Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 1995, volum 19, núm. 1, 2, 3
Segona època

Entitats patrocinadores:

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

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 1995, volum 19, núm. 1, 2, 3

Sumari

PAPER ECOLÒGIC

Editorial volum 19 (1995), Números 1, 2 i 3

Aquest volum únic està dedicat al professor C.R. Rao amb motiu de la seva investidura com a *doctor honoris causa* per la Universitat de Barcelona, en acte celebrat el 13 de març d'enguany amb la presència de les autoritats acadèmiques corresponents i d'un nombrós públic. De fet, la noble aula Paranimf de la Universitat estava plena de gom a gom. No era per a menys: professors, investigadors, estudiants i públic en general tenien l'oportunitat de veure i escoltar un dels estadístics més importants del moment.

Recentment, el professor B. Efron citava els homes que feren possible la maduresa de l'estadística, a saber: Fisher, Neyman, Pearson, Kolmogorov, Hotelling, Wald, Cramer i Rao. Tots aquests noms són autèntics mites en el camp de l'estadística, però hem de destacar que el professor C.R. Rao era, cronològicament, 25 anys més jove que qualsevol dels altres! Des d'aquesta perspectiva, i amb motiu del 75è. aniversari del professor C.R. Rao enguany (Madras, 20 de setembre de 1920), la revista **Qüestiió** vol retre també un merescut homenatge a l'insigne investigador i professor.

En aquest context, a més dels articles originals ordinaris, el Volum 19 conté els discursos pronunciats pels professors C.M. Cuadras i C.R. Rao durant l'acte d'investidura anteriorment esmentat. D'entre els articles ordinaris, hom disposa igualment d'un interessant article del mateix C.R. Rao en el que s'ofereix una visió generalitzada de dues tècniques estadístiques multivariants: l'anàlisi canònica i la representació de dades categòriques (anàlisi de correspondències). Finalment, la Secció Docent inclou un article del professor C.R. Rao, text que forma part del seu llibre *Estadística y Verdad*, una amena dissertació de l'estadística en el seu context social. Amb la publicació d'aquest notable conjunt de textos de l'homenatjat (gràcies a l'amable autorització del Servei d'Informació i Publicacions i de l'editorial PPU de la Universitat de Barcelona), **Qüestiió** posa a l'abast dels seus lectors un ampli recull força actualitzat de les seves consideracions sobre el paper i reptes amb què s'enfronta l'estadística avui dia.

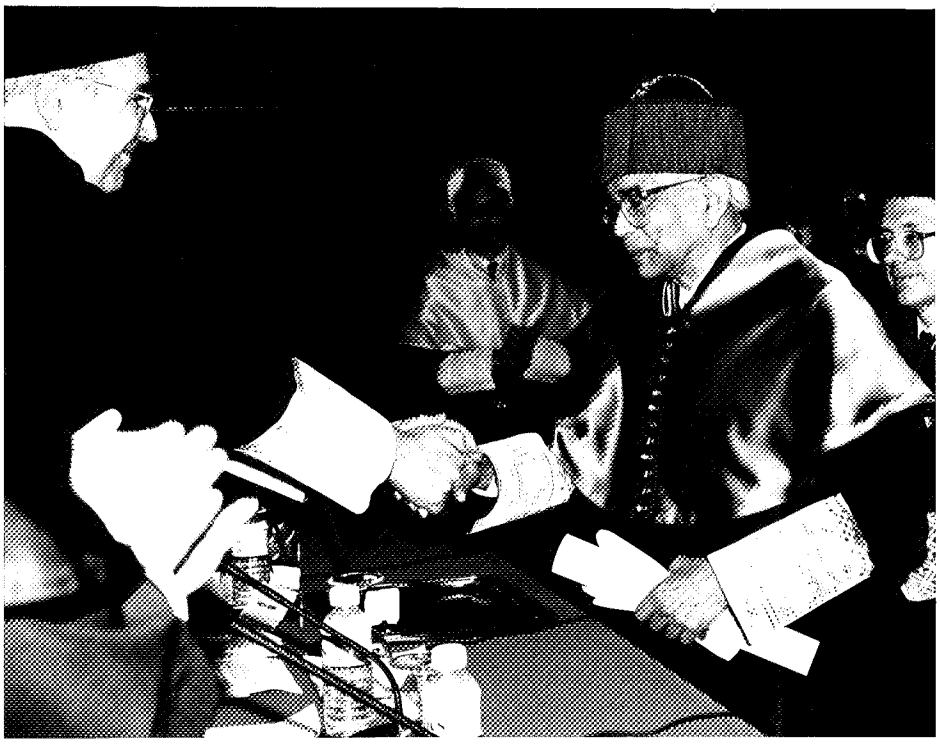
Per la seva banda, la secció d'Estadística Oficial s'adhereix igualment a l'homenatge al professor Rao en aquest volum extraordinari que li dedica **Qüestiió**. Ho fa amb la publicació de dos articles sobre el tema de la "metainformació estadística", és a dir, la informació sobre la informació estadística, temàtica que ha esdevingut en un aspecte estratègic de l'estadística oficial, apassionant i complex a la vegada.

Així, l'article d'Isidre Canals (de l'Institut d'Estadística de Catalunya), titulat "Introducció als sistemes de metainformació estadística", ofereix una visió general completa i n'analitza diversos aspectes: presenta la metainformació des del punt de vista de l'ús i de la mateixa producció estadística, tant des de la perspectiva de les microdades com de les macrodades; també mostra la construcció de diversos sistemes de

metainformació, realitzats per diverses institucions i amb diversos suports informàtics, tot i detallant-hi les experiències internacionals que es desenvolupen, amb els seus problemes de coordinació, coherència dels continguts i formats, i la compatibilitat dels sistemes d'acció. Finalment, procura una informació força exhaustiva de les conferències i grups de treball que actualment estudien el tema.

L'article titulat "Guidelines for the modelling of statistical data and metadata", del notable especialista suec Bo Sundgren (Statistics Sweden Institute), és una referència fonamental en aquest àmbit. Concretament, aquest treball és una versió de diversos estudis i informes que l'autor ha presentat en diverses reunions europees, moltes d'elles auspiciades per l'Oficina estadística de la Unió Europea (Eurostat). Bo Sundgren és mundialment reconegut com un iniciador i desenvolupador de l'estudi de la metainformació estadística, autor de la introducció del terme "metadata" fa més de vint anys i, amb el seu treball constant, ha consolidat i prestigiat aquesta àrea d'estudi, que ha adquirit un reconeixement científic i pràctic internacional, especialment al si dels instituts i organismes oficials d'estadística.

Barcelona, novembre de 1995



Acte de lliurament al professor C.R. Rao del diploma de doctor honoris causa de la Universitat de Barcelona, per part de l'Excm. i Mgfc. Rector d'aquesta universitat, Senyor A. Caparrós, el 13 de març del 1995.

DISCURS DE PRESENTACIÓ
DEL PROFESSOR
CARLES CUADRAS

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 7-12, 1995

C.R. RAO: UNA VIDA DEDICADA A L'ESTADÍSTICA

Carles M. Cuadras*
Departament d'Estadística

Excel·lentíssim i Magnífic Sr. Rector,
Excel·lentíssims Vice-rectors,
Il·lustríssims Degans, Autoritats,
Senyores i Senyors:

És per a mí un gran honor fer la presentació del Professor C.R. Rao com a Doctor Honoris Causa per aquesta universitat, i motiu d'agraïment el que la Facultat de Matemàtiques hagi proposat i el Comitè Acadèmic hagi acceptat aquesta proposta. L'honor és, per a mi, doble: com a catedràtic d'estadística i com a president de la Regió Espanyola de la International Biometric Society, de la que el Professor C.R. Rao va ésser president internacional i que actualment és membre honorari . És també motiu d'emoció parlar-vos en el Paranimf, lloc on vaig tenir el primer contacte amb aquesta casa l'any 1963, i recordar al meu pare, que hi va començar la carrera de Matemàtiques el 1928, i que va tenir al Prof. Josep M. Orts, fundador del Seminari Matemàtic, com a professor d'estadística.

És difícil destacar en breus paraules la immensa labor portada pel Professor C.R. Rao en el camp de l'estadística teòrica i aplicada. Els seus mèrits acadèmics són prou eloqüents:

Té dos doctorats (Ph.D., Sc.D., 1948, 1965) per la Universitat de Cambridge, i 15 doctorats honorífics. És *Fellow o Honorary Fellow* de nombroses institucions i societats internationals, entre las que cal esmentar: Royal Statistical Society, Institute of Mathematical Statistics, International Statistical Institute, American Statistical Association, International Biometric Society. A més ha estat president de les següents societats: Indian Econometric Society (1971-76), International Biometric Society (1973-75), Institute of Mathematical Statistics (1976-77), International Statistical Institute (1977-79) i Forum for Interdisciplinary Mathematics (1982-1984).

*Discurs de presentació del Professor C.R. Rao com a Doctor Honoris Causa per la Universitat de Barcelona. Publicat amb autorització del Servei d'Informació i Publicacions de la Universitat de Barcelona.

Actualment és titular de la Eberly Family Chair of Statistics, Director del Center for Multivariate Analysis a la Pennsylvania State University, *Indian National Professor* i *Adjunt Professor* de la Universitat de Pittsburgh.

Amb ell es dóna el fet irrepetible d'haver estat l'únic alumne de doctorat de R.A. Fisher (1890-1962), una de les personalitats més influents de l'estadística matemàtica i aplicada del present segle.

De Fisher probablement va aprendre a desenvolupar metodologia des de les pròpies dades, més que no pas des d'una perspectiva purament matemàtica.

> "Daily contact with the statistical problems which present themselves to the laboratory worker has stimulated the purely mathematical researches upon which are based the methods here presented. Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow!"
>
> (R.A. Fisher, 1926).

També P.C. Mahalanobis (1893-1972), fundador del Indian Statistical Institute, amb la seva àmplia visió de l'estadística, va tenir una forta influència sobre C. R. Rao. Com ell mateix ens diu en una recent nota biogràfica, P.C. Mahalanobis va ésser físic de formació, estadístic per instint i economista per convicció. Les seves contribucions a l'estadística varen ésser pioneres.

> "Statistics must have a clearly defined purpose, one aspect of which is scientific advance and the other, human welfare and national development"
>
> (P.C. Mahalanobis, 1956).

El Professor C.R. Rao va continuar durant molts anys la labor iniciada per P.C. Mahalanobis, contribuint a que el Indian Statistical Institute sigui una institució de gran prestigi internacional.

A més, les aportacions de C.R. Rao a l'anàlisi estadístic multivariant han contribuït al desenvolupament de l'antropologia en especial, i també de la biologia, geologia, planificació econòmica, demografia, etc.

> "I enjoyed doing my applied work, especially the applications in anthropology, but that work gave rise to the development of methodology in multivariate analysis"
>
> (C.R. Rao, 1987).

Pero és que el Professor C.R. Rao ha fet, a més a més, contribucions fonamentals a l'estadística matemàtica i a virtualment tots els camps importants de l'estadística. Molts resultats, considerats clàssics en estadística, porten el seu nom: Cota de Cramér-Rao, Teorema de Rao-Blackwell, Eficiència de segon ordre de Rao (teoria de l'estimació en petites mostres), Distància de Rao (distància geodèsica per problemes d'inferència estadística), Entropia quadràtica de Rao (anàlisi multivariant), estadístic "score" de Rao (test asimptòtic d'hipòtesi), Cota de Hamming-Rao (ordenacions combinatories anomenades matrius ortogonals), el U-test de Rao.

És també autor d'onze llibres d'estadística, alguns d'ells traduïts als principals idiomes, i font constant de referència en treballs de recerca. La seva labor científica es manifesta en més de 300 publicacions.

Ha estat Editor o membre del Consell Editorial de moltes revistes internacionals, entre les quals destaquem:

Sankhyā Series A and B
Journal of Multivariate Analysis
Communications in Statistics
Journal of Statistical Planning and Inference
Journal of Mathematical Modelling
Stochastic Analysis and Applications
International J. of Mathematical and Statistical Sciences
Combinatorics, Information and Statistical Sciences.

Molt recenment, el Professor B. Efron ens diu:

> "Karl Pearson's famous chi-square paper appeared in the spring of 1900, an auspicious beginning to a wonderful century for the field of statistics. The first half of the century was the golden age of statistical theory, during which our field grew from ad hoc origins, similar to the current state of computer science, into a firmly grounded mathematical science. Men of the intellectual calibre of Fisher, Neyman, Pearson, Kolmogorov, Hotelling, Wald, Cramer and Rao were needed to bring statistical theory to maturity"

(Bradley Efron, *RSS (Royal Statistical Society) News.* **Vol 22, 5**, gener 1995)

La primera relació del nostre departament amb el Prof. C.R. Rao s'inicia amb les tesis doctorals de mí mateix i del Prof. J.M. Oller, la primera sobre anàlisi canònica, seguint idees de Rao, i la segona sobre la distància de Rao. Aquests resultats varen ésser exposats personalment al Prof. C.R. Rao amb motiu de la 33 ISI Session, que

tinguè lloc a Madrid el 1983. D'ença, hom mantingué una relació epistolar, que va cristal·litzar amb una visita del Prof. C.R. Rao al nostre departament el maig de 1987, on va impartir un Seminari d'Anàlisi Multivariant i es varen discutir temes de mutu interès.

L'any 1987, els Professors J.M. Oller i el que us parla, varem tornar la visita viatjant al Center for Multivariate Analysis, University of Pittsburgh, visita que es va repetir, per part meva, el maig de 1992, amb el centre ubicat a la Pennsylvania State University.

La relació del Professor C.R. Rao amb la nostra universitat es reforça per la seva incorporació, des de 1987, al Consell Editor de la revista **QÜESTIIÓ**, editada conjuntament per la Universitat de Barcelona, la Universitat Politècnica de Catalunya i l'Institut d'Estadística de Catalunya.

El punt culminant del nostre contacte va ésser l'organització el 1992 de la Seventh International Conference on Multivariate Analysis, per a donar continuació a les altres sis conferències que havien estat organitzades per P.R. Krisnaiah (1965, 1968, 1972, 1976, 1978 i 1983), mort el 1987. Donat l'alt grau de desenvolupament de l'anàlisi multivariant, el Prof. C.R. Rao va creure adient organitzar-la a tres llocs diferents i amb temes complementaris: University Park (USA), Barcelona i New Delhi.

Aleshores es va posar en contacte amb el Departament d'Estadística demanant l'organització de la sessió de Barcelona, proposta que el nostre Departament va acceptar. Aquesta sessió es celebrà del 21 al 24 de setembre del 1992, i va comptar amb 130 participants, molts d'ells primeres figures, que exposaren conferències i contribucions dins el marc de quatre temes d'actualitat sobre Anàlisi Multivariant.

Com a resultat d'aquesta conferència, es va publicar el llibre *Multivariate Analysis: Future Directions 2*, C.M. Cuadras i C. R. Rao, editors, North-Holland, Amsterdam, 1993.

El Professor C.R. Rao té el singular honor de ser, possiblement, l'únic gran científic viu del que es fa esment en la docència de primer, segon i tercer cicle de l'ensenyament de Matemàtiques. La cota de Cramér-Rao és un resultat fonamental en l'estimació Estadística de paràmetres (assignatura Estadística de segon curs), així com el Teorema de Rao-Blackwell (assignatura Estadística Matemàtica de tercer curs), i les tècniques d'anàlisi canònica i geometria diferencial basades amb la distància de Rao (assignatures d'especialitat i programa de Doctorat del Departament d'Estadística).

Podríem parlar molt i molt bé de la seva categoria humana, però jo em limitaré a destacar tres aspectes de la personalitat científica del Professor C.R. Rao. El primer és el seu punt de vista integrador de l'estadística amb la societat i les altres branques de la

ciència. Ha estat sempre partidari de construir els models després d'entendre les dades. Ha estat obert a totes les tendències, com és ben palés en el seu famós llibre *Linear Statistical Inference and its Applications*. Aquesta visió integradora, en un context ampli i assequible a tothom, cristal·litza en el seu últim llibre *Statistics and Truth*, recentment traduït al castellà, obra que tots els científics, estadístics i no estadístics hauríem de llegir per entendre el fascinant món de l'atzar com a forma d'expressió del coneixement. El segon aspecte és la seva vitalitat científica. He viscut la seva participació activa a molts congressos, i he vist com comentava les comunicacions dels participants i com sabia trobar una resposta encertada a les qüestions que es plantejaven sobre qualsevol tema. I finalment, el seu interés, demostrat en fets, d'ajudar a desenvolupar l'Estadística en els països del tercer món.

Vaig començar a ensenyar a aquesta universitat càlcul numèric el 1968, al temps que treballava com analista al Laboratori de Càlcul, aleshores dirigit pel Prof. Joan Augé. Possiblement vaig ésser el primer programador científic de la Universitat de Barcelona, mèrit més aviat petit al costat de C.R. Rao, que va ésser el primer programador de la India, l'any 1955!

Ben aviat m'en vaig adonar que molts investigadors, dels camps de la biologia, geologia, medicina, psicologia, etc, el que necessitaven era assessorament estadístic. I així, veient i intentant entendre les dades, em vaig introduir a l'estadística, en la seva vesant docent i investigadora, de la mà dels Professors Antoni Prevosti, Francesc d'A. Sales i Joaquim Torrens-Ibern.

No vaig trigar gaire a veure l'abisme que hi havia entre la teoria que havia aprés i les dades reals. Va ser aleshores per a mi vital la lectura de *Advanced statistical methods in biometric research*, obra també del Professor C.R. Rao, en la que teoria, dades i models sintonitzaven en perfecta harmonia.

I aprofitant l'oportunitat de dirigir-me a tant distingida audiència en un marc ple de tradició, voldria fer uns breus comentaris sobre la matèria que el nou Doctor Honoris Causa tant bé representa. L'estadística matemàtica és interessant i necessària, però pot derivar en un simple i meritós entreteniment matemàtic.

> "Not so many years ago it was fashionable to view a statistical analysis as one in which a hypothesis, or some formal model, was stipulated. Parameters were estimated and hypotheses tested. Although this may be the ideal, there has been an increasing acceptance that often much groundwork is required before a model or hypothesis can be formulated".
>
> (J.C. Gower, 1984)

En canvi, l'estadística que té com a finalitat l'estudi i interpretació de les dades, al connectar amb les ciències experimentals i socials, és molt més rica i interessant, i

sovint més difícil i desafiant.. Crec que l'augment del factor d'impacte de les revistes amb contingut metodològic, però amb caire aplicat, sobre les d'orientació més aviat teòrica, és prova d'una tendència general a valorar millor la metodologia aplicable. Tant de bo que les meves paraules en favor de l'estadística en connexió amb les demés ciències, donin el seu fruit algun dia!

> "That is the future of statistics? Statistics is now evolving as a meta-science. Its object is the logic and methodology of the other sciences —the logic of decision making and the logic of experimenting in them. The future of statistics is in communication of statisticians which research workers in other branches of learning it will depend on the way the principal problems are formulated in other fields of knowledge"
>
> (C.R. Rao, 1989).

DISCURS DEL PROFESSOR
C.R. RAO

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 15-19, 1995

STATISTICS: A NATIONAL RESOURCE IN IMPROVING THE QUALITY OF HUMAN LIFE

C.R. Rao*
Pennsylvania State University

Magnificient Rector and
Distinguished Faculty Members,

It is, indeed, a rare and unique honor to receive the highest academic distinction, an Honorary Doctorate from the University of Barcelona which is one of the leading centers of learning in the world. For giving me this award and also the status of an outstanding member of the university body, I am grateful to the Rector and faculty members of the university. This is a great moment of my life, and I shall ever remember and cherish the colorful ceremony held today for awarding the Honorary Doctorate.

My association with the University of Barcelona started about fifteen years ago when I first met Professor C.M. Cuadras and Professor J.M. Oller. We discovered that we have some common interest, and since then I have watched with great admiration the fundamental contributions to statistics that are being made by the Barcelona school of statisticians. Two years ago Professor Cuadras and I organized a series of conferences around the world on *Future Directions of Research in Multivariate Analysis*, one at the Pennsylvania State University, USA, the second at the University of Barcelona, Spain and the third at the University of Delhi, India. These conferences were attended by specialists from all over the world and resulted in the publication or two volumes covering some of the frontier areas of research in multivariate analysis. The award of an Honorary Doctorate degree with the status of an outstanding member of the university body provides me opportunities to make closer contacts with the Spanish statisticians and collaborate in research work on new areas of statistical theory and practice.

I am not a stranger to Barcelona. The city with Gaudi's walls, Domenech i Montaner's concert hall, the Palace of Catalan Music and the wide avenues buzzing with activity has fascinated me. It is, indeed, a city worth visiting again and again.

---

*Discurs presentat durant l'Acte d'Investidura de Doctor Honoris Causa.
Publicat amb autorització del Servei d'Informació i Publicacions de la Universitat de Barcelona.

In less than a decade, we will be entering into the third millennium with exciting possibilities and challenges. In the present century, mostly in our life time, there has been phenomenal progress in science and technology and also unprecedented transformation in the socio-economic-political structure all over the globe. All these developments have, no doubt, altered our life style and generally enhanced the well being of our society. Judging from the current trends the developments in science and technology will be more far reaching in the next century than in the present. The exact impact of these developments on individuals and society may be difficult to foresee, but we can have a broad glimpse of the scenario which can guide us in preparing to meet the new challenges that will arise in the future.

Judging from the current trends of scientific activity one can broadly foresee the dominant technologies in the next several decades. The first that comes to my mind is genetic engineering of **biotechnology** based on the science of genetics and cellular physiology at the molecular level. The second is perhaps **space technology** involving the exploration of the solar system and the physical environment of the earth. The third which become an essential tool in all operations and research work is **informatics** encompassing the whole of communications, interactive intelligent systems, massive data bases and complex information processing networks. Each one is revolutionary in character and bound to have a profund effect on the economic, social, physical and psychological aspects of human society. It is belived that with these technologies, it would be possible to stamp out hunger, feed people adequately, contain epidemics, render strenuous physical labor unnecessary, and above all enable people to lead substantially more comfortable life. It is also belived that these technologies being more knowledge based and less dependent on availability of conventional energy, fabrication of complex machinery and a variety of resources, will eventually remove the existing disparities between countries.

On the negative side, the expansion of technology over surface of the globe has some effects on nature, like depletion of the ozone layer exposing us to greater amounts of radiation from the sum and green house effect leading to increasing global temperatures. The affluence and leisure provided by these new technologies will, no doubt have some effect on our life styles; the attitudes of people towards religion and arts may change in such a way that moral values are eroded and population at large will not have any fulfilled creative content of life.

An important issue that faces us is the proper management of science and technology to provide the maximum possible benefit to mankind avoiding the possible dangers of material pollution. There is also a need to improve the ability of ordinary people to cope up with the new scientific and technological culture permeating the world.

The main key for success in these endeavours is acquiring information and processing it in a useful form for the policy makers to make decesisions on the utilization

of existing technologies and to promote further research for the benefit of mankind. This is where statistics comes in as the science and technology of collecting relevant data expeditiously and analyzing them to extract the desired information and communicating the results to the users.

Statistics has a long antiquity but a short history. It emerged as a separate discipline of study and research only in the second quarter of the present century. Its subject matter is even now not well defined and it may appear to have no definite classification as science, technology or art. Is seems to involve the salient aspects of all these areas. However, it is generally recognized as a logical method of reasoning under uncertainty for predicting future events and taking decisions. As such its scope extends to the whole gamut of natural and social sciences, engineering and technology, management and economic affairs, legal matters, and even arts and literature.

Not long ago, there were misconceptions and skepticisms about statistics, which were mainly due to unplanned and biased collection of data and misuse of figures. We have heard statements like

> "It is easy to lie with figures".
> And answer to this, as Frederik Mosteller says, is
> "It is easier to lie without figures"
> or more appropriately
> "Figures never lie but liers can figure".

Now with the development of statistical methods for systematic collection of data through scientifically designed sample surveys and experiments, and processing of data in an objective manner to extract the desired information, the perception of statistics among the public, the policy makers and the scientistics has changed. Statistics is now used with some confidence by individuals in taking decisions in daily life, by the governments in making day to day policy decisions and long range plans for socio-economic development and by scientists in testing of hypotheses and interpreting experimental results. Statistical methods are routinely used in industry to improve productivity and quality of manufactured goods, in medicine for diagnosis of disease and evaluating the efficacy of new drugs, in agriculture for increasing food production and so on. Accurate weather forecasts are made possible through statistical research. Statistics has also become a major instrument in a democratic set up for assessing public opinion on issues under legislation by the government. In courts of law, statistical evidence in the form of probability of occurrence of certain events is used to supplement the traditional oral and circumstantial evidence in judging cases. The seems to be no human activity whose value cannot be enhanced by injecting statistical ideas in planning and by using statistical methods for efficient analysis of data and assessment of results for feedback and control. Referring to the ubiquity of statistics, Sir Ronald Fisher, the founder of modern statistics, said:

> "Statistical science is the peculiar aspect of human progress which gave the 20th century its special character,... it is to the statiscian that the present age turns for what is most essential in all its more important activities".

Sir Francis Galton, a pioneering statiscian, expressed his conviction of statistics as a gateway to knowledge as follows:

> "Some people hate the very name statistics but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated pehomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the science of man".

In this book, *The Social Functions of Science*, J.D. Bernal cautioned:

> "It is no use improving the knowledge that scientists have about each other's work, if we do not at the same time see that a real understanding of science becomes a part of the common life of our times".

Public understanding of science is important. However, we live in an uncertain world. It is far more important for an individual to know how to deal with uncertainty which confronts him in daily life than to acquire knowledge of modern scientific advances. Statistical knowledge is the best weapon for protecting oneself against false propaganda, dispelling superstition and fighting against wrong allegations. It also helps and individual in making wise decisions, taking advantage of weather forecasts, understanding natural disasters, protecting himself and his family against infection and scores of other things which effect him and on which he has no control. Prophesying the need for public understanding of statistics, H.G. Wells said:

> "A time may not be more remote when it will ve understood that for complete initiation as an efficient citizen... it is as necessary to be able to compute, to think in terms of averages and maxima and minima, as it is now to be able to read and write".

Statistical knowledge is a national resource by which individual and institutional efforts could be enhanced in exploiting the scientific and technological advances for improving the quality of human life. Statistics is a human science and needs to be propagated widely aming all sections of the people.

I consider the distinction awarded to me as a recognition given to statistics by the university as an important discipline which will play a major role as a key techonology for shaping a new world through a radically reoriented program of human resource development, not in the narrow managerial sense of the term, but in a broader sense of improving the quality of life of all individuals in the world and ensuring a commitment to peace and international solidarity.

I thank you once again, Magnificient Rector for giving me a chance to have an affiliation with the University of Barcelona and to interact with the faculty members.

Estadística General i Investigació Operativa

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 23-63, 1995

A REVIEW OF CANONICAL COORDINATES AND AN ALTERNATIVE TO CORRESPONDENCE ANALYSIS USING HELLINGER DISTANCE

C. RADHAKRISHNA RAO*
Pennsylvania State University

In this paper, a general theory of canonical coordinates is developed for reduction of dimensionality in multivariate data, assessing the loss of information and plotting higher dimensional data in two or three dimensions for visual displays. The theory is applied to data in two way tables with variables in one category and samples (individual or populations) in the other. Two types of data are considered, one with continuous measurements on the variables and another with frequencies of attributes. An alternative to the usual correspondence analysis of contingency tables based on Hellinger rather than the chisquare distance is suggested. The new method has some attractive features and does not suffer from some inherent drawbacks resulting from the use of the chi-square distance and variable sample sizes for the populations in the correspondence analysis. The technique of biplots where the populations and the variables are represented on the same chart is discussed.

Key words: Canonical coordinates, Chisquare distance, Contingency tables, Correspondence analysis, Hellinger distance, Matrix approximations, Power divergence tests, Principal component analysis.

AMS classification index: 62H30, 62H17.

*C. Radhakrishna Rao. Department of Statistics. Pennsylvania State University. University Park, PA 16802.

-Article rebut el gener de 1995.

-Acceptat el maig de 1995.

1. INTRODUCTION

The concept of canonical variates (coordinates) was introduced in an early paper by the author (Rao (1948)) for graphical representation of taxonomical units characterized by multiple measurements. This was, perhaps, the first attempt to reduce high dimensional data to two or three dimensions using an objective criterion for purposes of graphical displays. Since then, graphical representation of multivariate data for visual examination of clusters, outliers and other structures in the data has been an active field of research. Some of the developments are biplots (Gabriel (1971), Gifi (1990), Gower (1993), Greenacre (1993)) multidimensional scaling (Kruskal and Wish (1978)), correspondence analysis (Benzécri (1992), Greenacre (1984)), Chernoff's faces (Chernoff (1993)) and parallel coordinates (Mahalanobis, Mazumdar and Rao (1949), Wegman (1990)). Cavalli-Sforza (1991) uses canonical coordinates (variables) in interpreting the evolution of human populations.

The object of the present paper is to briefly review the concept of canonical coordinates as originally introduced in 1948 and later elaborated in Rao (1964, 1979, 1980, 1985) in the light of modern developments and present an alternative to the current practice of correspondence analysis, which seems to have some attractive properties.

In Section 2 we consider the general problem of transforming the points of a p-dimensional vector space endowed with a specified inner product to a lower dimensional Euclidean space with the usual definition of inner product and distance. The solution to the problem is considered in a more general set up than what is possible through the use of Eckart and Young (1936) theorem. In Section 3, some measures are introduced to assess the loss of information in reduction of dimensionality. The role of biplots and their interpretation are also discussed. An example with continuous measurements is considered in Section 4. An alternative to correspondence analysis applied to contingency tables based on Hellinger rather than the chisquare distance is also given in Section 4. Some results on optimization problems and chi-square goodness-of-fit tests are discussed in the Appendix.

2. REDUCTION OF DIMENSIONALITY

The problem we consider may be stated as follows. Let $X = (X_1 : \cdots : X_m)$ be a $p \times m$ data matrix, with the i-th column vector X_i representing measurements of p variables made on the i-th population (individual or unit). The column vector X_i will be referred to as the i-th population profile (PP). The PP's can be represented as m points in a p-dimensional vector space R^p with a specified inner product and the associated norm

$$(x,y) = x'My, \quad x,y \in R^p \tag{2.1}$$

$$||x|| = (x,x)^{1/2}, \quad x \in R^p \tag{2.2}$$

where M is a positive definite matrix. We may call this the Mahalanobis or M-space. In practical situations, it may be necessary to attach a weight $w_i \geq 0$ to the i-th PP, the exact use of which will be detailed in the following discussion. We represent the vector $(w_1,\ldots,w_m)'$ by w and the diagonal matrix with w_i as the i-th diagonal element by W. The M-space with weight as an additional dimension will be referred to as WM-space. [In our treatment we consider W as a general positive definite matrix to cover more general applications].

The problem is to find a $k \times m$ matrix

$$Y = (Y_1 : \cdots : Y_m) \tag{2.3}$$

with $k < p$ for representing the PP's in a k-dimensional Euclidean space (E^k) with the usual inner product, $x'y$ for $x,y \in E^k$, and the k-vector Y_i as the profile of the i-th population, in such a way that the relative positions of the PP's in the M-space (in terms of distances between profiles) are preserved to the extent possible in E^k. For this purpose, we need to have a criterion for measuring the loss of information in reducing the dimension of the profile space, by minimizing which we obtain an optimum solution for (2.3).

The relative positions of the PP's in the M-space can be described by what may called a *configuration matrix*

$$C = (X - \xi 1')'M(X - \xi 1') = ((X_i - \xi)'M(X_j - \xi)) = (c_{ij}) \tag{2.4}$$

where ξ is some chosen reference (profile) vector, 1 is the column vector of unities and the c_{ij}'s represent the distances and angles between profiles as explained in the Figure 1.

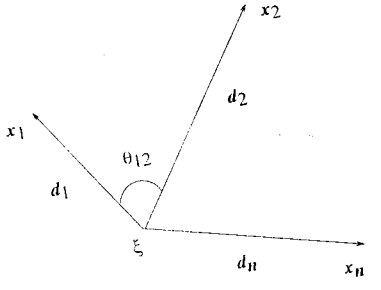


Figure 1.
Configuration of the profiles in M-space $(d_i = \sqrt{c_{ii}}, \quad d_i d_j \cos\theta_{ij} = c_{ij})$

The corresponding configuration about the origin in the reduced space E^k is $Y'Y$. The problem then reduces to minimizing

$$\|C - Y'Y\| \tag{2.5}$$

with respect to Y, a $k \times m$ matrix as defined in (2.3), for a suitably chosen matrix norm. The following theorem provides the solution.

Theorem 1

Consider the s.v.d. (singular value decomposition)

$$M^{1/2}(X - \xi 1')W^{1/2} = \lambda_1 U_1 V_1' + \ldots + \lambda_p U_p V_p' \tag{2.6}$$

with singular values $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, where $M^{1/2}$ and $W^{1/2}$ are symmetric square roots of M and W. Then the choice

$$Y = Y_{(k)} = \begin{pmatrix} \lambda_1 V_1' W^{-1/2} \\ \vdots \\ \lambda_k V_k' W^{-1/2} \end{pmatrix} \tag{2.7a}$$

or conventionally written in the transposed form

$$\lambda_1 W^{-1/2} V_1, \lambda_2 W^{-1/2} V_2, \ldots, \lambda_k W^{-1/2} V_k \tag{2.7b}$$

where the components of the i-th m-vector are the i-th canonical coordinates (i.e., the coordinates in the i-th dimension of the reduced space) for the different populations, minimizes (2.5) for any (W, W)-invariant norm as defined in Note 2.1. We call these coordinates the *canonical coordinates* for populations or profiles (CCP).

The result (2.7a) follows from Theorems A.1 and A.2 given in the Appendix.

Note 2.1. (A, B)-invariant norm of an $m \times n$ matrix is the usual norm (satisfying the postulates of a norm) with the additional property

$$\|C \cdot D\| = \|\cdot\| \quad \text{for any} \quad C, D \text{ such that} \quad C'AC = A, D'BD = B \tag{2.8}$$

where C is an $m \times m$ matrix, D is an $n \times n$ matrix, and A and B are positive definite matrices of orders m and n respectively. This is a generalization given in Rao (1980) of a unitarily invariant norm defined by von Neumann (1937) with A and B as unit matrices.

Note 2.2. In our applications, we indicate some choices of the reference vector ξ. However, we note that a further minimization of (2.5) with respect to ξ leads to the choice

$$\hat{\xi} = (1'W1)^{-1}XW1 \tag{2.9}$$

where 1 is the column vector of unities.

Note 2.3. Using the notation

$$\begin{aligned} \Lambda_{(i)} &= \text{Diag}(\lambda_1, \ldots, \lambda_i) \\ U_{(i)} &= (U_1 : \ldots : U_i), V_{(i)} = (V_1 : \ldots : V_i) \end{aligned}$$

we may write the solution Y given in (2.7a) in the concise form

$$Y_{(k)} = \Lambda_{(k)} V'_{(k)} W^{-1/2}. \tag{2.10}$$

Note 2.4. In the expression (2.6), a symmetric square root of a positive definite matrix is used. It can be computed in a simple way as follows. If A is a positive definite matrix of order p with the spectral decomposition

$$A = \Sigma \lambda_i^2 Q_i Q_i' = Q\Lambda^2 Q'$$

where $Q = (Q_1 : \ldots : Q_p)$, then

$$\begin{aligned} A^{1/2} &= \Sigma \lambda_i Q_i Q_i' = Q\Lambda Q' \\ A^{-1/2} &= \Sigma (\lambda_i)^{-1} Q_i Q_i' = Q\Lambda^{-1} Q'. \end{aligned} \tag{2.11}$$

We may look at the problem in a slightly different way by defining what is called the dispersion matrix between profiles

$$B = (X - \xi 1')W(X - \xi 1')' = (b_{ij}) \tag{2.12}$$

where b_{ii} is the weighted variance of the i-th variable and b_{ij} is the weighted covariance between the i-th and j-th variables across the profiles. Consider an approximation, $Z_i \in R^p$ to $(X_i - \xi)$, with the restriction that $Z_1, \ldots, Z_m$ lie in a k dimensional subspace of R^p, in which case we have the representation

$$Z = (Z_1 : \ldots : Z_m) = AC \tag{2.13}$$

where A is a $p \times k$ matrix whose columns span the subspace and C is a $k \times m$ matrix. Without loss of generality we may choose A to satisfy the condition $A'MA = I$ (i.e.,

the columns of A are orthonormal in the M-space). The dispersion matrix between profiles in the reduced space is $ACWC'A'$, and we choose A and C such that

$$||B - ACWC'A'|| \tag{2.14}$$

is a minimum for an appropriate norm of the matrix. The solution is given in Theorem 2.

Theorem 2

Consider the same s.v.d. as in Theorem 1

$$M^{1/2}(X - \xi 1')W^{1/2} = \lambda_1 U_1 V_1' + \ldots + \lambda_p U_p V_p'.$$

Then the optimum choice of AC which minimizes (2.14) for any (M,M)-invariant norm is

$$AC_{(k)} = M^{-1/2}(\lambda_1 U_1 V_1' + \ldots + \lambda_k U_k V_k')W^{-1/2} \tag{2.15}$$

where the suffix (k) is introduced to indicate the dimension of the reduced space. We may choose

$$\begin{aligned} A &= M^{-1/2}(U_1 : \ldots : U_k) = M^{-1/2}U_{(k)} \\ C_{(k)} &= \begin{pmatrix} \lambda_1 V_1' W^{-1/2} \\ \vdots \\ \lambda_k V_k' W^{-1/2} \end{pmatrix} = \Lambda_{(k)} V_{(k)}' W^{-1/2}. \end{aligned} \tag{2.16}$$

The results follow by applying Theorems A.1 and A.2 given in the Appendix.

Note 2.5. We may represent the profiles by plotting the columns of C in *a* k-dimensional Euclidean space, which is the same solution as that obtained in Theorem 1.

A geometric approach to the problem of reduction of dimensionality is to fit a k-dimensional plane to the data. A set of m points on a k-plane can be written as

$$\xi 1' + AC \tag{2.17}$$

where A is a $p \times k$ matrix and C is a $k \times m$ matrix. We determine A, C, ξ such that

$$||X - \xi 1' - AC|| \tag{2.18}$$

is a minimum for a suitably chosen norm. The solution is given in Theorem 3.

Theorem 3

Consider the same s.v.d. as in Theorem 1. Then the choices of A and C as in Theorem 2 and $\xi = (1'W1)^{-1}XW1$ as in (2.9) minimize any (M,W)-invariant norm of (2.18).

The result follows by a direct application of Theorem A.1 of the Appendix.

Note 2.6. We may also look at the problem in some other ways. Let T be a $k \times p$ matrix providing a transformation of the column vectors of X to $Y = TX$ in a k-dimensional space with the induced inner product matrix $TM^{-1}T'$. The squared distance between the i-th and j-th profiles is

$$D_{ij}^2 = (X_i - X_j)'M(X_i - X_j) \tag{2.19}$$

in the full space, and

$$D_{ij(k)}^2 = (X_i - X_j)'T'(TM^{-1}T')^{-1}T(X_i - X_j) \tag{2.20}$$

in the reduced space. By definition $D_{ij(k)}^2 \leq D_{ij}^2$. We may then choose T by minimizing some function of the differences or ratios of D_{ij}^2 and $D_{ij(k)}^2$.

One of the functions suggested by Rao (1948) was the difference in the weighted sum of all possible differences

$$\Sigma\Sigma w_i w_j (D_{ij}^2 - D_{ij(k)}^2) \tag{2.21}$$

which leads to the same solution for $Y = TX$ as in Theorems 1, 2 and 3.

Another method is to choose T by maximizing the minimum of $D_{ij(k)}^2$ over all i and j as suggested by Eslava-Gomez and Marriott (1993), or by maximizing the minimum of the ratios $D_{ij(k)}^2/D_{ij}^2$. Both these methods are computationally very complex, but can be managed when p is small.

Note 2.7. The choices of M and W as inputs in the analysis for canonical coordinates need some discussion. The choice of M is related to the distance measure between profiles appropriate to a given investigation. In taxonomical classification, M is generally chosen as the inverse of the variance-covariance (dispersion) matrix of the measurements on units within taxa leading to Mahalanobis distance (see Rao (1945, 1947)). The matrix W is taken to be diagonal with the i-th diagonal element w_i proportional to the number of individuals sampled from the i-th taxa to estimate its profile. For a chosen M, the configuration of the profiles in the reduced space will depend on W, but is likely to be robust provided the w_i's are not widely different. In the study reported in Rao (1948), all the w_i's were chosen as equal although the

sample sizes for different populations were different. However, the choice of w_i's as proportional to sample sizes enables us to test hypotheses on goodness of fit of lower dimensional planes to the observed profiles. For details, the reader is referred to Rao (1973, pp. 556–560, 1985).

If we desire that the configuration of a subset of profiles to be better preserved in the reduced space than the others, then we have to give bigger weights to those profiles.

Note 2.8. In many situations we have a data matrix X giving the measurements of p variables made on m individuals without any further information to guide us in the choices of the M and W matrices. In such cases, the usual choices of M and W are the unit matrices and the resulting canonical coordinate analysis is the Principal Component Analysis (PCA) introduced by Hotelling. Some characterizations of the principal components and their applications are given in papers by Rao (1958, 1964, 1987). It is also the practice to apply PCA on CX, i.e., after a suitable scaling of the measurements. One choice of C is a diagonal matrix with the i-th diagonal element $c_i = s_{ii}^{-1/2}$, where s_{ii} is the i-th diagonal element of the matrix

$$(X - \bar{X}1')(X - \bar{X}1')'.$$

This procedure is equivalent to using the canonical coordinate analysis choosing $M = C$ and $W = I$. Another possibility which has not been considered before is the choice, $c_i = 1/m_i$ where m_i is a measure of location such as the mean or median of the measurements on the i-th variable.

Note 2.9. A more general problem not considered in this paper is as follows. The basic space is somewhat general with a specified nonnegative proximity index between any two points. Given a set of points with the matrix of proximity indices between points, the problem is to transform the points to a low dimensional Euclidean space such that the inequality relationships between proximity indices are maintained to the extent possible in the corresponding Euclidean distances. Such a transformation is achieved through the algorithm for multidimensional scaling as developed by Kruskal and Wish (1978).

3. LOSS OF INFORMATION

The representation of the PP's in a lower dimensional space will entail some loss of information depending on the object of statistical analysis. However, we provide

some general criteria for assessing the amount of distortion in the configuration of the profiles due to reduction of dimensionality.

