;b_1,\ldots,b_q;X) &= \sum_{k=0}^{\infty}\sum_{\kappa}\frac{(a_1)_\kappa\cdots(a_p)_\kappa}{(b_1)_\kappa\cdots(b_q)_\kappa}\frac{C_\kappa(X)}{k!}\\
&= \sum_{k=0}^{\infty}\sum_{\kappa}\frac{(a_1)_\kappa\cdots(a_p)_\kappa}{(b_1)_\kappa\cdots(b_q)_\kappa}\frac{\sum_{\lambda\le\kappa}c_{\kappa,\lambda}M_\lambda(X)}{k!}\\
&= \sum_{k=0}^{\infty}\sum_{\kappa}\sum_{\lambda\le\kappa}\frac{1}{k!}\frac{(a_1)_\kappa\cdots(a_p)_\kappa}{(b_1)_\kappa\cdots(b_q)_\kappa}c_{\kappa,\lambda}M_\lambda(X)\\
&= \sum_{k=0}^{\infty}\sum_{\rho} f_\rho M_\rho(X),
\end{aligned}
$$

donde de nuevo la suma $\sum_\rho$ es en todas las particiones ρ de cada entero k. Una vez calculados los coeficientes f_ρ, la serie puede aproximarse mediante sus sumas parciales.

En orden a considerar la bondad de estas aproximaciones, hay que tener en cuenta que la cola de la serie truncada en el grado r, contiene tan sólo términos en las raíces características de X de grado superior a r.

En la Tabla 1 se muestran resultados aproximados de sumación referidos a la función bivariante ${}_2F_1$, para distintos valores de los parámetros cuando $X = I_2$. Todos ellos se comparan con el valor exacto de la suma, conocido gracias al Teorema de Sumación de Gauss en su versión multivariante. En general, podemos establecer una cota del error, ε, cometido mediante la suma de los $M+1$ primeros términos cuando $a_1, a_2, b_1 > 0, b_1 > a_1 + a_2$ y $\|X\| \leq 1$, dada por

$$\begin{aligned} \sum_{k=M+1}^{\infty} \sum_{\kappa} \frac{(a_1)_\kappa (a_2)_\kappa}{(b_1)_\kappa} \frac{C_\kappa(X)}{k!} &\leq \sum_{k=M+1}^{\infty} \sum_{\kappa} \frac{C_\kappa(X)}{k!} \\ &= \sum_{k=M+1}^{\infty} \frac{traza(X)^k}{k!} \\ &\leq \sum_{k=M+1}^{\infty} \frac{m^k}{k!} \\ &= e^m - \sum_{k=0}^{M} \frac{m^k}{k!} \end{aligned} \tag{9}$$

Para una discusión más detallada de la metodología de este método de sumación puede verse Gutiérrez *et al.* (2000).

Por otra parte, las funciones hipergeométricas clásicas de argumento escalar, fundamentalmente la función de Gauss ${}_2F_1$ y sus extensiones y particularizaciones, pueden emplearse para generar las distribuciones discretas más usuales (Johnson, Kotz y Kemp, 1992; Gutiérrez y Rodríguez, 1997; Rodríguez *et al.*, 2001*a*). También se han introducido extensiones bivariantes de algunas funciones hipergeométricas de argumento escalar, como la F_1 o la F_3, para generar distribuciones bivariantes discretas.

En este sentido, las funciones hipergeométricas de argumento matricial también pueden utilizarse para generar distribuciones discretas, en este caso multivariantes, que generalicen sus homólogas univariantes. Para ello tan sólo es necesario considerar el desarrollo en serie dado en (8), de manera que las probabilidades de estas distribuciones vienen dadas por las constantes f_ρ convenientemente normalizadas. La principal característica de estas distribuciones es que sus frecuencias son invariantes frente a permutaciones de las variables que la forman; es decir, si $X_{m\times 1}$ es un vector aleatorio generado por una función de este tipo, se verifica:

$$P[X = (i_1, \ldots, i_m)] = P[X = \sigma(i_1, \ldots, i_m)], \tag{10}$$

donde $(i_1, \ldots, i_m)$ es cualquier vector posible para X y σ es cualquier permutación. Esta propiedad hace que estas distribuciones sean modelos adecuados para fenómenos

multivariantes discretos cuyas componentes no son necesariamente independientes pero sí idénticamente distribuidas, de manera que las probabilidades de ocurrencia no dependan del orden de las variables. Así, se han utilizado con éxito, por ejemplo, en la modelización del número de ingresos diarios en dos boxes del Servicio de Urgencias del Hospital de San Agustín (Linares, España) (Rodríguez *et al*, 2001*b*). Otros ámbitos adecuados de aplicación aparecen, por ejemplo, en la modelización del número de fallos en varias máquinas indistinguibles en paralelo o el número de llamadas a varios servidores idénticos dispuestos también en paralelo.

En la Figura 1 se muestran ejemplos de poligonales de frecuencias generadas mediante funciones ${}_1F_1$ y ${}_2F_1$ bivariantes.

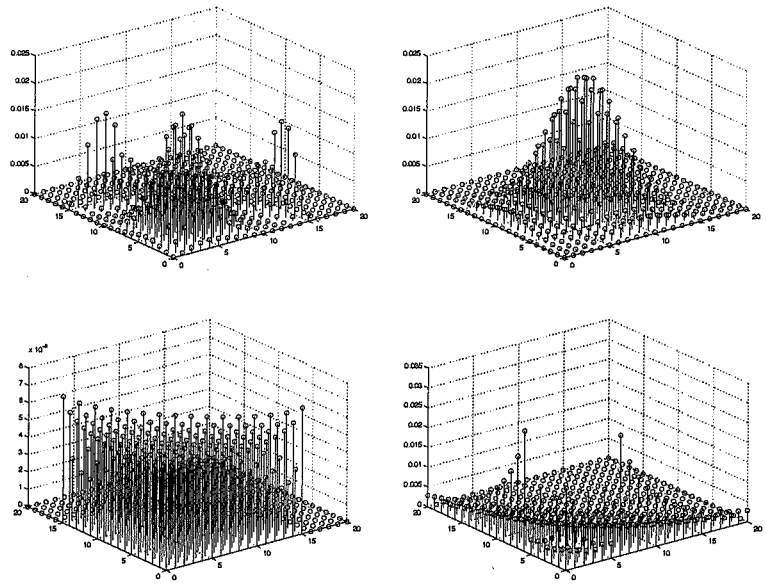


Figura 1. Poligonal de frecuencias para distribuciones generadas por funciones ${}_1F_1$ (*a.* y *b.*) y ${}_2F_1$ (*c.* y *d.*).

3.2. Distribución de la menor raíz característica de una matriz de Wishart en términos de polinomios zonales

Dentro de la Estadística Matemática, los polinomios zonales y las funciones hipergeométricas de argumento matricial han sido utilizados, entre otras aplicaciones, en la expresión de la distribución de formas cuadráticas asociadas a poblaciones normales. Así, por ejemplo, expresiones que involucran funciones hipergeométricas aparecen en

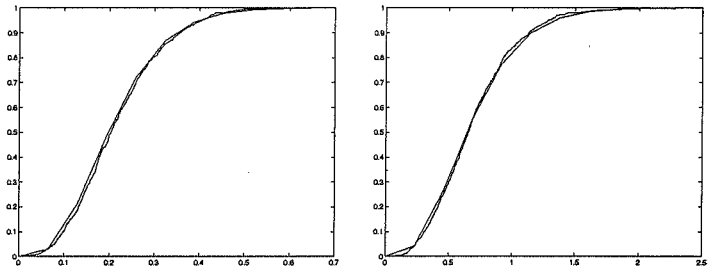


Figura 2. Funciones de distribución teórica y empírica para la menor raíz característica de $A^1 \longrightarrow W_3(10,\Sigma_1)$ (*a.*) y $A^2 \longrightarrow W_3(10,\Sigma_2)$ (*b.*).

la distribución χ^2 (Pearson, 1900), de Wishart central y no central (Wishart, 1928), de las raíces características de la matriz de covarianza de una distribución gaussiana (James, 1964), en la ratio de la varianza F (Fisher, 1924), de la F no central con p y n grados de libertad (Fisher, 1924), etc. En este apartado describimos, precisamente, una aplicación de la implementación realizada que nos permite evaluar la distribución de probabilidad de la menor raíz característica de una matriz de Wishart.

Concretamente (Muirhead, 1982), si l_m es la menor raíz característica de S, donde $A = nS$ es $W_m(n,\Sigma)$, y $r = \frac{1}{2}(n-m-1)$ es un entero positivo, entonces

$$P_\Sigma(l_m > x) = etr\left(-\frac{1}{2}nx\Sigma^{-1}\right)\sum_{k=0}^{mr}\sum_{\kappa}{}^{*}\frac{C_\kappa\left(\frac{1}{2}nx\Sigma^{-1}\right)}{k!}, \tag{11}$$

donde $\sum_\kappa{}^*$ denota la suma sobre todas las particiones $\kappa = (k_1,\ldots,k_m)$ de k tales que $k_1 \leq r$.

La puesta en práctica de este teorema tan sólo implica el desarrollo de una serie finita en términos de polinomios zonales, así que puede realizarse de manera exacta hasta distintos valores de m y n, aunque no muy elevados.

Nosotros hemos considerado una aplicación en donde $m = 3$ y $n = 10$. Se han simulado mediante el programa matemático MATLAB 1.000 valores de matrices aleatorias $X_{n\times m} \longrightarrow N(0, I_n \otimes \Sigma)$. A partir de estas 1.000 matrices, $\mathbf{x}^1_{n\times m},\ldots,\mathbf{x}^{1,000}_{n\times m}$, se calculan 1.000 valores simulados de sendas distribuciones de Wishart, mediante la expresión $\mathbf{A}^i = \mathbf{x}^{i\prime}\cdot\mathbf{x}^i$, $i = 1,\ldots,1{,}000$, de modo que la distribución teórica de estas matrices es la de $A \longrightarrow W_3(10,\Sigma)$. Posteriormente se han dividido estas matrices por $n = 10$ y a cada una de estas nuevas matrices se les ha calculado su menor raíz característica. Esto genera una muestra aleatoria simple, $l_1,\ldots,l_{1,000}$, de 1.000 valores de la variable aleatoria $\mathbf{l}_m$, menor raíz característica de A/n.

En la Figura 2 aparecen superpuestas la función de distribución empírica, calculada a partir de sendas muestras aleatorias simples, y la función de distribución teórica, dada

por (11), para matrices

$$\Sigma_1 = \begin{pmatrix} 1 & 0{,}58 & 0{,}61 \\ 0{,}58 & 1 & 0{,}58 \\ 0{,}61 & 0{,}58 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 6{,}83 & 6{,}40 & 2{,}84 \\ 6{,}40 & 8{,}02 & 2{,}79 \\ 2{,}84 & 2{,}79 & 3{,}98 \end{pmatrix}$$

Como puede comprobarse, son prácticamente idénticas, confirmando que los datos empíricos quedan perfectamente ajustados mediante la distribución teórica.

4. DISCUSIÓN

La metodología descrita en este trabajo, que involucra el cálculo explícito de los polinomios zonales, proporciona la evaluación, exacta en algunos casos y aproximada en otros, de cualquier serie hipergeométrica de argumento matricial. Esto permite la puesta en práctica de numerosos resultados que involucran a estas series y que hasta ahora, o no se habían aplicado a casos reales, o si se habían aplicado, tenían que limitarse a casos donde técnicas de análisis numérico permitieran algún resultado concreto. A su vez, también permite abrir nuevas líneas de investigación, como la de la génesis de distribuciones multivariantes de probabilidad mediante funciones hipergeométricas de argumento matricial.

Hay que decir, no obstante, que el éxito de esta metodología está supeditado a las limitaciones computacionales en el cálculo de los polinomios zonales y, sobre todo, de los monomios simétricos en términos de los cuales vienen definidos.

5. AGRADECIMIENTOS

Deseamos agradecer a los referees sus comentarios y sugerencias, los cuales han sido de gran ayuda a la hora de elaborar este manuscrito.

REFERENCIAS

Gutiérrez, R. and Rodríguez, J. (1997). «Family of Pearson Discrete Distributions Generated by the Univariate Hypergeometric function ${}_3F_2(\alpha_1, \alpha_2, \alpha_3; \gamma_1, \gamma_2; \lambda)$». *Appl. Stochastics Models and Data Anal*, 13, 115-125.

Gutiérrez R., Rodríguez J. and Sáez A. J. (2000). «Approximation of hypergeometric functions of matricial argument through their development in series of zonal polynomials». *Electronic Transactions in Numeric Analysis*, 11, 121-130.

James, A. T. (1961). «Zonal polynomials of the real positive definite symmetric matrices». *Annals of Mathematics*, 74, 456-469.

James, A. T. (1968). «Calculation of zonal polynomial coefficient by use of the Laplace-Beltrami operator». *Ann. Math. Statis.*, 39, 1711-1718.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*. Wiley, New York. Second Edition.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. Wiley. New York.

Rodríguez, J., Conde, A. and Sáez, A. (2001*a*). «A new class of discrete distributions with complex parameters». *Accepted in StatisticalPapers*.

Rodríguez, J., Conde, A., Olmo, M. J. and Sáez A. J. (2001*b*). «Discrete multivariate distributions symmetric in frequencies generated by the Gauss function ${}_2F_1$ of matricial argument». *(Submmited to Test)*.

Takemura, A. (1984). «Zonal Polynomials». *Institute of Mathematical Statistics*. Hayward, California.

ENGLISH SUMMARY

IMPLEMENTING CALCULUS OF ZONAL POLYNOMIALS AND APPLICATIONS IN MULTIVARIATE ANALYSIS

J. RODRÍGUEZ-AVI
A. J. SÁEZ-CASTILLO
A. CONDE-SÁNCHEZ
Universidad de Jaén*

The implementation of an algorithm that permits the calculation of zonal polynomials and two applications in multivariate analysis are described in this paper. This implementation provides summation results of hypergeometric functions of matricial argument that may be used in generating discrete multivariate distributions whose frequencies are symmetrical. Development of hypergeometric series of matricial argument also allows a well-known result that characterizes the distribution of the lowest eigenvalue of a random matrix with Wishart distribution to be put into practice.

Keywords: Zonal polynomials, discrete distributions, Wishart distribution, hypergeometric functions of matricial argument

AMS Classification (MSC 2000): 60E05, 62E15

* Dep. de Estadística e I.O. Despacho 147, Ed. D-3, Paraje Las Lagunillas, Universidad de Jaén, 23071 Jaén, España. Telf: + 34 953 012.207; fax: + 34 953 012.222; email: jravi@ujaen.es.

- Received December 2000.
- Accepted April 2002.

Zonal polynomials constitute a family of symmetrical and homogeneous polynomials. They may be considered to be a multivariate extension of powers of a single variable and are defined as eigenfunctions of the differential operator

$$\Delta_X = \sum_{i=1}^{m} x_i^2 \frac{\partial^2}{\partial x_i^2} + \sum_{i=1}^{m} \sum_{1=j\neq i}^{m} \frac{x_i^2}{x_i - x_j} \frac{\partial}{\partial x_i},$$

where

$$\alpha_\kappa = \sum_{i=1}^{m} k_i (k_i - i) + k(m-1),$$

in such a way that there is no a explicit general formula for their evaluation.

Alternatively, James (1968) establishes an algorithm that permits zonal polynomials to be express as linear combinations of symmetric monomials, which are a common base of symmetrical and homogeneous polynomials. This algorithm has been implemented employing MATLAB. Source files may be obtained from

http://www.ujaen.es/dep/estinv/pdi/ajsaez/software.html

Moreover, hypergeometric functions of matricial argument,

$${}_pF_q(a_1,\ldots,a_p;b_1,\ldots,b_q;X) = \sum_{k=0}^{\infty} \sum_{\kappa} \frac{(a_1)_\kappa \cdots (a_p)_\kappa}{(b_1)_\kappa \cdots (b_q)_\kappa} \frac{C_\kappa(X)}{k!},$$

where $\sum_\kappa$ denotes the sum over all partitions $\kappa = (k_1,\ldots,k_m)$ of k and $C_\kappa(X)$ is the zonal polynomial of X corresponding to κ, are multivariate extension of classical hypergeometric functions, where Pochammer symbols are replaced by generalized hypergeometric coefficients and powers of single variables by zonal polynomials. As neither have an exact formula, we propose a development of the hypergeometric series in terms of symmetric monomials using the James algorithm for zonal polynomials: in this manner we obtain approximate values of the series.

In addition, hypergeometric functions of matricial argument may be used to generate probability functions of multivariate discrete distributions; these distributions extend the family of well-known univariate distributions generated by classical hypergeometric functions (Johnson, Kotz and Kemp, 1992). In this context, the development of hypergeometric functions of matricial argument considered in this paper is suitable for easily obtaining frequencies of the distributions as coefficients of the series in terms of symmetric monomials. The main characteristic of these distributions is that frequencies are invariable when permutations of the variables in the random vector are considered: this property is due to the symmetry of zonal polynomials.

In the field of Mathematical Statistics, hypergeometric series of matricial argument have been mainly used in expressing multivariate distributions of quadratic forms in normal

samples that extend analogous results in a univariate environment. Many of these results involve hypergeometric series without explicit expression, so they cannot be put into practice. In this sense, we consider an example where the distribution of the lowest eigenvalue of a random matrix with Wishart distribution is expressed as a finite sum of zonal polynomials. Specifically, if l_m is the lowest eigenvalue of S, where $A = nS$ is $W_m(n,\Sigma)$, and $r = \frac{1}{2}(n-m-1)$ is an positive integer, then (Muirhead, 1982)

$$P_\Sigma(l_m > x) = etr\left(-\frac{1}{2}nx\Sigma^{-1}\right)\sum_{k=0}^{mr}\sum_{\kappa}{}^{*}\frac{C_\kappa\left(\frac{1}{2}nx\Sigma^{-1}\right)}{k!},$$

where $\sum_\kappa{}^*$ denotes sum over all partitions $\kappa = (k_1,\ldots,k_m)$ of k such that $k_1 \leq r$. We have expressed this result in terms of symmetric monomials in such a way that we can calculate exact probabilities of the distribution. The paper concludes with a simulation to show that the theoretical distribution function is close to the empirical distribution functions when we consider sample data for a Wishart sample.

QÜESTIIÓ, vol. 26, 3, p. 443-459, 2002

ANALYSIS ON THE INDIVIDUAL EFFICIENCY PREDICTION IN THE COMPOSED ERROR FRONTIER MODEL. A MONTE CARLO STUDY

R. DIOS-PALOMARES
A. RAMOS-MILLÁN
J. A. ROLDÁN-CASAS
University of Córdoba*

This study seeks to analyse some important questions related to the Stochastic Frontier Model, such as the method proposed by Jondrow et al *(1982) to separate the error term into its two components, and the measure of efficiency given by Timmer (1971). To this purpose, a Monte Carlo experiment has been carried out using the Half-Normal and Normal-Exponential specifications throughout the rank of the* γ *parameter. The estimation errors have been eliminated, so that the intrinsic variability of the conditional of u given* ε *can be evaluated. In addition, the behaviour of the mean and mode as point estimators of u is investigated. The results have yielded some interesting findings. We have observed that both the point estimates and the mean efficiency are more precise in cases of lower efficiency. This occurs when the variable that generates the inefficiency outweighs the one that picks up the errors out of the control.*
The change in order found between the estimated efficiency and its true value is misleadingly high especially for low γ*, which underlines the risk of estimating at these values.*

Keywords: Bias, MSE, point estimator, production frontier, Jondrow formulae

AMS Classification (MSC 2000): 62F12, 65C05

* Department of Statistics E.T.S.I.A.M. University of Córdoba. Apdo. 3048, 14080 Córdoba. Spain. Telephone number: 957 218 479. Email: ma1dipar@uco.es.

- Received July 2001.
- Accepted October 2002.

1. INTRODUCTION

A number of studies have already been carried out on the subject of productivity. Nowdays, the two methodologies most used to estimate efficiency by means of the Frontier Production are Linear Programming applying DEA, and the Stochastic Frontier Production (Henceforth SFP, also called 'Composed error'). The second of these will be the subject of the present work.

The Stochastic Frontier Production function was proposed independently by Aigner, Lovell and Schmidt (1977) and by Meeusen and Van den Broeck (1977). The general model of an SFP is as follows:

$$y_i = X_i * \beta + \varepsilon_i \qquad (i = 1,2,\ldots N) \quad \varepsilon_i = v_i - u_i$$

where y_i denotes the output for observation i, X_i the vector of inputs for observation i, β is a vector of parameters, N is the sample size, and the variable error ε_i collects the difference between the systematic part of the model noted v_i and the observed values u_i.

The Composed Error presumes that the error variable not only catches the effect of the inefficiency but also that another error exists which is not controllable by the firms. This last is also included in the composed error. It is therefore assumed that the ε variable is generated as the difference between a stochastic variable v (not controllable, symmetric, and defined between $-\infty$ and ∞) and the inefficiency variable u which will always be positive and assymmetric. The v variable is assumed to be Normally distributed with mean zero and variance σ_v^2, and the component u is either Exponential, Half-normal, Truncated Normal or Gamma.

Once a specification is presumed for the u component, the estimation of the production frontier model is accomplished by assigning the corresponding distribution to ε as the difference between v and u. The residues are attained as an immediate result from the estimation ($\varepsilon_i = y_i - X_i * \hat{\beta}$). These may be considered as estimates of the error terms. However, the problem of decomposing these estimates into separate estimates of the constituent parts has remained unsolved for some time. Of course, the average technical inefficiency –the mean of the distribution of u_i– is easily calculated, but the main issue is how to get the relative importance of the elements and thus, to be able to match the results.

The study of efficiency by means of the SFP model was significantly developed thanks to the contribution of J. Jondrow, C. A. K. Lovell, I. S. Materov, and P. Schmidt (Henceforth, JLMS) who in 1982 proposed a formulae which made it possible to separate the two components of the ε variable.

The above mentioned formula constituted a milestone, and has remained widely used ever since, regardless of its possible weaknesses. Concerning this, we must point out

that the estimation of a particular value of u by means of the mode or the expectation of the whole conditional distribution introduces a concentration effect into the distribution of the estimated inefficiency.

Although Olson *et al.* (1980) conducted a noteworthy experiment dealing with the Stochastic Frontier Production, the objective of the research focussed on the mean efficiency estimation using Corrected Least Squares.

The results of recent Monte Carlo experiments in which the JLMS performance was analysed can be found in Coelli (1995) and Kumbhakar and Löthgren (1998). In these papers the bias and the variance of the mean efficiency estimate were studied, considering the gamma parameter as a source of variation. The conclusion was that both the bias and the precision present their worst results when gamma takes central values: i.e. near 0.5.

Nevertheless, nothing in relation with the accuracy of the individual estimation of the efficiency has been taken into consideration in previous research. We consider that special attention should be paid to the change in the order in efficiency that occurs in the estimation process. It would be worthwhile to carry out further studies into the evaluation of the error introduced by the JLMS method.

We have conducted a Monte Carlo experiment in order to investigate the effect of the main sources of variation on the estimation error and the change in the efficiency order. These sources are: The value of the gamma parameter, the inefficiency distribution, the sample size and the employed formula. In respect of this last, either the mode or the expectation may be used.

The present paper is organised as follows. In Section 2 the JLMS method is briefly introduced, in section 3 the design of the Monte Carlo study is described. Section 4 reports the main results of the experiment and finally, Section 5 shows the conclusions of the study.

The JLMS formulation

The JLMS method widely develops the use of SFP models in empirical applications. The researchers arrived at the formulae for the separation of the u_i component from the residues ε_i which last contains information on u_i. They proceeded to consider the conditional distribution of u_i, given ε_i, arguing that this distribution contains whatever information ε_i yields about u_i. Thus, either the mean or the mode of this distribution can be used as a point estimate of u_i. They explicitly dealt with the commonly assumed cases of half-normal and exponential for u_i, expressed in the two following theorems:

The half-normal case

Theorem 1. *The conditional distribution of u given ε is that of a $N(\mu_*, \sigma_*^2)$ variable truncated at zero.*

$$v_i \approx N(0, \sigma_v^2) \qquad u_i \approx |N(0, \sigma_u^2)|$$

$$\sigma^2 = \sigma_u^2 + \sigma_v^2 \qquad \mu_* = -\frac{\sigma_u^2 \varepsilon}{\sigma^2} \qquad \sigma_*^2 = \frac{\sigma_u^2 \sigma_v^2}{\sigma^2}$$

where «≈» means «distributed as» and μ_* and σ_*^2 are the mean and the variance of the conditional distribution of u given ε, respectively.

Hereafter, we will drop the subscript (i) for the sake of simplicity. Jondrow et al defined λ as $\lambda = \frac{\sigma_u}{\sigma_v}$ and produced the following two point estimators:

$$E(u|\varepsilon) = \sigma_* \left[\frac{f\left(\frac{\varepsilon\lambda}{\sigma}\right)}{1 - F\left(\frac{\varepsilon\lambda}{\sigma}\right)} - \left(\frac{\varepsilon\lambda}{\sigma}\right) \right]$$

$$M(u|\varepsilon) = -\varepsilon\left(\frac{\sigma_u^2}{\sigma^2}\right) \quad \text{if} \quad \varepsilon \leq 0$$

$$= 0 \quad \text{if} \quad \varepsilon > 0$$

where the sign «|» means «conditioned to» and $f(.)$ and $F(.)$ are the probability density and distribution functions, respectively, of a standard normal random variable.

The exponential case

Theorem 2. *The conditional distribution of u given ε is that of a $N(-\sigma_v A, \sigma_v^2)$ variable truncated at zero.*

The parameter A was defined as $A = \frac{\varepsilon}{\sigma_v} + \frac{\sigma_v}{\sigma_u}$, deducing the point estimators for u:

$$E(u|\varepsilon) = \sigma_v \left[\frac{f(A)}{1 - F(A)} - A \right]$$

$$M(u|\varepsilon) = -\varepsilon - \frac{\sigma_v^2}{\sigma_u} \quad \text{if} \quad \varepsilon \leq -\frac{\sigma_v^2}{\sigma_u}$$

$$= 0 \quad \text{if} \quad \varepsilon > -\frac{\sigma_v^2}{\sigma_u}$$

where $f(.)$ and $F(.)$ have been defined above.

The estimation of the u component is carried out by replacing the appropriate parameters in the above formulation. Obviously, as JLMS noted, the intrinsic variability of the conditional distribution of u given ε is independent of sample size, given that ε contains rudimentary information on u.

3. DESIGN OF THE MONTE CARLO STUDY

The main parameters of the experiment are: σ^2(the variance of ε), γ (variance ratio) and the sample size (N). Due to the invariance results noted by OSW (1980) only one value of the variance ($\sigma^2 = 1$) is considered.

The variance ratio γ is taken to reflect the percentage contribution of the variance of u to the total variance of the error term in the data generating process.

The random terms v_i, $i = 1,\ldots,N$, are drawn from a Normal distribution $N(0,\sigma_v^2)$ and the technical inefficiency terms u_i, from either a Normal truncated at zero from below or an Exponential distribution. ε_i is obtained by means of the expression $\varepsilon_i = v_i - u_i$ $(i = 1,2,\ldots,N)$.

No regression is involved and we assume that the usually unobservable ε and its two components are known beforehand.

Four sample sizes ($N = 25,50,100,200$) and nine variance ratios γ ($\gamma = 0,1, 0,2,\ldots, 0,9$) are considered in the study. 2000 Monte Carlo replications are involved in each combination. This gives 80000 generated data sets.

We calculated three estimates of u_i, namely, two using the JLMS expressions for the expected value ($\hat{u}_{i(E)}$), and the mode ($\hat{u}_{i(M)}$) and the third by the use of the Timmer method ($\hat{u}_{i(T)}$). As a result, three estimates of efficiency were obtained, calculated as

$$e^{-\hat{u}_{i(E)}}, \quad e^{-\hat{u}_{i(M)}}$$

and

$$e^{\hat{u}_{i(T)}} \quad \text{respectively.}$$

$\hat{u}_{i(T)}$ are attained as:

$$\hat{u}_{i(T)} = \varepsilon_i - (\text{máx}\,\varepsilon) \quad (i = 1,2,\ldots,N)$$

where (máx ε) denotes the maximum ε value of the considered sample, which contains $\hat{u}_{i(T)}$.

Thus, $\hat{u}_{i(T)}$ represents the estimated efficiency of the i-th element calculated as a deterministic frontier approach.

In each Monte Carlo replicate, the following measures are observed:

- *Disparity (D)*

  We propose this statistic in order to provide more detail on the reliability of the estimates. D is defined as follows:

$$D(\theta) = \frac{1}{N}\sum_{i=1}^{N} \left|\hat{\theta}_i - \theta_i\right| \quad (i = 1,2,\ldots,N)$$

  where θ_i denotes the i-th true value of θ, $\hat{\theta}_i$ the estimation of θ_i and N the sample size. Thus, D can be applied to $D(\hat{u}_E)$ and $D\left(e^{-\hat{u}_E}\right)$

$$D\left(\hat{u}_{(E)}\right) = \frac{1}{N}\sum_{i=1}^{N} \left|\hat{u}_{i(E)} - u_{i(E)}\right|$$

$$D\left(e^{-\hat{u}_{(E)}}\right) = \frac{1}{N}\sum_{i=1}^{N} \left|e^{-\hat{u}_{i(E)}} - e^{-u_{i(E)}}\right| .$$

- *Mean Square Error of the mean estimated efficiency. (MSEE)*

  This measure is obtained differently from that in Kumbhakar and Löthgren (1998). It is calculated using the following expression:

$$MSEE = \frac{1}{2000}\sum_{j=1}^{2000}\left(\sum_{i=1}^{N}\frac{e^{-\hat{u}_i}}{N} - E\left(e^{-u_\gamma}\right)\right)^2 + \left(\frac{\sum_{j=1}^{2000}\sum_{i=1}^{N} e^{-\hat{u}_i}}{2000} - E\left(e^{-u_\gamma}\right)\right)^2$$

  $(i = 1,2,\ldots,N)$

  $(N = 25,50,100,200)$

  where $E(e^{-u_\gamma})$ is the expected value of the efficiency distribution, and will depend on the considered γ.

- *Variance of the estimated efficiency (V)*

  This statistic is calculated as:

$$V = \frac{1}{2000}\sum_{j=1}^{2000}\left(\sum_{i=1}^{N}\frac{e^{-\hat{u}_i}}{N} - E\left(e^{-u_\gamma}\right)\right)^2$$

- *Bias of the mean estimated efficiency (B)*

  This measure is once again calculated, with relation to $E\left(e^{-u_Y}\right)$.

$$B = \left(\frac{\sum_{j=1}^{2000} \sum_{i=1}^{N} e^{-\hat{u}_i}}{2000} - E\left(e^{-u_Y}\right) \right)$$

  This is expressed as $MSEE = V + B^2$.

- *Order changes of efficiency (O)*[1].

  O is defined as: $O = \frac{1}{N} \sum_{i=1}^{N} \left| O_i - \hat{O}_i \right|$

  Where O_i indicates the order corresponding to the i-th element within its original efficiency ranking and $\hat{O}_i$ indicates the same aspect on the base of the estimated values of u[2]

- *Inter-sample category changes (C)*

  Once we have O_i and $\hat{O}_i (i = 1, 2, \ldots, N)$ both series are ranked and divided into 3 categories –low, medium and high efficiency– in order to compute the elements whose categories have changed after the estimation, that is (C).

Additionally, several correlations between different elements are reported for each reiteration.

4. RESULTS

In this section we present the main results of the Monte Carlo experiment, which reveal a similar performance by the assigned specifications for ε (Normal-Halfnormal (N-HN) and Normal-Exponential (N-E)).

We present all the relevant tables at the end of the paper.

[1]This measure will not be applied to $\hat{u}_{i(M)}$, because this estimator produces several zero values, i.e. full efficiencies, and the order would be distorted.

[2]Note that O values will be identical when using either $e^{\hat{u}_{i(E)}}$ or $e^{\hat{u}_{i(T)}}$.

Tables 1 and 2 illustrate the average of D results obtained for $\hat{u}_{i(E)}$, $\hat{u}_{i(M)}$ and $e^{\hat{u}_{i(E)}}$, $e^{\hat{u}_{i(M)}}$ respectively. They show the data yielded at $\gamma = 0{,}5$, which is taken as an intermediate point in order to compare their performances through different sample sizes. Looking at tables 1 and 2, we see that D is always greater for the mode than for the expected value, which bears out the hypothesis that a high estimated error is introduced by the use of the JLMS expression for the mode as a point estimator. This fact leads us to drop $\hat{u}_{i(M)}$ from further consideration.

The figures indicate that the sample size does not affect D.

Turning our attention to the disparity of $\hat{u}_{i(E)}$ $\left(D\left(\left(\hat{u}_{i(E)}\right)\right)\right)$, we see in table 3 that a maximum was reached, located at the central γ value, and the minimum at the extremes, underlining the fact that this measure of disparity depends on σ_*^2, which, as is shown in table 4, has a maximum at intermediate values of γ.

In contrast to this, we see that the behaviour of $\left(D\left(e^{-\hat{u}_{i(E)}}\right)\right)$ is slightly different from the last one, having a maximum at $\gamma \approx 0{,}3$.

Also, it is interesting to note that D is unaffected by the sample size (see tables 3 and 5).

We found a different type of behaviour with both taken specifications of ε. Thus, the N-HN shows higher disparity than the N-E case. Their $\left(D\left(e^{-\hat{u}_{i(E)}}\right)\right)$ values show a crossing point, having the N-E lower D values than the N-HN from $\gamma = 0{,}3$ to $\gamma = 0{,}9$.

The overall performance of V and MSEE is similar to that presented by Kumbhakar and Löthgren (1998), in spite of the fact that our methodological procedure differs from theirs. Both measures have a comparable pattern (tables 6 and 7), having a maximum at approximately $\gamma = 0{,}4$, and a decreasing trend until it reaches the value of 0.9. The causes of this pattern are the same as those mentioned for D (High values of σ_*^2 at central γ's) behaviour.

Both statistics improve as sample size increases. Thus, for $N = 200$ the above mentioned decreasing trend is more marked, especially in the variance case.

The above noted suggests that more reliable results will be attained for high N and, moreover, at extreme values of γ.

With regard to the bias (B) (table 8), it is noteworthy that B is not affected by the sample size. B turns out to be always negative, following the same pattern as V, given that the highest absolute values of B are ranged between $\gamma = 0{,}3$ and $0{,}5$.

The study of the O statistic is crucial. The ranking of individuals or firms could be more relevant than efficiency in itself. Consider, for instance, the cases of applications see-

king causes of inefficiencies by the 2-stage method; under this procedure the elements are generally divided into 3 categories, depending on their ranking. This indicates the importance of O.

Table 9 shows that the predominant factor is γ. Thus, O gradually decreases from 28 (27) at $\gamma = 0{,}1$ down to 10 (8) at $\gamma = 0{,}9$ in the N-HN (N-E) case. This bears out the assumption that the greater the γ, the greater variance of u and, therefore, it will be more difficult for the generated $u_{i's}$ to have their orders altered after the estimation.

Inspection of the data on C (Tables 10 and 11) shows that the N-HN distribution presents a lower C than N-E for any γ value.

A high C can be seen especially for small and medium γ's. Consequently, about 30 % of the individuals belonging to the high efficiency category fall, after the estimation, into the group of medium efficiency, and 20 % into the lowest category. Obviously, this fact reveals a severe weakness of the JLMS method and may lead to erroneous conclusions.

The sample size does not affect the total number of category changes C, although the composition of the changes is modified. Tables 12 and 13 show the total changes between the extreme categories and the total changes among the nearest categories respectively.

It can also be observed that a significant decrease of C occurs at γ's very close to 0.9.

Moving onto the correlation study, one thing is immediately noticeable: the relationships between u and v with ε respectively (table 14), reflect that ε is indeed made up of both components. Therefore, the higher the γ, the higher (lower) the correlation of $u(v)$ with ε.

Since the components u and v have been generated independently, we observe that the correlation u and v is close to zero.

The correlation of u with $\hat{u}$ (table 15) shows that the estimation is improved as γ rises, and that we must be cautious when applying the JLMS expressions at low values of γ. The JLMS method produces a negative correlation of $\hat{u}$ with $\hat{v}$, which decreases as γ increases. Also, a close relationship appears between ε and $\hat{u}_{(E)}$, yielding an almost exact correlation, but negative.

The existing correlation between the true efficiency and $e^{-\hat{u}_{i(E)}}$ is one of the main results. As can be seen in table 16 the higher the γ value, the greater the correlation. However, this correlation is very weak at low γ's, which underlines the risk of estimating at these values.

The correlation between the efficiency estimated by Timmer and the true efficiency, when v is different from zero, tends to be higher as γ rises. Obviously, the error is

caused by the fact that the Timmer method assumes $v = 0$. Therefore, the lower the v, the higher the correlation. This correlation is always lower than that existing between e^{-u_i} and $e^{-\hat{u}_{i(E)}}$, demonstrating that from this point of view, the JLMS method, using the expected value, outperforms the Timmer one for any γ considered.

The estimates of the efficiency given by JLMS and Timmer are correlated. However, the lower the γ, the lower the correlation obtained ($r = 0{,}7$-$0{,}95$).

5. CONCLUSIONS

The present study investigates the behaviour of the Jondrow *et al.* method to estimate individual technical efficiency by means of a Monte Carlo experiment.

The following conclusions can be drawn from the results:

- The expectation estimator of u conditional on epsilon outperforms the one calculated using the mode.
- The theoretical variance of the estimator proposed by JLMS has been empirically contrasted and has been shown to be superior for intermediate values of gamma, but inferior for the extreme values. The bell-shaped performance influences considerably the disparity measures as well as the variance and MSE.
- The bias is negative in all cases, implying an underestimation of the mean efficiency.
- The sample size affects the MSE of the efficiency, improving the precision the higher the size. However, the bias and the disparity measures are not affected.
- Slight differences in the disparity and the mean variation of orders among the Normal-Half-normal and Normal-Exponential are found.
- The deterministic frontier function approach yields worse results for any value of gamma different from one, being more imprecise when gamma is lower.
- The mean variation of orders is unaffected by the variance of the point estimator of u. Thus, the former decreases monotonically when the variance ratio rises.
- It can be concluded that both the point estimates and the mean efficiency are more precise in cases of lower efficiency. That is, when the variable that generates the inefficiency outweighs the one that picks up the errors out of the control.
- The change in order found between the estimated efficiency and its true value is misleadingly high, especially for low γ. More precisely, approximately 40 % of the firms change from one group to another. Therefore, the picture that emerges is rather disappointing, since the potential conclusions drawn can be distorted.

6. REFERENCES

Aigner, D., Lovell, C. A. K. and Schmidt, P. (1977). «Formulation and Estimation of Stochastic Frontier Production Function Models», *Journal of Econometrics*, 6, 21-37.

Battese, G. E. and Coelli, T. J. (1993). «A Stochastic Frontier Production Function Incorporating a Model for Technical Inefficiency Effects», *Working Papers in Econometrics and Applied Statistics*, 69, University of New England, Armidale.

Coelli, T. J. (1995). «Estimators and Hypothesis Tests for a Stochastic Frontier Function: A Monte Carlo Analysis», *Journal of Productivity Analysis*, 6, 247-268.

Hjalmarsson, L., Kumbhakar, S. C. and Heshmati, A. (1996). «DEA, DFA and SFA: A Comparison», *Journal of Productivity Analysis*, 7, 303-327.

Horrace, W. C. and Schmidt, P. (1996). «Confidence Statements for Efficiency Estimates from Stochastic Frontier Models», *Journal of Productivity Analysis*, 7, 257-282.

Jondrow, J., Lovell, C. A. K., Materov, I. S. and Schmidt, P. (1982). «On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model», *Journal of Econometrics*, 19, 233-238.

Kumbhakar, S. and Löthgren, M. (1998). «A Monte Carlo Analysis of Technical Inefficiency Predictors», *Working Paper Series in Economics and Finance*, 229, March 1998. Stockholm School of Economics.

Meeusen, W. and van den Broeck, J. (1977). «Efficiency Estimation from Cobb-Douglas Production Function with Composed Errors», *International Economic Review*, 18, 435-444.

Olson, J. A., Schmidt, P. and Waldman, D. W. (1980). «A Monte Carlo Study of Estimators of Stochastic Frontier Production Functions», *Journal of Econometrics*, 13, 67-82.

Timmer, C. P. (1971). «Using a Probabilistic Frontier Production Functions to Measure Technical Efficiency», *Journal of Political Economy*, 79(4), 776-794.

Table 1. Disparity of u estimates based on JLMS point estimators.

gamma = 0.5	Size	N-HN.(E)	N-E.(E)	N-HN(M)	N-E.(M)
	25	0.38179651	0.34505828	0.42633239	0.43844921
	50	0.38180968	0.34532386	0.42599261	0.43909834
	100	0.38155264	0.34387603	0.42533376	0.43793732
	200	0.38143671	0.34412331	0.42505889	0.43927048

Table 2. Disparity of efficiency estimates based on JLMS point estimators.

gamma = 0.5	Size	N-HN.(E)	N-E.(E)	N-HN(M)	N-E.(M)
	25	0.156645445	0.1668595	0.194236911	0.248554577
	50	0.156475167	0.166527231	0.193707896	0.248804567
	100	0.156298243	0.165771439	0.193260251	0.248286296
	200	0.156290733	0.165932704	0.193123864	0.249090726

Table 3. Disparity of u estimates based on JLMS point estimators.

γ	N-HN.N = 25	N-HN.N = 50	N-HN.N = 100	N-HN.N = 200	NE.N = 25	NE.N = 50	NE.N = 100	NE.N = 200
0.1	0.2393	0.2376	0.2374	0.2383	0.2196	0.2195	0.2192	0.2191
0.2	0.3147	0.3136	0.3144	0.3150	0.2886	0.2869	0.2868	0.2873
0.3	0.3556	0.3565	0.3572	0.3566	0.3252	0.3259	0.3231	0.3241
0.4	0.3773	0.3766	0.3790	0.3764	0.3404	0.3415	0.3408	0.3411
0.5	0.3818	0.3818	0.3816	0.3814	0.3451	0.3453	0.3439	0.3441
0.6	0.3731	0.3722	0.3707	0.3714	0.3370	0.3363	0.3356	0.3361
0.7	0.3443	0.3460	0.3460	0.3456	0.3146	0.3149	0.3148	0.3146
0.8	0.3001	0.3005	0.3011	0.3018	0.2771	0.2762	0.2774	0.2769
0.9	0.2274	0.2277	0.2287	0.2274	0.2134	0.2138	0.2122	0.2129

Table 4. Variance of u given epsilon distribution.

γ	σ_*^2
0.1	0.2108
0.2	0.3261
0.3	0.3788
0.4	0.3883
0.5	0.3667
0.6	0.3220
0.7	0.2596
0.8	0.1833
0.9	0.0961

Table 5. Disparity of efficiency estimates based on JLMS point estimators.

γ	N-HN N = 25	N-HN N = 50	N-HN N = 100	N-HN N = 200	NE N = 25	NE N = 50	NE N = 100	NE N = 200
0.1	0.1529	0.1523	0.1523	0.1527	0.1495	0.1497	0.1496	0.1493
0.2	0.1704	0.1704	0.1704	0.1707	0.1713	0.1708	0.1706	0.1708
0.3	0.1709	0.1720	0.1719	0.1716	0.1752	0.1762	0.1750	0.1754
0.4	0.1658	0.1655	0.1664	0.1656	0.1719	0.1730	0.1730	0.1731
0.5	0.1566	0.1565	0.1563	0.1563	0.1669	0.1665	0.1658	0.1659
0.6	0.1444	0.1440	0.1440	0.1436	0.1567	0.1565	0.1564	0.1560
0.7	0.1282	0.1279	0.1285	0.1282	0.1427	0.1422	0.1428	0.1426
0.8	0.1085	0.1089	0.1087	0.1092	0.1246	0.1242	0.1241	0.1241
0.9	0.0813	0.0816	0.0820	0.0816	0.0964	0.0963	0.0957	0.0959

Table 6. Variance of efficiency estimates based on JLMS point estimators.
(N-NH specification)

γ	N = 25	N = 50	N = 100	N = 200
0.1	0.0009	0.0008	0.0008	0.0008
0.2	0.0019	0.0017	0.0016	0.0016
0.3	0.0027	0.0023	0.0021	0.0020
0.4	0.0030	0.0025	0.0022	0.0021
0.5	0.0030	0.0024	0.0021	0.0019
0.6	0.0029	0.0021	0.0018	0.0016
0.7	0.0028	0.0019	0.0015	0.0012
0.8	0.0026	0.0016	0.0011	0.0008
0.9	0.0026	0.0014	0.0008	0.0005

Table 7. MSEE of efficiency estimates based on JLMS point estimators.