In Theorems 1 and 2 of Section 2, it is shown that the best approximation to X in the reduced space is

$$\hat{X} = \xi 1' + M^{-1/2} U_{(k)} \Lambda_{(k)} V'_{(k)} W^{-1/2} \tag{3.1}$$

so that the matrix

$$D_1 = X - \hat{X} = M^{-1/2}(\lambda_{k+1} U_{k+1} V'_{k+1} + \ldots + \lambda_p U_p V'_p) W^{-1/2}. \tag{3.2}$$

gives a complete account of the errors in individual profiles due to reduction.

The configuration of the profiles in the reduced space is

$$C_{(k)} = W^{-1/2} V_{(k)} \Lambda^2_{(k)} V'_{(k)} W^{-1/2} \tag{3.3}$$

so that the matrix

$$D_2 = C_{(p)} - C_{(k)} = W^{-1/2}(\lambda^2_{k+1} V_{k+1} V'_{k+1} + \ldots + \lambda^2_p V_p V'_p) W^{-1/2}. \tag{3.4}$$

measures the distortion in the configuration, where $C_{(p)} = C$ as defined in (2.4). An overall (weighted) measure of loss of information is the ratio of

$$\text{trace} \quad W^{1/2} D_2 W^{1/2} = \lambda^2_{k+1} + \ldots + \lambda^2_p, \tag{3.5}$$

to the total variation $(\lambda^2_1 + \ldots + \lambda^2_p)$, which can be written as

$$1 - (\sum_{1}^{k} \lambda^2_i)/(\sum_{1}^{p} \lambda^2_i). \tag{3.6}$$

It is more important to assess the distortions in the inter profile squared distances. The matrix of these squared distances denoted by S can be computed from the configuration matrix C using the formula (A 1.16) given in the Appendix as

$$S = c1' + 1c' - 2C \tag{3.7}$$

where c is the vector of the diagonal elements of C. The corresponding matrix in the reduced space is

$$S_{(k)} = c_{(k)} 1' + 1c'_{(k)} - 2C_{(k)} \tag{3.8}$$

so that the matrix

$$D_3 = S - S_{(k)} = (d^*_{ij}) \tag{3.9}$$

measures the deficiencies in the distances due to reduction of dimensionality. An over all measure of deficiency is

$$\Sigma\Sigma w_i w_j d_{ij}^* = \lambda_{k+1}^2 + \ldots + \lambda_p^2 \tag{3.10}$$

which is the same as in (3.5).

The dispersion matrix between profiles in the whole space, as introduced in (2.12) is

$$B = (X - \xi 1')W(X - \xi 1')' = (b_{ij}) \tag{3.11}$$

while the corresponding matrix in the reduced k-dimensional space is

$$B_{(k)} = M^{-1/2} U_{(k)} \Lambda_{(k)}^2 U_{(k)}' M^{-1/2} = (b_{ij(k)}). \tag{3.12}$$

The proportion of the between profile variance in the i-th variable explained by the first k canonical variates (coordinates) is

$$b_{ii(k)}/b_{ii}, i = 1, \ldots, p. \tag{3.13}$$

For an interpretation of the canonical coordinates in different dimensions it would be useful to compute the proportion of variance in each variable explained by each of the canonical variates, i.e., to obtain a decomposition of (3.13) in terms of canonical variates. For this purpose, we introduce the matrices

$$E_1 = M^{-1/2}(\lambda_1 U_1 : \ldots : \lambda_p U_p) = (e_{ij}) \tag{3.14}$$

$$E_2 = (e_{ij}/\sqrt{b_{ii}}) = (f_{ij}) \tag{3.15}$$

where b_{ii} is as defined in (3.11). Let $E_{i(k)}$ be the matrix obtained by retaining only the first k columns in E_i for i=1,2. Then it is seen that

$$E_1 E_1' = B, E_{1(k)} E_{1(k)}' = B_{(k)}. \tag{3.16}$$

Let us consider the matrix $E_{1(k)}$ and define what may be called canonical coordinates for variables (CCV) in k dimensions as follows.

Table 3.1.

Canonical coordinates for variables

Variable	dim 1	dim 2	$\cdots$	dim k
1	e_{11}	e_{12}	$\cdots$	e_{1k}
2	e_{21}	e_{22}	$\cdots$	e_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
p	e_{p_1}	e_{p_2}	$\cdots$	e_{p_k}

If we plot the variables as points in E^k using the row coordinates in different dimensions, then the scalar products of the vectors representing the variables are the elements of $B_{(k)}$, the best k-dimensional approximation to B.

There is some advantage in plotting the variables using the standardized coordinates (f_{ij}) defined in (3.15) as shown in Table 3.2.

Table 3.2.

Standardized CCV's and the variance explained by each canonical variate

Variable	Standardized coordinates dim 1 ... dim k	Proportion of variance explained dim 1 ... dim k	total
1	$f_{11} \cdots f_{1k}$	$f_{11}^2 \cdots f_{1k}^2$	$\sum f_{1i}^2$
$\vdots$	$\vdots \ddots \vdots$	$\vdots \ddots \vdots$	$\vdots$
p	$f_{p1} \cdots f_{pk}$	$f_{p1}^2 \cdots f_{pk}^2$	$\sum f_{pi}^2$

The magnitudes in the right hand block of Table 3.2 indicate the influence of different variables in each dimension (canonical variate) in the reduced space. This may enable us to associate each dimension with certain variables.

We may plot the variables using the standardized CCV's in the same chart as the canonical coordinates for the profiles. It is seen that all variable points lie inside the unit sphere in E^k, and the variables close to the surface of the sphere have greater influence on the canonical variates.

It may also be mentioned that it is the usual practice in a biplot to represent the i-th variable as a directed line using the direction cosines proportional to the i-th row elements in the matrix

$$E_{1(k)} = M^{-1/2}(U_1 : \ldots : U_k) \tag{3.17}$$

in which case the projections of a profile point in these directions are proportional to the approximate coordinates of the profile in the original space (see Gabriel (1971) and Greenacre (1993)).

Note 3.1. We may consider the k columns in Table 3.1 of the CCV's as k points in the p-dimensional variable space. These points were termed as *typical profiles* in Rao (1964), in the sense that the variance-covariance matrix of the variables computed from them provides the best approximation to that computed from all the original profiles.

Table 4.1.1.

Mean value by groups and characters

Caste		Sample	Stature	Sitting height	Nasal depth	Nasal height	Head length	Frontal breadth	Bizygomatic breadth	Head breadth	Nasal breadth
groupe	Symbol	size	st	sh	nd	nl	hl	fb	bb	hb	nb
1. Brahmin	(B_1)	85	164.51	86.43	25.49	51.24	191.92	104.74	133.86	139.88	36.55
2. Brahmin	(B_2)	92	165.07	86.25	24.74	50.40	191.35	104.46	132.68	139.50	36.13
3. Chatri	(C)	139	163.33	82.25	24.73	52.72	192.58	103.98	131.70	131.72	35.64
4. Muslim	(M)	168	162.45	81.83	24.49	51.38	190.78	103.28	131.52	137.40	36.36
5. Bhatru	(Bh)	149	163.38	84.49	25.09	52.06	186.10	99.34	133.55	138.58	35.65
6. Habru	(H)	124	164.91	85.53	24.19	50.30	186.94	100.18	131.16	137.40	35.82
7. Dom	(D)	113	166.53	84.19	25.33	50.34	186.40	104.16	132.64	137.52	38.11
8. Ahir	(A_1)	67	161.37	84.35	24.29	48.98	187.45	102.76	131.70	138.12	35.60
9. Kurmi	(A_2)	94	161.35	83.41	24.03	49.22	188.86	102.62	131.82	137.86	36.21
10. Artesan	(A_3)	172	161.34	83.09	23.73	48.72	187.69	102.44	131.30	136.84	36.27
11. Kahar	(A_4)	57	160.53	81.47	23.84	48.62	188.83	101.68	130.70	136.28	36.61
12. Tharu	(Th)	191	163.33	83.57	21.72	49.94	187.78	100.00	131.68	135.90	37.90
13. Chamar	(Ch)	158	161.88	80.39	23.27	48.50	186.85	102.62	130.28	136.40	36.35
14. Chero	(T_1)	100	162.03	79.67	22.64	47.88	187.09	101.40	130.00	135.48	38.88
15. Majhi	(T_2)	153	162.96	79.89	22.37	46.70	187.12	102.56	131.34	135.92	39.24
16. Panika	(T_3)	156	159.76	78.71	22.40	47.22	186.34	102.14	130.44	135.10	38.05
17. Kharwar	(T_4)	197	160.79	80.31	22.19	48.34	185.92	102.34	131.22	136.74	38.47

Note 3.2. The standardized CCV's are not the coordinates for row profiles. They are used for interpreting the CC's of column profiles. If a representation of row profiles is needed, we consider the matrix X' with appropriate choice of the M and W matrices (which may be different from those used for column profiles) and repeat the analysis indicated in (2.6) — (2.7b).

4. APPLICATIONS

4.1. Continuous measurements

In practical applications, we have at least two types of data matrices, one where the profiles are vectors of measurements on continuous variables and another where the profiles are vectors of relative frequencies of attributes. We discuss these two cases with some examples.

Table 4.1.1 gives the sample sizes and mean values of 9 anthropometric measurements made on samples of individuals from 17 populations. This is our data matrix X' of order $m \times p$ with m=17 and p=9. The mean values in each row of Table 4.1.1 represent a population profile which can be plotted as a point in a p-dimensional vector space. (Note that in the Table 4.1.1, rows represent populations and columns the variables). We would like to represent the populations in a lower (2 or 3) dimensional space for a visual examination to find clusters and outliers and other structures in the data. Besides the data, we need some other inputs to apply the methods of data reduction developed in the previous sections.

First is the choice of a metric or an inner product in the vector space of profiles, in terms of which the configuration of a set of profiles can be described. The choice naturally depends on the object of investigation. In the context of anthropometric measurements a natural choice is the Mahalanobis distance (see Mahalanobis (1936) and Rao (1948)). If X_i' and X_j' are the row vectors (mean values) representing the profiles of the i-th and j-th populations respectively, then the square of the Mahalanobis distance between populations i and j is

$$(X_i - X_j)'\Sigma^{-1}(X_i - X_j)$$

where Σ is the common (or the average) variance-covariance (dispersion) matrix of the variables within a population. We may choose $M = \Sigma^{-1}$. In the above example we have samples from each population with a total number of 2215 individuals. We use MANOVA (or Analysis of dispersion, Rao (1973, p. 570)) to obtain the decomposition of T the total sum of squares and products matrix of order 9x9 as

between B and within W. Dividing W by the degrees of freedom (2215-17=2198), we have an estimate of Σ as shown in Table 4.1.2.

Table 4.1.2.

The estimated variance-covariance matrix $\hat{\Sigma}$ within populations

	st	*sh*	*nd*	*nl*	*hl*	*fb*	*bb*	*hb*	*nb*
st	32.9476	10.7434	1.7820	3.9658	10.2211	4.8894	7.6002	3.6472	3.1023
sh		10.2400	1.1726	2.4304	5.5989	3.7782	4.3895	2.9794	0.9721
nd			3.0625	1.7824	1.7752	0.8527	1.2624	1.0301	0.5123
nh				12.2500	4.0610	1.5627	2.9687	2.7326	0.3940
hl					43.5600	5.8729	8.4397	5.8865	3.2737
fb						15.3664	8.8511	7.8692	1.8446
bb							20.9764	11.1438	3.2122
hb								20.2500	1.6341
nb									6.6049

Second is the choice of the weights to be attached to the profiles. A natural choice of the weight to be attached to the i-th population is its sample size n_i expressed as a proportion of the total sample size n ($w_i = n_i/n$). Such a choice may not always be the best, especially when the sample sizes are widely different, although there is an advantage when tests of significance are also considered. (In an earlier analysis, Rao (1948) used equal weights although the sample sizes were different. Usually the reduced configuration is robust to different choices of weights).

Choosing $M^{-1} = \hat{\Sigma}$ as in Table 4.1.2. and $w_i = n_i/n$, i.e.,

$$2215W = \text{diag}(85, 92, 139, 168, 149, 124, 113, 67, 94, 172, 57, 191, 158, 100, 153, 156, 197)$$

where the numbers are the sample sizes, we compute the s.v.d. of

$$\hat{\Sigma}^{-\frac{1}{2}}(X - \xi 1')W^{1/2} = \lambda_1 U_1 V_1' + \ldots + \lambda_p U_p V_p'$$

with $\xi = XW1/1'W1$. The first three canonical coordinates as defined in (2.7) are as shown in Table 4.1.3. The three columns in the table are the vectors $\lambda_1 W^{-1/2} V_1$, $\lambda_2 W^{-1/2} V_2$ and $\lambda_3 W^{-1/2} V_3$.

The squares of the singular values are 2165.6, 1012.2, 478.1, 265.5, 207.6, 93.9, 64.5, 38.6 and 17.0 with a total of 4343.1. The first three canonical coordinates explain about 84% of variation between populations. Figure 2 gives a plot of the first two canonical coordinates

Table 4.1.3

Canonical coordinates in three dimensions

Group	dim 1	dim 2	dim 3
B_1	1.505584255	0.151968922	0.49580195
B_2	1.268677020	−0.008997802	0.43108765
C	0.376888549	1.864144347	−1.05099680
M	0.271688813	0.739815783	0.04093115
Bh	1.497884705	−0.638438490	−0.23425360
H	1.188939615	−0.741844231	−0.08910818
D	0.458655511	0.268433534	0.63562071
A_1	0.962270474	−0.186013502	0.41809923
A_2	0.570135707	−0.064868650	0.24783429
A_3	0.352468454	−0.066568893	0.18233164
A_4	−0.006945018	0.191180399	0.17306344
Th	−0.301523213	−1.119103947	−0.96252679
Ch	−0.503912465	0.368649078	0.29020494
T_1	−1.306812703	−0.088500525	0.11166126
T_2	−1.525798049	−0.125746569	0.20196936
T_3	−1.412895439	0.197121678	0.11349265
T_4	−1.125431499	−0.338872721	0.01751354
% Variance explained	49.86	23.30	11.00

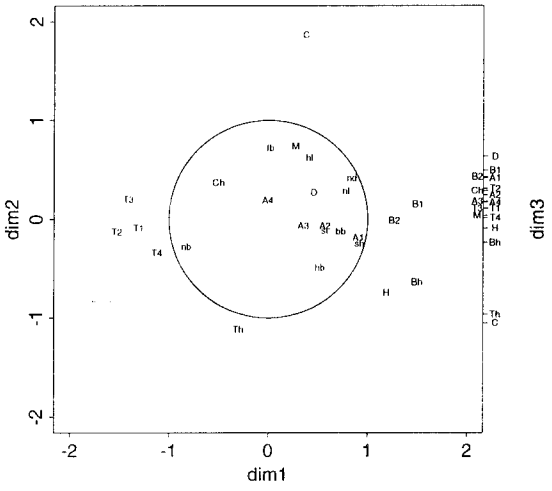


Figure 2.
Configuration of canonical coordinates for caste groups and variables. The coordinates in the third dimension for the caste groups are represented on a separate line.

The standardized first three canonical coordinates for the variables as defined in (3.15) and the between group variance in each variable explained are given in Table 4.1.4. They are the column vectors $\lambda_1\Delta^{-1}\Sigma^{1/2}U_1, \lambda_2\Delta^{-1}\Sigma^{1/2}U_2$ and $\lambda_3\Delta^{-1}\Sigma^{1/2}U_3$ where Δ is the diagonal matrix formed by the square roots of the diagonal elements of the matrix $(X-\xi 1')W(X-\xi 1')'$.

Table 4.1.4.

Standardized CC's for variables and between group variance explained

Variable	dim 1	dim 2	dim 3	% variance
st	0.56447484	−0.04964491	−0.02136700	32.15530
sh	0.91478661	−0.24173510	−0.03048027	89.61995
nd	0.83822736	0.41848655	0.25011364	94.03129
nl	0.77824697	0.28621522	−0.43125314	87.35668
hl	0.40934124	0.62262950	−0.22874285	60.75510
fb	0.02029308	0.72195951	0.43410258	71.00824
bb	0.72821282	−0.12153872	−0.02456127	54.56688
hb	0.51704122	−0.48947063	0.60153126	86.87530
nb	−0.82527584	−0.27455517	0.11745159	77.02556

The first two CC's for the variables are plotted in the same Figure 2. It is seen that they all lie within the unit circle. The nasal measurements (nl, nd and hb) and sitting height (sh) are close to the circumference of the circle indicating that they are well represented in the first two CC's for the populations.

Note 4.1. A two dimensional representation may not show the exact relative positions of the populations in the original space but the distortion may be large in particular cases. In general it will be wise to take into account the differences in the third as well as in the higher dimensions while interpreting the distances in the two dimensional representation. An innovation in Figure 2 is the representation of the coordinates in the third dimension on a separate line, which shows the additional differences among the groups. [Note that the square of the distance between any two groups in the three dimensional representation is the square of the distance in the two dimensional representation plus the square of the difference in the coordinates of the third dimension.]

A general recommendation is to consider a two dimensional plot supplemented by representation of the points in other dimensions on separate lines as parallel coordinates (see Wegman (1990) for details).

4.2. Two way contingency tables

We consider dichotomous categorical data with s rows and m columns and n_{ij} observations in the (i,j)-th cell. Define

$$N = (n_{ij}), n_{i.} = \sum_{j=1}^{m} n_{ij},\ n_{.j} = \sum_{i=1}^{s} n_{ij},\ n = \sum_{1}^{s}\sum_{1}^{m} n_{ij}$$

$$R = \text{Diag}\,(n_{1.}/n, \ldots, n_{s.}/n),\ C = \text{Diag}\,(n_{.1}/n, \ldots, n_{.m}/n)$$

$$P = n^{-1}NC^{-1} = \begin{pmatrix} p_{1|1} & \cdots & p_{1|m} \\ \vdots & \ddots & \vdots \\ p_{s|1} & \cdots & p_{s|m} \end{pmatrix}, \quad \text{column profiles,} \tag{4.2.1}$$

$$Q = n^{-1}R^{-1}N = \begin{pmatrix} q_{1|1} & \cdots & q_{m|1} \\ \vdots & \ddots & \vdots \\ q_{1|s} & \cdots & q_{m|s} \end{pmatrix}, \quad \text{row profiles} \tag{4.2.2}$$

$$p = R1,\ q = C1.$$

The problem is to represent the column (row) profiles as points in $E^k, k < s$, such that the Euclidean distances between points reflect specified affinities between the corresponding column (row) profiles.

The technique developed for this purpose by Benzécri (1992) is known as correspondence analysis (CA) which can be identified as canonical coordinate analysis. For instance, for representing the column profiles by this method, one chooses

$$X = P,\ M = R^{-1},\ W = C \tag{4.2.3}$$

and applies the analysis described in Theorem 1 (equation (2.6)). Thus one finds the s.v.d. of

$$R^{-1/2}(P - p1')C^{1/2} = \lambda_1 U_1 V_1' + \cdots + \lambda_s U_s V_s' \tag{4.2.4a}$$

giving the coordinates for the column profiles in E^k

$$\lambda_1 C^{-1/2} V_1, \lambda_2 C^{-1/2} V_2, \ldots, \lambda_k C^{-1/2} V_k \tag{4.2.4b}$$

where the components of i-th vector are the coordinates of the profiles in the i-th dimension. The standardized canonical coordinates in E^k for the rows, as described in (3.15), obtained from the same s.v.d. as in (4.2.4a) are

$$\lambda_1 \Delta^{-1} R^{1/2} U_1, \lambda_2 \Delta^{-1} R^{1/2} U_2, \ldots, \lambda_k \Delta^{-1} R^{1/2} U_k \tag{4.2.4c}$$

where the components of the i-th vector are the coordinates of the rows in the i-th dimension and Δ is a diagonal matrix with the i-th diagonal element as the square root of the i-th diagonal element of the matrix

(4.2.4d)
$$R^{1/2}(\lambda_1^2 U_1 U_1' + \cdots + \lambda_s^2 U_s U_s')R^{1/2} = (P - p1')C(P - p1')'.$$

The coordinates (4.2.4c) do not represent the row profiles but are useful in interpreting the different dimensions of the column profiles. The coordinates for representing the row profiles in correspondence analysis are given in (4.2.6c).

Implicit in this analysis is the choice of measure of affinity between the i-th and j-th profiles as the squared distance

(4.2.5)
$$d_{ij}^2 = \frac{(p_{1|i} - p_{1|j})^2}{p_1} + \cdots + \frac{(p_{s|i} - p_{s|j})^2}{p_s}$$

which is the chisquare distance. The squared Euclidean distance in E^k, the reduced space, between the points representing the i-th and j-th profiles is an approximation to (4.2.5). Thus the clusters we see in the Euclidean representation is based on the affinities measured by the chisquare distance (4.2.5).

Why should one choose the chisquare distance to measure the affinities between profiles? Some of the advantages mentioned by Benzécri and Greenacre are as follows.

1. Note that the expression in (4.2.4a)

(4.2.6a)
$$R^{-1/2}(P - p1')C^{1/2} = R^{1/2}(Q - 1q')C^{-1/2} = T$$

so that if we need a representation of the row (as population) profiles in E^k, we use the same s.v.d. as in (4.2.4a)

(4.2.6b)
$$R^{1/2}(Q - 1q')C^{-1/2} = \lambda_1 U_1 V_1' + \cdots + \lambda_s U_s V_s'$$

leading to the row (population) coordinates

(4.2.6c)
$$(\lambda_1 R^{-1/2} U_1 : \ldots : \lambda_k R^{-1/2} U_k)$$

so that no extra computations are needed if we want a representation of the row profiles also. In correspondence analysis it is customary to plot the points (4.2.4b) and (4.2.6c) in the same chart. Then the standardized coordinates for the columns (as variables) are

(4.2.6d)
$$\lambda_1 \Delta_1^{-1} C^{1/2} V_1, \ldots \lambda_k \Delta_1^{-1} C^{1/2} V_k$$

where $\triangle_1$ is a diagonal matrix with the i-th diagonal element as the square root of the i-th diagonal element of $(Q - 1q')'R(Q - 1q')$.

2. It is easy to see that

$$\begin{aligned} n(\lambda_1^2 + \cdots + \lambda_k^2) &= n \text{ trace } TT', \quad \text{with } T \text{ as in (4.2.6a)} \\ &= \sum\sum \frac{(n_{ij} - np_iq_j)^2}{np_iq_j} \end{aligned}$$

which is the Pearson chisquare statistic for testing independence between the attributes in a contingency table. Thus the computations involved in CA automatically allow us to test for independence, and also test for the dimensionality of the space of profiles using statistics of the type

$$n(\lambda_i^2 + \cdots + \lambda_s^2), i = 1, 2, \ldots \tag{4.2.7}$$

as discussed in Rao (1973, pp. 556-560).

3. CA is only an exploratory data analysis to examine the configuration of row and column profiles in a general way, so that a particular convenient choice of the distance measure can serve the purpose.

On the other hand, there seem to be some drawbacks in using the chisquare distance.

1. The chisquare distance (4.2.5) is not a function of the i-th and j-th column profiles only. It involves the marginal profile which is a weighted average of the individual column profiles. The weights depend on the observed numbers of individuals in the column categories. These numbers may not have any relevance to the problem under study, especially when the columns represent different populations from each of which some individuals are chosen and classified according to row categories. In such a case the marginal profile depends on the actual sample sizes chosen or realized for different populations. The examples discussed in the sequel show that the derived configurations in the reduced space may be sensitive to the sample numbers.

2. The marginal profile depends on the set of populations included in CA. The CA's based on a given set of populations (S_1) and an extended set of populations (S_1, S_2) may provide different configurations to the subset S_1 as shown in Example 4.2.1.

3. There is no particular advantage in plotting the row and column profiles in the same chart. Indeed one could use different distance measures for column and row profiles and study configurations of the column and row profiles separately.
4. Since the chisquare distance uses the marginal proportions in the denominator, undue emphasis is given to the categories with low frequencies in measuring affinities between profiles.

An alternative to the chisquare distance which has some advantages is the Hellinger Distance (HD) between the i-th and j-th column profiles defined by

$$d_{ij}^2 = (\sqrt{p_{1|i}} - \sqrt{p_{1|j}})^2 + \cdots + (\sqrt{p_{s|i}} - \sqrt{p_{s|j}})^2 \tag{4.2.8}$$

which depends only on the i-th and j-th column profiles. In such a case, the Euclidean distance in the reduced space between the i-th and j-th column profiles is an approximation to (4.2.8). For the derivation of canonical coordinates of the column profiles (considered as population) we choose

$$X = \begin{pmatrix} \sqrt{p_{1|1}} & \cdots & \sqrt{p_{1|m}} \\ \vdots & \ddots & \vdots \\ \sqrt{p_{s|1}} & \cdots & \sqrt{p_{s|m}} \end{pmatrix}$$

$$M = I, \qquad W = C = \text{Diag}\,(n_{.1}/n, \ldots, n_{.m}/n) \tag{4.2.9}$$

and consider the s.v.d.

$$(X - \xi 1')C^{1/2} = \lambda_1 U_1 V_1' + \cdots + \lambda_s U_s V_s'. \tag{4.2.10}$$

We may choose $\xi' = (\xi_1, \ldots, \xi_s)$ as

$$\xi_i = \sqrt{p_{i.}} = \sqrt{n_{i.}/n}, \quad \text{or} \tag{4.2.11}$$

$$= n^{-1}(n_{.1}\sqrt{p_{i|1}} + \cdots + n_{.m}\sqrt{p_{i|m}}). \tag{4.2.12}$$

The canonical coordinates in E^k for the column profiles choosing ξ as in (4.2.11) or (4.2.12) are

$$\lambda_1 C^{-1/2} V_1, \lambda_2 C^{-1/2} V_2, \ldots, \lambda_k C^{-1/2} V_k \tag{4.2.13a}$$

where the components of the i-th vector are the coordinates of the m column (population) profiles in the i-th dimension. The standardized coordinates in E^k for the variables, i.e., the row categories, obtained as described in (3.15) from the same s.v.d. as in (4.2.10) are

$$\lambda_1 \Delta^{-1} U_1, \lambda_2 \Delta^{-1} U_2, \ldots, \lambda_k \Delta^{-1} U_k \tag{4.2.13b}$$

where Δ is a diagonal matrix with the i-th diagonal element as the square root of the i-th diagonal element of

(4.2.13c) $$\lambda_1^2 U_1 U_1' + \cdots + \lambda_s^2 U_s U_s' = (X - \xi 1')C(X - \xi 1')'.$$

The s components of $\lambda_i \triangle^{-1} U_i$ in (4.2.13b) are the coordinates of the s variables in the i-th dimension.

It can be shown that the statistic

(4.2.14) $$4n(\lambda_1^2 + \cdots + \lambda_s^2)$$

is distributed asymptotically as chisquare on $(s-1)(m-1)$ degrees of freedom to test independence in the two way contingency table. Further, hypotheses specifying the dimensions of the subspace in which the profiles can be represented can also be tested in the same way as in (4.2.7) using the residual singular values.

The advantages in using HD between profiles are the following.

1. The measure depends only on the profiles of the concerned pair. It is not altered when an extended set of profiles is considered.
2. The measure does not depend on the sample sizes on which the profiles are estimated.
3. If a representation of the row profiles is also needed we take $X = sqrt(Q')$, i.e., the elements of X are the square roots of the elements of Q' where Q is the matrix defined in (4.2.2) and compute the s.v.d.

(4.2.15a) $$(X - \eta 1')R^{1/2} = \mu_1 A_1 B_1' + \cdots + \mu_s A_s B_s'$$

leading to the canonical coordinates for row profiles (considered as populations)

(4.2.15b) $$\mu_1 R^{-1/2} B_1, \mu_2 R^{-1/2} B_2, \ldots, \mu_k R^{-1/2} B_k.$$

The corresponding standardized coordinates for the columns considered as variables are

(4.2.15c) $$\mu_1 \triangle_c^{-1} A_1, \mu_2 \triangle_c^{-1} A_2, \ldots$$

where $\triangle_c$ is a diagonal matrix with the i-th diagonal element as the square root of the i-th diagonal element of

(4.2.15d) $$\mu_1^2 A_1 A_1' + \cdots + \mu_s^2 A_s A_s'.$$

4. If we choose ξ as in (4.2.11), then the matrix in (4.2.10) is

$$(X - \xi 1')C^{1/2} = \left(\sqrt{\frac{n_{ij}}{n}} - \sqrt{\frac{n_{i.}}{n}\frac{n_{.j}}{n}}\right)$$

which is symmetric in i and j. Then, the same s.v.d. as in (4.2.10) could be used for computing the canonical coordinates

$$\lambda_1 R^{-1/2} U_1, \lambda_2 R^{-1/2} U_2, \ldots, \lambda_k R^{-1/2} U_k$$

for the row profiles, as in the case of CA.

Example 4.2.1

Consider for example the data on smoking habits (s_1, s_2, s_3) of different categories of employees (A, B, C, D, E) given in Tables 4.2.1 and 4.2.2, one of which has data on two additional categories of employees.

Table 4.2.1.

	Frequencies				Profiles			
	A	B	C	Total	A	B	C	Average
s_1	5	0	0	5	1	0	0	5/11
s_2	0	5	0	5	0	1	0	5/11
s_3	0	0	1	1	0	0	1	1/11
Total	5	5	1	11	1	1	1	1

Table 4.2.2.

	Frequencies						Profiles					
	A	B	C	D	E	Total	A	B	C	D	E	Average
s_1	5	0	0	4	1	10	1	0	0	.8	.2	10/21
s_2	0	5	0	1	4	10	0	1	0	.2	.8	10/21
s_3	0	0	1	0	0	1	0	0	1	0	0	1/21
Total	5	5	1	5	5	21	1	1	1	1	1	1

The two dimensional representations of the employee categories obtained through correspondence analysis using the formula (4.2.4b) for the data in the above tables are given in Figure 3, where the lower case letters are used for the data of Table 4.2.2. We make the following observations.

1. The relative positions of A, B and C change when additional employee categories are introduced, although their individual profiles remain the same.
2. The profiles for A, B, C in Table 4.2.1 suggest that they are equally distant between each other, and the correspondence analysis would have revealed this if the sample size for category C had been the same as that for A and B.

Thus, with chisquare distances, there is instability of the configuration of the populations in the reduced space with changes in the sample sizes and the addition or deletion of populations.

The two dimensional representation of the employee categories based on Hellinger distance using the formula (4.2.13a) is given in Figures 3 and 4. It is seen that the graph based on Table 4.2.1 shows that A, B and C are equidistant from each other and the graph based on Table 4.2.2 shows that the positions of A, B, C represented by lower case letters are not changed when the additional categories *d* and *e* are introduced. Thus, the use of Hellinger distance seems to provide a more satisfactory solution.

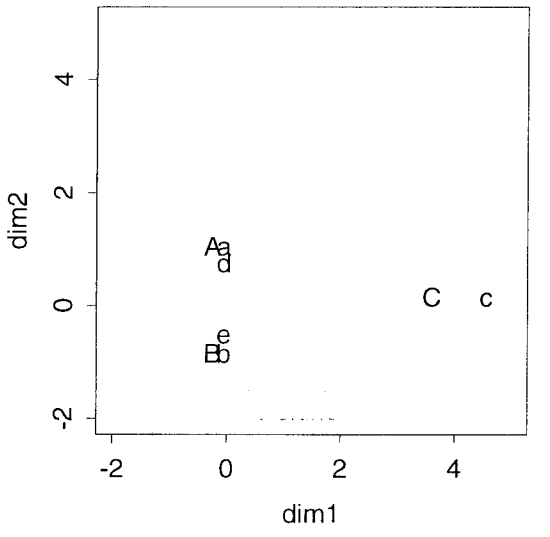


Figure 3.
Configuration of canonical coordinates based on chisquare distance (Lower case letters are used when the analysis is based on all the categories).

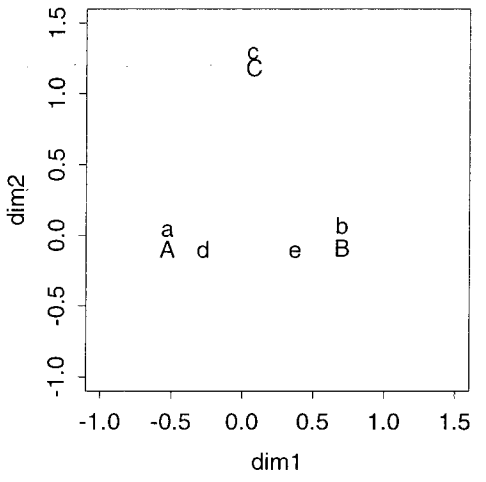


Figure 4.
Configuration of canonical coordinates based on Hellinger distance (Lower case letters are used when the analysis is based on all the categories).

Example 4.2.2

We consider the data (from Greenacre (1993)) on 796 scientific researchers classified according to their scientific discipline (as populations) and funding category (as variables) as shown in Table 4.2.3.

Table 4.2.3.

Scientific disciplines by research funding categories

		Funding category					
Scientific discipline		a	b	c	d	e	Total
Geology	G	3	19	39	14	10	85
Biochemistry	B_1	1	2	13	1	12	29
Chemistry	C	6	25	49	21	29	130
Zoology	Z	3	15	41	35	26	120
Physics	P	10	22	47	9	26	114
Engineering	E	3	11	25	15	34	88
Microbiology	M_1	1	6	14	5	11	37
Botany	B_2	0	12	34	17	23	86
Statistics	S	2	5	11	4	7	29
Mathematics	M_2	2	11	37	8	20	78
Total		31	128	310	129	198	796

The canonical coordinates for the scientific disciplines (considered as populations) in the first three dimensions and percentage of variance explained by each are given in Table 4.2.4 for the analyses based on the chisquare distance (correspondence analysis) and the Hellinger distance (alternative). The formula (4.2.6b) is used for the analysis based on chisquare and the formula (4.2.15a) for that based on Hellinger distance. For Hellinger distance analysis, the central point is chosen according to the formula (4.2.12).

Table 4.2.4.

Canonical coordinates for the scientific disciplines in the first three dimensions

Subjects	Chisquare Distance			Hellinger distance		
	Dim1	Dim2	Dim3	Dim1	Dim2	Dim3
G	.076401	.302569	−.087749	−.031140	.167408	−.048245
B_1	.179892	−.454996	−.151716	−.129374	−.242174	−.077614
C	.037644	.073353	.042371	−.021144	.040433	.028254
Z	−.327365	.102283	.064515	.138850	.045255	.056894
P	.315552	.026997	.108688	−.165340	.010679	.023844
E	−.117495	−.291712	.107330	.049451	−.129906	.082901
M_1	.012766	−.109656	−.041435	−.004913	−.052588	−.008439
B_2	−.178695	−.038501	−.129055	.151404	−.036559	−.108025
S	.124638	.014162	.107190	−.066639	.011763	.052571
M_2	.106751	−.061316	−.175688	−.050307	−.037572	−.078006
% var.	47.20	36.66	13.11	45.87	34.10	16.57

The plots of the scientific disciplines (subjects) using the canonical coordinates based on the chisquare and Hellinger distances are given in Figures 5 and 6 respectively. The coordinates in the third dimension are plotted on a line on the right hand side of the two dimensional plot. This will be of help in visualizing the plot in three dimensions and in interpreting the distances in the two dimensional plot. Thus, although B_2 and E appear to be close to each other in the two dimensional chart, they are clearly separated in the third dimension. No additional distances in the third dimension are involved in the case of P, C, S, Z and E.

It is of interest to note in this example that the configuration of the scientific disciplines in three dimensions obtained by both the methods are very similar. The percentage of variance explained in each dimension is nearly the same for both the methods.

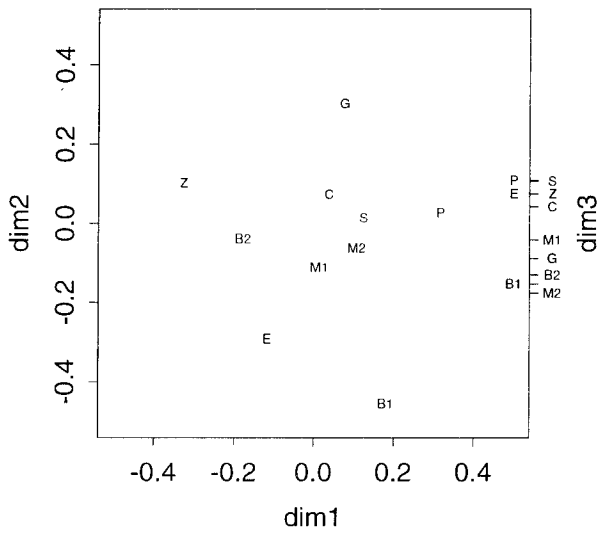


Figure 5.
Configuration of scientific disciplines using Chisquare distance (Correspondence Analysis).

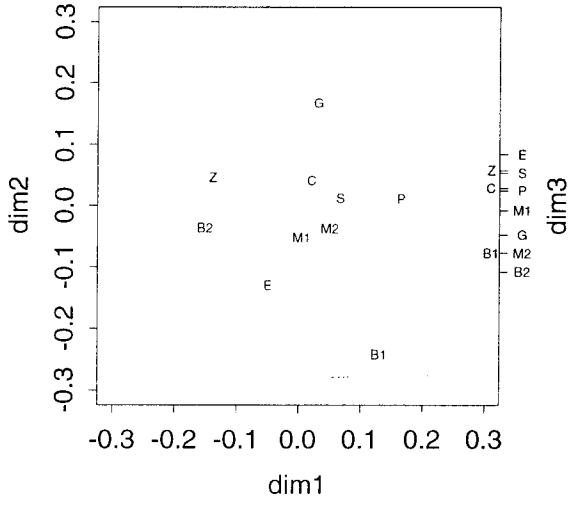


Figure 6.
Configuration of scientific disciplines using Hellinger distance (Alternative to Correspondence Analysis).

The standardized canonical coordinates for the funding categories (considered as variables) are computed using the formula (4.2.6d) for the chisquare analysis and the formula (4.2.15b) for the Hellinger distance analysis. These are obtained from the *same s.v.d.* used to compute the canonical coordinates for the scientific disciplines. Table 4.2.5 gives the standardized canonical coordinates for the funding categories, a, b, c, d, e, using the two methods.

Table 4.2.5.

Standardized canonical coordinates for funding categories (variables) in the first three dimensions

Funding category	Chisquare Distance				Hellinger Distance			
	dim1	dim2	dim3	%var	dim1	dim2	dim3	%var
a	.758	.114	−.619	97.1	.796	.164	−.573	98.9
b	.535	.728	−.137	83.5	.438	.766	−.008	77.9
c	.583	.352	.694	94.6	.501	.327	.759	93.4
d	−.426	.331	−.172	99.8	−.888	.358	−.285	99.7
e	−.108	−.909	−.081	99.6	−.088	−.978	−.159	98.9

The standardized canonical coordinates for the funding categories are plotted in Figure 7 (for chisquare distance) and in Figure 8 (for Hellinger distance). It may be noted that all the points lie within the unit circle. It is customary to represent the canonical coordinates for the subjects and variables in one chart. We are using separate charts in order to explain the salient features of the configuration of the variables. The following interpretations emerge from the study of Table 4.2.5 and Figures 7 and 8.

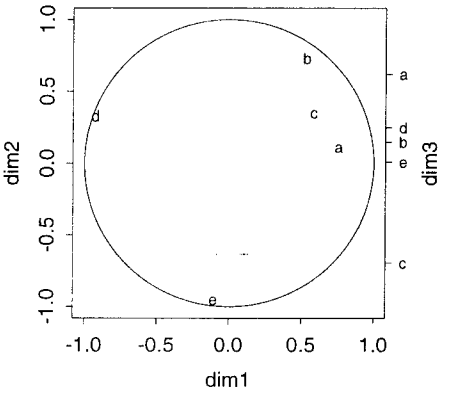


Figure 7.
Configuration of funding categories using standardized canonical coordinates based on Chisquare distance.

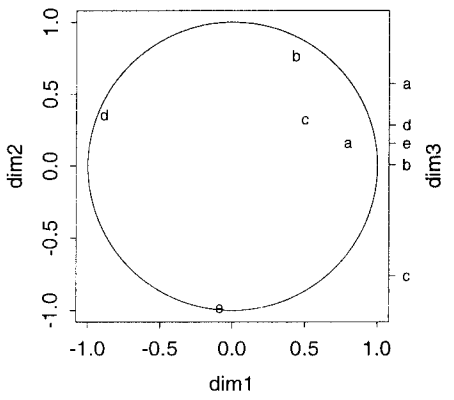


Figure 8.
Configuration of funding categories using standardized canonical coordinates based on Hellinger distance.

1. The configurations of the funding categories as exhibited by Figures 7 and 8 obtained by using chisquare and Hellinger distances are very similar.

2. All most all the variation in the funding categories a, d and e is captured in the first three canonical coordinates of the scientific disciplines. A large percentage of variation in b and c is explained by the first three coordinates.

3. The first dimension is strongly influenced by a, d the second dimension by b, e and the third dimension by a, c.

Thus the use of standardized coordinates for variables enables us to interpret the different dimensions in terms of observed variables. There are other ways of plotting the coordinates of the variables as mentioned in the paragraphs below Table 3.2. Such biplots having a different interpretation are discussed in Gabriel (1971), Gifi (1990), Gower (1993) and Greenacre (1993).

Note 4.2. In computing the canonical coordinates based on Hellinger distance (HD) using the formula (4.2.10), we chose the relative sample sizes as the weights to be attached to the populations. We could have chosen an alternative set of weights if we wanted distances between a specified subset of populations to be better preserved in the reduced space than the others. In particular, we could have chosen uniform weights for all populations. In fact such an option could be exercised if the sample sizes of different populations were widely different. Unfortunately no such options are available in correspondence analysis.

Example 4.2.3

In the example 4.2.2, there was a perfect match between the plots based on CA and HD. This probably demonstrates that the method of derivation of canonical coordinates is somewhat robust to the choice of the distance measure as well as to the weights. However the choice of HD provides as insurance against possible distortion due to variations in sample sizes for the populations as the following example shows.

Table 4.2.6, reproduced from Gifi (1981), gives the distributions of the pages devoted to different topics denoted by A, B, C, D, E, F and G in 20 books on Multivariate analysis designated as $a, b, \ldots, t$. Gifi (1981) did correspondence analysis on the data and drew some conclusions based on the first three canonical coordinates which explain a high percentage of variation.

Table 4.2.6.