(N-NH specification)

γ	N = 25	N = 50	N = 100	N = 200
0.1	0.0016	0.0015	0.0015	0.0015
0.2	0.0035	0.0031	0.0031	0.0031
0.3	0.0046	0.0042	0.0041	0.0039
0.4	0.0051	0.0044	0.0042	0.0041
0.5	0.0048	0.0041	0.0038	0.0037
0.6	0.0042	0.0035	0.0032	0.0031
0.7	0.0038	0.0029	0.0025	0.0022
0.8	0.0031	0.0021	0.0016	0.0013
0.9	0.0028	0.0016	0.0009	0.0006

Table 8. Bias of efficiency estimates based on JLMS point estimators.

(N-NH specification)

γ	N = 25	N = 50	N = 100	N = 200
0.1	−0.0269	−0.0266	−0.0269	−0.0271
0.2	−0.0389	−0.0381	−0.0390	−0.0389
0.3	−0.0442	−0.0437	−0.0440	−0.0436
0.4	−0.0457	−0.0442	−0.0447	−0.0447
0.5	−0.0419	−0.0418	−0.0420	−0.0423
0.6	−0.0369	−0.0372	−0.0377	−0.0380
0.7	−0.0323	−0.0320	−0.0321	−0.0316
0.8	−0.0229	−0.0233	−0.0237	−0.0229
0.9	−0.0134	−0.0133	−0.0136	−0.0130

Table 9. Mean order variation of efficiency estimates based on JLMS point estimators.

γ	N.HN	N.E
0.1	28.0008	27.3777
0.2	25.5301	24.6233
0.3	23.4530	22.3785
0.4	21.7847	20.2589
0.5	19.8739	18.1641
0.6	18.0790	16.1377
0.7	16.0540	13.9268
0.8	13.5566	11.3914
0.9	10.2587	8.2966

Table 10. Changes of category after the estimation of efficiency. N-HN error specification.

γ	Total Changes	Changes from 1 to 2	Changes from 1 to 3	Changes from 2 to 1	Changes from 2 to 3	Changes from 3 to 1	Changes from 3 to 2
0.1	57.564	9.89	7.336	9.342	11.83	7.884	11.282
0.2	53.274	9.64	5.516	8.986	11.808	6.17	11.154
0.3	49.562	9.234	4.374	8.8	11.39	4.808	10.956
0.4	45.928	8.682	3.424	8.302	11.048	3.804	10.668
0.5	42.574	8.034	2.496	7.912	10.818	2.618	10.696
0.6	38.246	7.42	1.764	7.242	10.028	1.942	9.85
0.7	33.856	6.7	0.928	6.628	9.336	1	9.264
0.8	28.412	5.672	0.402	5.692	8.122	0.382	8.142
0.9	20.256	4.172	0.092	4.14	5.88	0.124	5.848

Table 11. Changes of category after the estimation of efficiency. N-E error specification.

γ	Total Changes	Changes from 1 to 2	Changes from 1 to 3	Changes from 2 to 1	Changes from 2 to 3	Changes from 3 to 1	Changes from 3 to 2
0.1	58.572	9.844	7.734	9.344	11.958	8.234	11.458
0.2	55.378	9.398	6.288	8.848	12.278	6.838	11.728
0.3	51.706	8.948	4.876	8.234	12.386	5.59	11.672
0.4	48.88	8.496	4.032	7.816	12.252	4.712	11.572
0.5	45.738	8.214	3.076	7.516	11.928	3.774	11.23
0.6	42.632	7.666	2.388	7.146	11.522	2.908	11.002
0.7	38.348	6.826	1.544	6.442	10.996	1.928	10.612
0.8	33.696	5.996	0.924	5.836	10.008	1.084	9.848
0.9	25.474	4.464	0.242	4.386	8.07	0.32	7.992

Table 12. Total number of changes between categories 1-3, 3-1.

γ	N = 25	N = 50	N = 100	N = 200
0.1	15.22	13.936	14.4415	14.111
0.2	11.686	10.748	11.214	10.927
0.3	9.182	8.32	8.7545	8.39075
0.4	7.228	6.145	6.507	6.2395
0.5	5.114	4.378	4.6	4.41525
0.6	3.706	2.81	3.027	2.8325
0.7	1.928	1.57	1.6335	1.4355
0.8	0.784	0.546	0.549	0.47975
0.9	0.216	0.07	0.052	0.03875

Table 13. Total number of changes between categories 1-2, 2-1, 2-3, 3-2.
(N-NH specification)

γ	N = 25	N = 50	N = 100	N = 200
0.1	42.344	43.734	43.426	43.574
0.2	41.588	42.716	42.389	42.4065
0.3	40.38	41.34	40.902	41.1335
0.4	38.7	40.09	39.548	39.613
0.5	37.46	37.91	37.578	37.6075
0.6	34.54	35.628	35.189	35.233
0.7	31.928	31.936	31.998	31.986
0.8	27.628	27.214	27.084	27.142
0.9	20.04	19.918	19.957	19.496

Table 14. Correlation u, v, epsilon. Sample Size = 100.

γ	$u-\varepsilon$.(N-HN)	$u-\varepsilon$.(N-E)	$u-v$.(N-HN)	$u-v$.(N-E)	$u-u_T$.(N-NH)	$u-u_T$.(N-E)	$v-\varepsilon$.(N-HN)	$v-\varepsilon$.(N-E)
0.1	−0.3124	−0.3075	0.0000	0.0050	0.3124	0.3075	0.9489	0.9484
0.2	−0.4452	−0.4401	−0.0016	0.0012	0.4452	0.4401	0.8942	0.8945
0.3	−0.5423	−0.5449	0.0040	−0.0068	0.5423	0.5449	0.8352	0.8380
0.4	−0.6280	−0.6230	0.0037	0.0009	0.6280	0.6230	0.7723	0.7764
0.5	−0.7038	−0.7006	−0.0011	−0.0026	0.7038	0.7006	0.7071	0.7094
0.6	−0.7704	−0.7656	0.0028	−0.0011	0.7704	0.7656	0.6313	0.6379
0.7	−0.8339	−0.8289	0.0008	0.0021	0.8339	0.8289	0.5468	0.5514
0.8	−0.8927	−0.8893	−0.0029	0.0008	0.8927	0.8893	0.4493	0.4505
0.9	−0.9472	−0.9467	0.0031	−0.0003	0.9472	0.9467	0.3145	0.3176

Table 15. Correlation between true components and estimates based on JLMS. Sample Size = 100.

γ	$u-\hat{u}_{(E)}$ (N-HN)	$u-\hat{u}_{(E)}$ (N-E)	$\hat{u}_{(E)})-\hat{v}_{(E)}$ (N-HN)	$\hat{u}_{(E)})-\hat{v}_{(E)}$ (N-E)	$u-\varepsilon$. (N-HN)	$u-\varepsilon$ (N-E)
0.1	0.3182	0.3287	−0.9773	−0.9162	−0.9816	−0.9330
0.2	0.4581	0.4835	−0.9556	−0.8535	−0.9717	−0.9087
0.3	0.5606	0.6019	−0.9342	−0.7998	−0.9678	−0.9048
0.4	0.6484	0.6826	−0.9117	−0.7602	−0.9680	−0.9138
0.5	0.7244	0.7562	−0.8874	−0.7217	−0.9712	−0.9273
0.6	0.7892	0.8118	−0.8603	−0.6846	−0.9762	−0.9426
0.7	0.8487	0.8639	−0.8262	−0.6428	−0.9824	−0.9594
0.8	0.9026	0.9114	−0.7796	−0.5932	−0.9890	−0.9755
0.9	0.9517	0.9562	−0.7028	−0.5187	−0.9954	−0.9901

Table 16. Correlation between true efficiency and estimates based on JLMS. Sample Size = 100.

γ	$e^{-u}-e^{-\hat{u}_{(E)}}$ (N-HN)	$e^{-u}-e^{-\hat{u}_{(E)}}$ (N-E)	$e^{-u}-e^{-\hat{u}_{(T)}}$ (N-HN)	$e^{-u}-e^{-\hat{u}_{(T)}}$ (N-E)
0.1	0.3100	0.3116	0.2282	0.2078
0.2	0.4347	0.4391	0.3214	0.2869
0.3	0.5219	0.5290	0.3906	0.3528
0.4	0.5967	0.5918	0.4604	0.4077
0.5	0.6628	0.6554	0.5272	0.4740
0.6	0.7212	0.7085	0.5981	0.5391
0.7	0.7798	0.7639	0.6768	0.6195
0.8	0.8418	0.8253	0.7653	0.7148
0.9	0.9076	0.8963	0.8689	0.8350

QÜESTIIÓ, vol. 26, 3, p. 461-482, 2002

APLICACIÓN DE REDES NEURONALES ARTIFICIALES A LA PREVISIÓN DE SERIES TEMPORALES NO ESTACIONARIAS O NO INVERTIBLES

R. PINO*
D. DE LA FUENTE
J. PARREÑO
P. PRIORE
Universidad de Oviedo

En los últimos tiempos se ha comprobado un aumento del interés en la aplicación de las Redes Neuronales Artificiales a la previsión de series temporales, intentando explotar las indudables ventajas de estas herramientas. En este artículo se calculan previsiones de series no estacionarias o no invertibles, que presentan dificultades cuando se intentan pronosticar utilizando la metodología ARIMA de Box-Jenkins. Las ventajas de la aplicación de redes neuronales se aprecia con más claridad, cuando se trata de pronosticar sistemas multivariantes no estacionarios.

The application of artificial neural networks to forecasting non-stationary or non-invertible time series

Palabras clave: Series Temporales, Previsión, Redes Neuronales Artificiales

Clasificación AMS (MSC 2000): 37M10 Time Series Analysis
62M10 Time Series
62M45 Neural Networks
82C32 Neural Networks

*Dpto. Admon. de Empresas y Contabilidad. Escuela Politécnica de Ingenieros de Gijón. Campus de Viesques, s/n. 33204 GIJÓN (Spain). E-Mail: pino@etsiig.uniovi.es.

–Recibido en mayo de 2001.
–Aceptado en octubre de 2002.

1. INTRODUCCIÓN

Es un hecho conocido que para que el modelo de una serie temporal ARIMA pueda ser utilizado correctamente, se deben cumplir las condiciones de estacionariedad e invertibilidad (Box and Jenkins, 1970). En los modelos autorregresivos (AR), estas condiciones se cumplen si todas las raíces del polinomio característico están situadas fuera del círculo unidad; la misma condición asegura la invertibilidad de los modelos media móvil (MA). Para los modelos autorregresivos y media móvil (ARMA), existen varias condiciones que garantizan la estacionariedad e invertibilidad del modelo para determinadas localizaciones de las raíces de los polinomios característicos de los factores AR y MA (Huang, 1990; Huang and Anh, 1993).

Si estas condiciones no se cumplen, la metodología de Box-Jenkins llevaría al cálculo de previsiones fuera de unos límites razonables, por lo que, en el caso de series temporales no estacionarias o no invertibles, es necesario realizar transformaciones previas que garanticen estas dos condiciones antes de pasar a la estimación del modelo y al posterior cálculo de las previsiones.

Una de las características fundamentales de las Redes Neuronales Artificiales (RNAs) es su capacidad de «aprender» a partir de los ejemplos que se le proporcionan, sin hacer suposiciones a priori sobre los modelos y relaciones que subyacen en la serie temporal (Zhang *et al.*, 1998). Esta propiedad hace que sea posible calcular previsiones de cualquier serie temporal sin tener la necesidad de asegurar previamente ninguna de las condiciones comentadas anteriormente (estacionariedad e invertibilidad). Por tanto, el tratamiento de los datos, no dependerá de si es estacionaria o no, sino que será el mismo que se lleva a cabo con el resto de las series temporales (normalización de los datos, eliminación de la tendencia, etc.)

Hemos querido demostrar estos hechos, utilizando RNAs para calcular previsiones de varias series temporales no estacionarias o no invertibles. En el apartado 2 se hace una breve introducción a las redes neuronales artificiales, destacando las dos arquitecturas que se utilizan en este estudio; posteriormente se pronostican dos modelos autorregresivos no estacionarios de órdenes 1 y 2 (en los apartados 3 y 4 respectivamente); en el apartado 5 se calculan previsiones de un modelo media móvil no invertible de orden 2. Hasta aquí, todos los sistemas pronosticados son univariantes, pero donde verdaderamente se comprueba la utilidad de las RNAs, es al calcular previsiones de sistemas multivariantes, como se puede comprobar en el apartado 6, en el que se han aplicado las RNAs en el cálculo de previsiones de un sistema multivariante no estacionario. Terminamos el estudio resumiendo las conclusiones alcanzadas.

2. REDES NEURONALES ARTIFICIALES

En este apartado describiremos brevemente la estructura de una red neuronal (especialmente el Perceptrón Multicapa (MLP) y las redes de Elman, que son las que utilizaremos en este estudio), pero antes intentaremos establecer lo que se entiende por Computación Neuronal. Con este término definimos el estudio de estructuras de cálculo paralelo basadas en procesadores elementales adaptables (neuronas) ampliamente interconectados que, a partir del aprendizaje mediante ejemplos, almacenan conocimiento experimental y lo hacen disponible para su uso.

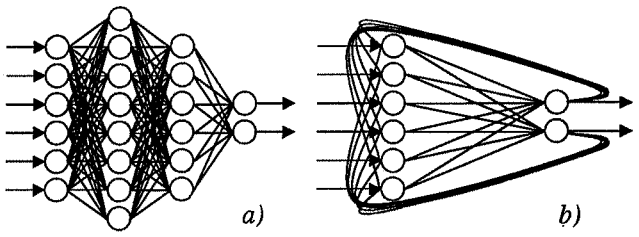


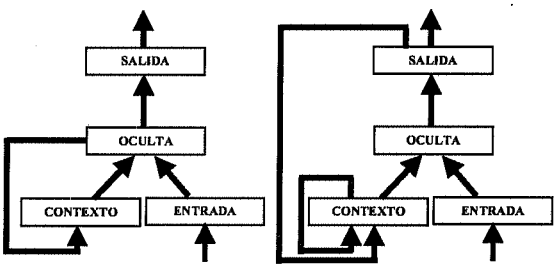
Figura 2.1. Ejemplos de Redes Neuronales «Feedforward» *(a)* y «Feedback» *(b)*.

El Perceptrón Multicapa (Multilayer Perceptron, MLP), también conocido como Red Backpropagation (Backpropagation Net, BPN) es uno de los modelos de Red Neuronal Artificial más conocidos y utilizado en la práctica como clasificadores de patrones y aproximadores de funciones (Lippman, 1987; Freeman and Skapura, 1991). Pertenece a la clase de las denominadas redes no realimentadas o «feedforward», su topología es la de un sistema neuronal estructurado en capas (Fig. 2.1.*a*), donde la información siempre fluye desde la capa de entrada, cuyo único papel es el de enviar los datos de entrada al resto de la red, hacia la de salida, atravesando la capa o capas ocultas. Esencialmente son las capas internas las encargadas de realizar el procesamiento de la información, extrayendo rasgos característicos de los datos de entrada. Aunque existen multitud de variantes, normalmente cada neurona de una capa se conecta a todas las neuronas de la capa siguiente; sin embargo, no existe conexión, ni por tanto interacción, entre las neuronas de una misma capa.

El algoritmo de entrenamiento más habitualmente utilizado en el MLP es el denominado «Retro-propagación» (Backpropagation, BP; Rumelhart and McClelland, 1986; Rumelhart *et al.*, 1986). Se trata de un aprendizaje de tipo *supervisado* en el que se muestra a la red neuronal tanto los patrones de entrada como su salida deseada, la red aprenderá a asociarlos por medio de la regla de aprendizaje, modificando los pesos sinápticos. Mediante este tipo de aprendizaje una red neuronal puede aproximar una compleja función a partir de muestras de ella o clasificar patrones a partir de ejemplos correctamente clasificados.

El BP, pese a ser de lenta convergencia y no garantizar el alcance de un mínimo global de la función de error, es un algoritmo muy utilizado por su relativa simplicidad y eficiencia. Se han propuesto en la literatura muchas variaciones para resolver estos problemas (Freeman and Skapura, 1991). Uno de los más utilizados para acelerar la convergencia y eliminar posibles oscilaciones en torno a mínimos locales, es la adición, a la hora de variar los pesos, de un determinado término denominado *momento*, que tiene en cuenta la anterior variación de los pesos.

El segundo tipo de redes que se utilizan en este estudio son las redes parcialmente recurrentes (Recurrent Neural Networks, RNN). Estas redes tienen una arquitectura basada en el MLP, pero con alguna modificaciones que las hacen especialmente interesantes para aplicarlas a predicción. En ellas existe una capa con neuronas especiales (denominada capa contextual) en la que se guarda la activación del instante anterior de las neuronas de otra capa (la de salida en las Redes de Jordan, o la oculta en las de Elman). En la figura 2.2. se pueden observar las dos estructuras más típicas de redes parcialmente recurrentes (Elman, 1990; Jordan, 1986).



Red Parcialmente Recurrente de Elman Red Parcialmente Recurrente de Jordan

Figura 2.2. Arquitecturas típicas de Redes Parcialmente Recurrentes.

Las recurrencias aparecen, si consideramos que la copia de la activación de la capa de neuronas normales a la capa contextual se realiza a través de unas conexiones con pesos fijos con valor 1. La particularidad de las redes parcialmente recurrentes de poder captar la evolución temporal de una variable en instantes consecutivos las hacen especialmente interesantes y apropiadas para predicción (Debar *et al.*, 1992; Dorffner *et al.*, 1994; Gordon *et al.*, 1991; Lee and Park, 1992; Kamijo and Tanigawa, 1993).

3. MODELO AR(1) NO ESTACIONARIO

Si tenemos un modelo autorregresivo de orden uno:

$$(1 - \phi_1 B) \cdot x_t = \varepsilon_t$$

Para que el proceso sea estacionario, se debe cumplir que el parámetro ϕ_1 satisfaga la condición $|\phi_1| < 1$ (Box and Jenkins, 1970), lo que equivale a decir que la raíz de la ecuación $1 - \phi_1 B = 0$ debe estar situada fuera del círculo unidad (figura 3.1).

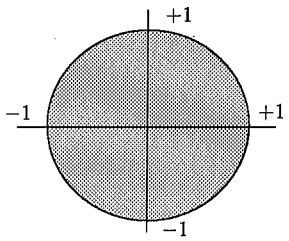


Figura 3.1. Círculo unidad.

Para comprobar la capacidad de las RNAs para pronosticar series temporales no estacionarias, hemos simulado la serie temporal llamada AR1 que tiene el siguiente modelo:

$$(1 - 1{,}01 \cdot B) \cdot x_t = \varepsilon_t$$

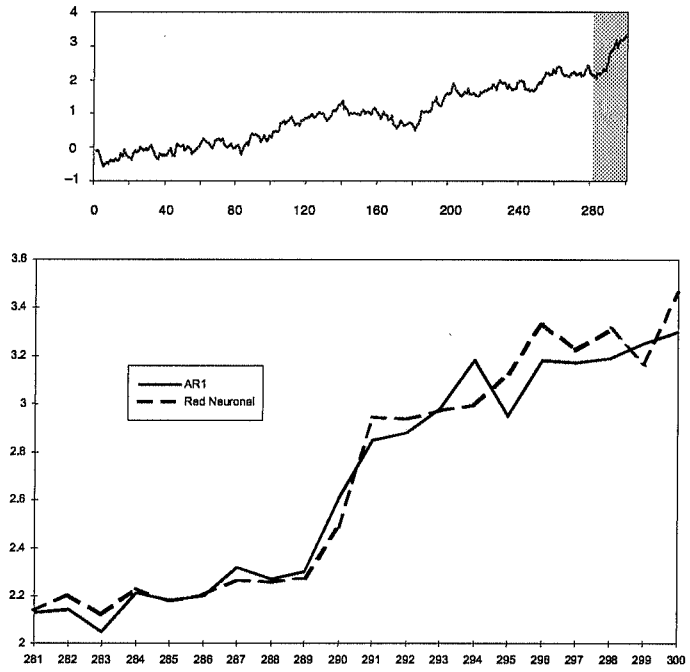


Figura 3.2. Serie AR1 (modelo AR(1) no estacionario).

Se puede comprobar que no se cumple la condición de estacionariedad, pues $\phi_1 = 1,01 > 1$. En la figura 3.2, se ha representado la serie AR1 simulada (la zona sombreada representa la zona que se pronostica), y a su derecha, se muestran las previsiones calculadas empleando un perceptrón multicapa 14-7-1 (14 neuronas en la capa de entrada, 7 en la oculta y 1 en la salida).

4. MODELO AR(2) NO ESTACIONARIO

Supongamos ahora el modelo autorregresivo de orden 2:

$$(1 - \phi_1 B - \phi_2 B^2) \cdot x_t = \varepsilon_t$$

Como ya se ha dicho en el caso anterior, para que el proceso sea estacionario, se debe cumplir que las raíces del polinomio característico estén fuera del círculo unidad, lo cual implica que los parámetros del modelo ϕ_1 y ϕ_2 cumplan las siguientes condiciones (Box and Jenkins, 1970):

$$\begin{cases} \phi_2 + \phi_1 < 1 \\ \phi_2 - \phi_1 < 1 \\ -1 < \phi_2 < +1 \end{cases}$$

Estas tres condiciones se pueden representar gráficamente en unos ejes ϕ_1 y ϕ_2, de forma que si el punto designado por (ϕ_1, ϕ_2) está situado dentro del triángulo, se puede demostrar que el proceso es estacionario, y si está situado sobre los bordes o fuera del triángulo, el proceso será no estacionario (figura 4.1).

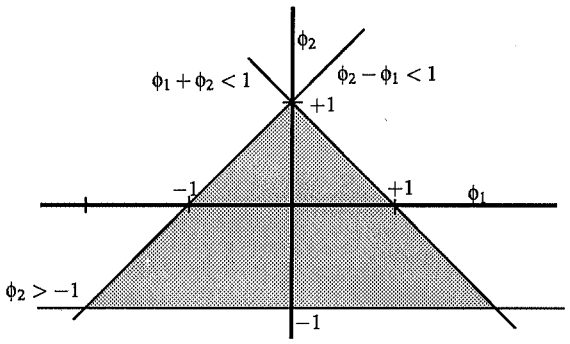


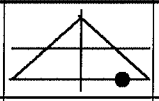
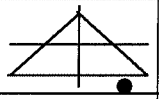
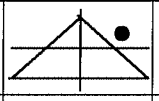
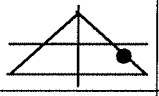
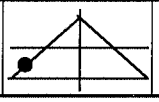
Figura 4.1. La zona sombreada implica estacionariedad.

Hemos simulado seis combinaciones de los parámetros ϕ_1 y ϕ_2 que hacen no estacionario un proceso AR(2), obteniendo las series temporales llamadas AR21, ..., AR26. En

la tabla 4.1, se muestran los valores de los parámetros de las seis series y su situación con respecto al triángulo. Por ejemplo, el modelo de Box-Jenkins de la serie AR21 será:

$$x_t - 1,2 \cdot x_{t-1} + x_{t-2} = \varepsilon_t$$

Tabla 4.1. Series No Estacionarias AR(2).

Serie	Situación	Parámetros	Serie	Situación	Parámetros
AR21		$\phi_1 = 1,2$ $\phi_2 = -1,0$	AR22		$\phi_1 = 1,2$ $\phi_2 = -1,1$
AR23		$\phi_1 = 1,2$ $\phi_2 = 0,5$	AR24		$\phi_1 = 1,2$ $\phi_2 = -0,2$
AR25		$\phi_1 = -1,7$ $\phi_2 = -0,7$	AR26		$\phi_1 = -1,7$ $\phi_2 = 0,5$

En las figuras 4.2 hasta 4.7, se presentan las series y las previsiones que se han calculado para cada una de ellas.

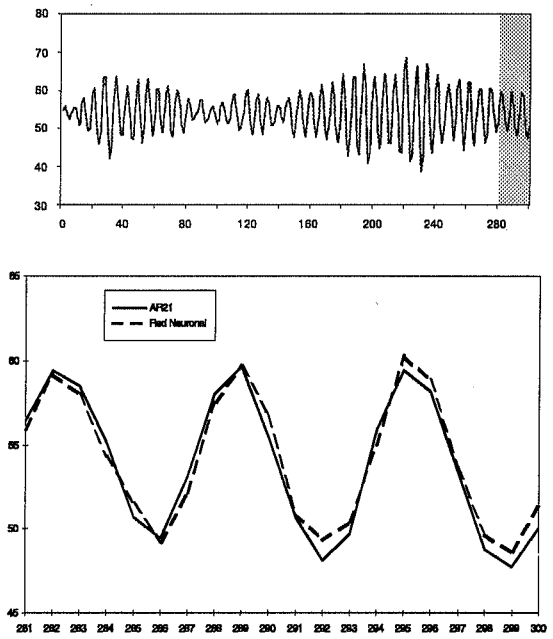


Figura 4.2. Serie AR21 (modelo AR(2) no estacionario).

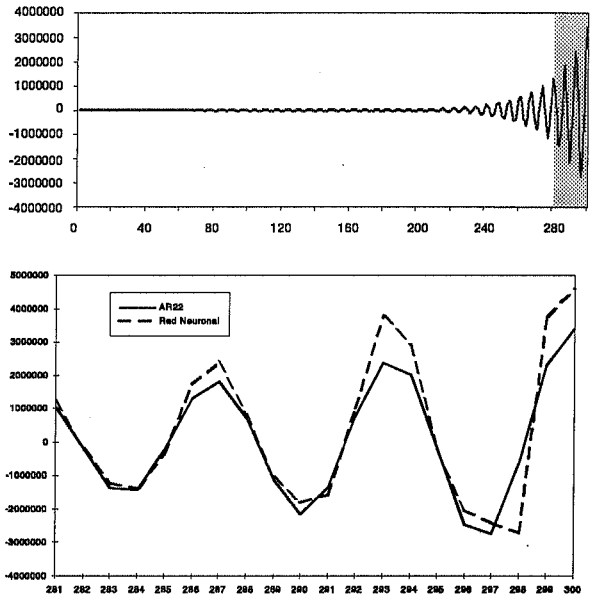


Figura 4.3. Serie AR22 (modelo AR(2) no estacionario).

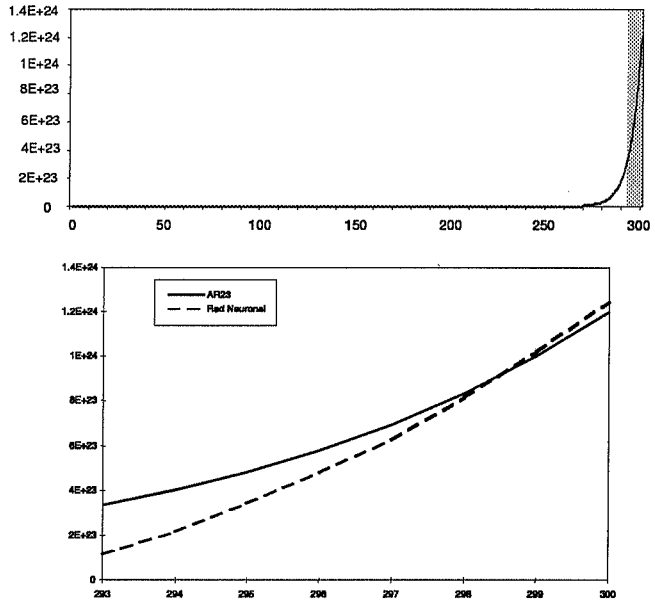


Figura 4.4. Serie AR23 (modelo AR(2) no estacionario).

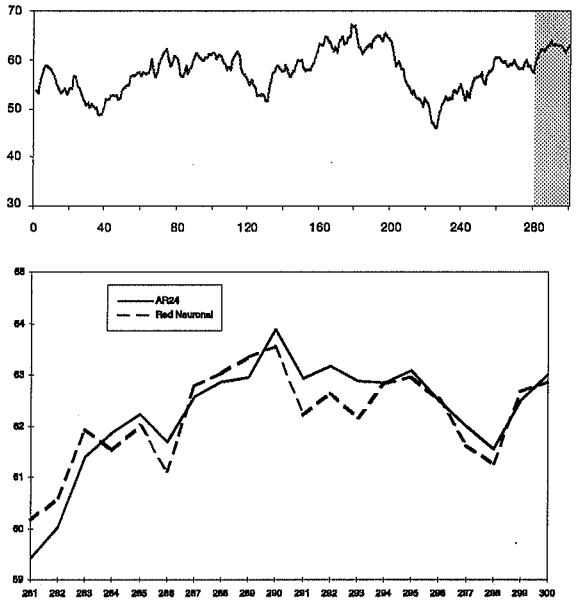


Figura 4.5. Serie AR24 (modelo AR(2) no estacionario).

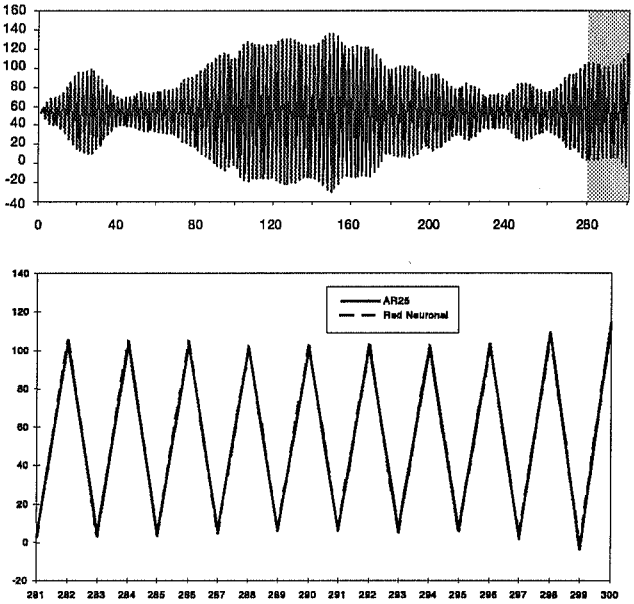


Figura 4.6. Serie AR25 (modelo AR(2) no estacionario).

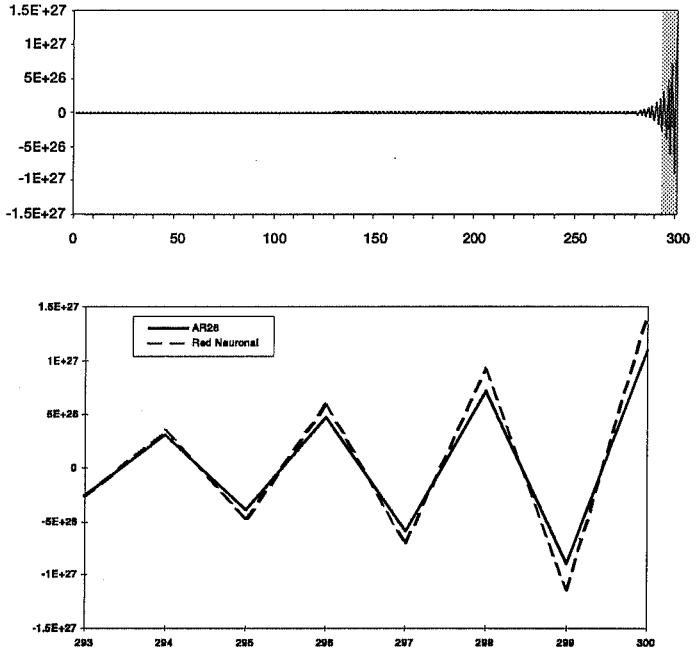


Figura 4.7. Serie AR26 (modelo AR(2) no estacionario).

5. MODELO MA(2) NO INVERTIBLE

Si tenemos ahora un proceso que presenta un modelo Media Móvil de orden 2:

$$x_t = (1 - \theta_1 B - \theta_2 B^2) \cdot \varepsilon_t$$

Las condiciones de estacionariedad de los modelos autorregresivos, aquí se convierten en condiciones de invertibilidad, por lo tanto los parámetros θ_1 y θ_2 deben cumplir (Box and Jenkins, 1970):

$$\begin{cases} \theta_2 + \theta_1 < 1 \\ \theta_2 - \theta_1 < 1 \\ -1 < \theta_2 < +1 \end{cases}$$

Al igual que ocurría en los procesos autorregresivos, se pueden representar gráficamente estas condiciones, de tal manera que si el punto designado por (θ_1, θ_2) está situado dentro del triángulo, el proceso es invertible mientras que si está fuera o sobre el borde, no será invertible (figura 5.1).

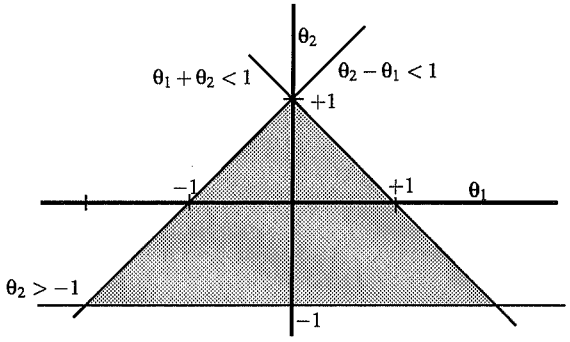
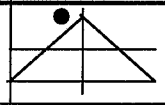
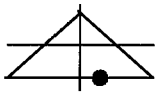
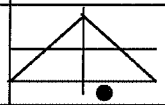
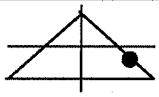
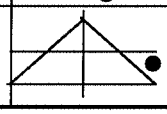
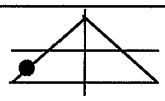


Figura 5.1. La zona sombreada implica invertibilidad.

Hemos simulado las seis combinaciones de los parámetros θ_1 y θ_2, que hacen no invertible el proceso, obteniendo las series MA20, MA21, ..., MA25. En la tabla 5.1, se muestran los valores de los parámetros, que hemos utilizado para la obtención de cada una de las series, y la situación del punto (θ_1, θ_2) con respecto al triángulo.

Tabla 5.1. Series No Invertibles MA(2).

Serie	Situación	Parámetros	Serie	Situación	Parámetros
MA20		$\theta_1 = -0,5$ $\theta_2 = 1,0$	MA21		$\theta_1 = 0,5$ $\theta_2 = -1,0$
MA22		$\theta_1 = 0,5$ $\theta_2 = -1,5$	MA23		$\theta_1 = 1,6$ $\theta_2 = -0,6$
MA24		$\theta_1 = 2,0$ $\theta_2 = -0,6$	MA25		$\theta_1 = -1,6$ $\theta_2 = -0,6$

En las figuras 5.2 hasta 5.7, se han representado las seis series temporales y las previsiones que se han calculado en cada uno de los casos.

En todas las series anteriores, se han calculado previsiones utilizando el programa de simulación de redes neuronales SNNS 4.2 (Stuttgart Neural Network Simulator). Las arquitecturas que se han probado son el Perceptrón Multicapa (MLP) y las Redes Parcialmente Recurrentes de Elman (ELM). En cada caso se ha seguido un proceso de prueba y error para determinar el número adecuado de neuronas tanto en la capa de entrada como en la capa oculta.

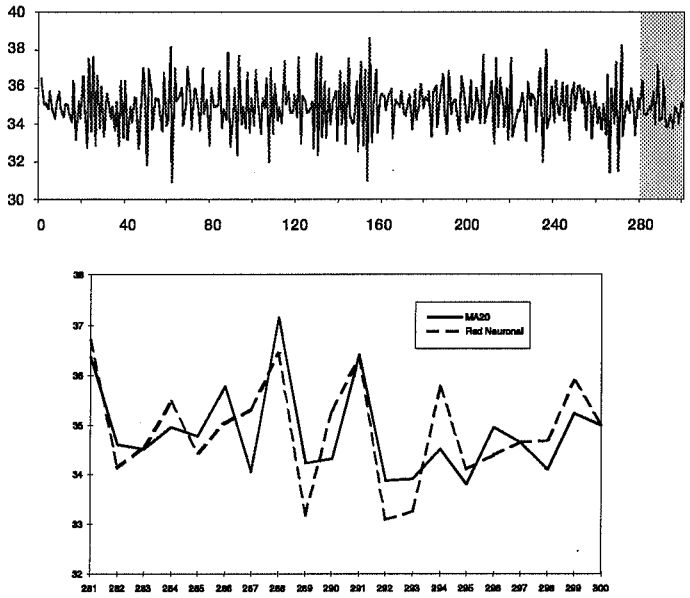


Figura 5.2. Serie MA20 (modelo MA(2) no invertible).

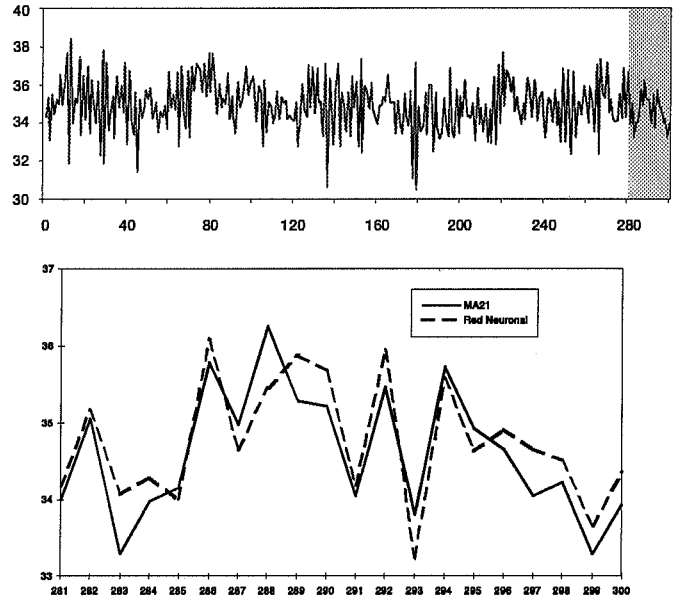


Figura 5.3. Serie MA21 (modelo MA(2) no invertible).

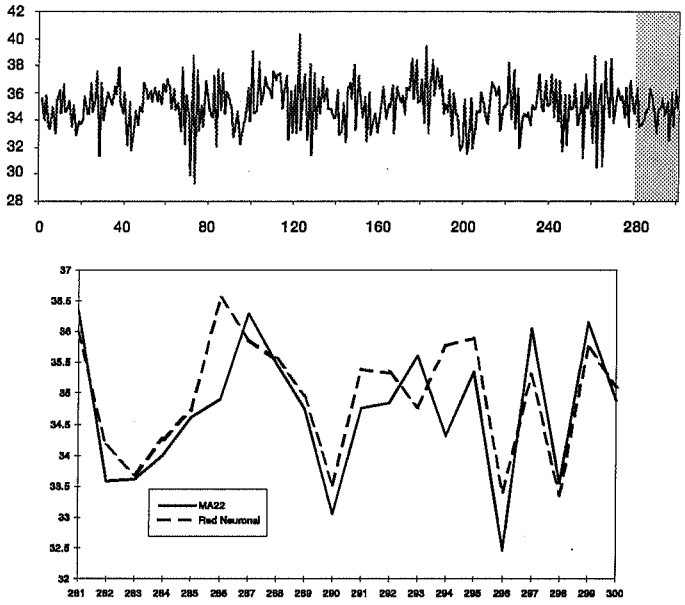


Figura 5.4. Serie MA22 (modelo MA(2) no invertible).

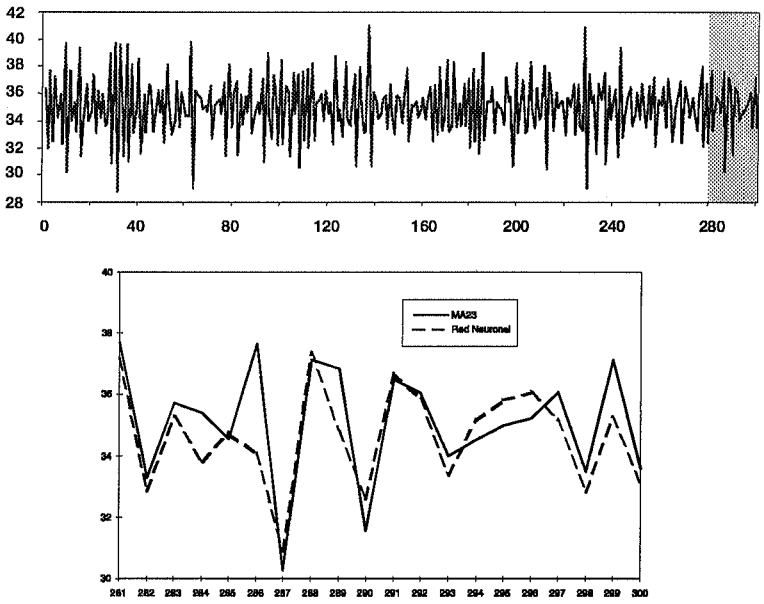


Figura 5.5. Serie MA23 (modelo MA(2) no invertible).

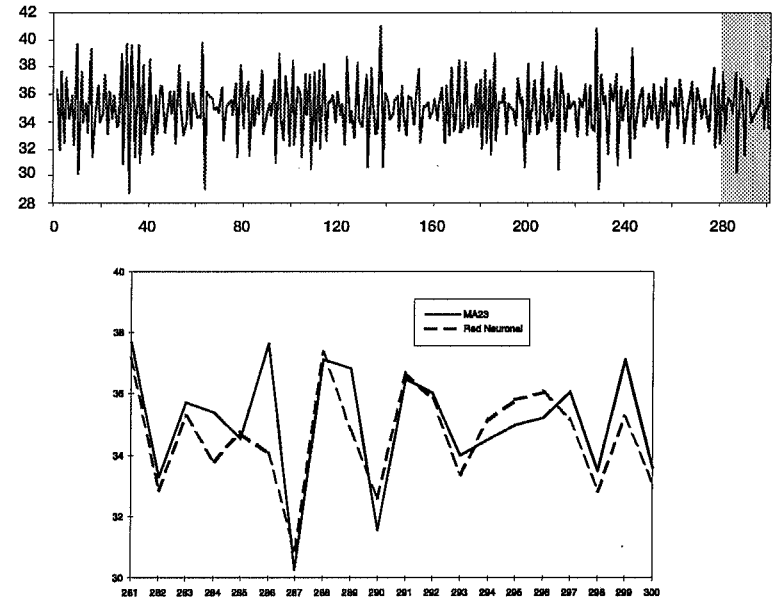


Figura 5.5. Serie MA23 (modelo MA(2) no invertible).

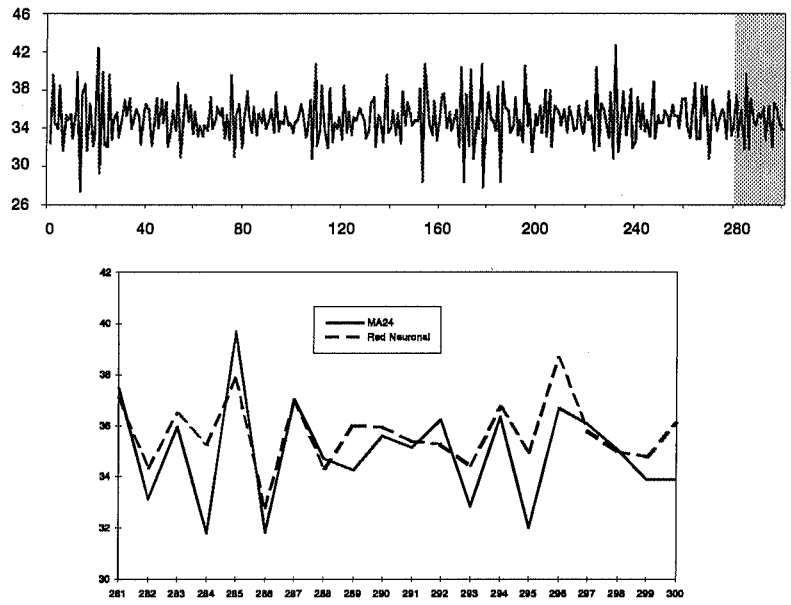


Figura 5.6. Serie MA24 (modelo MA(2) no invertible).

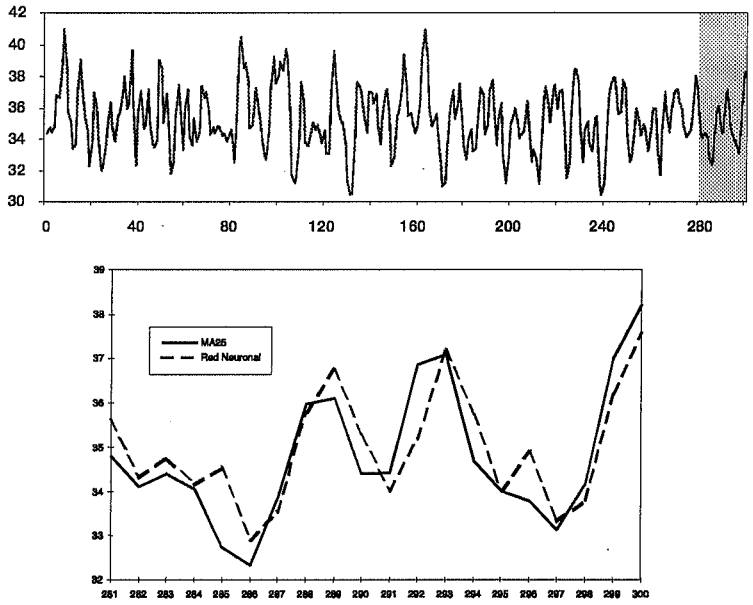


Figura 5.7. Serie MA25 (modelo MA(2) no invertible).

Tabla 5.2. Resultados del cálculo de previsiones.

Serie Temporal	Arquitectura y Configuración		MAD (1)	MSE (2)	MAPE (3)
AR1	MLP	14-7-1	0.074267	0.009043	2.59 %
AR21	MLP	8-3-1	0.686762	0.601951	1.30 %
AR22	MLP	6-4-1	519103.7	5.93E+11	28.84 %
AR23	MLP	10-6-1	9.98E+22	1.50E+46	28.62 %
AR24	MLP	13-8-1	0.359097	0.179917	0.58 %
AR25	MLP	8-4-1	0.982656	1.360244	13.94 %
AR26	MLP	6-4-1	1.38E+26	3.05E+52	17.58 %
MA20	ELM	16-9-1	0.565675	0.461447	1.63 %
MA21	MLP	11-7-1	0.378792	0.185006	1.09 %
MA22	ELM	8-5-1	0.531488	0.449266	1.52 %
MA23	ELM	10-6-1	0.894204	1.442150	2.56 %
MA24	ELM	10-6-1	1.120516	2.158522	3.21 %
MA25	MLP	12-7-1	0.617539	0.622427	1.77 %

(1) MAD (Mean Absolute Deviation).
(2) MSE (Mean Squared Error).
(3) MAPE (Mean Absolute Percentage Error).

En la tabla 5.2, se muestran las configuraciones de las redes seleccionadas para calcular las previsiones de todas estas series no estacionarias o no invertibles, y los estadísticos empleados para medir la calidad de las previsiones.

Es destacable el hecho de que en todos los modelos autorregresivos estudiados, la mejor arquitectura encontrada fue el MLP, mientras que en los modelos de media móvil, en la mayoría de los casos se obtuvieron mejores resultados con redes recurrentes de Elman. Esto confirma los estudios de Dorffner (1996), en los que se identifican los modelos *autorregresivos* de Box-Jenkins con el Perceptrón Multicapa y los modelos *media móvil* con la filosofía del funcionamiento de las redes parcialmente recurrentes (de Jordan o de Elman).

6. SISTEMA MULTIVARIANTE NO ESTACIONARIO

Un caso particularmente interesante es el de los sistemas multivariantes no estacionarios. A la dificultad propia de los sistemas multivariantes, se suma el hecho de que alguna o todas las series que componen el sistema, sean no estacionarias.

Presentamos como ejemplo el sistema denominado ARUMA33 (Huang and Anh, 1993), que se puede describir mediante la ecuación:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} B \\ B \end{bmatrix}^2 \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \end{bmatrix}$$

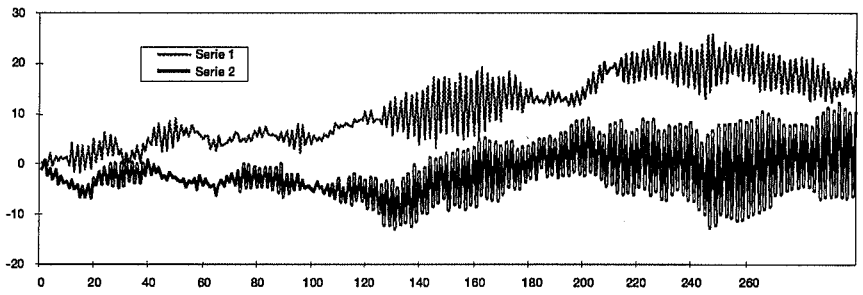


Figura 6.1. Sistema Multivariante ARUMA33.

En la figura 6.1, se han representado las dos series que componen el sistema multivariante y en la figura 6.2, las previsiones calculadas para cada una de ellas. En este caso se han utilizado dos redes neuronales: un perceptrón multicapa 12-8-1 para pronosticar la Serie 1, y otro perceptrón 12-9-1 para calcular las previsiones de la Serie 2. Igual que

en los casos anteriores, se han llevado a cabo distintas pruebas para encontrar el número adecuado de neuronas en la capa oculta de cada una de las redes.

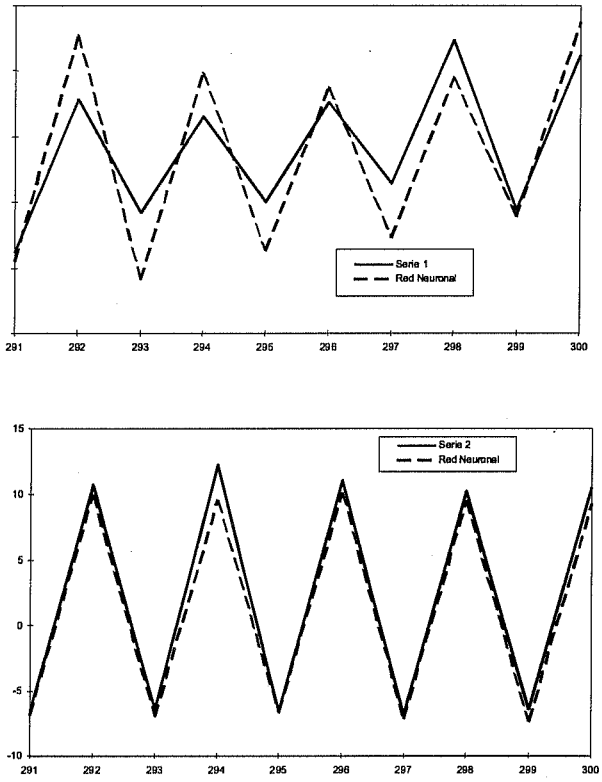


Figura 6.2. Previsiones para el Sistema Multivariante ARUMA33.

7. CONCLUSIONES

En este trabajo se ha demostrado el potencial de las Redes Neuronales Artificiales como herramienta para el cálculo de previsiones de series temporales.

En aquellos casos en los que aplicar la metodología clásica de Box-Jenkins es difícil o lleva a resultados inaceptables debido a las características de la serie temporal que se desea pronosticar, por ejemplo cuando se trata de series no estacionarias y/o no invertibles, es necesario realizar una serie de transformaciones de la serie temporal para que se asegure la estacionariedad y/o la invertibilidad.

Las Redes Neuronales gracias a su capacidad de «aprender» a partir de los ejemplos que se le proporcionan, sin hacer suposiciones a priori sobre los modelos y relaciones que subyacen en la serie, pueden ser aplicadas directamente, lo que facilita considerablemente el proceso de cálculo de previsiones.

Hemos querido demostrar este hecho, calculando previsiones de varias series temporales: 7 modelos autorregresivos no estacionarios, 6 modelos media móvil no invertibles, y un sistema multivariante no estacionario de 2 entradas y 2 salidas.

En el caso de las series univariantes, la arquitectura que se demostró más adecuada para los modelos autorregresivos fue el MLP, mientras que para los modelos de media móvil, los mejores resultados se obtuvieron con redes recurrentes de Elman. Los problemas de la aplicación de la metodología de Box-Jenkins, son aún mayores cuando se aplican al cálculo de previsiones de sistemas multivariantes, mientras que en el caso de las redes neuronales, su tratamiento no difiere del necesario para los sistemas univariantes, como hemos podido demostrar calculando pronósticos de un sistema multivariante no estacionario.

REFERENCES

Huang, D. and Anh, V. V. (1993). «Estimation of the Non-Stationary Factor in ARUMA Models». *Journal of Time Series Analysis*, 14(1), 27-46.

Box, G. E. and Jenkins, G. M. (1970). *Time Series Analysis*, Holden-Day, San Francisco.

Debar, H. and Dorizzi, B. (1992). «An Application of a Recurrent Network to an Intrusion Detection System». *Proceedings of the IEEE International Joint Conference on Neural Networks*, Baltimore.

Dorffner, G. (1996). «Neural Networks for Time Series Processing», *Neural Network World*, 6(4), 447-468.

Dorffner, G.; Leitgeb, E. and Koller, H. (1994). «Towards Improving Exercise ECG for Detecting Ischemic Heart Disease with Recurrent and Feedforward Neural Nets». *Neural Networks for Signal Processing IV*, Vlontzos, J. *et al.* (eds.), IEEE, New York, 499-508.

Elman, J. L. (1990). «Finding structure in time». *Cognitive Science*, 14, 179-211.

Freeman, J. A. and Skapura, D. M. (1991). *Neural Networks: Algorithms, Applications, and Programming Techniques*. Addison-Wesley Publishing Company, Inc., Reading, MA.

Gordon, A.; Steele, J. P. H. and Rossmiller, K. (1991). «Predicting Trajectories Using Recurrent Neural Networks». *Intelligent Engineering Systems through Artificial Neural Networks*, Dagli, C.H. *et al.* (eds.), ASME Press, New York, 365-370.

Huang, D. (1990). «Levinson-type recursion for least squares estimation». *Journal of Time Series Analysis*, 11, 295-315.

Jordan, M. I. (1986). «Attractor dynamics and parallelism in a connectionist sequential machine». *Proceedings of the Eight Annual Conference of the Cognitive Science Society*, 531-546, Erlbaum.

Kamijo, K. and Tanigawa, T. (1993). «Stock Price Pattern Recognition: A Recurrent Neural Network Approach». *Neural Networks in Finance and Investing*, Trippi, R. R. *et al.* (eds.), 357-370.

Lee, C. H. and Park, K. C. (1992). «Prediction of Monthly Transition of the Composition Stock Price Index using Recurrent Back-Propagation». *Artificial Neural Networks 2*, Aleksander, I. *et al.* (eds.), 1629-1632.

Lippmann, R. P. (1987). «An Introduction to Computing with Neural Networks». *IEEE ASSP Magazine*, 3(4), 4-22.

Rumelhart, D. E. and McClelland, J. (1986). *Parallel Distributed Processing*. MIT Press, Cambridge.

Rumelhart, D. E.; Hinton, G. E. and Wiliams, R. J. (1986). «Learning representations by back-propagating errors». *Nature*, 323, 533-536.

SNNS (1998). *Stuttgart Neural Network Simulator. User Manual, Version 4.2*. University of Stuttgart.

Zhang, G.; Patuwo, B. E. and Hu, M. Y. (1998). «Forecasting with artificial neural networks: The state of the art». *International Journal of Forecasting*, 14, 35-62.

ENGLISH SUMMARY

THE APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO FORECASTING NON-STATIONARY OR NON-INVERTIBLE TIME SERIES

R. PINO*
D. DE LA FUENTE
J. PARREÑO
P. PRIORE
Universidad de Oviedo

There has recently been noticeably increased interest in the application of Artificial Neural Networks to forecasting times series, and attempts have been made to exploit the undoubted advantages of these tools. This paper deals with non-stationary or noninvertible times series that are problematical if forecasting uses the Box-Jenkins' ARIMA methodology. The advantages of applying neural networks can be appreciated more clearly in the case of multivariate non-stationary system forecasting.

Keywords: Time Series, Forecasting, Artificial Neurl Networks

AMS Classification (MSC 2000): 37M10 Time Series Analysis
62M10 Time Series
62M45 Neural Networks
82C32 Neural Networks

*Dpto. Admon. de Empresas y Contabilidad. Escuela Politécnica de Ingenieros de Gijón. Campus de Viesques, s/n. 33204 GIJÓN (Spain). E-Mail: pino@etsiig.uniovi.es.

-Received May 2001.
-Accepted October 2002.

1. INTRODUCTION

It is well known that a time series has to meet the stationarity and invertibility conditions in order to be correctly modelled by an ARIMA model. AutoRegresssive (AR) models meet these conditions if all the roots of the characteristic polynomial are outside the unit circle; the same condition also ensures the invertibility of Moving Average (MA) models. For autoregressive and moving average (ARMA) models, several conditions exist that guarantee stationarity and invertibility for certain locations of the roots of the characteristic polynomial of AR and MA factors.

If these conditions are not met, Box-Jenkins methodology would lead to forecasts beyond reasonable limits, so that it would be necessary to undertake some prior transformations to the time series before estimating the model and then making the forecast.

One of the characteristics of Artificial Neural Nets (ANNs) resides in their ability to «learn» from the examples they are provided with, without making a priori suppositions about the models and underlying correlations in a given time series. This fact makes possible to calculate forecasts for any time series without having to previously test the stationarity and invertibility.

The aim of this job is to demonstrate the affirmations made in the previous paragraph by using ANNs to calculate forecasts for several non-stationary or non-invertible time series.

2. ARTIFICIAL NEURAL NETS

This section will briefly describe the structure of Multilayer Perceptron (MLP) and Recurrent Neural Nets (RNNs).

MLP, also known as Backpropagation Net (BPN) is one of the most used ANN models as pattern classifier and function approximator. It belongs to the category of feedforward nets, and is structured in layers where the information flows from the input layer, through the hidden layers, to the output layer. The information processing is responsibility of the hidden layers.

The most commonly training algorithm for MLP is the Backpropagation (BP), an algorithm that is widely used because of its simplicity and efficiency.

The second type of net used in this study are Recurrent Neural Nets (RNN). These nets have an architecture based on MLP, but with certain modifications, as the presence of a contextual layer. The ability of RNNs for detecting the time evolution of a variable during consecutive time periods makes them very interesting and adequate for forecasting.

3. RESULTS

Forecasts were calculated for several time series: seven non-stationary AR models, six non-invertible MA models and a 2-input 2-output non-stationary multivariate system. For all these series, Box-Jenkins methodology showed little forecasting accuracy.

For the univariate cases, MLP proved to be the more adequate architecture for AR models, whilst best results for MA models were obtained with RNNs. For the multivariate case, the best results were provided by MLP.

QÜESTIIÓ, vol. 26, 3, p. 483-501, 2002

ANÀLISI DE MATRIUS QUADRADES NO SIMÈTRIQUES: UN ENFOCAMENT INTEGRAL USANT ANÀLISI DE CORRESPONDÈNCIES

J. DAUNIS I ESTADELLA*
T. ALUJA BANET**
S. THIÓ FDEZ. DE HENESTROSA*

En aquest article s'introdueix una proposta integrada d'anàlisi de matrius quadrades no simètriques mitjançant anàlisi de correspondències. En ella s'han integrat les dues principals famílies de metodologies que proposen solucions a la problemàtica inherent d'aquest tipus de taules: les diagonals sobrecarregades i els zeros estructurals. S'aplica a l'exemple de l'estudi dels moviments de població deguts al treball entre les 41 comarques catalanes.

Analysis of square skew-symmetric tables: An integral vision by correspondence analysis

Paraules clau: Matrius quadrades, diagonal sobrecarregada, anàlisi de correspondències, zeros estructurals, descomposició en valors singulars, dades mancants

Classificació AMS (MSC 2000): 62H17, 62H25

* Departament d'Informàtica i Matemàtica Aplicada. Universitat de Girona. Av. Lluís Santaló s/n. 17071 GIRONA. Email: josep.daunis@udg.es santiago.thio@udg.es.
** Departament d'Estadística i Investigació Operativa. Universitat Politècnica de Catalunya. C/Pau Gargallo, núm. 5. 08028 BARCELONA. Email: tomas.aluja@upc.es.

– Rebut al desembre de 2001.
– Acceptat a l'octubre de 2002.

1. INTRODUCCIÓ

En diferents àmbits científics, ens trobem davant la presència de taules quadrades, on les fileres i les columnes es refereixen al mateix conjunt d'objectes o categories. Aquesta és una tipologia de taula d'ús molt freqüent i en les que ens centrarem en el cas de taules de tipus no simètric, de més ampli ús, on no hi ha d'haver una relació simètrica entre fileres i columnes. Anem a il·lustrar la problemàtica amb les dades objecte del nostre estudi: els moviments migratoris obligats pel treball, entre les comarques de Catalunya corresponents al cens de 1996 (Institut d'Estadística de Catalunya, 2001).

Trobem a la Taula 1 les dades, on les columnes són els orígens dels treballadors i les fileres les destinacions, dels moviments entre les 41 comarques catalanes deguts al treball. Clarament ens trobem amb una matriu quadrada no simètrica. Per a ajudar a interpretar més bé els moviments, ho referirem al mapa adjunt de les 41 comarques catalanes (Figura 1).

Figura 1. Les 41 comarques catalanes. (Origen Web de l'Idescat. Disseny propi)

Habitualment, aquestes taules són analitzades mitjançant anàlisi de correspondències (AC), però en l'anàlisi d'aquests tipus de taules, la principal problemàtica sorgeix en la inèrcia de la diagonal. En aquestes diagonals, els efectius habitualment poden ser o molt

grans amb relació amb la resta d'efectius (diagonals carregades), com podem trobar en l'exemple de mobilitat donat per Kazmierczak (1978), o nuls per l'estructura de la taula (zeros estructurals). Aquesta darrera problemàtica, els zeros estructurals, també es pot extrendre, però no tan freqüentment, fora de la diagonal.

L'explicació d'aquesta problemàtica es troba en l'anàlisi de la inèrcia i l'estudi de la descomposició de la inèrcia en aquests tipus de taules per a cadascuna de les caselles. En l'AC la *inèrcia total*, mesura de la variació de la taula **P** de terme general p_{ij} és:

$$I(T) = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

on p_{ij} és la freqüència relativa de cada casella i r_i i c_j les marginals de fileres i columnes respectivament, essent r i c els vectors marginals. La inèrcia deguda a cada casella (i, j) és $I_{ij} = (p_{ij} - r_i c_j)^2/(r_i c_j)$.

Es pot descomposar la inèrcia total de la taula en la inèrcia de la diagonal $I(D)$ i en la de fora de la diagonal $I(\overline{D})$, com a la suma de totes les inèrcies de les caselles pertanyents a aquesta o aquella. En la nostra taula, la importància de la inèrcia de la diagonal és del 97.8 % sobre una inèrcia total de 28.65. Per tant la representació gràfica factorial obtinguda és bàsicament el moviment dels treballadors que tenen l'origen i el destí dins la pròpia zona. Aquest percentatge tan elevat és un tret característic de les matrius quadrades no simètriques objecte d'estudi.

L'anàlisi d'aquestes taules és problemàtica, per la qual cosa existeixen principalment dues famílies de mètodes de tractament. El nostre objectiu serà analitzar les metodologies existents i proposar una metodologia d'anàlisi integradora que superi la influència d'aquestes caselles productores d'inèrcies estructurals per a una millor anàlisi de les relacions globals entre les modalitats.

2. ANTECEDENTS: METODOLOGIES EXISTENTS DE SOLUCIÓ

La gran influència de la diagonal, en aquestes taules objecte d'estudi, ha portat a la necessitat d'estudiar propostes de modificació de les anàlisis en el sentit de treure la influència d'aquestes caselles. Aquestes propostes de modificació es basen principalment en reconstituir els elements de les diagonals, amb metodologies diferents, atenent a diferents propostes de reconstitució, però també s'ha abordat el problema des d'una metodologia d'intentar descomposar l'anàlisi en referència a la simetria o no simetria de les dades. A les següents seccions tractarem les propostes que diferents autors han realitzat, les compararem breument i finalment proposarem la nostra metodologia de resolució.

2.1. Reemplaçament dels elements de la diagonal sota hipòtesi d'independència

Aquesta proposta, així com les següents, està sustentada en base als *mètodes d'anàlisi factorial en referència a un model* proposats per Escofier (1984), ja que la taula objecte d'estudi és la taula original menys el model d'independència imposat als elements diagonals.

Per anul·lar la component inercial dels elements de la diagonal De Leeuw i van der Heijden (1988) suggereixen redefinir els elements de la diagonal com a producte dels marges, és a dir, sota un model d'independència: $\hat{p}_{ii} = r_i c_i$. Aquesta reconstitució de la matriu de dades sota el supòsit d'independència comporta que les caselles reconstituïdes tinguin, en principi, una inèrcia nul·la: $I(D) = 0$, però no es conserven els marges de les taules i, a més, la reconstitució està influïda pels propis valors de les caselles influents.

L'anàlisi de correspondències de la taula reconstituïda sota la hipòtesi d'independència ens dóna que encara es conserva la gran relació d'un mateix lloc com a origen i com a destinació. Aquesta relació és deguda al fort paper que tenen les caselles de la diagonal en les marginals corresponents.

2.2. Tractament com a *dada mancant* de les caselles diagonals

La tècnica usada en aquest apartat s'inclou en les tècniques per al tractament de taules incompletes proposades per Nora (1974), també tractades per Benzécri (1973), així com per Greenacre (1984) i de Leeuw i van der Heijden (1988) entre d'altres, essent ara aplicada al tractament de dades influents.

La influència de les caselles de la diagonal en les marginals porta a una proposta de solució basada en el tractament de les dades de la diagonal com a *dades mancants* (*missing values*), és a dir, considerarem a tots els efectes que aquestes dades són mancants i per tant la seva reconstitució a partir de les marginals no serà influïda per aquestes. Es redefineix cada element de la diagonal com:

$$\hat{p}_{ii} = \frac{r_i^0 c_i^0}{1^0 - (r_i^0 + c_i^0)}$$

on el superíndex zero significa que l'element de la diagonal ha estat exclòs dels respectius marginals.

Es pot veure que, aplicant aquesta metodologia, es perd l'estructura deguda a la presència de grans quantitats en les caselles de la diagonal i només resten les quantitats degudes a les caselles de fora de la diagonal.

Aquesta reconstitució de la matriu per la imputació d'efectius com a dades mancants comporta que les cel·les reconstituïdes tenen una inèrcia gairebé nul·la, no es conserven

les marginals de la taula però la reconstitució ara no està influïda pels valors diagonals influents. Aquesta mateixa tècnica pot ser usada per al tractament de zeros estructurals fora de la diagonal o de les subtaules diagonals en anàlisi de correspondències múltiples (Greenacre, 1988).

2.3. Reemplaçament dels elements de la diagonal mitjançant la fórmula de reconstitució

Basada en el tractament tipus mancant aplicat a la diagonal i proposat per Mutombo (1973) i Nora (1974), per a dades mancants, hi ha el tractament k-EM. Aquest es basa en un algorisme tipus EM, Dempster (1976), però amb la introducció d'un nou paràmetre, l'ordre k de reconstitució inherent a les anàlisis factorials.

Aquest algorisme està basat en fer una primera reconstitució tipus mancant, on aquest serà el valor inicial de l'algorisme. Llavors s'aplica l'algorisme EM, on les etapes d'estimació (E) i maximització (M) es basen en la imputació dels valors obtinguts en l'etapa anterior i la reconstitució d'ordre k d'aquests valors mitjançant la fórmula inherent a les anàlisis factorials. En aquesta metodologia es conserven les marginals havent-ne exclòs d'elles els elements de la diagonal.

Altres propostes en aquesta mateixa línia, d'imposar un model sobre les dades, han estat proposades, vegi's per exemple Foucart (1985), però també hi ha metodologies alternatives com la de la següent secció.

2.4. Descomposició de la matriu com l'addició d'una matriu simètrica i d'una d'antisimètrica

En un altre vessant totalment diferent, trobem la proposta de Constantine i Gower (1978), revisada per Greenacre (2000), de solució basada en la descomposició de la matriu com a suma d'una matriu simètrica i d'una altra d'antisimètrica. La matriu objecte d'estudi $\mathbf{P}\text{-}\mathbf{r}\mathbf{c}^{\mathrm{T}}$ pot ésser descomposada en $\mathbf{Q}+\mathbf{R}$ on $\mathbf{Q}$ és una matriu simètrica i $\mathbf{R}$ una matriu antisimètrica. Si anomenem $\mathbf{A}=\mathbf{P}\text{-}\mathbf{r}\mathbf{c}^{\mathrm{T}}$, llavors podem trobar $\mathbf{Q}$ **i** $\mathbf{R}$ de la següent manera:

$$\mathbf{Q}=\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{A}+\mathbf{A}^{\mathrm{T}}) \qquad \text{i} \qquad \mathbf{R}=\frac{\mathbf{1}}{\mathbf{2}}(\mathbf{A}-\mathbf{A}^{\mathrm{T}})$$

Utilitzant les propietats de les matrius $\mathbf{Q}$ -amb marginals $\mathbf{w}$- i $\mathbf{R}$, i com que:

$$\mathbf{P}-\mathbf{w}\mathbf{w}^{\mathrm{T}}=\mathbf{Q}-\mathbf{w}\mathbf{w}^{\mathrm{T}}+\mathbf{R}$$

Això ens permet fer una descomposició de la inèrcia en la part deguda a la simetria i a la no-simetria. L'anàlisi de $\mathbf{Q}\text{-}\mathbf{w}\mathbf{w}^{\mathrm{T}}$ reflexa la inèrcia de la simetria (valors diagonals i mo-

viments recíprocs) i l'anàlisi de **R** reflexa la no-simetria (els balanços dels moviments recíprocs).

Llavors en comptes de l'anàlisi estàndard, això és la descomposició en valors singulars generalitzada (DVSG) de la tripleta $(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathrm{T}}, \mathbf{D}_r^{-1}, \mathbf{D}_c^{-1})$, les anàlisis proposades són les DVSG:

$$(\mathbf{Q} - \mathbf{w}\mathbf{w}^{\mathrm{T}}, \mathbf{D}_{\mathbf{w}}^{-1}, \mathbf{D}_{\mathbf{w}}^{-1})$$

$$(\mathbf{R}, \mathbf{D}_{\mathbf{w}}^{-1}, \mathbf{D}_{\mathbf{w}}^{-1})$$

És a dir se centra en referència a les marginals de **Q** i es prenen també les mètriques de la part simètrica.

3. PROPOSTA D'ANÀLISI DE MATRIUS QUADRADES NO SIMÈTRIQUES

Atenent a la problemàtica descrita i a les diferents propostes realitzades pels diferents investigadors, l'anàlisi global que suggerim agafarà part d'aquestes propostes (la descomposició en part simètrica i no-simètrica, la reconstitució EM), però també el càlcul de la importància de l'efectiu de la diagonal en relació al global i la problemàtica de la reconstitució influïda per la diagonal. L'anàlisi la dividirem en tres etapes que, per la seva relació de similitud amb els moviments migratoris, seran anomenades:

1. Anàlisi migració intra
2. Anàlisi migració entre i recíproca
3. Anàlisi saldos migratoris

Passarem ara a descriure cadascuna d'aquestes tres parts i posteriorment farem la seva aplicació sobre les nostres dades.

3.1. Anàlisi migració intra

La primera etapa és l'anomenada *anàlisi de la migració intra* o l'anàlisi dels moviments dins de la mateixa zona, és a dir l'estudi de la importància dels valors de la diagonal en referència als seus respectius marges, això ens donarà quins percentatges d'individus romanen dins la mateixa categoria, o fent el símil, quins individus romanen dins la pròpia zona en els seus desplaçaments laborals.

Per a aquesta primera etapa el que farem serà calcular els percentatges del moviment intern de la zona en referència al moviment global vers la mateixa zona, aquest ja serà un primer indicador de l'anàlisi, ja que tota la part no recollida en l'etapa Intra serà susceptible de ser recollida en les etapes posteriors. És una primera part d'anàlisi purament descriptiva.

3.2. Anàlisi migració entre i recíproca

La segona etapa és l'anomenada *anàlisi de la migració entre i recíproca* o l'anàlisi dels moviments simètrics entre cadascuna de les parelles de categories, és a dir, l'estudi de la component simètrica, definida ja en apartats anteriors, però a la qual li traurem la influència dels valors de la diagonal, ja que aquests influeixen en la part simètrica, i la seva importància ja ha quedat expressada en l'apartat anterior.

Per a treure la influència de la diagonal, en l'estudi de la part simètrica el que farem és aplicar l'algorisme de la reconstitució tipus k-EM a la matriu $\mathbf{Q}\text{-}\mathbf{w}\mathbf{w}^{\mathrm{T}}$, matriu de la component simètrica, reconstitució ja exposada en l'apartat 2.3.

Les etapes de l'algorisme són:

a) S'assignen els valors inicials de $\hat{q}_{ii}$.

b) Es realitza una AC de la matriu amb els valors reconstituïts tipus *mancant* (etapa M, o de maximització, de l'algorisme EM), per a trobar els eixos i les coordenades factorials.

c) S'utilitza la fórmula de reconstitució de les anàlisis factorials, expressada en termes dels valors propis λ i de les coordenades estàndards ψ i ϕ, d'ordre $k = 1$ del valor $\hat{q}_{ii}$ (etapa E, o d'estimació, de l'algorisme EM).

d) Es reiteren els passos *(b)* i *(c)* fins a la convergència dels valors reconstituïts.

Així:

$$\hat{q}_{ii}^{(1,m)} = r_i^{(1,m-1)} c_i^{(1,m-1)} \left\{ 1 + \sqrt{\lambda_\alpha^{(1,m-1)}} \psi_{\alpha i}^{(1,m-1)} \phi_{\alpha i}^{(1,m-1)} \right\}$$

on els superíndexs (1,*m*) representen l'ordre de reconstitució i l'ordre d'iteracció respectivament.

e) Una vegada obtinguda aquesta convergència, s'incrementa l'ordre de reconstitució, essent ara $k = 2$, prenent com a valor inicial l'obtingut en l'etapa (d) del procés anterior. Es reiteren les etapes de maximització i d'estimació fins a assolir de nou la convergència, amb ordre de reconstitució $k = 2$.

Es finalitza el procés augmentant l'ordre de reconstitució, utilitzant a l'inici de cada nova etapa com a valor inicial el valor obtingut al final del procés d'iteració d'ordre inferior. La fórmula general de l'etapa d'estimació és:

$$\hat{q}_{ii}^{(k,m)} = r_i^{(k,m-1)} c_i^{(k,m-1)} \left\{ 1 + \sum_{\alpha=1}^{k} \sqrt{\lambda_\alpha^{(k,m-1)}} \psi_{\alpha i}^{(k,m-1)} \phi_{\alpha i}^{(k,m-1)} \right\}$$

S'acaba el procés d'acord, per exemple, amb els valors propis significatius en l'anàlisi de la taula, indicador habitual de la qualitat de representació de cada eix.

3.3. Anàlisi saldos migratoris

Finalment, la darrera etapa és l'anomenada *anàlisi dels saldos migratoris* o l'anàlisi de la diferència de moviments no recíprocs entre categories, és a dir, l'estudi de la component no simètrica dels moviments entre objectes.

Per a realitzar aquesta darrera part realitzarem l'estudi ja presentat de la matriu **R**, matriu corresponent a la no-simetria, on realitzarem sobre aquestes dades la descomposició en valors singulars generalitzada de la tripleta:

$$(\mathbf{R}, \mathbf{D}_\mathbf{w}^{-1}, \mathbf{D}_\mathbf{w}^{-1})$$

És a dir ponderem en referència a les marginals i prenem també les mètriques de la part simètrica **Q**.

4. EXEMPLE D'APLICACIÓ

En aquesta secció anem a aplicar la proposta d'anàlisi efectuada a l'exemple dels moviments deguts al treball entre les comarques catalanes, el qual el tractarem en base als tres ítems presentats: (1) anàlisi migració intra, (2) anàlisi migració entre i recíproca i (3) anàlisi saldos migratoris.

4.1. Anàlisi migració intra

Començarem amb l'estudi de la migració intra, és a dir la importància dels moviments de treballadors de la pròpia zona sobre el total d'efectius que arriben a la zona. En la Figura 2 trobem una representació gràfica d'aquests moviments. Podem observar que, del total de moviments sobre cadascuna de les zones, trobem que com a mínim un 61.2 % d'aquests moviments pertanyen a gent de la pròpia zona (Baix Llobregat) mentre que arriba a uns màxims de més d'un 90 % a la Val d'Aran, Segrià, Osona, Alt Urgell, Alt Empordà i Garrotxa. És a dir, molt més de la meitat de les migracions sobre una determinada zona són internes. Es dóna la circumstància que les zones amb una migració Intra més altes es corresponen a zones més perifèriques i a zones nord, mentre que aquelles on la migració intra és més baixa són les que estan situades en zona d'influència de grans capitals, sobretot de Barcelona. Ho podem visualitzar globalment en la Figura 3.

4.2. Anàlisi migració entre i recíproca

L'anàlisi de la migració recíproca entre zones, comporta fer l'anàlisi de la matriu de moviments simètrics, però sense la influència dels valors de la diagonal, per la qual cosa es realitzarà una anàlisi tipus k-EM de la matriu de simetria.

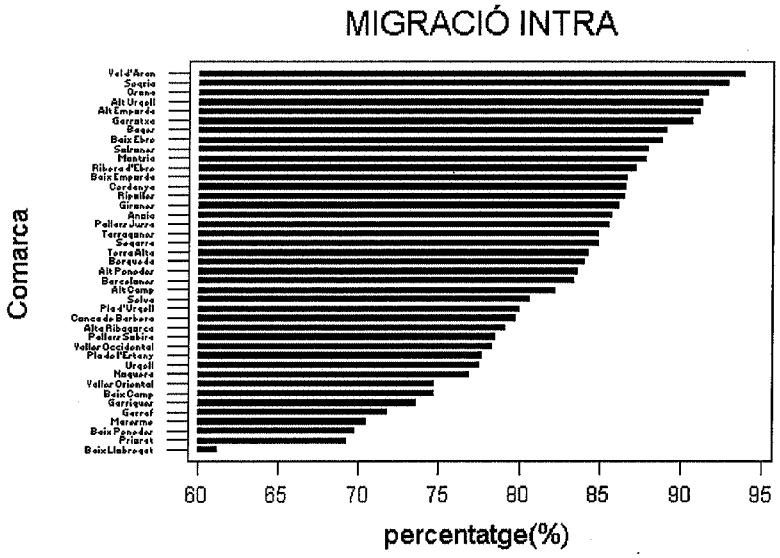


Figura 2. Anàlisi de la migració intra.

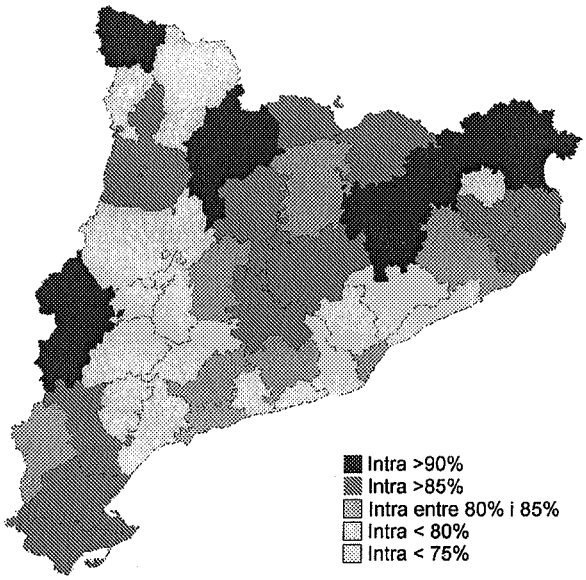


Figura 3. Mapa de la migració intra.

En l'anàlisi dels dos primers eixos factorials (qualitat = 33.7 %) podem trobar en el primer eix les zones on els intercanvis són importants situades totes elles a les comarques meridionals amb dues comarques de transició. Queden configurades a l'extrem de l'eix pel Baix Camp i el Tarragonès, diferenciades del Montsià i Baix Ebre, per una banda, i més cap a l'origen de l'eix Alt Camp, Priorat, Ribera d'Ebre i Terra Alta. Les comarques de transició són la Conca de Barberà i el Baix Penedès. En l'origen de l'eix hi ha un gran cúmul de punts.

Mentre que en el segon eix trobem un altre grup, format per les comarques de la regió occidental amb la mateixa estructura de dues zones i comarques de transició cap a l'origen. Garrigues, Noguera, Pla d'Urgell, Segarra, Segrià, i Urgell formarien la primera zona, i més propera a l'origen, una segona zona formada per Alt Urgell, Alta Ribagorça, els dos Pallars i la Val d'Aran, i de transició cap al gran centre el Solsonès i l'Anoia.

Més explícit és el gràfic amb els tres primers eixos factorials (qualitat = 46.4 %) (Figura 4) on es reprodueix l'estructura que inclou un tercer eix on les comarques nord-orientals ens repeteixen l'estructura zonal (una zona i comarques de transició) i es veu clarament que destaca la posició central del Barcelonès respecte els tres eixos independents de desplaçaments. L'anàlisi ens permet també elaborar un mapa de Catalunya classificat amb comarques amb gran relació entre i recíproca (Figura 5).

4.3. Anàlisi saldos migratoris

Finalment s'efectua l'estudi dels saldos migratoris, és a dir, les diferències entre moviments rebuts d'altres zones i moviments vers altres zones. La realització d'aquesta fase comporta l'estudi de la matriu corresponent a la no simetria. Trobarem en la Figura 6 la representació gràfica d'aquests moviments.

En aquest darrer gràfic dels dos primers eixos factorials (qualitat = 31.15 %) podem veure les relacions de saldos migratoris, podent comprovar primerament que donat el caràcter antisimètric de la matriu analitzada, les representacions dels individus com a productors de saldos positius o negatius, és a dir com a receptors o com a productors de moviments de treballadors, està subjecte a una rotació de 90 graus.

Per altra banda, podem veure relacions de saldos entre les comarques del Baix Camp i el Tarragonès, amb un gran saldo a favor d'aquesta darrera comarca, Uns moviments amb origen al Priorat i destins a l'Alt i Baix Camp i al Tarragonès i finalment d'origen a la Terra Alta i destí a la Ribera d'Ebre. Es pot observar que aquestes zones amb grans diferències de saldos migratoris òbviament corresponen a zones que no tenen una gran migració intra. El tercer eix factorial ens aporta unes inter-relacions triangulars entre les Garrigues, la Noguera i el Pla d'Urgell. Aquestes relacions les podem trobar representades en la Figura 7.

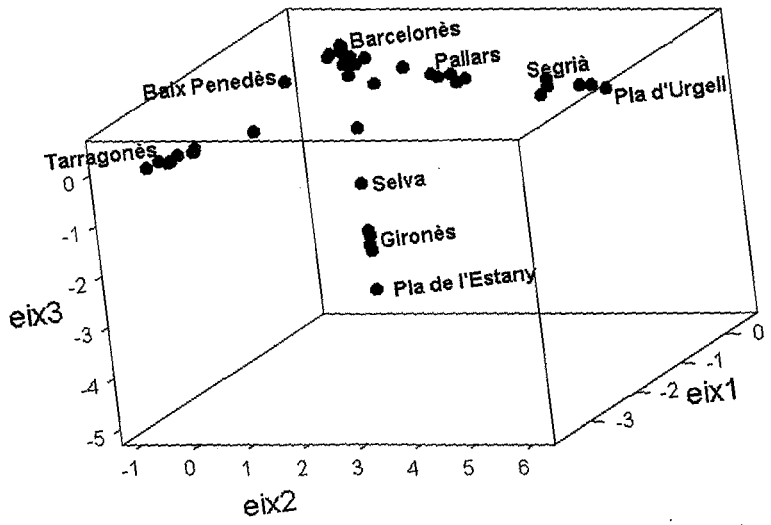


Figura 4. Gràfic factorial 3-EM de l'anàlisi de la migració entre i recíproca.

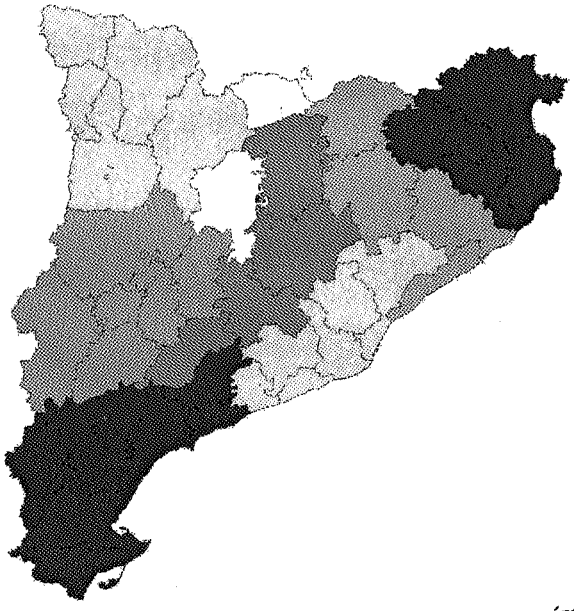


Figura 5. Mapa de les zones de migració entre i recíproca.

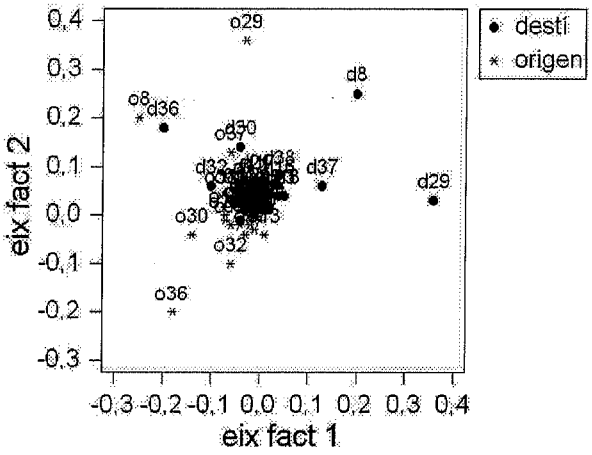


Figura 6. Gràfic factorial de l'anàlisi dels saldos migratoris.

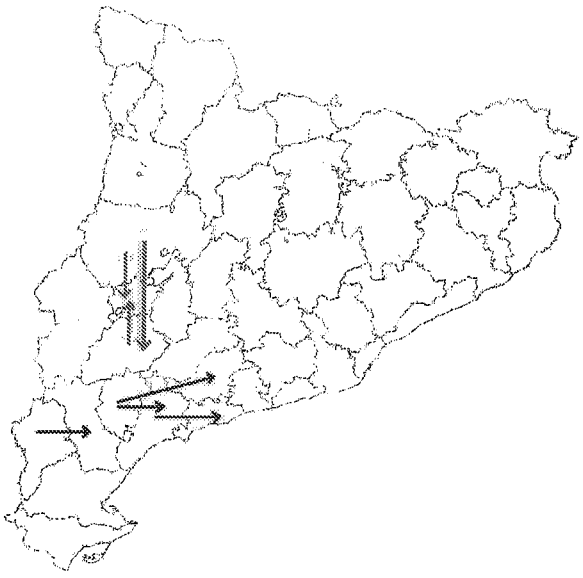


Figura 7. Representació dels principals saldos migratoris.