Number of pages by topics

Books	A	B	C	D	E	F	G
a	31	0	0	0	0	164	11
b	0	16	54	18	27	13	14
c	0	40	32	10	42	60	0
d	19	0	35	19	28	163	52
e	14	7	35	22	17	0	56
f	20	69	72	33	55	0	32
g	74	0	86	14	0	84	48
h	78	0	80	5	17	105	60
i	74	19	33	12	26	0	0
j	80	68	67	15	29	0	0
k	108	48	4	10	46	108	0
l	109	13	5	17	39	32	46
m	16	35	69	24	0	26	41
n	26	86	60	6	48	48	28
o	290	10	6	0	8	0	2
p	184	48	82	42	134	0	0
q	29	0	0	0	41	211	32
r	0	19	56	0	39	75	0
s	0	22	45	42	60	230	59
t	30	128	90	28	48	0	0

The first three canonical coordinates for the profiles of the books based on CA and HD approaches are given in Table 4.2.7.

Table 4.2.7.

	Chisquare Distance			Hellinger Distance		
	dim1	dim2	dim3	dim1	dim2	dim3
a	−1.10857	−0.61445	−0.33902	0.64632	0.36299	0.12879
b	0.07397	0.70254	0.25265	−0.01661	−0.48923	−0.12388
c	−0.21153	0.46054	−0.49228	0.10998	−0.42185	0.32822
d	−0.77795	−0.11074	0.15556	0.46658	−0.01597	−0.10284
e	0.02781	0.40651	1.06135	−0.19193	−0.15180	−0.45570
f	0.35780	0.69602	0.09284	−0.37016	−0.29359	−0.14451
g	−0.16412	−0.15719	0.46353	0.23979	0.16911	−0.35829
h	−0.25023	−0.19626	0.39002	0.26103	0.14730	−0.23804
i	0.72788	−0.19452	−0.04749	−0.50899	0.14292	0.02936
j	0.68403	0.24337	−0.17956	−0.53320	−0.01242	0.04724
k	0.02729	−0.36648	−0.44297	0.03996	0.21098	0.36189
l	0.26802	−0.44749	0.28287	−0.00524	0.27070	−0.06481
m	0.02188	0.50893	0.51719	0.01506	−0.19266	−0.34080
n	0.12052	0.48459	−0.19476	−0.04555	−0.20966	0.04945
o	1.08308	−1.32602	0.03206	−0.39476	0.66357	−0.00090
p	0.64959	−0.07081	−0.13268	−0.49299	0.09097	0.08510
q	−0.98347	−0.39273	−0.25019	0.58910	0.21379	0.19442
r	−0.40006	0.32919	−0.33826	0.21605	−0.35929	0.28764
s	−0.74726	0.08101	−0.00508	0.43349	−0.30134	0.03139
t	0.56547	0.81454	−0.35256	−0.51167	−0.27874	0.10162

It may be noted that the total number of pages of a book depends on the font size of the print, while its profile in terms of proportions of pages used on different topics remain the same for all sizes. Table 4.2.8 gives the data on books having the same profiles as in Table 4.2.6 with the total number of pages altered for the books d, f, g, h, j and n.

Table 4.2.8.

Number of pages by topics

Books	*A*	*B*	*C*	*D*	*E*	*F*	*G*
a	31	0	0	0	0	164	11
b	0	16	54	18	27	13	14
c	0	40	32	10	42	60	0
d	190	0	350	190	280	1630	520
e	14	7	35	22	17	0	56
f	10	34	36	17	28	0	16
g	740	0	860	140	0	840	480
h	780	0	800	50	170	1050	600
i	74	19	33	12	26	0	0
j	40	34	33	8	15	0	0
k	108	48	4	10	46	108	0
l	109	13	5	17	39	32	46
m	16	35	69	24	0	26	41
n	13	43	30	3	24	24	14
o	290	10	6	0	8	0	2
p	184	48	82	42	134	0	0
q	29	0	0	0	41	211	32
r	0	19	56	0	39	75	0
s	0	22	45	42	60	230	59
t	30	128	90	28	48	0	0

The three dimensional canonical coordinates based on CA and HD approaches are given in Table 4.2.9. Using the coordinates one can obtain the mutual distances between the books in the three dimensional reduced Euclidean space. Figure 9 gives a plot comparing the squared distances between books based on CA using the data of Tables 4.2.6 and 4.2.8. Figure 10 gives the corresponding plot for the squared distances based on the HD approach. It is seen that the three dimensional representation of the data of Tables 4.2.6 and 4.2.8 are more similar under HD analysis than that under CA. The relative positions of the books are influenced by the font size in printing when CA is used, although the profiles of the books are not altered. There appears to be greater stability with the HD analysis which provides insurance against different choice of sample sizes. Further, one can exercise the option of using a common weight for all the books in the HD analysis when the differences in book sizes are large.

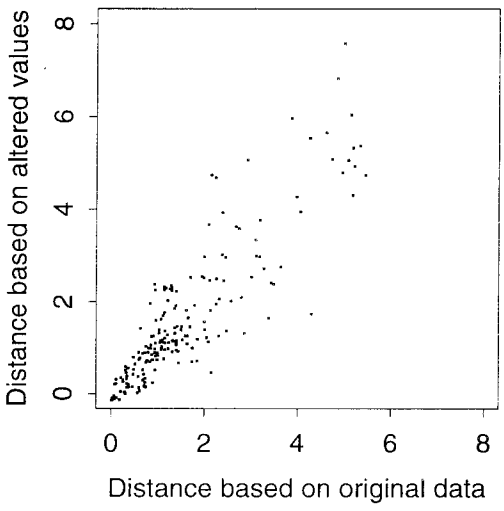


Figure 9.
A comparative plot of squared distances between all pairs of books in the reduced spaces based on correspondence analysis.

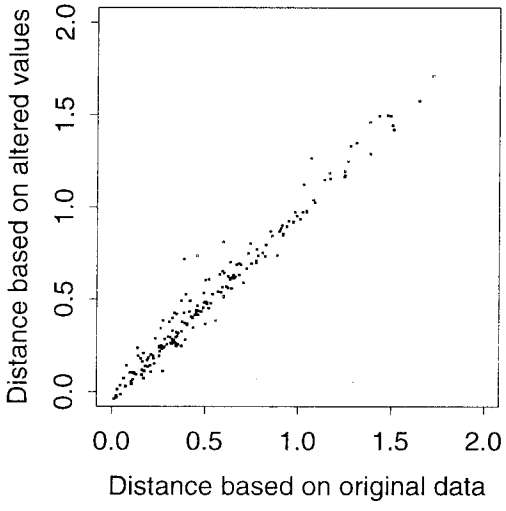


Figure 10.
A comparative plot of squared distances between all pairs of books in the reduced spaces based on Hellinger distance analysis.

Table 4.2.9.

	Chisquare Distance			Hellinger Distance		
	dim1	dim2	dim3	dim1	dim2	dim3
a	−0.62310	0.30413	−0.53463	−0.35082	−0.04632	0.42925
b	0.63345	0.41500	0.44316	0.25096	−0.37625	−0.40565
c	0.90486	0.78379	−0.17802	0.22985	−0.60374	−0.08540
d	−0.36427	0.36470	−0.12611	−0.23454	−0.16742	0.01739
e	0.20621	0.05647	0.54414	0.34853	0.01585	−0.32928
f	1.23299	0.45214	0.27035	0.59591	−0.20402	−0.28683
g	−0.16974	−0.27626	0.25537	−0.08445	0.24917	−0.09573
h	−0.18352	−0.18729	0.09783	−0.07908	0.11818	0.00148
i	0.85607	−0.49586	−0.21672	0.73058	0.07206	0.07026
j	1.35943	−0.06808	0.07459	0.77422	−0.01788	−0.04783
k	0.61122	0.01365	−0.60327	0.28646	−0.20830	0.40764
l	0.32350	−0.41447	−0.39537	0.23929	0.02213	0.24724
m	0.58680	0.20860	0.65792	0.17707	0.01371	−0.36016
n	1.20448	0.51816	0.07444	0.32243	−0.27238	−0.09222
o	0.47616	−1.60929	−0.73075	0.61206	0.41339	0.46017
p	0.87199	−0.26044	−0.37985	0.72806	−0.03491	0.09187
q	−0.44497	0.44631	−0.57047	−0.27936	−0.25639	0.41829
r	0.34209	0.56913	−0.04981	0.10278	−0.49844	−0.07991
s	−0.10036	0.60181	−0.15749	−0.14779	−0.45659	−0.10069
t	1.87687	0.57376	0.28038	0.78353	−0.24748	−0.19823

5. CONCLUDING REMARKS

A general theory is developed for plotting high dimensional "population by variable" data, i.e. measurements made on a set of characteristics of given populations, in a low dimensional Euclidean space. A first step in such a problem is the specification of the basic metric space in which the populations can be represented as points using

the entire data, and a characterization of the configuration of the points in terms of distances between points. The second is the development of methodology for transforming the points from the basic space to a low dimensional Euclidean space with the usual definition of distance preserving the configuration of points to the extent possible. The choices of the basic space and the distance function between points have to be made on practical considerations depending on the problem under investigation. A closed form solution is obtained when the basic space is a vector space endowed with an inner product and the associated norm. Some examples are given involving measurements on continuous and discrete variables.

When we have data in the form of frequencies of individuals of a population under different categories of an attribute, a well known method for dimensionality reduction for representing, say the populations, is correspondence analysis. The basic space in this case is a vector space where each population is represented by the vector of relative frequencies of the different categories of an attribute and distance between vectors is defined by a chisquare type formula. Such a distance function is not an intrinsic measure of difference between two populations as it depends not only on the differences between their relative frequencies, but also on the average relative frequencies computed from the set of populations under study. Thus the configuration of any subset of populations depends on what other populations are included in the analysis, and also on the relative numbers of individuals observed from each population. An alternative approach of representing a population by the vector of the square roots of relative frequencies and defining distance between two populations by the Hellinger formula does not have the drawbacks associated with the chisquare type formula. In addition, the new analysis has the same advantage of providing tests of significance for homogeneity of the populations as in correspondence analysis based on the chisquare formula.

It may be contended that CA is meant to be used for the analysis of contingency tables with dichotomized data using two attributes like hair color and eye color (as originally demonstrated by R.A. Fisher), and not for the analysis of population by variable data where anomalies of the type described in the paper may occur. However, one finds in published literature more examples of the latter type of data analyzed through CA. Further, even with attribute data, if the configurations of the column (or row) profiles for two different populations (with possibly different marginal distributions) are to be compared, HD analysis is more appropriate than the CA. It is the author's opinion that the choice of a distance measure between populations (row or column profiles) must depend on the nature of the data and the purpose of analysis. Prescription to use a particular distance as in the CA may be misleading. Distance measures other than the chisquare and Hellinger types may be more appropriate in some situations. For a purely exploratory data analysis, it is possible that a wide variety of distance measures reveal similar configurations of the populations in terms of clustering and inter cluster relationships.

Between the choices of chisquare and Hellinger distances, the latter seems to offer some advantages, as the latter has similar theoretical properties as the former and in addition it is defined as an intrinsic function of two population profiles independent of what other populations are included in a study.

A recent technical report by Rios, Villarroya and Oller (1994) discusses the same problem as in the present paper, viz., simultaneous representation of populations and random variables, under the assumption of an underlying parametric model.

The method, referred there as *Intrinsic Data Analysis*, is based on the Riemannian structure given by the Fisher information metric and its corresponding distance, the *Rao distance*. The statistical populations are viewed as points on a Riemannian manifold and the random variables with finite expectation, as vector fields, namely, the gradient of the random variable mean value, or, by integration, a bundle of curves on the manifold.

Then, assuming certain additional regularity conditions, a reference point on the manifold is selected as the statistical populations Riemannian center of mass, and the points representing the populations and the curves representing the variables are mapped, through the inverse of the Riemannian exponential map, into the tangent space at the center of mass, which has a Euclidean vector space structure. Then, classical dimension reduction techniques such as principal component analysis can be used to obtain a low dimensional Euclidean space which allows an optimal population representation. Finally, the curves in the tangent space are projected into the low dimensional space obtained.

This method is applied to multivariate normal and multinomial distributions. In the multinomial case, the Rao distance, ρ, between two populations $p_1,\ldots,p_n$ and $q_1,\ldots,q_n$, is proportional to the Bhattacharyya distance

$$\rho = 2arc\cos\sum_{j=1}^{n}\sqrt{p_j q_j}$$

which is a monotone transformation of the Hellinger distance, and thus this method will share some properties with the latter.

APPENDIX

A1. Theorems on optimization.

Let A be an $r_1 \times r_2$ matrix, X_1 be an $r_1 \times s_1$ matrix of rank s_1, X_2 be an $r_2 \times s_2$ matrix of rank s_2 and Σ_1 and Σ_2 be positive definite matrices of orders r_1 and r_2

respectively. Further let

$$P_1 = X_1(X_1'\Sigma_1 X_1)^{-1}X_1'\Sigma_1 \quad \text{(A 1.1)}$$

$$P_2 = X_2(X_2'\Sigma_2 X_2)^{-1}X_2'\Sigma_2 \quad \text{(A 1.2)}$$

be orthogonal projection matrices and

$$F = \Sigma_1^{1/2}(I - P_1)A(I - P_2')\Sigma_2^{1/2}$$

with the s.v.d. (singular value decomposition)

$$F = \lambda_1 U_1 V_1' + \ldots + \lambda_m U_m V_m', \quad \lambda_i > 0, i = 1, \ldots, m \quad \text{(A 1.3)}$$

Theorem A.1

Let A, X_1, X_2, Σ_1 and Σ_2 be as defined above. Then, for any (Σ_1, Σ_2)-invariant norm $\|\cdot\|$, as defined in (2.8),

$$\min_{\Psi_1, \Psi_2, \Gamma} \|A - X_1\Psi_1 - \Psi_2 X_2' - \Gamma\| \quad \text{(A 1.4)}$$

with rank $\Gamma = r \leq \min(r_1 - s_1, r_2 - s_2)$ is attained at

$$X_1\Psi_1 = P_1 A, \Psi X_2' = (I - P_1)AP_2' \quad \text{(A 1.5)}$$

$$\Gamma = \Sigma_1^{-1/2}(\lambda_1 U_1 V_1' + \ldots + \lambda_r U_r V_r')\Sigma_2^{-1/2} \quad \text{(A 1.6)}$$

where λ_i, U_i and V_i are as defined in (A 1.3).

For details of the proof, the reader to referred to Rao (1980, 1985, p.175), where a number of applications of the above theorem are given.

Theorem A.2

Let B an $r_1 \times r_2$ matrix. Then

$$\min_C \|B'\Sigma_1 B - C'C\| \quad \text{(A 1.7)}$$

with rank $C = s \leq$ rank $B'\Sigma_1 B$ is attained for any (Σ_2, Σ_2)-invariant norm at

$$C' = \Sigma_2^{-1/2}(\lambda_1 V_1 : \ldots : \lambda_s V_s) \quad \text{(A 1.8)}$$

where λ_i and V_i are the singular values and vectors from the s.v.d.

$$\Sigma_1^{1/2} B \Sigma_2^{1/2} = \lambda_1 U_1 V_1' + \ldots + \lambda_m U_m V_m' \quad \text{(A 1.9)}$$

Proof

A direct application of Theorem A.1 gives the optimum choice of $C'C$ as

$$C'C = \Sigma_2^{-1/2}(\lambda_1^2 V_1 V_1' + \ldots + \lambda_s^2 V_s V_s')\Sigma_2^{-1/2} \tag{A 1.10}$$

where λ_i^2 and V_i are from the spectral decomposition

$$\Sigma_2^{1/2} B' \Sigma_1 B \Sigma_2^{1/2} = \lambda_1^2 V_1 V_1' + \ldots + \lambda_m^2 V_m V_m' \tag{A 1.11}$$

which can be computed from the s.v.d. (A 1.9). From (A 1.10), it is seen that one choice of C' is as given in (A 1.8).

■

Theorem A.3

Let $X = (X_1 : \ldots : X_m)$ be a $p \times m$ matrix with the $m \times m$ configuration matrix

$$F = (X - \xi 1')' M (X - \xi 1') = (f_{ij}) \tag{A 1.12}$$

and the $m \times m$ inter square distance matrix

$$D = (d_{ij}),\ d_{ij} = (X_i - X_j)' M (X_i - X_j) \tag{A 1.13}$$

where M is a positive definite matrix and

$$\xi = \Sigma w_i X_i,\ w_i \geq 0 \quad \text{and} \quad \Sigma w_i = 1 \tag{A 1.14}$$

Then

(i)

$$\begin{aligned} -2F &= (d_{ij} - d_{.j} - d_{i.} + d_{..}) \\ \text{(A 1.15)} \quad \text{where} \quad d_{i.} &= \Sigma w_j d_{ij},\ d_{.j} = \Sigma w_i d_{ji} \quad \text{and} \quad d_{..} = \Sigma\Sigma w_i w_j d_{ij}. \end{aligned}$$

(ii)

$$D = f1' + 1f' - 2F \tag{A 1.16}$$

where f is the vector of diagonal elements of F.

(iii)

$$\begin{aligned} \text{Trace}\ W^{1/2} F W^{1/2} &= \Sigma w_i (X_i - \xi)' M (X_i - \xi) \\ \text{(A 1.17)} \qquad &= \Sigma\Sigma w_i w_j d_{ij} = d_{..} \end{aligned}$$

where

$$W = \text{Diag}(w_1, \ldots, w_m)$$

The results (A 1.15), (A 1.16) and (A 1.17) are easy to establish.

A2. Homogeneity tests in a contingency table

Let n_{ij} be the number of individuals belonging to category A_i of an attribute out of $n_{.j}$ individuals sampled from population j for $i = 1, \ldots, p$ and $j = 1, \ldots, m$. Then the data can be exhibited as a contingency table (population by attribute) with fixed column totals.

Population by attribute frequencies

Attribute	Populations				Total
	1	2	$\cdots$	m	
A_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{.1}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_p	n_{p1}	n_{p2}	$\cdots$	n_{pm}	$n_{.m}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.m}$	$n_{..}$

Let $\pi_{i|j}$ be the true proportion of individuals with attribute A_i in population j. We would like to test the homogeneity hypothesis

(A 2.1) $$\pi_{i|1} = \pi_{i|2} = \ldots = \pi_{i|m}, \; i = 1, \ldots, p$$

using the statistic

(A 2.2) $$H = 4 \sum_i \sum_j n_{.j} (\sqrt{p_{i|j}} - \sqrt{\hat{\pi}_i})^2$$

where $p_{i|j} = n_{ij}/n_{.j}$ and $\hat{\pi}_i$ is an estimate of the common value π_i of (A 2.1), $i = 1, \ldots, p$. The statistic (A 2.2) based on Hellinger distance (see Freeman and Tukey (1950), and Matusita (1965)) falls in the category of the power divergence tests discussed in great detail by Read and Cressie (1988). The estimates of $\hat{\pi}_i$ generally recommended are

(i) the maximum likelihood (ML) estimate

(A 2.3) $$\hat{\pi}_i = \sum n_{.j} p_{i|j} / n_{..} = n_{i.}/n_{..}$$

and

(ii) the minimum distance (MD) estimate

$$\hat{\pi}_i = \frac{(\sum_j n_{.j}\sqrt{p_{i|j}})^2}{\sum_i(\sum_j n_{.j}\sqrt{p_{i|j}})^2} \tag{A 2.4}$$

which ensure asymptotic χ^2 distribution on $(p-1)(m-1)$ degrees of freedom for H. We suggest another estimate

(iii) the unrestricted MD estimate

$$\hat{\pi}_i = (\sum_j n_{.j}\sqrt{p_{i.j}}/n_{..})^2 \tag{A 2.5}$$

which although does not satisfy the natural restriction $\sum\hat{\pi}_i = 1$, ensures the same asymptotic distribution for H as in (i) and (ii) above. The only condition needed is that $n_{.j} \to \infty$ for each j. When $\hat{\pi}_i$ as in (A 2.5) is used in H of (A 2.2), H ceases to be in the class of power divergence tests.

All the stated results follow from the general theorems proved in Rao (1973, pp. 382-388) and Read and Cressie (1988), or simply by noting that

$$(\sqrt{p} - \sqrt{\hat{\pi}})^2 = \frac{(p-\hat{\pi})^2}{\hat{\pi}}\left(1 + \frac{\sqrt{p}}{\sqrt{\hat{\pi}}}\right)^{-2} \tag{A 2.6}$$

which has the Pearson χ^2 component as the dominant factor with the other factor tending to 1/4 as $p/\hat{\pi} \to 1$. We need only use the expression (A 2.6) for each term in (A 2.2) to show the equivalence of H with Pearson's χ^2.

Note 1. With the choice of the ML estimate (A 2.3) for $\hat{\pi}$, the statistic H can be written as

$$H = \sum_i\sum_j(\sqrt{n_{ij}} - \sqrt{n_{i.}n_{.j}/n_{..}}) \tag{A 2.7}$$

which is symmetric with respect to interchange of rows and columns. The H based on the estimates (A 2.4) or (A 2.5) does not have this property.

Note 2. The choice of $\hat{\pi}_i$ as in (A 2.5), which is used in the analysis of Section 4.2 seems to be appropriate even in tests of significance. For instance, a natural test for homogeneity when $m = 2$ is

$$4\frac{n_{.1}n_{.2}}{n_{.1}+n_{.2}}\sum_{i=1}^{p}(\sqrt{p_{i|1}} - \sqrt{p_{i|2}})^2 \tag{A 2.8}$$

which is equal to

$$H = 4[n_{.1}\sum_{i=1}^{p}(\sqrt{p_{i|1}} - \hat{\pi}_i)^2 + n_{.2}\sum_{i=1}^{p}(\sqrt{p_{i|2}} - \sqrt{\hat{\pi}_i})^2] \tag{A 2.9}$$

only when $\hat{\pi}_i$ is as chosen in (A 2.5).

The research work of this paper is sponsored by the Army Research Office under Grant DAAH04-93-G-0030.

REFERENCES

[1] **Benzécri, J.P.** (1992). *Correspondence Analysis Handbook.* Marcel Dekker, Inc., New York.

[2] **Cavallis-Sforza, L.L.** (1991). "Genes, peoples and languages". *Scientific American*, **265**, 104–110.

[3] **Chernoff, H.** (1973). "The use of faces to represent points in k-dimensional space graphically". *J. Amer. Statist. Assoc.*, **68**, 361–368.

[4] **Cressie, N.A.C.** and **Read, T.R.C.** (1984). "Multinomial goodness-of-fit tests". *J. Roy. Statist. Soc.*, **46**, 440–464.

[5] **Eckart, C.** and **Young, G.** (1936). "The approximation of one matrix by another of lower rank". *Psychometrika*, **1**, 211–308.

[6] **Eslava-Gomez, G.** and **Marriott, F.H.C.** (1993). "Criteria to represent groups in the plane when the grouping is unknown". *Biometrics*, **49**, 1088–1098.

[7] **Freeman, M.F.** and **Tukey, J.W.** (1950). "Transformations related to the angular and the square root". *Ann. Math. Stat.*, **21**, 607–611.

[8] **Gabriel, K.R.** (1971). "The biplot graphical display of matrices with applications to principal component analysis". *Biometrika*, **58**, 453–467.

[9] **Gifi, A.** (1981, 1990). *Nonlinear Multivariate Analysis.* New York: John Wiley.

[10] **Gower, J.C.** (1993). "Recent advances in biplot methodology". In *Multivariate Analysis: Future Directions 2* (Eds. C.M. Cuadras and C.R. Rao), North Holland, 295–325.

[11] **Greenacre, M.J.** (1984). *Theory and Applications of Correspondence Analysis.* London: Academic Press.

[12] **Greenacre, M.J.** (1993). "Biplots in correspondence analysis". *J. Applied Statistics*, **20**, 251–269.

[13] **Kruskal, J.B.** and **Wish, M.** (1978). *Multidimensional Scaling.* Sage Publications.

[14] **Mahalanobis, P.C.** (1936). "On the generalized distance in statistics". *Proc. Nat. Inst. Sci. India*, **12**, 49–55.

[15] **Mahalanobis, P.C., Mazumdar, D.N.** and **Rao, C.R.** (1949). "Anthropometric survey of United Provinces, 1941. A statistical study". *Sankhya*, **9**, 90–324.

[16] **Matusita, Kameo** (1955). "Decision rules, based on the distance, for problems of fit, two samples, and estimation". *Ann. Math. Stat.*, **26**, 631–640.

[17] **Nishisato, S.** (1980). *Analysis of Categorical Data: Dual Scaling and its Applications*. University of Toronto Press, Toronto, Canada.

[18] **Rao, C.R.** (1945). "Information and accuracy attainable in the estimation of statistical parameters". *Bull. Cal. Math. Soc.*, **37**, 81–91.

[19] **Rao, C.R.** (1947). "The problem of classification and distance between two populations". *Nature*, **159**, 30.

[20] **Rao, C.R.** (1948). "The utilization of multiple measurements in problems of biological classification (with discussion)". *J. Roy. Statist. Soc.*, **Series B10**, 159–193.

[21] **Rao, C.R.** (1958). "Some statistical methods for comparison of growth curves". *Biometrics*, **14**, 1–17.

[22] **Rao, C.R.** (1964). "The use and interpretation of principal component analysis in applied research". *Sankhya*, **26**, 329–357.

[23] **Rao, C.R.** (1973). *Linear Statistical Inference and its Applications*, 2nd Edition, New York : Wiley.

[24] **Rao, C.R.** (1979). "Separation theorems for singular values of matrices and their applications in multivariate analysis". *J. Multivariate Analysis*, **9**, 362–377.

[25] **Rao, C.R.** (1980). "Matrix approximations and reduction of dimensionality in multivariate statistical analysis". In *Multivariate Analysis V* (Ed. P.R. Krishnaiah), Amsterdam: North Holland, 3–22.

[26] **Rao, C.R.** (1985). "Tests for dimensionality and interaction of mean vectors under general and reducible covariance structures". *J. Multivariate Analysis*, **16**, 173–184.

[27] **Rao, C.R.** (1987). "Prediction of future observations in growth curve type models". *J. Statistical Science*, **2**, 434–471.

[28] **Read, T.R.C.** and **Cressie, N.A.C.** (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

[29] **Rios, M., Villarroya, A.** and **Oller, J.M.** (1994). "Intrinsic Data Analysis: A method for the simultaneous representation of populations and variables". *Mathematics preprint series*, **160**. Universitat de Barcelona.

[30] **Von Neumann, J.** (1937). "Some matrix inequalities and metrization of metric spaces". *Tomsk. Univ. Rev.*, **1**, 286–299.

[31] **Wegman, E.J.** (1990). "Hyperdimensional data analysis using parallel coordinates". *J. Amer. Statist. Assoc.*, **85**, 664–675.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 65-91, 1995

A REVIEW OF THE RESULTS ON THE STEIN APPROACH FOR ESTIMATORS IMPROVEMENT

V.G. VOINOV* and M.S. NIKULIN†

Since 1956, a large number of papers have been devoted to Stein's technique of obtaining improved estimators of parameters, for several statistical models. We give a brief review of these papers, emphasizing those aspects which are interesting from the point of view of the theory of unbiased estimation.

Key words: Admissible Estimator, Berger's Estimator, Equivariant Estimator, James-Stein Estimator, Minimax Estimator, Quadratic Loss Function, Shrinkage Estimator.

1. INTRODUCTION

Let we need not an unbiased estimator but an estimator with the smallest risk for a prescribed loss function. Much help in the situation gives the well known Stein phenomenon discovered by Charles Stein (1956). He was the first showed that the best unbiased estimator for the mean of a multivariate normal distribution can be improved upon by so-called shrinkage estimators. Since that time, a large number of papers have been published, proposing classes of improved estimators for different probability models (see e.g. a monograph by Hoffmann (1992)). We shall give a brief account of these works emphasizing those questions which are related to the unbiased estimation point of view.

* Voinov, V.G. Physical Technical Institute, Alma-Ata, Kazakhstan.

† Nikulin, M.S. Mathématiques Stochastiques, Université Bordeaux 2, France.
Steklov Mathematical Institute, St Petersbourg, Russie.

– Article rebut el juny de 1994.

– Acceptat el setembre de 1994.

2. ESTIMATING THE MEAN VECTOR

Let us start with the problem of estimating the mean vector $\mu = (\mu_1, \cdots, \mu_p)^T$ of a p-dimensional normal distribution with known covariance matrix equal to identity, **I**, under the loss function

$$L(\mu, \hat{\mu}) = (\hat{\mu} - \mu)^T \mathbf{A} (\hat{\mu} - \mu), \tag{1}$$

where **A** is a positive-semidefinite matrix, from a single observation $\mathbf{X} = (X_1, X_2, \cdots, X_p)^T$.

It is known that the observation **X** itself is a maximum likelihood, minimum variance unbiased, invariant for a wide class of loss functions and minimax estimator for μ. Nevertheless, this estimator is inadmissible in the sense that there exist other estimators whose risks are everywhere smaller than the risk of **X**. To consider such estimators $\hat{\mu}(\mathbf{X})$ let us follow for a while the Stein's lectures (1977, 1986). Without loss of generality the matrix **A** may be considered to be a diagonal matrix $\mathbf{A} = \operatorname{diag}(\alpha_1, \cdots, \alpha_p)$, where

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_p \geq 0.$$

Estimators $\hat{\mu}(\mathbf{X})$ may be constructed with the help of the following identity

$$\mathbf{E}_\mu \sum_{i=1}^{p} \alpha_i [X_i + g_i(\mathbf{X}) - \mu_i]^2 = \mathbf{E}_\mu \sum_{i=1}^{p} \alpha_i [1 + g_i^2(\mathbf{X}) + 2g_{ii}(\mathbf{X})], \tag{2}$$

where $\mathbf{g} = (g_1, \cdots, g_p)^T$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^n$,

$$g_{ii}(\mathbf{X}) = \frac{\partial}{\partial X_i} g_i(\mathbf{X}), \quad g_i : \mathbb{R}^n \to \mathbb{R}$$

and

$$\mathbf{E}_\mu |g_{ii}(\mathbf{X})| < \infty, \quad \mathbf{E}_\mu g_i^2(\mathbf{X}) < \infty.$$

A remarkable feature of equality (2) is that the expectand on the right hand side of it does not involve the unknown parameter μ.

The proof of identity (2) is based on

Lemma 1

Let Y be a standard normally distributed random variable and $g : \mathbb{R} \to \mathbb{R}$ be an absolutely continuous on $\mathbb{R}$ function such that $\mathbf{E}|g'(Y)| < \infty$. Then

$$\mathbf{E} g'(Y) = \mathbf{E} Y g(Y). \tag{3}$$

Proof

$$\begin{aligned}
\mathbf{E}g'(Y) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g'(y)e^{-\frac{y^2}{2}}dy = \\
&= \frac{1}{\sqrt{2\pi}}\left\{\int_0^{\infty} g'(y)dy\int_y^{\infty} xe^{-\frac{x^2}{2}}dx - \int_{-\infty}^{0} g'(y)dy\int_{-\infty}^{y} xe^{-\frac{x^2}{2}}dx\right\} = \\
&= \frac{1}{\sqrt{2\pi}}\left\{\int_0^{\infty} xe^{-\frac{x^2}{2}}dx\int_0^{x} g'(y)dy - \int_{-\infty}^{0} xe^{-\frac{x^2}{2}}dx\int_x^{0} g'(y)dy\right\} = \\
&= \frac{1}{\sqrt{2\pi}}\left\{\int_0^{\infty} (g(x)-g(0))xe^{-\frac{x^2}{2}}dx + \int_{-\infty}^{0} (g(x))-g(0))xe^{-\frac{x^2}{2}}dx\right\} = \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xg(x)e^{-\frac{x^2}{2}}dx = EYg(Y).
\end{aligned}$$

To prove (2) we have to prove that

$$\mathbf{E}_\mu[X_1 + g_1(\mathbf{X}) - \mu_1]^2 = 1 + \mathbf{E}_\mu[g_1^2(\mathbf{X}) + 2g_{11}(\mathbf{X})].$$

Assuming the function $g_1 : \mathbb{R}^n \to \mathbb{R}$ to be absolutely continuous by x_1 and

$$\mathbf{E}_\mu|g_{11}(\mathbf{X})\ | < \infty,$$

then by Lemma 1

$$\mathbf{E}_\mu g_{11}(\mathbf{X}) = \mathbf{E}_\mu(X_1 - \mu_1)g_1(\mathbf{X})$$

and

$$\mathbf{E}_\mu[X_1 + g_1(\mathbf{X}) - \mu_1]^2 = 1 + \mathbf{E}_\mu[g_1^2(\mathbf{X}) + 2g_{11}(\mathbf{X})]. \tag{4}$$

It follows from (2) that if

$$\sum_{i=1}^{p} \alpha_i[g_i^2(\mathbf{X}) + 2g_{ii}(\mathbf{X})] < 0, \tag{5}$$

then

$$\mathbf{E}_\mu \sum_{i=1}^{p} \alpha_i[X_i + g_i(\mathbf{X}) - \mu_i]^2 < \sum_{i=1}^{p} \alpha_i \le \mathbf{E}_\mu \sum_{i=1}^{p} \alpha_i(X_i - \mu_i)^2.$$

Since inequality (5) does not depend on the unknown parameter μ, the estimator

$$\hat{\mu} = \mathbf{X} + \mathbf{g}(\mathbf{X})$$

is everywhere in the parameter space better (in the sence of the smaller risk) than the usual estimator

$$\hat{\mu}^0 = \mathbf{X}.$$

Thus, the problem reduces to the appropriate choice of the function g satisfying differential inequality (5).

The first result in this direction has been obtained by James and Stein (1960).

Denoting

$$\mathbf{x}^T\mathbf{y} = \sum_{i=0}^{p} x_i y_i, \qquad |\mathbf{x}|^2 = \mathbf{x}^T\mathbf{x}$$

and

$$\nabla = \left(\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_p}\right)^T$$

for the special case $\mathbf{A} = \mathbf{I}$ we may rewrite identity (2) as follows

$$\mathbf{E}_\mu|\mathbf{X} + \mathbf{g}(\mathbf{X}) - \mu|^2 = p + \mathbf{E}_\mu[|\mathbf{g}(\mathbf{X})| + 2\nabla^T\mathbf{g}(\mathbf{X})], \tag{6}$$

where

$$\nabla^T\mathbf{g}(\mathbf{X}) = \sum_{i=1}^{p} \frac{\partial g_i(\mathbf{X})}{\partial X_i}.$$

Consider the case when $\hat{\mu}$ is equivariant with respect to the group of orthogonal transformations of $\mathbb{R}^n$, i.e. there exists a function $h : \mathbb{R}^n \to \mathbb{R}$ such that

$$\mathbf{g}(\mathbf{X}) = h(|\mathbf{X}|^2)\mathbf{X}.$$

In this case (6) gives

$$\begin{aligned} \mathbf{E}_\mu|\mathbf{X} + \mathbf{g}(\mathbf{X}) - \mu|^2 &= p + \mathbf{E}_\mu[h^2(|\mathbf{X}|^2)|\mathbf{X}|^2 + \\ &+ 2ph(|\mathbf{X}|^2) + 4|\mathbf{X}|^2 h'(|\mathbf{X}|^2)]. \end{aligned} \tag{7}$$

Let

$$h(t) = -\frac{c}{t},$$

where $c > 0$, then for $p \geq 3$

$$\mathbf{E}_\mu\left|\left(1 - \frac{c}{|\mathbf{X}|^2}\right)\mathbf{X} - \mu\right|^2 = p + \mathbf{E}_\mu \frac{c^2 - 2(p-2)c}{|\mathbf{X}|^2}. \tag{8}$$

The right hand side of (8) is minimized by

$$c = p - 2$$

and the minimum value is

$$\mathbf{E}_\mu \left| \left(1 - \frac{c}{|\mathbf{X}|^2}\right)\mathbf{X} - \mu \right|^2 = p - (p-2)^2 \mathbf{E}_\mu \frac{1}{|\mathbf{X}|^2}.$$

This will be less than the constant risk p of the estimator $\mathbf{X}$ provided that $p > 2$. Thus the James and Stein estimator reads

$$\hat{\mu} = \left(1 - \frac{p-2}{|\mathbf{X}|^2}\right)\mathbf{X}. \tag{9}$$

James and Stein showed that the risk of $\hat{\mu}$ at $\hat{\mu} = 0$ under the loss function

$$L(\hat{\mu}, \mu) = |\hat{\mu} - \mu|^2$$

is equal to 2. For $p > 2$ this is a substantial improvement over the constant risk p of the estimator $\hat{\mu}^0$. Even for $|\mu|^2/p = 2$ the estimator (9) gives the risk about 25 per cent lower than that of $\hat{\mu}^0 = \mathbf{X}$ at $\mu = 0$. Nevertheless, statisticians and users do not "rush to embrace this considerable improvement ... " as Efron (1975) said. He considered several reasons for this fact. The most interesting reason in our opinion is the following one. "If the different μ_i refer to obviously disjoint problems (e.g., μ_1 is the speed of light, μ_2 is the price of tea in China, μ_3 is the efficacy of new treatment for psoriasis, etc.) combining the data can produce a definitely uncomfortable feeling in the statistician". The situation seems to be paradoxical, but "if we really aren't interested in $\mu_2, \mu_3, \cdots, \mu_p$, just μ_1, then

$$L_1(\hat{\mu}, \mu) = (\hat{\mu}_1 - \mu_1)^2$$

seems to be a more reasonable loss function than $|\hat{\mu} - \mu|^2$. It is not true that

$$\mathbf{E}_\mu(\hat{\mu}_1 - \mu_1)^2 < \mathbf{E}_\mu(X_1 - \mu_1)^2$$

for all μ. As a matter of fact

$$\mathbf{E}_\mu(\hat{\mu}_1 - \mu_1)^2 / \mathbf{E}_\mu(X_1 - \mu_1)^2$$

can be as large as about $\sqrt{p}$ for certain configurations of $\mu_2, \mu_3, \cdots, \mu_p$ (namely $\mu_1 = \sqrt{p}, \mu_2 = \cdots = \mu_p = 0$)". The paradox seems to be disappeared. Since it is not obvious that the loss function $(\hat{\mu_1} - \mu_1)^2$ is the best one for the situation, the "uncomfortable feeling" remains. That is why many attempts to resolve the paradox and to reduce the maximum component risk

$$R^*(\mu) = \max_i \mathbf{E}_\mu(\hat{\mu}_i - \mu_i)^2$$

have been undertaken.

3. THE BAYESIAN APPROACH

The best heuristic explanation of the paradox contains a Bayesian argument. If the μ_i are a priori independent $N(0,\tau^2)$, then (9) can be considered as an empirical Bayes estimator (Efron and Morris (1973)). More details about the relation of the Stein's effect to the empirical Bayes approach and related problems readers may find in Efron and Morris (1976b), Berger and Srinivasan (1977), Haff (1980), Berger (1982a), Chen (1988) and others. Another explanation interpretes (9) as a smoothed version of a weighted mean of 0 and $\hat{\mu}^0$, i.e. one uses $\hat{\mu} = 0$ or $\hat{\mu}_i = X_i$ depending on the outcome of the test verifying the null hypothesis that $\mu = 0$ (Lehmann (1983)). An interesting approach to the problem gave Stigler (1990) who had considered Stein estimation as a regression problem. A simple geometrical argument of the possibility to improve upon the best invariant estimator via shrinkage estimation have been given by Brandwein and Strawderman (1990). It is worth to note the work of Beran (1992) who showed that under quadratic loss the Stein and the positive-part Stein estimators of μ are both approximately admissible and approximately minimax on large compact balls about the shrinkage point as $p \to \infty$.

4. MINIMIZING THE RISK

To this end we would also like to say that actually one should choose such an estimator which corresponds a goal he pursues. This depends also on what one intends to do with an estimator. If he needs an unique point estimator it is preferable to use one which minimizes the risk. If on the contrary he will perform averaging of many estimators the unbiased one might be the best (see §1, Ch.3 in Voinov and Nikulin (1993)).

Let $\mathbf{X} \sim N(\hat{\mu}, \sigma^2\mathbf{I})$, σ^2 known. In this case

$$\hat{\mu} = \left(1 - \frac{(p-2)\sigma^2}{|\mathbf{X}|^2}\right)\mathbf{X}. \tag{10}$$

If σ^2 is unknown and one possesses an estimator S_n of σ^2 independent of $\mathbf{X}$ and distributed like $\sigma^2\chi^2_n$, then (James & Stein (1960))

$$\hat{\mu} = \left(1 - \frac{(p-2)S_n}{(n+2)|\mathbf{X}|^2}\right)\mathbf{X}. \tag{11}$$

In the most realistic situation $\mathbf{X} \sim N(\mu, \Sigma)$, Σ being unknown. Let there is an independent estimator $\mathbf{S}$ for Σ, which is distributed as $W_{n-1}(\mathbf{s};\Sigma)$, a Wishart distribution.

James & Stein proposed the estimator

$$\hat{\mu} = \left(1 - \frac{(p-2)}{(n-p+3)\mathbf{X}^T\mathbf{S}^{-1}\mathbf{X}}\right)\mathbf{X}. \tag{12}$$

An explicit formula for the risk of the James and Stein estimator is available in Egerton and Laycock (1982).
Baranchik (1964, 1970) showed that for $\mathbf{X} \sim N(\mu, \mathbf{I})$ the estimator

$$\hat{\mu} = \left(1 - \frac{r(|\mathbf{X}|^2)}{|\mathbf{X}|^2}\right)\mathbf{X} \tag{13}$$

under the loss function

$$L(\hat{\mu}, \mu) = |\hat{\mu} - \mu|^2$$

dominates $\hat{\mu}^0$ if

$$0 < r(|\mathbf{X}|^2) < 2(p-2)$$

and $r(.)$ is a nondecreasing function.

A more general minimax condition for $r(.)$ has been given by Efron & Morris (1976a). Let $\mathbf{X}$ have a p-variate normal distribution $N(\mu, \mathbf{D})$, where μ is unknown and $\mathbf{D}$ is a known nonsingular covariance matrix. Let d_L be the largest eigenvalue of $\mathbf{D}$. Bock (1975) has shown that for $p > 2$

$$\hat{\mu} = \left(1 - \frac{cr(\mathbf{X}^T\mathbf{D}^{-1}\mathbf{X})}{\mathbf{X}^T\mathbf{D}^{-1}\mathbf{X}}\right)\mathbf{X} \tag{14}$$

is a minimax spherically symmetric estimator for μ if $0 < c \leq 2((tr\mathbf{D})d_L^{-1} - 2)$ and $r(.)$ is a monotone non-decreasing function. Some generalizations of the James and Stein estimator (9) have been summarized by Akai (1988). Considering estimators

$$\hat{\mu} = \left(1 - \frac{b}{|\mathbf{X}|^2}\right)\mathbf{X}, \tag{15}$$

where $0 < b < 2(p-2)$, Akai (1988) showed that the estimator

$$\hat{\mu} = \left(1 - \frac{b}{|\mathbf{X}|^2} + \frac{g(|\mathbf{X}|)}{|\mathbf{X}|^{2r}}\right)\mathbf{X} \tag{16}$$

has under the same loss function as above the smaller risk than that of (15), if r is an integer such that

$$r > 1 \text{ and } p > 2(2r-1) \geq b$$

and $g(s)$ satisfies either condition (i) or (ii):

(i) $g(s)s^{-1/\alpha}$ is nonincreasing for $\alpha > 0$, and $g(s)$ is nondecreasing and satisfies

$$0 < g(s) \leq 2^r(b-(p-2r+2/\alpha))\Gamma(\tfrac{p}{2}-r)/\Gamma(\tfrac{p}{2}-2r+1)$$

for $\alpha > 0$ and $b > p-2r+2/\alpha$,

(ii) $g(s)$ is nonincreasing and satisfies

$$0 > g(s) > 2^r/(b-(p-2r))\Gamma(\tfrac{p}{2}-r)/\Gamma(\tfrac{p}{2}-2r+1)$$

for $b < p-2r$.