5. CONCLUSIONS

En aquest treball es presenta una proposta d'anàlisi de matrius quadrades no simètriques. En ella s'han integrat les dues principals metodologies de tractament, la reconstitució de caselles i la descomposició en simetria-antisimetria, aprofitant els avantatges de cadascun dels dos mètodes. Aquesta integració pretén donar una nova visió, més clarificadora, del tractament d'aquests tipus de matrius. S'ha integrat en la metodologia un primer apartat de l'anàlisi dels moviments dins la pròpia zona, ja que és un dels principals indicadors de moviment. En segon lloc, l'anàlisi dels moviments simètrics sense ser influïts pels moviments interns, ens permet detectar regions en el mapa de forta relació mútua i finalment els moviments no simètrics ens permeten detectar els diferents pols d'atracció sobre el territori.

REFERÈNCIES

Benzécri, J. P. (1973). *L'analyse des Données*. Dunod, Paris.

Constantine, A. G. i Gower, J. C. (1978). «Graphical representation of asymmetry». *Applied Statistics*, 27, 297-304.

Daunis-i-Estadella, J. i Aluja-Banet, T. (1997). «Análisis de correspondencias de matrices cuadradas no simétricas. Problemática y tratamientos». *Actas XXIII Congreso Nacional de Estadística e Investigación Operativa*, 40.3-40.4.

Daunis-i-Estadella, J., Aluja-Banet, T. i Thió-Henestrosa, S. (1997). «Reconstitución de datos influyentes en análisis de correspondencias». *NGUS'97 IV International Meeting of Multidimensional Data Analysis*, 108-111.

Dempster, A. P., Laird, N. M. i Rubin, D. R. (1976). «Maximum likelihood from incomplete data via EM algorithm». *Journal of the Royal Statistical Society-Series B*, 93, 1-38.

Escofier, B. (1984). «Analyse factorielle en reference a un modèle. Application a l'analyse de tableaux d'échanges». *Revue de Statistique Appliquée*, 32 25-36.

Foucart, T. (1985). «Tableaux symmétriques et tableaux d'échanges». *Revue de Statistique Appliquée*, 33, 37-54.

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

Greenacre, M. J. (1988). «Correspondence analysis of multivariate categorical data by weighted least-squares». *Biometrika*, 75, 457-467.

Greenacre, M. J. (2000). «Correspondence analysis of square asymmetric matrices». *Applied Statistics*, 49, 297-310.

Institut d'Estadística de Catalunya (2001). «Estadística de població 1996». *Vol 12. Fluxos de mobilitat obligada pel treball o estudi. Dades Comarcals i municipals*, 21-25.

Kazmierczak, J. B. (1978). «Migrations interurbaines dans la banlieue sud de Paris». *Cahiers de l'Analyse des Données*, 3, 203-218.

De Leeuw, J. i van der Heijden, P. G. M. (1988). «Correspondence analysis of incomplete contingence tables». *Psychometrika*, 53, 223-233.

Mutombo, F. K. (1973). «Traîtement des données manquantes et rationalisation d'un réseau de stations de mesures». *PhD thesis, Université Paul et Marie Curie, Paris VI.*

Nora, C. (1974). «Une métode de reconstitution et d'analyse de données incomplétes». *PhD thesis, Université Paul et Marie Curie, Paris VI.*

Taula 1. Moviment de treballadors a les 41 comarques catalanes.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A. Camp-1	10119	4	28	1	0	7	6	417	9	1	33	95	225	0	0	289	10	2	0	2	1
A. Emporda-2	0	32542	7	0	0	16	9	2	2	453	67	4	753	4	3	0	3	0	127	1244	38
A. Penedes-3	17	6	23368	1	0	262	31	15	3	10	983	487	1879	42	1	4	343	5	5	12	21
A. Urgell-4	0	3	2	5508	3	3	13	0	0	2	21	1	162	7	70	0	3	0	2	2	3
A. Ribagorça-5	0	0	0	0	1029	5	1	1	0	4	6	0	91	0	0	0	1	0	0	2	0
Anoia-6	17	19	411	7	0	27531	180	5	1	14	1115	13	1773	12	1	100	36	0	12	6	54
Bages-7	0	15	54	16	0	252	48169	15	3	7	804	15	2130	436	20	4	17	0	8	31	56
B. Camp-8	579	4	41	1	1	20	17	37061	134	7	130	170	1194	9	0	95	33	23	16	30	31
B. Ebre-9	34	2	8	0	0	1	2	234	19487	1	31	23	380	1	1	4	9	4	0	7	10
B. Emporda-10	1	524	7	3	0	7	19	2	1	31730	122	9	1164	2	11	0	3	0	82	2179	65
B. Llobregat-11	14	36	968	3	3	693	464	47	16	28	142346	131	74642	19	9	10	646	1	17	100	486
B. Penedes-12	137	1	889	1	0	29	18	89	12	3	430	12050	2087	3	0	16	469	2	2	9	13
Barcelones-13	61	226	1080	32	5	586	806	206	73	186	52938	241	613094	85	45	24	1040	14	69	563	6693
Bergueda-14	1	4	18	9	0	15	760	5	1	7	62	4	501	10802	70	0	16	0	5	10	21
Cerdanya-15	0	2	3	71	0	1	9	0	0	3	25	0	358	10	4354	0	0	0	2	32	13
Conca Barb.-16	395	2	14	0	1	64	1	110	8	1	37	9	237	1	0	5101	0	17	0	1	4
Garraf-17	17	5	855	3	2	44	33	39	5	0	1272	547	5563	6	6	11	23481	0	2	13	51
Garrigues-18	14	0	7	5	1	10	11	20	0	4	28	8	173	0	0	71	4	4767	0	9	3
Garrotxa-19	0	148	3	1	0	1	5	3	0	53	22	0	262	1	4	0	0	0	16914	656	27
Girones-20	2	683	6	4	1	40	21	6	0	1065	150	6	1072	9	6	1	7	0	265	44429	107
Maresme-21	5	52	59	0	0	51	65	11	10	69	1577	8	25345	10	7	2	43	1	25	319	81561
Montsia-22	19	2	3	1	0	3	10	74	1468	2	24	12	252	0	0	2	4	0	1	8	4
Noguera-23	3	3	2	40	8	39	10	9	2	5	45	5	387	3	3	7	1	16	0	8	8
Osona-24	5	18	12	3	3	26	229	5	4	22	184	1	1672	61	26	0	15	0	27	45	48
Pallars J.-25	1	0	4	13	7	2	4	5	0	4	23	1	268	1	4	0	0	0	1	3	14
Pallars S.-26	3	1	1	10	0	1	7	4	0	1	24	3	198	1	1	0	3	0	3	5	5
Pla d'Urgell-27	5	1	4	6	1	32	8	9	3	1	37	2	190	0	4	11	4	78	3	2	6
Pla Estany-28	0	179	0	0	0	1	0	0	0	67	8	1	131	0	1	0	2	1	300	1268	11
Priorat-29	7	1	3	0	0	3	3	245	11	0	9	7	120	0	0	2	3	27	1	2	5
Ribera Ebre-30	7	0	7	1	0	0	1	175	102	5	10	8	163	0	2	3	7	3	0	0	4
Ripolles-31	0	30	8	0	1	2	8	2	1	25	40	0	382	9	52	0	0	0	159	129	24
Segarra-32	1	2	3	11	1	162	19	6	2	1	42	2	273	1	3	20	3	7	1	1	9
Segria-33	11	10	21	61	18	61	38	52	12	8	155	12	1345	4	7	25	11	245	4	36	22
Selva-34	0	114	8	1	0	9	15	5	4	241	105	10	1393	10	3	1	5	1	167	3331	1608
Solsones-35	0	0	0	54	0	17	141	7	0	0	13	1	123	36	0	4	0	0	0	3	4
Tarragones-36	1000	8	109	5	0	29	19	3918	101	11	287	688	2205	1	1	140	101	19	2	21	34
Terra Alta-37	2	0	1	1	0	2	1	30	67	0	15	9	94	0	1	0	2	0	0	0	3
Urgell-38	7	4	6	4	1	56	9	9	1	0	43	2	270	0	1	30	3	23	0	6	15
Val d'Aran-39	0	2	0	3	7	0	1	0	0	0	9	0	77	0	1	0	2	0	0	1	2
Valles Oc.-40	16	54	292	12	1	199	677	25	16	63	6364	76	37886	25	4	7	109	7	23	132	578
Valles Or.-41	25	25	41	0	0	46	170	12	2	59	1417	16	15007	20	3	0	44	1	9	149	790

	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
A. Camp-1	5	1	2	0	0	1	0	6	3	0	13	7	1	0	991	1	1	2	24	6
A. Emporda-2	1	1	28	0	1	7	139	0	2	17	7	11	107	0	9	0	0	5	84	49
A. Penedes-3	5	2	7	0	1	0	14	0	4	0	2	5	7	4	71	2	1	1	279	70
A. Urgell-4	0	22	10	9	3	2	0	0	0	1	11	72	2	42	3	2	6	2	32	10
A. Ribagorça-5	1	2	0	23	0	2	0	0	0	0	1	69	0	1	3	1	2	47	6	3
Anoia-6	0	5	13	1	0	11	1	1	2	2	331	34	6	9	14	0	20	2	300	86
Bages-7	1	4	272	1	5	3	7	1	3	10	40	28	19	159	35	1	3	1	1185	272
B. Camp-8	29	7	12	1	3	10	0	50	263	2	2	69	6	4	9449	9	8	9	119	47
B. Ebre-9	1180	0	3	4	2	2	1	4	61	4	7	23	6	0	331	34	2	0	22	22
B. Emporda-10	0	0	21	3	1	0	49	0	0	15	2	6	319	0	10	0	2	2	179	86
B. Llobregat-11	13	5	126	5	5	6	10	0	12	6	20	56	81	3	210	0	16	6	9424	1912
B. Penedes-12	50	0	9	0	0	6	7	6	4	2	0	14	7	5	640	1	2	1	232	45
Barcelones-13	34	32	564	24	9	30	30	12	36	62	77	243	663	24	618	1	32	17	38554	16455
Bergueda-14	0	3	278	0	3	1	2	0	1	36	4	9	5	90	6	0	0	1	90	32
Cerdanya-15	0	5	10	0	0	1	1	0	0	22	3	5	20	6	1	0	0	0	63	10
Conca Barb.-16	4	2	4	0	1	2	0	3	1	0	17	29	3	0	262	0	30	2	30	10
Garraf-17	3	2	6	1	0	2	1	2	7	122	0	14	9	0	127	0	1	2	328	135
Garrigues-18	0	14	6	7	2	303	0	6	12	1	40	830	1	1	43	3	49	3	20	8
Garrotxa-19	0	0	24	0	0	0	208	0	0	92	0	1	186	0	4	0	1	0	33	21
Girones-20	1	1	33	0	0	0	707	0	0	33	2	7	2620	0	19	0	1	2	180	101
Maresme-21	10	6	99	1	0	10	25	0	3	8	0	13	1908	1	61	0	4	4	2011	2505
Montsia-22	15782	3	5	2	0	1	0	0	34	0	0	16	6	1	184	21	1	1	30	7
Noguera-23	0	9383	2	20	11	212	0	1	0	5	307	1265	8	22	26	0	277	14	36	19
Osona-24	2	1	45904	0	1	2	12	0	2	307	8	15	98	9	18	0	0	1	256	1065
Pallars J.-25	2	24	0	3890	42	6	0	0	2	0	19	131	4	3	16	0	7	5	38	8
Pallars S.-26	0	10	3	48	1839	5	0	0	1	0	3	94	2	0	3	0	3	19	37	7
Pla d'Urgell-27	1	170	2	5	0	8507	1	0	1	4	208	866	3	1	18	0	401	7	27	17
Pla Estany-28	0	0	8	0	1	0	7345	0	0	4	0	2	91	0	2	0	0	0	23	13
Priorat-29	2	0	1	0	1	0	0	2095	217	0	1	13	1	0	224	2	1	0	19	1
Ribera Ebre-30	9	4	0	0	3	0	0	87	6113	0	2	18	2	0	181	59	1	1	26	8
Ripolles-31	0	0	362	1	0	0	13	0	0	8980	0	2	18	0	6	0	0	0	75	59
Segarra-32	0	49	3	1	0	35	0	0	0	1	5895	117	0	26	3	0	198	1	45	7
Segria-33	3	451	18	62	24	656	3	5	41	14	157	55346	11	24	136	1	214	43	148	61
Selva-34	0	1	87	0	0	1	62	0	1	5	4	10	34022	3	11	0	0	3	284	700
Solsones-35	0	8	3	5	3	4	1	0	0	2	22	35	5	3871	1	0	1	4	28	8
Tarragones-36	44	2	13	0	3	1	1	25	115	3	6	73	9	2	52114	11	9	1	210	70
Terra Alta-37	11	2	3	1	0	0	0	4	280	0	5	2	1	0	95	3434	2	0	8	1
Urgell-38	1	220	6	10	3	380	1	0	0	0	907	466	5	4	26	0	8858	5	41	17
Val d'Aran-39	0	3	1	6	4	1	0	0	0	0	0	53	1	0	4	0	0	2913	11	2
Valles Oc.-40	14	12	205	4	7	7	11	1	10	13	21	54	159	6	129	1	9	10	194624	6843
Valles Or.-41	3	0	561	0	3	1	12	0	3	10	15	11	632	2	42	0	4	1	8989	82507

ENGLISH SUMMARY

ANALYSIS OF SQUARE NON-SYMMETRIC TABLES: AN INTEGRAL VISION BY CORRESPONDENCE ANALYSIS

J. DAUNIS I ESTADELLA*
T. ALUJA BANET**
S. THIÓ FDEZ. DE HENESTROSA*

In this paper a new integrated approach to the analysis of square non-symmetric tables is introduced by means of correspondence analysis. The application of correspondence analysis to such tables is not successful due to the strong role played by the diagonal values: overloaded diagonals and structural zeros. Two main families of methods of resolution are integrated in this paper. The resulting method is applied to a study of commuting between the 41 catalan counties.

Keywords: Square matrices, overloaded diagonals, correspondence analysis, structural zeros, singular value decomposition, missing data

AMS Classification (MSC 2000): 62H17, 62H25

* Departament d'Informàtica i Matemàtica Aplicada. Universitat de Girona. Av. Lluís Santaló s/n. 17071 GIRONA. Email: josep.daunis@udg.es santiago.thio@udg.es.
** Departament d'Estadística i Investigació Operativa. Universitat Politècnica de Catalunya. C/Pau Gargallo, núm. 5. 08028 BARCELONA. Email: tomas.aluja@upc.es.

- Received December 2001.
- Accepted October 2002.

In many scientific areas we deal with square tables, where rows and columns refer to the same set of objects or categories. Such non-symmetric square tables are very common, where there is not a symmetric relation between rows and columns. We illustrate our problem in the context of migratory movements due to work (*commuting*) between the 41 catalan counties, using data from the 1996 census.
Often these tables are analyzed by means of correspondence analysis (CA), but the application of CA to such tables is not successful due to the strong role played by the diagonal values in the total inertia. On this diagonal the entries are often greater than the non-diagonal entries (overloaded diagonals) or empty due to the table structure (structural zeros).
The table's total inertia can be decomposed as the sum of diagonal inertia $I(D)$ and non-diagonal inertia $I(\overline{D})$.
The analysis of these tables is problematic and there are two families of methodologies to do it, based mainly on the reconstitution of diagonal elements, or trying to decompose the analysis referring to the symmetry and skew-symmetry of data. Our goal is to analyze both methodologies and to integrate them, avoiding the influence of those elements that produce structural inertia and analyzing the joint relationship existing between row and column categories separately from the polarities established among them.
These existing methodologies are:

- Replacement of diagonal elements under independence hypothesis
- Treatment of diagonal elements as *missing values*
- Replacement of diagonal elements using the reconstitution formula
- Decomposition of the matrix as the sum of a symmetric matrix and a skew-symmetric matrix

The proposed analysis can be divided into the following three steps, named by analogy withmmigratory movements:

1. Intra migration analysis
2. Between and reciprocal migration analysis
3. Migratory balance analysis

The first step, called intra migration analysis, studies the movements inside the same zone, that is it analyzes the importance of the diagonal in reference to the margins.
The second step, called between and reciprocal migration analysis, is the analysis of the symmetric movements between every two pairs of categories, that is, the study of the symmetric component without the diagonal influence using the k-EM reconstitution algorithm on the matrix that contains the symmetric component.

Finally, the last step, called migratory balance analysis, consists of the analysis of the difference of non-reciprocal movements between categories, that is the study of the skew-symmetric component of movements between the objects.
As a conclusion, this work presents a new proposal for the analysis of non-symmetric square matrices. Here we integrate the reconstitution methodology and the decomposition in both symmetric and skew-symmetric components, gaining the advantages of both methods. This integration tries to give a new vision, a clearer vision, of treatment of this kind of data matrix.

QÜESTIIÓ, vol. 26, 3, p. 503-513, 2002

LA DISTÀNCIA DE l'EIXAMPLE

M. GREENACRE
Universitat Pompeu Fabra*

Inspirant-nos en el disseny de l'Eixample de Barcelona per Ildefons Cerdà, definim una nova distància estadística, anomenada la distància de l'eixample. Estudiarem algunes propietats d'aquesta funció de distància i proposem algunes possibles aplicacions.

The Eixample Distance

Paraules clau: Statistical distance, Manhattan distance, Euclidean distance

Classificació AMS (MSC 2000): 62H17, 62H25

* Universitat Pompeu Fabra. Barcelona.

– Rebut al desembre de 2001.
– Acceptat a l'octubre de 2002.

1. INTRODUCCIÓ

Ildefons Cerdà i Sunyer (1815-1876), el dissenyador de l'Eixample de Barcelona, no era només enginyer, urbaniste i catalanista. El 1856, tres anys abans de presentar *El Proyecto de Ensanche de 1859,* va publicar el seu primer llibre: *Una Monografía Estadística de la Classe Obrera de Barcelona.* Era una obra molt detallada, plena de taules i llistes, basada en fets recollits pel mateix Cerdà, i era el primer estudi de l'espai urbà de Barcelona, dels serveis de transport i de sanitat i de les condicions de treball dels seus ciutadans. Mai no s'havia donat a aquests assumptes una base estadística, i aquesta base sòlida de dades el va conduir a una visió global de les necessitats de la ciutat renovada i reconstruïda.

Figura 1. Ildefons Cerdà i Sunyer (1815-1876).

No entrarem en més detalls del Pla Cerdà, a part de citar, segons l'exposició del Pla Cerdà i de *Urbs Quadrada*, organitzada a la Universitat Pompeu Fabra per l'Institut d'Estudis Territorials l'any 1999, que Cerdà va deduir «*la dimensió de la mansana... a través d'una fórmula matemàtica que considera com variables l'amplada del carrer, la longitud de la façana, la fondària del solar, les característiques de la unitat d'habitació*

definida el 1855 i el nombre de metres quadrats per habitant». Sembla que Cerdà va tenir habilitat també per a l'anàlisi multivariant.

En aquest article ens inspirem en el disseny bàsic de les mansanes de l'eixample, proposada per en Cerdà, per donar a la llum una nova mesura de distància estadística. Els que han caminat per l'eixample s'han adonat que és possible escurçar la distància entre dos punts aprofitant dels xamfrans[1] a cada cantonada d'una mansana. Definirem una distància suggerida d'aquesta observació, que anomenarem la ***distància de l'eixample***, i n'estudiarem algunes propietats i possibles aplicacions.

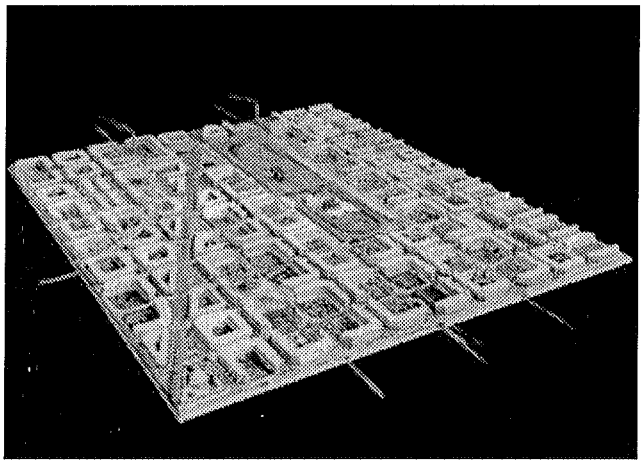


Figura 2. La visió d'en Cerdà.

2. DISTÀNCIA DE L'EIXAMPLE: DEFINICIÓ

Per al nostre objectiu és suficient de considerar un eixample sense carrers ni voreres entre les mansanes. Això no és cap restricció, perquè podem suposar que aquesta amplada està continguda en la longitud de les façanes. És a dir, si el carrer i les voreres tenen un amplada total de 35 m (que era, de fet, l'amplada recomanada per Cerdà), es pot comptar aquesta longitud en la de la façana de cada edifici.

[1] **xamfrà** Xamfrà constituïda per un pla que forma angles obtusos, especialment de 135°, amb cadascuna de les dues façanes que la determinen.

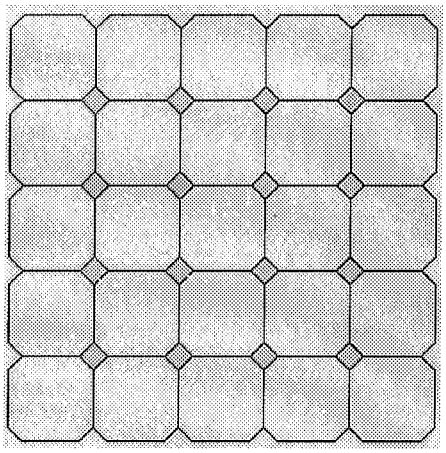


Figura 3. Un Eixample sense Carrers ni Voreres.

Aleshores considerem un eixample com el dibuixat a la Figura 3. Cada mansana té vuit costats, quatre façanes i quatre xamfrans (Figura 4), i suposem que l'angle del xamfrà és de 135°. Sigui a la longitud de la façana i b la longitud del seu xamfrà. Anomenem la longitud del quadre format per la mansana la *longitud exterior*. La longitud exterior de la mansana és igual a $a + \sqrt{2}b$.

Suposem que hem de calcular la distància entre un punt amb coordenades (x_1, y_1) i un altre amb coordenades (x_2, y_2). Siguin les diferències absolutes entre les coordenades:

$$x = |x_1 - x_2| \qquad y = |y_1 - y_2|$$

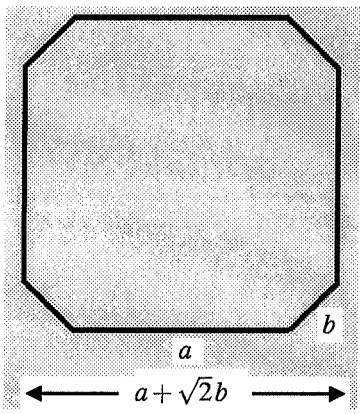


Figura 4. Longitud de la façana i de la xamfrà, i longitud exterior.

Provisionalment suposem que els números positius x i y tenen un factor comú. Un exemple senzill seria $x = 2$ i $y = 3$ amb factor comú 1. Llavors suposem que la longitud exterior de cada façana és igual a 1, és a dir $a + \sqrt{2}b = 1$.

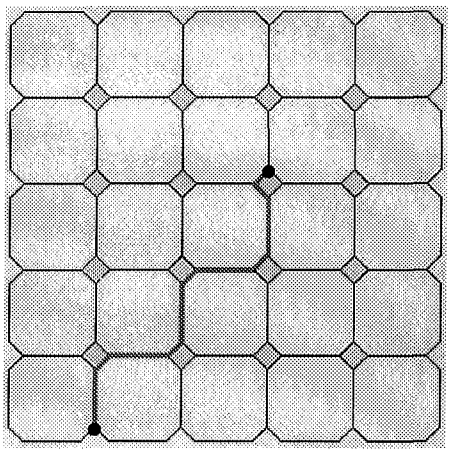


Figura 5. La distància de l'eixample (en traç gruixut).

A la Figura 5 els dos punts en negre tenen la posició relativa desitjada (comencem i acabem sempre a la mateixa posició en una mansana, en aquest cas al punt nord de la cruïlla). Per caminar una distància mínima entre aquests dos punts hauríem de passar per 5 façanes i 6 xamfrans. És fàcil deduir que el número de façanes és igual a la suma de x i y, i que el número de xamfrans és igual a la suma més la diferència absoluta. En aquest exemple $2+3=5$ i $2+3+|2-3|=6$. Llavors la distància a caminar és igual a $5a+6b$.

Un altre exemple més concret és la distància a caminar entre el *Forum Vergés*, actualment l'Institut d'Educació Continua (IdEC) de la Universitat Pompeu Fabra, a la cruïlla de Balmes i Rosselló, i *La Pedrera* de Gaudí a la cruïlla de Provença i Passeig de Gràcia. Aquests edificis tenen la mateixa orientació al sud-oest de les mansanes respectives, i tenen una illa de diferència en la direcció nord-sud i dues en la direcció est-oest. Aleshores, la distància a caminar és igual a $3a+4b$, on a és en aquest cas la longitud de la façana més l'amplada del carrer (suposant, és clar, que cada mansana és perfectament quadrada).

Aquest resultat no depèn de la mida de la mansana, però sí que depèn de la relació entre a i b. Si enlloc de mansanes de longitud exterior 1, considerem mansanes de longitud exterior $1/n$, on n és un enter, el nombre de façanes i xamfrans a caminar està multiplicat per n, però la longitud de cada façana i xamfrà està dividida per n, llavors la distància a

caminar queda igual. En general, per a nombres x i y reals, no trobarem necessàriament un factor comú per dividir el pla en mansanes. Però, per a nombres arrodonits sempre podem escriure:

$$x = m \times 10^k \qquad y = n \times 10^k$$

on m, n i k són enters, m, n positius i k tant gran en valor absolut com volguem. Aleshores la longitud exterior de la façana serà 10^k i el punt tindrà m mansanes de diferència a l'eix horitzontal i n a l'eix vertical.

Les longituds a i b de la façana amb longitud exterior de 10^k han de satisfer la igualtat: $a + \sqrt{2}b = 10^k$. Si definim la relació entre b i a com $c = b/a$, llavors:

$$a = \frac{1}{1+\sqrt{2}c} = 10^k \qquad b = \frac{c}{1+\sqrt{2}c} 10^k$$

i la distància a caminar és igual a:

$$\begin{aligned} d &= (m+n)a + (m+n+|m-n|)b \\ &= (x+y)\frac{1}{1+\sqrt{2}c} + (x+y+|x-y|)\frac{c}{1+\sqrt{2}c} \end{aligned}$$

Introduïm algunes definicions:

- Una **mansana de Cerdà** és un octàgon que té la forma de la Figura 4: tots els angles interiors són de 135° i els costats i els xamfrans poden tenir longituds diferents.
- El **número de Cerdà** c és la relació b/a entre les longituds del xamfrà i de la façana d'una mansana.
- Un **eixample** és una matriu de mansanes de Cerdà.

I, finalment, la definició de la distància de l'eixample:

- Sigui x i y la diferència absoluta entre les coordenades respectives dels punts (x_1, y_1) i (x_2, y_2): $x = |x_1 - x_2|$, $y = |y_1 - y_2|$; sigui c el número de Cerdà de l'eixample. La **distància de l'eixample** entre els dos punts està definida com:

$$\begin{aligned} &\frac{1}{1+\sqrt{2}c}(x+y) + \frac{c}{1+\sqrt{2}c}(x+y+|x-y|) \\ &= \frac{1+c}{1+\sqrt{2}c}(x+y) + \frac{c}{1+\sqrt{2}c}(|x-y|) \end{aligned}$$

3. PROPIETATS

- És evident que la distància de l'eixample és una mètrica, és a dir que satisfà la desigualtat triangular. Per anar d'un punt A a un punt B, és impossible que el camí de A a un tercer punt C i de C a B sigui més curt.
- No és evident si la distància de l'eixample es pot representar en un espai euclidià o no. És un problema obert provar que podem fer una «immersió» («embedding») de les distàncies entre un conjunt de punts dins d'un espai euclidià.
- La distància de l'eixample és sempre més llarga que la distància Pitagoriana (norma L_2).
- Comparada amb la distància de Manhattan (norma L_1), la distància de l'eixample pot ser més curta o més llarga depenent de la «diagonalitat» del trajecte. Per fer un trajecte de x mansanes horitzontals i y mansanes verticals, la distància de Manhattan D_M i la distància de l'eixample D_E són:

$$D_M = (x+y)(a+\sqrt{2}b)$$
$$D_E = (x+y)a+(x+y+|x-y|)b$$

  Es pot demostrar que $D_E < D_M$ quan el pendent y/x és entre els valors $\sqrt{2}-1 = 0,4142$ i $1/(\sqrt{2}-1) = 2,4142$. La curiositat d'aquest resultat és que no depèn del número de Cerdà. Quan s'ha de caminar entre dos punts i el trajecte és «menys diagonal» que aproximadament dues illes en un sentit i cinc illes en l'altre, per tant la distància de l'eixample és més llarga que la distància de Manhattan. Més concretament si heu de caminar entre el Forum Vergés i La Pedrera (pendent 0,5), el camí és més curt gràcies al disseny de l'eixample. Però per caminar entre el Forum Vergés i la cruïlla Rambla de Catalunya i València (pendent 3), heu de caminar una mica més a causa dels xamfrans extres que cal passar baixant Balmes o la Rambla.

4 APLICACIONS

L'idea de parametritzar la mesura de distància és nova i té aplicacions potencials. Jugant amb el paràmetre c de la distància podrem introduir més informació en la distància entre objectes que finalment representarem visualment en un gràfic. Per exemple, podrem estimar c per ajustar les distàncies millor al pla bidimensional de representació, i després podrem donar una interpretació de la estimació. És també possible de generalitzar la distància de l'eixample a illes rectangulars, que té sentit quan tenim dues variables d'escales diferents. En aquest cas la xamfrà no tindrà necessàriament un angle interior de 135°.

Una generalització multidimensional serà igualment interessant. Inspirant-nos en la distància de l'eixample, podem definir distàncies que són combinacions lineals de distàncies conegudes, p.e. les distàncies L_1 i L_2:

$$D(\mathbf{x},\mathbf{y},\lambda) = \lambda d_1(\mathbf{x},\mathbf{y}) + (1-\lambda)d_2(\mathbf{x},\mathbf{y})$$

on $\mathbf{x}$ i $\mathbf{y}$ són vectors multidimensionels, d_1 és la distància L_1 de Manhattan, d_2 és la distància L_2 euclideana, i λ és un paràmetre entre 0 i 1 que es pot definir o estimar, segons l'aplicació.

AGRAÏMENTS

Voldria donar gràcies a en Daniel Serra, de l'Institut d'Estudis Territorials, d'haver portat a la Universitat Pompeu Fabra l'exposició «Urbs Quadrada» sobre el projecte de l'Eixample, i pels seus comentaris útils. S'agraeix l'ajut BFM2000-1064 del Ministeri de Ciència i Tecnologia.

REFERÈNCIES

Magrinyà, F. & Tarragó, S. (1994). *Cerdà. Urbs i Territori, una visió del futur*. Catàleg de l'exposició.

ENGLISH SUMMARY

THE EIXAMPLE DISTANCE

M. GREENACRE
Universitat Pompeu Fabra*

Taking our inspiration from the design of the Eixample district in Barcelona by Ildefons Cerdà, we define a new statistical distance function, called the Eixample Distance. We study some properties of this distance and propose some possible applications.

Keywords: Statistical distance, Manhattan distance, Euclidean distance

AMS Classification (MSC 2000): 62H17, 62H25

* Universitat Pompeu Fabra. Barcelona.

- Received December 2001.
- Accepted October 2002.

1. INTRODUCTION

The Catalan engineer and town planner, Ildefons Cerdà (1815-1876) was commissioned with the remodelling of the district of Barcelona called the Eixample (literally: «broad axis»). Cerdà based his design partially on a detailed statistical analysis of the urban area of Barcelona, its public transport and the needs of the residents. The key idea was an urban layout in the form of blocks, each of which has an interior courtyard and a characteristic octagonal shape with its four oblique «cut-off» corners providing spacious intersections. The design of the Eixample has inspired a distance function, called the Eixample distance, which imitates a path traced by a pedestrian in this district trying to minimize the distance between two points by taking maximum advantage of the oblique corners.

2. THE EIXAMPLE DISTANCE: DEFINITION

The Eixample distance is a particular combination of a distance in a «north-south» or «east-west» sense and a distance in a diagonal sense, like the sides and oblique corners respectively of a typical Eixample block in Barcelona.

This distance depends on the ratio between the length of the diagonal corner of the block (called a *xamfrà* in catalan) and the length of the façade (or *façana*), where we incorporate the width of the street into the façade without lack of generality. We call this ratio the Cerdà index, and denote it by c.

The Eixample distance between two points (x_1,x_2) and (y_1,y_2) , where $x = |x_1 - x_2|$ and $y = |y_1 - y_2|$ are the absolute differences between the respective coordinates, can then be shown to be equal to:

$$\frac{1}{1+\sqrt{2}c}(x+y)+\frac{c}{1+\sqrt{2}c}(x+y+|x-y|)$$
$$=\frac{1+c}{1+\sqrt{2}c}(x+y)+\frac{c}{1+\sqrt{2}c}(|x-y|)$$

3. PROPERTIES AND APPLICATIONS

The Eixample distance is a metric, but it is not clear whether it is Euclidean embeddable or not. It is always larger than the Pythagorean (or L_2) distance. Compared to the Manhattan (or city-block, or L_1) distance it will be shorter or longer depending on the «diagonality» of the trajectory between the two points. When the slope between the

two points is between the values $\sqrt{2}-1$ and $1/(\sqrt{2}-1)$, the Eixample distance will be shorter than the Manhattan distance.

The novelty of this new distance measure is that it contains a parameter, the Cerdà index. This parameter can be varied between 0 and 1 according to some external criterion, for example the quality of the visualization of the distances in a multidimensional scaling. Thus the introduction of the parameter coul lead to an improved fit of distance representations in multidimensional scaling displays. Parametrizing the distance function allows additional information to be included in the distance and in the eventual analysis of distance.

Since the Eixample distance is a type of combination of different distances, this also suggests defining distance functions that are combinations of well-known distances, such as a convex linear combination of the L_1 and L_2 distances, leading to a more flexible class of distance measures.

Investigació Operativa

QÜESTIIÓ, vol. 26, 3, p. 517-536, 2002

UN ALGORITMO PARA EL PROBLEMA DE BIFLUJO MÁXIMO SIMÉTRICO NO DIRIGIDO

A. SEDEÑO-NODA
C. GONZÁLEZ-MARTÍN
Universidad de La Laguna*

En este trabajo proponemos un algoritmo de $O(nm\log U)$ para resolver el problema de biflujo máximo simétrico en una red no dirigida. Para resolver este problema se introduce un cambio de variable que permite dividir el problema original en dos problemas de flujo máximo. De esta manera se obtiene un algoritmo sencillo y eficiente donde se utilizan las herramientas computacionales propias de la resolución del clásico problema de maximizar un único flujo.

An algorithm for the undirected symmetric maximum biflow problem

Palabras clave: Problema de biflujo máximo simétrico, algoritmo de escalado en las capacidades, redes no dirigidas

Clasificación AMS (MSC 2000): 90C27

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

- Recibido en agosto de 2001.
- Aceptado en mayo de 2002.

1. INTRODUCCIÓN

Son numerosos los casos reales en los que aparecen problemas de flujos en redes (ver, por ejemplo, Ahuja *et al.* (1993)). Entre estos problemas, unos de los más estudiados son los *problemas de flujo máximo*, cuya versión clásica considera un único flujo sobre una red conexa dirigida con un único nodo fuente y un único nodo sumidero. Ford y Fulkerson (1956) estudian y resuelven este problema proponiendo el conocido Método de los Caminos Incrementales, que toma como base el denominado Teorema de Flujo-Máximo Corte-Mínimo.

Cuando se consideran, simultáneamente, varios tipos de flujos aparecen los correspondientes problemas de flujos múltiples. Ejemplos de estos casos aparecen en problemas de tráfico urbano, problemas de transporte ferroviario, problemas de sistemas de comunicaciones de diverso tipo, etc. (ver, por ejemplo, Ahuja *et al.* (1993)).

Hemos de señalar que, en general, para los problemas en los que se maximizan tres o más flujos no se verifica el equivalente del Teorema de Flujo-Máximo Corte-Mínimo. Sin embargo, para el caso de dos flujos (*problema de biflujo máximo*), Hu (1963) demuestra el Teorema de Biflujo-Máximo Corte-Mínimo. Una demostración del dual del teorema de Hu, es dada por Seymour (1978).

En este trabajo consideramos el *problema de biflujo máximo simétrico no dirigido* (BFMS). Para ilustrar el caso en estudio, consideremos el siguiente ejemplo. Dos ciudades, de características similares, deben estar interconectadas por una red de tráfico terrestre. Señalemos cuatro vértices de la red, distinguiendo los dos vértices por los que se accede a la red de tráfico desde cada una de las dos ciudades y los otros dos vértices por los que se sale de la citada red hacia la ciudad contraria a la de partida. Si, como es lógico, diferenciamos el tráfico desde la primera ciudad hacia la segunda y desde ésta a la primera, tendremos un problema de dos flujos cuya suma, para hacer eficiente la gestión de la red, habrá que maximizar. Teniendo presente, además, que los desplazamientos desde una ciudad a otra deben producirse (dentro de una determinada unidad de tiempo) en sentido contrario, la similitud de características de las dos ciudades impondrá la igualdad de la magnitud de los dos flujos considerados.

Este problema es un caso particular del problema de biflujo máximo no dirigido (BFM). Sin embargo, la resolución de este problema requiere un tratamiento específico ya que, al contrario de lo que ocurre para el problema BFM, su solución no queda caracterizada por el teorema de Hu. Hemos de señalar que, en la literatura que hemos consultado, no existe ningún tipo de estudio referido al problema BFMS.

Recordemos que la formulación del problema BFM se realiza habitualmente sobre una red dirigida donde los flujos pueden tomar valores negativos. Esta formulación aparece en los trabajos de Hu (1963), Rothschild y Whinston (1966), Sakarovitch (1973), Itai (1978), Seymour (1978) y Sedeño-Noda *et al.* (2001). En el presente trabajo, partimos

de una formulación similar para el problema BFMS, realizando un cambio de variable introducido en Sedeño-Noda *et al.* (2001) que posibilita separar el problema original en dos problemas de flujo máximo. Esta forma de actuar permite desarrollar un algoritmo que resuelve el problema estudiado de forma eficiente.

Después de esta introducción, en la segunda sección establecemos la formulación del problema de biflujo máximo simétrico. En la sección tercera, recordamos los conceptos y definiciones asociados al problema de flujo máximo necesarios para el entendimiento del algoritmo propuesto posteriormente. En la cuarta sección, introducimos la formulación equivalente del problema. En la quinta sección presentamos el algoritmo que resuelve el problema estudiado, incluyendo su ejecución sobre un ejemplo, y su complejidad teórica. En la sexta sección, el trabajo finaliza con una serie de conclusiones.

2. NOTACIÓN Y FORMALIZACIÓN DEL PROBLEMA

Sea $G = (V,A)$ una red dirigida, donde V es el conjunto de n nodos y A es el conjunto de m arcos. Distinguimos cuatro nodos especiales en G: los nodos fuentes s^1, s^2y los nodos sumideros t^1, t^2. Al igual que en Sakarovitch (1973), consideraremos que la red es antisimétrica, es decir, si $(i,j) \in A \Rightarrow (j,i) \notin A$. Dado cualquier nodo i, definimos los conjuntos $\text{Pred}(i) = \{j \in V/(j,i) \in A\}$ y $\text{Suc}(i) = \{j \in V/(i,j) \in A\}$ (respectivamente, los vértices predecesores y sucesores del vértice dado). $\forall (i,j) \in A$, $u_{ij} \geq 0$ denota la mayor cantidad de flujo (capacidad) que puede soportar (i,j). Denominaremos U al $\max\{u_{ij}/(i,j) \in A\}$ y $\text{A}(i) = \{(i,j) \in A : j \in \text{Suc}(i)\}$ al conjunto de arcos que salen del nodo i.

Un *biflujo* es un par $x = (x^1,x^2)$ donde $x^k : A \longrightarrow R$ con $k = 1,2$, que satisface:

$$\sum_{j \in \text{Suc}(i)} x_{ij}^k - \sum_{j \in \text{Pred}(i)} x_{ji}^k = \begin{cases} f^k, & \text{si } i = s^k \\ 0, & \text{si } i \in V - \{s^k,t^k\} \\ -f^k, & \text{si } i = t^k \end{cases} \tag{1.1}$$

$$|x_{ij}^1| + |x_{ij}^2| \leq u_{ij}, \qquad \forall (i,j) \in A \tag{1.2}$$

para algún $f = (f^1,f^2)$, con $f^k \geq 0$ para $k = 1,2$. El problema de biflujo máximo (BFM) consiste en encontrar un biflujo x tal que la suma $f^1 + f^2$ sea máxima. En otras palabras, el problema BFM consiste en enviar simultáneamente la mayor cantidad posible de flujo de la fuente s^1 al sumidero t^1 y de la fuente s^2 al sumidero t^2, satisfaciendo las restricciones de capacidades (1.2) y las de conservación del flujo (1.1). Notar que las restricciones (1.2) implican que no se satisfaga la propiedad de unimodularidad exhibida en los problemas de redes normalizadas en las que circula un único flujo. Por lo tanto,

el óptimo del *patrón de biflujo* puede ser un vector no entero. Evidentemente, $s^1 \neq t^1$ y $s^2 \neq t^2$. Asumiremos, sin pérdida de generalidad, que $s^1 \notin \{s^2, t^2\}$ y $t^1 \notin \{s^2, t^2\}$.

El problema de biflujo máximo simétrico (BFMS) se plantea en los mismos términos que el problema BFM, pero con la restricción añadida de que la cantidad enviada entre los nodos s^1 y t^1 debe coincidir con la que se envíe entre los nodos s^2 y t^2, es decir, se ha de cumplir que $f^1 = f^2$.

Debemos notar que las restricciones (1.2) permiten la posibilidad de valores negativos en el biflujo. Esto significa que si el valor del k-ésimo flujo x_{ij}^k es negativo para el arco $(i,j) \in A$, entonces el flujo atraviesa el arco en sentido opuesto.

3. CONCEPTOS Y PROPIEDADES BÁSICOS

Los siguientes conceptos y propiedades son conocidos para los problemas en los que se maximiza un único flujo entre la fuente s y el sumidero t de una red dirigida G. Un *corte* es una partición del conjunto de nodos V en dos subconjuntos Sy $\bar{S} = V - S$. Un corte define un subconjunto en A denotado por $[S,\bar{S}]$: el conjunto de arcos con un extremo en S y el otro extremo en $\bar{S}$. La *capacidad de un corte* viene dada por la suma de las capacidades de los arcos que salen del conjunto S y llegan a $\bar{S}$, es decir, los arcos del conjunto que denotaremos por $(S,\bar{S})$. Esta capacidad es igual a $u[S,\bar{S}] = \sum_{(i,j)\in(S,\bar{S})} u_{ij}$.

Debemos notar que $[S,\bar{S}] = (S,\bar{S}) \cup (\bar{S},S)$.

Dado un flujo x, la *capacidad residual* de cualquier arco $(i,j) \in A$ es la máxima cantidad adicional de flujo que se puede enviar desde el nodo i al nodo j mediante los arcos (i,j) y (j,i), cuyo valor es $r_{ij} = u_{ij} - x_{ij} + x_{ji}$. La red $R(x)$ que contiene los arcos con capacidades residuales positivas se denomina *red residual.* Por tanto, la red residual $R(x)$ tiene el mismo conjunto de nodos V y, por cada arco $(i,j) \in A$ se introducen los arcos (i,j) y (j,i) con capacidades residuales r_{ij} y r_{ji} respectivamente.