Let, for example, $g(s) = d$, where d is a constant, and $b = p-2$. Then the risk of the estimator

$$\hat{\mu} = \left(1 - \frac{p-2}{|\mathbf{X}|^2} + \frac{d}{|\mathbf{X}|^{2r}}\right)\mathbf{X} \tag{17}$$

for

$$L(\hat{\mu},\mu) = |\hat{\mu}-\mu|^2$$

is uniformly smaller than that of $\hat{\mu}^0$ if

$$1 < r < (p+2)/4$$

and

$$0 < d \leq 2^{r+1}(r-1)\Gamma(\tfrac{p}{2}-r)/\Gamma(\tfrac{p}{2}-2r+1).$$

Consider the special form

$$\hat{\mu} = \left(1 - \frac{b}{a+|\mathbf{X}|^2}\right)\mathbf{X}, \qquad a > 0, \quad 0 < b < 2(p-2), \tag{18}$$

of Baranchik's estimator (13). Let $p > 6$ and $b = p-2$, then (Akai (1988)) the risk of (18) under the squared error loss function is uniformly smaller than that of $\hat{\mu}^0$ if

$$0 < a < 4(p-6)/(p-2).$$

Akai also showed that for $p > 6$ the risk of

$$\hat{\mu} = \left(1 - \frac{b}{a+|\mathbf{X}|^2} + \frac{c}{|\mathbf{X}|^4}\right)\mathbf{X} \tag{19}$$

for

$$L(\hat{\mu},\mu) = |\hat{\mu}-\mu|^2$$

is uniformly smaller than that of (18) if the constants a and c satisfy the conditions:

(i) for $b > p-4$,

$$0 \le a < \frac{[b-(p-4)](p-6)}{p-4}$$

and

$$0 < c \le 2\frac{[b-(p-4)](p-6)^2 - a(p-4)(p-6)}{a+p};$$

(ii) for $b < p-4$

$$a \ge 0 \text{ and } 0 > c \ge \frac{[b-(p-4)](p-6)^2 - a(p-4)(p-6)}{a+p}.$$

Suppose now that $p > 6$. Then the risk of

$$\hat{\mu} = \left(1 - \frac{b}{a+|\mathbf{X}|^2} + \frac{h(|\mathbf{X}|^2)}{(d_1+|\mathbf{X}|^2)^2}\right)\mathbf{X} \tag{20}$$

for

$$L(\hat{\mu},\mu) = |\hat{\mu}-\mu|^2$$

is uniformly smaller than that of (18) if $h(s)$, a and d_1 satisfy either condition (i) or (ii):

(i) $h(s)(d_1+s)^{-1/\alpha}$ is nonincreasing for $\alpha > 0$ and $d_1 > 0$ and $h(s)$ is nondecreasing and satisfies

$$0 < h(s) \le 2\{(b-(p-4+2/\alpha))(p-6) - d_1 p\} \quad \text{for } \alpha > 0,$$

$$b > p-4+2/\alpha \quad \text{and} \quad a \le d_1 < (b-(p-4+2/\alpha))(p-6)/p,$$

(ii) $h(s)$ is nonincreasing and satisfies

$$0 > h(s) \ge 2\{(b-(p-4))(p-6) - d_1 p\} \quad \text{for } a \ge d_1 \ge 0 \quad \text{and } b < p-4.$$

The estimator (9) modifies the usual estimator $\hat{\mu}^0$ by "shrinking" it towards the origin, the more so the closer $\mathbf{X}$ is itself to the origin. For $|\mathbf{X}|^2 < (p-2)$ the estimator $\hat{\mu}$ is actually passed the origin in the direction opposite $\mathbf{X}$. If such a behaviour is undesirable one may introduce the positive-part James & Stein estimator

$$\hat{\mu} = \left(1 - \frac{p-2}{|\mathbf{X}|^2}\right)_+ \mathbf{X}, \tag{21}$$

where $a_+ = \max(a,0)$. Baranchik (1964) showed that the estimator (21) dominates $\hat{\mu}$, defined by (9). The same is true for the estimator (see (11))

$$\hat{\mu}_1 = \left(1 - \frac{(p-2)S_n}{(n+2)|\mathbf{X}|^2}\right)_+ \mathbf{X}. \tag{22}$$

Both estimators are not admissible, but no estimators uniformly better than (21) and (22) are known. An explicit formula for the risk of the estimator (22) is available in Robert (1988).

Remark 1. Since all James-Stein like estimators are all inadmissible, there are different approaches for their improvement. The last example of such an improvement has been given e.g. by Berry (1994), who used for this the improved variance estimator of variance estimator. The estimator constructed is

$$\hat{\mu} = \left(1 - \frac{r(F)}{F}\right) X,$$

where

$$r(F) = \begin{cases} \frac{p-2}{n+2}, & F \geq \frac{p}{n+2}, \\ \frac{p-2}{n+p+2}(1+F), & F < \frac{p}{n+2} \end{cases}$$

and $F = |X|^2 / S_n$.

This estimator dominate the traditional James-Stein estimator (22), but the domination does not hold for the positive part version of these estimators.

5. OTHER ASPECTS ON THE IMPROVEMENT OF THE RISK

The improvement in the risk obtained by Stein estimators is significant only if μ_i are close to the point towards which these estimators shrink. For a heavy-tailed prior distribution the Stein estimators may give little improvement over $\hat{\mu}^0$. Stein (1981), Dey & Berger (1983) and Alam and Mitra (1986) gave some modifications of (10) which help to overcome the difficulty. For example, Alam and Mitra (1986) instead of (9) proposed the estimator which is given component-wise by

$$\eta_i(\mathbf{X}) = \left(1 - \frac{(p-2)}{T_i}\right) X_i, \tag{23}$$

where

$$T_i = (\alpha - 1)X_i^2 + |\mathbf{X}|^2, \quad \alpha \leq p/2.$$

This estimator is minimax and reduces to (9) for $\alpha = 1$. Due to the term $(\alpha - 1)X_i^2$ the estimator naturally reduces the maximum component risk

$$R^*(\mu) = \max_i \mathbf{E}_\mu(\eta_i(\mathbf{X}) - \mu_i)^2.$$

Alam and Mitra (1986) have shown that the positive part estimator

$$\eta_i^+(\mathbf{X}) = \left(1 - \frac{(p-2)}{T_i}\right)_+ X_i \tag{24}$$

dominates $\eta_i(\mathbf{X})$, component-wise.

When σ^2 is not known and there is an estimator S_n of σ^2 independent of $\mathbf{X}$ and distributed as $\sigma^2\chi_n^2$ the estimator

$$\eta_i(\mathbf{X}) = \left(1 - \frac{(p-2)S_n}{(n+2)T_i}\right) X_i$$

is again minimax for $\alpha \leq p/2$. Alam and Mitra (1986) have given explicit expressions for component risks of (23) and (24).

Let $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ be i.i.d. $\mathbf{N}(\mu, \sigma^2\mathbf{V})$ random vectors, where μ and σ^2 are unknown and $\mathbf{V}$ is a $p \times p$ known positive definite matrix. Suppose the loss function to be

$$L(\hat{\mu}_n, \mu) = (\hat{\mu}_n - \mu)^T \mathbf{Q} (\hat{\mu}_n - \mu), \tag{25}$$

where $\mathbf{Q}$ is a known positive definite matrix and $\hat{\mu}_n = \hat{\mu}_n(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ is an estimator for μ. Hudson (1974) and Berger (1976) introduced the following class of James-Stein estimators

$$\hat{\mu}_n = \lambda + \left(\mathbf{I} - \frac{r(F_n)}{F_n}\mathbf{Q}^{-1}\mathbf{V}^{-1}\right)(\overline{\mathbf{X}}_n - \lambda), \tag{26}$$

where $\lambda \in \mathbb{R}^p$ is a preassigned constant to which $\hat{\mu}_n$ shrinks the estimator $\overline{\mathbf{X}}_n$,

$$\begin{aligned} F_n &= n(\overline{\mathbf{X}}_n - \lambda)^T \mathbf{V}^{-1} \mathbf{Q}^{-1} \mathbf{V}^{-1} (\overline{\mathbf{X}}_n - \lambda) / S_n^2, \\ S_n^2 &= [(n-1)p + 2]^{-1} \sum_{j=1}^{n} (\mathbf{X}_j - \overline{\mathbf{X}}_n)^T \mathbf{V}^{-1} (\mathbf{X}_j - \overline{\mathbf{X}}_n), \quad n \geq 2. \end{aligned}$$

Nickerson (1989) has shown that if $r(.)$ is absolutely continuous,

$$0 < r(F_n) < 2(p-2)$$

and

$$\psi_n(F_n) = F_n^{(p-2)/2} r(F_n) / (2(p-2) - r(F_n))^{1+2b_n}, \quad b_n = (p-2)/((n-1)p+2),$$

is nondecreasing in F_n, then (26) dominates $\overline{\mathbf{X}}_n$ under the loss (25). For every absolutely continuous $r(.)$ he introduced the positive-part estimator

$$\hat{\mu}_{n1} = \lambda + \left(\mathbf{I} - \frac{r(F_n)}{F_n}\mathbf{Q}^{-1}\mathbf{V}^{-1}\right)_+ (\overline{\mathbf{X}}_n - \lambda), \tag{27}$$

where

$$\left(\mathbf{I}-\frac{r(F_n)}{F_n}\mathbf{Q}^{-1}\mathbf{V}^{-1}\right)_+ = \mathbf{P}^{-1}\left[\text{diag}\left(\left(1-\frac{r(F_n)}{F_n}d_1^{-1}\right)_+,\right.\right.$$
$$\left.\left.\cdots,\left(1-\frac{r(F_n)}{F_n}d_p^{-1}\right)_+\right)\right]\mathbf{P},$$

$\mathbf{P}$ is a nonsigular matrix such that $\mathbf{PQ}^{-1}\mathbf{P}^T=\mathbf{I}$ and $\mathbf{PVP}^T=\mathbf{D}=$ diag $(d_1,\cdots,d_p)$. The estimator (27) dominates $\overline{\mathbf{X}}_n$ under the loss function (25).

One may substitute λ in (26) and (27) by another estimator $\hat{\mu}_2$ of μ, then estimators (26) and (27) will shrink $\bar{\mathbf{X}}_n$ towards this arbitrary estimators. The problem of selecting the $\hat{\mu}_2$ has been considered among others by George (1986) and Sengupta (1991).

Let $\mathbf{X}\sim N(\mu,\Sigma)$, where $N(\mu,\Sigma)$ is a p-variate normal distribution with unknown μ and unknown positive-definite covariance matrix Σ. Let there is an independent estimator $\mathbf{S}$ for Σ, which is distributed as $W_{n-1}(\mathbf{s};\Sigma)$. Using the loss function

$$L(\hat{\mu},\mu)=(\hat{\mu}-\mu)^T\mathbf{Q}(\hat{\mu}-\mu),$$

where $\mathbf{Q}$ is a given positive definite matrix, Alam(1977) considered a class of estimators $\hat{\mu}$ of μ of the form

$$\hat{\mu}=\Phi(\mathbf{X}^T\mathbf{S}^{-1}\mathbf{X})\mathbf{X}$$

for a certain class of functions Φ. In particular he has shown that an estimator $\hat{\mu}$ given by

$$\hat{\mu}=\frac{2\nu}{p}\frac{{}_2F_1(\nu+1,\frac{n}{2}+\frac{p}{2}+1;\frac{p}{2}+1;\frac{x}{1+x})}{{}_2F_1(\nu,\frac{n}{2}+\frac{p}{2}+1;\frac{p}{2};\frac{x}{1+x})}\mathbf{X},\quad 0<\nu<1,$$

where $x=\mathbf{X}^T\mathbf{S}^{-1}\mathbf{X}$, is minimax if

$$\nu\geq\frac{(n-p+2)n}{2(2\alpha_0 p+n-p-2)}-\frac{n-p}{2},\quad \alpha_0>\frac{1}{2}+\frac{p}{2},$$

with

$$\alpha_0=\frac{tr\Sigma^{1/2}\mathbf{Q}\Sigma^{1/2}}{p\max(\alpha_1,\ldots,\alpha_p)},$$

where $\alpha_1,\ldots,\alpha_p$ denote the characteristic roots of the matrix $\Sigma^{1/2}\mathbf{Q}\Sigma^{1/2}$.

Berger & Haff (1983) for the loss function

$$L(\hat{\mu},\mu)=(\hat{\mu}-\mu)^T\mathbf{Q}(\hat{\mu}-\mu)/tr(\mathbf{QS})$$

have considered a class of minimax estimators of μ

$$\hat{\mu}=\left(\mathbf{I}-c\alpha(\mathbf{Q}^{1/2}\mathbf{S}\mathbf{Q}^{1/2})h(\mathbf{X}^T\mathbf{S}^{-1}\mathbf{X})\mathbf{Q}^{-1}\mathbf{S}^{-1}\right)\mathbf{X},\tag{28}$$

where α and h are positive real valued functions and $c \geq 0$. Berger & Haff (1983) have given conditions on c, α and h under which estimators (28) are minimax and hence

$$R(\hat{\mu}, \mu) \leq R(\hat{\mu}^0, \mu).$$

Simple cases of the above problem under the same loss function have been considered by Menjoge & Rao (1985).

We see that the class of estimators better than $\hat{\mu}^0$ is very large, so it is difficult to choose the needed one. Berger (1980a) has shown that each estimator better than $\hat{\mu}^0$ is significantly better only for μ in a small region of the parameter space. Berger (1982b) proposed a rather simple minimax estimator which allows the user to select the region where significant improvement over $\hat{\mu}^0$ is achieved. He described the region of preferency by an ellipse

$$\{\mu : (\mu - \lambda)^T \mathbf{A}^{-1} (\mu - \lambda) \leq p\},$$

where λ is the center of the region and $\mathbf{A}$ determines axes and an orientation of the ellipse.

Let $\mathbf{X} = (X_1, \cdots, X_p)^T$ be distributed as $N(\mu, \Sigma)$, Σ being known. It is desired to estimate μ under the loss function

$$L(\hat{\mu}, \mu) = (\hat{\mu} - \mu)^T \mathbf{Q} (\hat{\mu} - \mu),$$

where $\mathbf{Q}$ is a known positive-definite matrix. Assume for simplisity that $\mathbf{Q}, \Sigma$ and $\mathbf{A}$ are diagonal with elements q_i, σ_i^2 and A_i respectively. The proposed minimax estimator is given, coordinate-wise, by

$$\hat{\mu}_i = X_i - \frac{\sigma_i^2}{\sigma_i^2 + A_i}(X_i - \lambda_i)\left[\frac{1}{q_i^*}\sum_{j=1}^{p}(q_j^* - q_{j+1}^*)\min_j\left\{1, \frac{2(j-2)_+}{||\mathbf{x}^j - \lambda^j||^2}\right\}\right],$$

where

$$q_i^* = q_i\sigma_i^4/(\sigma_i^2 + A_i), ||\mathbf{x}^j - \lambda^j||^2 = \sum_{i=1}^{j}(x_i - \lambda_i)^2/(\sigma_i^2 + A_i), q_{p+1}^* = 0$$

and X_i are indexed so that $q_1^* \geq q_2^* \geq \cdots \geq q_p^*$.

Naturally, the point λ may be thought to be a prior for μ and $\mathbf{A}$ to be a prior covariance matrix for Σ.

Problems of estimating matrices of normal mean where the Stein's method may be applied have been considered by Zheng (1986) and Konno (1991).

It is worth to note that the minimax property of Stein's rule is preserved not only with respect to the quadratic loss function but with respect to a generalised loss function

$$L(\hat{\mu},\mu) = |\hat{\mu}-\mu|^{2m}$$

too (Alam and Hawkes (1979)).

The Stein effect of improving upon usual estimators of mean of a multivariate normal distribution is also valid under the classical Pitman closeness criterion. Let $\mathbf{X}$ be $N(\mu,\sigma^2\mathbf{V})$ random vector, where $\mathbf{V}$ is a known positive definite matrix while μ and σ^2 are both unknown. For a given positive-definite matrix $\mathbf{Q}$ define the norm

$$||\mathbf{x}-\mathbf{y}||_{\mathbf{Q}} = (\mathbf{x}-\mathbf{y})^T\mathbf{Q}(\mathbf{x}-\mathbf{y}).$$

One may say that an estimator $\hat{\mu}_1$ is closer to μ than $\hat{\mu}_2$ is in the Pitman sense (in the norm $||.||_{\mathbf{Q}}$) if

$$P\{||\hat{\mu}_1-\mu||_{\mathbf{Q}} \le ||\hat{\mu}_2-\mu||_{\mathbf{Q}}\} \ge \frac{1}{2} \tag{29}$$

for all μ and σ^2.

Consider shrikage estimators of the form

$$\hat{\mu} = \mathbf{X} - \frac{\varphi(\mathbf{X},S_m)S_m\mathbf{Q}^{-1}\mathbf{V}^{-1}\mathbf{X}}{\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Q}^{-1}\mathbf{V}^{-1}\mathbf{X}}, \tag{30}$$

where φ is a nonegative function bounded from above by

$$(p-1)(3p+1)/(2p)$$

for every $(\mathbf{X},S_m)$ and a statistic S_m is independent of $\mathbf{X}$ and distributed like $m^{-1}\sigma^2\chi^2_m$. Sen, Kubokawa and Saleh (1989) have shown that the estimator (30) is closer to μ than $\mathbf{X}$ in the Pitman sense (29) (see also Keating and Mason (1988) and Strivastava (1993)). James and Stein (1992) have formuated the general problem of admissible estimation with quadratic loss and considered some problems of admissibility of Pitman's estimators. They have formulated also some unsolved problems.

The Stein approach is useful not only for estimating the unknown vector of means μ if $\mathbf{X} \sim N(\mu,\Sigma)$ but for estimating the unknown covariance matrix Σ as well. The usual best invariant unbiased and minimax estimator of Σ is $\hat{\Sigma}_0 = \mathbf{S}/(n-1)$. It is known that the sample eigenvalues of $\mathbf{S}$ tend to be more spread out than the population eigenvalues of Σ. This fact suggests to search estimators of Σ whose risks are smaller than that of $\hat{\Sigma}_0$. Dey and Srinivasan (1985, 1986) (see also reeferences there in) obtained some improved estimators of Σ under the Stein loss function

$$L(\hat{\Sigma},\Sigma) = tr(\hat{\Sigma}\Sigma^{-1}) - \log\det(\hat{\Sigma}\Sigma^{-1}) - p. \tag{31}$$

They considered the class of orthogonally invariant estimators

$$\hat{\Sigma} = R\varphi(\mathbf{L})R^T, \tag{32}$$

where $\mathbf{S} = R\mathbf{L}R^T$ with R the matrix of normalized eigenvectors ($RR^T = R^TR = \mathbf{I}$), $\mathbf{L} =$ diag $(l_1, l_2, \cdots, l_p)$ is the diagonal matrix of corresponding eigenvalues with

$$l_1 \geq l_2 \geq \cdots \geq l_p \text{ and } \varphi(\mathbf{L}) = \text{ diag } (\varphi_1(\mathbf{L}), \varphi_2(\mathbf{L}), \cdots, \varphi_p(\mathbf{L})).$$

For an estimator $\hat{\Sigma}$ of the form (32) the loss function (31) reduces to

$$L(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}\Sigma^{-1}) - \sum_{i=1}^{p} \log \varphi_i(\mathbf{L}) + \log\det(\Sigma) - p. \tag{33}$$

Since the last two terms in (33) are constant it is sufficient to consider the risk as follows

$$R^*(\hat{\Sigma}, \Sigma) = \mathbf{E}_\Sigma \left\{ tr(\hat{\Sigma}\Sigma^{-1}) - \sum_{i=1}^{p} \log \varphi_i(\mathbf{L}) \right\}. \tag{34}$$

Having obtained the unbiased estimator of $R^*(\hat{\Sigma}, \Sigma)$ and its upper bound, Dey and Srinivasan (1985) have shown that the estimator (32) for $p \geq 3$ and

$$\varphi_i(\mathbf{L}) = \frac{l_i}{n} - \frac{(l_i log l_i)\tau(u)}{b+u}, \qquad i = 1, 2, \ldots, p,$$

where

$$u = \sum_{i=1}^{p} \log^2 l_i, \qquad b > 144(p-2)^2/25k^2,$$

and $\tau(u)$ is a function satisfying:

(i) $0 < \tau(u) < 2(p-2)/k^*$, $k^* = 5k^2/6$;

τ is monotone nondecreasing in u and $\mathbf{E}[\tau'(u)] < \infty$, dominates $\hat{\Sigma}_0$ for the loss function (33). Dey and Srinivasan (1985) gave another estimators of Σ whose risks are smaller than that of the considered one. For orthogonally invariant estimators (32) the order relation of components of $\mathbf{L}$ and $\varphi(\mathbf{L})$ is important. If their components follow the same order relation the estimator (32) will dominate an estimator which does not preserve this ordering (Sheena and Takemura (1992)). Perron (1992) proposed a minimax orthogonally equivariant estimator $\hat{\Sigma}$ of Σ which satisfies the ordering and the shrinkage properties. The two sample analogue of the above problem has been considered by Loh (1991b) (see Remark 1).

Kubokawa (1989) gave an improved estimator of Σ under the quadratic loss function

(35) $$\mathbf{L}(\hat{\Sigma},\Sigma) = tr(\hat{\Sigma}\Sigma - \mathbf{I})^2.$$

He has shown that

(36) $$\hat{\Sigma} = \begin{cases} b(\mathbf{S}+\mathbf{X}\mathbf{X}^T) & \text{if } \mathbf{X}^T\mathbf{S}^{-1}\mathbf{X} \le (a_0-b)/b, \\ a_0\mathbf{S} & \text{otherwise}, \end{cases}$$

where

$$(n+2)a_0/(n+3) \le b < a_0, \qquad a_0 = 1/(n+p+1),$$

dominates $\hat{\Sigma}_0$ under the loss function (35). Dey, Ghosh and Srinivasan (1990) have considered the problem under a loss introduced by Efron and Morris (1976b).

In the multiple linear regression model containing more than two explanatory variables with normally distributed disturbances, the least squares estimator for the coefficient vector is inadmissible under a quadratic loss function.

Srivastava & Srivastava (1993) using the Stein-rule proposed the estimator which dominates the least squares estimator under a general convex loss function.

The Stein's effect is useful not only when estimating parameters of a multivariate normal distribution but for estimating parameters of some univariate and many other multivariate distributions including their mixtures too.

Bravo and MacGibbon (1988a) proposed James-Stein estimators for the parameters of an inverse Gaussian distribution. Bravo and MacGibbon (1988b) and Srivastava and Bilodeau (1989) constructed minimax Stein's like estimators for the mean vector μ under loss functions of the type (1) if an observation $\mathbf{X}$ is distributed as a scale mixture of normal distributions. The spherically symmetric case has been considered by Brandwein and Strawderman (1990), Ralescu, Brandwein and Strawderman (1992) and Cellier and Fourdrinier (1992).

Consider the problem of estimation of a common location of several exponential distributions with unknown scale parameters. Let $\{X_{ij}\}$, $j = 1,2,\ldots,n_i$, $i = 1,2,\ldots,p$, be a sample where

(37) $$X_{ij} \sim \frac{1}{\sigma_i} exp\left[-\frac{x_{ij}-\mu}{\sigma_i}\right], \qquad x_{ij} \ge \mu, \qquad \forall i,j.$$

The problem of estimating the common location μ of the model (37) arises, e.g., in life testing, survival analysis, etc. It is known that the set of complete sufficient statistics for $\mu,\sigma_1,\sigma_2,\ldots,\sigma_p$ is $Z,T_1,T_2,\ldots,T_p$, where

$$Z = \min_{\substack{1 \le i \le p \\ 1 \le j \le n_i}} X_{ij} = \min_{1 \le i \le p} X_{i(1)},$$

$$T_i = \sum_{j=1}^{n_i} (X_{ij} - Z), \quad i = 1, 2, \ldots, p.$$

The **MLE** $\hat{\mu}$ and the **MVUE** $\tilde{\mu}$ are given by (Ghosh and Razmpour (1984), see also, Bordes, Nikulin, Voinov (1994))

$$\hat{\mu} = Z, \quad \tilde{\mu} = Z - \left(\sum_{i=1}^{p} \frac{n_i(n_i - 1)}{T_i} \right)^{-1}. \tag{38}$$

Introducing instead of T_i the statistics

$$T_i^* = \sum_{j=1}^{n_i} (X_{ij} - X_{i(1)}), \quad i = 1, 2, \ldots, p,$$

assuming that the sample sizes $n_1, \ldots, n_p$ are all equal $n > 2$ and denoting

$$\hat{a}_c^* = nc \sum_{i=1}^{p} (1/T_i^*)$$

Pal and Sinha (1990) proved that the estimator

$$\hat{\mu}_c^* = Z - (\hat{a}_c^*)^{-1} \tag{39}$$

dominates the **MLE** $\hat{\mu} = Z$ for every $c \ge n/2$ under the squared loss

$$L(\hat{\mu}_c^*, \mu) = (\hat{\mu}_c^* - \mu)^2.$$

It dominates $\hat{\mu}$ whenever

$$p \ge 4 \text{ and } n > 5, \text{ or } p = 3 \text{ and } n > 4, \text{ or } p = 2 \text{ and } n > 3.$$

The same is true for the Pitman's criterion of nearness. All the same hold in the case of unequal sample sizes (Pal and Sinha (1990)).

Berger (1980b) developed a technique for improving upon inadmissible estimators of parameters for a class of continuous exponential distributions.

Let $\mathbf{X} = \{X_1, \ldots, X_p\}$ be a sample, where $\mathbf{X} \sim f(\mathbf{x};\theta)$ and the loss function in estimating $\mathbf{g}(\theta)$ by $\delta(\mathbf{X})$ is $L(\delta(\mathbf{X}), \mathbf{g}(\theta))$. Following Stein Berger used for the risk

$$R(\delta(\mathbf{X}), \mathbf{g}(\theta)) = \mathbf{E}_\theta L(\delta(\mathbf{X}), \mathbf{g}(\theta))$$

the representation

$$R(\delta, \theta) = \int L(\delta(\mathbf{x}), \mathbf{g}(\theta)) f(\mathbf{x};\theta)\mathbf{dx} = \int \mathcal{D}(\delta(\mathbf{x})) f(\mathbf{x};\theta)\mathbf{dx}, \tag{40}$$

where a function $\mathcal{D}(\delta(\mathbf{x}))$ involves $\delta(\mathbf{x})$ and its derivatives but not θ. Comparing an estimator $\delta^*(\mathbf{X})$ with the usual one $\delta^0(\mathbf{X})$ one should solve the inequality

$$R(\delta^*, \theta) - R(\delta, \theta) = \mathbf{E}_\theta[\mathcal{D}(\delta^*(\mathbf{X})) - \mathcal{D}(\delta^0(\mathbf{X}))] < 0$$

for all θ. The problem thus reduces to solving the differential inequality

$$\mathcal{D}(\delta^*(\mathbf{x})) - \mathcal{D}(\delta^0(\mathbf{x})) < 0. \tag{41}$$

Let X_i, $i = 1, 2, \ldots, p$, be independent random Gamma variables with probability density function

$$f(x_i;\theta_i) = \theta_i^{\alpha_i} x_i^{(\alpha_i - 1)} e^{-x_i\theta_i}/\Gamma(\alpha_i), \qquad \alpha_i > 0, 0 < \theta_i < \infty, \tag{42}$$

and we want to estimate the vector

$$(\theta_1^{-1}, \theta_2^{-1}, \ldots, \theta_p^{-1})^T \text{ by } \delta(\mathbf{X}) = (\delta_1(\mathbf{X}), \delta_2(\mathbf{X}), \ldots, \delta_p(\mathbf{X}))^T$$

with the loss function

$$L(\delta, \theta) = \sum_{i=1}^{p} \theta_i^m (1 - \delta_i(\mathbf{X})\theta_i)^2, \qquad m \in Z, \quad \delta = (\theta_1, \ldots, \theta_p)^T. \tag{43}$$

Solving inequality (41) for $m = -2, -1, 0$ and 1, Berger (1980b) obtained estimators $\delta^*(\mathbf{X})$, which for $p \geq 2$ or 3 under the loss function (42) are uniformly better than the standard estimator

$$\delta_i^0(\mathbf{X}) = \frac{X_i}{\alpha_i + 1}$$

of θ_i^{-1} for the loss

$$\theta_i^m (1 - \delta_i\theta_i)^2, \qquad i = 1, 2, \ldots, p.$$

In the case $m = -2$ e.g., which corresponds to the usual sum of squares error loss the estimator $\delta^*(\mathbf{X})$ is given component-wise by

$$\delta_i^*(\mathbf{X}) = \frac{X_i}{\alpha_i + 1}\left(1 + \frac{c(\alpha_i - 1)(\alpha_i + 1)^2/X_i^2}{[b + \sum_{j=1}^{p}(\alpha_j + 1)^4/X_j^2]} + \frac{2c(\alpha_i + 1)^6/X_i^4}{[b + \sum_{j=1}^{p}(\alpha_j + 1)^4/X_j^2]^2}\right), \tag{44}$$

$0 < c < 4(p-1)$, $p \geq 2$, $b \geq (c^2/4)(1 + 1/\min\alpha_i)$.

The Berger's estimators $\delta^*(\mathbf{X})$ for different values of m are completely different in their functional form. Das Gupta (1986) and Bilodeau (1988) obtained a new class of estimators for $(\theta_1^{-1}, \theta_2^{-1}, \ldots, \theta_p^{-1})^T$ under the more general weighted loss function

$$L(\delta, \theta) = \sum_{i=1}^{p} w_i \theta_i^{m_i} (1 - \delta_i(\mathbf{x})\theta_i)^2, \qquad (w_i > 0, m_i \neq 0).$$

All these estimators possess the same functional form. Bilodeau (1988) showed that for $p \geq 2$ the estimator

$$\delta_i(\mathbf{X}) = \frac{X_i}{\alpha_i + 1}(1 + \phi_i(\mathbf{X})), \tag{45}$$

where

$$\phi_i(\mathbf{X}) = -k(\text{sign}\Delta_i) X_i^{m_i/2} \prod_{j=1}^{p} X_j^{-m_j/2p}$$

and

$$\Delta_i = \frac{\Gamma(\alpha_i + 2 + \frac{m_i}{2} - \frac{m_i}{2p})}{(\alpha_i + 1)\Gamma(\alpha_i)} - \frac{\Gamma(\alpha_i + 1 + \frac{m_i}{2} - \frac{m_i}{2p})}{\Gamma(\alpha_i)}$$

uniformly dominates

$$\delta_i^0(\mathbf{X}) = \frac{X_i}{\alpha_i + 1} \text{ for } 0 < k < 2MDp/KB,$$

where

$$M = \prod_{i=1}^{p} \frac{\Gamma(\alpha_i - \frac{m_i}{2p})}{\Gamma(\alpha_i)}, \qquad D = \min_i \frac{w_i \mid \Delta_i \mid \Gamma(\alpha_i)}{(\alpha_i + 1)\Gamma(\alpha_i - \frac{m_i}{2p})},$$

$$K = \sum_{i=1}^{p} \frac{w_i \Gamma(\alpha_i + 2 + m_i - \frac{m_i}{p})}{(\alpha_i + 1)^2 \Gamma(\alpha_i - \frac{m_i}{p})} \text{ and } B = \prod_{i=1}^{p} \frac{\Gamma(\alpha_i - \frac{m_i}{p})}{\Gamma(\alpha_i)}.$$

Dey, Ghosh and Srinivasan (1987) using the Berger (1980b) technique obtained for the probability model (42) different classes of shrinkage estimators dominating for $p \geq 3$ the **UMVUE**

$$\delta^0(\mathbf{X}) = (X_1/\alpha_1, X_2/\alpha_2, \ldots, X_p/\alpha_p)^T$$

which is usual for the loss function

$$L(\delta, \theta) = \sum_{i=1}^{p} (\delta_i(\mathbf{X})\theta_i - \log(\delta_i(\mathbf{X})\theta_i) - 1)$$

being taken by Dey *et al.*

A remarkable Stein's approach for estimators improvement proves to be useful for simultaneous estimation of location parameters of distributions with finite support too. Akai (1986) considered probability density functions which are positive on a finite interval, symmetric about θ_i and have the same variance. He obtained a class of shrinkage estimators of the location vector $\theta = (\theta_1, ..., \theta_p)^T$ under the squared error loss function. In particular, explicit dominating estimators where the distributions of X_i's are mixture of two uniform distributions were given.

A large amount of work is devoted to the application of Stein's ideas for shrinkage parameters estimation in multivariate discrete distributions. Ghosh, Hwang and Tsui (1983) constructed estimators which dominate the **UMVUE** under a weighted squared error loss function shrinking the **UMVUE** towards a prescribed nonzero point or a data-based point. Let, for example, X_i, $i = 1, 2, ..., p$, be independent negative binomial random variables and

$$P(X_i = x_i; \theta_i) = \begin{pmatrix} r_i + x_i - 1 \\ r_i - 1 \end{pmatrix} \theta_i^{x_i} (1 - \theta_i)^{r_i}, \quad x_i = 0, 1, ...$$

The **UMVUE** $\delta^0(\mathbf{X}) = (\delta_1^0(\mathbf{X}), ..., \delta_p^0(\mathbf{X}))^T$ of vector $\theta = (\theta_1, ..., \theta_p)^T$ is (see §A25 in Voinov and Nikulin (1993)) the vector with components

$$\delta_i^0(\mathbf{X}) = X_i/(X_i + r_i - 1).$$

Denoting $N(\mathbf{X}) = \#\{i : X_i > \lambda_i\}$, where $\lambda = (\lambda_1, ..., \lambda_p)^T$ is above mentioned prescribed point, Ghosh, Hwang and Tsui (1983) showed that estimators

$$\delta_i(\mathbf{X}) = \delta_i^0(\mathbf{X}) - c[N(\mathbf{X}) - 2]_+ H_i(X_i)/D, \tag{46}$$

where

$$D = \sum_{j=1}^{p} d_j(X_j), \quad h_i(X_i) = \sum_{j=1}^{X_i} (r_i + j - 1)/j, \quad H_i(X_i) = h_i(X_i) - h_i(\lambda_i),$$

$$d_i(X_i) = \begin{cases} H_i^2(X_i) + b_i H_i(X_i) & \text{if} \quad X_i \geq \lambda_i, \\ H_i^2(X_i) + a_i & \text{if} \quad X_i < \lambda_i, \end{cases}$$

$$a_i = r_i\{3h_i(\lambda_i)/2 - 1\}_+ \text{ and } b_i = (r_i + \lambda_i + 1)/(\lambda_i + 2),$$

for $0 < c \leq 2$ and $p > 2$ dominate the **UMVUE** $\delta^0(\mathbf{X})$ under

$$L(\theta, \delta(\mathbf{X})) = \sum_{i=1}^{p} (\theta_i - \delta_i(\mathbf{X}))^2.$$

Many results conserning the improvement upon **MVUE**s for discrete probability distributions parameters and functions of them including densities with dependent marginals

may be found in Tsui (1986), Kant and Sharma (1986), Chou (1991) and Dey and Chung (1991).

In order to finish this review, we would like to add two more remarks concerning this topic.

Remark 2. The main feature of Stein's technique for estimators improvement consists of a representing a risk function as an expectation of a function which does not involve unknown parameters. There exists another approach to the problem with the same idea to work with expressions independent on parameters. If there exists a sufficient statistic for unknown parameters we may construct, if exists, the unbiased estimator for a risk function and then to use it for estimators improvement. This technique has been used among others by Haff (1980) and Loh (1991a). Loh (1991b) considered the following problem.

Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be two independent $p \times p$ Wishart matrices where $\mathbf{S}_1 \sim W_{n_1-1}(\mathbf{s}_1;\Sigma_1)$ and $\mathbf{S}_2 \sim W_{n_2-1}(\mathbf{s}_2;\Sigma_2)$. The problem is to estimate (Σ_1,Σ_2) under the loss function

$$L(\hat{\Sigma}_1,\hat{\Sigma}_2;\Sigma_1,\Sigma_2) = \sum_{i=1}^{2}\{tr(\Sigma_i^{-1}\hat{\Sigma}_i) - \log\det(\Sigma_i^{-1}\hat{\Sigma}_i) - p\}. \tag{47}$$

Loh (1991b) considered estimators invariant under the group of transformations

$$\Sigma_i \to \mathbf{A}\Sigma_i\mathbf{A}^T, \quad \mathbf{S}_i \to \mathbf{A}\mathbf{S}_i\mathbf{A}^T, \quad \forall \mathbf{A} \in GL(p,R), \quad i = 1,2, \tag{48}$$

where $GL(p,R)$ denotes the set of all $p \times p$ nonsingular matrices. He proved that the estimator $(\hat{\Sigma}_1,\hat{\Sigma}_2)$ is an equivariant under transformations (48) estimator of (Σ_1,Σ_2) iff

$$\begin{aligned} &\hat{\Sigma}_1(\mathbf{S}_1,\mathbf{S}_2) = \mathbf{B}^{-1}\Psi(\mathbf{I}-\mathbf{F})\mathbf{B}^{(T)^{-1}}, \\ &\hat{\Sigma}_2(\mathbf{S}_1,\mathbf{S}_2) = \mathbf{B}^{-1}\Phi(\mathbf{F})\mathbf{B}^{(T)^{-1}}, \\ &\mathbf{B}(\mathbf{S}_1+\mathbf{S}_2)\mathbf{B}^T = \mathbf{I}, \quad \mathbf{B}\mathbf{S}_2\mathbf{B}^T = \mathbf{F}, \end{aligned} \tag{49}$$

where Ψ and Φ are diagonal matrices, $\mathbf{F} = \mathrm{diag}(f_1,\ldots,f_p)$ with $f_1 \geq \ldots \geq f_p$. Denote

$$\nabla^{(i)} = (\nabla^{(i)}_{jk})_{p\times p}, \quad 1 \leq j,k \leq p, \quad i = 1,2,$$

where

$$\nabla^{(i)}_{jk} = \frac{1}{2}(1+\delta_{jk})\frac{\partial}{\partial S^{(i)}_{jk}}, \quad \delta_{jk} = \begin{cases} 1, & j = k \\ 0, & j \neq k, \end{cases} \qquad \mathbf{S}_i = (S^{(i)}_{jk}),$$

$S^{(i)}_{jk}$, $1 \leq j, k \leq p$, being elements of $\mathbf{S}_i$, $\quad i = 1,2$. If matrices

$$\Psi = \mathrm{diag}(\psi_1,\ldots,\psi_p) \text{ and } \Phi = \mathrm{diag}(\phi_1,\ldots,\phi_p)$$

satisfy the Wishart identity in the sense that

$$\text{(50)} \qquad \mathbf{E}\ \text{tr}\ (\Sigma_i^{-1}\hat{\Sigma}_i) = \mathbf{E}\ \text{tr}\ [2\nabla^{(i)}(\hat{\Sigma}_i) + (n_i - p - 1)\mathbf{S}_i^{-1}\hat{\Sigma}_i], \qquad i = 1,2,$$

then the risk of $(\hat{\Sigma}_1, \hat{\Sigma}_2)$ with respect to the loss (47) is given by

$$R(\hat{\Sigma}_1, \hat{\Sigma}_2; \Sigma_1, \Sigma_2) = \mathbf{E}\{\sum_{i=1}^{p}[\frac{n_1 - p - 1}{1 - f_i}\psi_i - 2\psi_i \sum_{i \neq j} \frac{f_j}{f_i - f_j} + 2\psi_i + 2f_i \frac{\partial \psi_i}{\partial(1 - f_i)} -$$

$$\log \frac{\psi_i}{1 - f_i} + \frac{n_2 - p - 1}{f_i}\phi_i + 2\phi_i \sum_{i \neq j} \frac{1 - f_j}{f_i - f_j} + 2\phi_i + 2(1 - f_i)\frac{\partial \phi_i}{\partial f_i} - \log \frac{\phi_i}{f_i} -$$

$$\text{(51)} \qquad \log \chi^2_{n_1 - i + 1} - \log \chi^2_{n_2 - i + 1} - 2]\} = \mathbf{E}\hat{R}(\hat{\Sigma}_1, \hat{\Sigma}_2; \Sigma_1, \Sigma_2).$$

The estimator $\hat{R}(\hat{\Sigma}_1, \hat{\Sigma}_2; \Sigma_1, \Sigma_2)$ being dependent on complete sufficient for parameters Σ_1, Σ_2 statistic $(\mathbf{S}_1, \mathbf{S}_2)$ (see (49)) is the **UMVUE** of the risk $R(\hat{\Sigma}_1, \hat{\Sigma}_2; \Sigma_1, \Sigma_2)$.