Un *camino incremental* en $R(x)$ es un camino dirigido de s a t en $R(x)$. Definimos la *capacidad residual* de un camino como la menor de las capacidades residuales de los arcos en el camino. El resultado que indica que *un flujo x es máximo si la red residual $R(x)$ no contiene caminos incrementales* es ampliamente conocido y constituye la regla de parada del ya clásico *Algoritmo de Caminos Incrementales* de Ford y Fulkerson (1956). Estos autores demuestran que *la mayor cantidad de flujo que se puede enviar desde el nodo fuente s al nodo sumidero t coincide con la menor de las capacidades de todos los $s-t$ cortes de la red* (Teorema del *Flujo-Máximo Corte-Mínimo*).

Entenderemos que $g = (V,E)$ es la red no dirigida obtenida cuando no se consideran direcciones en los arcos de G y que $G' = (V,A')$ es la versión dirigida de g, donde

$A' = \{(i,j),(j,i)/\{i,j\} \in E\}$. Necesitaremos particularizar algunos de los conceptos anteriores para la red no dirigida g. Así, debemos significar que la capacidad de un corte en g, $u^*\left[S,\bar{S}\right]$[1], coincide con la suma de las capacidades de los arcos $(S,\bar{S})$ y $(\bar{S},S)$ en la red dirigida antisimétrica G, es decir, $u^*\left[S,\bar{S}\right] = \sum_{(i,j)\in(S,\bar{S})} u_{ij} + \sum_{(i,j)\in(\bar{S},S)} u_{ij}$.

Para el problema de biflujo máximo, Hu (1963) da el siguiente resultado para el caso de un grafo no dirigido:

Teorema 1. *(Teorema del Biflujo-Máximo Corte-Mínimo). El valor Máximo de* $f^1 + f^2 = \min\left(u\left[S^1,\bar{S}^1\right], u\left[S^2,\bar{S}^2\right]\right)$, *donde* $\left[S^1,\bar{S}^1\right]$ *recorre los posibles cortes tales que* $s^1,s^2 \in S^1$ *y* $t^1,t^2 \in \bar{S}^1$ *y* $\left[S^2,\bar{S}^2\right]$ *recorre los posibles cortes tales que* $s^1,t^2 \in S^2 y t^1, s^2 \in \bar{S}^2$ *en la red no dirigida g.*

4. FORMULACIÓN EQUIVALENTE DEL PROBLEMA DE BIFLUJO MÁXIMO SIMÉTRICO

La resolución del problema BFMS parte de una formulación equivalente del problema 1, realizando un cambio de variables que permite separarlo en dos problemas de flujo máximo con un único bien. Este cambio de variables, introducido por Sedeño-Noda *et al.* (2001), es el siguiente:

$$\begin{aligned} x^1_{ij} + x^2_{ij} &= z^1_{ij} - z^1_{ji} \\ x^1_{ij} - x^2_{ij} &= z^2_{ij} - z^2_{ji} \end{aligned}$$

Por tanto:

$$x^1_{ij} = \frac{z^1_{ij} - z^1_{ji} + z^2_{ij} - z^2_{ji}}{2}; \quad x^2_{ij} = \frac{z^1_{ij} - z^1_{ji} - z^2_{ij} + z^2_{ji}}{2}$$

Realizando los cambios de variables, y sumando y restando las restricciones de conservación de flujo (1.1) se obtiene para $k = 1,2$:

$$\sum_{j\in \mathrm{Suc}(i)} \left(z^k_{ij} - z^k_{ji}\right) - \sum_{j\in \mathrm{Pred}(i)} \left(z^k_{ji} - z^k_{ij}\right) = \begin{cases} f^1, & \text{si } i = s^1 \\ (-1)^{k-1} f^2, & \text{si } i = s^2 \\ 0, & \text{si } i \in V - \{s^1,s^2,t^1,t^2\} \\ -f^1, & \text{si } i = t^1 \\ (-1)^k f^2, & \text{si } i = t^2 \end{cases}$$

[1]Denotaremos u^* a la capacidad del corte en la red no dirigida g

Recordemos que, por ser G una red antisimétrica, si $j \in \text{Suc}(i)$ entonces $j \notin \text{Pred}(i)$ y, si $j \in \text{Pred}(i)$, entonces $j \notin \text{Suc}(i)$. Por tanto, si queremos determinar los vértices predecesores y sucesores de uno dado sobre la red $G' = (V, A')$, encontramos que $\text{Pred}'(i) = \text{Suc}'(i) = \text{Pred}(i) \cup \text{Suc}(i)$. Por tanto, para esta última red podemos rescribir para $k = 1, 2$, las restricciones de conservación de flujo de la manera siguiente:

$$\sum_{j \in \text{Suc}'(i)} z_{ij}^k - \sum_{j \in \text{Pred}'(i)} z_{ji}^k = \begin{cases} f^1, & \text{si} \quad i = s^1 \\ (-1)^{k-1} f^2, & \text{si} \quad i = s^2 \\ 0, & \text{si} \quad i \in V - \{s^1, s^2, t^1, t^2\} \\ -f^1, & \text{si} \quad i = t^1 \\ (-1)^k f^2, & \text{si} \quad i = t^2 \end{cases}$$

Por otro lado, veamos cómo se modifican las restricciones (1.2) cuando se introducen los cambios de variables anteriores. Para ello, podemos escribir las restricciones (1.2) de la manera siguiente:

$$(1.2') \qquad \left|x_{ij}^1\right| + \left|x_{ij}^2\right| \leq u_{ij} \quad \Leftrightarrow \quad \begin{cases} -u_{ij} \leq x_{ij}^1 + x_{ij}^2 \leq u_{ij} \\ -u_{ij} \leq x_{ij}^1 - x_{ij}^2 \leq u_{ij} \end{cases}$$

Debemos notar que $-u_{ij} \leq z_{ij}^k - z_{ji}^k \leq u_{ij}$ para $k = 1, 2$ y que esas restricciones se pueden escribir como $0 \leq z_{ij}^k \leq u_{ij}$ y $0 \leq z_{ji}^k \leq u_{ij}$ (solución no solapada). Es decir, si $z_{ij}^k - z_{ji}^k = \alpha \geq 0$, hacemos $z_{ij}^k = \alpha$ y $z_{ji}^k = 0$; y, si $z_{ij}^k - z_{ji}^k = \alpha < 0$, hacemos $z_{ij}^k = 0$ y $z_{ji}^k = -\alpha$. Las variables z_{ij}^k se identifican usando las capacidades residuales r_{ij}^k tal que $z_{ij}^k = \text{máx}\left(0, u_{ij} - r_{ij}^k\right)$.

Por conveniencia en la notación, sea u'_{ij} la capacidad del arco $(i, j) \in A'$ definida por:

$$u'_{ij} = \begin{cases} u_{ij}, & \text{si} \quad (i, j) \in A \\ u_{ji}, & \text{si} \quad (j, i) \in A \end{cases}$$

En definitiva, para $k = 1, 2$, tenemos las cotas $0 \leq z_{ij}^k \leq u'_{ij}$ para todo arco $(i, j) \in A'$.

Obtenemos, por tanto, la siguiente formulación alternativa de (1):

maximizar $f^1 + f^2$
sujeto a:

$$(2.1a) \qquad \sum_{j \in \text{Suc}'(i)} z_{ij}^1 - \sum_{j \in \text{Pred}'(i)} z_{ji}^1 = \begin{cases} f^1, & \text{si} \quad i = s^1 \\ f^2, & \text{si} \quad i = s^2 \\ 0, & \text{si} \quad i \in V - \{s^1, s^2, t^1, t^2\} \\ -f^1, & \text{si} \quad i = t^1 \\ -f^2, & \text{si} \quad i = t^2 \end{cases}$$

$$\sum_{j\in \mathrm{Suc}'(i)} z_{ij}^2 - \sum_{j\in \mathrm{Pred}'(i)} z_{ji}^2 = \begin{cases} f^1, & \text{si } i = s^1 \\ -f^2, & \text{si } i = s^2 \\ 0, & \text{si } i \in V - \{s^1, s^2, t^1, t^2\} \\ -f^1, & \text{si } i = t^1 \\ f^2, & \text{si } i = t^2 \end{cases} \qquad \textbf{P2} \tag{2.1b}$$

$$0 \le z_{ij}^k \le u'_{ij}, k = 1,2 \quad \forall (i,j) \in A' \tag{2.2}$$

$$f^1 = f^2 \tag{2.3}$$

En el problema *P2* se puede observar que las restricciones pueden ser separadas en aquellas que únicamente afectan a cada una de las variables z^k con $k = 1, 2$. Sin embargo, el valor de la función objetivo no es separable. Separaremos estos dos problemas, llamando *P2a* al problema con las restricciones que únicamente contiene las variables z^1 y denominando *P2b* al relacionado con las variables z^2.

maximizar $f^1 + f^2$

sujeto a:

$$\sum_{j\in \mathrm{Suc}'(i)} z_{ij}^1 - \sum_{j\in \mathrm{Pred}'(i)} z_{ji}^1 = \begin{cases} f^1, & \text{si } i = s^1 \\ f^2, & \text{si } i = s^2 \\ 0, & \text{si } i \in V - \{s^1, s^2, t^1, t^2\} \\ -f^1, & \text{si } i = t^1 \\ -f^2, & \text{si } i = t^2 \end{cases}$$

$$0 \le z_{ij}^1 \le u'_{ij}, \forall (i,j) \in A'$$

$$f^1 = f^2$$

P2a

maximizar $f^1 + f^2$

sujeto a:

$$\sum_{j\in \mathrm{Suc}'(i)} z_{ij}^2 - \sum_{j\in \mathrm{Pred}'(i)} z_{ji}^2 = \begin{cases} f^1, & \text{si } i = s^1 \\ -f^2, & \text{si } i = s^2 \\ 0, & \text{si } i \in V - \{s^1, s^2, t^1, t^2\} \\ -f^1, & \text{si } i = t^1 \\ f^2, & \text{si } i = t^2 \end{cases}$$

$$0 \le z_{ij}^2 \le u'_{ij}, \forall (i,j) \in A'$$

$$f^1 = f^2$$

P2b

En el problema $P2a$ se maximiza un único flujo entre las fuentes s^1, s^2 y los sumideros t^1, t^2 con la restricción adicional de que $f^1 = f^2$. De manera similar, en el problema $P2b$ se maximiza un único flujo desde las fuentes s^1, t^2 a los sumideros t^1, s^2 verificando que $f^1 = f^2$. Evidentemente, si la solución óptima de estos dos problemas es la misma, hemos solucionado el problema $P2$.

Los siguientes resultados, y sus correspondientes demostraciones, aparecen en Sedeño-Noda *et al.* (2001). Aunque en dicha referencia están planteados para el problema de biflujo máximo sin considerar la restricción de simetría $f^1 = f^2$, son también válidos para el caso simétrico en la forma en que se exponen a continuación:

Lema 1. *El valor óptimo de P2 coincide con el mínimo de los valores óptimos de P2a y P2b.*

Teorema 1. *El valor óptimo de P2a(P2b) coincide con la menor de las capacidades de cualquier corte* $\left[S^1, \bar{S}^1\right]$ $\left(\left[S^2, \bar{S}^2\right]\right)$ *en la red no dirigida g tal que* $s^1, s^2 \in S^1$ $\left(s^1, t^2 \in S^2\right)$ *y* $t^1, t^2 \in \bar{S}^1$ $\left(t^1, s^2 \in \bar{S}^2\right)$.

Teorema 2. *(Biflujo-Máximo Corte-Mínimo). El máximo biflujo en una red G coincide con la menor de las capacidades de los cortes en la red no dirigida g de la forma* $\left[S^1, \bar{S}^1\right]$ *y* $\left[S^2, \bar{S}^2\right]$, *donde* $s^1, s^2 \in S^1$ *y* $s^1, t^2 \in S^2$. *Es decir,* $f^1 + f^2 = \text{mín}\left(u\left[S^1, \bar{S}^1\right], u\left[S^2, \bar{S}^2\right]\right)$ *en g.*

Supongamos que un algoritmo envía la mayor cantidad de flujo posible, que denotamos por f^{1*}, entre los nodos s^1 y t^1. Sea entonces $f^{2'}$ la mayor cantidad de flujo que se puede enviar de s^2 a t^2 cuando se envían f^{1*} unidades de flujo entre los nodos s^1 y t^1. Evidentemente, $\left(f^{1*}, f^{2'}\right)$ se corresponde con una solución óptima del problema de biflujo máximo.

Observemos que, al ser un caso particular, podemos establecer la siguiente relación entre las soluciones óptimas del problema de biflujo máximo simétrico y del problema de biflujo máximo:

Lema 2. *Sea* $\left(f^1, f^2\right)$ *el biflujo máximo simétrico, entonces se tiene que* $f^1 + f^2 \leq f^{1*} + f^{2'}$.

Demostración

Es evidente por el teorema de Biflujo-Máximo Corte-Mínimo, que cualquier biflujo $\left(f^1, f^2\right)$ tal que $f^1 = f^2$ debe satisfacer $f^1 + f^2 \leq f^{1*} + f^{2'}$.

□

Entonces, cabe plantear si a partir del biflujo máximo $\left(f^{1*}, f^{2'}\right)$ es posible obtener el biflujo máximo simétrico.

Teorema 3. *Dado el biflujo máximo* $\left(f^{1*}, f^{2'}\right)$ *se puede obtener un biflujo máximo simétrico* (f^1, f^2).

Demostración

Los posibles casos que se pueden dar son:

a) Si $f^{1*} = f^{2'}$, hemos encontrado trivialmente el biflujo máximo simétrico con $f^k = f^{1*}$ con $k = 1,2$. En este caso, se tiene que $f^1 + f^2 = f^{1*} + f^{2'}$. (f^1, f^2) es máximo por el teorema 2.

b) Si $f^{1*} < f^{2'}$, tenemos que enviar $-\left(f^{2'} - f^{1*}\right)$ unidades de flujo s^2 a t^2. De esta manera $f^k = f^{1*}$ con $k = 1,2$. Por tanto, $f^1 + f^2 = 2f^{1*} < f^{1*} + f^{2'}$. (f^1, f^2) es máximo, pues $f^1 = f^{1*}$ es el máximo valor de flujo que se puede enviar de s^1 a t^1.

c) Si $f^{1*} > f^{2'}$, tenemos que enviar $-\left(f^{1*} - f^{2'}\right)/2$ unidades de flujo s^1 a t^1. Por el teorema de Biflujo-Máximo Corte-Mínimo, a lo sumo se pueden enviar $\left(f^{1*} - f^{2'}\right)/2$ unidades de flujo s^2 a t^2. Existen dos posibles subcasos:

*c*1) Se pueden enviar las $\left(f^{1*} - f^{2'}\right)/2$ de s^2 a t^2. Con lo que se obtiene un biflujo máximo simétrico tal que $f^k = \left(f^{1*} - f^{2'}\right)/2$ con $k = 1,2$. En este caso se tiene que $f^1 + f^2 = f^{1*} + f^{2'}$. (f^1, f^2) es máximo por el teorema 2.

*c*2) Se pueden enviar ϕ unidades de s^2 a t^2 con $\phi < \left(f^{1*} - f^{2'}\right)/2$. En este caso, con el fin de obtener un biflujo máximo simétrico se han de enviar $-\left(\left(f^{1*} - f^{2'}\right)/2 - \phi\right)$ de s^1 a t^1. De esta forma se obtiene un biflujo máximo simétrico tal que $f^k = f^{2'} + \phi$ con $k = 1,2$. Además, se tiene que $f^1 + f^2 = 2f^{2'} + 2\phi < f^{1*} + f^{2'}$. Entonces, (f^1, f^2) es máximo, pues $f^2 = f^{2'} + \phi$ es la mayor cantidad de flujo que se puede enviar de s^2 a t^2 cuando $f^1 = f^{2'} + \phi$.

□

5. ALGORITMO PARA OBTENER UN BIFLUJO MÁXIMO SIMÉTRICO

En los problemas *P2a* y *P2b* se impone el envío de la mayor cantidad de flujo posible desde s^1 a t^1. Parece lógico que, en una primera instancia, se trate de obtener el flujo

máximo entre estos nodos utilizando el adecuado algoritmo de flujo máximo. Una vez resuelto este primer paso, consideraremos dos redes residuales cada una de ellas asociadas a las variables z^k con $k = 1,2$. Sean $R^1(z^1)$ y $R^2(z^2)$ las correspondientes redes residuales una vez se han enviado las f^1 unidades de flujo. Para obtener la solución del problema de biflujo, necesitamos enviar la misma cantidad de flujo f^2 desde s^2 a t^2 en $R^1(z^1)$ y desde t^2 a s^2 en $R^2(z^2)$, teniendo en cuenta los posibles casos señalados en el teorema 3. Por tanto, debemos identificar caminos incrementales en ambas redes, determinar sus correspondientes capacidades y enviar el mínimo de ambas a través de los correspondientes caminos. Esto último conlleva una actualización de las capacidades residuales de los arcos de ambos caminos.

Para identificar los respectivos caminos, proponemos el uso de dos etiquetas distancias (Goldberg (1985)). Sea una *función distancia* $d : V \longrightarrow Z^+ \cup \{0\}$ con respecto a las capacidades residuales r_{ij} de una red residual $R(x)$. Se dice que es *válida* si satisface las siguientes dos condiciones:

$$a)\ d(t) = 0$$

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

donde t es el nodo sumidero (s es el nodo fuente) del correspondiente problema de flujo máximo. Llamaremos *etiqueta distancia* de un nodo i a $d(i)$. Mediante inducción, se puede demostrar que $d(i)$ es una cota inferior de la longitud del camino mínimo desde el nodo i al nodo t en $R(x)$, donde la longitud del camino viene dada por el número de arcos que este utiliza. Si para cada nodo i, $d(i)$ coincide con la longitud del camino mínimo desde el nodo i al nodo t en $R(x)$, entonces diremos que las etiquetas distancias son *exactas*. Si $d(s) \geq n$, entonces no hay camino incremental en $R(x)$.

Un arco (i,j) en la red residual se denomina *admisible* si satisface $d(i) = d(j) + 1$. Un camino incremental en $R(x)$ que consiste de arcos admisibles es denominado *camino admisible*. Un camino admisible es un camino incremental mínimo de s a t.

Denotaremos por $d^1(i)$ a la etiqueta distancia del nodo i en $R^1(z^1)$ y por $d^2(i)$ a la etiqueta distancia del mismo nodo en $R^2(z^2)$.

Con el fin de no tener que calcular las capacidades residuales de los caminos incrementales y, posteriormente, actualizar las capacidades residuales de sus arcos, proponemos escalar las capacidades.

Para definir el escalado en las capacidades introducimos un parámetro Δ y nos referimos a la red Δ-*residual*, $R(\Delta)$, como a la red que únicamente contiene arcos cuya capacidad residual es al menos Δ. De esta manera, todos los caminos incrementales en $R(\Delta)$ tienen una capacidad residual de al menos Δ.

Como hemos introducido una escala, diremos que un arco (i,j) es Δ-*admisible* si $d(i) = d(j)+1$ y $r_{ij} \geq \Delta$. Por lo tanto, en una fase Δ se identificarán caminos incrementales Δ-*admisibles* en $R^1(z^1)$ y en $R^2(z^2)$ de manera simultánea.

Los procedimientos que proponemos trabajan en fases de escalado. Cada fase tiene un Δ fijo y, cuando finaliza una fase, comienza la siguiente con $\Delta = \Delta/2$. Al principio $\Delta = 2^{\lceil \log 2U \rceil}$, es decir, es igual al menor entero potencia de dos que es mayor o igual que $2U$. La última fase ocurre cuando $\Delta = 1$. En esta última fase $R^k(\Delta) = R^k(z^k)$ con $k = 1,2$. De esta manera, cada procedimiento realiza $\lceil \log 2U \rceil$ fases de escalado.

Debemos tener en cuenta que, cuando tengamos que devolver flujo, nos plantearemos identificar caminos de t^1 a s^1. Las anteriores ideas permiten desarrollar el siguiente algoritmo:

```
Algoritmo Biflujo_Máximo_Simétrico;
  begin
  Sea Z¹ = Z² :=0; ∀(i, j) ∈ A Sea r¹ij = r²ij := uij y r¹ji = r²ji := uij;
  Obtener el flujo máximo f¹ entre s¹ y t¹;
  f² = 0; φ := f¹; Actualizar_flujo(s², t², t², s², R¹, R², φ);
  if φ>0 then {En cualquier otro caso tenemos el biflujo máximo simétrico}
    begin
    φ := φ/2; θ = φ;
    Actualizar_flujo(t¹, s¹, t¹, s¹, R¹, R², θ); f¹ := f¹-φ; (caso c)
    Actualizar_flujo(s², t², t², s², R¹, R², φ);
    if φ>0 then; Actualizar_flujo(t¹, s¹, t¹, s¹, R¹, R², φ); f¹ := f¹-φ (caso c2)
    end;
  f² := f¹
  end.
```

```
Procedure Avanzar(i,d,R,θ,pred);
   Sea (i,j) un arco Δ-admisible;
   rij = rij - θ; rji = rji + θ;
   pred(j) := i; i := j
end
```

```
Procedure Retroceder(i,d,R,θ,pred,s);
   d(i) := min{d(j) + 1 : (i, j) ∈ A(i) y rij ≥ Δ};
   if i ≠ s then
     begin
     j := i; i := pred(i);
     rij = rij + θ; rji = rji - θ
     end;
```

```
Procedure Restablecer(i,s,R,θ,pred);
  while (i≠s) do
    begin
    j := i; i := pred(i);
    rij = rij + θ; rji = rji - θ
    end;
```

```
Procedure Actualizar_flujo(p, q, v, w, R¹ , R² , φ );
 begin
 Δ := 2^⌈log 2U⌉ ;
 while (Δ ≥ 1) and (φ>0) do
  begin
  d¹(q) := 0; Obtener d¹ mediante un recorrido en amplitud hacia atrás en R¹ a
  partir de q; Sea i¹ := p;
  d²(w) := 0; Obtener d² mediante un recorrido en amplitud hacia atrás en R² a
  partir de w; Sea i² := v;
  while (d¹(p) < n ) and (d²(v) < n ) and (φ>0) do
   begin
   θ :=min(φ , Δ );
   if (i¹ ≠ q) then
    if (i¹ tiene un arco Δ-admisible) then Avanzar(i¹ , d¹ , R¹ , θ , pred¹ )
    else Retroceder(i¹ , d¹ , R¹ , θ , pred¹ ,p);
   if (i² ≠ w) then
    if (i² tiene un arco Δ-admisible) then Avanzar(i² , d² , R² , θ , pred² )
    else Retroceder(i² , d² , R² , θ , pred² , v);
   if (i¹ =q) and (i² =w) then i¹ := p; i² := v; φ :=φ - θ
   end;
  if (i¹ ≠ p) then Restablecer(i¹ ,p, R¹ , θ , pred¹ );
  if (i² ≠ v) then Restablecer (i² , v, R² , θ , pred² );
  Δ := Δ / 2
  end
 end;
```

El algoritmo para el biflujo máximo simétrico, comienza obteniendo el flujo máximo entre s^1 y t^1 enviando $f^1 = f^{1*}$. A continuación, en la primera llamada al procedimiento *Actualizar_flujo*, se intentan enviar $f^2 = f^1$ unidades de flujo de s^2 a t^2 en $R^1(z^1)$ y de t^2 a s^2 en $R^2(z^2)$. Esta forma de proceder impide que el caso $f^1 < f^2$ pueda darse. Por ello, en la aplicación del algoritmo únicamente se pueden dar los casos *a*) y *c*), nunca el *b*) del teorema 3. En consecuencia, después de la primera llamada a *Actualizar_flujo*, o bien es $f^1 = f^2$ (con lo que se ha obtenido el biflujo máximo simétrico) o bien f^1 sigue siendo mayor que f^2. En este último caso, al realizar la segunda llamada a *Actualizar_flujo*, $\phi = (f^1 - f^2)/2$ unidades de flujo son devueltas desde t^1 a s^1 tanto en $R^1(z^1)$ como en $R^2(z^2)$. Luego, cuando se llama por tercera vez al procedimiento *Actualizar_flujo*, se intentan enviar esas ϕ unidades de flujo de s^2 a t^2 en $R^1(z^1)$ y de t^2 a s^2 en $R^2(z^2)$. Si se han enviado, el algoritmo termina y $f^1 = f^2$; en otro caso, todavía se tiene que $f^1 > f^2$ y se envían $\phi = f^1 - f^2$ unidades desde t^1 a s^1, tanto en $R^1(z^1)$ como en $R^2(z^2)$. Esta operación se realiza en la cuarta llamada a *Actualizar_flujo*, obteniendo, en su caso, el biflujo máximo simétrico.

Para identificar simultáneamente un camino Δ-*admisible* en cada red residual $R^k(\Delta)$ con $k = 1, 2$, se procede de la manera siguiente. En cada fase de escalado se calculan las etiquetas distancias exactas d^k con $k = 1, 2$. El algoritmo realiza, iterativamente, pasos *Avanzar* o *Retroceder* sobre el último nodo de cada camino admisible parcial. Si del

nodo actual i^k sale un arco admisible (i^k, j^k), entonces realiza un paso *Avanzar* y añade este arco al camino parcial actual. Cada paso *Avanzar* actualiza las capacidades residuales del arco Δ-*admisible* (i^k, j^k) por $r_{ij} = r_{ij} - \theta$ y $r_{ji} = r_{ji} + \theta$, es decir, θ unidades de flujo son enviadas a través de dicho arco. Si no se identifica un arco Δ-*admisible*, se realiza un paso *Retroceder* que incrementa la etiqueta distancia del nodo i^k, haciendo que el arco $(\text{pred}^k(i^k), i^k)$ sea no admisible y, por lo tanto, retrocediendo un arco sobre el camino parcial $\text{pred}^k(i^k)$ almacena el nodo anterior al nodo i^k en el actual camino Δ-*admisible* de $R^k(\Delta)$). Además, se deshace el envío de θ unidades de flujo anterior. Así, si (i^k, j^k) está en un camino Δ-*admisible*, ya hemos enviado las θ unidades de flujo y, si retrocedemos sobre él, deshacemos ese envío. Si en este cometido se alcanzan los respectivos nodos sumideros, el proceso se repite comenzando en las respectivas fuentes. Por la forma en que se desarrolla el algoritmo, es claro que esto únicamente sucede si se han detectado, a la vez, un camino Δ-*admisible* en $R^1(\Delta)$ y en $R^2(\Delta)$, aumentando f^2, ó disminuyendo f^1, en θ unidades de flujo. Una fase de escalado acaba cuando la etiqueta distancia de al menos un nodo fuente sea mayor o igual que n. Si esto ocurre, al menos un i^k puede ser distinto de su nodo fuente, lo que significa que se han enviado θ unidades de flujo a lo largo del camino parcial identificado para la correspondiente red residual. El procedimiento *Restablecer* se encarga de devolver esas θ unidades de flujo. Finalmente, para obtener el patrón de biflujo máximo se deshace el cambio de variable, es decir:

$$\forall (i,j) \in A, \quad z_{ij}^k - z_{ji}^k = u_{ij} - r_{ij}^k \quad \Rightarrow \quad x_{ij}^1 = \frac{z_{ij}^1 + z_{ij}^2}{2} - u_{ij} x_{ij}^2 = \frac{z_{ij}^1 - z_{ij}^2}{2}$$

5.1. Ejemplo

A continuación aplicaremos el algoritmo al ejemplo de la figura 1. El proceso para obtener el patrón de biflujo máximo simétrico dado en la figura 2, se desarrolla en la figura 3.

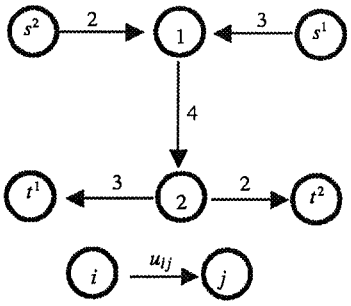


Figura 1

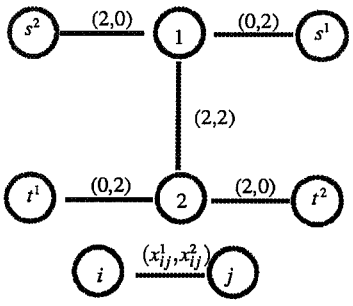


Figura 2

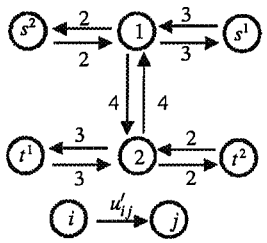


Fig. 3*a*

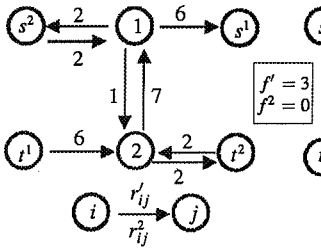


Fig. 3*b*

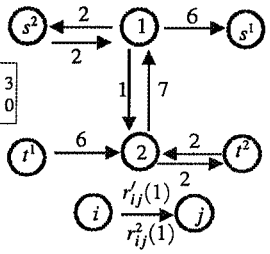


Fig. 3*c*

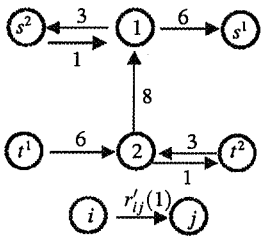


Fig. 3*d*

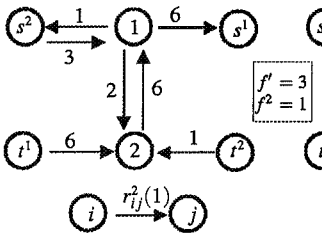


Fig. 3*e*

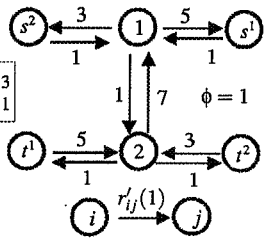


Fig. 3*f*

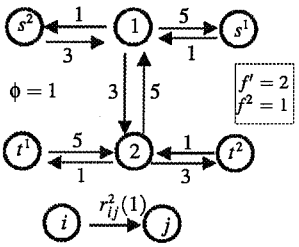


Fig. 3*g*

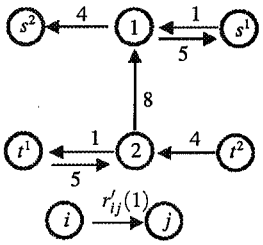


Fig. 3*h*

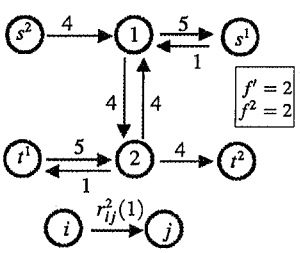


Fig. 3*i*

La figura 3*a*, representa la red G' con las capacidades u'. La figura 3*b*, muestra la red residual $R^1(z^1)$ que coincide con $R^2(z^2)$ cuando se han enviado $f^1 = 3$ unidades de flujo desde s^1 a $t^1 (f^2 = 0)$. Se puede observar en esta figura que únicamente hay un camino de s^2 a t^2 en $R^1(z^1)$ de capacidad 1 y otro de t^2 a s^2 en $R^2(z^2)$ de capacidad 2. Por ello, al llamar por primera vez a *Actualizar_flujo*, se identifican ambos caminos cuando el procedimiento se encuentre en la fase $\Delta = 1$ (figura 3*c*). Tras enviar 1 unidad de flujo a través de estos caminos en sus respectivas redes incrementales (figuras 3*d* y 3*e*), se puede observar que en $R^1(1)$ no hay camino incremental (aunque si lo hay en $R^2(1)$). En este punto tenemos que $f^1 = 3$ y $f^2 = 1$. Por lo tanto, $f^1 > f^2$, lo que implica que $\phi = (f^1 - f^2)/2 = 1$. Entonces tenemos que la segunda llamada al procedimiento

Actualizar_flujo envía 1 unidad de flujo de t^1 a s^1, tanto en $R^1(z^1)$ y en $R^2(z^2)$. El resultado de esta operación se muestra en las figuras $3f$ y $3g$ respectivamente. Ahora tenemos que $f^1 = 2$ y $f^2 = 1$, pero lo importante es que podemos enviar 1 unidad de flujo de s^2 a t^2 en $R^1(z^1)$ y de t^2 a s^2 en $R^2(z^2)$. Esta operación se realiza cuando se llama por tercera vez al procedimiento *Actualizar_flujo*. El resultado de esta operación se muestra en las figuras $3h$ y $3i$, de donde se tiene que $f^1 = 2$ y $f^2 = 2$. En este punto no es necesario llamar de nuevo al procedimiento *Actualizar_flujo*. Finalmente, el patrón de biflujo máximo simétrico se obtiene deshaciendo los cambios de variables (Figura 2).

5.2. Complejidad del algoritmo

En esta sección demostraremos que la complejidad del algoritmo en el caso peor es $O(nm \log U)$. El algoritmo se inicia con la obtención del flujo máximo entre s^1 y t^1. Por tanto, se debe elegir un método de flujo máximo con la complejidad computacional adecuada para realizar este cometido.

Por otra parte, el algoritmo anterior ejecuta entre una y cuatro veces el procedimiento *Actualizar_flujo*. Ambos procedimientos requieren el mismo esfuerzo computacional. Cada uno de ellos realiza $\lceil \log 2U \rceil$ fases de escalado. Cada fase de escalado Δ es llevada a cabo hasta que alguna de las etiquetas distancia de las correspondientes fuentes en $R^1(\Delta)$ ó en $R^2(\Delta)$ es mayor o igual que n.

Lema 3. *Cada fase de escalado de Actualizar_flujo requiere un esfuerzo computacional de O(nm).*

Demostración

Supongamos que nos fijamos en la etiqueta distancia $d^1(p)$ en $R^1(\Delta)$ y que ésta es la que determina la condición de parada de cada fase de escalado cuando $d^1(p) \geq n$. Así, el procedimiento actualiza la etiqueta distancia $d^1(i)$ en $R^1(\Delta)$ de un nodo i a lo sumo n veces ya que, si un nodo tiene una distancia mayor o igual que n, este nodo no es alcanzable mediante un camino elemental desde la fuente. Además, en el peor de los casos, la etiqueta distancia de un nodo puede incrementarse cada vez en una unidad. Por tanto, el número de operaciones *Retroceder* está acotado por $O(n^2)$, lo que implica que el esfuerzo computacional necesario para encontrar arcos admisibles y actualizar las etiquetas de los vértices es $O(mn)$.

Veamos ahora cuál es el número de envíos de flujo. Para ello diremos que un envío es Δ-*saturante* a través de (i, j) si, después de realizarlo, la capacidad residual del arco (i, j) es estrictamente menor que Δ. Un arco puede ser Δ-*saturado* a lo sumo $n/2$ veces. Esto es así debido a que entre dos envíos Δ-*saturantes* consecutivos, a través de (i, j), las

etiquetas distancias de los nodos i y j deben haber incrementado en al menos dos unidades. Así, sumando para todos los arcos, se tiene que el número de envíos Δ-*saturantes* es $O(nm)$. Esto implica que el número de envíos de flujo es a lo sumo $O(nm)$. Como el esfuerzo computacional de cada envío de flujo es $O(1)$, el esfuerzo total vuelve a ser $O(nm)$. El número de llamadas a *Avanzar* está acotado por el número de envíos de flujo más el número de operaciones *Retroceder*, es decir, $O(nm+n^2)$. Finalmente, la operación *Restablecer* es llamada a lo sumo una vez, requiriendo un esfuerzo de $O(n)$. Por lo tanto, la complejidad de una fase, considerando que la etiqueta distancia $d^1(p)$ en $R^1(\Delta)$ determina la condición de parada, es $O(nm)$.

La misma complejidad en el peor caso se obtendría si se considerara que es la etiqueta distancia $d^2(v)$ en $R^2(\Delta)$ la que determina la condición de parada. Como, en cada fase, alguna de las dos posibilidades debe ocurrir, la complejidad de cada fase coincide con el máximo de las complejidades de ambas. Por lo tanto, la complejidad es $O(nm)$.

□

Teorema 4. *El algoritmo tiene una complejidad* $O(nm\log U)$.

Demostración

Por el lema 3, cada fase de escalado emplea un tiempo de $O(nm)$, como el número de fases de escalado es $\lceil \log 2U \rceil$, obtenemos que cada procedimiento se ejecuta en un tiempo $O(nm\log U)$. Como, a lo sumo, se ejecutan cuatro veces el procedimiento *Actualizar_flujo*, el algoritmo requiere un tiempo $O(nm\log U)$.

□

6. CONCLUSIONES

En este trabajo, estudiamos el problema de Biflujo Máximo Simétrico sobre redes no dirigidas. Este problema no ha sido considerado previamente en la literatura que hemos consultado. Para su resolución, introducimos una nueva formulación, resultante de la consideración de cambios de variables, que permiten resolverlo utilizando las herramientas clásicas desarrolladas para problemas de flujos de un solo tipo. Después de exponer los conceptos y propiedades necesarios, introducimos un algoritmo que resuelve eficientemente el problema con una complejidad computacional teórica de $O(nm\log U)$.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado a través del proyecto 180221024 de la Universidad de La Laguna.

7. REFERENCIAS

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

Ford, L. R. and Fulkerson, D. R. (1956). «Maximal Flow through a Network», *Canadian Journal of Mathematics*, 8, 399-404.

Goldberg, A. V. (1985). «A New Max-Flow Algorithm», *Technical Report MIT/LCS/TM-291*, Laboratory for Computer Science, MIT, Cambridge, MA.

Hu, T. C. (1963). «Multi-Commodity Network Flows», *Operations Research*, 11, 344-360.

Itai, A. (1978). «Two-Commodity flow», *J. Assoc. Comput. Mach.*, 25 (4), 596-611.

Rothschild, B. and Whinston, A. (1966). «On Two Commodity Network Flows», *Operations Research*, 14, 377-387.

Sakarovitch, M. (1973). «Two Commodity Network Flows and Linear Programming», *Mathematical Programming*, 4, 1-20.

Sedeño-Noda, A., González-Martín, C. and Alonso-Rodríguez, S. (2001). «The Maximum Biflow Problem: A study of the undirected case», en proceso de referee en *Naval Research Logistics*.

Seymour, P. D. (1978). «A Two Commodity cut Theorem», *Discrete Mathematics*, 23, 177-181.

ENGLISH SUMMARY

AN ALGORITHM FOR THE UNDIRECTED SYMMETRIC MAXIMUM BIFLOW PROBLEM

A. SEDEÑO-NODA
C. GONZÁLEZ-MARTÍN
Universidad de La Laguna*

In this paper, an $O(nm \log U)$ algorithm to solve the maximum undirected symmetric biflow problem is proposed. We introduce changes of variables in the initial mathematical formulation of the problem, which let us split this problem in two maximum flow problems. In this way, we obtain an easy algorithm that uses the computational tools, which are associated with the resolution of the maximum flow problem.

Keywords: Symmetric Maximum Biflow problem, Capacity Scaling Algorithm, Undirected Networks

AMS Classification (MSC 2000): 90C27

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

- Received August 2001.
- Accepted Mai 2002.

Many realistic problems are solved using network flows modelling (see, for example, Ahuja *et al.* (1993)). Among them, one of the network flows most intensely studied is the maximum flow problem. The original version, with a single commodity, is studied and solved in Ford and Fulkerson (1956), through the well-known *Augmenting Path* algorithm, based on the max-flow minimum-cut theorem. From this starting point, several works have improved the computational complexity given.

When different kinds of commodities over a same network are considered, multi-commodity flow problems arise. In general, for three or more kinds of flows, the analogous max-flow minimum-cut theorem is not verified. However, for the two-commodity problem in undirected case, Hu (1963) proves the max-biflow min-cut theorem. An analogous result is given by Seymour (1978).

In this work, we consider the symmetric maximum biflow (SMBF) problem on undirected network. This problem is a particular case of the maximum biflow (MBF) problem. The formulation of the MBF problem uses a directed network where the quantities of flow can be negative. This formulation is common with Hu (1963), Rothschild and Whinston (1966), Sakarovitch (1973), Itai (1978), Seymour (1978) and Sedeño-Noda *et al.* (2001).

The MBF problem is stayed in the next way: Given a directed network $G = (V,A)$, let V be the set of n nodes and A be the set of m arcs. We distinguish four special nodes in G: the source nodes s^1, s^2 and the sink nodes t^1, t^2. We assume that the network is antisymmetric, that is, if $(i,j) \in A \Rightarrow (j,i) \notin A$. Given any node i, we define the sets $\text{Pred}(i) = \{j \in V/(j,i) \in A\}$ and $\text{Suc}(i) = \{j \in V/(i,j) \in A\}$. Let $u_{ij} \geq 0$ be the capacity associated with the arc $(i,j) \in A$ and $U = \max\{u_{ij}/(i,j) \in A$. For each node i, we also define the *arc adjacency list* $\text{A}(i) = \{(i,j) \in A : j \in \text{Suc}(i)\}$.

A *biflow* in G is a pair $x = (x^1, x^2)$ that satisfies:

$$\sum_{j \in \text{Suc}(i)} x_{ij}^k - \sum_{j \in \text{Pred}(i)} x_{ji}^k = \begin{cases} f^k, & \text{if } i = s^k \\ 0, & \text{if } i \in V - \{s^k, t^k\} \\ -f^k, & \text{if } i = t^k \end{cases} \qquad k = 1,2 \tag{1.1}$$

$$\left|x_{ij}^1\right| + \left|x_{ij}^2\right| \leq u_{ij}, \qquad \forall (i,j) \in A \tag{1.2}$$

for some $f = (f^1, f^2)$ with $f^k \geq 0$, for $k = 1,2$. The MBF problem consists in finding a biflow x such that the summation $f^1 + f^2$ is maximal. In other words, in MBF we will try to send the maximum amount of flow from the source s^1 to the sink t^1, and from the source s^2 to the sink t^2, verifying the constraints (1.1) and (1.2). Note that constraints (1.2) allow the values of the biflow to be negative. This means that if the value of the kth flow x_{ij}^k is negative for the arc $(i,j) \in A$, we will interpret that the flow travels in

the opposite direction. This corresponds with the undirected case of the MBF problem. The SMBF problem equals the MBF problem with the additional constraint $f^1 = f^2$.

In these conditions, we introduce new changes of variables that let us split the problem in two maximum flow problems with the additional constraint $f^1 = f^2$. Furthermore, this new development leads us to build an efficient algorithm to solve the problem. The changes of variables used in this paper, allow all lower bounds to be zero and the resolution of the problems using classic tools. Then, the implementation of the proposed algorithm results simple and with a complexity of $O(nm \log U)$.

QÜESTIIÓ, vol. 26, 3, p. 537-559, 2002

PROBLEMA DE CONTRATACIÓN DE CARRETILLEROS PARA UN ALMACÉN DE PRODUCTOS MANUFACTURADOS

C. R. DELGADO
S. CASADO
J. F. ALEGRE
Universidad de Burgos*

En este trabajo se analiza un problema planteado recientemente a sus autores por una empresa fabricante de componentes de automóviles. Dicha empresa almacena sus productos manufacturados hasta que los clientes (compradores) pasan a recogerlos. Los clientes solicitan sus productos con una frecuencia conocida. Se trata de determinar en función de dichas frecuencias, en qué fechas y a que horas o slots, han de pasar los clientes a recoger sus pedidos. Fijado el horizonte temporal objeto de estudio, el objetivo para la empresa es minimizar en ese conjunto de días, el número de carretilleros necesarios para cargar los pedidos en los camiones de los clientes. El número de carretilleros diario viene determinado por el slot más ocupado. En este problema se han de tomar decisiones a dos niveles: elaboración del calendario de entrega para cada uno de los pedidos, y asignación diaria de pedidos a slots. No se ha encontrado en la literatura ningún trabajo que estudie o pueda ajustarse a este problema. Se diseñan y analizan 3 sencillos Metaheurísticos: uno sigue un proceso de Búsqueda Tabú *básico, otro es un procedimiento de* Búsqueda en Entorno Variable *y el tercero es un* Algoritmo Evolutivo. *Se realizan pruebas con instancias ficticias y por último se resuelven las instancias reales planteadas a los autores de este trabajo.*

Labor requeriments for vehicle loading problem in a store of manufactured products

Palabras clave: Logística, Multiprocessor Scheduling Problem (MSP), Búsqueda Local, Búsqueda Tabú, Búsqueda en Entorno Variable, Algoritmos Evolutivos

Clasificación AMS (MSC 2000): 90B06

* Facultad C. Económicas y Empresariales. Pza. Infanta Elena s/n. 09001 Burgos.

- Recibido en octubre de 2001.
- Aceptado en junio de 2002.

1. INTRODUCCIÓN

Se estudia el caso de una empresa fabricante de componentes de automóviles que almacena sus productos manufacturados a la espera de que los clientes, o compradores, vayan a recogerlos. Los clientes acuden a recoger determinadas cantidades de productos $q(i)$ con una frecuencia conocida. La empresa contrata diariamente los carretilleros necesarios para cargar los productos en los camiones de los clientes, teniendo en cuenta la necesidad del slot (hora) más ocupado.

Se trata de asignar a cada pedido i un conjunto de fechas o *calendario* C_r^i –de entre los que tiene disponibles T_i– y en cada día del calendario un slot u hora de entrega; el objetivo es minimizar el número de carretilleros contratados en un determinado horizonte temporal.

Si un pedido tiene una frecuencia de *1 vez cada 4 días* los posibles conjuntos de fechas de recogida de ese pedido, o calendarios, son los siguientes:

$$\Gamma_i = \{\{1,5,9,13,17,\ldots\}, \quad \{2,6,10,14,18,\ldots\}, \quad \{3,7,11,15,19,\ldots\}, \\ \{4,8,12,16,20,\ldots\}\}$$

(con algunas modificaciones por los días festivos). El horizonte temporal para la planificación es habitualmente de un mes, pero en algunos casos podría ser mayor (2 o 3 meses).

Calendars	Days	1	2	3	4	5	6	7	8	9	10	11	12	13	....
1	{1, 5, 9, 13, .}	▒				▒				▒				▒	....
2	{2, 6, 10, 14, .}		▒				▒				▒				....
3	{3, 7, 11, 15, .}			▒				▒				▒			....
4	{4, 8, 12, 16, .}				▒				▒				▒		....

Figura 1. Posibles calendarios para un producto con una frecuencia de 1 pedido cada 4 días.

El problema implica tomar decisiones a 2 niveles: determinar las fechas para preparar los pedidos, y determinar para cada día las horas de entrega que minimizan el número de carretilleros necesarios. Dicho número, para cada slot, se considera proporcional al número de palets que hay que cargar. Por tanto, sin pérdida de generalidad, se puede identificar número de carretilleros necesarios en cada slot con número total de palets asignados a ese slot. Los contratos de carretilleros son diarios; el número de carretilleros que se contrata diariamente viene determinado por el slot más ocupado. Por tanto, la asignación diaria de slots a pedidos se ajusta al conocido *Multiprocessor Scheduling Problem (MSP)* o al *k-Partition problem.*

Se trata de un problema NP-hard (su problema de decisión asociado es NP-completo). Piénsese que su segundo nivel, el MSP, es un problema NP-completo, como se puede ver por ejemplo en Garey y Johnson (1979).

El MSP es un problema muy estudiado en la literatura y muy estrechamente relacionado con el *Bin Packing*. Algoritmos constructivos para este problema son los conocidos LPT (*largest processing time first*) de Graham (1969) y MULTIFIT de Coffman *et al.* (1978). En este trabajo se utilizará el LPT. Básicamente consiste en ordenar los pedidos de manera decreciente respecto al número de palets que los constituyen, para posteriormente iniciar un proceso de inserción en el slot menos ocupado. Métodos de mejora son los conocidos intercambios de Finn and Horowitz (1979), o el de Langston (1982). Más recientemente en Franca *et al.* (1994) se propone un eficaz heurístico en 3 fases que combina un método constructivo, e intercambios de tipo 0-1 («trasferencia») y 1-1, («intercambio»). Por otra parte en Hübscher and Glover (1994) y Thesen (1998) se proponen estrategias basadas en Búsqueda Tabú. Algoritmos exactos se pueden encontrar en los trabajos de Sahni (1976), Blazewicz (1987), Dell'Amico and Martello (1990) o Harche F. and Seshadri S. (1995). Asimismo, en Babel *et al.* (1998) se dan herramientas para su resolución tanto exacta como aproximada. Un reciente «survey» sobre este problema se puede encontrar en Fujita y Yamashita (2000). La dificultad del MSP, en general, aumenta cuando los pedidos son de diferente tamaño, como en el modelo que tratamos.

Una posible formulación del problema que nos ocupa en este trabajo es la siguiente:

$$min \sum_{d=1}^{ndias} \left\{ \max_{j=1\ldots nslots} \sum_{i=1}^{np} q(i) X_{idj} \right\}$$

sujeto a:

$$\sum_{r=1}^{T_i} Y_{ir} = 1 \qquad i = 1, \ldots, np$$

$$\sum_{j=1}^{nslots_j} X_{idj} = Y_{ir} \qquad d \in C_r^i;\ r = 1, \ldots, T_i;\ i = 1, \ldots, np$$

donde:

- *ndias*: número de días;
- *n_slots*: número de slots;
- *np*: número de pedidos;
- $q(i)$: palets del cliente i;
- $\Gamma_i = \left\{ C_1^i, C_2^i, \ldots, C_{T_i}^i \right\}$: conjunto de calendarios del pedido i;

- C_r^i: calendario r del pedido i;
- T_i: número de calendarios del pedido i;

con las siguientes variables de decisión:

- Y_{ir}: variable binaria igual a 1 si el pedido i tiene asignado el calendario r y 0 en caso contrario;
- X_{idj}: variable binaria igual a 1 si el pedido i se ha asignado el día d al slot j y 0 en caso contrario.

En principio se trata de un problema particular planteado por una empresa concreta. Sin embargo, entendemos que es un interesante problema, y que problemas muy similares se plantean todos los días a los responsables de diferentes empresas logísticas. No hemos encontrado en la literatura referencias de trabajos que aborden este problema.

Nuestro objetivo será diseñar una estrategia sencilla y adecuada, que dé buenos resultados con datos reales. Así, para la resolución del problema se utilizan tres algoritmos metaheurísticos: uno sigue un procedimiento de *Búsqueda Tabú Básico*, el segundo sigue una estrategia de *Búsqueda en Entorno Variable* (VNS) y el tercero es un *Algoritmo Evolutivo*. Compararemos los resultados obtenidos con cada estrategia, para obtener conclusiones acerca de la eficiencia de cada uno de los tres algoritmos.

El trabajo se estructura como sigue: en el siguiente apartado se propone una forma de representación de soluciones y una definición de movimientos simples entre soluciones; en el tercero se muestra una estrategia basada en un procedimiento muy básico de Búsqueda Tabú, en el cuarto el diseño de un algoritmo basado en Búsqueda en Entorno Variable y en el quinto se describe una estrategia basada en un Algoritmo Evolutivo; en el apartado sexto se muestran los resultados de experimentos computacionales realizados con una serie de instancias ficticias (adaptadas de instancias de la literatura para otros modelos) y otra serie de instancias con datos reales. Finalmente, el último apartado se dedica a las conclusiones.

2. REPRESENTACIÓN DE LAS SOLUCIONES Y DEFINICIÓN DE MOVIMIENTOS

2.1. Representación de soluciones

Las soluciones se representan en dos niveles: en el primero se consideran los calendarios que se asignan a cada pedido, y en el segundo la asignación diaria de pedidos a slots. La asignación de calendarios se representa mediante un vector o *np-upla S*, en el que se indica que calendario corresponde a cada pedido. Por ejemplo, si al pedido p_i le

corresponde el Calendario $C^i_{j_i}$, el correspondiente elemento en el vector es j_i. Por tanto la *np-upla* será de la forma siguiente:

$$S = (j_1, j_2, j_3, \ldots, j_{np}) \,/\quad \forall i \in \{1,2,\ldots,np\}, \quad Y_{ij_i} = 1$$

$$\downarrow \quad \downarrow \quad \downarrow \quad \quad \downarrow$$

$p_1 \;\; p_2 \;\; p_3 \;\; \ldots \;\; p_{np}$ Corresponde al pedido

A la componente i-ésima de la solución S, la designaremos $S(i)$. Así en S, $S(i) = j_i$.

Con **R** se representa, en el conjunto de días, la distribución de pedidos por slots:

$$\mathbf{R} = \{R(d)/d = 1,\ldots,ndias\},$$

siendo $R(d)$ un vector, desde 1 hasta *np*, que indica el slot donde está cada pedido.

Slots	Pedidos $(q(i))$
1	1 -11-5
2	22
3	13-6
4	14-16

Figura 2. Distribución de pedidos para un día «d».

Supongamos un día con cuatro slots u horas de carga: de 9 a 10 (1), de 10 a 11 (2), de 11 a 12 (3) y de 12 a 13 (4); supongamos la distribución de pedidos $R(d)$, propuesta en la figura 2, para un día cualquiera, donde para simplificar la ilustración, se indica de cada pedido i directamente la cantidad demandada, es decir $q(i)$. Para cada slot, la suma de palets de cada pedido, indica el número de carretilleros necesarios o nivel de ocupación. El número de carretilleros a contratar cada día viene determinado por el slot más ocupado. En el ejemplo propuesto viene determinado por el slot 4, con dos pedidos de 30 palets en total (es decir 30 carretilleros). El número total de carretilleros a contratar en el conjunto de días se denota como $f(\mathbf{R})$.

2.2. Definición de vecindarios o entornos

Dada una *np-upla S* definimos su vecindario N(S) de la siguiente forma:

$$S' \in N(S) \;\Leftrightarrow\; \exists i' \in \{1,2,\ldots,np\}/S'(i') \neq S(i') \quad \text{y} \quad \forall i \neq i' \quad S'(i) = S(i)$$

es decir, soluciones generadas cuando un pedido (y sólo uno) cambia su *calendario*.

2.3. Movimiento vecinal

El movimiento a una solución vecina puede provocar una disminución parcial Δ, y un aumento parcial Δ' en la función objetivo. La disminución se produce cuando al cambiar de calendario a p_i, éste se retira, en alguno de los días, del slot con mayor ocupación; en este caso, debemos calcular y actualizar, en los días que esto suceda, el valor del slot más ocupado. La inserción del pedido p_i en los nuevos días de entrega, se realiza en el slot con menor ocupación, de forma que la función objetivo aumente lo menos posible. Dada una solución $S = (j_1, j_2, j_3, \ldots, j_i, \ldots, j_{np})$, para cada pedido $i \in \{1,2,\ldots,np\}$, se denota g_{ij}, a la disminución del valor de la función objetivo, producida al cambiar a p_i, el calendario C_{j_i} por el C_j. Los valores de g_{ij} se calculan como la diferencia $g_{ij} = \Delta' - \Delta$.

Dada la solución (S, $\boldsymbol{R}$), se define:

$$g_{i*j*} = \min\{g_{ij}/i \in \{1,2,\ldots,np\}, \quad j \in \{1,2,\ldots,T_i\}, \quad j \neq j_i\};$$

de forma que $i*$ y $j*$ determinan el movimiento a la mejor solución vecina de S.

Una vez realizado el movimiento se mejoran los días modificados utilizando un procedimiento de Búsqueda Local. Se proponen dos movimientos para dicho procedimiento —ver figura 3—, usados frecuentemente en múltiples trabajos para problemas combinatorios como por ejemplo, Barnes y Laguna (1993), y para el MSP en Hubscher y Glover (1994), y Thesen (1998):

- Movimiento 0-1: Cambio de un pedido, a un slot diferente al suyo.
- Movimiento 1-1: Intercambio de pedidos, pertenecientes a distintos slots.

Los movimientos 0-1 son fácilmente programables a partir de los movimientos 1-1. Para ello es necesario considerar en cada slot un pedido «dummy» de valor 0.

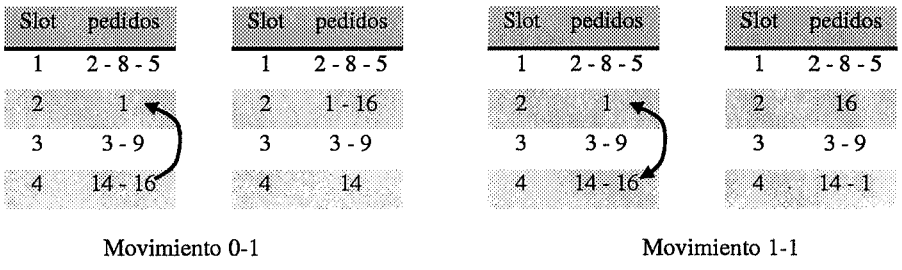


Figura 3. Movimientos vecinales para el MSP.

Para realizar los movimientos Hubscher y Glover (1994) consideran intercambios entre el slot más ocupado y los que tienen una ocupación inferior a la ocupación media; Thesen (1998) considera intercambios entre el slot más ocupado y el menos ocupado. En nuestro caso, ya que tanto el número de pedidos como el de slots no es muy alto, consideramos intercambios entre el slot más ocupado y el resto.

Sea el pedido i perteneciente al slot j, con una ocupación de P_j palets, y el pedido k perteneciente al slot l con una ocupación de P_l palets, la bondad del intercambio entre los pedidos i y k viene dada por:

$$max\{P_j, P_l\} - max\{P_j - q(i) + q(k), P_l - q(k) + q(i)\};$$

nótese que este criterio de bondad es similar a los usados en los trabajos anteriormente señalados. En cada paso se ejecuta el mejor movimiento.

Para acelerar la ejecución de este procedimiento de Búsqueda Local se adapta la estrategia de Búsqueda Local Rápida propuesta por Bentley (1992). Esta estrategia consiste básicamente en lo siguiente: se divide el vecindario, o conjunto de movimientos posibles en subvecindarios; en nuestro caso cada subvecindario, o subconjunto de movimientos, viene definido por un par de slots, es decir, está compuesto por todos los intercambios entre pedidos de ese par de slots. A cada subvecindario se le asocia una variable binaria que indica si está activo o no, de forma que en cada iteración sólo se exploran los subvecindarios activos. Inicialmente todos los subvecindarios están activos. Si tras una exploración se comprueba que en un subvecindario, ningún intercambio de pedidos entre los dos slots que lo componen produce mejora, dicho subvecindario pasa a ser inactivo. Por otro lado, tras la ejecución de un movimiento, todos los subvecindarios alterados, es decir, subvecindarios donde alguno de los slots que los definen se hayan modificado, pasan a ser activos siendo necesario explorarlos de nuevo.

Se han realizado pruebas con problemas generados aleatoriamente cuyo número de pedidos (np) asciende a 3.000 y con un número de slots que varía desde 20 hasta 200. Los resultados obtenidos muestran que la Búsqueda Local Rápida reduce el tiempo de computación medio en el conjunto de pruebas en un 12-13 % aproximadamente sin que varíe la solución obtenida. Dicha reducción es mayor cuanto mayor es el número de slots considerado.

2.4. Propuesta de un procedimiento de Búsqueda Local

Teniendo en cuenta lo expuesto hasta ahora, se describe un sencillo algoritmo de búsqueda local para el problema que se analiza en este trabajo:

Procedimiento Búsqueda_Local_Log(var S, **R**)

Repetir

Determinar $g_{i*j*} = min\{g_{ij}/i \in \{1,2,\ldots,np\}, \quad j \in \{1,2,\ldots,T_i\}, j \neq j_i\}$;

Si $g_{i*j*} < 0$ *entonces:*

– *Hacer* $S(i*) = j*(oj_{i*} = j*)$

– *Modificar y mejorar los elementos de* **R** *según se ha explicado en 2.3*

Hasta $g_{i*j*} \geq 0$

3. DISEÑO DE UN ALGORITMO DE BÚSQUEDA TABÚ

En general, en los problemas de optimización combinatoria, la utilización de procedimientos de Búsqueda Local suele llevar a óptimos de mala calidad, debido a la dependencia que existe de la solución inicial. En el problema que estamos tratando se ha observado también esta situación.

El procedimiento de Búsqueda Local puede ser modificado de forma muy sencilla para dar lugar a un procedimiento basado en *Búsqueda Tabú* (muy básico) que puede ayudarnos a escapar de estos pobres óptimos locales, permitiéndonos alcanzar mejores soluciones. La *Búsqueda Tabú* es una estrategia dada a conocer en los trabajos de Glover (1989) y (1990), 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)). Una vez que se llega a un óptimo local se permiten movimientos *hacia arriba* o que empeoran la solución, para salir de dicho óptimo. Simultáneamente, los últimos movimientos realizados se califican como *tabús* durante las siguientes iteraciones; de esta forma se evita volver a soluciones anteriores y que el algoritmo cicle. Recientes y amplios tutoriales sobre Búsqueda Tabú que incluyen todo tipo de aplicaciones pueden encontrarse en Glover y Laguna (1997) y (2001).

En nuestro caso se define:

Matriz_tabú (i, j) = número de la iteración en la que el pedido p_i dejó de tener asignado el calendario C_j^i.

Se denota por (S, **R**) a la solución actual y por (S*, **R***) a la mejor solución encontrada; el valor de la función objetivo para dichas soluciones será, respectivamente, f y $f*$; el algoritmo de Búsqueda Tabú se puede escribir de la forma siguiente:

Algoritmo Búsqueda_Tabú_Log(var S, **R**; S*, **R***)

Hacer Matriz_tabú $(i,j) = -Tabu\_Tenure, i \in \{1,2,\ldots,np\}, \quad j \in \{1,2,\ldots,T_i\}$;
Hacer niter = 0, iter_mejor = 0, S = S, y* $\boldsymbol{R^* = R}$
Repetir
  niter = niter + 1
  Determinar
    $g_{i*j*} = min\{g_{ij}/i \in \{1,2,\ldots,np\}, j \in \{1,2,\ldots,T_i\}; j \neq j_i$ *verificando que*
      [niter >
    Matriz_tabú (i,j) *+ Tabu_Tenure o* $f + g_{ij} < f*$ *(«criterio de aspiración»)]}*
  Hacer Matriz_tabú $(i*, j_{i*}) = niter$ *y* $S(i*) = j*$ $(o j_{i*} = j*)$
  Modificar y Mejorar los elementos de **R** *según se explicó en 2.3*
  Actualizar f
  Si $f < f*$ *entonces hacer:* $S* = S$, $\boldsymbol{R^* = R}$, $f* = f$ *e iter_mejor = niter*
Hasta (niter − iter_mejor > max_iter) u otro criterio de parada

El parámetro *Tabu_Tenure* indica el número de iteraciones en las que no se permitirá a un pedido volver a tener asignado un calendario que acaba de dejar. Esta restricción o *condición tabú* puede ser superada por el *criterio de aspiración*: una solución o movimiento tabú puede ser aceptado, si da lugar a una solución mejor que la mejor encontrada hasta el momento. El parámetro *max_iter* indica el máximo número de iteraciones sin mejoras.

4. DISEÑO DE UN ALGORITMO BASADO EN BÚSQUEDA EN ENTORNO VARIABLE

Búsqueda en Entorno Variable (*Variable Neighborhood Search* o VNS) es un reciente metaheurístico diseñado para resolver problemas combinatorios, propuesto y descrito en los trabajos de Mladenovic (1995), Mladenovic y Hansen (1997), y Hansen y Mladenovic (1998). La idea básica es combinar la aplicación de un procedimiento de búsqueda local con un sistemático cambio del entorno de búsqueda. El algoritmo aplica la búsqueda local en alguna solución del entorno de la mejor solución obtenida hasta el momento (*solución actual*). Si no se consigue mejorar esta *solución actual* se considera un entorno mayor. Cuando se obtiene una solución mejor se reinicia el proceso. Recientes tutoriales con ejemplos se encuentran en Hansen y Mladenovic (2000) y (2001).

En nuestro caso para obtener una solución en un entorno de la solución actual se perturba esta de la siguiente manera. Sea (S, **R**) una determinada solución, y $k \in$ IN se define:

Procedimiento Perturbar (k,S)

Seleccionar k pedidos aleatoriamente, $p_{i1}, p_{i2}, \ldots, p_{ik}$, *con* $T_{il} \geq 2, l = 1 \ldots k$,
Para $l = 1, \ldots, k$:
Reasignar a p_{il} *un calendario* $C_{l'}^{il} \in \Gamma_i / l' \neq S(il)$ *escogido aleatoriamente*

Sea $(S, \boldsymbol{R})$ una solución con valor f, el algoritmo VNS que se propone en este trabajo queda de la forma siguiente:

Algoritmo VNS (*kmax, S,* $\boldsymbol{R}$)

Hacer niter $= 0$
Repetir
Hacer: niter = niter + 1 y k = 0
Repetir
Hacer k = k+1;
$S' = S$
Ejecutar Perturbar (k,S')
Construir R' : $\forall d = 1, \ldots,$ *ndias formar* $\boldsymbol{R}(d)$ *con un algoritmo constructivo para el MSP (LTP de Graham* et al. *(1969)) y mejorar con intercambios 0-1,1-1*
Determinar f' *el valor de* $(S', \boldsymbol{R}')$
Ejecutar Búsqueda_Local_Log $(S', \boldsymbol{R}')$
Hasta $(f' < f)$ *o (k = kmax)*
Si $f' < f$ *hacer:* $S = S', R = R'$ *y* $f = f'$ *e iter_mejor = niter*
Hasta (niter-iter_mejor = maxiter) u otro criterio de parada

El parámetro *kmax* indica el máximo número de puntos que se van a cambiar y *max_iter* el máximo número de iteraciones sin mejora. El uso de un procedimiento de perturbación aleatorio tiene como objeto evitar ciclos.

5. PROPUESTA DE UN ALGORITMO EVOLUTIVO

Los Algoritmos Evolutivos se inspiran en los principios básicos de la evolución de los seres vivos y los modifican para obtener sistemas eficientes para la resolución de diferentes problemas. Un Algoritmo Evolutivo es un proceso estocástico e iterativo que opera sobre un conjunto P de *individuos* que constituyen lo que se denomina *población.* La población inicial es generada aleatoriamente o con la ayuda de algún heurístico de

construcción. Cada individuo es evaluado a través de una función de bondad (fitness). Estas evaluaciones se usan para predisponer la selección de cromosomas de forma que los superiores (aquellos con mayor evaluación) se reproduzcan más a menudo que los inferiores.

El algoritmo se estructura en tres fases principales que se repiten de forma iterativa, lo que constituye el ciclo reproductivo básico o generación. Dichas fases son: selección, reproducción y reemplazo. El algoritmo evolutivo diseñado, emplea heurísticos para mejorar las soluciones antes de recombinarlas, introduciendo así, un mayor conocimiento del problema.

En nuestro caso, el Algoritmo Evolutivo sigue el siguiente esquema básico:

Algoritmo Evolutivo_Log

Generar una población inicial de soluciones
Mejorarlas mediante un método preestablecido
Repetir
- *Seleccionar aleatoriamente un subconjunto de elementos (par) de la población con una probabilidad proporcional a su bondad*
- *Cruce Reproducción: Emparejar o cruzar estas soluciones (Padres) para dar lugar a nuevas soluciones (Hijos). De cada pareja de padres se ha de generar una nueva pareja de hijos*
- *Mutación: Las soluciones hijas pueden cambiar, con una pequeña probabilidad, alguno de sus elementos (genes)*
- *Aplicar el procedimiento de mejora a los hijos*
- *Sustituir las peores soluciones de la población por las nuevas soluciones hijas*

Hasta conseguir algún criterio de parada

A continuación se hacen las siguientes consideraciones. Las operaciones de cruce y mutación se realizan sobre los vectores S de cada solución. El operador de cruce sigue el esquema «one-point crossing», es decir, sean S y S' dos vectores «padres» seleccionados:

$$S = (j_1, j_2, j_3, \ldots, j_{np}) \quad \text{y} \quad S' = (j'_1, j'_2, j'_3, \ldots, j'_{np})$$

se genera aleatoriamente un «punto de cruce» (*pto_cruce*) entre 1 y *np*, de forma que los nuevos vectores hijos se definen como:

$$S'' = (j_1, j_2, \ldots, j_{pto\_cruce}, j'_{pto\_cruce+1}, \ldots, j'_{np})$$
$$S''' = (j'_1, j'_2, \ldots, j'_{pto\_cruce}, j_{pto\_cruce+1}, \ldots, j_{np}).$$

Los componentes (calendarios) de cada nuevo vector hijo pueden cambiar o *mutar* con una pequeña probabilidad, *p_mut*. Para decidir si un determinado componente cambia, se genera un valor aleatorio uniformemente en (0, 1); si este valor es menor que *p_mut* se realiza el cambio. Éste consiste en elegir aleatoriamente un calendario diferente al actual y asignárselo (*mutación*). Este proceso de mutación tiene por objeto dar diversidad al proceso y evitar que éste se encajone en una región entorno a un mínimo local.

Una vez que se genera un nuevo vector S y pasa por el proceso de mutación, la segunda componente de la solución, $\boldsymbol{R}$, se define de la forma siguiente: $\forall d = 1, \ldots,$ *ndias* se consideran todos los pedidos que se han de entregar el día d, según lo establecido por S; la composición de los slots de $\boldsymbol{R}(d)$ se genera de acuerdo al siguiente esquema (al igual que en el apartado anterior con el algoritmo VNS):

- Diseño de $\boldsymbol{R}(d)$ mediante el algoritmo *LTP de Graham* et al. *(1969) para el MSP.*
- Mejora con intercambios 0-1 y 1-1, según se expuso en el subapartado 2.3.

Una vez generada $\boldsymbol{R}$, la solución hija (S, $\boldsymbol{R}$), es mejorada aplicando el procedimiento *Búsqueda_Local_Log*(S, $\boldsymbol{R}$).

Se definen los siguientes parámetros: *n_pob* = número de elementos de la población, *n_sel* = número de padres seleccionados, *p_mut* = probabilidad de mutación, y *max_gen* = número de «generaciones» (iteraciones) sin mejora como criterio de parada.

6. EXPERIENCIAS COMPUTACIONALES

Para contrastar y comparar la eficacia de los 3 procedimientos propuestos en este trabajo se han realizado una serie de pruebas, tanto con instancias ficticias (adaptadas de instancias de la literatura para otros problemas), como con instancias tomadas de los datos reales facilitados.

Se han tomado de adaptaciones de las instancias número 1, 2, 5 y 10 Christofides y Beasley, (1984) (instancias para el PVRP, *Periodic Vehicle Routing Problem*, correspondientes a *np* = 51, 50, 75 y 100), para 30, 60 y 90 días, (pues son tiempos habituales para los que se realizan las planificaciones de producción) con *n_slots* = 4 –media jornada– y 8 –jornada completa–. (En el apéndice se describe la forma en que se han adaptado estas instancias a nuestro problema).

Para cada caso se ha generado una población de 100 soluciones aleatorias[1]. Esta población sirve de punto de partida del algoritmo Evolutivo –Evol–. Por otra parte, la mejor

[1]Asignación de calendarios aleatoria y composición de slots, en cada día, usando LTP y mejorando con intercambios 0-1, 1-1.

de estas soluciones (*SI*) sirve de punto de partida para los otros 2 algoritmos, Búsqueda Tabú –BT– y VNS. En el caso de VNS, dado que la ejecución no es determinista, se presentan los resultados medios de 10 ejecuciones para cada instancia. Los parámetros han utilizado los siguientes valores: *Tabu_Tenure* = $np \cdot ndias$ (Búsqueda Tabú), *pmut* = 0.05, *npob* = 100 y *nsel* = 20 (Algoritmo Evolutivo). Como criterios de parada se han utilizado los siguientes:

- En el caso de la Búsqueda Tabú el criterio de parada es que pasen ($20 \cdot np \cdot n\_dias$) iteraciones sin mejora (*max_iter* = ($20 \cdot np \cdot n\_dias$)).
- En el caso de Búsqueda en el Entorno Variable es que *kmax* (número máximo de pedidos que se seleccionan para perturbar) = 20, y *max_iter* (número de iteraciones sin mejora) = *np*.
- Para el Algoritmo Evolutivo *max_gen* = *np* (generaciones sin mejorar).

Los resultados obtenidos para las instancias ficticias cuando *n_slots* = 4 son los que se muestran en la tabla 1 (en negrita se resalta la mejor solución).

Tabla 1. Resultados de las instancias ficticias (n_slots=4).

Instancia	n_dias	SOL. INICIAL	BT	VNS	EVOL
			Tiempo de Computación Total/T.C. hasta encontrar la mejor solucion		
1	30	1097	1010 ***14,89/0,00***	1009,7 *47,52/22,00*	**1007** *49,6/35,43*
1	60	1800	1696 *114,85/**56,96***	1686,1 ***107,48**/71,48*	**1676** *137,09/113,64*
1	90	2648	2545 ***75,19/0,05***	2531,7 *126,95/79,74*	**2511** *161,1/125,29*
2	30	953	**945** *28,06/13,45*	946,3 *36,42/**12,16***	946 ***27,13**/13,73*
2	60	1636	1619 *106,5/49,98*	**1617,8** *77,82/35,81*	1620 ***32,52/9,29***
2	90	2451	**2420** *69,53/**0,16***	2421,9 *107,63/51,23*	2425 ***39,93/6,15***
5	30	1747	**1742** *40,86/6,81*	1742,6 *93,12/29,28*	1743 ***38,5/9,78***
5	60	2958	**2951** *339,66/211,74*	2952,7 *157,54/57,36*	2954 ***58,94/9,29***
5	90	4501	4491 *239,36/77,77*	**4488,2** *218,42/83,80*	4489 ***107,6/34,05***
10	30	1501	**1500** *88,59/23,95*	1500,05 *107,76/17,98*	1501 ***38,89/0,00***
10	60	2475	**2462** *285,07/36,69*	2463,4 *282,82/125,80*	2467 ***90,9/33,06***
10	90	3755	3738 *426,99/**100,89***	**3734,1** *514,51/305,75*	3744 ***171,48**/ 104,41*
TOTALES		27522	27119 *1829,55/578,45*	27094,55 *1878,02/892,43*	**27083** ***953,68/494,12***

Los resultados obtenidos para las instancias ficticias cuando *n_slots* = 8 son los que se muestran en la tabla 2.

Tabla 2. Resultados de las instancias ficticias (n_slots = 8).

Instancia	n_dias	SOL. INICIAL	BT	VNS	EVOL
			Tiempo de Computación Total/T.C. hasta encontrar la mejor solucion		
1	30	763	616 ***32,52/16,20***	622,4 *53,47/30,18*	**610** *64,48/48,44*
1	60	1215	**981** ***60,75/0,77***	998,3 *69,68/35,05*	992 *70,8/40,65*
1	90	1705	1450 *83,7/**6,15***	1480,6 ***81,49/33,95***	**1444** *128,91/85,62*
2	30	613	**515** ***31,03/15,1***	520 *58,63/26,73*	518 *48,11/32,19*
2	60	1037	888 ***63,83/0,61***	892,8 *122,89/69,85*	**866** *119,46/94,8*
2	90	1531	**1329** ***71,73/0,99***	1342,1 *126,01/70,14*	1358 *71,96/37,74*
5	30	912	**887** ***51,85/0,05***	889,6 *110,47/23,78*	893 *58,22/19,33*
5	60	1612	1516 *178,95/**0,88***	1522,1 *193,71/68,12*	**1515** ***153,25/91,35***
5	90	2393	2299 *238,7/**0,49***	2308,8 ***220,03/69,78***	**2295** *391,07/300,33*
10	30	818	**760** *86,73/**1,48***	761,4 *186,5/64,82*	763 ***70,31/25,21***
10	60	1358	**1257** *323,89/**2,19***	1261,1 *421,97/234,62*	1264 ***205,59/130,56***
10	90	2019	**1907** *435,51/**9,12***	1911,9 *576,44/54,26*	1933 ***97,16/26,09***
TOTALES		15976	**14405** *1659,19/**54,03***	14511,1 *221,30/1081,31*	14451 ***1479,32/932,31***

Búsqueda Tabú aporta los mejores resultados para un mayor número de instancias. En el caso de las instancias de menor número de pedidos (1 y 2), las mejores soluciones las aporta, en general, el Algoritmo Evolutivo, mientras que en las instancias de mayor número de pedidos (5 y 10) las mejores soluciones las aporta Búsqueda Tabú. Con respecto al tiempo de computación, se observa que el Algoritmo Evolutivo es en general, el que menos tiempo de computación total emplea para aportar su mejor solución, aunque en las instancias de menor tamaño y sobre todo para *n_slots* = 8 BT aporta soluciones más rápidamente. VNS es el que peores soluciones aporta, sobre todo en el caso de *n_slots* = 8.

Dada la dificultad para obtener la solución óptima, se han obtenido unas, entendemos que, buenas aproximaciones a dicho óptimo. Para ello se han ejecutado todos los algoritmos para cada instancia, con diferentes soluciones y poblaciones iniciales, y un

tiempo de parada de una hora, y se ha tomado la mejor solución. Los resultados se muestran en la tabla 3.

Tabla 3. Instancias ficticias. Comparación con la mejor solución encontrada.

n_solts	Inst.	np	n_dias	BT	VNS	EVOLU.	Mejor Sol. Conocida	% Desviac.
4	1	51	30	1010	1009,7	**1007**	**1001**	0,60
			60	1696	1686,1	**1676**	**1668**	0,48
			90	2545	2531,7	**2511**	**2495**	0,64
	2	50	30	**945**	946,3	946	**937**	0,85
			60	1619	**1617,8**	1620	**1599**	1,18
			90	**2420**	2421,9	2425	**2396**	1,00
	5	75	30	**1742**	1742,6	1743	**1737**	0,29
			60	**2951**	2952,7	2954	**2940**	0,37
			90	4491	**4488,2**	4489	**4467**	0,47
	10	100	30	**1500**	1500,05	1501	**1494**	0,40
			60	**2462**	2463,4	2467	**2452**	0,41
			90	3738	**3734,1**	3744	**3721**	0,35
8	1	51	30	616	622,4	**610**	**574**	6,27
			60	**981**	998,3	992	**938**	4,58
			90	1450	1480,6	**1444**	**1392**	3,74
	2	50	30	**515**	520	518	**499**	3,21
			60	888	892,8	**866**	**856**	1,17
			90	**1329**	1342,1	1358	**1289**	3,10
	5	75	30	**887**	889,6	893	**880**	0,80
			60	1516	1522,1	**1515**	**1498**	1,13
			90	2299	2308,8	**2295**	**2264**	1,37
	10	100	30	**760**	761,4	763	**755**	0,66
			60	**1257**	1261,1	1264	**1237**	1,62
			90	**1907**	1911,9	1933	**1877**	1,60

Si comparamos la mejor solución aportada por los algoritmos iniciales con la «*Mejor Solución Conocida*» vemos que se consiguen resultados satisfactorios para $n\_solts = 4$ (la desviación es sólo del 0,58 %) y aceptables para $n\_solts = 8$ (la desviación es del 2,43 %).

Finalmente se va a analizar el comportamiento de las estrategias desarrolladas en este trabajo, para la resolución de una serie de instancias generadas a partir de los datos reales facilitados por los responsables de la empresa fabricante de componentes de automóviles. Se tienen 45 pedidos con las frecuencias y cargas que se indican en la tabla 4.

Tabla 4. Datos de los pedidos.

n°ped	*freq*	*q*	*n°ped*	*freq*	*q*	*n°ped*	*freq*	*q*
1	1D	12	16	1M	1	31	1S	2
2	1M	1	17	1D	15	32	1S	6
3	1M	2	18	1M	2	33	1M	4
4	1M	1	19	2S	4	34	1M	1
5	2S	20	20	2S	6	35	1D	3
6	1D	3	21	2M	3	36	1D	8
7	1S	3	22	1S	1	37	2M	2
8	1M	1	23	1S	2	38	2S	3
9	1D	6	24	2M	1	39	1S	2
10	1S	6	25	2M	1	40	1S	8
11	1M	1	26	1M	1	41	1S	10
12	1M	3	27	2S	6	42	1M	1
13	1S	8	28	2S	12	43	2S	6
14	1S	8	29	1S	1	44	1S	10
15	1D	3	30	1M	1	45	1S	2

El significado de las frecuencias es el siguiente: 1D = una recogida cada día; 1S = una recogida a la semana; 2S = una vez cada 2 semanas; en el primer caso sólo hay un calendario posible $\{(1,2,3,\ldots,ndias)\}$; en el segundo el conjunto de calendarios sería de la siguiente forma $\{(1,8,15,22,29,\ldots), (2,9,16,23,30,\ldots), \ldots\}$; en el tercero el conjunto sería $\{(1,15,29,\ldots), (2,16,30,\ldots), \ldots\}$.

Para estos datos se consideran horizontes de 30, 60 y 90 días y $n\_slots = 4$ (media jornada) y 8 (jornada completa), como en las anteriores instancias. Suponemos que el día 1 es lunes y se van a considerar festivos los domingos (es decir, los días 7, 14, 21, 28, ...). Si al diseñar un calendario algún día resulta ser festivo se cambia por el siguiente día (día 8 en vez de día 7, por ejemplo), sin variar el resto del calendario.

Para la resolución de estos problemas reales, Búsqueda Tabú (BT) y VNS, usan como solución inicial $-SI-$ la aportada por una empresa logística[2], mientras que el Algoritmo Evolutivo –Evol– usa una población de soluciones iniciales aleatorias como en las instancias anteriores.

Los resultados obtenidos para la instancia real son los que aparecen en la tabla 5 (en negrita la mejor solución). El tamaño del problema es en este caso similar al de las

[2]Grupo SAAT, (Sociedad Anónima de Aduanas y Transportes), Aduana Interior Villafría, 09192, Burgos, Spain.

instancias 1 y 2 anteriores. Al igual que en dichas instancias el mejor resultado lo aporta el Algoritmo Evolutivo.

Tabla 5. Resultados para las instancias reales.

slots	n_dias	SOLUCIÓN INICIAL	BT	VNS	EVOL
			Tiempo de Computación Total/T.C. hasta encontrar la mejor solucion		
4	30	516	448 *10,38/2,75*	443,6 *20,68/7,96*	**440** *24,67/17,58*
4	60	999	892 *51,73/19,71*	889,1 *34,17/12,96*	**887** ***22,52/8,51***
4	90	1484	1338 *96,06/25,21*	1336,8 *61,91/30,86*	**1332** ***46,69/26,31***
8	30	419	**400** *8,29/0,22*	**400** *12,93/0,19*	**400** ***5,77/0,00***
8	60	827	**800** *31,31/0,44*	**800** *20,02/0,31*	**800** *10,55/0,00*
8	90	1235	**1200** *68,16/0,71*	**1200** *21,67/0,42*	**1200** *14,94/0,00*

Podemos afirmar teniendo en cuenta estos resultados lo siguiente:

- Los tres algoritmos propuestos consiguen mejorar significativamente las soluciones propuestas por la empresa logística.
- Entre los tres algoritmos no hay ninguna diferencia para $n\_slots = 8$ (jornada completa), sin embargo si que se aprecian algunas diferencias para $n\_slots = 4$ (media jornada).
- En este último caso, el Algoritmo Evolutivo consigue llegar a soluciones mejores que los otros dos. Búsqueda Tabú es el que aporta peores soluciones.

7. CONCLUSIONES

En este trabajo se ha analizado un problema real planteado a los autores de este trabajo por los responsables de una fábrica de componentes de automóviles.

Se trata de un problema a dos niveles: en el primer nivel se trata de asignar un calendario de recogida a cada pedido; en el segundo nivel se trata de asignar, en cada día, una hora de recogida a cada pedido. El objetivo es racionalizar (minimizar) el número de contratos del personal encargado de preparar y cargar los pedidos periódicos que realizan los clientes en los vehículos de dicha empresa (carretilleros).

Del problema conjunto no existen referencias previas anteriores; aunque el segundo nivel está muy relacionado con el *Multiprocessor Scheduling Problem (MSP)*, problema muy estudiado en la literatura y relacionado con el *Bin Packing*.

Para su resolución se proponen tres sencillos algoritmos Metaheurísticos basados en movimientos vecinales: uno sigue un procedimiento de *Búsqueda Tabú* básico, el segundo sigue una estrategia de *Búsqueda en Entorno Variable* (VNS), y el tercero es un *Algoritmo Evolutivo*.

Búsqueda Tabú consigue para un mayor número de instancias mejores soluciones que los otros dos. Se observa sin embargo, que las soluciones aportadas por el Algoritmo Evolutivo son de mayor calidad que las de Búsqueda Tabú para las instancias de menor tamaño; en las instancias de mayor tamaño ocurre justamente lo contrario.

Finalmente, las estrategias analizadas mejoran sustancialmente para los problemas basados en datos reales, las propuestas realizadas por la empresa logística con la que trabaja habitualmente la fábrica de componentes de automóviles.

En cualquier caso, este trabajo supone una primera aproximación, ya que las estrategias analizadas son muy básicas. En futuros trabajos se espera completarlas con otros elementos (memoria a medio y largo plazo, renovación de la población), además de analizar otras estrategias u otros tipos de movimientos vecinales.

8. APÉNDICE: GENERACIÓN DE INSTANCIAS

Como se ha comentado en el apartado 6, se han generado diferentes instancias para este problema tomando los datos de conocidas instancias para el VRP o PVRP. La forma de realizar las adaptaciones es la siguiente:

- Se leen de dichas instancias las unidades o pallets $-q(i), i = 1, \ldots, np-$ demandadas de los pedidos;
- Se ordenan los pedidos según los valores de $q(i)$, de menor a mayor, para agruparlos: en el primer grupo los de menor valor de q, en el segundo los siguientes, ..., y en el último los de mayor valor.
- Se asigna la misma frecuencia y calendarios a todos los pedidos del mismo grupo.
- Se generan 3 instancias por cada instancia original del VRP o PVRP leída, correspondientes a horizontes temporales de 30, 60 y 90 días.
- Para las instancias de 30 días se consideran 6 grupos: al primer grupo se le asigna una frecuencia de entrega diaria, (i.e., 30 entregas), al segundo una entrega cada 3 días (10 entregas), al tercero cada 5 días (6 entregas), al cuarto cada 10 días, al quin-

to cada 15 días y al sexto cada 30 días (una única entrega).

- Para las instancias de 60 días se toman 7 grupos y las siguientes frecuencias de entrega: 1 día (60 entregas), 2 (30), 6, 10, 20, 30 y 60.
- Para las instancias de 90 días se toman 8 grupos y las siguientes frecuencias de entrega: 1 día (90 entregas), 2 (45), 3, 6, 10, 30, 45 y 60.

Una vez que se asignan las frecuencias, los calendarios se generan automáticamente; por ejemplo, en un horizonte de 30 días los pedidos que tengan una entrega cada 3 días (segundo grupo), tendrán el siguiente conjunto de calendarios: $\{1,4,7,10,13,16,\ldots,28\}$, $\{2,5,8,11,14,17,\ldots,29\}$ y $\{3,6,9,12,15,18,\ldots,30\}$.

9. BIBLIOGRAFÍA

Babel, L., Kellerer, H. and Kotov, V. (1998). «The k-partitioning Problem», *Mathematical Methods of Operations Research*, 47 (1), 59-82.

Barnes, J. W. and Laguna, M. (1993). «A Tabu Search Experience in Production Scheduling», *Annals of Operations Research*, 41, 141-156.

Bentley, J. L. (1992). «Fast Algorithms for Geometric Traveling Salesman Problems», *ORSA Journal on Computing*, 4, 387-411.

Blazewicz, J. (1987). «Selected Topics in Scheduling Theory», *Annals Discr. Math.*, 31, 1-60.

Coffman Jr., E. G., Garey, M. R. and Johnson, D. S. (1978). «An Application to the Bin-Packing to Multiprocessor Scheduling», *SIAM J. Comput.*, 7, 1-17.

Christofides, N. and Beasley, J. E. (1984). «The Period Routing Problem», *Networks*, 14, 237-256.

Dell' Amico, M. and Martello, S. (1990). «Optimal Scheduling of Tasks on Identical Parallel Processors», Research Report OR/90/7. *DEIS*, University of Bologna.

Finn, G. and Horowitz, E. (1979). «A Linear Time Approximation Algorithms for Multiprocessor Scheduling», *BIT*, 19, 312-320.

Franca, P., Gendrau, M., Laporte, G. and Muller, F. M. (1994). «A Composite Heuristic for the Identical Parallel Machine Scheduling Problem with Minimun Makespan Objective», *Computers Ops. Res.*, 21, 2, 205-210.