To construct the Stein-type estimator for (Σ_1, Σ_2) one should now to minimize the **UMVUE** $\hat{R}(\hat{\Sigma}_1, \hat{\Sigma}_2; \Sigma_1, \Sigma_2)$ to obtain elements $\psi_1^*, ..., \psi_p^*$ and $\phi_1^*, ..., \phi_p^*$ of matrices Ψ and Φ respectively which give minimum to $\hat{R}(.)$. By ignoring the derivative terms in $\hat{R}(.)$ and solving equations

$$\frac{\partial \hat{R}(.)}{\partial \psi_i} = 0, \qquad \frac{\partial \hat{R}(.)}{\partial \phi_i} = 0, \qquad \forall i,$$

Loh (1991) obtained approximate values of $\psi_i^*, \phi_i^*, \quad i = 1,2,...p$, as follows

$$\psi_i^* = (1 - f_i) / \left[n_1 - p + 1 - 2 \sum_{i \neq j} \frac{f_j(1 - f_i)}{f_i - f_j} \right],$$

$$\phi_i^* = f_i / \left[n_2 - p - 1 + 2 \sum_{i \neq j} \frac{f_i(1 - f_j)}{f_i - f_j} \right], \qquad i = 1, ..., p,$$

where $0 \leq \psi_1 \leq ... \leq \psi_p$, $\phi_1 \geq ... \geq \phi_p \geq 0$. Substituting ψ_i^* and ϕ_i^*, $i = 1, ..., p$, into (49) we obtain a Stein-type estimator $(\hat{\Sigma}_1(\mathbf{S}_1, \mathbf{S}_2), \hat{\Sigma}_2(\mathbf{S}_1, \mathbf{S}_2))$ for (Σ_1, Σ_2). Loh (1991a) have also shown that natural ordering of $\phi_1, ..., \phi_p$ may be altered by Stein's isotonic regression technique.

Remark 3. The uniformly minimum variance unbiased estimation technique is useful not only for a total risk estimating to improve estimators, but for estimating the mean-squared-error matrix as well, since the last estimator is helpful in identifying the components of $\hat{\mu}$ which contribute most to the total risk.

Let $\mathbf{X} \sim N(\mu, \Sigma)$, μ and Σ being unknown. Let $\mathbf{S}$ be an estimator of Σ independent of $\mathbf{X}$ and $\mathbf{S} \sim W_{n-1}(\mathbf{s}; \Sigma)$. Considering the class of estimators

$$\hat{\mu}_s = \left(1 - \frac{k}{\mathbf{X}^T \mathbf{S}^{-1} \mathbf{X}}\right) \mathbf{X}, \quad 0 < k < \frac{2(p-2)}{n-p+3},$$

Bilodeau and Srivastava (1988) have shown that the **MVUE** of the mean-squared matrix

$$M(\hat{\mu}_s) = \mathbf{E}(\hat{\mu}_s - \mu)(\hat{\mu}_s - \mu)^T$$

is given by

$$\hat{M}(\hat{\mu}_s) = \left(1 - \frac{2k}{\mathbf{X}^T \mathbf{S}^{-1} \mathbf{X}}\right) \frac{\mathbf{S}}{n} + k \left(k + \frac{4(n+1)}{n(n-p+3)}\right) \frac{\mathbf{X}\mathbf{X}^T}{(\mathbf{X}^T \mathbf{S}^{-1} \mathbf{X})^2}.$$

It is this estimator which allows to determine "risky" components of $\hat{\mu}_s$.

ACKNOWLEDGMENT

The authors wish to thank Professors R.Beran and C.Cuadras for their advice and encouragement and referees for their careful reading of the work.

REFERENCES

[1] **Akai, T.** (1988). "Further generalization of Stein estimators". *Keio Science and Technology reports*, **40**, 4, 41–54.

[2] **Alam, K.** (1977). "Minimax estimators of a multivariate normal mean". *Scand. J. Statist.*, **4**, 125–130.

[3] **Alam, K.** and **Mitra A.** (1986). "Component risk in multiparameter estimation". *Ann. Inst. Statist. Math.*, **38**, Part A, 399–410.

[4] **Alam, K.** and **Hawkes, J.S.** (1979). "Minimax property of Stein's estimator". *Commun. Statist.-Theor. Meth.*, **A8(6)**, 581–590.

[5] **Baranchik, A.J.** (1964). "Multiple regression and estimation of the mean of a multivariate normal distribution". *Stanford Univ. Technical Report #51.*

[6] **Baranchik, A.J.** (1970). "A family of minimax estimators of the mean of a multivariate normal distribution". *Ann. Math. Statist.*, **41**, 642–645.

[7] **Beran, R.** (1992). "Stein estimation in high dimensions and the bootstrap". *Preprint.* pp. 1–21.

[8] **Berger, J.** (1976). "Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss". *Ann. Statist.*, **4**, 223–226.

[9] **Berger, J.O.** and **Srinivasan, C.** (1977). "Generalized Bayes estimators in multivariate problems". *Purdue Univ., Depart. of Statist., Mimeograph Series #481.*

[10] **Berger, J.** (1980a). "A robust generalized Bayes estimator and confidence region for a multivariate normal mean". *Ann. Statist.*, **8**, 716–761.

[11] **Berger, J.** (1980b). "Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters". *Ann. Statist.*, **8**, 545–571.

[12] **Berger, J.** (1982a). "Bayesian robustness and the Stein effect". *JASA*, **77**, 358–368.

[13] **Berger, J.** (1982b). "Selecting a minimax estimator of a multivariate normal mean". *Ann. Statist.*, **10**, 81–92.

[14] **Berger, J.** and **Haff. L.R.** (1983). "A class of minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix". *Stat. & Decisions.*, **1**, 105–129.

[15] **Berry, J.C.** (1994). "Improving the James-Stein estimator using the Stein variance estimator". *Statist. & Prob. Letters.*, **20**, p. 241–245.

[16] **Bilodeau, M.** and **Srivastava, M.S.** (1982). "Estimation of the MSE matrix of the Stein estimator". *Can. J. Statist.*, **16**, 153–159.

[17] **Bilodeau, M.** (1988). "On the simultaneous estimation of scale parameters". *Can. J. Statist.*, **16**, 169–174.

[18] **Bock, M.E.** (1975). "Minimax estimators of the mean of a multivariate normal distribution". *Annals of Statist.*, **3**, 209–218.

[19] **Bordes, L., Nikulin, M.** and **Voinov, V.** (1994). "Unbiased estimators' tables of parameter functions for a several exponential with a common shift". *Preprint 94-07 de Bordeaux 2*, 22p.

[20] **Bravo, G.** and **Mac Gibbon, B.** (1988a). "Improved shrinkage estimators for the mean vector of a scale mixture of normals with unknown variance". *Can. J. Statist.*, **16**, 237–245.

[21] **Bravo, G.** and **Mac Gibbon, B.** (1988b). "Improved estimation for the parameters of an inverse Gaussian distribution". *Commun. Statist.- Theory Meth.*, **17**, 4285–4299.

[22] **Brandwein, A.C.** and **Strawderman, W.E.** (1990). "Stein estimation: The spherically symmetric case". *Statistical Science*, **5**, 356–369.

[23] **Cellier, D.** and **Fourdrinier, D.** (1992). "Estimation du paramètre de position d'une loi à symétrie sphérique: une condition générale de domination de l'estimateur des moindres carrés". *C.R. Acad. Sci. Paris*, **315**, 1203–1206.

[24] **Chen, S.Y.** (1988). "Restricted risk Bayes estimation for the mean of the multivariate normal distribution". *J. Multivariate Analysis.* **24**, 207–217.

[25] **Chou, J.P.** (1991). "Simultaneous estimation in discrete multivariate exponential families". *Ann. Statist.*, **19**, 314–328.

[26] **Das Gupta, A.** (1986). "Simultaneous estimation in the multiparameter gamma distribution under weighted quadratic losses". *Ann. Statist.*, **14**, 206–219.

[27] **Dey, D.K.** and **Berger, J.O.** (1983). "On truncation of shrinkage estimators in simultaneous estimation of normal means". *JASA*, **78**, #384, 865–869.

[28] **Dey, D.K.** and **Srinivasan, C.** (1985). "Estimation of a covariance matrix under Stein's loss". *Annals of Statist.*, **13**, 1581–1591.

[29] **Dey, D.K.** and **Srinivasan, C.** (1986). "Trimmed minimax estimator of a covariance matrix". *Ann. Inst. Statist. Math.*, **38**, part A, 101–108.

[30] **Dey, D.K., Ghosh, M.** and **Srinivasan, C.** (1987). "Simultaneous estimation of parameters under entropy loss". *J. Statist. Plann. Imfer.*, **15**, 347–363.

[31] **Dey, D.K., Ghosh, M.** and **Srinivasan, C.** (1990). "A new class of improved estimators of a multinormal precision matrix". *Statistics & Decisions*, **8**, 141–151.

[32] **Dey, D.K.** and **Chung, Y.** (1991). "Multiparameter estimation in truncated power series distributions under the Stein's loss". *Commun. Statist. -Theory Meth.*, **20**, 309–326.

[33] **Efron, B.** and **Morris, C.** (1973). "Stein's estimation rule and its competitors. An empirical Bayes approach". *JASA*, **68**, 117–130.

[34] **Efron, B.** (1975). "Biased versus unbiased estimation". *Advances in Math.*, **16**, 259–277.

[35] **Efron, B.** and **Morris, C.** (1976a). "Families of minimax estimators of the mean of a multivariate normal distribution". *Ann. of Statist.*, **4**, 11–21.

[36] **Efron, B.** and **Morris, C.** (1976b). "Multivariate empirical Bayes and estimation of covariance matrices". *Ann. Statist.*, **4**, 22–32.

[37] **Egerton, M.** and **Laycock, P.** (1982). "An explicit formula for the risk of James-Stein estimators". *Canad. J. Statist.*, **10(3)**, 199–205.

[38] **Ghosh, M.** and **Razmpour, A.** (1984). "Estimation of the common location parameter of several exponentials". *Sankhyá*, **48A**, 383–394.

[39] **Ghosh, M., Hwang, J.T.** and **Tsui, K.W.** (1983). "Construction of improved estimators in multiparameter estimation for discrete exponential families". *Ann. Statist.*, **11**, 351–367.

[40] **George, E.** (1986). "Minimax multiple shrinkage estimators". *Ann. Statist.*, **14**, 188–205.

[41] **Haff, L.R.** (1980). "Empirical Bayes estimation of the multivariate normal covariance matrix". *Ann. Statist.*, **8**, 586–597.

[42] **Hoffmann, K.** (1992). "Improved estimation of distribution parameters: Stein-type estimators". *B.G. Teubner Verlagsgesellschaft mbH, Stuttgart.*

[43] **Hudson, H.M.** (1974). "Empirical Bayes estimation". *Technical Report #58, Departament of Statistics, Stanford University.*

[44] **James, W.** and **Stein, C.** (1960). "Estimation with quadratic loss". *Proc. 4rd Berkeley Symp. Math. Statist. Prob.* , **Vol 1**, *University of California Press, Berkeley, Calif.*, 361–379.

[45] **James, W.** and **Stein, C.** (1992). "Estimation with quadratic loss". *(in "Breakthroughs in Statistics.* **Vol.1***", eds. S.Kotz & N.L.Johnson, Springer)*, 443–460.

[46] **Kant, V.** and **Sharma, D.** (1986). "Simultaneous estimation of λ_i^s for Poisson distributions". *Statistics*, **17**, 539–547.

[47] **Keating, J.P.** and **Mason, R.L.** (1988). "James-Stein estimation from an alternative perspective". *Amer. Statist.*, **42**, 160–164.

[48] **Konno, Y.** (1991). "On estimation of a matrix of normal means with unknown covariance matrix". *J. Multivariate Analysis*, **36**, 44–55.

[49] **Kubokawa, T.** (1989). "Improved estimation of a covariance matrix under quadratic loss". *Statist. & Prob. Letters*, **8**, 69–71.

[50] **Lehmann, E.L.** (1983). *Theory of Point Estimation*. Wiley, New York.

[51] **Loh, W.** (1991a). "Estimating covariance matrices". *Ann. Statist.*, **19**, 283–296.

[52] **Loh, W.** (1991b). "Estimating the common mean of two multivariate normal distributions". *Ann. Statist.*, **19**, 297–313.

[53] **Menjoge, S.S.** and **Rao, P.S.R.S.** (1985). "Simultaneous estimation of parameters using compound weighted loss". *Sankhyá*, **47B**, 338–345.

[54] **Nickerson, D.M.** (1989). "Dominance of the positive-part version of the James-Stein estimator". *Statist. & Prob. Letters*, **7**, 97–103.

[55] **Pal, N.** and **Sinha, B.K.** (1990). "Estimation of a common location of several exponentials". *Statist. & Decisions*, **8**, 27–36.

[56] **Perron, F.** (1992). "Minimax estimators of a covariance matrix". *J. Multivar. Anal.*, **43**, 16–28.

[57] **Ralescu, S., Brandwein, A.C.** and **Strawderman, W.E.** (1992). "Stein estimation for non-normal spherically symmetric location families in three dimensions". *J. Multivar. Analys.*, **42**, 35–50.

[58] **Robert, C.** (1988). "An explicit formula for the risk of the positive-part James-Stein estimator". *Canad. J. of Statist.*, **16**, 161–168.

[59] **Sen, P.K., Kubokawa, T.** and **Saleh, A.K.Md.E.** (1989). "The Stein paradox in the sense of the Pitman measure of closeness". *Ann. Statist.*, **17**, 1375–1386.

[60] **Sheena, Y.** and **Takemura, A.** (1992). "Inadmissibility of non-order-preserving orthogonally invariant estimators of the covariance matrix in the case of Stein's loss". *J. Multivariate Anal.*, **41**, 117–131.

[61] **Srivastava, M.S.** and **Bilodeau, M.** (1989). "Stein estimation under elliptical distributions". *J. of Multivariate Analysis*, **28**, 247–259.

[62] **Srivastava, A.K.** and **Srivastava, V.K.** (1993). "Pitman closeness for Stein-rule estimators of regression coefficients". *Statist. & Prob. Letters*, **18**, 85–89.

[63] **Stein, C.** (1956). "Inadmissibility of the usual estimator for a mean of multivariate normal distribution". *Proc. 3rd Berkeley Symp. Math. Statist. Prob.,* **Vol 1**, *University of California Press, Berkeley, Calif.*, 197–206.

[64] **Stein, C.** (1977). "Lectures on the theory of estimation of many parameters (in Russian)". *Studies in the Statistical Theory of Estimation, Part I* (I.A. Ibragimov and M.S.Nikulin. Eds.) Proceedings of Scientific Seminars of the Steklov Institute, Leningrad Division, **7**, pp.4–65.

[65] **Stein, C.** (1981). "Estimation of the mean of a multivariate normal distribution". *The Annals of Statistics*, **9**, 6, 1135–1151.

[66] **Stein, C.** (1986). "Lectures on the theory of estimation of many parameters". *J.Sov. Math.* **34**, 1, 1373–1403.

[67] **Stigler, S.M.** (1990). "The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators". *Statistical Science* **5**, 1, 147–155.

[68] **Tsui, K.W.** (1986). "Multiparameter estimation for some multivariate discrete distributions with possibly dependent components". *Ann. Inst. Statist. Math.*, **38A**, 45–56.

[69] **Voinov, V.G.** and **Nikulin, M.S.** (1993). "Unbiased estimators and their applications". **V.1**, Univariate case, *Kluwer.*

[70] **Zheng, Z.** (1986). "On estimation of matrix of normal mean". *J. Multivariate Analysis*, **18**, 70–82.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 93-106, 1995

ON THE PROBLEM OF THE MEANS OF WEIGHTED NORMAL POPULATIONS

M. S. NIKULIN * and V. G. VOINOV‡

An analitical problem, which arises in the statistical problem of comparing the means of two normal distributions, the variances of which as their ratio are unknown, is well-known in the mathematical statistics as the Behrens-Fisher problem. One generalization of the Behrens-Fisher problem and different aspects concerning the estimation of the common mean of several independent normal distributions with different variances are considered and one solution is proposed.

Key words: Behrens-Fisher problem; Minimum Variance Unbiased Estimator; Point and Interval estimators; efficient estimator, test.

AMS Subject Classification: 62FXX.

1. THE BEHRENS-FISHER PROBLEM

Let X_{ij} $(i = 1,\ldots,k;\ j = 1,\ldots,n_i)$ be mutually independent normally distributed random variables

$$\mathrm{E}X_{ij} = a_i \quad \text{and} \quad \mathrm{Var}\,X_{ij} = \sigma_i^2$$

In this model we have a minimal sufficient statistic

$$\mathbf{T}_{2k} = (X_1,\ldots,X_k,S_1^2,\ldots,S_k^2)^{\mathrm{T}},$$

*M. S. Nikulin. Laboratory of Statistical Methods. Steklov Mathematical Institute. St.Petersburg, Russia. UFR MI25. Université Bordeaux 2, France.

‡V. G. Voinov. Laboratory of Statistics. The Institute of Theoretical and Applied Mathematics. Alma-Ata, 480082. Kazakhstan.

-Article rebut el setembre de 1992.

-Acceptat el maig de 1995.

the components of which

$$X_i = \frac{1}{n}\sum_{j=1}^{n_i} X_{ij} \quad \text{and} \quad S_i^2 = \sum_{j=1}^{n_i} (X_{ij} - X_i)^2, \quad (i = 1, \ldots, k)$$

are mutually independent statistics such that,

$$X_i \quad \text{being normal} \quad \mathrm{N}\left(a_i, \frac{\sigma_i^2}{n_i}\right) \quad \text{and} \quad \frac{S_i^2}{\sigma_i^2} = \chi_{f_i}^2, \quad f_i = n_i - 1.$$

Let $s_i^2 = \frac{1}{f_i} S_i^2$, $i = 1, \ldots, k$. Since

$$\mathrm{E}\{X_i\} = a_i \quad \text{and} \quad \mathrm{E}\{s_i^2\} = \sigma_i^2 \quad (i = 1, \ldots, k)$$

and since the sufficient statistic $\mathbf{T}_{2k}$ is complete, we can affirm that the statistic $\mathbf{U} = (X_1, \ldots, X_k, s_1^2, \ldots, s_k^2)$ is the best (minimum variance) unbiased estimator (MVUE) for the parameter $(a_1, \ldots, a_k, \sigma_1^2, \ldots, \sigma_k^2)^{\mathrm{T}}$.

The well-known Behrens-Fisher problem consists in testing the hypothesis

$$H_0 : a_1 = a_2 = \cdots = a_k = a$$

or in constructing a confidence interval for the unknown common mean a under the assumption that the solution should be written in term of the minimal sufficient statistic $\mathbf{T}_{2k}$. We note that even for $k = 2$ the exact solution of this problem has not been obtained as yet.

Remark 1. The statistical problem of comparing the mathematical exectations a_1 and a_2 of two normal distributions, the variances of which σ_1^2, σ_2^2 and their ratio σ_1^2/σ_2^2 are unknown, was posed by Behrens (1929). The modern formulation of this problem is due to Sir R. Fisher and is based on the concept of sufficiency, since a sufficient statistic contains the same information about the unknown parameters $a_1, \sigma_1^2, a_2, \sigma_2^2$ as the initial data $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$,

$$\mathrm{E}X_{ij} = a_i \quad \text{and} \quad \mathrm{Var}X_{ij} = \sigma_i^2 \quad (i = 1,2; \quad j = 1, \ldots, n_i;\ n_i \geq 2),$$

and hence only the sufficient statistic $\mathbf{T}_4 = (X_1, X_2, S_1^2, S_2^2)^{\mathrm{T}}$ needs be considered in testing hypotheses about the values of unknown parameters $a_1, \sigma_1^2, a_2, \sigma_2^2$. In particular, this approach was used by Linnik, Sudakov and Romanovsky (1964) to test the hypothesis

$$H_\delta : a_1 - a_2 = \delta,$$

where δ is a given number. In this situation the Behrens-Fisher problem reduces to construct a critical set $K_\alpha \subset \mathfrak{X}_3 \subset \mathbb{R}^3$ in the sample space $\mathfrak{X}_3$ of the three-dimentional (under the hypothesis H_δ) sufficient statistic

$$\mathbf{V}_3 = (X_1 - X_2, S_1^2, S_2^2)^{\mathrm{T}}$$

such that the probability $\mathcal{P}\{\mathbf{V}_3 \in K_\alpha | H_\delta\}$ does not depend on the unknown parameters a_i, σ_i^2 and the unknown ratio σ_1^2/σ_2^2 and

$$\mathcal{P}\{\mathbf{V}_3 \in K_\alpha | H_\delta\} = \alpha, \quad (0 < \alpha < 0.5)$$

The question of the existence of such K_α was discussed during many years in the statistical literature. In 1964 Linnik, Romanovsky and Sudakov have shown (see e.g. Linnik (1966), (1968)) that if n_1 and n_2 are of different parities, then a set K_α to the Behrens-Fisher problem exists. But if n_1 and n_2 are of the same parities, the existence of a solution to the Behrens-Fisher problem remains an open question even today. There are many others modifications and generalizations of the Behrens-Fisher problem. For example, A. Wald posed the problem of existence of a critical set K_α in the sample space $\mathfrak{X}_2 \subset \mathbb{R}^2$ of the two-dimentional statistic

$$\mathbf{V}_2 = ((X_1 - X_2)/(S_1^2, S_1^2/S_2^2)^{\mathrm{T}}.$$

The solution of this problem has not been obtained as yet. However, it is effectively possible to find a set $K_\alpha^* \subset \mathfrak{X}_2$ such that

$$\mathcal{P}\{\mathbf{V}_2 \in K_\alpha^* | H_\delta\} \cong \alpha,$$

but in this case the probability $\mathcal{P}\{\mathbf{V}_2 \in K_\alpha^* | H_\delta\}$ depends on the unknown ratio σ_1^2/σ_2^2. This approach is used very often in statistical applications for the practical construction of tests for the comparison of a_1 and a_2. But the statistics of these tests are not expressed in terms of sufficient statistic $\mathbf{T}_4$ and hence usually these tests are less powerful than test based on the solution of the Behrens-Fisher problems and its generalizations.

Remark 2. We note here that many papers have been devoted to the problem of testing the hypothesis $H_0 : a_1 = a_2 = a, \quad (\delta = 0)$. For example, Welch (1947), James (1956, 1959), Brown and Forsythe (1974), Linnik (1966), Dijkstra and Werter (1981) and many others analysed the different approximations and some of them gave tables of critical values of the appropriate test statistics. Another approaches are also known. Gans (1981), for example, investigated the use of first a test for equality of the variances $\sigma_1^2 = \sigma_2^2$ and then Student's t-statistic or the alternate test, described in the above-mentioned papers, if equality is rejected.

2. ESTIMATION OF THE UNKNOWN COMMON MEAN

Evidently, in the case where the hypothesis H_0 is accepted, the problem of obtaining the best estimator for the unknown common mean arises. Naturally, we may use

both interval and point estimators of the parameter a. The problem of approximate interval estimator of a has been considered by James (1956), Pagurova and Gursky (1979), Marič and Graybill (1979), Khatri and Shah (1981) and others. It is interesting to note the main result of Marič and Graybill (1979), which consists in setting in some sense the "exact" confidence interval for a with confidence coefficient equal to any preselected value $1-\alpha$.

The problem of point estimation of the common mean a is also not simple. This is apparently due to the incompleteness of the minimal sufficient statistic. Usually the parameter a is estimated by the weighted mean

$$\hat{a} = \sum_{i=1}^{k} w_i X_i, \tag{1}$$

where

$$W_i = \frac{\frac{1}{Y_i}}{\sum_{i=1}^{k} \frac{1}{Y_i}} \quad \text{and} \quad Y_i = \frac{1}{n_i(n_i-1)} \sum_{j=1}^{n_i} (X_{ij} - X_i)^2 = \frac{S_i^2}{n_i(n_i-1)} = \frac{1}{n_i} s_i^2,$$

which is an **unbiased** but **nonefficient estimator** of a, see Hinckley (1979). We note that $n_i Y_i = S_i^2/f_i = s_i^2$ is the **MVUE for** σ_i^2, $i = 1,\ldots,k$. Because of this, many authors search for more efficient estimators of a, see Zacks (1966), Hinckely (1979), Bhattacharya (1980) and others. If the estimator (1) is accepted, we want also to obtain its variance and the unbiased estimator of this variance. These problems have been considered by, among others, Meier (1953), Cochran and Carroll (1953), Levi and Mantel (1974), Bement and Willams (1969), Norwood and Hinkelmann (1977). Norwood and Hinkelmann showed the non-efficient estimator (1) is in some sense optimal. More specifically they showed that the variance of $\hat{a}$ is less than any of the variance σ_i^2/n_i $(i = 1,\ldots,k)$ if and only if either $n_i > 9$ $(i = 1,\ldots,k)$ or, for some i, $n_i = 9$ and $n_j > 17$ $(j = 1,\ldots,k;\ j \neq i)$.

In view of the great practical importance of the problem of combining of estimators in the normal case and because of the artificiality of the estimator (1) the following approach seems interesting.

Let $X_1,\ldots,X_k, s_1^2,\ldots,s_k^2$ be mutually independent statistics, where

$$X_i \quad \text{is normal} \quad \mathrm{N}\left(a, \frac{\sigma_i^2}{n_i}\right) \quad \text{and} \quad s_i^2 = \frac{\sigma_i^2}{f_i}\chi_{f_i}^2,$$

$f_i = n_i - 1$, $i = 1,\ldots,k$, the values $f_1,\ldots,f_k$ being fixed and parameters $a, \sigma_1^2,\ldots,\sigma_k^2$ being all unknown.

Consider the class of functions $A_k(\cdot), B_k(\cdot), C_k(\cdot), \quad k \in \mathbb{N}$, mapping $\mathbb{R}^k \times \mathbb{R}^k_+ \times \mathbb{R}^k_+$ onto $\mathbb{R}^1$, where $\mathbb{R}^k$ is the Euclidean space of k dimensions, and

$$\mathbb{R}^k_+ = \{x = (x_1, \ldots, x_k) \in \mathbb{R}^k, \quad x_1 > 0, \ldots, x_k > 0\}.$$

That is, for any matrix of order $3 \times k$

$$\left\| \begin{matrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \\ z_1 & z_2 & \cdots & z_k \end{matrix} \right\| \tag{2}$$

with elements satisfying the following conditions

$$|x_i| < \infty, \quad y_i > 0, \quad z_i > 0, \quad i = 1, 2, \ldots, k,$$

functions $A_k(\cdot), B_k(\cdot), C_k(\cdot)$ give real numbers

$$A_k = A_k\left(\left\| \begin{matrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \\ z_1 & z_2 & \cdots & z_k \end{matrix} \right\|\right), \quad B_k = B_k\left(\left\| \begin{matrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \\ z_1 & z_2 & \cdots & z_k \end{matrix} \right\|\right),$$

$$C_k = C_k\left(\left\| \begin{matrix} x_1 & x_2 & \cdots & x_k \\ y_1 & y_2 & \cdots & y_k \\ z_1 & z_2 & \cdots & z_k \end{matrix} \right\|\right),$$

from $\mathbb{R}^1$. Suppose, further, that functions that $A_k(\cdot), B_k(\cdot), C_k(\cdot)$ satisfy the following conditions.

(1) $A_k(\cdot), B_k(\cdot), C_k(\cdot)$ are symmetric with respect to rearrangement of the columns of (2).

(2) If $k = 1$, then,

$$A_1 = A_1\left(\left\| \begin{matrix} x_1 \\ y_1 \\ z_1 \end{matrix} \right\|\right) = x_1, B_1 = B_1\left(\left\| \begin{matrix} x_1 \\ y_1 \\ z_1 \end{matrix} \right\|\right) = y_1, C_1 = C_1\left(\left\| \begin{matrix} x_1 \\ y_1 \\ z_1 \end{matrix} \right\|\right) = z_1,$$

from where one can see the sense of these functions A_k, B_k, C_k.

(3) For any real numbers α and β

$$A_k\left(\left\| \begin{matrix} \alpha x_1 + \beta & \alpha x_2 + \beta & \cdots & \alpha x_k + \beta \\ \alpha^2 y_1 & \alpha^2 y_2 & \cdots & \alpha^2 y_k \\ z_1 & z_2 & \cdots & z_k \end{matrix} \right\|\right) = \alpha A_k\left(\left\| \begin{matrix} x_1 & \cdots & x_k \\ y_1 & \cdots & y_k \\ z_1 & \cdots & z_k \end{matrix} \right\|\right) + \beta,$$

$$B_k\left(\left\|\begin{matrix}\alpha x_1+\beta & \alpha x_2+\beta & \cdots & \alpha x_k+\beta\\ \alpha^2 y_1 & \alpha^2 y_2 & \cdots & \alpha^2 y_k\\ z_1 & z_2 & \cdots & z_k\end{matrix}\right\|\right) = \alpha^2 B_k\left(\left\|\begin{matrix}x_1 & \cdots & x_k\\ y_1 & \cdots & y_k\\ z_1 & \cdots & z_k\end{matrix}\right\|\right),$$

$$C_k\left(\left\|\begin{matrix}\alpha x_1+\beta & \alpha x_2+\beta & \cdots & \alpha x_k+\beta\\ \alpha^2 y_1 & \alpha^2 y_2 & \cdots & \alpha^2 y_k\\ z_1 & z_2 & \cdots & z_k\end{matrix}\right\|\right) = C_k\left(\left\|\begin{matrix}x_1 & \cdots & x_k\\ y_1 & \cdots & y_k\\ z_1 & \cdots & z_k\end{matrix}\right\|\right).$$

(4) For $k=n+m$ and for

$$A_n^* = A_n\left(\left\|\begin{matrix}x_1 & \cdots & x_n\\ y_1 & \cdots & y_n\\ z_1 & \cdots & z_n\end{matrix}\right\|\right), A_m^{**} = A_m\left(\left\|\begin{matrix}x_{n+1} & \cdots & x_{n+m}\\ y_{n+1} & \cdots & y_{n+m}\\ z_{n+1} & \cdots & z_{n+m}\end{matrix}\right\|\right),$$

$$B_n^* = B_n\left(\left\|\begin{matrix}x_1 & \cdots & x_n\\ y_1 & \cdots & y_n\\ z_1 & \cdots & z_n\end{matrix}\right\|\right), B_m^{**} = B_m\left(\left\|\begin{matrix}x_{n+1} & \cdots & x_{n+m}\\ y_{n+1} & \cdots & y_{n+m}\\ z_{n+1} & \cdots & z_{n+m}\end{matrix}\right\|\right),$$

$$C_n^* = C_n\left(\left\|\begin{matrix}x_1 & \cdots & x_n\\ y_1 & \cdots & y_n\\ z_1 & \cdots & z_n\end{matrix}\right\|\right), C_m^{**} = C_m\left(\left\|\begin{matrix}x_{n+1} & \cdots & x_{n+m}\\ y_{n+1} & \cdots & y_{n+m}\\ z_{n+1} & \cdots & z_{n+m}\end{matrix}\right\|\right),$$

the following relations hold:

$$A_k = A_2\left(\left\|\begin{matrix}A_n^* & A_m^{**}\\ B_n^* & B_m^{**}\\ C_n^* & C_m^{**}\end{matrix}\right\|\right), B_k = B_2\left(\left\|\begin{matrix}A_n^* & A_m^{**}\\ B_n^* & B_m^{**}\\ C_n^* & C_m^{**}\end{matrix}\right\|\right), C_k = C_2\left(\left\|\begin{matrix}A_n^* & A_m^{**}\\ B_n^* & B_m^{**}\\ C_n^* & C_m^{**}\end{matrix}\right\|\right)$$

The proposed class is not empty, since there are functions $A_(\cdot), B_k(\cdot), C_k(\cdot)$ of matrices (2) satisfying conditions (1)-(4). Indeed, define functions $A_k(\cdot), B_k(\cdot), C_k(\cdot)$ as follows

(3)
$$\left.\begin{array}{l}\left\|\begin{matrix}x_1 & \cdots & x_k\\ y_1 & \cdots & y_k\\ z_1 & \cdots & z_k\end{matrix}\right\| \xrightarrow{A_k(\cdot)} A_k = \dfrac{\sum\limits_{i=1}^{k}\dfrac{x_i}{y_i}}{\sum\limits_{i=1}^{k}\dfrac{1}{y_i}},\\ \left\|\begin{matrix}x_1 & \cdots & x_k\\ y_1 & \cdots & y_k\\ z_1 & \cdots & z_k\end{matrix}\right\| \xrightarrow{B_k(\cdot)} B_k = \dfrac{1}{\sum\limits_{i=1}^{k}\dfrac{1}{y_i}},\\ \left\|\begin{matrix}x_1 & \cdots & x_k\\ y_1 & \cdots & y_k\\ z_1 & \cdots & z_k\end{matrix}\right\| \xrightarrow{C_k(\cdot)} C_k = \sum\limits_{i=1}^{k} z_i.\end{array}\right\}$$

Evidently, these functions satisfy condition (1). If $k=1$, then $A_1=x_1, B_1=y_1, C_1=z_1$. Hence, condtion (2) is also satisfied. Condition (3) holds, since

$$\frac{\sum_{i=1}^{k}\frac{\alpha x_i+\beta}{\alpha^2 y_i}}{\sum_{i=1}^{k}\frac{1}{\alpha^2 y_i}}=\alpha\cdot\frac{\sum_{i=1}^{k}\frac{x_i}{y_i}}{\sum_{i=1}^{k}\frac{1}{y_i}}+\beta,$$

$$\frac{1}{\sum_{i=1}^{k}\frac{1}{\alpha^2 y_i}}=\alpha^2\cdot\frac{1}{\sum_{i=1}^{k}\frac{1}{y_i}}$$

Finally, condition (4) holds since

$$\frac{\frac{\sum_{i=1}^{n}\frac{x_i}{y_i}}{\sum_{i=1}^{n}\frac{1}{y_i}}+\frac{\sum_{i=n+1}^{n+m}\frac{x_i}{y_i}}{\sum_{i=n+1}^{n+m}\frac{1}{y_i}}}{\frac{1}{\sum_{i=1}^{n}\frac{1}{y_i}}+\frac{1}{\sum_{i=n+1}^{n+m}\frac{1}{y_i}}}=\frac{\sum_{i=1}^{k=n+m}\frac{x_i}{y_i}}{\sum_{i=1}^{k=n+m}\frac{1}{y_i}},$$

$$\frac{1}{\frac{1}{\sum_{i=1}^{n}\frac{1}{y_i}}+\frac{1}{\sum_{i=n+1}^{n+m}\frac{1}{y_i}}}=\frac{1}{\sum_{i+1}^{k=n+m}\frac{1}{y_i}},$$

$$\sum_{i=1}^{n}z_i+\sum_{i=n+1}^{n+m}z_i=\sum_{i=n+1}^{k=n+m}z_i.$$

From this it follows that functions $A_k(\cdot), B_k(\cdot), C_k(\cdot)$ defined by (3) satisfy the conditions (1)-(4). Note that if $\alpha=-1$ and $\beta=0$,

$$A_k\left(\left\|\begin{matrix}-x_1 & \cdots & -x_k\\ y_1 & \cdots & y_k\\ z_1 & \cdots & z_k\end{matrix}\right\|\right)=-A_k\left(\left\|\begin{matrix}x_1 & \cdots & x_k\\ y_1 & \cdots & y_k\\ z_1 & \cdots & z_k\end{matrix}\right\|\right).$$

This follows from condition (3).

For convenience we denote here $\mathbf{V}_i = s_i^2 \quad (i = 1,\ldots,k)$ and let consider the statistic

$$\hat{a} = A_k\left(\left\|\begin{array}{ccc} X_1 & \cdots & X_k \\ \mathbf{V}_1 & \cdots & \mathbf{V}_k \\ f_1 & \cdots & f_k \end{array}\right\|\right),$$

which by virtue of (3) may be rewritten as

$$\hat{a} = a + A_k\left(\left\|\begin{array}{cccc} \sigma_1\mathbf{U}_1/\sqrt{n_1} & \sigma_2\mathbf{U}_2/\sqrt{n_2} & \cdots & \sigma_k\mathbf{U}_k/\sqrt{n_k} \\ \mathbf{V}_1 & \mathbf{V}_2 & \cdots & \mathbf{V}_k \\ f_1 & f_2 & \cdots & f_k \end{array}\right\|\right) = a + \xi$$

where

$$\mathbf{U}_i = \sqrt{n_i}\,\frac{X_i - a}{\sigma_i}, \quad i = 1,\ldots,k,$$

are independent standard normal variables. We now find the expected value $\mathrm{E}\xi$ of the random variable ξ. Under the assumption that

$$\mathbf{V}_i = v_1, \mathbf{V}_2 = v_2, \ldots, V_k = v_k, \quad (\mathbf{V}_i = s_i^2)$$

the conditional expectation $\mathrm{E}\{\xi|\mathbf{V}_1 = v_1, \mathbf{V}_2 = v_2, \ldots, \mathbf{V}_k = v_k\}$ is equal to

$$\frac{1}{(2\pi)^{k/2}}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} A_k\left(\left\|\begin{array}{ccc} \sigma_1 u_1/\sqrt{n_1} & \cdots & \sigma_k u_k/\sqrt{n_k} \\ v_1 & \cdots & v_k \\ f_1 & \cdots & f_k \end{array}\right\|\right) e^{-\frac{1}{2}\sum_{i=1}^{k} u_i^2}\, du_1 \ldots du_k =$$

$$\frac{(-1)^k}{(2\pi)^{k/2}}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} A_k\left(\left\|\begin{array}{ccc} -\sigma_1 t_1/\sqrt{n_1} & \cdots & -\sigma_k t_k/\sqrt{n_k} \\ v_1 & \cdots & v_k \\ f_1 & \cdots & f_k \end{array}\right\|\right) e^{-\frac{1}{2}\sum_{i=1}^{k} t_i^2}\, dt_1 \ldots dt_k =$$

$$(-1)^{2k+1}\mathrm{E}\{\xi|\mathbf{V}_1 = v_1, \mathbf{V}_2 = v_2, \ldots, \mathbf{V}_k = v_k\} =$$

$$-\mathrm{E}\{\xi|\mathbf{V}_1 = v_1, \mathbf{V}_2 = v_2, \ldots, \mathbf{V}_k = v_k\}.$$

Hence, if the expectation $\mathrm{E}\{\xi|\mathbf{V}_1 = v_1, \mathbf{V}_2 = v_2, \ldots, \mathbf{V}_k = v_k\}$ exists, it is identically equal to zero. From this fact we may conclude that $\mathrm{E}\xi = 0$ and $\mathrm{E}\hat{a} = a$. That is,

$$\mathrm{E}A_k\left(\left\|\begin{array}{ccc} X_1 & \cdots & X_k \\ s_1^2 & \cdots & s_k^2 \\ f_1 & \cdots & f_k \end{array}\right\|\right) = a.$$

We may set up here the following problems.

1. Define the class of all functions $A_k(\cdot), B_k(\cdot), C_k(\cdot)$ satisfying conditions (1)-(4) and find in this class an estimator a which has the minimal variance. Possibly, there are no other functions except $A_k(\cdot), B_k(\cdot), C_k(\cdot)$ defined by (3). These functions in such a case give the solution of the problem.

2. Find the variance $\mathbf{D}\hat{a}$ and its unbiased estimator.

Now we return to our problem of the estimation of the common mean of several independent normal distributions with different variances. In particular we shall obtain the variance of (1) and its **MVUE**.

3. THE VARIANCE OF THE WEIGHTED MEAN AND ITS MVUE

Using the fact that

$$\sum_{i=1}^{k} w_i = 1 \quad \text{and} \quad \mathrm{E}X_i^2 = a^2 + \sigma_i^2/n_i,$$

the variance of the weighted mean (1) may be written as

$$\mathbf{D}\hat{a} = \mathrm{E}\hat{a}^2 - \{\mathrm{E}\hat{a}\}^2 = \sum_{i=1}^{k} \mathrm{E}\{w_i^2\}\frac{\sigma_i^2}{n_i}. \tag{4}$$

Since the probability density function of the random variable $\mathbf{T}_i = 1/Y_i$ is

$$f(t_i) = \frac{b_i^{\frac{n_i-1}{2}}}{\Gamma\left(\frac{n_i-1}{2}\right)} \cdot t_i^{-\frac{n_i+1}{2}} \cdot e^{-b_i/t_i}, \quad t_i \geq 0,$$

where

$$b_i = \frac{n_i f_i}{2\sigma_i^2},$$

we have

$$\mathrm{E}\{w_j^2\} = \left\{\prod_{i=1}^{k} \frac{b_i^{\frac{n_i-1}{2}}}{\Gamma\left(\frac{n_i-1}{2}\right)}\right\} \int_0^\infty \cdots \int_0^\infty \frac{t_j^2}{\left(\sum_{i=1}^{k} t_i\right)^2} \cdot \prod_{i=1}^{k} t_i^{-\frac{n_i+1}{2}} \exp\left\{-\sum_{i=1}^{k} \frac{b_i}{t_i}\right\} dt_1 \ldots dt_k. \tag{5}$$

It is well known that

$$\left(\sum_{i=1}^{k} t_i\right)^{-2} = \int_0^\infty u \exp\left\{-u\sum_{i=1}^{k} t_i\right\} du.$$

Substituting this expression into (5) and rearranging the order of integration, we get

$$\begin{aligned}
\mathrm{E}\{w_j^2\} &= \left\{\prod_{i=1}^{k} \frac{b_i^{\frac{n_i-1}{2}}}{\Gamma\left(\frac{n_i-1}{2}\right)}\right\} \int_0^\infty u du \int_0^\infty \cdots \int_0^\infty t_j^{-\frac{n_j-3}{2}} e^{-ut_j - \frac{b_j}{t_j}} \cdot \\
&\cdot \prod_{i\neq j} \left(t_i^{-\frac{n_j+1}{2}} e^{-ut_i - \frac{b_i}{t_i}} \right) dt_1 \ldots dt_k = \\
&= b_j 2^{\frac{n-3k}{2}-1} \left\{\prod_{i=1}^{k} \frac{b_i^{\frac{n_i-1}{4}}}{\Gamma\left(\frac{n_i-1}{2}\right)}\right\} \int_0^\infty x^{\frac{n-k}{2}+1} \mathcal{K}_{\frac{n_i-5}{2}}\left(x\sqrt{b_i}\right) \cdot \\
&\cdot \prod_{i\neq j} \mathcal{K}_{\frac{n_i-1}{2}}\left(x\sqrt{b_i}\right) dx,
\end{aligned}$$

where $n = n_1 + n_2 + \cdots + n_k$ and $\mathcal{K}_i(\cdot)$ is the modified Bessel function. Using this result and formula (4), one may express the variance $\mathbf{D}\hat{a}$ of the weighted mean $\hat{a}$ as

$$\begin{aligned}
\mathbf{D}\hat{a} &= \frac{1}{2^{\frac{n-3k}{2}+1}} \left\{\prod_{i=1}^{k} \frac{b_i^{\frac{n_i-1}{4}}}{\Gamma\left(\frac{n_i-1}{2}\right)}\right\} \sum_{j=1}^{k} \left(\frac{n_j-1}{2}\right) \int_0^\infty x^{\frac{n-k}{2}+1} \cdot \\
&\cdot \mathcal{K}_{\frac{n_j-5}{2}}\left(x\sqrt{b_j}\right) \cdot \prod_{i\neq j} \mathcal{K}_{\frac{n_i-1}{2}}\left(x\sqrt{b_j}\right) dx. \qquad (6)
\end{aligned}$$

If $k = 2$, for example, see Gradshteyn and Ryzhik (1980),

$$\mathbf{D}\hat{a} = \frac{b_1^{\frac{n_1-1}{4}} b_2^{\frac{n_1-1}{4}}}{2^{\frac{n-4}{2}} \Gamma\left(\frac{n_1-1}{2}\right)\Gamma\left(\frac{n_2-1}{2}\right)} \left\{ \frac{b_1\sigma_1^2}{n_1} \int_0^\infty x^{\frac{n}{2}} \mathcal{K}_{\frac{n_1-5}{2}}\left(x\sqrt{b_1}\right) \cdot \right.$$

$$\left. \cdot \mathcal{K}_{\frac{n_2-1}{2}}\left(x\sqrt{b_2}\right) dx + \frac{b_2\sigma_2^2}{n_2} \int_0^\infty x^{\frac{n}{2}} \mathcal{K}_{\frac{n_1-1}{2}}\left(x\sqrt{b_1}\right) \mathcal{K}_{\frac{n_2-5}{2}}\left(x\sqrt{b_2}\right) dx \right\} =$$

$$= \frac{b_2^{\frac{n_1-1}{2}} \Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n_2+3}{2}\right)\sigma_1^2}{b_1^{\frac{n_2-1}{2}} \Gamma\left(\frac{n+2}{2}\right)\Gamma\left(\frac{n_2-1}{2}\right) n_1} {}_2F_1\left(\frac{n-2}{2}, \frac{n_2+3}{2}; \frac{n+2}{2}; 1-\frac{b_2}{b_1}\right) +$$

$$+ \frac{b_2^{\frac{n_1-1}{2}} \Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n_1+3}{2}\right)\sigma_2^2}{b_1^{\frac{n_2-1}{2}} \Gamma\left(\frac{n+2}{2}\right)\Gamma\left(\frac{n_1-1}{2}\right) n_2} {}_2F_1\left(\frac{n-2}{2}, \frac{n_2-1}{2}; \frac{n+2}{2}; 1-\frac{b_2}{b_1}\right).$$

Here ${}_mF_1(\alpha,\beta;\gamma;x)$ is the Gauss hypergeometric function. One may easily verify that this expression is identical to one given by Nair (1980) who used another approach.

Now we may find an unbiased estimate of (6). This problem reduces to constructing a statistic the expected value of which is exactly equal to $\mathbf{D}\hat{a}$. Since $\mathbf{D}\hat{a}$ does not depend on a, the complete sufficient statistic of our problem is the vector $(Z_1, Z_2, \ldots, Z_k)^{\mathrm{T}}$, where $Z_i = f_i/2Y_i \quad (i = 1,2,\ldots,k)$.

The probability density function of this random variable Z_i is

$$f(z_i) = \frac{[(n_i-1)b_i]^{\frac{n_i-1}{2}}}{2^{\frac{n_i-1}{2}} \Gamma\left(\frac{n_i-1}{2}\right)} z_i^{-\frac{n_i+1}{2}} \exp\left\{-\frac{(n_i-1)b_i}{2z_i}\right\}, \quad z_i \geq 0, \quad i = 1,\ldots,k.$$

Using tables of Gradshteyn and Ryzhik (1980), it is easy to verify that the unbiased estimates of

$$(7) \qquad b_i^{\frac{n_i-1}{4}} \mathcal{K}_{\frac{n_i-1}{2}}\left(x\sqrt{b_i}\right) \quad \text{and} \quad \frac{n_i-1}{2} b_i^{\frac{n_i-1}{2}} \mathcal{K}_{\frac{n_i-5}{2}}\left(x\sqrt{b_i}\right)$$

are

$$\frac{2^{\frac{n_i-3}{2}}\Gamma\left(\frac{n_i-1}{2}\right)}{x^{\frac{n_i-1}{2}}} \exp\left\{-\frac{x^2 z_i}{2(n_i-1)}\right\}$$

and

$$(8) \qquad \frac{2^{\frac{n_i-7}{4}}\Gamma\left(\frac{n_i-1}{2}\right) Z_i^{\frac{n_i-3}{4}}}{x^{(n_1-1)\frac{n_i-7}{4}}} W_{-\frac{n_i-3}{4},\frac{n_i-5}{4}}\left(\frac{x^2 Z_i}{2(n_i-1)}\right) e^{-\frac{x^2 Z_i}{4(n_i-1)}}$$

respectively, where $W_{\mu,\nu}(\cdot)$ is the Whittaker function. Substituting in (6) unbiased estimates (8) instead of (7) and performing the integration with respect to x, we get (see, also, Nikulin and Voinov, (1983), Voinov and Nikulin (1993))

$$\mathbf{D}\hat{a} = \Sigma_{j=1}^{k} \frac{Z_j^{\frac{n_j-3}{2}}}{2^{(n_j-1)\frac{n_j-3}{4}}} \left(\sum_{i=1}^{k} \frac{Z_i}{n_i-1}\right)^{-\frac{n_j-3}{2}} {}_2F_1\left(\frac{n_j-1}{2}, \frac{n_j-3}{2}; \frac{n_j+1}{2}; \frac{\Sigma_{i\neq j} \frac{z_i}{n_i-1}}{\Sigma_{i=1}^{k} \frac{z_i}{n_i-1}}\right) =$$

$$= \Sigma_{j=1}^{k} \frac{1}{Y_j^{\frac{n_j-3}{2}}} \left(\sum_{i=1}^{k} \frac{1}{Y_i^2}\right)^{-\frac{n_j-1}{2}} {}_2F_1\left(\frac{n_j-1}{2}, \frac{n_j-3}{2}; \frac{n_j+1}{2}; \frac{\Sigma_{i\neq j} \frac{1}{Y_i^2}}{\Sigma_{i=1}^{k} \frac{1}{Y_i^2}}\right).$$

Since

$${}_2F_1(a,b;c;x) = (1-x)^{c-a-b}\, {}_2F_1(c-a, c-b; c; x),$$

the unique minimum variance unbiased estimate of the variance of the weighted mean may be written as

$$(9) \qquad \mathbf{D}\hat{a} = \sum_{j=1}^{k} \frac{c^2}{Y_j^2} {}_2F_1\left(1,2; \frac{n_j+1}{2}; 1-\frac{c}{Y_j^2}\right), \quad \text{where} \quad c = \frac{1}{\Sigma_{i=1}^{k} \frac{1}{Y_i^2}}.$$

If

$$\min_{1\le i\le k} n_i \longrightarrow \infty, \quad \text{then} \quad {}_2F_1\left(1,2;\frac{n_i+1}{2};x\right) \longrightarrow 1$$

and expression (9) converges to $\left(\sum_{i=1}^{k} n_i/\sigma_i^2\right)^{-1}$ which is the variance of the best linear unbiased estimate of the weighted mean, if σ_i^2 are known. One may easily verify that expression (9) to, within terms of order $O(1/n_i^2)$, coincides with the approximately unbiased estimate of Meier (1953).

ACKNOWLEDGEMENTS

The authors wish to thank Prof. Terry Smith, Department of Mathematics and Statistics, Queen's University, Kingston, Canada for his helpful comments which helped to prepare this manuscript.

REFERENCES

[1] **Behrens W.U.** (1929). *Landwirtsch. Jahresber.* **68, 6**, 807–837.

[2] **Bement, T.R.** and **Williams, J.S.** (1969). "Variance of weighted regression estimators when sampling errors are independent and heteroscedastic". *JASA*, **64**, 1369–1382.

[3] **Bhattacharya, C.G.** (1980). "Estimation of a common mean and recovery of interblock information". *Ann. Statisti.*, **8, 1**, 205–211.

[4] **Brown, M.B.** and **Forsythe, A.B.** (1974). "The small sample behavior of some statistics which test the equality of several means". *Technometrics*, **16**, 129–132.

[5] **Cochran, W.G.** and **Carroll, S.P.** (1953). "A sampling investigation of the efficiency of weighting inversely as the estimated variance". *Biometrics*, **9**, 447–459.

[6] **Dijkstra, J.B.** and **Werter, P.S.P.J.** 1(1981). "Testing the equality of several means when the population variances are unequal". *Commun. in. Statist - Simula. Computa*, **B 10, 6**, 557–569.

[7] **Gans, D.J.** (1981). "Use of a preliminary test in comparing two sample means". *Commun. in Statist. - Simula. Computa*, **B 10, 2**, 163–174.

[8] **Gradshteyn, I.S.** and **Ryzhik, I.M.** (1980). "Table of integrals, series, and products". *Academic Press*, New York.

[9] **Hinckley, D.V.** (1979). "A note on the weighted means problem". *Scand. J. Statist.*, **6, 1**, 37–40.

[10] **James, G.S.** (1959). "The BehrensFisher distribution and weighted means". *Journal Royal Statist. Soc*, **B 21**, 73–90.

[11] **James, G.S.** (1956). "On the accuracy of weighted means and ratios". *Biometrics*, **43**, 304–321.

[12] **Khatri, C.G.** and **Shah, K.R.** (1981). "Interval estimation of the common mean". *Commun. in Statist. - Simula. Computa.*, **B 10, 2**, 99–107.

[13] **Levy, P.S.** and **Mantel, N.** (1974). "Combining unbiased estimates - A further examination of some old estimators". *J. Statist. Comput. Simul.*, **3**, 147–160.

[14] **Linnik, Yu.V., Romanovsky, I.V., Sudakov, V.N.** (1964). "A norandomized homogeneous test in the Behrens-Fisher problem". *Dokl. Acad. Nauk SSSR*, **155, 6**, 1262–1264.

[15] **Linnik, Yu.V.** (1966). "Randomized homogeneous tests for the Behrens-Fisher problem". In: *Selected Trans.in Math.Stat. and Prob.*, **6**, 207–217.

[16] **Linnik, Yu.V.** (1968). "Statistical problems with nuisance parameters". *Amer. Math. Soc.*, (translated from Russian).

[17] **Marič, N.** and **Graybill, F.A.** (1979). "Small sample confidence intervals on the common mean of two normal distributions with unequal variances". *Commun. in Statist. - Theor. Math.*, **A 8, 13**, 1255–1269.

[18] **Meir, P.** (1953). "Variance of a weighted mean". *Biometrics*, **9, 1**, 59–73.

[19] **Nair, K.A.** (1980). "Variance and distribution of the Graibill-Deal estimator of the common mean of two normal population". *Ann. of Statist.*, **8, 1**, 212–216.

[20] **Nikulin, M.S.** and **Voinov, V.G.** (1983). "On the problem of the weighted mean of several normal populations". *Preprint LOMI AN SSSR*, E183, Leningrad.

[21] **Norwood, T.E.** and **Hinkelmann, K.** (1977). "Estimating the common mean of several normal populations". *Ann. of Statist.*, **5, 5**, 1047–1050.

[22] **Pagurova, V.I.** (1968). "On a comparison of means of two normal samples". *Theory of probability and its applications*, **13, 3**, 527–534.

[23] **Pagurova, V.I.** and **Gursky, V.V.** (1979). "A confidence interval for the common mean of several normal distributions". *Theory of Probability and its applications*, **24, 4**, 882–888.

[24] **Voinov, V.G.** and **Nikulin, M.S.** (1993). "Unbiased estimators and their applications". **1**. Univariate case. *Kluwer*, Holland.

[25] **Welch, B.L.** (1947). "The generalization of "Student's" problem when several different population variances are involved". *Biometrika*, **34**, 28–35.

[26] **Zacks, S.** (1966). "Unbiased estimation of the common mean of two normal distributions based on small samples of equal size". *JASA*, **61**, 467–476.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 107-121, 1995

FITTING A LINEAR REGRESSION MODEL BY COMBINING LEAST SQUARES AND LEAST ABSOLUTE VALUE ESTIMATION

SIRA ALLENDE*, CARLOS BOUZA* and ISIDRO ROMERO†

Robust estimation of the multiple regression is modeled by using a convex combination of Least Squares and Least Absolute Value criterions. A Bicriterion Parametric algorithm is developed for computing the corresponding estimates. The proposed procedure should be specially useful when outliers are expected. Its behavior is analyzed using some examples.

Key words: Outliers in regression, L_1 regression, bicriterion parametric algorithm.

1. INTRODUCTION

The main objective of many applications is to obtain a function which should describe the relationship between a vector of known variables $\underline{x}^T = (X_1, \ldots, X_m)$ with $X_1 = 1$ and a response variable Y. That is the case when the abundance of phytoplankton (Y) is studied. The salinity, temperature and other variables can be measured. The biologist's aims is to establish a functional and to use the adjusted model for studying the effect of the explanatory variables in the abundance of phytoplankton. A classic approach is to suppose that

$$Y_i = \sum_{j=1}^{m} X_{ij} B_j + \varepsilon_1, \qquad i = 1, \ldots, n \tag{1.1}$$

* Sira Allende and Carlos Bouza. Universidad de La Habana.

† Isidro Romero. Universidad Autónoma de Puebla

– Article rebut el desembre de 1992.

– Acceptat el desembre de 1994.

is an adequate model. The parametric vector $\underline{B}^T = (B_1, \ldots, B_m)$ should be estimated. We usually want to compute an estimate $\hat{\underline{B}}$ of $\underline{B}$ which minimizes a certain function of the vector $(\mathfrak{E}_1, \ldots, \mathfrak{E}_n)^T$, where $\mathfrak{E}_i = Y_i - \underline{x}\underline{B}$ for $i = 1, \ldots, n$ and $\underline{x}_i^T = (X_{i1}, \ldots, X_{im})$. The classic approach is to minimize the squared deviations which leads to Least Squares (LS) estimation. The optimization problem to be solved here is:

$$\text{P1}: \quad \min_{\underline{B}} \sum_{i=1}^{n} \mathfrak{E}_i^2$$

$$\text{Subject to}: \quad \sum_{j=1}^{m} X_{ij}\hat{B}_j + \mathfrak{E}_i = Y_i; \quad i = 1, \ldots, n \tag{1.2}$$

However LS estimates are severly affected by outliers because of the weight given to each point in the solution of P1 is the same. Then, if a heavy tailed distribution generates some residuals, the influence of them may be very important. Therefore P1 does not provide adequate estimations of $\underline{B}$ in the study of phytoplankton because it is frequent to obtain some extreme data points. Several studies, see Dielman (1984), Dielman and Pfaffenberger (1988) and Efron (1988), have shown that Least Absolute Value (LAV) regression criterion provides improved estimates. The corresponding optimization problem is:

$$\text{P2}: \quad \min_{\underline{B}} \sum_{i=1}^{n} |\mathfrak{E}_i|$$

$$\text{Subject to (1.2)}$$

From a numerical point of view P2 is solved by computing the solution of the Linear Programming Problem

$$\text{P3}: \quad \min_{\underline{B}} \sum_{i=1}^{n} (d_i^+ + d_i^-)$$

$$\text{Subject to}: \quad \sum_{j=1}^{m} B_j X_{ij} + d_i^+ - d_i^- = Y_i; \quad i = 1, \ldots, n \tag{1.3}$$

$$d_i^+ \geq 0, \quad d_i^- \geq 0; \quad i = 1, \ldots, n \tag{1.4}$$

We will denote $d_i^+ + d_i^- = e_i$.

The equivalence between LAV problems and Linear Programming was pointed out by Charnes *et al.* (1955). Special purpose algorithms, based on modifications of the Simplex Method (SM), increased the possibilities of using LAV. The consistency and asymptotic efficiency of this method with respect to LS was established, see Basset and Koenker (1978) and Dielman and Pfaffenberg (1988). The analysis of the behavior of LAV estimation plays a key role in the evaluation of regression equation fitting. LAV estimates are recommended as a good starting point in the search for

a robust estimate of $\underline{B}$. As an example we can mention the results of Antonch and Bartkoviak (1988) in the study of the robutness of $\alpha-$trimmed and $\alpha-$winsorised estimators of $\underline{B}$. LAV and LS estimates may coincide but it follows from Jenssen's inequality that $\|\underline{e}\|_2 \leq \|\underline{e}\|_1$, where $\underline{e}^T = (e_1, \ldots, e_n)$. The idea of combining LAV and LS in a convex function is present in different previous publications. Arthanari and Dodge (1981) quoted the possibility of combining them for deriving a compromise estimator of $\underline{B}$. Dodge (1984) analysed the robustness of this procedure establishing that it provides the solution for a certain $M-$type estimator. Using the structure of the problem posed in that paper we have the optimization problem $Q(u)$ which is denoted as follows:

$$Q(u): \qquad \min_{\underline{B}} u \sum_{i=1}^{n} e_i^2 + (1-u) \sum_{i=1}^{n} e_i$$
$$\text{Subject to (1.3)}-\text{(1.4)}$$

The present paper follows a research of Allende and Bouza (1991). They studied the related optimization problem for $u \in [0,1]$ as appearing in $Q(u)$. Dodge (1984) solved a similar problem when u was known. The parameter u characterizes the contamination between an assumed normal distribution of the residuals and a heavy tailed one. We solve this problem formulating a Bicriterion Optimization Problem (BOP). A set of linear regression models is generated and one of them is selected on the basis of the analysis of values of a proposed coefficient. It has the property to increase with the goodness of the approximation of the equation to the observed data.

The Mathematical Model is discussed in the second section. The structure of the optimization problem is characterized by fixing the Karush-Kuhn-Tucker Conditions (KKTC). In the third section a model selection criterium is proposed. An algorithm for solving the problem, for a fixed value of u, is proposed. In Section 4 the results of a classic regression problems are used for evaluating the proposed procedure. A Monte Carlo experiment is performed and the behavior of the algorithm is discussed. The method is also used for fitting the regression equation of a set of data of the abundance of phytoplankton in the bay of Monterrey.

2. A CONVEX COMBINATION OF LS AND LAV

We consider the classic linear regression model $\underline{Y} = \underline{X}\underline{B} + \underline{\mathcal{E}}$. $\underline{Y} \in \mathbb{R}^n$ is an observable random vector and $\underline{\mathcal{E}} \in \mathbb{R}^n$ a random unobservable vector with $\mathrm{E}(\underline{\mathcal{E}}) = \underline{0}$ and $\underline{B} \in \mathbb{R}^m$ is unknown.

A compromise between LS and LAV for estimating $\underline{B}$ was expressed by Allende and Bouza (1991) in terms of a Bicriterion Optimization Problem (BOP) given by:

$$\min_{\underline{\hat{B}}} \left\{ \sum_{i=1}^{n} e_i^2, \quad \sum_{i=1}^{n} e_i \right\}$$
$$\text{Subject to: } (1.3)-(1.4)$$

The definition of e_i fixes that $\underline{e}^T = (e_1, \ldots, e_n) \geq \underline{0}$.

An efficient point of BOP represents a regression model such that there does not exist another offering a smaller LAV and LS errors simultaneously. The set of efficient points of BOP coincides with the set of optimal solutions of $Q(m)$. See Lommatsch (1979) for a detailed discussion of this result. If $m \neq 1$ then $Q(m)$ can be reformulated as

$$Q(\theta): \quad \min_{\underline{\hat{B}},\theta} \left\{ \sum_{i=1}^{n} e_i^2 + e \sum_{i=1}^{n} e_i : \quad \theta = m/1 - m \right\}$$
$$\text{Subject to: } (1.3)-(1.4)$$

θ will be called weight coefficient. $QP(\theta)$ is an one parameter Parametric Programming Problem and its special structure will be analyzed for obtaining a more efficient solution. The Karush-Kuhn-Tucker Conditions (KKTC) lead to a set of matrix equations which will be refered as the LCP−model or simply LCP. Taking $\underline{1}_m$ as a vector of m components equal to one, $\underline{g}(j)$, $j = 1, \ldots, 5$, and $\underline{B}(h)$, $h = 1,2$, as nonnegative vectors, $\underline{L}(h)$, $h = 1,2$, $\underline{g}(j)$, $j = 3,4,5$, and $\underline{e}$ belong to $\mathbb{R}^n$, $\underline{B}(h)$ and $\underline{g}(h)$, $h = 1,2$, belong to $\mathbb{R}^m$ and

$$(2.1) \qquad \underline{L}^T(1)\underline{g}(4) = \underline{L}^T(2)\underline{g}(5) = \underline{e}^T\underline{g}(3) = \underline{B}^T(1)\underline{g}(1) = \underline{B}^T(2)\underline{g}(2) = 0$$

The LC model is described as follows:

$$\begin{aligned}
\underline{X}^T\underline{L}(1) - \underline{X}^T\underline{L}(2) &= \underline{g}(1) \\
-\underline{X}^T\underline{L}(1) + \underline{X}^T\underline{L}(2) &= \underline{g}(2) \\
\theta\underline{1}_m + 2\underline{e} - \underline{L}(1) - \underline{L}(2) &= \underline{g}(3) \\
\underline{Y} - \underline{XB}(1) + \underline{XB}(2) + \underline{e} &= \underline{g}(4) \\
-\underline{Y} + \underline{XB}(1) - \underline{XB}(2) + \underline{e} &= \underline{g}(5)
\end{aligned}$$

Allende and Bouza (1991) proved that to solve $QP(\theta)$ is equivalent to obtain the solution of the LCP. The convexity of the objective function plays a key role in the proof.

An algorithm for estimating $\underline{B}$ should be able to determine, for each $\theta \geq 0$, an element of the set $G(\theta) = \left\{ (\underline{x}, \underline{g}) \in \mathbb{N}(\theta) : \underline{x}^T\underline{g} = 0 \right\}$, where $\underline{x} = (\underline{L}^T(1), \underline{L}^T(2), \underline{e}^T, \underline{b})^T$,

$\underline{b} = (\underline{B}^T(1), \underline{B}^T(2))$, $\mathbb{N}(\theta) = \{(\underline{x}, \underline{g}) : \underline{g} - \underline{W}\underline{x} = \underline{r}_0 + \theta \underline{r}_1\}$ and $\underline{g} = (\underline{g}^T(4), \underline{g}^T(5), \underline{g}^T(3), \underline{g}^T(1), \underline{g}^T(2))^T$. The vectors $\underline{r}_0$ and $\underline{r}_1$ are such that

$$r_{0i} = \begin{cases} 0 & \text{if} \quad i = 1, \ldots, 3n \\ Y_i & \text{if} \quad i = 3n+1, \ldots, 3n+m \\ -Y_i & \text{if} \quad i = 3n+m+1, \ldots, 3n+2m \end{cases} \tag{2.2}$$

$$r_{1i} = \begin{cases} 0 & \text{if} \quad i = 1, \ldots, 2n \\ 1 & \text{if} \quad i = 2n+1, \ldots, 3n \\ 0 & \text{if} \quad i = 3n+1, \ldots, 3n+m \end{cases} \tag{2.3}$$

Take $\underline{0}_{KH}$ as a matrix of zeros and $\underline{I}$ as the $m \times m$ identity matrix. (2.1) is a complementarity condition because, when the product of two variables is zero, each of them is called complement of the other one. The vector of the coefficient of the regression can be expressed by $\underline{B} = \underline{B}(1) - \underline{B}(2)$. The matrix $\underline{W}$ is given by:

$$\underline{W} = \begin{bmatrix} \underline{0}_{nn} & \underline{0}_{nn} & \underline{0}_{nm} & \underline{X}^T & -\underline{X}^T \\ \underline{0}_{nn} & \underline{0}_{nn} & \underline{0}_{nm} & -\underline{X}^T & \underline{X}^T \\ \underline{0}_{mn} & \underline{0}_{mn} & 2\underline{I} & -\underline{I} & -\underline{I} \\ -\underline{X} & \underline{X} & \underline{I} & \underline{0}_{mm} & \underline{0}_{mm} \\ \underline{X} & -\underline{X} & \underline{I} & \underline{0}_{mm} & \underline{0}_{mm} \end{bmatrix} \tag{2.4}$$

Using $\underline{W}$ the solution algorithm determines an element $(\underline{x}^*, \underline{g}^*)$ of the polyhedron $\mathbb{N}(\theta)$ with the property $\underline{x}^{*T}\underline{g}^* = 0$. Note that the points of $\mathbb{N}(\theta)$ that satisfy the constraint $-\underline{X}\underline{B} - \underline{e} \leq -\underline{Y}$ conforms $G(\theta)$.

Now we can formulate the following theorem:

Theorem 2.1

i) $G(\theta) \neq \Phi$ for all $\theta \geq 0$

ii) $G(\theta)$ is a closed face of the convex polyhedron $\mathbb{N}(\theta)$ or is equal to $\mathbb{N}(\theta)$.

Proof:

Take an arbitrary $\theta \geq 0$ and a particular vector $(\underline{x}_0, \underline{g})$ such that

$$\begin{aligned} \underline{e}_0 &= \underline{1}_n \theta/2 \\ \underline{L}(1)_0 &= \underline{L}(2)_0 = \underline{g}(3)_0 = \underline{g}(4)_0 = \underline{g}(5)_0 = \underline{0} \end{aligned}$$

fixing

$$B(1)_{0i} = \left\{ \begin{array}{ll} Y_i - \theta/2 & \text{if } \theta \leq 2Y_i \\ 0 & \text{otherwise} \end{array} \right\}$$

$$B(2)_{0i} = \left\{ \begin{array}{ll} -Y_i - \theta/2 & \text{if } \theta > 2Y_i \\ 0 & \text{otherwise} \end{array} \right\}$$

As $\underline{g}(1)_0 = \underline{g}(2)_0 = \underline{0}$ the constraints of $G(\theta)$ are satisfied. For obtaining ii) consider the partition of $\underline{W}$ given by:

$$\underline{W} = \left[\begin{array}{ccc} \underline{0}_{(2n)2n} & \underline{0}_{(2n)m} & \underline{A}_{(2n)2m} \\ \underline{0}_{(m)2n} & \underline{0}_{(m)m} & \underline{I}^*_{(m)2m} \\ -\underline{A}^T_{(2m)2n} & -\underline{I}^{*T}_{(2m)m} & \underline{0}_{(2m)2m} \end{array} \right]$$

where $\underline{I}^* = [-\underline{I}_{mm} \quad -\underline{I}_{mm}]$ and

$$\underline{A} = \left[\begin{array}{cc} \underline{X}^T & -\underline{X}^T \\ -\underline{X}^T & \underline{X}^T \end{array} \right]$$

Taking $\underline{v}^T = [\underline{v}^T(1), \underline{e}^T, \underline{v}^T(3)] \in \mathbb{R}^{3n+2m}$ the relation $\underline{v}^T \underline{W} \underline{v} = 2\underline{e}^T \underline{I} \underline{e} \geq 0$ permits to establish that $\underline{W}$ is positive semidefinite. From Theorem A.3.3 of Bank *et. al.* (1982) follows ii).

3. ALGORITHM FOR SELECTING THE REGRESSION EQUATION

We denote the prediction of Y_i, for a fixed weight coefficient θ, by $Y_{i\theta}$. An estimate of $\underline{B}$ is computed for each θ. Denote by $\underline{B}_\theta$ the corresponding estimator. The statistician needs to evaluate the behavior of the estimator with respect to various values of θ. The goodness of the regression equation obtained by the formula $B = \underline{X}\underline{B}_\theta + \underline{e}$ is measured by

$$\frac{\sum_{i=1}^{n}(Y_{i\theta} - \overline{Y})^2 + \theta \sum_{i=1}^{n} |Y_{i\theta} - \overline{Y}|}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 + \theta \sum_{i=1}^{n} |Y_i - \overline{Y}|} = R(\theta)$$

When $Y_{i\theta} \approx Y_i$, for any $i = 1, \ldots, n$, the coefficient $R(\theta)$ will be near to one and it will decrease with the $Y_{i\theta}$'s when they are very different from Y_i. Note that $R(\theta)$ plays a role similar to the determination coefficient in common LS regression problems.

The decision maker of the statistician can fix a threshold value of R_0 for judging whether a fitted regression equation is admissible or not. This criterion permits to

obtain a set with the regression equations which can be used for the decision making. Take $\underline{B}_\theta$ as the estimate of $\underline{B}$ obtained for a θ fixed. Then $M = \underline{X}_\theta \underline{B}_\theta R(\theta)$ is such set.

Consider the sets

$$\mathbb{Z} = \left\{ (\underline{x}, \underline{g}, \theta) \in \mathbb{R}^{3m+2n} \times \mathbb{R}^{3m+2n} \times \mathbb{R} : (\underline{x}, \underline{g}) \in \mathbb{N}(\theta) \right\}$$

and

$$\mathbb{C} = \left\{ (\underline{x}, \underline{g}, \theta) \in \mathbb{Z} : \underline{x}^T \underline{g} \geq 0 \right\}$$

We will obtain a description of the vertices in $\mathbb{Z}$ contained in $\mathbb{C}$ using the results of Bank *et al.* (1982). Take $V = \left\{ \underline{V}_n = {}^h\underline{x}, {}^h\underline{g}, {}^h\theta) \in \mathbb{Z} \cap \mathbb{C}, \quad h = 0, 1, \ldots, p \right\}$.

The parametric problem $QP(\theta)$ is solved by taking into account those vertices. The statistician has a set $\Theta = \{\theta_1, \ldots, \theta_s\}, s > 0$ is an arbitrary integer, of weight coefficients and the solution of (3.2) for each $\theta_j \in \Theta$ permits to determine $\underline{B}_{\theta j}$. The 'best' equation is selected analyzing the corresponding $R(\theta_j)$'s. If $M = \Phi$ we consider that (1.1) is an inadequate model. The least permissible value R_0 of the $R(\theta_j)$, $j = 1, \ldots, s$, is fixed by the statistician. The appropiate value of θ is determined by the following procedure:

Procedure for selecting the best equation

START: Give $\Theta = \{\theta_1, \ldots, \theta_s\}, R_0$ and the observations $(y_i, x_{i1}, \ldots, x_{im})$, $i = 1, \ldots, n$
BEGIN {main}
(1) Construct a vertex $\underline{v}_0$ of the polyhedron $\mathbb{Z}$ such that $\underline{v}_0 \in \mathbb{C}$
Repeat:=true
i:=0
(2) while repeat do
begin
If there is $\underline{v}_{i+1} \in \mathbb{Z}$ adjacent to $\underline{v}_i$ with $\theta_{i+1} > \theta_i$ and $\underline{v}_{i+1} \in \mathbb{C}$
then
i:=i+1
else
begin
Repeat:=false
{describe the unbounded edge of $\mathbb{Z}$ contained in $\mathbb{C}$:
$(\underline{d}_1, \underline{d}_2, \underline{d}_3) \in \mathbb{R}^{3m+2n} \times \mathbb{R}^{3m+2n} \times \mathbb{R}$}
end
end; while (2)

$$\underline{v}_i = ({}^i\underline{x}, {}^i\underline{g}, {}^i\theta), \quad i = 1, \ldots, p \text{ was calculated}$$
$$\underline{B}(i) = {}^i\underline{B}(1) - {}^i\underline{B}(2)$$

Repeat:=true

$j = 1; \quad \mathbb{R}^* = -\infty$

(3) While repeat do

begin

Calculation of $\mathbb{R}(\theta_j)$ for each $j = 1, \ldots, s$

(4) While $j \leq s$ do

Determine:

$$k = \max_{0 \leq i \leq p} \{i | {}^i\theta \leq \theta_j\}$$

If $k < p$ then

begin

$$a = 1 - \frac{\theta_j - k_\theta}{k + 1_\theta - k_\theta}$$

$$\underline{B}(j) = a\underline{B}(k) + (1-a)\underline{B}(k+1)$$

$$\underline{e}(j) = a^k \underline{e} + (1-a)^{k+1}\underline{e}$$

end;

else

begin

$$a = \theta_j / p_\theta$$

$$\underline{B}(j) = a\underline{B}(p)$$

$$\underline{e}(j) = a^p \underline{e}$$

end;

Calculate $\mathbb{R}(\theta_j)$

If $\mathbb{R}(\theta_j) > \mathbb{R}^*$ then $\mathbb{R}^* = \mathbb{R}(\theta_j)$

$j := j + 1$

end: {while (1)}

If $\mathbb{R}^* \geq \underline{R}_0$ then $\underline{B}_0$ describes the linear regression model

else

begin

write "do you want to give another value of θ?"

if (answer="not") then repeat:=false

else

$s = s + 1$

end: {while (3)}

END {main}

This procedure permits to determine a set of adequate regression equations.

Since $\underline{r}_1$, defined in (2.3), is non null: $\mathbb{Z}$ is not empty.

Considering $\theta = 0$ the initial point $\underline{v}_0$ is obtained by applying the algorithm of Lemke (1970). It is an algorithm based in a Simplex tableau and a complementarity

pivot technique. An artificial variable $\mathbb{Z}_0$ is introduced and a starting tableau is constructed by eliminating linear dependent rows and columns from $\underline{W}$. Then $\underline{W}'$ is obtained and an initial basic solution of the system $\underline{y}^* - \underline{W}'\underline{x} - \underline{x}_0 = \underline{r}_0$ is constructed with $\underline{y}$ as a basic value. By entering $\mathbb{Z}_0$ into the basis an index k is fixed by $r_{0k} = \min_i r_{0i}$ and y_k leaves the basis. The basic variables $y_j \quad (j \neq k)$ and $\mathbb{Z}_0$ have positive values. At each iteration the complement of the variable which lefts the basis in the previous iteration is entered. For the regression problem we can specify this algorithm in the following way:

T_0 : Starting Tableau

Basic variable	Non Basic Variable	$\mathbb{Z}_0$	
		-1	
		$\bullet$	
$\underline{g}'$	$\underline{W}'$	$\bullet$	$\underline{r}'_0$
		$\bullet$	
		-1	

The solution of the initial vertex is obtained by using the following procedure:

Procedure for obtaining the initial solution (PIS)

Begin

$\{T_0$ is not primal feasible because $r'_0 \not\geq 0\}$

EV:=$\mathbb{Z}_0$ (variable $\mathbb{Z}_0$ enters into the basis)

Determine an index such that

$|y_{i0}| = \max_{i=1,\ldots,m} |y_i|$

If $y_{i0} < 0$ then RV = $\underline{g}_{i_0}(4)$; $\{\underline{g}_{i_0}(4)$ leaves the basis$\}$

else

RV = $\underline{g}_{i_0}(5)$; $\{\underline{g}_{i_0}$ leaves the basis $\}$

T':= Pivoting T: $\{T'$ tableau is obtained after pivoting $T\}$

Repeat:=true

While Repeat do

begin

EV:=comp (RV): { The complement variable of RV is introduced into the basis}

$q :=$ index of EV

(4) Determine an index i_0 such that

$$t'_{i_0} = \min_{t_{i_q} > 0} \frac{t'_{i_0}}{t_{i_q}}$$

RV:= Basic Variable (i_0)\{ The basic variable (BV) corresponding to the index i_0 leaves the basis \}

If RV = $\mathbb{Z}_0$ then

Repeat:=false

$T' :=$ Pivoting T'

end; {while}

(5) Calculate

$t_\theta = f^{-1}\underline{r}_1$

{f denotes the basis corresponding to the tableau T' and f^{-1} its inverse}

Add the column t_θ to T'

{t_θ is the column corresponding to θ}

(6) Determine a Basic Index i such that $(i \in \mathrm{IB}(T'))$

$$t_{q0}/t_{q\theta} = \min_{\substack{t'_{i\theta}>0 \\ i\in \mathrm{IB}(T')}} t'_{i0}/t_{i\theta}$$

$T'' =$ Pivoting $T'(t_{q\theta} =$ pivot)

Exchange rows of T' in such a way that the last row corresponds to θ

END

This procedure computes an initial solution of LCP. The following theorem establishes the importance of PIS.

Theorem 3.1

A solution of LCP(0) is obtained by using PIS.

Proof:

PIS is a specification of Lamke's algorithm. It solves the linear complementary problem, for a finite number of iterations, for any right hand side vector if the matrix is copositive plus. Then we need to prove that $\underline{W}'$ is copositive plus. That is to derive if the following results hold:

a) $\underline{v}^T\underline{W}'\underline{v} \geq \underline{0}, \quad \forall \quad \underline{v} \in \mathbb{R}_+^{n+3m}$

b) $\underline{v}^T\underline{W}'\underline{v} = \underline{0} \quad \Rightarrow \quad (\underline{W}'^T + \underline{W}')\underline{v} = \underline{0}$

a) is obtained by using a procedure similar to that applied in Theorem 2.1 for deriving that $\underline{W}$ is a semidefinite positive matrix.

Taking

$$\underline{W}'^T + \underline{W}' = \begin{bmatrix} \underline{0}_{nm} & \underline{0}_{nm} & \underline{0}_{n(2m)} \\ \underline{0}_{mn} & 4\underline{I}_{mm} & \underline{0}_{m(2n)} \\ \underline{0}_{(2m)n} & \underline{0}_{(2m)n} & \underline{0}_{(2m)2n} \end{bmatrix}$$

$$\underline{v}^T \underline{W}' \underline{v} = 2\underline{Ie} = \underline{0} \quad \Rightarrow \quad (\underline{W}'^T + \underline{W}')\underline{v} = \underline{0}.$$

Therefore b) is satisfied and $\underline{W}'$ is copositive plus.

In order to obtain a vertex of $\mathbb{Z}$ belonging to $\mathbb{C}$, θ is introduced in the tableau as a variable in Step 5. The coefficients is the column entering into the basis are not smaller or equal to zero, since LCP(0) has a feasible solution. Hence θ is exchanged with one of the variables of the basis. According to the rules of the SM exactly one index k^*, with $(\underline{x}_{k^*}, \underline{g}_{k^*})$, fixes the non basic variables in the obtained solution. The solution obtained by using PIS corresponds to a vertex of $\mathbb{Z} \cap \mathbb{C}$, see Bank *et al.* (1982).

For solving the parametric dependent LCP it is necessary to modify the parametric principal pivoting algorithm refered by Pang and Lee (1981), because θ and $\underline{r}_1$ has null components.

The procedure for solving parameter dependent LCP needs, as initial information, the tableau T'', the index sets IB and IS of the basic and non basic variables respectively. The following procedure solves parameter dependent LCP:

Procedure for solving Parameter Dependent LCP

Begin
$T := T''$
cont:=1
EV:=$\underline{x}_{k^*}$
$h := K^*$
While cent ≤ 2 do
begin
$Q_0 = +\infty$
Determine an index $d \in \{1, \ldots, 3m+n\}$ such that
$Q_0 = t_{Q_0} = \min_{t_{i_h} > 0} t_{i_0}/t_{i_h}$
If $Q_0 < \infty$ and BV$(Q) \neq \theta$ then
begin
T'' = Pivoting (T)
$({}^{k+1}\underline{x}, {}^{k+1}\underline{g}, {}^{k+1}\theta)$: solution corresponding to T''

If BV$(Q) = x_j$ then
EV = g_j
$h = j + n$
If BV$(Q) = g_j$ then
EV = x_j
$h = j$

$T := T''$
$k = k + 1$
end

If $Q_0 = +\infty$ then
begin

$$z_h = \frac{\theta - k_\theta}{t_{3m+n,0}}$$

$\mathrm{BV}(i) = \mathrm{BV}(i) - t_{ih}, \quad i \in \mathrm{IB}$
$\mathrm{NBV}(j) = d_1, \quad j \neq h$

END

4. BEHAVIOR OF THE PROPOSED ALGORITHM

The proposed procedures were programmed in Fortran 77. The computer program uses the facilities of a standard Numerical Analysis Package (NAP) as MatLab for some calculations. Then, the accuracy of the approximations and the size of the solvable problem, depend of the characteristics of the MAP used. The code and more technical information is available from Dr. Romero. It can be implemented in an IBM 486 PC.

Dodge (1984) studied the robustness of the estimates derived by using $Q(u)$ for a fixed u. This procedure permits to compute $M-$type estimates of $\underline{B}$. The method proposed in this paper computes the best regression equation for a given set of values of $\theta = u/1 - u$. The values of θ reflect the ideas on the possible degree of contamination of the distribution of the residuals. For example, if it is close to a normal distribution function, with zero mean and variance σ^2, the selected model will have $\theta \approx 0$.

The proposed procedure is a generalization of the model of Dodge (1984). Its construction grants that the behavior of its solution is not worse than the use of LS or LAV as optimization criteria. Say that the estimates will not have a larger Residual Sum of Squares (RSS) and a Sum of Absolute Deviations (SAD) than any other derived by LS or LAV.

The data from the experiment of Hald (1952) used by Draper and Smith (1981) are reanalyzed. This is a classic example where LS is a good method. Table 4.1 gives the

corresponding results. The optimal convex combination was obtained for $u = 0.05$. Note that observation number 6 is considered as an extreme data point by the three methods. For LAV observation number 8 is another outlier. The regression equation fitted by the algorithm has a smaller RSS than the equation derived by the use of LAV and its SAD is smaller than the corresponding to LS method. This is expected because of the theoretical properties of the methods. As the degree of contamination is small the results of the recomended model are similar to the obtained using LS. The computing time for this example was 18.24 seconds.

Table 4.1

Analysis of Hald's data

Estimated Coefficients	LS	LAV	Convex Combination
B_1	62.206	62.405	62.300
B_2	1.551	1.551	1.551
B_3	0.510	0.499	0.511
B_4	0.102	0.102	0.101
B_5	−0.144	−0.130	−0.140
Item		Residuals	
1	0.005	0.548	0.149
2	1.511	1.103	1.395
3	−1.671	−1.326	−1.685
4	−1.727	−2.040	−1.830
5	0.251	0.368	0.185
6	3.925*	4.557*	3.890*
7	−1.449	−0.741	−1.411
8	−3.175	−3.408	−3.250
9	1.375	1.671	1.366
10	0.281	0.443	0.248
11	1.991	1.990	1.946
12	0.973	1.542	0.984
13	−2.294	−1.703	−2.286
RSS	45.037	50.919	47.784
SAD	20.628	18.321	20.625

A study of the abundance of phytoplankton (Y) produced 120 measurements in the analysis of the effect of pollution in the bay of Monterrey, Mexico. The independent variables were salinity (X_1), temperature of the water (X_2), Ph (X_3), intensity of the light (X_4), $X_5 = X_1X_2$, $X_6 = X_1X_3$, $X_7 = X_1X_4$, $X_8 = X_2X_3$, $X_9 = X_2X_4$ and $X_{10} = X_3X_4$. The behavior of the three approaches, in the presence of outliers was analyzed using Monte Carlo experiments. Twelfe points of the collected data, were considered as outliers by the experts. A percent of the generated sample was selected from the set of outliers. In each Monte Carlo experiment 0, 6, 12 or 24 experiment points were generated by means of a random sampling with replacement mechanism from the outliers. The rest of the units were selected from the not outliers. One hundred samples of size fifty were generated for each percent of outliers. Table 4.2 contains the means of the calculated RSS and SAD.