Fujita, S. and Yamashita, M. (2000). «Approximation Algorithms for Multiprocessor Scheduling Problem», *IEICE Trans. Inf. & Syst.*, E83-D, 3, 503-509.

Garey, M. R. y Johnson, D. S. (1979). «Computers and Intractability». Freeman.

Graham, R. L. (1969). «Bounds on Multiprocessing Timing Anomalies», *SIAM J. Appl. Math.*, 17, 416-429.

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, 105-143.

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

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

Glover, F. y Laguna, M. (2002). «Tabu Search», *Handbook of Applied Optimization*, P. M. Paradalos and M. G. S. Resende (eds). Oxford University Press, 194-208.

Hansen, P. and Mladenovic, N. (1998). «An Introduction to Variable Neigborhood Search», in S. Voss *et al.* (eds.), *Metaheuristics Advances and Trends in Local Search Paradigms for Optimization*, 433-458, MIC-97, Kluwer, Dordrecht.

Hansen, P. and Mladenovic, N. (2000). «Recherche a Voisinage Variable», *Les Cahiers de GERAD*, G-2000-08, Montreal, Canada.

Hansen, P. and Mladenovic, N. (2001). «An Introduction to Variable Neigborhood Search». To appear in *Handbook of Applied Optimization*, P. M. Pardalos and M. G. C. Resende (eds.), Oxford Academic Press.

Harche, F. and Seshadri, S. (1995). «An LPT-Bound for a Parallel Multiprocessor Scheduling Problem», *Journal of Mathematical Analysis and Applications*, 196 (1), 181-195.

Hubscher, R. and Glover, F. (1994). «Applying Tabu Search with Influential Diversification to Multiprocessor Scheduling», *Computers and Operations Research*, 21, 877-844.

Langston, M. A. (1982). «Improved 0/1 Interchange Scheduling», *BIT*, 22, 282-290.

Mladenovic, N. (1995). «A Variable Neighborhood Algorithm-A New Metaheuristic for Combinatorial Optimization». Abstracts of papers presented at *Optimization Days*, Montreal, 112.

Mladenovic, N. and Hansen, P. (1997). «Variable Neighborhood Search», *Computers and Operations Research*, 24, 1097-1100.

Moscato, P. (1989). «On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms», *Caltech Concurrent Computation Program*, C3P Report 826.

Sahni, S. (1976). «Algorithms for Scheduling Independent Tasks», *J. ACM*, 23, 1, 116-127.

Thesen, A. (1998). «Design and Evaluation of Tabu Search Algorithms for Multiprocessor Scheduling», *Journal of Heursitics*, 4, 141-160.

ENGLISH SUMMARY

LABOR REQUERIMENTS FOR VEHICLE LOADING PROBLEM IN A STORE OF MANUFACTURED PRODUCTS

C. R. DELGADO
S. CASADO
J. F. ALEGRE
Universidad de Burgos*

This paper examines a problem proposed by a local autopart enterprise. This firm stores its outputs until the customers go to collect them. The clients apply for their products with a known frequency. The problem is to determine when the clients have to pick up their orders (the dates and the hours or slots).
With a planning horizon fixed, the aim of the firm is to minimize, for the joint of days, the number of «wheelbarrow mans» that are necessary for loading the orders in the customers' trucks; this number is determined each day for the busiest slot.
In this problem the decisions are made in two levels: doing a calendar of delivery for each order and the daily assignation of slots to the orders. There isn't any work in the literature that studies a problem like this.
In this work are designed and analyzed three simple Metaheuristics Procedures: the first is a basic Tabu Search, the second is an algorithm based on Variable Neighborhood Search and the third is a Evolutionary Algorithm. Computational experiences with fictitious and real instances are made.

Keywords: Logistic, Multiprocessor Scheduling Problem (MSP), Local Search, Tabu Search, Variable Neighborhood Search, Evolutionary Algorithm

AMS Classification (MSC 2000): 90B06

*Facultad C. Económicas y Empresariales. Pza. Infanta Elena s/n. 09001 Burgos.

-Received October 2001.
-Accepted June 2002.

In this paper, we address a logistical problem that a manufacturer of auto parts described to the authors. The manufacturer stores the products in its warehouse until customers pick them up. The customers and the manufacturer agree upon an order pickup frequency. For example, a customer may require a frequency of one pickup every other day, while another customer may require a frequency of one pickup every 4 working days. If the required frequency is of one visit to the manufacturer every 4 working days, then the possible pickup calendars in a 20-day planning horizon are:

$$\{1,5,9,13, \text{ and } 17\}$$
$$\{2,6,10,14, \text{ and } 18\}$$
$$\{3,7,11,15, \text{ and } 19\}$$
$$\{4,8,12,16, \text{ and } 20\}$$

where the numbers indicate the index of the day when the customer visits the manufacturer.

The problem is to find the best pickup schedule, which consists of the days and times during the day that each customer is expected to pick up his/her order. For a given planning horizon, the optimization problem is to minimize the labor requirements to load the vehicles that the customers use to pick up their orders. We approach this situation as a decision problem with two levels:

1. To assign each customer to a set of days in the planning horizon such that the frequency requirements are satisfied. That is, the problem is to assign each customer to one of its feasible pickup calendars.
2. To assign each customer to a time slot within each day such that the labor requirements are minimized.

Clearly, the decisions in the first sub-problem condition the best possible solutions that can be found when solving the second sub-problem. That is, the second sub-problem tries to minimize the labor requirements in each day, but the set of customers assigned to each day constraints the solution space. While, to the best of our knowledge, the entire problem has not been addressed in the literature, the second sub-problem is equivalent to the *multiprocessor scheduling problem* (MSP), or also known as the *k partition problem*, and to the *bin packing problem* that seeks to minimize the heaviest bin.

The objective function minimizes the sum of the daily labor requirements in the planning horizon. The maximum number of workers needed in any slot gives the labor requirements for the day. In the environment that we studied, workers cannot be hired on an hourly basis. That is, a full day is the minimum amount of time that a worker can be hired to load auto parts onto customer vehicles.

In this work are designed and analyzed three simple Metaheuristics Procedures: the first is a basic *Tabu Search*, the second is an algorithm based on *Variable Neighborhood Search* and the third is an *Evolutionary Algorithm*. Computational experiences with fictitious and real instances are performed.

Tabu Search arises the best solutions for the major of the instances. Note however, the Evolutionary Algorithm solutions are better than Tabu Search, for the smallest instances. For the biggest ones occurs the opposite. Finally the analyzed strategies improve the logistic enterprise solutions for the real problem.

In any case, this paper is a first approach, because of the analyzed strategies are basic. In future works we hope improve this strategies using another elements such as short-term memory, long-term memory and solutions set update methods. Also, we'll analyze another strategies and other types of neighborhood moves.

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. D'altra banda, a cada número de *Qüestiió* s'inclouen d'un a tres problemes i les solucions corresponents es publiquen en el número següent. Atès que el número 3 del volum 26 és el darrer número de la segona etapa de *Qüestiió*, a partir del qual deixaràn de publicar-se nous enunciats de problemes en la seva versió impresa, les solucions corresponents als problemes 93, 94, 95, 96, 97 i 98 podràn consultar-se únicament en el lloc web de la revista.

Els lectors poden proposar problemes amb les solucions pertinents i enviar-los a *Qüestiió,* que farà una selecció i en publicarà a la versió electrònica 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. La revista es reservarà, però, el dret a publicar-les.

SOLUCIÓ AL PROBLEMA PROPOSAT AL VOLUM 26 N. 1 i 2

PROBLEMA N. 92

1) Clearly $\mathbf{H} = \mathbf{H}^2$, hence $\mathbf{H}$ is (symmetric) idemptotent. Its eigenvalues are therefore 0 or 1. Hence $r(\mathbf{H}) = \mathrm{tr}(\mathbf{H}) = n - 1$, as $\mathbf{H}$ has $n - 1$ unit eigenvalues and one zero-eigenvalue. As usual $r(\mathbf{H})$ and $\mathrm{tr}(\mathbf{H})$ denote rank and trace of $\mathbf{H}$, respectively.

2) $\mathbf{B} = \mathbf{H} = \mathbf{H} \cdot \mathbf{H} = \mathbf{H}(\mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n')\mathbf{H}$, where $\mathbf{H}\mathbf{1}_n = 0$, $\mathbf{1}_n'\mathbf{H} = 0$. Hence

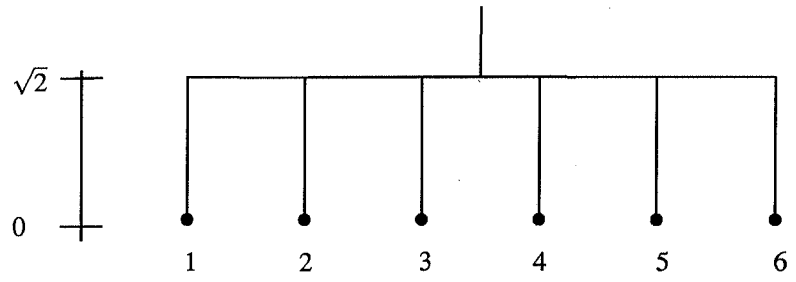
$$\mathbf{B} = -\frac{1}{2}\mathbf{H}\left(2(\mathbf{1}_n \mathbf{1}_n' - \mathbf{I}_n)\right)\mathbf{H}$$

and the (Euclidean) distance matrix is

$$\mathbf{D} = \sqrt{2}(\mathbf{1}_n \mathbf{1}_n' - \mathbf{I}_n)$$

i.e., $d_{ij} = \sqrt{2}$, $i \neq j$, and $d_{ii} = 0$, $i, j = 1, \ldots, n$.

3) Clearly, all distances are equal to $\sqrt{2}$. A dendrogram for the n objects, assumming $n = 6$, is:



Heinz Neudecker
Cesaro
The Netherlands

PROBLEMES PROPOSATS

PROBLEMA N. 93

Let the $(m \times m)$ random matrix S follow a central Wishart distribution with scale matrix Σ and n degrees of freedom. Find the expected value of $(\mathrm{tr}S)^{-1}S$, exact or approximate.

Heinz Neudecker
Cesaro
The Netherlands

PROBLEMA N. 94

Find the expected value of $SAS^{-1}BS^2$ when $S \sim W_m(\Omega, n)$.

Heinz Neudecker
Cesaro
The Netherlands

PROBLEMA N. 95

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be iid observations from continuous distribution functions F and G, respectively. To test the hypothesis $F = G$ the following statistic can be used

$$B(m,n) = \frac{mn}{(m+n)} \int (F_m(x) - G_n(x))^2 \, dx,$$

where F_m and G_n are the corresponding empirical distribution functions. The hypothesis may be rejected for large values of $B(m,n)$. Now suppose

$$F(x) = \frac{1}{1+e^{-x}} \quad -\infty < x < +\infty,$$

the logistic distribution. Prove that if $F = G$ then

$$B(m,n) \xrightarrow[m,n \to \infty]{} \sum_{i=1}^{\infty} \frac{Z_i^2}{i(i+1)}$$

where $Z_1, Z_2, \ldots$ are iid $N(0,1)$ random variables and the convergence is in distribution.

C. M. Cuadras
Universitat de Barcelona

PROBLEMA N. 96

Let $\mathbf{U} = (u_{ij})$ be a $n \times n$ ultrametric distance matrix, i.e.,

$$u_{ij} \leqslant \max\{u_{ik}, u_{jk}\} \qquad i,j,k = 1,\ldots,n,$$

where $n \geqslant 3$ and $u_{ij} > 0$ for $i \neq j$. It is well-known that $\mathbf{U}$ is a Euclidean distance matrix. Thus, if $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'$ is the $n \times n$ centring matrix and $\mathbf{A} = -\frac{1}{2}(u_{ij}^2)$, then the centred inner product matrix $\mathbf{B} = \mathbf{HAH}$ has non-negative eigenvalues

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{n-1} > \lambda_n = 0.$$

Prove that the smallest positive eigenvalue is given by

$$\lambda_{n-1} = \min\left\{\frac{1}{2}u_{ij}^2 \,|\, i \neq j = 1,\ldots,n\right\}.$$

C. M. Cuadras
Universitat de Barcelona

PROBLEMA N. 97

Let $F(x)$ be a continuous cumulative distribution function. Define

$$K(x) = F(x)\big(1 - F(x)\big).$$

Find $F(x)$ in terms of $K(x)$.

C. M. Cuadras
Universitat de Barcelona

PROBLEMA N. 98

Let $a_i > 0$, $i = 1,\ldots,n$, be n positive integer numbers. Use the well-known property that the likelihood ratio lies between 0 and 1, to prove the inequality

$$a_1^{a_1} \cdots a_n^{a_n} \geqslant \left(\frac{a_1 + \cdots + a_n}{n}\right)^{a_1 + \cdots + a_n}$$

with equality if $a_1 = \cdots = a_n$.

C. M. Cuadras
Universitat de Barcelona

QÜESTIIÓ, vol. 26, 3, p. 569-585, 2002

ESTADÍSTICA, SOCIETAT I VERITAT

C. M. CUADRAS
Universitat de Barcelona*

Fem una exposició general de la incidència de l'estadística en la societat, des d'una perspectiva històrica i actual. Els mètodes i resultats de l'estadística representen una forma actual i imprescindible del pensament, que abasta tots els camps del coneixement i totes les activitats humanes.

Statistics, society and truth

Paraules clau: Applications of statistics, determinism, chaos, chance, statistics of populations, quirks and paradoxes, discoveries due to statistics

Classificació AMS (MSC 2000): 62-01, 62A01

* Department d'Estadística, Universitat de Barcelona.

– Rebut a l'octubre de 2001.
– Acceptat al setembre de 2002.

INTRODUCCIÓ

L'Estadística té cada vegada més influència en la societat. Els periòdics porten cada dia resultats estadístics sobre economia, salut, opinió, política. Què hi ha de veritat en una estadística? Podem desvirtuar la veritat amb estadístiques, però podem especular més sense estadístiques. L'estadística és necessària quan els fenòmens objecte d'estudi no es poden predir amb exactitud, doncs tenen una component d'atzar, d'incertesa. Però l'atzar ja no és el producte de la nostra ignorància, sinó una forma d'expressió del nostre coneixement. La incertesa es pot controlar i quantificar gràcies a l'estadística.

Quan les estadístiques estan basades en dades certes, poden donar una informació molt valuosa a la societat. En aquesta dissertació veurem com ha evolucionat la població humana des dels seus orígens fins ara, com ha canviat Espanya en els últims cent anys, algunes dades estadístiques històriques, com el coneixement estadístic ens informa de com viure més i millor, com l'anomenat caos determinista es pot descriure en termes estadístics, com desmitificar alguns miracles i resultats sorprenents amb estadística, la diferència entre determinisme i aleatorietat, com esbrinar qui era el vertader autor d'una obra. Recordarem què pensaven els grans científics sobre el determinisme i l'atzar. Il·lustrarem i comentarem la regularitat estadística, la demagògia política, les curiositats i paradoxes estadístiques, alguns resultats polítics impossibles, l'estadística familiar i algunes característiques de certes poblacions descobertes amb estadística. En definitiva, provarem que l'estadística és una eina potent i imprescindible, utilitzada en tots els camps de l'activitat humana. Intentarem distingir entre la veritat estadística i la veritat.

POBLACIÓ MUNDIAL

Potser l'estadística més bàsica i més global, fa referència al nombre de persones que viuen i han viscut a la Terra. La Taula 1 conté aquest nombre d'habitants.

Taula 1. Població abans i ara.

Any	Població en milions
10000 a.C.	5
1 d.C.	250
1100 d.C.	500
1800 d.C.	1000
1930 d.C	2000
1960 d.C.	3000
1987 d.C.	5000
1998 d.C	6000

És realment curiós el fet que mai hi ha hagut dues persones idèntiques, llevat dels bessons univitelins. El nombre de genotips diferents és tan gran (de l'ordre de trilions) que els 6000 milions d'habitants actuals no són suficients per a trobar alguna coincidència.

UN SEGLE EN XIFRES

Concretem ara les estadístiques a Espanya i a Barcelona. Si comparem el final del segle XIX amb el del segle XX, podem observar que estem molt millor ara que abans

Taula 2. Xifres estadístiques comparades.

	1898	1998
Esperança de vida	34,8 anys	80 anys
Població	18,5 milions	39,6 milions
Mortaldat infantil	172/mil	7,6/mil
Analfabetisme	63 %	2,6 %
Jornada laboral	11-15 hores	8 hores
Universitaris	17.000	1.543.805
Habitants Barcelona	500.000	1.501.805

L'any 1898 va presenciar el fi de l'imperi colonial espanyol. Espanya va perdre Cuba i Filipines en una desigual guerra contra els Estats Units.

Una curiosa estadística propagandística de l'exèrcit nord-americà, a fi de reclutar voluntaris per anar a lluitar a Cuba, deia que mentre a la ciutat de New York morien 16 de cada mil habitants, a la guerra morien només 9 de cada mil. Per tant, era més segur allistar-se que romandre a la ciutat. Aquest seria un exemple de resultat que no té en compte altres variables, com l'edat. A New York hi vivien joves i vells. A l'exèrcit només s'hi allistaven joves.

ESCONS

L'estadística dels escons al parlament espanyol a finals del segle XIX i principis del XX dóna sorpreses, com ens mostra la Taula 3. El partit dels conservadors guanyava als liberals de manera alternada. Passava del poder a l'oposició amb un nombre d'escons similar. És això possible? No en condicions democràtiques normals. Resultava que els dos principals partits s'alternaven en el poder de forma pactada.

Taula 3. Escons al parlament espanyol.

Any	Conservadors	Liberals
1896	269	88
1898	68	266
1899	222	93
1901	79	233
1903	230	93

FRASES SOBRE L'ESTADÍSTICA

La història de l'estadística és recent i des que es va començar a aplicar ha rebut crítiques. Són conegudes les frases:

- Mentides, grans mentides i estadístiques.
- Conec la resposta, doneu-me una estadística i la justificaré.
- Una mostra estadística convenientment torturada confessa el que vulguis.

Però l'estadística no menteix, menteixen els que en fan un mal ús, sigui per falsificació, sigui per omissió. Per tant:

- Podem mentir amb estadístiques però podem mentir més sense estadístiques.

Aleshores l'estadística, ben utilitzada i ben interpretada pot ser molt útil a la societat. Les següents frases ho demostren:

- Segons les estadístiques els homes casats viuen més anys.
- Una enquesta estadística revela que la presa d'aspirina en dies alterns redueix el risc d'infart.
- Estadísticament parlant, els pares alts tenen fill alts.
- L'estadística demostra que fumar perjudica la salut.
- Prendre vitamina C cada dia pot prolongar la vida 6 anys.

DETERMINISME, INDETERMINISME, CAOS I ATZAR

Determinisme, indeterminisme

L'estadística apareix quan el grau de coneixement d'un fenomen és imprecís. Hi havia un temps en que es pensava que tot estava ben determinat. Grans científics i pensadors així ho creien. Es va dir:

- Grans, eternes i immutables lleis determinen els camins que tots recorrem sense rumb fix (Goethe).
- Déu no juga a daus amb l'univers (Einstein).
- L'atzar és potser el pseudònim de Déu quan no vol signar (A. France).
- Els coneixements de les lleis de Newton ens determina el passat, present i futur del sistema solar (Laplace).

Malauradament aquest optimisme seria trencat per Gödel, un matemàtic austríac. Gödel va provar que hi havia resultats que no es podien demostrar. Per exemple, la conjectura de Goldbach, que diu que tot nombre parell és suma de nombres primers ($8 = 3 + 5$, $12 = 5 + 7$, etc.), és segurament certa, però som incapaços de demostrar-la. Gödel va provar que:

- No hi ha cap sistema axiomàtic complet i coherent en matemàtiques.

Hi ha propietats que no podem demostrar, i si les acceptem com axiomes, aleshores apareixeran noves propietats que seguirem sense poder demostrar. Aquest indeterminisme, aquestes limitacions del determinisme, van propiciar les ciències que tenen com a base les probabilitats i l'estadística.

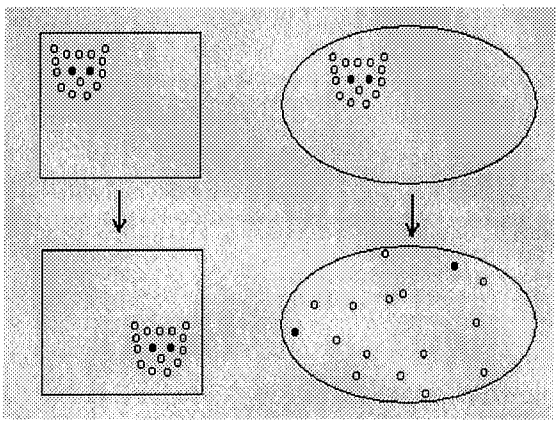


Figura 1. Comportament caòtic d'un procés determinista.

Caos

Posteriorment, en veure la impossibilitat de predir el temps atmosfèric, va sorgir el concepte de caos determinista. No tan sols hi havia veritats que no es podien demostrar, sinó que a més hi havia sistemes aparentment molt simples que podien evolucionar d'una manera molt complicada. Sistemes clarament deterministes però que eren impossibles de controlar. La Figura 1 és un exemple. Les boles dibuixen la cara d'un gat. Si es mouen en la mateixa direcció, rebotant a les bandes d'un rectangle, la cara del gat es manté. Però si les boles reboten a les bandes d'un recinte en forma d'estadi, les boles segueixen direccions diferents i al cap d'una estona, pràcticament ja no podem preveure on estarà cada bola.

Un exemple molt senzill de sistema caòtic ve donat per la fórmula iterativa

$$T \longrightarrow 1 - 2T^2$$

que consisteix en començar per un nombre real T entre -1 i $+1$, i seguidament calcular $T' = 1 - 2T^2$. Al resultat obtingut T' li tornem a aplicar la mateixa fórmula i així successivament. Si aquest interval, multiplicant per 30, el traslladem a l'interval que va de -30 a $+30$, com si fos una temperatura, resulta que si comencem amb 20 graus, després de 50 iteracions arribem a 21,1, però si el valor inicial és 20,1, el final serà $-29{,}7$. Així, en 50 passes:

partint de 20 arribem a 21,1; partint de 20,1 arribem a $-29{,}7$.

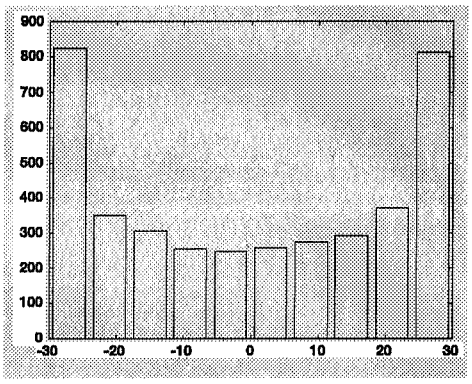


Figura 2. Distribució estadística de valors generats per una fórmula iterativa.

Una petita variació en l'estat inicial ens proporciona un valor final molt diferent. Què podem dir dels valors finals després de moltes iteracions? Matemàticament el podem

trobar, però a la pràctica, per limitacions de computació, no podem predir amb precisió aquest valor final, només la seva distribució estadística. La Figura 2 representa l'histograma de $n = 4000$ valors consecutius generats seguint aquest algorisme. Aquesta distribució segueix una llei de probabilitats coneguda. Podem observar que els valors extrems són més probables que els valors centrals.

Atzar

L'estadística estudia els fets que tenen una component aleatòria. Podem entendre l'atzar com el comportament d'un fenomen que no és ni determinista ni caòtic, sinó aleatori. Hi ha diferents possibilitats i cadascuna té un grau de probabilitat.

L'estadística, en sentit ampli, seria la disciplina que tractaria de la recol·lecció, estudi i interpretació de les dades (socials o de la natura), quantificant la seva incertesa.

L'estadística descriptiva tindria la missió de «descriure» les dades mitjançant histogrames, com el de la Figura 2, i altres gràfics i diagrames, i calculant certes mesures (moda, mediana, mitjana) que permetrien resumir la tendència de les dades. L'estadística matemàtica estaria basada en la probabilitat i proporcionaria models i mètodes per treure conclusions sobre una població a partir d'una mostra.

DESCOBRIMENTS I ACTIVITATS ON INTERVÉ L'ESTADÍSTICA

La metodologia estadística ha intervingut en gairebé tots els camps del coneixement.

- Economia i política. Saber l'estat social i econòmic d'un país permet als governs fer prediccions a llarg i a curt termini. La prosperitat d'un país es pot mesurar per la qualitat de les seves estadístiques.
- Evolució. Estudiant les similituds entre espècies podem construir un arbre evolutiu. Fent el mateix amb paraules corrents (ull, mà, mare, u, ...) de diferents llengües podem estimar el temps de separació entre dues llengües i seguidament construir l'arbre evolutiu de les llengües.
- Origen comú de les anguiles. Observant que mostres d'anguiles, malgrat trobar-les a llocs molt distants, presentaven mitjanes i desviacions típiques semblants, es va concloure que procedien d'una mateixa àrea de criança de l'oceà, que posteriorment va ser localitzada.
- Autor d'una obra. Estudiant la distribució de les paraules diferents utilitzades per Shakespeare (884647 paraules en total, de les quals 31534 eren diferents) i altres escriptors clàssics, es va poder esbrinar que Shakespeare havia escrit un nou poema que no tenia autor.

- Control de qualitat. Tècniques estadístiques relativament senzilles permeten millorar la qualitat dels productes manufacturats.
- Negocis. Els models estadístics són útils per predir la demanda, controlar els estocs i planificar la producció.
- Medicina i farmacologia. Tècniques estadístiques permeten diagnosticar de manera objectiva, certes malalties a partir de símptomes, de dissenyar i comparar nous medicaments, d'estudiar la supervivència de grups de malalts, de quantificar els factors de risc.
- Plets a tribunals. L'evidència estadística amb l'ajuda de la probabilitat permet complementar les declaracions orals i documentals en els plets, com per exemple, quan es discuteix un tema de paternitat.
- Estructura de la personalitat i de la intel·ligència. Tècniques estadístiques multivariants permeten explorar i quantificar les dimensions de la personalitat i de la intel·ligència.
- Dietes. El consum de peix blau (com la sardina) redueix el colesterol i els problemes de cor.

D'altra banda hi ha fets dels quals es desconeix la causa o no es disposa d'una explicació clara, però que van sortir a la llum gràcies a l'estadística. L'estadística ens revela que:

- Falten nenes a la Xina. El naixement de nens i nenes a la Xina no segueix la proporció natural per causes provocades, resultant que hi ha un dèficit de cents de milers de nenes.
- Inutilitat de la pena capital. Un estudi estadístic revela que l'aplicació de la pena capital no redueix la criminalitat a curt termini als Estats Units.
- Relació oli de colza-síndrome tòxic. La pneumònia atípica, una malaltia d'origen desconegut que va afectar a milers d'espanyols a finals dels anys setanta del segle XX, va ser estadísticament relacionada amb el consum d'oli industrial no apte per al consum.
- Suïcidis. La tassa de suïcidis al cos de carrabiners de Xile és 9 vegades superior a la normal. Les autoritats han pres mesures per prevenir aquesta tendència.
- Quetelet (segle XIX), ajustant la campana de Gauss o corba normal a les talles dels mossos en edat militar, va descobrir que 2000 joves havien al·legat una talla inferior a la mínima establerta per eludir el servei militar.

PARADOXES DE L'ESTADÍSTICA

L' estadística, com les matemàtiques, la lògica i altres branques de la ciència, no és lliure de les paradoxes i resultats sorprenents o difícils d'entendre.

- Regressió de les estatures. La famosa regressió a la mitjana, descoberta per Galton, ens prova que pares alts tenen fills alts però, en mitjana, més baixos que el pare.
- Percentatges parcials i globals. Podem tenir uns percentatges totals favorables a un grup i en canvi els percentatges parcials resultar favorables a un altre grup. Per exemple, pot succeir que la proporció de noies admeses a una universitat sigui inferior als nois, però que a cada facultat per separat passi el contrari. Coneguda com paradoxa de Simpson, s'explica (en aquest cas) pel fet que moltes noies optin a facultats més selectives.
- Suposem que hi ha 3 candidats A, B, C a un càrrec polític. Si 1/3 de l'electorat ordenen els candidats segons preferències ABC, un altre 1/3 BCA i els restants CAB, aleshores tots 3 candidats «són els millors». En efecte, 2/3 parts prefereixen A sobre B i C, 2/3 parts prefereixen B sobre A i C i 2/3 parts també prefereixen C sobre A i B. Aquestes i altres incoherències del sistema democràtic són conseqüència del Teorema d'Arrow.

CURIOSITATS ESTADÍSTIQUES

- Quan transmeten un partit de futbol per televisió disminueixen les urgències a l'Hospital Clínic de Barcelona.
- Només coneixem l'1 % dels virus, el 0,13 % dels bacteris.
- Només veiem el 2 % de l'univers.
- En el món hi ha 4,3 arbres per persona.
- Hi ha prop de 40000 noies per 25000 nois a la Universitat de Barcelona.
- En el Nou Testament apareixen 215 noms propis d'home i també 215 noms propis de dona.
- El 23 % dels lectors d'un diari comencen a llegir-lo per darrera.
- Els malalts ingressats a la UVI tenen més probabilitat de sobreviure si poden veure la finestra.
- Els jugadors que van als casinos en dies de lluna plena guanyen un 2 % més.
- El 100 % dels votants van donar suport al govern de l'Irak. Molts anys abans, a Espanya el 103 % dels votants van dir sí a un referèndum franquista.

REGULARITAT ESTADÍSTICA

La regularitat, la periodicitat dels fenòmens naturals, forma part del sistema solar, del món i dels seus habitants. Però l'estadística, malgrat descriure fets aleatoris, també té una certa regularitat. La Taula 4 ens ho mostra. El nombre mitjà de gossos que van mossegar algú per dia a New York es manté quasi constant al llarg de 5 anys. La mitjana d'assassins a Anglaterra i Gales varia molt poc al llarg de 5 dècades. Les freqüències de 0 defuncions, 1 defunció, 2 defuncions, 3 defuncions i 4 defuncions a causa d'una guitza de cavall (dades de 1898) a 200 cossos l'exèrcit Prusià , així com la freqüència de bombes volants caigudes sobre 575 quadrícules de Londres durant la segona guerra mundial (és adir, a 229 quadrícules van caure 0 bombes, a 211 quadrícules van caure 1 bomba, etc.), i el nombre de platets voladors observats a Estats Units a diferents llocs, segueixen una mateixa llei estadística: la llei de Poisson. No és fàcil explicar aquestes regularitats, sobretot si estan causades per animals. Semblen reflectir unes probabilitats latents, unes lleis de l'atzar ocultes, que les estadístiques posen de manifest.

Taula 4. Regularitat estadística.

Gossos mossegadors/dia N. York	**75,3**	**73,6**	**73,5**	**74,5**	**72,4**
Any	1955	1956	1957	1958	1959
Assassins/milió Anglaterra-Gales	**3,84**	**3,27**	**3,92**	**3,30**	**3,50**
Dècada	1920-29	1930-39	1940-49	1950-59	1960-69
Nombre	**0**	**1**	**2**	**3**	**4**
Defuncions per guitza de cavall	109	65	22	3	1
Bombes volants sobre Londres	229	211	93	35	7
Platets voladors	75	152	167	98	41

ESTADÍSTICA FAMILIAR

Clíniques de maternitat convertides en geriàtrics, descens del nombre d'alumnes a les guarderies, escoles i universitats, agències especialitzades en viatges per a jubilats, i altres fets socials que són conseqüència del descens de la natalitat i l'envelliment de la població. L'estadística no pot esbrinar les causes, però pot descriure les dades i estudiar la seva evolució. La Taula 5 ens mostra l'augment de l'edat de contraure matrimoni i el descens del nombre de fills per dona. La Figura 3 il·lustra l'evolució de les edats dels joves i grans.

Taula 5. Edat de casament d'homes i dones (Catalunya) i mitjana de fills per dona.

Any	Homes	Dones	Catalunya	Europa
1975	26,2	23,7	2,7	1,9
1981	25,7	23,4	1,7	1,8
1986	26,6	24,5	1,4	1,6
1991	27,8	25,8	1,2	1,5
1995	28,7	26,8	1,1	1,4
1996	29,0	27,0	1,2	1,4

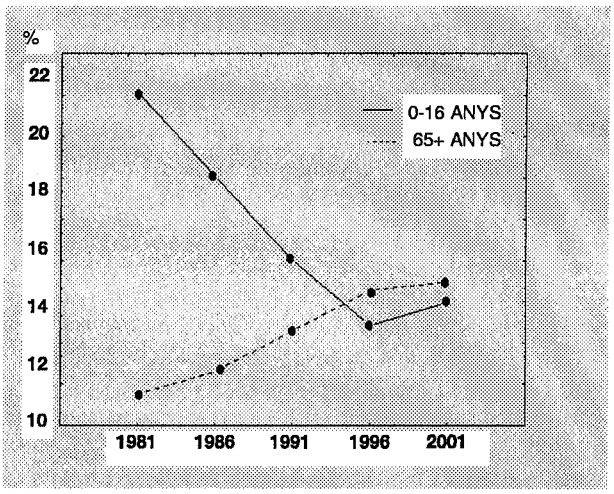


Figura 3. Evolució del percentatge dels joves respecte els grans.

ESTADÍSTICA VITAL

Morts no naturals

A tots ens arribarà el moment de deixar aquest món, però alguns no moren voluntàriament. L'estadística ens pot donar la probabilitat aproximada de morir per causa no natural. Aquesta probabilitat, que no seria la mateixa a cada país, s'aproxima trobant la proporció de morts en cada cas.

Taula 6. Probabilitats de mort no natural.

Causa	Probabilitat
Accident de cotxe	1/100
Homicidi	1/300
Incendi	1/800
Arma de foc	1/2500
Llampec	1/5000
Meteorit	1/10000
Accident aeri	1/20000
Inundació	1/30000
Serp verinosa	1/100000

Espera fins l'aniversari

Una curiositat estadística sobre la mort de personatges cèlebres: el 44 % moren en els mesos abans de l'aniversari, i el 56 % moren en els mesos després de l'aniversari. La majoria sembla que esperi a complir anys abans de morir

Violència

La violència amb resultat de mort es pot mesurar estadísticament pel nombre de morts per homicidi per cada 100000 habitants. Espanya amb 3,3 és el més violent d'Europa, que té una mitjana d'1,7. D'altra banda, resulta sorprenent que els països més violents són catòlics i els menys violents són musulmans, o al menys era així fa pocs anys.

Taula 7. Tasses d'homicidis.

Catòlics	Morts/cent mil	Musulmans	Morts/cent mil
Colòmbia	77,5	Egipte	1
Brasil	23,5	Jordània	1
Panamà	21,5	Kuwait	1
Mèxic	20,5	Bangla Desh	2,5
Veneçuela	16,5	Malasia	2,5

Disminució de l'esperança de vida

La Taula 8 ens descriu la disminució mitjana del nombre de dies sobre l'esperança de vida i que cal interpretar correctament. Els solters viuen, en mitjana, prop de 10 anys menys que els casats.

Taula 8. Disminució mitjana de dies de vida.

Causa	Dies
Solter	3500
Soltera	1600
Obesitat	1300-900
Fumar (homes)	2250
Fumar (dones)	800
Fumar en pipa	220
Alcohol	130

Fenòmens espontanis i miracles

Hi ha fets inexplicables com per exemple curacions espontànies. Si cents de milers d'individus demanen una curació miraculosa a un sant, és gairebé segur que alguns d'ells es curaran. Si ha intercedit o no el sant és un tema que no es pot demostrar, llevat de la fe i l'acceptació oficial de l'autoritat religiosa. Però l'estadística pot donar xifres totalment objectives. Per exemple, al Santuari de Lourdes han acudit més de 100 milions de pelegrins. La taxa de curació espontània de càncer va des de 1 entre 100000 fins a 1 entre 10000. Si el 5 % de pelegrins patien càncer, caldria esperar una curació espontània d'un mínim de 50 i un màxim de 500 afectats. Malgrat aquestes estadístiques, l'Església Catòlica només reconeix 3 curacions miraculoses.

Anàlogament, si un il·lusionista a través de la televisió demana que tothom que vulgui agafi un rellotge que no funciona i el sacsegi, i això ho fan milers de persones (com va demanar Uri Geller fa molts anys), alguns rellotges es posaran en marxa, però no com a conseqüència dels poders del mag.

Una altra estadística sorprenent es va obtenir separant, a l'atzar, 524 i 466 pacients amb problemes de cor. El primer grup rebia només tractament mèdic, però cada pacient del segon grup rebia a més l'ajuda de persones que pregaven per ells, sense que el pacient ho sabés. Les curacions en el grup que rebia les pregaries van ser superiors.

En general, l'estadística pot predir quants estaran afectats per algun fet, però no podrà predir quins. Si tenim una gran mostra de fumadors i bevedors d'alcohol, podem predir quants patiran càncer de laringe, però sobre una persona particular l'estadística només pot quantificar el risc que té un fumador, que pot a ser 50 vegades el d'una persona no fumadora.

LES SOCIETATS, ELS INSTITUTS I LES ESCOLES D'ESTADÍSTICA

Tots els països moderns tenen societats professionals que recullen i defensen els interessos del estadístics, i instituts d'estadística, que elaboren estadístiques d'interès econòmic i social.

La Sociedad de Estadística e Investigación Operativa (SEIO www.seio.es), fundada el 1962, té més de mil socis, edita les revistes TEST i TOP, organitza un congrés nacional cada 18 mesos, i està estructurada en grups de treball.

L'American Statistical Association té més de vint mil socis, edita un butlletí mensual i moltes revistes científiques d'alt nivell.

El Instituto Nacional de Estadística (INE) elabora les estadístiques nacionals i calcula indicadors econòmics, com l'índex de preus al consum, la taxa d'atur i la inflació. Cada comunitat autònoma té una institució similar, més especialitzada en les estadístiques pròpies. La catalana és l'Institut d'Estadística de Catalunya (IDESCAT www.idescat.es), que depèn de la Generalitat de Catalunya i publica estadístiques sobre demografia, població, economia, comerç, indústria, sanitat i ensenyament. Edita la revista *Qüestiió*, transformada recentment en la revista SORT.

L'organització internacional més important d'estadística és el International Statistical Institute (ISI www.cbs.nl/isi/). Fundat el 1885, promou la metodologia estadística i les seves aplicacions a nivell de cooperació internacional. Té moltes seccions i organitza un congrés cada dos anys, en el que participen més de mil persones.

La importància de l'estadística es manifesta per ser una especialitat a totes les carreres de matemàtiques, ciències econòmiques, agrícoles, biològiques i mèdiques. Això comporta la creació d'Escoles i Diplomatures d' Estadística. La primera Escuela de Estadística es va fundar a Madrid el 1952. Actualment moltes universitats tenen Diplomatures d'Estadística. A Catalunya imparteixen aquest ensenyament la UAB, UB i UPC, i algunes universitats, com la UPC, imparteixen a més la Llicenciatura d'Estadística de segon cicle.

EPÍLEG

La metodologia i la tècnica estadística s'aplica a tots els camps del saber, fins i tot a camps com el dret i la lingüística, a les empreses, als centres de trànsit, als hospitals, als laboratoris, a les fàbriques, etc. Cap metodologia posseeix aquesta ubiqüitat. On hi ha dades que cal tabular, representar, veure quina distribució segueixen i finalment interpretar, és necessària la intervenció d'un estadístic. La millor estadística a favor de l'estadística és potser que gairebé el 100 % dels estadístics troben feina.

REFERÈNCIES

Rao, C. R. (2002). «Estadística y Verdad». 1ª ed. PPU, Barcelona, 1994, 2ª ed. EUB, Barcelona.

Tamur, J. M. (1992). «La Estadística. Una guía de lo desconocido». Alianza Editorial, Madrid.

ENGLISH SUMMARY

STATISTICS, SOCIETY AND TRUTH

C. M. CUADRAS
Universitat de Barcelona*

We show how Statistics has influenced society, both from a historical and present-day perspective. Statistical methods and results obtained from these methods involve a modern and indispensable way of thinking, covering all fields of knowledge and human activities.

Keywords: Applications of statistics, determinism, chaos, chance, statistics of populations, quirks and paradoxes, discoveries due to statistics.

AMS Classification (MSC 2000): 62-01, 62A01

* Department d'Estadística, Universitat de Barcelona.

- Received October 2001.
- Accepted September 2002.

In this paper we present statistics on human population, politics, economics, industry, biomedical sciences and religion. We discuss and illustrate concepts such as determinism, chaos and chance.

Statistics is necessary when the subject under study is not deterministic, having a random component. All the possible results can be quantified thanks to the probability. In general, statistics alows us to describe, represent and quantify data coming from social and experimental sciences.

Several favourable sentences are compared to other contrary sentences to statistics and it is discussed what is true and what is false in statistics. The statistical regularity as well as several paradoxes and curiosities in statistics are also commented.

Some discoveries due to statistics are presented. Statistics on life and family is specially commented. Improbable events or difficult to explain, even miracles, are analyzed under the statistical perspective.

It is shown that chance is not the consequence of our ignorance but a natural fact that can be controlled and studied thanks to statistics, which plays an important role in the modern society.

Finally the activities of some official statistical societies as well as the schools and careers of statistics are commented.

Comentaris de llibres

Análisis de datos multivariantes

Daniel Peña

Editorial McGraw Hill, Madrid, 2002
539 pàgines

En los últimos años han aparecido bastantes libros y monografías sobre análisis multivariante, pero el nuevo libro del Profesor Daniel Peña es una novedad por la cantidad y calidad de su contenido. A lo largo de 16 capítulos, el autor no tan sólo aborda los temas clásicos, como el análisis de componentes principales, el análisis factorial, el análisis de correspondencias, los tests multivariantes, sino que además presenta nuevas técnicas como la minería de datos, el reconocimiento de patrones, el modelo logit, la discriminación con mixturas de distribuciones y métodos avanzados de inferencia.

Después de un repaso del álgebra lineal, con aspectos incluso avanzados, una primera parte del libro expone diversos métodos para la descripción multivariante de los datos y puede leerse independientemente. Incluye un tratamiento de los datos atípicos. También se aborda el escalado multidimensional, el biplot y el análisis de conglomerados. Numerosos ejemplos, basados en datos reales, y gráficos de calidad, ilustran correctamente las técnicas.

La segunda parte empieza con una exposición de las definiciones y principales propiedades de las distribuciones multivariantes, como la normal, multinomial, Dirichlet, elípticas y esféricas, así como las que aparecen en muestreo multivariante como la Wishart y Hotelling, con una referencia a la Wilks. Seguidamente aborda la inferencia clásica, como el análisis multivariante de la varianza, y la inferencia avanzada (datos faltantes, algoritmo EM, robustez, métodos bayesianos, selección de modelos). El análisis factorial y discriminante son tratados desde una perspectiva clásica y muy bien presentados, desviando, como en los otros capítulos, los detalles técnicos a unos apéndices. En la misma línea de nivel superior, el autor presenta métodos avanzados de clasificación a partir del modelo logístico, multilogit, árboles de clasificación, redes neuronales, máquinas de vector soporte. El difícil tema de clasificación con mixturas de distribuciones, los métodos de proyección y otros, son también abordados con claridad. El capítulo 16 y último está dedicado a la correlación canónica y otros métodos relacionados, como los modelos de ecuaciones estructurales.

El tratamiento es riguroso, por lo que la obra puede ser útil tanto para ingenieros y economistas, como para matemáticos. En general, se trata de un libro de interés para una amplia audiencia de profesores e investigadores.

Los conjuntos de datos vienen en un apéndice, pero se pueden obtener vía internet. Otro acierto del libro consiste en iniciar cada capítulo con una fotografía y breve reseña biográfica de estadísticos destacados: Hotelling, Cayley, Mahalanobis, Nightingale, Pearson, Benzecri, Tukey, Wilks, Jeffreys, Spearman, Fisher, Efron, Rao y otros.

Tanto la presentación general como la bibliografía (libros y artículos) son muy buenas. El índice es también acertado. Se trata, en definitiva, de una excelente novedad editorial que puede ser de gran utilidad para usuarios, alumnos, profesores e investigadores de la estadística multivariante.