Table 4.2

Average of RSS and SAD in 100 Monte Carlo experiments with the abundance of phytoplankton data for $\theta \in \{0, 0.01, \ldots, 0.99\}$

	% of	Methods		
	outliers	LS	LAV	Convex Combination
R	0	105.8	124.4	110.3
S	6	135.9	188.7	177.6
S	12	158.0	188.6	185.3
	24	274.6	198.5	179.5
S	0	75.4	51.3	80.5
A	6	93.5	53.2	66.7
D	12	137.5	53.9	66.5
	24	245.6	85.8	100.5

Note that the existence of outliers does not affect seriously the accuracy of the regressions obtained by the proposed Convex Combination of LS and LAV, in the Montecarlo experiments. The mean computing time for deriving the solutions was 46.33 seconds.

ACKNOWLEDGMENT

Contracts with CONACYT supported this research. The computation was developed at Colegio de Computación of Benmérita Universidad Autónoma de Puebla,

México. The authors would like to thank helpful comments of the editor and referees which yield improvements in the quality of the original draft.

REFERENCES

[1] **Allende, S.M.** and **Bouza, C.N.** (1991). *A parametric programming approach to the estimation of the coefficients of the linear regression model.* Presented at "Parametric Optimization and Related Topics III" and "23 Jahrestagung Mathematische Optimisrung". Güstrow, Germany.

[2] **Antoch, J.** and **Barthowiak, A.** (1988). "$L-$estimation in linear models: Foundations and computational aspects. An application". *Biometrical Letters,* **25**, 3–24.

[3] **Arthanari, T.S.** and **Dodge, Y.** (1981). *Mathematical Programming in Statistis.* J. Wiley. N. York.

[4] **Bank, B., Guddat, J., Klatte, B.** and **Tammer, K.** (1982): *Nonlinear Parametric Optimization.* Akademie Verlag. Berlin.

[5] **Charnes, A., Cooper, W.W.** and **Ferguson, R.D.** (1955). "Optimal estimation of executive compensation by linear programming". *Management Sc.,* **1**, 138–151.

[6] **Dielman, T.E.** (1984). "Computational algorithms for calculating the Least Absolute Value and Chebyshov estimates for multiple regression". *American J. of Mathematical and Management Sc.,* **4**, 169–197.

[7] **Dielman, T.E.** and **Pfaffenberger, R.** (1988). "Least Absolute Value regression necessary sample size to use normal theory inference procedures". *Decision Sc.,* **19**, 734–743.

[8] **Dodge, Y.** (1984). "Robust estimation of regression coefficients by minimizing the convex combination of Least Squares and Least Absolute Deviation". *Computational Statistics Quart.,* **1**, 139–153.

[9] **Draper, N.R.** and **Smith, R.** (1981). *Applied Regression Analysis.* J. Wiley. N. York.

[10] **Efron, B.** (1978). "Computer intensive methods in statistical regression". *SIAM,* **30**, 421–449.

[11] **Hald, A.** (1952). *Statistical Theory Engineering Applications.* J. Wiley. N. York.

[12] **Lemke, C.E.** (1970). "Recent results on complementarity problems". In Rosen, J.B. *et al.* (eds.): *Non Linear programming Proceedings Symposium.* University of Wincosin, Madison. Academic Press. N. York, 349–384.

[13] **Lommatsch, K.** (1979). *Anwendung del Linearen Parametrichen Optimierung.* Akademie Verlag. Berlin.

[14] **Pang, J.S.** and **Lee, P.S.C.** (1981). "A parametric complementary technique for the computation of equilibrium prices in a single commodity spatial model". *Math. Programming,* **20**, 81–102.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 123-130, 1995

EMPIRICAL SIGNIFICANCE TEST OF THE GOODNESS-OF-FIT FOR SOME PYRAMIDAL CLUSTERING PROCEDURES

CARLES CAPDEVILA MARQUÈS* and ANTONI ARCAS PONS‡

Through a series of simulation tests by Monte Carlo methods, some aspects related to the inference concerning pyramidal trees build by the maximum and minimum methods are considered. In this sense, the quantiles of the γ-Goodman-Kruskal statistic allow us to tabulate a significance test of the goodness-of-fit of a pyramidal clustering procedure. On the other side, the pyramidal method of maximum is observed to be clearly better (more efficient) than that of the minimum in terms of the expected value for the gamma statistic. Both for the maximum and minimum methods, a relation between the number of objects to classify and the gamma distribution is observed.

Key words: Goodnes-of-fit, Pyramidal Clustering, Sample distribution of the γ-statistic, Simulation test.

1. INTRODUCTION

Ultrametric trees are the most studied representations using discrete models. The aim of this model is to achieve a family of partitions that can be interpreted as a set of "natural" classifications of the population to classify, Ω.

Pyramidal trees, introduced by E. Diday, are a logical generalization of ultrametric trees. They are less restrictive structures where recovering replaces the concept of

*Carles Capdevila Marquès. Departament de Matemàtica. Universitat de Lleida. C/. Bisbe Messeguer, s/n. 25003. Lleida.

‡Antoni Arcas Pons. Departament d'Estadística. Universitat de Barcelona. Avda. Diagonal, 645. 08028 Barcelona.

-Article rebut el juny de 1994.

-Acceptat el novembre de 1994.

partition, obtaining a representation which bears more information and is closer to the initial dissimilarities. Diday (1986), Fichet (1984), Durand (1986, 1988) have studied some interesting topics about this model.

The pyramidal–representation models intend to detect the presence of a pyramidal structure of the population, starting from a dissimilarity matrix concerning the population. The process having this aim consist in transforming the initial dissimilarity into a pyramidal dissimilarity by means of some known pyramidal clustering procedure algorithm; this pyramidal dissimilarity is equivalent to an indexed pyramid.

In applied problems it is necessary to measure the fitting between the pyramidal tree obtained from some algorithm and the initial structure. In this sense, the most used parameters are the γ–Goodman–Kruskal coefficient (1954) and the cophenetic correlation coefficient ρ (Farris J.S., 1969).

In spite of having these coefficients as a measure of the fitting between the initial structure and the pyramidal tree obtained, it is difficult to determine exactly up to which point these coefficients are really significant for some particular case.

For example, let $\Omega = \{\omega_1, \ldots, \omega_6\}$ and let δ be a dissimilarity on Ω given by the matrix

$$\delta = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ & 0 & 2 & 1 & 3 & 4 \\ & & 0 & 3 & 2 & 1 \\ & & & 0 & 2 & 3 \\ & & & & 0 & 4 \\ & & & & & 0 \end{pmatrix}$$

If we now carry out a pyramidal clustering procedure by the methods of the minimum and the maximum and calculate the value of the gamma and rho coefficients, we obtain: $\gamma = 0.92$ and $\rho = 0.89$ in the case of the maximum, and $\gamma = 0.80$ and $\rho = 0.59$ in the case of the minimum. Although these values are near to unity, there is nothing allowing us to assure whether they are really significant, i.e. up to which point they reflect the fitting between the initial structure and the pyramidal tree obtained. Nevertheless, it is necessary to find an objective criterion showing if the fitting is good. This becomes a characteristic problem in Statistical Inference.

On the other side, it would be also convenient to be able to evaluate the power of the pyramidal representation methods.

Generally, it is very difficult to find the exact distribution function for γ and ρ. Therefore, we shall develop our study from an empirical point of view, using some simulation techniques by means of Monte Carlo methods. For this purpose, it was necessary to programme a pyramidal clustering procedure algorithm and to create a simulation programme which made possible to obtain the sample distribution of

gamma and to tabulate a goodness-of-fit test of the pyramidal representation with regard to the γ statistic, using the methods of the maximum and the minimum.

Also, the γ distribution as a function of the number of objects to classify is being studied. Finally, some results refering to the power–efficiency of the methods of the maximum and the minimum, and no commented in this paper, are obtained.

2. SIMULATION TESTS

Two simulation tests have been carried out, which we have called S1 and S2, by means of the programmes SIMULU and SIMULN respectively, made up for this purpose.

Basing on a value n, which represents the number of objects in the population to classify, $\Omega = \{\omega_1, \ldots, \omega_n\}$, the S1 test consists in generating N random dissimilarities U(0,1), δ_i^n with i = 1, ..., N. Starting from each one of them, a pyramidal representation is carried out using the methods of the maximum and the minimum. By this means, we shall obtain N pyramidal dissimilarities $d_{M,i}^n$ and $d_{m,i}^n$ for each one of both methods; we shall then compare each pyramidal dissimilarity with the respective initial dissimilarity through the gamma coefficient and obtain N gamma values $\gamma_{M,i}^n$ in the case of the maximum method, and other N gamma values $\gamma_{m,i}^n$ in the case of the minimum method. From these N values the mean M_γ, the standard deviation S_γ and the quantiles Q_α ($\alpha = 0.05,\ 0.10,\ 0.50,\ 0.75,\ 0.90,\ 0.95$) of the γ statistic are calculated. This results are shown in Table 1 and Table 3.

In our study we have considered populations with $n = 4,5,\ldots,18,10,25$ objects. The number of simulations carried out was N = 200 for $n = 25$, and N = 1000 for the other values of n.

The S2 test was set out in the same terms as the S1 test, but replacing the random dissimilarities U(0,1) by values with a distribution N(0,1) adding the constant 10 in order to avoid negative values in the dissimilarity matrix. The results obtained in this test are shown in Table 2 and Table 4.

The aim of setting out a second test was mainly to see whether the distribution of the dissimilarities generated randomly had an effect anyhow on the results. As it can be seen from Tab. 1 and Tab. 2 as well as in Tab. 3 and Tab. 4, the results obtained in both cases (uniform or normal distribution) for the means, standard deviations and quantiles of gamma are virtually the same. As a conclusion, these results seem to point out that in case of a random assignation of dissimilarities, the relationship between the number of objects and the expected value of gamma does not depend on the distribution used for generating the random dissimilarities. Anyway, in order to confirm this result, which we are just suggesting here, it would be convenient to

carry out further tests with other distributions for the initial dissimilarities, which we leave for a later work.

3. GOODNESS-OF-FIT TEST IN A PYRAMIDAL CLUSTERING PROCEDURE

Table 1 and Table 2, show that the sample means of gamma and its standard deviations decrease as n increases. In addition, the results obtained by the minimum pyramidal clustering procedure coincide with those obtained by L. Hubert (1974), where a relationship between n and the sample mean of gamma, as well as between n and the standard deviation of gamma, in the case of the minimum hierarchical clustering procedure.

From theoretical hypothesis about Ω and Π (algorithm), it is very difficult to find the exact distribution function for γ and ρ. We have found an approximation of these functions by means of the quantiles obtained in the simulation (Tab. 3 and Tab. 4).

In a more concrete way, if $\Omega = \{\omega_1, \ldots, \omega_n\}$ and $\Delta = \{\delta_\Omega\}$, i.e. the dissimilarities family defined on Ω, we can consider Π the algorithm that transforms a dissimilarity defined on Ω into a pyramidal form.

$$\text{Let the function} \qquad \begin{array}{rccc} \gamma_\Pi & : & \Delta & \longrightarrow & \mathbb{R} \\ & & \delta_\Omega & \longrightarrow & \gamma_\Pi(\delta_\Omega) \end{array}$$

where $\gamma_\Pi(\delta_\Omega)$ is the Goodman–Kruskal coefficient between δ_Ω and the pyramidal dissimilarity obtained from Π.

If δ_Ω is pyramidal, then $\Pi(\delta_\Omega)$ is pyramidal too and coincides with δ_Ω, so $\gamma_\Pi(\delta_\Omega) = 1$. On the contrary, if δ_Ω is obtained by random generation, $\gamma_\Pi(\delta_\Omega)$ would be close to zero. In this way, if we consider a population Δ with random distances generated by some method, it could be possible to tabulate the distribution function for γ.

If we want to know if some dissimilarity obtained could be represented by a pyramidal structure, we are forced to consider a null hypothesis $\mathbf{H}_0$ that represents randomness in the sense that Δ contains pyramidal dissimilarities obtained using the process Π from random distances generated by some method. In this way, from the quantiles table of γ (n= $4, \ldots, 18, 20, 25$ objects), we obtain a goodness–of–fit test of the pyramidal representation using the minimum method and the maximum method. We also obtain that maximum method works better than minimum method in the above sense.

In practice, if the gamma value after applying a pyramidal clustering procedure is greater than the quantile Q_α, we can reject the randomness hypothesis for the initial dissimilarity at a significance level of $1 - \alpha$.

Relationship between the number n of objects of Ω and the sample mean (M_γ) and sample standard deviation (S_γ) for gamma, for the methods of the minimum and the maximum. Mean and Standard Deviation based on a sample of N = 200 for n = 25 and N=1000 for $n = 4, \ldots, 18, 20$.

Table 1

Initial random dissimilarity U(0,1)

	Maximum		Minimum	
n	M_γ	S_γ	M_γ	S_γ
4	0.96	0.07	0.78	0.21
5	0.85	0.10	0.63	0.19
6	0.74	0.10	0.54	0.17
7	0.66	0.09	0.47	0.14
8	0.59	0.09	0.40	0.13
9	0.53	0.08	0.36	0.11
10	0.49	0.07	0.33	0.10
11	0.45	0.07	0.30	0.09
12	0.42	0.07	0.28	0.09
13	0.39	0.06	0.25	0.08
14	0.36	0.06	0.23	0.08
15	0.34	0.06	0.22	0.07
16	0.32	0.05	0.21	0.06
17	0.30	0.05	0.20	0.06
18	0.28	0.05	0.18	0.06
20	0.26	0.05	0.16	0.05
25	0.21	0.04	0.13	0.04

Table 2

Initial random dissimilarity N(0,1)+10

	Maximum		Minimum	
n	M_γ	S_γ	M_γ	S_γ
4	0.96	0.07	0.78	0.21
5	0.85	0.10	0.63	0.19
6	0.75	0.11	0.54	0.16
7	0.66	0.10	0.46	0.15
8	0.59	0.09	0.41	0.13
9	0.54	0.09	0.36	0.12
10	0.49	0.08	0.33	0.10
11	0.45	0.07	0.30	0.09
12	0.42	0.07	0.28	0.09
13	0.39	0.07	0.25	0.08
14	0.36	0.06	0.24	0.07
15	0.34	0.06	0.21	0.07
16	0.32	0.06	0.20	0.06
17	0.30	0.05	0.19	0.06
18	0.28	0.05	0.18	0.05
20	0.26	0.05	0.16	0.05
25	0.21	0.04	0.13	0.04

Table 3

Relationship between n and the quantiles of the γ statistic. Initial random dissimilarity U(0,1)

	Maximum						
n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.86	0.86	0.86	1.00	1.00	1.00	1.00
5	0.68	0.71	0.78	0.85	0.91	1.00	1.00
6	0.56	0.61	0.68	0.75	0.81	0.86	0.90
7	0.49	0.53	0.60	0.66	0.72	0.78	0.82
8	0.43	0.47	0.52	0.60	0.65	0.71	0.74
9	0.39	0.43	0.47	0.53	0.58	0.64	0.67
10	0.37	0.39	0.44	0.49	0.53	0.58	0.61
11	0.33	0.35	0.40	0.45	0.50	0.55	0.58
12	0.29	0.32	0.37	0.42	0.47	0.51	0.54
13	0.28	0.30	0.34	0.39	0.43	0.46	0.48
14	0.26	0.28	0.32	0.36	0.40	0.44	0.46
15	0.24	0.26	0.30	0.34	0.38	0.42	0.44
16	0.23	0.25	0.28	0.32	0.36	0.39	0.41
17	0.21	0.23	0.26	0.30	0.34	0.37	0.39
18	0.21	0.22	0.25	0.28	0.32	0.35	0.37
20	0.17	0.19	0.22	0.26	0.29	0.32	0.33
25	0.13	0.15	0.18	0.21	0.23	0.25	0.26
	Minimum						
n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.45	0.45	0.64	0.82	1.00	1.00	1.00
5	0.31	0.37	0.48	0.64	0.77	0.88	0.94
6	0.24	0.32	0.43	0.54	0.64	0.76	0.81
7	0.24	0.28	0.37	0.46	0.56	0.66	0.72
8	0.19	0.23	0.31	0.41	0.49	0.57	0.61
9	0.17	0.21	0.28	0.36	0.44	0.51	0.55
10	0.16	0.19	0.26	0.32	0.40	0.46	0.50
11	0.14	0.17	0.23	0.30	0.36	0.41	0.46
12	0.13	0.16	0.22	0.27	0.33	0.38	0.41
13	0.12	0.15	0.20	0.25	0.30	0.35	0.38
14	0.12	0.14	0.18	0.23	0.29	0.33	0.36
15	0.10	0.13	0.17	0.21	0.26	0.31	0.34
16	0.10	0.12	0.16	0.20	0.25	0.29	0.31
17	0.10	0.12	0.16	0.20	0.24	0.28	0.30
18	0.08	0.11	0.14	0.18	0.22	0.26	0.28
20	0.08	0.10	0.13	0.16	0.20	0.23	0.25
25	0.06	0.09	0.10	0.13	0.16	0.19	0.20

Table 4

Relationship between n and the quantiles of the γ statistic. Initial random dissimilarity N(0,1) + 10

	Maximum						
n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.86	0.86	0.86	1.00	1.00	1.00	1.00
5	0.67	0.71	0.79	0.86	1.00	1.00	1.00
6	0.56	0.60	0.67	0.75	0.83	0.88	0.91
7	0.50	0.53	0.59	0.66	0.73	0.79	0.82
8	0.44	0.48	0.54	0.60	0.65	0.71	0.74
9	0.40	0.42	0.48	0.53	0.60	0.64	0.68
10	0.36	0.39	0.44	0.49	0.55	0.60	0.62
11	0.32	0.35	0.41	0.45	0.51	0.55	0.57
12	0.30	0.33	0.37	0.42	0.46	0.50	0.53
13	0.28	0.30	0.34	0.39	0.43	0.47	0.50
14	0.25	0.27	0.32	0.36	0.40	0.45	0.47
15	0.24	0.26	0.30	0.33	0.37	0.41	0.43
16	0.22	0.25	0.28	0.32	0.35	0.39	0.41
17	0.21	0.23	0.26	0.30	0.34	0.37	0.39
18	0.20	0.21	0.24	0.28	0.31	0.35	0.37
20	0.18	0.19	0.22	0.26	0.29	0.32	0.33
25	0.13	0.15	0.18	0.21	0.23	0.25	0.26
	Minimum						
n	$Q_{.05}$	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$	$Q_{.95}$
4	0.45	0.45	0.64	0.82	1.00	1.00	1.00
5	0.31	0.37	0.48	0.60	0.77	0.88	0.93
6	0.27	0.32	0.43	0.55	0.67	0.75	0.81
7	0.21	0.27	0.36	0.45	0.56	0.66	0.70
8	0.19	0.24	0.33	0.41	0.50	0.58	0.63
9	0.17	0.22	0.29	0.36	0.44	0.52	0.56
10	0.16	0.20	0.26	0.33	0.40	0.46	0.50
11	0.15	0.19	0.24	0.30	0.36	0.42	0.45
12	0.14	0.17	0.22	0.27	0.34	0.39	0.42
13	0.12	0.15	0.20	0.25	0.31	0.36	0.38
14	0.12	0.14	0.18	0.23	0.28	0.33	0.37
15	0.10	0.13	0.16	0.21	0.26	0.30	0.32
16	0.10	0.12	0.16	0.20	0.25	0.29	0.31
17	0.10	0.12	0.15	0.20	0.23	0.27	0.30
18	0.09	0.11	0.14	0.18	0.21	0.25	0.27
20	0.08	0.10	0.13	0.16	0.20	0.23	0.25
25	0.06	0.07	0.10	0.13	0.16	0.18	0.20

Finally, we have studied the efficiency of the maximum method and the minimum method through other simulation tests, in the sense of establishing which one best recovers a possible pyramidal structure underlying the initial data. In this sense, we would just point out that the results obtained in these tests show that the pyramidal method of the maximum generally is more efficient than the pyramidal method of the minimum.

REFERENCES

[1] **Baker, F.B., Hubert L.J.** (1975). "Measuring the power of hierarchical cluster analysis". *Journal of the American Statistical Association*, **Vol. 70 ,349**.

[2] **Bertrand P., Diday E.** (1985). "A visual representation of the compatibility between an order and a dissimilarity index: the pyramids". *Computational Statistics Quarterly*, **Vol. 2, Issue 1**, 1985, 31–41.

[3] **Bock, H.H.** (1984). "Statistical testing and evaluation methods in cluster analysis". *Proceedings of the Indian Statistical Institute Golden Jubilee*, International Conference on Statistics: Applications and New Directions.

[4] **Diday, E.** (1986). "Une représentation visuelle des classes empietantes: Les pyramides". *Rairo. Analyse des données.* **52,** 475–526.

[5] **Durand, C.** (1986). "Sur la représentation pyramidale en analyse des données". *Mémoire de DEA en Mathématiques Appliquées.* Université de Provence. Marseille.

[6] **Durand, C.** (1988). "Une approximation de Robinson inférieur maximale". *Rapport de Recherche. Laboratoire de Mathématiques Appliquées et Informatique.* **N88-02**. Université de Provence. Marseille.

[7] **Farris, J.S.** (1969). "On the cophenetic correlation coefficient". *Syst. Zoology*, **18(3)**, 279–285.

[8] **Fichet, B.** (1984). "Sur une extension de la notion de hiérarchie et son equivalence avec certaines matrices de Robinson". *Journées de Statistique.* Montpellier.

[9] **Goodman, L.A., Kruskal, W.H.** (1954). "Measures of association for cross-classification". *Journal of the American Statistical Association.* **Vol. 49, Dec.1954**, 732–764.

[10] **Hubert, L.J.** (1974). "Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures". *Journal of the American Statistical Association*, **Vol.69, 347.**

[11] **Jardine, N., Sibson, R.** (1971). *Mathematical Taxonomy.* Wiley, New York.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 131-151, 1995

PIRÁMIDES INDEXADAS Y DISIMILARIDADES PIRAMIDALES

CARLES CAPDEVILA MARQUÈS* y ANTONI ARCAS PONS‡

En este trabajo se pretende una formalización de las bases matemáticas sobre las que se amparan los modelos de clasificación y representación piramidal, estableciéndose para ello una equivalencia entre los conceptos de pirámide indexada y disimilaridad piramidal. Asimismo se precisa el concepto de información redundante, contenida en una estructura piramidal y se dan criterios objetivos para poder prescindir de ella sin que ello suponga una modificación de la estructura piramidal.

Indexed Pyramids and Pyramidal Dissimilarities.

Key words: Pyramid, Pyramidal dissimilarity, Spare groups.

1. INTRODUCCIÓN

Los modelos piramidales de clasificación y representación de datos multidimensionales, introducidos por Bertrand y Diday (1985) y estudiados más recientemente por otros autores, (Fichet, Durand, etc.) pretenden ser una generalización de los modelos jerárquicos clásicos, en el sentido de permitir la existencia no sólo de grupos disjuntos o encajados, sino también la existencia de grupos solapados y por tanto permitir clasificaciones en las que, a cada nivel, los grupos en que queda dividida la población no tengan que constituir forzosamente particiones de la misma, como ocurre en el caso de las clasificaciones inducidas por jerarquías, sino que puedan

*Carles Capdevila Marquès. Departament de Matemàtica. Universitat de Lleida. C/. Bisbe Messeguer, s/n. 25003. Lleida.

‡ Antoni Arcas Pons. Departament d'Estadística. Universitat de Barcelona. Avda. Diagona, 645. 08028 Barcelona.

-Article rebut el juny de 1994.

-Acceptat el maig de 1995.

constituir recubrimientos. En este caso puede suceder que un mismo individuo, o grupo de individuos, pertenezcan a dos grupos diferentes de una misma clasificación, con lo cual estos individuos podrán caracterizarse por las propiedades de los diversos grupos a los que pertenezcan. En este sentido podemos intuir que estos nuevos modelos de clasificación y representación de datos, deberán ser más próximos a la realidad que los modelos jerárquicos clásicos.

En este trabajo, pretendemos dar una visión rigurosa, desde el punto de vista de la formalización matemática, de las bases teóricas sobre las que se amparan los modelos piramidales de clasificación y representación. Más concretamente, se formaliza la relación existente entre pirámide indexada y disimilaridad piramidal.

Desde otro punto de vista se plantea también el problema de la información redundante contenida en una estructura piramidal y que, como tal, puede ser eliminada de la estructura simplificándola. En este sentido se precisa el concepto de información redundante y se dan criterios objetivos para prescindir de ella sin tener que modificar la pirámide correspondiente.

2. PRELIMINARES

En esta sección introducimos los conceptos básicos para el estudio de los modelos piramidales de clasificación y representación de datos, cuales son los de disimilaridad piramidal, pirámide, pirámide indexada, preorden compatible, etc. así como algunas propiedades de los mismos.

Definición 2.1

Dada una disimilaridad d sobre un conjunto finito Ω, diremos que es una disimilaridad piramidal, si verifica las condiciones:

$D.1$ $d(\omega_i, \omega_j) \geq 0, \quad \forall\, \omega_i, \omega_j \in \Omega$
$D.2$ $d(\omega_i, \omega_i) = 0, \quad \forall\, \omega_i \in \Omega$
$D.3$ $d(\omega_i, \omega_j) = d(\omega_j, \omega_i), \quad \forall\, \omega_i, \omega_j \in \Omega$
$D.4$ Existe un preorden total ($\leq$) sobre Ω, tal que para todo $\omega_i, \omega_j, \omega_k \in \Omega$, con $\omega_i \leq \omega_j \leq \omega_k$, $d(\omega_i, \omega_k) \geq \max\{d(\omega_i, \omega_j),\ d(\omega_j, \omega_k)\}$

Proposición 2.1

Una matriz simétrica de orden n, de términos reales positivos (d_{ij}), es una matriz asociada a una disimilaridad ultramétrica d, sobre un conjunto $\Omega = \{\omega_1, \ldots, \omega_n\}$, sii

existe una ordenación total de los individuos $(\omega_1 \leq \cdots \leq \omega_n)$, *de modo que la matriz* $M(d,\leq)$[1] *verifique:*

a) $d_{ij} \leq d_{ij+1}$ *para todo* $i \leq j$.

b) *Para todo* $k \in \{1,\ldots,n\}$, *si* $d_{kk+1} = \ldots = d_{kk+s+1} < d_{kk+s+2}$, *entonces:*
$d_{k+1j} \leq d_{kj}$ *para* $k+1 < j \leq k+s+1$
$d_{k+1j} = d_{kj}$ *para* $j > k+s+1$

La demostración puede encontrarse en Lerman (1981).

Proposición 2.2

Toda disimilaridad ultramétrica es también piramidal.

Demostración

Si d es una disimilaridad ultramétrica sobre un conjunto finito Ω, podemos considerar la ordenación sobre Ω establecida por la proposición 2.1.

Sean $\omega_i, \omega_j, \omega_k \in \Omega$ tales que $\omega_i \leq \omega_j \leq \omega_k$. Las condiciones a) y b) de la proposición anterior implican, respectivamente: $d(\omega_i,\omega_j) \leq d(\omega_i,\omega_k)$ y $d(\omega_j,\omega_k) \leq d(\omega_i,\omega_k)$. Por tanto, de ambas desigualdades obtendremos: $d(\omega_i,\omega_k) \geq \max\{d(\omega_i,\omega_j), d(\omega_j,\omega_k)\}$, con lo cual d será piramidal.

■

Definición 2.2

Dada una disimilaridad d, y un preorden $\leq$, definidos sobre Ω, diremos que d y $\leq$ son compatibles, si y solamente si, para cualesquiera elementos de Ω, tales que $\omega_i \leq \omega_j \leq \omega_k$, se verifica, $d(\omega_i,\omega_k) \geq \max\{d(\omega_i,\omega_j), d(\omega_j,\omega_k)\}$.

Definición 2.3

Sea Ω un conjunto finito preordenado y $h \in \wp(\Omega)$. Diremos que h es una parte conexa de Ω respecto al preorden, sii $\forall\, \omega_i, \omega_j \in h$, con $\omega_i \leq \omega_j$, $\{\omega \in \Omega / \omega_i \leq \omega \leq \omega_j\} \subset h$.

Definición 2.4

Diremos que un preorden definido sobre Ω es compatible con una familia $\mathrm{H} \subset \wp(\Omega)$, sii $\forall\, h \in \mathrm{H}$, h es una parte conexa respecto al preorden.

[1] Si d es una disimilaridad sobre un conjunto finito Ω, cualquier ordenación total $(\leq)$ de sus elementos, dará lugar a una matriz de disimilaridad para d, en la que la i-ésima columna (fila) representará al i-ésimo elemento de Ω respecto a la ordenación dada. Esta matriz se representa por $\mathrm{M}(d,\leq)$.

Proposición 2.3

Una disimilaridad d es una disimilaridad piramidal sobre Ω*, sii existe un preorden total sobre* Ω *compatible con d.*

La proposición es evidente a partir de las definiciones.

Definición 2.5

Sea Ω un conjunto finito y $\mathbf{P} \subset \wp(\Omega)$. Diremos que $\mathbf{P}$ es una pirámide, si se verifican las condiciones siguientes:

P.1 $\Omega \in \mathbf{P}$.
P.2 $\forall\, \omega \in \Omega, \quad \{\omega\} \in \mathbf{P}$.
P.3 $\forall\, h, h' \in \mathbf{P}, \quad h \cap h' = \emptyset \quad$ ó $\quad h \cap h' \in \mathbf{P}$.
P.4 Existe un preorden total sobre Ω, compatible con $\mathbf{P}$.

Definición 2.6

Sea $\mathbf{P}$ una pirámide sobre Ω.

Diremos que una función $i : \mathbf{P} \to \mathbb{R}^+$ es un índice asociado a $\mathbf{P}$, si verifica las condiciones:

I.1 $i(\{\omega\}) = 0, \quad \forall\, \omega \in \Omega$
I.2 $\forall\, h, h' \in \mathbf{P}, \quad h \subset h' \quad \Rightarrow \quad i(h) \geq i(h')$

Si $\mathbf{P}$ es una pirámide sobre Ω e i un índice asociado a $\mathbf{P}$, diremos que $(\mathbf{P}, i)$ es una pirámide indexada.

Definición 2.7

Si $(\mathbf{P}, i)$ es una pirámide indexada tal que:

$$\text{I.3} \quad \forall\, h, h' \in \mathbf{P}, \quad h \subsetneq h' \quad \Rightarrow \quad i(h) < i(h')$$

diremos que $(\mathbf{P}, i)$ es una pirámide indexada en sentido estricto.

Si en lugar de la condición I.3 se verifica:

I.3′ $\forall\, h, h' \in \mathbf{P}$ tales que $h \subsetneq h'$ y $i(h) = i(h')$,
existen $h_1, h_2 \in \mathbf{P}$, distintos de h, y tales que $h = h_1 \cap h_2$.

Entonces diremos que $(\mathbf{P}, i)$ es una pirámide indexada en sentido amplio.

De las definiciones anteriores resulta inmediato:

Proposición 2.4

Toda pirámide indexada en sentido estricto, lo es también en sentido amplio.

A los elementos de una pirámide los llamaremos "*grupos*" de la pirámide, y si una pirámide es indexada, el valor del índice sobre cada grupo, se denomina "*índice del grupo*".

Es interesante observar, tal como se demuestra en Diday (1986), que toda jerarquía total indexada es también una pirámide indexada, con lo cual puede concluirse que las pirámides son una generalización natural de las jerarquías.

En este sentido, puede verse que cada grupo de una pirámide puede tener hasta dos predecesores, mientras que en una jerarquía cada clase tiene un único predecesor.

Del mismo modo que los dendrogramas proporcionan una representación visual de las Jerarquías Indexadas, es posible construir otro tipo de grafos conexos, a los que llamaremos "*grafos piramidales*" o "*pirámides*", que proporcionen una representación visual de las Pirámides Indexadas.

Su construcción consiste en expresar gráficamente las propiedades características de las pirámides, es decir:

Dos grupos, no encajados, pueden tener intersección no vacía.
Cada grupo puede tener hasta dos predecesores.
Existencia de un preorden total para el que cada grupo sea conexo.

El grafo piramidal expresará entonces, en un solo diagrama, la imbricación de unos grupos en otros y el índice de la pirámide. Su representación visual se efectuará asociando a cada grupo un segmento horizontal situado, sobre una escala, a una altura igual a la de su índice y unido a sus predecesores mediante unas líneas oblicuas, que llamaremos "*aristas*", de manera que cada arista una el centro del segmento asociado a un grupo con uno de los extremos de los segmentos asociados a sus predecesores.

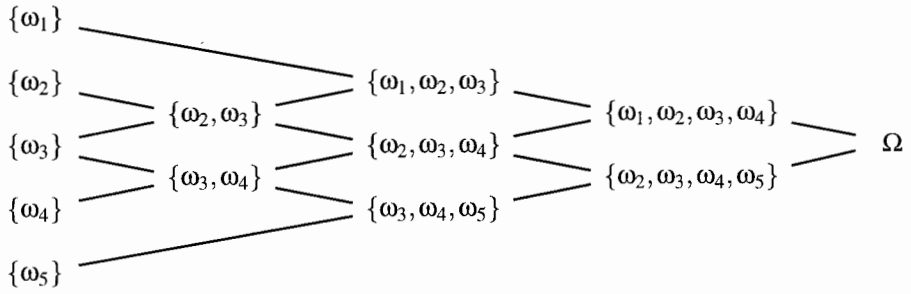
Ejemplo 2.1

Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$, ordenado por $\omega_1 \leq \omega_2 \leq \omega_3 \leq \omega_4 \leq \omega_5$ y sea

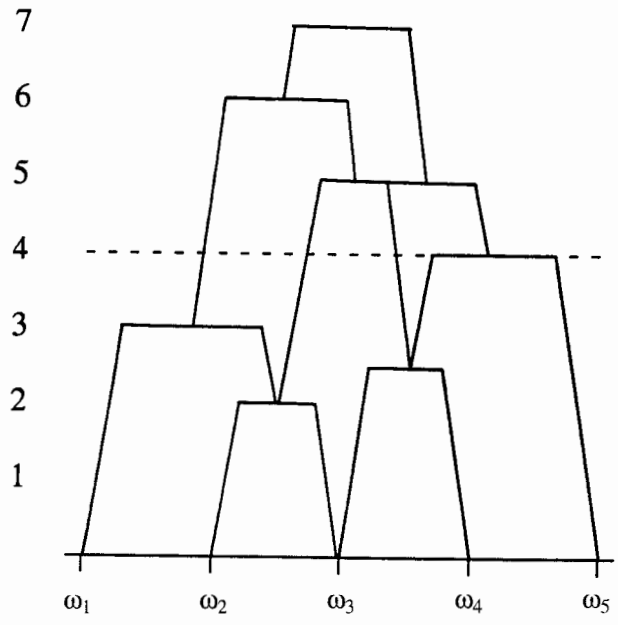
$\mathbf{P} = \{\{\omega_1\}, \ldots, \{\omega_5\}, \{\omega_2, \omega_3\}, \{\omega_3, \omega_4\}, \{\omega_1, \omega_2, \omega_3\}, \{\omega_3, \omega_4, \omega_5\}, \{\omega_2, \omega_3, \omega_4\}, \{\omega_2, \omega_3, \omega_4, \omega_5\}, \{\omega_1, \omega_2, \omega_3, \omega_4\}, \Omega\}$ una pirámide indexada por:

$i(\{\omega_j\}) = 0$, para $j \in \{1, \ldots, 5\}$
$i(\{\omega_2, \omega_3\}) = 2$
$i(\{\omega_3, \omega_4\}) = 2.5$
$i(\{\omega_1, \omega_2, \omega_3\}) = 3$
$i(\{\omega_3, \omega_4, \omega_5\}) = 4$
$i(\{\omega_2, \omega_3, \omega_4\}) = 5$
$i(\{\omega_2, \omega_3, \omega_4, \omega_5\}) = 5$
$i(\{\omega_1, \omega_2, \omega_3, \omega_4\}) = 6$
$i(\Omega) = 7$

Si partimos de los singletones, $\{\omega_1\},\ldots,\{\omega_5\}$ y determinamos los sucesivos predecesores, obtendremos:



La representación visual de la pirámide $(\mathbf{P}, i)$ será:



Obsérvese como, a cada nivel, la clasificación correspondiente constituye un recubrimiento de la población Ω. A nivel 4, por ejemplo, la clasificación piramidal obtenida viene dada por los grupos: $\{\omega_1,\omega_2,\omega_3\},\{\omega_2,\omega_3\},\{\omega_3,\omega_4\},\{\omega_3,\omega_4,\omega_5\}$, que obviamente forman un recubrimiento de Ω.

3. RELACIÓN ENTRE PIRÁMIDES INDEXADAS Y DISIMILARIDADES PIRAMIDALES

3.1. Planteamiento del problema

Teniendo en cuenta que la base matemática de las representaciones jerárquicas se fundamenta en la relación existente entre la obtención de una disimilaridad ultramétrica sobre una población Ω y una jerarquía indexada de $\wp(\Omega)$, en esta sección demostraremos la existencia de una correspondencia biunívoca, entre las pirámides indexadas y las disimilaridades piramidales, de modo que los grafos piramidales puedan interpretarse como la representación visual de dichas disimilaridades.

Para ello, tendremos en cuenta el paralelismo existente con el caso de las jerarquías y las ultramétricas antes mencionado.

Sea Ω un conjunto finito y d una disimilaridad piramidal sobre Ω. Nuestro objetivo será el de asociar a la disimilaridad d una pirámide indexada de $\wp(\Omega)$.

Sea $A = \{\alpha \in \mathbb{R} / \exists\ \omega_i, \omega_j \in \Omega \quad y \quad d(\omega_i, \omega_j) = \alpha\}$, para cualquier $\alpha \in A$, definimos la siguiente relación sobre Ω:

$$\forall\, \omega, \omega' \in \Omega; \qquad \omega R_\alpha \omega' \quad \Leftrightarrow \quad d(\omega, \omega') \leq \alpha$$

Si d fuese una disimilaridad ultramétrica, $\forall\, \alpha \in A$ las relaciones R_α serían de equivalencia. En nuestro caso, si d es piramidal, las propiedades reflexiva y simétrica se satisfacen de forma inmediata, sin embargo ello no es suficiente para asegurar la transitividad de dichas relaciones, por lo que no serán de equivalencia, y por tanto no darán lugar, en general, a particiones de Ω, sino a ciertos recubrimientos, a los que denominaremos P_α. Así pues, $\forall\, \alpha \in A$ el recubrimiento P_α estará formado por grupos $h \in \wp(\Omega)$ tales que $d(\omega, \omega') \leq \alpha, \quad \forall\, \omega, \omega' \in h$.

A pesar de todo, podríamos pensar que la unión de todos estos recubrimientos, $P' = \bigcup_{\alpha \in A} P_\alpha$ es la pirámide inducida por la disimilaridad d. Pues bien, esto no siempre es cierto, puesto que la intersección de dos grupos de un tal P', no siempre es de P' o vacía, tal como puede verse en el ejemplo 3.1.

Ejemplo 3.1

Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ y d una disimilaridad piramidal sobre Ω, dada por la matriz:

$$\begin{bmatrix} 0 & 1 & 1 & 3 \\ & 0 & 1 & 2 \\ & & 0 & 2 \\ & & & 0 \end{bmatrix}$$

En este caso, $A = \{0,1,2,3\}$ y construyendo los distintos recubrimientos de Ω, para los distintos valores de α, tendremos:

$$\begin{aligned} P_0 &= \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\}\} \\ P_1 &= \{\{\omega_1,\omega_2,\omega_3\},\{\omega_4\}\} \\ P_2 &= \{\{\omega_1,\omega_2,\omega_3\},\{\omega_2,\omega_3,\omega_4\}\} \\ P_3 &= \{\Omega\} \end{aligned}$$

con lo que $P' = \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\},\{\omega_1,\omega_2,\omega_3\},\{\omega_2,\omega_3,\omega_4\},\Omega\}$ y puesto que existe una intersección que no es de P', $\omega_1,\omega_2,\omega_3\} \cap \{\omega_2,\omega_3,\omega_4 = \{\omega_2,\omega_3\} \notin P'$, P' no es una pirámide.

Así pues deberemos reconsiderar el método de construcción de una pirámide a partir de una disimilaridad piramidal.

En este empeño, para cualquier $\alpha \in A$, consideraremos las siguientes familias de $\wp(\Omega)$:

$$\begin{aligned} \mathbf{C} &= \{h \in \wp(\Omega)/\exists\, \omega_i,\omega_j \in \Omega; \quad d(\omega_i,\omega_j) = \alpha, \quad h = \{\omega \in \Omega/\omega_i \leq \omega \leq \omega_j\}\}^2 \\ \mathbf{C} &= \{h \in \mathbf{C}_\alpha / h\text{es maximal en } \mathbf{C}_\alpha\text{, en el sentido de la inclusión}\} \end{aligned}$$

Precisando un poco más, $h \in \mathbf{C}_\alpha$ sii existen $\omega_i,\omega_j \in \Omega$, a distancia α, y $h = \{\omega \in \Omega/\omega_i \leq \omega \leq \omega_j\} = [\omega_i,\omega_j]_\alpha$. Por consiguiente, $h \in \mathbf{C}^*_\alpha$ sii $\exists\, \omega_i,\omega_j \in \Omega/d(\alpha_i,\alpha_j) = \alpha$ y $h = [\omega_i,\omega_j]_\alpha$ maximal. Es decir, los grupos $h \in \mathbf{C}^*_\alpha$ son intervalos maximales, en el sentido de que $\forall\, \omega' \notin h, \quad \exists \omega \in h$ tal que $d(\omega,\omega') > \alpha$.

Obsérvese que en el caso particular de la población Ω y de la disimilaridad piramidal d del ejemplo 3.1, estas familias serían:

$$\begin{aligned} \mathbf{C}_0 &= \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\}\} \\ \mathbf{C}_1 &= \{\{\omega_1,\omega_2\},\{\omega_1,\omega_2,\omega_3\},\{\omega_{2,3}\}\} \\ \mathbf{C}_2 &= \{\{\omega_2,\omega_3,\omega_4\},\{\omega_3,\omega_4\}\} \\ \mathbf{C}_3 &= \{\Omega\} \end{aligned} \qquad \begin{aligned} \mathbf{C} &= \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\}\} \\ \mathbf{C} &= \{\{\omega_1,\omega_2,\omega_3\}\} \\ \mathbf{C} &= \{\{\omega_2,\omega_3,\omega_4\}\} \\ \mathbf{C} &= \{\Omega\} \end{aligned}$$

A partir de estas nuevas familias de $\wp(\Omega)$, podríamos intentar definir la pirámide inducida por d como $P^* = \bigcup_{\alpha \in A} \mathbf{C}^*_\alpha$; pero P^*, en general, tampoco será una pirámide puesto que en el caso particular de la población y la disimilaridad piramidal considerados en el ejemplo 3.1, resulta ser $P^* = P'$, con lo cual P^* tampoco será la pirámide buscada.