Carles M. Cuadras
Universitat de Barcelona

Statistical consulting

J. Cabrera and A. McDougall

Springer-Verlag, New York, 2002
390 pàgines

The motivation for this book comes from the statistical consulting course that the authors have been teaching regularly for years. In this sense, the authors use a direct style, incorporating short sentences and a clear client-oriented strategy in the subjects, and focusing on communication and understanding issues. The book looks for the following goals: *(a)* Understanding the statistical consulting process, *(b)* developing effective communication skills, and *(c)* obtaining experience through case studies.

To achieve these objectives the book is divided into two parts. Part I (chapters 1 to 4) is devoted to the methodology of the statistical consulting and Part II (chapters 5 to 9) analyses in detail a wide rang of case studies of varying complexity that help the reader understand and appreciate the diversity of projects that can arise in statistical consulting. From «Job Promotion Discrimination» to «A Device to Reduce Engine Emissions», from «Does It Have Good Taste?» to «Expenditure in NY Municipalities» and from «Plastic Explosives Detection» to «Maria'S project (A Market Research Study)» is a motivating list of projects to be solved. The complexity goes from simple case studies requiring only standard statistical methods to research-oriented case studies that often require multivariate methods or specialized statistical methods for the analysis. Finally, three appendices including valuable information on resources, SAS and S-PLUS software packages, and useful reference statistical tables complete the book.

Chapter 1 is an introduction to statistical consulting. From the history of the scientific method and the development of statistics the authors present the need and characteristics of the statistical consultant and they identify the role of the statistician in diverse environments (pharmaceutical, telecommunication, business, government and university).

A detailed discussion on verbal and written communication skills involved in an effective consulting environment is presented in Chapter 2. The interaction consultant-client is considered here. Information on how to write reports, how to make effective presen-

tations as well as basic guidelines for writing and the importance of quality graphics are provided.

The aim of Chapter 3 is to provide the reader with an overview of some techniques that are commonly used in statistical consulting. In keeping with spirit of the book the emphasis of the presentation is on the client's perspective rather than the technical details on the methods. Short description of standard methods (exploratory data analysis, contingency tables, $t-$test, analysis of variance and regression) and general methods (nonparametric tests, general linear models, multivariate analysis, categorical data analysis and specialized procedures) are included. The chapter also contains a section devoted to design of experiments issues as well as a section where the capabilities of several statistical software packages are presented.

Chapter 4 reproduces the entire consultation process for a particular project from the initial contact with the client (belonging to the university environment) to the final written report and postcompletion followup. The chapter presents information related to the prior arrangements, the financial issues, the first meeting, the documentation needed, almost the full project analysis (including SAS code for the analysis and S-PLUS code for generating costumized graphs), the presentation of the results and the final written report, as well as the comments received from the client. Interesting remarks and advises are included. Specific proposals for forms like contract, project summary and invoice are also shown. The chapter ends with some questions left to the reader to complete the analysis.

Under the perspective that *the best way to learn about statistical consulting is to do it!*, the authors present in Part II a very interesting variety of real projects. Twelve case studies have been analyzed in three groups according on the level of complexity and the type of statistical analysis required. The methodologies illustrated in the case studies are as follows. Group I: Contingency tables, sample survey, $t-$test and analysis of variance, and summary statistics; Group II: Probit analysis, factorial designs, regression methods, and time series analysis; Group III: Mixed models, discriminant analysis, factor analysis, and data mining-multistage analysis.

The format for each case study is the following: (1) The context of the problem, questions to be solved and the main statistical issues, (2) The data. Description and format of the database and properties of the relevant variables, (3) Methods. Details on the statistical procedures that can be used for analysis, (4) Some preliminary results, including SAS and S-PLUS outputs and (5) Questions left to the reader as a exercise.

A very important and pedagogical issue is that the reader can reproduce the analysis of the projects because all the data are available via the Springer-Verlag or the authors' websites: `www.springer-ny.com`, `www.rci.rutgers.edu/~cabrera` and `www.csam.montclair.edu/~mcdougal`.

At the end of Part II, Chapter 9 proposes eight open case studies to be solved for the reader.

Appendix A presents a very useful reference information (in particular on the world wide web) and the outline for a fifteen weeks course on statistical consulting. The course is addressed to second-year graduate students and the proposal uses this book as a textbook and real case studies (presented by invited speakers) for the analysis. Appendix B is a introductory and comprehensive SAS and S-PLUS guide for beginners (but not only for) that includes very practical material. Finally, in Appendix C, the reader can found a Statistical Addendum consisting of univariate and multivariate distributions, standard tests and sample size tables.

In summary, this is a very accessible book, easy reading and very practical. The material is presented in a well-organized, itemized and structured way and is clearly motivated by the case studies experience of the authors. Although the reader would need to complement the book with some other more technical references in statistical methodology for a consultancy in depth, the book perfectly achieves the proposed objectives. This is the reason I would strongly recommend this book for library purchase.

Carles Serrat
Universitat Politècnica de Catalunya

Análisis estadístico de textos

Ludovic Lebart, André Salem, Mónica Bécue Bertaut

Editorial Milenio, Lleida, Spain, 2000
215 pàgines

The statistical analysis of texts is a quite new discipline that is becoming more and more important, both for the social scientists who want to study and compare any kinds of texts using quantitative methods and for the statisticians who deal with open questionnaires. In this interdisciplinary field the book «*Análisis Estadístico de Textos*» is a remarkable text, the first of its characteristics published in Spanish, which is written with an accesible language for a large range of human science researchers, and which presents, in a systematic way, the main subjects, concepts, methods and techniques of this subject. For these qualities, it constitutes a practical tool for research and also a basic contribution for the consolidation and legitimation of this new speciality.

T. Kuhn, in the papers collected in «The Road Since 'Structure'» (2000), claimed that he had always continued to think on the concept of paradigm and that he would finally relate it with the creation of new specialities of a discipline. We can emphasize that statistics, when applied to an specific matter such as economics take into account at the one hand theories of probability and algebra, that are integrated in the mathematical statistics and, on the other hand, the specific theories of the subject on which it works, Moreover it requires computer facilities and a large set of practical knowledge that is adquired in a long professional experience. For these reasons, modern statistics have been and continue to be very productive in the creation of new specialities or paradigms such as econometry, biometry, psychometry, and more recently, chemiometry. I think that we can look at the statistical analysis of texts from that point of view and consider that it is in a stage of social acceptance and consolidation.

The histoy of this kind of analysis, as professor Daniel Peña points out in his Foreworld of the book, has two periods. In the first period it only used univariate technics, focused on the analysis of frequences of words, and achieved some brilliant results, as those reported by an study of Mosteller and Wallace on the anonimous papers published in 1787-1788 in order to induce the citizens of New York to ratify the Constitution. In the second period, the statistical analysis of texts introduces multivariative technics. Now it is known that all present texts are elaborated and stored in computer supports, that

our personal computers have a huge capacity for multivariative analysis and that this method can answer relevant questions about text. These facts justify that we expect a brilliant future for the statistical analysis of texts.

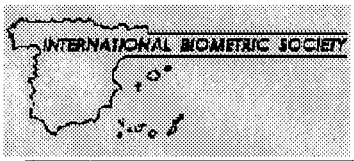
The book presents the following subjects: a short introduction to the notions of text and to statistical analysis, which focus on the exploratory analysis of textual data. The codification of open questionnaires. Lexical units and segmentation of texts. Lexicographical documents. Correspondence analysis of texts. Authomatic classification of tables and texts. Visualization of textual data. Searching for characteristic elements. Analysis of chronological corpus.

The quality of the book reflects the high scientific value and experience of its authors. Ludovic Lebart is research director at the CNRS (Centre National de la Recherche Scientifique) and professor of the ENST (Ecole Nationale Supérieur de Télécommunications) in Paris. He is one of the father founders of the French Correspondence Analysis School, has worked on the analysis of open questionnaires and has created the software pack «SPAD». André Salem is full professor on Language Sciences at the University Paris-Sorbonne. Mónica Bécue is professor of Statistics at the Universitat Politècnica de Catalunya and has collaborated in the development of «SPAD-T». All of them have relevant contributions to the statistical analysis of texts.

Eduard Bonet
Esade (Universitat Ramon Llull)

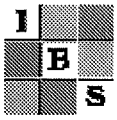
Ressenyes d'activitats institucionals

Sociedad Española de Biometría



Región Española
de la Sociedad Internacional de Biometría

Spanish Region
of the International Biometric Society



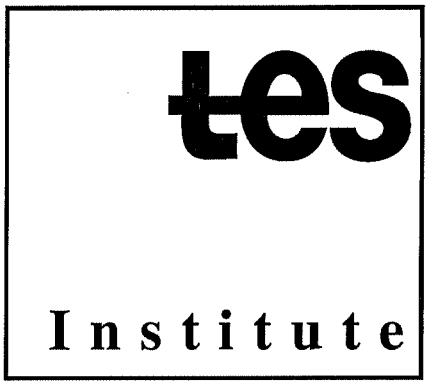
http://www.iata.csic.es/ibsresp

La **Sociedad Española de Biometría/Región Española de la Sociedad Internacional de Biometría** (abreviadamente **SEB** o ***REsp***) tiene como objetivos promover, impulsar y difundir el desarrollo y la aplicación de los métodos matemáticos y estadísticos a la biología, medicina, psicología, farmacología, agricultura y otras ciencias afines (ciencias relacionadas con los seres vivos). Cualquier profesional o alumno de estas disciplinas puede ser miembro de la SEB.

Consejo Directivo

Presidente:	María Jesús Bayarri García (Medicina)
Vicepresidente:	Guadalupe Gómez Melis (Biología)
Secretario y Tesorero:	Fernando López Santoveña (Agronomía)
Vocales en calidad de Miembros del Consejo de la IBS:	Guadalupe Gómez Melis (Biología) Emilio A. Carbonell Guevara (Agronomía)
Vocales:	Amparo Oliver Germes (Psicología) Antonio López Quílez (Medicina) Juan Ferrándiz Ferragud (Biología) Purificación Galindo Villardón (Medicina) Eduardo García Cueto (Psicología) José Luis González Andújar (Agronomía) Alex Sánchez Plá (Biología)
Corresponsal de la REsp en el «Biometric Bulletin» de la IBS:	María Luz Calle Rosingana

IX CONFERENCIA ESPAÑOLA DE BIOMETRÍA
se celebrará en A CORUÑA los días 28, 29 y 30 de mayo de 2003
más información en:
http://www.udc.es/dep/mate/biometria2003/index.htm



The TES Institute
Training of European Statisticians

GENERAL INTRODUCTION

The TES Institute is a non-profit making association of 17 European National Statistical Institutes and the University Centre of Luxembourg. It offers a post-graduate vocational training programme for statisticians and other groups working in the statistical environment. The TES Institute today boasts a solid foundation of statistical training experience, built up since the beginning of the nineties.

Our mission is to provide world-wide cutting edge lifelong international training for professional statisticians. The training emphasises the inter country know-how transfer and focuses on the dissemination of best practices in terms of providing statistics compliant with both European and International Statistical System standards.

The training programmes of the TES Institute are designed for both public and private sector statisticians in the broadest sense of the word i.e. university graduates of any discipline involved in statistical work within Statistical Institutes, European Institutions, Government Agencies, National or private Banks, Enterprises, ...

Training methods are problem-oriented and based on a twofold track approach combining both teaching of theoretical concepts and practical sessions using real life cases. Moreover, the «learning-by-doing» process provides an opportunity to respond in an effective way to recent scientific and technological developments as well as new requirements from the labour market.

The training programmes offered by the TES Institute provide both theoretical and practical background but the courses have all a very strong applied character. These programmes also offer participants the opportunity to meet colleagues form all over Europe and other countries since the TES Institute has extended its activities to the Central European, Mediterranean Basin and TACIS countries.

The above characteristics represent the basic conditions to acquire sharper competence in their work environment and highlight the European dimension of their activity.

After ten years of existence, the programme became entire part of the statistical world. For the time being, around 500 participants coming from more than 30 countries are trained every academic year. Such an interest is mainly due to the large number of courses on offer. Indeed, the TES portfolio comprises more than 80 courses of short duration all at post-graduate level.

The TES Institute offers a coherent set of interrelated courses in the following domains:

- Official Statistics: General Issues (OSG)
- Official Economic Statistics (ECO)
- Official Social Statistics (SOC)
- Data Collection and Survey Methodology (DAT)
- Applied Statistical Analysis (ASA)
- Publication and Dissemination of Statistics (PDS)
- Management in a Statistical Institute (MSI)
- Statistical Information Systems (SIS)

To reach the widest audience possible, the TES Institute currently offers courses in English, French, German, Arabic and Russian.

After a few years of co-operation with the Central European countries, the TES Institute has recently extended the co-operation to the MEDSTAT and TACIS region. Such an internationalisation is the direct result of the growing importance of training as a part of the current technological and intellectual development. Therefore, as far as statistics and economics are concerned, it is of the utmost importance to extend the best national practices to an international level. It is obvious that the TES programmes should be considered as a complement and not a substitute to the training provided at national level.

In brief, one may say that by offering training opportunities that are complementary to the ones provided at national level, the TES Institute is offering a new approach of the subsidiarity concept.

TRAINING

The Core Programme of Eurostat

The Core Programme 2003 offering 28 courses will run from January to November 2003. The overview of the courses can be found at the end of the present paper.

The Special Courses Programme

Beside the execution of the subsidised annual vocational training programme –which is only one of our many activities– the TES Institute also offers Special Courses that may be repetition of courses in the Core Programme or tailor-made and organised at the request of a country or a group of countries.

In this context, the TES Institute is active through the PHARE programme for Central European Countries, through the MEDSTAT programme in the countries of the Mediterranean Basin and through the TACIS programme in a number of CIS countries.

CONSULTING

In addition to the vocational training activities, we will continue to develop and extend our consulting activities in the above mentioned regions and elsewhere in order to maintain and solidify our position in the market of professional training and staff development for statisticians. These consulting activities consist in either *competence building* by the training of «in-country» trainers with multiplicative effects of the training or *institutional building* by the development of training centres, improvement of vocational training system and curricula. For the time being, the TES Institute has been involved in consulting activities with the Russian Federation, Ukraine and Kazakhstan.

RESEARCH

In the context of the 5th Framework Programme of the Commission, the TES Institute is participating in a consortium which will focus on the creation and the development of on-line training courses. On the basis of possibilities offered by existing tools, the VL-CATS project will provide access over the Internet, to teaching and educational material with a special focus on Official Statistics. The main objectives of VL-CATS Project are:

- To create a virtual library of all reference material available in Official Statistics;
- To define and implement standards for the development of distance courses in the various fields of Official Statistics;

- To develop a controlled environment that will support the virtual library and give selective access to training courses;
- To develop a service for update and quality assurance of VL-CATS.

The VL-CATS Project foresees the design and the implementation of the following courses:

- Sampling techniques
- Official Statistics and the European Statistical System
- Confidentiality
- Time Series Analysis
- Statistical Modelling
- Strategic Management

The TES Institute is also involved in a consortium which started in January 2001 the implementation of the ASSO project (Analysis System of Official Symbolic Data). The general objective of the ASSO project is to design methods, methodology and software tools for extracting knowledge from multidimensional complex data coming from very large databases in Statistical Offices and Administrations. The aim of the project is to use Symbolic Data Analysis to solve problems of Statistical Offices, in particular problems of confidentiality and missing data.

The previous project SODAS has developed a software prototype for the building and the analysis of symbolic data and has shown the interest of such methods for administrative data. ASSO will improve the first project in order to render it more operational and attractive following users requests. It will also add new innovative methods and demonstrate that these techniques meet needs of Statistical Offices with a particular interest given to business registers, «new economy» data, environment and employment.

PUBLICATIONS

The TES Institute is regularly publishing articles on its current activities in periodic Newsletter of several statistical institutes and some statistical journals as *Qüestiió* (Quaderns d'Estadística i Investigació Operativa), edited by the Institut d'Estadística de Catalunya (Idescat).

The TES Institute has started the production of TES Manuals on subjects covered by the vocational training programme:

- The first manual available at the TES Institute presents «The role of statistics in a Democracy».
- The second manual to be published soon will cover the Index numbers for spatial and temporary comparisons.

Further manuals will cover topics of sampling techniques, seasonal adjustment methods.

TES NEWSLETTER

The TES Institute will issue its Newsletter «Facts & Visions» on a quarterly basis. This newsletter should be considered as an important information tool between the TES Institute and its partners from all over Europe focusing on matters related to training and staff development in the statistical world. So, may we encourage you to send us any papers or information that might be of interest to be published.

GENERAL INFORMATION

For further details on any of the above-mentioned topics, please contact directly Ms. Valérie Vandewalle (*Head of Curriculum Development*):

Phone: (352) 29.85.85.34
Fax: (352) 29.85.29
E-mail: vvandewalle@tes-institute.lu

European Statistical Training Programme

Core Programme 2003

Codes	Titles	Course Leader	Location	From
DAT-102	Survey Non-response: Reduction, Weighting and Imputation	Lynn	Neuchâtel	27-31 January 2003
ECO-204	European System of Accounts (ESA 95)-Goods and Services	Konijn	Luxembourg	27-29 January 2003
ECO-155	Road Freight Transport Statistics	Collings	Luxembourg	3-5 March 2003
OSG-001	The European Statistical System	Nanopoulos	Luxembourg	12-14 March 2003
ASA-201	Seasonal Adjustment Methods	Maravall	Luxembourg	24-28 March 2003
PDS-101	Towards a User-friendly Statistical Reporting	Wright	London	31 March-2 April 2003
ECO-101	Nomenclatures, Classifications and their Harmonisation	Langkjaer	Luxembourg	31 March-3 April 2003
SOC-104	Concepts and Measurements of Inequality and Poverty	Guzman	Libourne	31 March-4 April 2003
PDS-103	Techniques of Electronic Data Dissemination	Argueso	Madrid	7-10 April 2003
PDS-001	Basic Principles of Publication and Dissemination of Statistical Products	Swires-Hennessy	London	7-11 April 2003
SIS-001	An Overview of Statistical Confidentiality in Official Statistics	Nanopoulos	Luxembourg	7-9 May 2003
ECO-107	Regional Accounts: Theory and Practice	Behrens	Luxembourg	12-14 May 2003
PDS-105	Marketing and Sales of Statistical Products and Services	Munch Haagensen	Copenhagen	12-14 May 2003
DAT-105	Introduction to the Use of Administrative Sources for Statistical Purposes	Vale	Helsinki	21-23 May 2003
MSI-150	Quality Management in Statistics	Bergdahl	Luxembourg	2-4 June 2003
SIS-250	Statistical Metadata	Hustoft	Oslo	2-4 June 2003
DAT-002	Sampling Techniques and Practice	Smith	Southampton	16-27 June 2003
DAT-103	New Advanced Technologies for Data Collection	Kunzler	Luxembourg	8-10 September 2003
ECO-103	Enterprise Statistics	Willeboordse	Barcelona	15-19 September 2003
SOC-003	Living Conditions, Social Indicators Social Reporting	Everaers	Heerlen	15-24 September 2003
SOC-102	Labour Costs Statistics	Clarke	Luxembourg	22-24 September 2003
ECO-203	European System of Accounts (ESA 95) - Financial Accounts	Glatzel	Luxembourg	6-8 October 2003
ECO-001-E	National Accounts Statistics in Practice	van Nunspeet	Voorburg	13-24 October 2003
ECO-201	Environmental Expenditure Statistics and Accounts	Johansson	Luxembourg	20-23 October 2003
DAT-205	Design of Experiments within Surveys	van den Brackel	Heerlen	3-5 November 2003
SIS-201	Introduction to Statistical Disclosure Control Methods	de Wolf	Voorburg	3-6 November 2003
SIS-152	New Tools and Methods of Data Transmission for Statisticians	Vlietinck	Luxembourg	10-13 November 2003
ECO-124	Short-term Indicators	van den Bos	Luxembourg	17-21 November 2003

La Sociedad de Estadística e Investigación Operativa y el Departament de Matemàtica de la Universitat de Lleida, anuncian la celebración del **27 Congreso Nacional de Estadística e Investigación Operativa** que tendrá lugar en Lleida, durante los dias **del 8 al 11 de abril del 2003**, en el **Edificio del Rectorado de la Universidad de Lleida**, Plaza Víctor Siurana, 1.

CONTRIBUCIONES Y SESIONES

Por lo que se refiere al programa científico, en esta ocasión, además de las sesiones de comunicaciones orales, pósters y conferencias plenarias, tanto del ámbito de la Estadística como de la Investigación Operativa, estamos organizando unas sesiones temáticas entorno a la **Calidad en la Industria y los Servicios** y a la **Estadística Oficial**, que se desarrollarán en paralelo a las sesiones de comunicaciones orales y que tratarán, entre otros, temas como «*Modelos de gestión*», «*Planificación de la producción*», «*Control de procesos*», «*Logística*», «*Experiencias empresariales en control y mejora*», «*Estimación en pequeñas áreas*», «*Estadística de la Sociedad de la Información*», *etc.*

Las contribuciones de los participantes podrán exponerse oralmente, en sesiones temáticas paralelas, o en forma de póster. Las sesiones temáticas en las que podrán presentarse trabajos serán las siguientes:

Análisis de Supervivencia
Aplicaciones de la Estadística i la I.O.
Control y Mejora de la Calidad
Decisión Multicriterio
Didáctica de la Estadística
Diseño de Experimentos
Econometría
Estadística Computacional
Estadística Espacial
Estadística Oficial
Fuzzy Sets
Inferencia y Decisión
Inferencia no Paramétrica
Localización
Métodos Bayesianos
Métodos Multivariantes
Modelos y Aplicaciones en I.O.
Muestreo
Predicción Dinámica
Probabilidad
Procesos Estocásticos
Programación Matemática
Series Temporales
Simulación
Técnicas de Remuestreo
Teoría de Colas
Teoría de Juegos
Teoría de la Decisión
Teoría de la Información
Otra:.....

SESIONES SOBRE ESTADÍSTICA OFICIAL

A lo largo del día 9 de abril se llevarán a cabo las sesiones dedicadas a la Estadística Oficial, organizadas conjuntamente por el Instituto Nacional de Estadística y el Institut d'Estadística de Catalunya (Idescat). El programa de esta sesión cuenta con tres ponencias invitadas, bajo la coordinación del moderador y, a continuación, la exposición de contribuciones libres que se sometan en este ámbito temático.

Sesión 1: Estimación en Áreas Pequeñas
Moderador: Jorge Saralegui (INE)

- *El problema de la estimación en áreas pequeñas para la estadística oficial. Recientes progresos en España*
  Montserrat Herrador (INE)
  Jorge Saralegui (INE)
- *La estimación de pequeñas áreas en la estadística oficial sobre empresas*
  Àlex Costa (Idescat)
  Albert Satorra (Universitat Pompeu Fabra, Barcelona)
  Eva Ventura (Universitat Pompeu Fabra, Barcelona)
- *Modelos estándard para estimaciones en áreas pequeñas. Extensión a diseños complejos*
  Domingo Morales (Universidad Miguel Hernández, Elche)

Sesión 2: Estadística de la Sociedad de la Información
Moderador: Enric Ripoll (Idescat)

- *Indicadores de la nueva economía en la perspectiva de los hogares y de las empresas*
  Fernando Cortina (INE)
  Fernando Celestino (INE)
- *Modelos de comportamiento en el uso de las TIC de los internautas catalanes*
  Montserrat Guillén (Universitat de Barcelona)
  Àlex Costa (Idescat)
  Maribel García (Idescat)
- *Modelización econométrica de la difusión regional de las nuevas tecnologías*
  Antonio Pulido (CEPREDE-Universidad Autónoma de Madrid)

INSCRIPCIÓN

La inscripción debe **realizarse a través de la hoja de inscripción electrónica que se encuentra en la página Web del Congreso.** En caso de imposibilidad de utilizar la hoja de inscripción electrónica, se admitirá también el envío, por correo postal o Fax, del Boletín de Inscripción que se adjunta.

En cualquier caso será **imprescindible** enviar por correo postal el **resguardo de la transferencia bancaria**, además de los justificantes acreditativos y fotocopia del DNI para los estudiantes y titulados en paro a:

27 Congreso SEIO
Departament de Matemàtica - Universitat de Lleida
Campus Cappont. C/ Jaume II, 69
25001 Lleida

Fax: 973 702716
e-mail: **seio2003@matematica.udl.es**
Web: **http://www.matematica.udl.es/seio2003**

ALOJAMIENTO Y RESERVAS

Las **reservas de alojamiento** podrán tramitarse a través de la Central de Reservas **INDÍBIL**. Para ello es necesario rellenar el **boletín de alojamiento,** que se encuentra en la página Web (sección Alojamiento) y remitirlo a: Central de Reservas INDIBIL . C/ Major, 31 Bis (25007) Lleida, o bien enviarlo por fax a **(973) 248120**, o por e-mail a **indibil@paeria.es**.

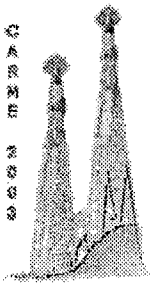
Una vez confirmada la reserva por parte de la Central, será imprescindible enviar por fax, el resguardo de la transferencia bancaria o autorización firmada de cargo a su tarjeta de crédito.

Aquellos congresistas que deseen bonos de descuento para Renfe, deberán solicitarlos por e-mail a **seio2003@matematica.udl.es**

International Conference on

CORRESPONDENCE ANALYSIS AND RELATED METHODS

Universitat Pompeu Fabra, Barcelona,
29 June – 2 July 2003



SECOND CALL FOR PAPERS

The international conference on Correspondence Analysis and Related Methods (CARME 2003) will be held at the the new campus of the Universitat Pompeu Fabra, close to the Olympic port and beaches of Barcelona, from 29 June to 2 July 2003. The organisers, Michael Greenacre (Barcelona) and Jörg Blasius (Bonn), assisted by local organisers Anna Cuxart, Clara Riba and Frederic Udina, look forward to welcoming you to the fourth in a series of successful conferences centred around the analysis of categorical data. Previous conferences, held at the Zentralarchiv für Empirische Sozialforschung in Cologne, were:

(1991) Correspondence Analysis in the Social Sciences
(1995) Visualisation of Categorical Data
(1999) Large Scale Data Analysis

The timing of the conference, at the end of June 2003, is designed to optimise your chances of visiting Barcelona, the Mediterranean coastline and the spectacular interior of Catalonia before the tourist season, so take the opportunity of extending your stay by a few days to enjoy the local scenery and gastronomy.

CARME 2003 will spotlight the very latest research in correspondence analysis and discuss future developments. We aim to bring together theoretical and applied researchers in all the areas where correspondence analysis is currently used, notably in the social sciences, ecology and environmental science, biomedical and health sciences, archaeology, marketing and management.

Themes of the conference include all forms of correspondence analysis: simple, multiple, joint, multiway, canonical and nonsymmetrical correspondence analysis, as well as the related fields of dual and optimal scaling, homogeneity analysis, multidimensional scaling and biplots of categorical data and compositional data. Interdisciplinary contributions are particularly encouraged.

For more details, including names of invited speakers, see:

http://www.econ.upf.es/carme

You can add your name to the mailing list or contact the conference organisers at the e-mail address:

carme2003@upf.es

Proposals for papers may be made before 31 March 2003, preferably by e-mail to the above address, or by post to:

Michael Greenacre
Department of Economics and Business
Universitat Pompeu Fabra
Ramon Trias Fargas, 25-27
08005 Barcelona
Spain

The document should be no longer than one page length, and should include the following:

- title
- author(s)
- affiliations
- postal address
- e-mail address
- abstract
- references (optional)

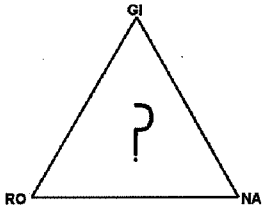
The abstract should contain as little mathematical formulation as possible and should be submitted in Word or plain text format. An example in .rtf (rich text) format (written in Wordpad and readable by all versions of Word), can be downloaded from the website www.econ.upf.es/carme.

Contributors will be notified about acceptance of their proposals by 30 April 2003.

At the conference the organisers will make a selection of papers to be included eventually in a group publication, similar to the previous successful books:

- *Correspondence Analysis in the Social Sciences.* Academic Press, London. 1994.
- *Visualization of Categorical Data.* Academic Press, San Diego. 1998.

COMPOSITIONAL DATA ANALYSIS WORKSHOP



CODAWORK'03
OCTOBER 15-17, 2003
GIRONA, SPAIN

*

FIRST ANNOUNCEMENT
AND
CALL FOR PAPERS

*

VISIT OUR WEBSITE
http://ima.udg.es/Activitats/CoDaWork03/

ABOUT CODAWORK'03

The Workshop on Compositional Data is intended as a forum for discussion of hot points related to the statistical treatment and modelling, as well as applications and interpretation, of compositional data. The goal of such discussions is to get some insight into the most appealing future lines of research in the field.

In order to attain this general but clear goal, the Network on Compositional Data, headed by the Girona Compositional Data Group, tries to meet with a significant number of specialists, users and interested people to collect critical contributions and open a stimulating brainstorming.

The contributions and discussions are intended to move around the following points:

- Geometry and statistics in the simplex
- Design of teaching and computing tools
- Applications to archaeology
- Applications to geology and environment
- Other fields of application

The workshop will consist in 2 hour sessions on the above tentative points. The chair of each session will present a summary of the contributions and will stimulate an open discussion on the concerned topics. Written contributions on these topics, and particularly on applications, are welcome (see abstract submission).

THE COMPOSITIONAL DATA ANALYSIS NETWORK AND THE GIRONA GROUP

Research on compositional data analysis, following ideas introduced in the eighties by J. Aitchison, was introduced in Spain through the Technical University of Catalonia (Barcelona). It initiated at the University of Girona around 1990 with a single PhD-student, growing up to a group which now includes six full time researchers. Present scientific cooperation with other universities and institutions, mainly in Europe (Spain, UK, Italy, Germany) but also in other countries (USA, Argentina, Canada, Russia and Ukraine) has motivated the organization of a network. It is integrated by groups from the following universities: UPC, UB (Barcelona), FUB (Berlin), USF (Firenze), UdG (Girona), UJ (Jena), NTU (Nottingham).

ABOUT GIRONA

More than a thousand years of history watch over this small, harmonious city, which has managed with the utmost of care to preserve the legacy of all the eras and sensitivities that have left their mark on the city: in an old quarter that takes pride in each and every one of the stones that dress its medieval facades, the most notable attractions include the Jewry, an exemplar of the Jewish quarter that is unique in the world, the cathedral, a singular example of Gothic architecture, its museum, with the Creation Tapestry and the Beatus, the wall enclosing the old quarter, and the recently inaugurated Cinema Museum, the only one of its kind in Europe.

HONORARY CHAIR

John Aitchison

WORKSHOP CHAIR

Vera Pawlowsky-Glahn * Carles Barcelo-Vidal

ORGANISER

Network on Compositional Data Analysis
Girona Compositional Data Group
Dept. of Computer Science and Applied Mathematics. University of Girona.

SCIENTIFIC COMMITTEE

Chair: Josep A. Martin-Fernandez (UdG) Michael J. Baxter (Nottingham Trent University)
* Antonella Buccianti (U. Studi di Firenze) *
Jaume Buxeda (UB) * Juan Jose Egozcue (UPC) *
Hilmar von Eynatten (U. Jena)

LOCAL ORGANISING COMMITTEE

Chair: Santi Thio-Henestrosa
Josep Daunis-Estadella * Gloria Mateu-Figueras

WORKSHOP SECRETARY

CoDaWork'03 - workshop Secretary
Universitat de Girona. Av Lluis Santalo. s/n
17071 - Girona, Spain
phone +34-972418417
fax +34-972418792
e-mail: codawork03@ima.udg.es

VISIT OUR WEBSITE

http://ima.udg.es/Activitats/CoDaWork03/

LANGUAGE

The language of the conference will be English.

IMPORTANT DATES

- January 31, 2003
- February 28, 2003 Author notification
- May 31, 2003 Electronic paper due
- June 30, 2003 Early registration

ABSTRACT SUBMISSION

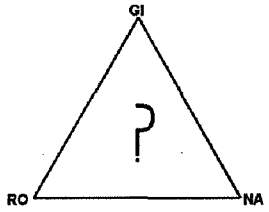
The Organising Committee welcomes abstracts of contributions on the workshop topics. The abstract should include:

- The title of the proposed paper
- Sufficient detail (200-400 words) to allow the Scientific Committee to judge the contents of the proposed paper
- The name, affiliation, mailing address, phone number and e-mail address of the author(s)

Abstracts should be sent to the workshop secretariat by January 31, 2003. Notification of acceptance will be mailed by February 28, 2003. The final electronic paper will be due before May 31, 2003, to allow the Scientific Committee final acceptance. Only papers of participants registered before June 30, 2003 will be included in the proceedings CD.

↦ ONLINE ABSTRACT SUBMISSION
(from October 2002)
http://ima.udg.es/Activitats/CoDaWork03/
TEX-LATEX, MSWord, PDF allowed

Final versions of contributed papers will be sent in electronic format. They will be published in the CoDaWork'03 CD (with ISBN) and will be available in the sessions of the workshop.



Compositional Data Analysis Workshop

CoDaWork'03

October 15-17, 2003

Girona, Spain

Pre-Registration Form

Name:.....

Organisation:.....

.....

Address:.....

.....

.....

Code/City:.....

Country:.....

Phone:.....

Fax:.....

E-mail:.....

Please complete and return this pre-registration form to the workshop secretariat so that your name will be on the mailing list for further announcements.

↔ ONLINE PRE-REGISTRATION

(from October 2002)

http://ima.udg.es/Activitats/CoDaWork03/

Informació pels autors i lectors

Normes per a la presentació d'articles

La revista accepta originals no sotmesos a cap altra revista dins els àmbits de l'estadística, la investigació operativa, l'estadística oficial i la biometria. Els articles poden ser teòrics o aplicats, i poden incloure aspectes computacionals i/o de caire docent, i es publicaran exclusivament en anglès, acompanyats d'un breu resum en català a càrrec de la pròpia revista.

Tots els originals destinats a les seccions temàtiques seran sotmesos sistemàticament a una avaluació prèvia a càrrec d'especialistes independents i/o membres del Consell Editor, llevat dels articles convidats i de les reimpressions d'articles. El resultat de l'avaluació es comunicarà a l'autor principal en el cas que es proposin correccions formals o de contingut.

Quan es tramet un original, la revista emet un justificant de recepció, la data del qual figura com a «data de rebuda» en la publicació de l'article, mentre que la «data d'acceptació» és la data de recepció de la seva versió definitiva.

Per a la presentació d'un article, l'autor l'haurà de trametre en format PDF a l'adreça electrònica de la revista (sort@idescat.es), en la primera pàgina del qual hi farà constar necessàriament el títol, el nom de l'autor o autors, l'afiliació professional i l'adreça completa, un resum de 75-100 paraules, les paraules clau i l'adscripció a la classificació Mathematics Subject Classification 2000 de l'American Mathematical Societ (MSC 2000). Abans de presentar els articles a la revista, s'aconsella els autors que assegurin la correcció lingüística dels textos. Les referències bibliogràfiques es faran indicant el cognom de l'autor seguit de l'any de la publicació entre parèntesis [i.e.: Mahalanobis (1936), Rao (1982b)] i es llistaran alfabèticament al final de l'article. Les referències múltiples d'un mateix autor s'ordenaran cronològicament. Les notes explicatives es numeraran correlativament i apareixeran al peu de la pàgina corresponent. Les taules i figures també es numeraran correlativament en el text i es reproduiran directament dels originals tramesos en cas que no sigui possible fer-ne l'autoedició.

Una vegada acceptat l'article, caldrà que l'autor trameti la versió electrònica definitiva, d'acord amb les instruccions que li seran indicades pel responsable de l'avaluació del seu article. Es recomana que la versió final es trameti mitjançant el processador de textos LaTeX 2ε.

La Secretaria de la revista posa a disposició dels autors que ho sol·licitin plantilles en format LaTeX 2ε per fer-ne l'edició, així com les referències adients de la classificació MSC2000 de l'American Mathematical Society. Per a més informació sobre l'ús i els criteris per a l'adopció d'aquesta classificació es pot consultar directament la Mathematics Subject Classification 2000 (MSC 2000) en el web de l'Institut d'Estadística de Catalunya.

Statistics & Operations Research Transactions (SORT)
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
Tel.: +34-93 412 09 24 o +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.es

Guidelines for submiting articles

The journal accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as «reception date». The date in which its final version is received will be the «acceptance date».

To submit an article, the author must send it in PDF format to the electronic address of journal (sort@idescat.es). The title page must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75-100 words) followed by the keywords and MSC2000 classification of the American Mathematical Society. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English. Bibliographic references within the text must follow this format: author surname followed by the year of publication in parentheses [i.e., Mahalanobis (1936), Rao (1982b)], with complete reference citations listed alphabetically at the end of the article. Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using the LaTeX 2ε.

In any case, upon request the journal secretary will provide authors with LaTeX 2ε templates and appropiate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya.

Statistics & Operations Research Transactions (SORT)
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
Tel.: +34-93 412 09 24 o +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.es

Butlleta de subscripció a *Statistics & Operations Research Transactions (SORT)*

Nom i cognoms _____

_____

Empresa/Institució _____

_____

Adreça _____

_____

Codi postal _____ Ciutat _____

Tel. _____ Fax _____ NIF _____

Correu electrònic _____

Data _____

Signatura

Desitjo subscriure'm a ***Statistics & Operations Research Transactions (SORT)*** per a l'any 2003 (volum 27)

Preu de subscripció vigent:

– Estat espanyol: 22 €/any (IVA inclòs)
– Estranger: 25 €/any (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:

Statistics & Operations Research Transactions (SORT)
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Preu de números solts (actuals i endarrerits):

– Estat espanyol: 9 €/exemplar (IVA inclòs)
– Estranger: 11 €/exemplar (IVA inclòs)

Exemplar per a l'entitat bancària

Autorització de domiciliació bancària per al pagament de les subscripcions anuals de ***Statistics & Operations Research Transactions (SORT)***

El/la sotasignat/ada _____

autoritza el Banc/Caixa _____

Adreça _____

Codi postal _____ Ciutat _____

a abonar les subscripcions a ***Statistics & Operations Research Transactions (SORT)*** amb càrrec al seu compte

número ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Data _____

Signatura

Statistics & Operations Research Transactions (SORT)
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Novetats editorials en matèria estadística de la Generalitat de Catalunya
gener-desembre 2002

- **Estadística de biblioteques 2000**
  Característiques bàsiques
  Idescat
  Gener 2002, 8,50 €, 131 pp.
  ISBN 84-393-5614-5

- **Mercat de treball 2001**
  Ampliació de resultats anuals de l'enquesta de població activa
  Idescat
  Maig 2002, 11,00 €, 326 pp.
  ISBN 84-393-5702-8

- **Classificació catalana de productes per activitats (CCPA-96)**
  Adaptació de la CNPA-96
  Idescat
  Març 2002, 12,50 €, 435 pp.
  ISBN 84-393-5637-4

- **Localització de l'activitat econòmica 1999**
  Empreses, professionals, establiments i superfícies
  Idescat
  Maig 2002, 10,30 €, 365 pp.
  ISBN 84-393-5738-9

- **Estadístiques de la societat de la informació. Catalunya 2001**
  Departament d'Universitats, Recerca i Societat de la Informació
  Secretaria de Telecomunicacions i Societat de la Informació
  Observatori de la Societat de la Informació
  2002, 9,02 €, 176 pp.
  ISBN 84-393-5429-0

- **Migracions internacionals i població jove de nacionalitat estrangera a Catalunya**
  Departament de la Presidència
  Secretaria General de Joventut
  2002, 13,00 €, 140 pp.
  ISBN 84-393-5511-4

- **Cens agrari 1999. Vol. 1**
  Aprofitament de la terra i ramaderia
  Idescat
  Juny 2002, 11,50 €, 238 pp.
  ISBN 84-393-5754-0

- **Moviment natural de la població 2000**
  Dades comarcals i municipals
  Idescat
  Juny 2002, 8,70 €, 106 pp.
  ISBN 84-393-5755-9

- **Classificació catalana d'educació 2000 (CCED-2000).**
  Adaptació de la CNED-2000
  Idescat
  Juny 2002, 10,00 €, 140 pp.
  ISBN 84-393-5761-3

- **Localització de l'activitat econòmica 2000**
  Empreses, professionals, establiments i superfícies
  Idescat
  Maig 2002, 10,30 €, 365 pp.
  ISBN 84-393-5739-7

- **Projeccions de llars de Catalunya 2010**
  Comarques i àmbits del Pla territorial
  Idescat
  Juny 2002, 9,50 €, 310 pp.
  ISBN 84-393-5762-1

- **Les dones a Catalunya: dades estadístiques**
  Departament de la Presidència
  Institut Català de la Dona
  2001, 22,80 €, 246 pp.
  ISBN 84-393-5571-8

- **Anuari de dades meteorològiques 2000**
  Departament de Medi Ambient
  Direcció General de Qualitat Ambiental
  2001, 10,22 €, 112 pp.
  ISBN 84-393-5467-3

■ **Estadística de finançament i despeses de l'ensenyament privat Curs 1999-2000**
Idescat
Juliol 2002, 8,50 €, 184 pp.
ISBN 84-393-5797-4

■ **Cens agrari 1999**
Vol. 2 Maquinària, emmagatzematge, comercialització, mà d'obra, marge brut i orientació tecnicoeconòmica
Idescat
Setembre 2002, 11,50 €, 240 pp.
ISBN 84-393-5853-9

■ **7. Material de transport**
Àngel Hermosilla Pérez i Natàlia Ortega Gómez
Idescat, Departament de Treball, Indústria Comerç i Turisme
Octubre 2002, 7,20 €, 125 pp.
ISBN: 84-393-5888-1

■ **Anuari estadístic de Catalunya 2002**
Idescat
Octubre 2002, 18,00 €, 760 pp.
ISBN 84-393-5923-3

■ **Xifres de Catalunya 2002**
(català, castellà, anglès, francès, alemany)
quadríptic

■ **Moviments migratoris 2000 Dades comarcals i municipals**
Idescat
Juliol 2002, 9,60 €, 184 pp.
ISBN 84-393-5798-2

■ **Estadística, producció i comptes de la indústria 2000**
Idescat
Setembre 2002, 9,00 €, 415 pp
ISBN 84-393-5854-7

■ **Cens agrari 1999**
Vol. 3 Estructura del sector agrari
Idescat
Octubre 2002, 11,50 €, 225 pp.
ISBN 84-393-5928-4

■ **Anuari estadístic de Catalunya 1992-2002 CD-ROM**
Idescat
Desembre 2002, 12,00 €
ISBN 84-393-5972-1

■ **Estadístiques culturals de Catalunya 2002**
Col·lecció Estadístiques Culturals, 5
Departament de Cultura
Gabinet Tècnic
2002, 12,00 €, 166 pp.

LLIBRERIES DE LA GENERALITAT

Barcelona
Rambla dels Estudis, 118 (tel. 93 302 64 62)
llibrbcn@correu.cattel.com

Girona
Gran Via de Jaume I, 38 (tel. 972 22 72 67)
llibrgi@ibernet.com

Lleida
Rambla d'Aragó, 43 (tel. 973 28 19 30)
llibrlle@ibernet.com

Madrid
Blanquerna. Llibreria catalana.
Serrano, 1 (tel. 91 431 00 22)
blanquerna@nauta.es

PUNT DE VENDA

Puigcerdà
Plaça del Rec, 5 (tel. 972 88 05 14)

VENDA PER CORREU

Apartat 2800, 08080 Barcelona
eadop@correu.gencat.es
Plaça del Rec, 5 (tel. 972 88 05 14)

Publicacions de la Generalitat. Apartat de correus 2800, 08080 Barcelona

Nom i cognoms _____

Empresa / Institució _____

Professió _____ E-mail _____

Adreça _____

Població _____ CP _____

NIF / DNI _____ Telèfon _____

Desitjo rebre els volums

_____

_____

○ Carregueu l'import a la meva targeta de crèdit Signatura

☐ American Express ☐ 6000

☐ Master Charge ☐ Visa

○ Contra reemborsament

Núm. de targeta |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|

Data de caducitat |_|_|_|_|

DOGC