[2]Esta definición tiene sentido puesto que al ser d una disimilaridad piramidal, debe existir sobre Ω un preorden total, lo cual permite, dados dos individuos $\omega_i,\omega_j \in \Omega$, a distancia α, considerar los individuos $\omega \in \Omega$ tales que $\omega_i \leq \omega \leq \omega_j$.

Por otra parte, el resultado $P^* = P'$, que hemos obtenido en el caso particular del ejemplo 3.1, puede generalizarse, como se demuestra en la siguiente proposición.

Proposición 3.1.1

Si d es una disimilaridad piramidal sobre Ω, $A = \{\alpha \in \mathbb{R} / \exists\ \omega_i, \omega_j \in \Omega$ *y* $d(\omega_i, \omega_j) = \alpha\}$ *y* $\forall\ \alpha \in A$, $\mathbf{C}^*_\alpha$ *y* P_α, *las familias de* $\wp(\Omega)$ *definidas anteriormente, entonces fijado* $\alpha \in A$, *se tiene* $\mathbf{C}^*_\alpha = P_\alpha$.

Demostración

Veamos en primer lugar que fijado $\alpha \in A, \mathbf{C}^*_\alpha \subset P_\alpha$.

Sea $h \in \mathbf{C}^*_\alpha$, esto significa que existirán $\omega_i, \omega_j \in \Omega$ a distancia α, tales que $h = [\omega_i, \omega_j]_\alpha$ maximal, y por tanto $\forall\ \omega' \notin h$, $\exists\ \omega \in h$ con $d(\omega, \omega') > \alpha$, lo cual equivale a decir que $\forall\ \omega \in h$ que esté a distancia menor o igual que α de cualquier otro $\omega' \in \Omega$, deberá ser $\omega' \in h$, con lo cual tendremos que $h = \{$individuos de Ω con interdistancias $\leq \alpha\}$ y por tanto será $h \in P_\alpha$.

Si por el contrario partimos de un grupo $h \in P_\alpha$, será $h = \{$individuos de Ω con interdistancias $\leq \alpha\}$. Puesto que $\alpha \in A$, existirán $\omega_i, \omega_j \in \Omega$ tales que $d(\omega_i, \omega_j) = \alpha$ y supongamos que ω_i y ω_j sean los individuos de Ω extremos, (en términos del preorden total existente en Ω, asociado a d) a distancia α. En estas condiciones, $\forall\ \omega \in h$ deberá ser $\omega_i \leq \omega \leq \omega_j$, puesto que si fuese $\omega < \omega_i \leq \omega_j$, por ser d piramidal tendríamos $d(\omega, \omega_j) \geq \alpha$. Por otra parte, puesto que $d(\omega_i, \omega_j) = \alpha$, ω_i y ω_j serán de h, y puesto que ω también es de h, será $d(\omega, \omega_j) \leq \alpha$. Por tanto tendríamos $d(\omega, \omega_j) = \alpha$ y $\omega < \omega_i \leq \omega_j$, con lo cual ω_i y ω_j no serían los individuos extremos a distancia α.

Así pues, $h \subset \{x \in \Omega / \omega_i \leq x \leq \omega_j\} = [\omega_i, \omega_j]_\alpha$.

Por otra parte, $\forall\ x \in \Omega$ tal que $\omega_i \leq x \leq \omega_j$ con $d(\omega_i, \omega_j) = \alpha$, deberá ser $x \in h$, puesto que $\forall\ x' \in [\omega_i, \omega_j]_\alpha$, $d(x, x') \leq d(\omega_i, \omega_j)$ por ser d piramidal, y por tanto $d(x, x') \leq \alpha$, con lo cual tendremos que $x \in h$.

De ambas inclusiones resultará $h = [\omega_i, \omega_j]_\alpha$. Además, al ser ω_i y ω_j los individuos extremos a distancia α, el intervalo $[\omega_i, \omega_j]_\alpha$ será maximal en $\mathbf{C}_\alpha$ respecto a la inclusión.

Así pues, $h = [\omega_i, \omega_j]_\alpha$ maximal $\Rightarrow$ $h \in \mathbf{C}^*_\alpha$.

Si hubiésemos supuesto $\omega_i \leq \omega_j < \omega$, habríamos llegado a la misma conclusión.

■

3.2. Teoremas de Existencia y Unicidad

Teorema 3.2.1

Toda disimilaridad piramidal d sobre Ω, define una pirámide indexada en sentido amplio de $\wp(\Omega)$ que es única.

Demostración

Dada una disimilaridad piramidal d sobre Ω, definimos el siguiente subconjunto de $\wp(\Omega)$: $\mathbf{P} = \mathrm{P}^* \cup \{\text{intersecciones no vacías de dos grupos de } \mathrm{P}^*\}$

Con ello, tendremos que los conjuntos de $\mathbf{P}$ o bien son grupos de P^*, o bien son intersecciones de dos grupos de P^*, que son precisamente los que le faltan a éste para ser una pirámide.

Veamos, en primer lugar, que $\mathbf{P}$ es una pirámide.

P.1 $\Omega \in \mathbf{P}$.
En efecto, $\Omega \in \mathrm{P}^*$ puesto que será $\Omega \in \mathbf{C}^*_\alpha$, con $\alpha = \max(\mathrm{A})$.

P.2 $\forall\, \omega \in \Omega; \quad \{\omega\} \in \mathbf{P}$
$\forall\, \omega \in \Omega, \quad \{\omega\} \in \mathrm{P}^*$ puesto que $\{\omega\} \in \mathbf{C}^*_\alpha$ y $0 \in \mathrm{A}$.

P.3 $\forall\, h, h' \in \mathbf{P}, \quad h \cap h' = \emptyset$ ó $h \cap h' \in \mathbf{P}$.
Sean h, h' dos grupos de $\mathbf{P}$ tales que $h \cap h' \neq \emptyset$
Puesto que cualquier grupo de $\mathbf{P}$, o bien es de P^*, o es intersección no vacía de dos grupos de P^*, si h y h' son grupos de $\mathbf{P}$, existirán $h_i \in \mathrm{P}^*$, $i \in \{1,2,3,4\}$, tales que: $h = h_1 \cap h_2$ y $h' = h_3 \cap h_4$.
Puesto que d es piramidal, tiene sentido suponer la existencia de un preorden total sobre Ω. Consideremos, para cualquier $i \in \{1,2,3,4\}$, ω_i y ω'_i los individuos extremos de $h_i (\omega_i \leq \omega'_i)$.
Si consideramos $\quad j = \max\{\omega_i\}_{i=1,2,3,4}$
$\qquad\qquad\quad \omega'_k = \min\{\omega'_i\}_{i=1,2,3,4}$
resulta inmediato, a partir de la consideración de que el preorden sobre es total, que: $h \cap h' = h_j \cap h_k$, con $h_j, h_k \in \mathrm{P}^*$, (Ver Fig. 3.2.1).
Así pues, tendremos que $h \cap h' \in \mathbf{P}$.
$h \cap h' = (h_1 \cap h_2) \cap (h_3 \cap h_4) = h_1 \cap h_4$:

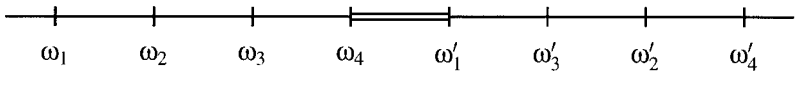


Figura 3.2.1

P.4 Existe un preorden total sobre Ω, compatible con $\mathbf{P}$.
Puesto que d es piramidal, existe un preorden total sobre Ω compatible con d, y este mismo preorden es compatible con $\mathbf{P}$, puesto que cualquier grupo de P^* es conexo respecto a este preorden por construcción, y la intersección no vacía de dos grupos conexos es otro grupo conexo (puesto que el preorden es total). Por tanto cualquier grupo de $\mathbf{P}$ será conexo respecto al preorden.
Finalmente vamos a ver que esta pirámide puede ser indexada en sentido amplio. Para ello, definamos la siguiente aplicación:

$$\begin{aligned} i : \mathbf{P} &\longrightarrow \mathbb{R}^+ \\ h &\longrightarrow i(h) = \max\{d(\omega,\omega'); \quad \omega,\omega' \in h\} \end{aligned}$$

Entonces se satisface:

I.1 $\forall\ \in \Omega, \quad i(\{\omega\}) = 0$.
Inmediato a partir de la definición.

I.2 $\forall\, h,h' \in \mathbf{P}, \quad$ si $h \subset h', \quad$ entonces $i(h) \leq i(h')$.
Si $h \subset h'$ es evidente que $\max\{d(x,y); \quad x,y \in h\} = i(h)$ es menor o igual que $\max\{d(x,y);\ x,y \in h'\} = i(h')$, con lo cual $i(h) \leq i(h')$.

I.3 $\forall\, h,h' \in \mathbf{P}$ con $h \subsetneq h'$ e $i(h) = i(h')$, existen $h_1,h_2 \in \mathbf{P}$, distintos de h, y tales que $h = h_1 \cap h_2$.
Si $h \subsetneq h'$ y son del mismo índice, entonces $h \notin \mathrm{P}^*$, puesto que h no es maximal para $\alpha = i(h)$.
Por tanto, si $h \in \mathbf{P}$ y $h \notin \mathrm{P}^*$, existirán h_1 y h_2 de $\mathrm{P}^* \subset \mathbf{P}$, distintos de h, y tales que $h = h_1 \cap h_2$.

Así pues, $(\mathbf{P},i)$ es una pirámide indexada en sentido amplio que además es única por construcción.

■

Teorema 3.2.2

Toda pirámide indexada $(\mathbf{P},i)$ de $\wp(\Omega)$, induce una única disimilaridad piramidal sobre Ω.

Demostración

Dada una pirámide indexada $(\mathbf{P},i)$, para cada par de individuos ω y ω' de Ω, definimos: $d(\omega,\omega') = i(h_0)$, siendo h_0, el mínimo grupo de $\mathbf{P}$, en el sentido de la inclusión, que contiene ω y ω'.

Esta definición es correcta puesto que si existe un único grupo de $\mathbf{P}$ que contiene ω y ω', entonces éste es el mínimo, y si existen varios, su intersección, que también

es de **P**, es el mínimo, con lo cual tiene sentido considerar el mínimo grupo de **P** que contiene ω y ω'.

Veamos pues que d es una disimilaridad piramidal:

D.1 $\forall\, \omega, \omega' \in \Omega, \quad d(\omega, \omega') \geq 0$.
Inmediato, puesto que para cada $h \in \mathbf{P}, i(h) \geq 0$.

D.2 $\forall\, \omega, \omega' \in \Omega, \quad d(\omega, \omega) = 0$.
El menor grupo de **P** que contiene ω, es $\{\omega\}$, e $i(\{\omega\}) = 0$.

D.3 $\forall\, \omega, \omega' \in \Omega, \quad d(\omega, \omega') = d(\omega', \omega)$.
Inmediato.

D.4 Existe un preorden total sobre Ω tal que, dados $\omega, \omega', \omega''$ de Ω, con $\omega \leq \omega' \leq \omega''$, $d(\omega, \omega'') \geq \max\{d(\omega, \omega'), d(\omega', \omega'')\}$.
Puesto que **P** es una pirámide, existe en Ω un preorden total compatible con **P**. Sean ω, ω' y ω'' elementos de Ω, con $\omega \leq \omega' \leq \omega''$.

Sea $h_{\omega\omega''}$ el mínimo grupo de **P** que contiene ω y ω'', puesto que $h_{\omega\omega''}$ es conexo respecto al preorden que **P** induce sobre Ω, tendremos que $\omega' \in h_{\omega\omega''}$.

Si $h_{\omega\omega'}$ es el mínimo grupo de **P** que contiene ω y ω', tendremos que tanto $h_{\omega\omega''}$ como $h_{\omega\omega'}$ contienen ω y ω'. Si son grupos conexos y $h_{\omega\omega'}$ es el mínimo, tendremos que $h_{\omega\omega'} \subset h_{\omega\omega''}$, por tanto $i(h_{\omega\omega'}) \leq i(h_{\omega\omega''})$ y $d(\omega, \omega') \leq d(\omega, \omega'')$.

Por otra parte, $h_{\omega\omega''}$ y $h_{\omega'\omega''}$, ambos contienen a ω' y ω'', por tanto, $h_{\omega'\omega''} \subset h_{\omega\omega''}$, así pues $i(h_{\omega'\omega''}) \leq i(h_{\omega\omega''})$ y $d(\omega', \omega'') \leq d(\omega, \omega'')$.
De ambas desigualdades, obtenemos: $d(\omega, \omega'') \geq \max\{d(\omega, \omega'), d(\omega', \omega'')\}$, que es lo que queríamos demostrar.

■

Puesto que, dados dos individuos cualesquiera ω y ω' de Ω, el mínimo grupo de **P** que los contiene es único, la disimilaridad d definida como el índice de dicho grupo, será también única.

Finalmente, cabe resaltar que el teorema anterior es válido, en particular, para pirámides indexadas en sentido amplio. Así pues, los teoremas 3.2.1 y 3.2.2 establecen una correspondencia biunívoca entre el conjunto de las pirámides indexadas en sentido amplio y el de las disimilaridades piramidales.

3.3. Otras Relaciones

Proposición 3.3.1

Sea $(\mathbf{P}, i)$ *una pirámide indexada de* $\wp(\Omega)$ *y sean* ω *y* ω' *dos individuos cualesquiera de* Ω. *El mínimo grupo de* **P** *que contiene a* ω *y* ω', *es también el de menor índice de entre todos los que los contienen.*

Proposición 3.3.2

Sea h un grupo cualquiera de una pirámide **P**. *Si* ω *y* ω' *son los individuos extremos de h, entonces h es el mínimo grupo de* **P** *que los contiene.*

Las demostraciones de las proposiciones 3.3.1 y 3.3.2 son inmediatas a partir de la definición de pirámide indexada y del hecho que cada grupo de la pirámide sea conexo respecto un cierto preorden total definido sobre la población Ω.

Proposición 3.3.3

Sea $(\mathbf{P}, i)$ *una pirámide indexada en sentido amplio sobre* Ω, *y d la disimilaridad piramidal inducida. Sea* $A = \{\alpha \in \mathbb{R} / \exists\, \omega_i, \omega_j \in \Omega \quad y \quad d(\omega_i, \omega_j) = \alpha\}$ *y sea* P' *la familia de* $\wp(\Omega)$ *definida a partir de la disimilaridad d en la sección 3.1. En estas condiciones:*

$$h \in \mathbf{P} \quad \Leftrightarrow \quad \begin{cases} h \in P' \\ o \\ \exists\, h_1, h_2 \in P', \quad \textit{distintos de h, tales que } h = h_1 \cap h_2 \end{cases}$$

Demostración

Sea h un grupo cualquiera de **P**. Puesto que **P** es una pirámide, tiene sentido considerar los individuos extremos de h respecto al preorden total correspondiente definido sobre Ω, sean éstos ω_i y ω_j y supongamos que $d(\omega_i, \omega_j) = \alpha_0$. En estas condiciones, la proposición 3.3.2 nos asegura que h es el mínimo grupo de **P** que contiene ω_i y ω_j.

Supongamos que $h \notin P'$, entonces $\forall\, \alpha \in A, h \notin P_\alpha$, en particular, $h \notin P_{\alpha 0}$, y por tanto, existirá $\omega_0 \in \Omega$, tal que:

$$(1) \qquad \omega_0 \notin h \quad y \quad \forall\, x \in h, \quad d(\omega_0, x) \leq \alpha_0 = d(\omega_i, \omega_j)$$

Si ω_i, ω_j son los individuos más alejados de h, y $\omega_0 \notin h$, entonces no puede ser $\omega_i \leq \omega_0 \leq \omega_j$, y por tanto tendrá que ser: $\omega_0 < \omega_i \leq \omega_j$ o bien $\omega_i \leq \omega_j < 0$.

Supongamos $\omega_0 < \omega_i \leq \omega_j$. Puesto que d es piramidal, tendremos: $d(\omega_0, \omega_j) \geq d(\omega_i, \omega_j)$.

Por otra parte, para $x = \omega_j$ la desigualdad (1) quedará: $d(\omega_0, \omega_j) \leq d(\omega_i, \omega_j)$.

Y combinando ambas desigualdades, tendremos:

$$(2) \qquad d(\omega_0, \omega_j) = d(\omega_i, \omega_j)$$

Consideremos, además, h_0, el mínimo grupo de **P** que contiene ω_0 y ω_j.

Puesto que la disimilaridad piramidal d, viene inducida por la pirámide **P**, resultará que: $d(\omega_0,\omega_j) = i(h_0)$ y $d(\omega_i,\omega_j) = i(h)$.

De lo cual, y teniendo en cuenta (2), se deduce que $i(h) = i(h_0)$.

Veamos que $h \subsetneq h_0$:

Puesto que $\omega_0 < \omega_i \leq \omega_j$ y h_0 es conexo, ω_i también ha de pertenecer a h_0.

Así pues, h es el mínimo grupo que contiene ω_i y ω_j y h_0 también los contiene, por tanto $h \subset h_0$.

Además $\omega_0 \in h_0$ y $\omega_0 \notin h$, con lo cual, $h \subsetneq h_0$.

Resumiendo pues, tendremos que: dado $h \in \mathbf{P}$, si $h \notin P'$, existirá $h_0 \in \mathbf{P}$ tal que $h \subsetneq h_0$ e $i(h) = i(h_0)$, y puesto que (**P**, i) es una pirámide indexada en sentido amplio, existirán $h_1, h_2 \in \mathbf{P}$, distintos de h, tales que $h = h_1 \cap h_2$.

Ahora bien, nuestro objetivo era demostrar que si $h \notin P'$, entonces es intersección de dos grupos de P', y no de **P**.

Supongamos pues que por lo menos uno de los dos grupos, h_1 por ejemplo, no es de P'. Entonces, existirán $h_1', h_1'' \in \mathbf{P}$, distintos de h_1 y tales que $h_1 = h_1' \cap h_1''$, con lo cual $h = h_1' \cap h_1'' \cap h_2$.

Si, a su vez, alguno de estos grupos no es de P', h_1' por ejemplo, existirán $\overline{h_1'}$ y $\hat{h_1'}$ de **P**, distintos de h_1', con $h_1' = \overline{h_1'} \cap \hat{h_1'}$, con lo cual $h = \overline{h_1'} \cap \hat{h_1'} \cap h_1'' \cap h_2$, y así sucesivamente.

En este proceso, obsérvese que $h \subsetneq h_1 \subsetneq h_1' \subsetneq \overline{h_1'} \subsetneq \cdots$ Puesto que las inclusiones son estrictas y $\wp(\Omega)$ es finito, o bien llegaremos a un grupo de P', o bien a un grupo $\tilde{h_1}$ tal que $h \subsetneq h_1 \subsetneq \cdots \tilde{h_1}$ y no podrá desdoblarse como intersección de dos grupos de **P**, entonces tendrá que ser $\tilde{h_1} \in P'$, puesto que de lo contrario, sería intersección de dos grupos estrictamente más grandes con lo cual $\tilde{h_1}$ no sería el último grupo de la cadena como estamos suponiendo.

Así pues, aplicando este proceso a los grupos que no sean de P', tendremos que h puede expresarse como $h = \tilde{h_1} \cap \tilde{h_2} \cap \cdots \cap \tilde{h_r}$, con $\tilde{h_i} \in P'$ y $\tilde{h_1} \neq h$, para todo $i \in \{1,2,\ldots,r\}$.

A partir de aquí, por ser el preorden sobre Ω total, es inmediato demostrar que $h = \tilde{h_k} \cap \tilde{h_j}$, con $k, j \in \{1,\ldots,r\}$, y $\tilde{h_k}, \tilde{h_j} \in P'$ que es lo que queríamos demostrar.

■

Sea $h \in P'$, entonces $\exists\, \alpha \in A / h \in P_\alpha$ y por tanto será $h = \{$individuos de Ω con interdistancias $\leq \alpha\}$.

Si $\alpha \in A \quad \Rightarrow \quad \exists\, \omega_i, \omega_j \in \Omega$ tales que $d(\omega_i, \omega_j) = \alpha$ y supongamos que ω_i y ω_j son los individuos extremos de h, en términos del preorden total sobre Ω asociado a d, puesto que si no fuese así, es decir si existiese un $\overline{\omega} \in h$ tal que $\overline{\omega} < \omega_i \leq \omega_j$ (ó bien $\omega_i \leq \omega_j < \overline{\omega}$) tendríamos que, por ser d piramidal, $d(\overline{\omega}, \omega_j) \geq d(\omega_i, \omega_j)$; y por ser $\overline{\omega}, \omega_j \in h$, $d(\overline{\omega}, \omega_j) \leq d(\omega_i, \omega_j)$, con lo cual $d(\overline{\omega}, \omega_j) = d(\omega_i, \omega_j)$ y por tanto, en la definición de h podríamos sustituir ω_i por $\overline{\omega}$.

Supongamos pues, ω_i, ω_j los extremos de h y sea h_0 el mínimo grupo de $\mathbf{P}$ que los contiene.

Entonces, $\forall\, \omega \in h$ será $\omega_i \leq \omega \leq \omega_j$ y puesto que h_0 es conexo, $\omega \in h_0$, con lo cual, $h \subset h_0$.

Si d es la disimilaridad piramidal inducida por la pirámide $(\mathbf{P}, i)$, la inclusión anterior no podrá ser estricta, por tanto $h = h_0$ y en consecuencia, $h \in \mathbf{P}$.

Finalmente, si $h = h_1 \cap h_2$, con $h_i \in P'$ y $h_i \neq h$, para $i \in \{1,2\}$, puesto que acabamos de demostrar que $P' \subset \mathbf{P}$, resultará que h será intersección de dos grupos de $\mathbf{P}$, por tanto será $h \in \mathbf{P}$.

Teorema 3.3.1

La pirámide indexada en sentido amplio inducida por una disimilaridad piramidal, induce esa misma disimilaridad piramidal.

Demostración

Para ello, vamos a ver que si una pirámide indexada en sentido amplio, $(\mathbf{P}, i)$ induce una disimilaridad piramidal d', y a su vez es la pirámide inducida por otra disimilaridad piramidal d, entonces $d = d'$.

Si d' es la disimilaridad piramidal inducida por $(\mathbf{P}, i)$, dados dos individuos cualesquiera $x, y \in \Omega$, $\quad d'(x,y) = i(h_{xy})$, siendo h_{xy} el mínimo grupo de $\mathbf{P}$ que contiene a x e y.

Si $(\mathbf{P}, i)$ es la pirámide inducida por d, entonces $i(h_{xy}) = \max\{d(\omega, \omega'); \omega, \omega' \in h_{xy}\}$, por consiguiente será $i(h_{xy}) \geq d(x,y)$.

De donde, $d'(x,y) \geq d(x,y)$.

Vamos a ver que esta desigualdad no puede ser estricta.

Supongamos que existan $x_{0,y}0 \in \Omega$ tales que $d'(x_0,y_0) > d(x_0,y_0)$. Si $d(x_0,y_0) = \alpha$ podemos considerar la familia de $\wp(\Omega)$, $\mathbf{C}^*_\alpha$ tal como la hemos definido en la sección 3.1. Sean $h^1_\alpha,\ldots,h^r_\alpha$ los grupos de dicha familia, que a su vez serán todos ellos grupos de $\mathbf{P}$, por ser ésta la pirámide inducida por d.

Por otra parte, es claro que existirá algún h^i_α, con $i \in \{1,\ldots,r\}$ tal que $x_0,y_0 \in h^i_\alpha$.

Así pues, si $(\mathbf{P},i)$ es la pirámide inducida por d, $\quad i(h^i_\alpha) = \alpha < d'(x_0,y_0) = i(h_0)$. Ahora bien, si h_0 es el mínimo grupo de $\mathbf{P}$ que contiene x_0 e y_0, en virtud de la proposición 3.3.1 es también el grupo de menor índice de entre todos los que contienen x_0 e y_0, por tanto no puede ser que $i(h^i_\alpha) < i(h_0)$. Por tanto no puede existir ningún par de individuos, $x_0,y_0 \in \Omega$ tales que $d'(x_0,y_0) > d(x_0,y_0)$, con lo cual tendrá que ser $d'(x,y) = d(x,y)$ para cualquier par, x e y, de individuos de Ω, y por tanto $d = d'$.

■

Teorema 3.3.2

La disimilaridad piramidal inducida por una pirámide indexada en sentido amplio, induce esa misma pirámide.

Demostración

Para ello, veremos que si la pirámide $(\mathbf{P},i)$ induce la disimilaridad piramidal d, y a su vez, ésta induce la pirámide $(\tilde{\mathbf{P}},\tilde{i})$, entonces $\mathbf{P} = \tilde{\mathbf{P}}$ y $i = \tilde{i}$.

Si $(\tilde{\mathbf{P}},\tilde{i})$ es la pirámide indexada inducida por d, por la demostración del teorema 3.2.1 sabemos que:

$$h \in \tilde{\mathbf{P}} \quad \Leftrightarrow \quad \begin{cases} h \in \tilde{\mathrm{P}}^* \\ \text{o} \\ \exists\, h_1,h_2 \in \tilde{\mathrm{P}}^*, \quad \text{distintos de } h, \text{ tales que } h = h_1 \cap h_2 \end{cases}$$

Por otra parte, puesto que gracias a la proposición 3.1.1, $\tilde{\mathrm{P}}^* = \mathrm{P}'$, resultará:

$$h \in \tilde{\mathbf{P}} \quad \Leftrightarrow \quad \begin{cases} h \in \mathrm{P}' \\ \text{o} \\ \exists\, h_1,h_2 \in \mathrm{P}', \quad \text{distintos de } h, \text{ tales que } h = h_1 \cap h_2 \end{cases}$$

Y por la proposición 3.3.3 esto equivale a que $h \in \mathbf{P}$.

Para ver que $i = \tilde{i}$, tendremos en cuenta que $\mathbf{P} = \tilde{\mathbf{P}}$ y demostraremos que $\forall\, h \in \mathbf{P}$, $i(h) = \tilde{i}(h)$.

Si $(\tilde{\mathbf{P}},\tilde{i})$ es la pirámide indexada inducida por d, para cualquier $h \in \mathbf{P} = \tilde{\mathbf{P}}$, $\tilde{i}(h) = \max\{d(\omega,\omega'),\ \forall\, \omega,\omega' \in h\}$, por tanto existirán $x,y \in h$ tales que $\tilde{i}(h) = d(x,y)$.

Por otra parte, sean ω_i y ω_j los individuos extremos de h. Puesto que $h \in \mathbf{P}$ y $\mathbf{P}$ es la pirámide que induce d, si ω_i y ω_j son los individuos más alejados de h, éste será el mínimo grupo de $\mathbf{P}$ que los contendrá (Proposición 3.3.2), y por tanto $i(h) = d(\omega_i, \omega_j)$.

Así pues, hasta el momento tenemos que: $x, y \in h \in \mathbf{P}$; ω_i, ω_j son los extremos de h; y d es la disimilaridad piramidal inducida por $(\mathbf{P}, i)$; con lo cual $d(x,y) \leq d(\omega_i, \omega_j)$.

Finalmente, puesto que $d(x,y) = \max\{d(\omega, \omega'), \quad \forall\, \omega, \omega' \in h\}$, tendremos que $d(x,y) \geq d(\omega_i, \omega_j)$.

De ambas desigualdades resultará: $d(x,y) = d(\omega_i, \omega_j)$ y por lo tanto $\forall\, h \in \mathbf{P}$, $i(h) = \tilde{i}(h)$.

■

4. GRUPOS SOBRANTES. DEPURACIÓN DE UNA PIRÁMIDE

Dada una pirámide indexada $(\mathbf{P}, i)$ construida por alguno de los algoritmos de clasificación piramidal existentes, Diday (1986), Bertrand (1986), Capdevila (1993), podemos preguntarnos si existen grupos "sobrantes", es decir, grupos que proporcionen información redundante sobre la clasificación y por tanto que puedan ser eliminados de la pirámide sin que varíe la clasificación piramidal.

Para responder a la pregunta con el máximo rigor, será necesario, en primer lugar, precisar el concepto de "*grupo sobrante*" en una pirámide indexada.

Si $(\mathbf{P}, i)$ es una pirámide indexada en sentido amplio, es evidente que ningún grupo puede ser sobrante, puesto que a dicha pirámide le corresponde biunívocamente una disimilaridad piramidal d, por tanto si eliminamos algún grupo h de $\mathbf{P}$, a pesar de que $\mathbf{P} - \{h\}$ continuase siendo una pirámide indexada en sentido amplio, induciría otra disimilaridad piramidal d' distinta a d y por tanto una clasificación también distinta. Si por el contrario, la pirámide es indexada en sentido estricto, también lo será en sentido amplio (Proposición 2.3) y por consiguiente tampoco habrá grupos susceptibles de ser eliminados.

Sea pues $(\mathbf{P}, i)$ una pirámide indexada pero no en sentido amplio. En este caso, existirán grupos $h, h' \in \mathbf{P}$; $h \subsetneq h'; i(h) = i(h')$ y h no será intersección no trivial de grupos de $\mathbf{P}$.

Consideremos el conjunto: $\Sigma_{\mathrm{P}} = \{h \in \mathbf{P} / \exists\, h' \in \mathbf{P}; h \subsetneq h'; \quad i(h) = i(h')$ y h no es intersección no trivial de grupos de $\mathbf{P}\}$

que será no vacío, por ser la pirámide indexada no en sentido amplio.

En estas condiciones, es evidente que $\mathbf{P}-\mathbf{\Sigma}_P$ será también una pirámide indexada y además, en sentido amplio.

Si finalmente demostramos que la disimilaridad piramidal inducida por $(\mathbf{P},i)$ es la misma que la inducida por $(\mathbf{P}-\mathbf{\Sigma}_P,i)$, quedará demostrado que los grupos de $\mathbf{\Sigma}_P$ son realmente los grupos sobrantes de la pirámide $\mathbf{P}$, puesto que $\mathbf{P}$ y $\mathbf{P}-\mathbf{\Sigma}_P$ serán pirámides equivalentes, en el sentido de tener asociada la misma disimilaridad piramidal, y al ser $(\mathbf{P}-\mathbf{\Sigma}_P,i)$ indexada en sentido amplio, no tendrá ningún grupo sobrante, con lo cual podremos asegurar que los grupos de $\mathbf{\Sigma}_P$ serán los únicos grupos sobrantes de $\mathbf{P}$.

Teorema 4.1

La pirámide indexada sobre $\Omega(\mathbf{P},i)$, induce la misma disimilaridad piramidal que la pirámide $(\mathbf{P}-\mathbf{\Sigma}_P,i)$.

Demostración

Sabemos que cualquier pirámide indexada de $\wp(\Omega)$, sea en sentido amplio o no, induce una disimilaridad piramidal sobre Ω, sin más que definir la distancia entre dos individuos como el índice del menor grupo que los contiene.

Sea pues Ω un conjunto finito y $(\mathbf{P},i)$ una pirámide indexada. Si lo es en sentido amplio, entonces $\mathbf{\Sigma}_P=\emptyset$ y no hay nada que demostrar.

Supongamos pues que $(\mathbf{P},i)$ es indexada pero no en sentido amplio, con lo cual $\mathbf{\Sigma}_P\neq\emptyset$ y $\mathbf{P}-\mathbf{\Sigma}_P\subsetneq\mathbf{P}$.

Sean d y d' las disimilaridades piramidales inducidas respectivamente por $(\mathbf{P},i)$ y $(\mathbf{P}-\mathbf{\Sigma}_P,i)$.

Sean x e y dos individuos cualesquiera de Ω, y sea h_0 el mínimo grupo de $\mathbf{P}$ que los contiene. Entonces será $d(x,y)=i(h_0)$.

Si $h_0\notin\mathbf{\Sigma}_P$, entonces $h_0\in\mathbf{P}-\mathbf{\Sigma}_P$, y por tanto $d'(x,y)=i(h_0)$, con lo cual tendremos $d(x,y)=d'(x,y)$, que es lo que queríamos ver.

Si por el contrario, $h_0\in\mathbf{\Sigma}_P, h_0\notin\mathbf{P}-\mathbf{\Sigma}_P$, y además existirán otros grupos de $\mathbf{P}$ que contendrán estrictamente h_0 y de su mismo índice. Estos otros grupos lo serán, naturalmente, en número finito, y no todos ellos podrán ser de $\mathbf{\Sigma}_P$ (por lo menos $\Omega\notin\mathbf{\Sigma}_P$).

Sea pues, h_0' el menor de los grupos de $\mathbf{P}$ tales que $h_0\subsetneq h_0'$, $i(h_0)=i(h_0')$ y $h_0'\notin\mathbf{\Sigma}_P$.

En estas condiciones pues, h_0' será el menor grupo de $\mathbf{P}-\mathbf{\Sigma}_P$ que contendrá h_0 y por tanto el menor grupo de $\mathbf{P}-\mathbf{\Sigma}_P$ que contendrá x e y. Así pues, $d'(x,y)=i(h_0')=i(h_0)$ y por tanto $d(x,y)=d'(x,y)$ que es lo que queríamos demostrar.

Resumiendo pues, la caracterización de los grupos sobrantes de una pirámide indexada será la siguiente: $h \in (\mathbf{P}, i)$ es un *grupo sobrante* sii $h \in \Sigma_P$. Es decir, h es un *grupo sobrante*, si no es intersección no trivial de grupos de $\mathbf{P}$ y existe otro grupo $h' \in (\mathbf{P}, i)$ tal que: $h \subsetneq h'$ e $i(h) = i(h')$.

La Depuración de una Pirámide consistirá en eliminar los grupos sobrantes. Una vez eliminados dichos grupos, desde el punto de vista de la representación visual, es posible que existan también *aristas sobrantes*, que deberán ser también eliminadas. Para ello, el criterio a seguir será el siguiente: Después de la depuración de una pirámide indexada, cualquier arista que no una un grupo con su predecesor es sobrante y por tanto debe ser también eliminada.

■

REFERENCIAS

[1] **Arcas, A.** y **Cuadras, C.M.** (1987). "Métodos Geométricos de Representación Mediante Modelos en Arbol". *Publicaciones de Bioestadística y Biomatemática.* Universitat de Barcelona.

[2] **Benzecri, J.P.** (1973). *L'Analyse des Données: La Taxonomie. Tome 1.* Dunod. París.

[3] **Bertrand, P.** (1986). "Etude de la Répresentation Pyramidale". *Thèse de Doctorat.* Université de Paris-Dauphine.

[4] **Bertrand, P.** y **Diday, E.** (1985). "A Visual Représentation of the Compatibility Between an Order and a Dissimilarity Index: The Pyramids". *Computational Statistics Quarterly*, **2, Issue 1**, 31–41.

[5] **Capdevila, C.** (1993). "Contribuciones al Estudio del Problema de la Clasificación Mediante Grafos Piramidales". *Tesis de Doctorado en Matemáticas.* Universitat de Barcelona.

[6] **Capdevila, C.** y **Arcas, A.** (1992). "Sobre un Algoritmo de Clasificación Mediante Grafos Piramidales". *Actas de la XX Reunion Nacional de Estadística e Investigación Operativa.* Cáceres 1992.

[7] **Capdevila, C.** y **Arcas, A.** (1993). "Some Aspects About Statistical Inference in Pyramidal Trees". *Comunication n° 41 a la 4th Conference of the International Federation of Classification Societies.* Paris 1993.

[8] **Diday, E.** (1986). "Une Représentation Visuelle des Classes Empietantes: Les Pyramides". *Rairo: Analyse des Données*, **52**, 475–526.

[9] **Diday, E.** (1986). "Ordres and Overlapping Clusters in Pyramides". *Multidimensional Data Analysis.* DSWO Press, Leiden, 201–234.

[10] **Durand, C.** (1986). "Sur la Représentation Pyramidale en Analyse des Données". *Mémoire de DEA en Mathématiques Appliquées.* Université de Provence, Marseille.

[11] **Durand, C.** (1988). "Une Approximation de Robinson Inférieur Maximale". *Rapport de Recherche*, 88–02, Laboratoire de Mathématiques Appliquées et Informatique, Université de Provence, Marseille.

[12] **Durand, C.** y **Fichet, B.** (1988). "One-to-one Correspondences in Pyramidal Representation: A Unified Approach". *Clasification and Related Methods of Data Analysis*. H.H. Bock, Ed. North-Holland. Amsterdam.

[13] **Durand, C.** (1989). "Ordres et Graphes Pseudo-Hierarchiques: Theorie et Optimisation Algoritmique". *Thèse de doctorat en Mathématiques Appliquées*. Universite de Provence. Marseille.

[14] **Fichet, B.** (1984). "Sur une Extension de la Notion de Hiérarchie et son Equivalence Avec Certaines Matrices de Robinson". *Comunication of the Journées de Statistique de Montpellier*.

[15] **Jardine, N.** y **Sibson, R.** (1971). *Mathematical Taxonomy*. Wiley, New York.

[16] **Lerman, I.C.** (1981). *Classification et Analyse Ordinale des Données*. Dunod. Paris.

ENGLISH SUMMARY:

INDEXED PYRAMIDS AND PYRAMIDAL DISSIMILARITIES

Carles Capdevila Marquès and Antoni Arcas Pons

Pyramidal trees, introduced by E. Diday, are a logical generalization of ultrametric trees. They are less restrictive structures where recovering replaces the concept of partition, obtaining a representation which bears more information and is closer to the initial dissimilarities.

The pyramidal representation processes for individuals of a finite population, consists in transforming the initial dissimilarity into a pyramidal dissimilarity by means of some known pyramidal clustering procedure algorithm; this pyramidal dissimilarity is equivalent to an indexed pyramid.

In this study we establish a rigorous mathematic formalization of the equivalence between the concepts of indexed pyramid and pyramidal dissimilarity. In addition, we study the problem of the redundant information contained in a pyramidal structure, and

we provide objective criteria to eliminate it without modificating the corresponding pyramid.

In section 2 you find the basic concepts for the study of the pyramidal models of classification and representation of data, and some elementary relationships among them are established.

We realize that all indexed total hierarchies are also indexed pyramids, and therefore, the pyramidal models constitute a natural generalization of the hierarchical models.

Bearing in mind that the mathematic bases of the hierarchical representations are based on the equivalence between the ultrametric dissimilarities and the indexed hierarchies of parts of a set, a similar equivalence is established between the pyramidal dissimilarities and indexed pyramids of parts of a set.

In section 3.2 it is proved that every pyramidal dissimilarity d defined in Ω, produces an unique indexed pyramid in broad terms of $\wp(\Omega)$, and conversely that all indexed pyramid $(\mathbf{P}, i)$ of $\wp(\Omega)$ induces an unique pyramidal dissimilarity in Ω.

In section 3.3 it is proved that, the indexed pyramid in broad terms of $\wp(\Omega)$ induced by a pyramidal dissimilarity induces the same pyramidal dissimilarity. Conversely, the pyramidal dissimilarity induced by a indexed pyramid in broad terms of $\wp(\Omega)$ induces the same indexed pyramid.

Finally, it is proved that in an indexed pyramid the spare groups can be characterized by: $h \in (\mathbf{P}, i)$ is a *spare group*, if h is a nontrivial intersection of groups of $\mathbf{P}$, and if there exists $h' \in (\mathbf{P}, i)$ such that: $h \subsetneq h'$ and $i(h) = i(h')$.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 153-169, 1995

DETERMINACIÓN DE LOS VALORES DE LOS FACTORES DE DISEÑO PARA LA OBTENCIÓN DE PRODUCTOS ROBUSTOS. APLICACIÓN AL CASO DE UN CIRCUITO CON BOBINA Y RESISTENCIA

ALBERT PRAT y PERE GRIMA*

Universitat Politècnica de Catalunya

Frente a la visión antigua de controlar la calidad mediante la inspección final, o la clásica del control estadístico del proceso, la forma que actualmente podríamos denominar moderna, y desde luego la más económica de asegurar la calidad, consiste en diseñar productos tales que sus prestaciones se mantengan en el nivel deseado aunque éstos no se hayan fabricado en las condiciones más adecuadas, o se hayan elaborado con materias primas que no reúnan las condiciones óptimas, o bien que sean utilizados en condiciones distintas a las previstas.

A estos productos se les denomina robustos, y en este artículo se presenta, a través del desarrollo de un caso concreto, una nueva metodología para la selección de los valores de sus parámetros.

Determination of Design Factors Values to Obtain Robust Products. Aplication to Circuits with Inductance and Resistance.

Key words: Producto Robusto, Método de Taguchi, Diseño Factorial, Varianza Mínima, Gráfico Distancia-Varianza.

* Albert Prat. y Pere Grima. E.T.S. Enginyers Industrials de Barcelona Dpt. Estadística i Investigació Operativa Universitat Politècnica de Catalunya.

Dirección para correspondencia: Albert Prat Bartés. Dpt. Estadística i Inv. Operativa. E.T.S. Enginyers Industrials de Barcelona. Av. Diagonal, 647. 08028 Barcelona.

- Article rebut el maig de 1994.

- Acceptat el febrer de 1995.

1. INTRODUCCIÓN

Las técnicas de diseño de productos robustos fueron introducidas en Occidente por el ingeniero japonés Genichi Taguchi (1984, 1986), que de esta forma dio nombre al método más conocido y utilizado. Uno de los aspectos del método de Taguchi es el que hace referencia a la determinación de los valores de los factores de diseño (parámetros), el cual consiste, en esencia, en el planteamiento de un plan de experimentación en el que se mide el valor de la característica de calidad (respuesta) en una cierta combinación de valores de los **factores de diseño** (contemplando la variabilidad de que estén afectados), y también de los llamados **factores de ruido** (ajenos al producto, pero que influyen en la respuesta y no pueden mantenerse fijos). A partir de los resultados obtenidos en la experimentación se determina cuales son los valores de los factores de diseño que consiguen un valor satisfactorio para la respuesta independientemente de los valores que tomen las fuentes de variabilidad que intervengan.

El método para elegir el valor de los factores de diseño que propone Taguchi se caracteriza por el uso de las llamadas **matrices producto** en el diseño del plan de experimentación y en la utilización de unos estadísticos un tanto "sui generis", denominados **ratios señal-ruido** en el análisis de los resultados de la experimentación.

A pesar del general reconocimiento sobre las aportaciones de Taguchi en torno al diseño de productos robustos, la metodología que utiliza presenta algunos aspectos controvertidos. Así, aunque en algunos casos el número de experimentos que se plantea es justamente el necesario para obtener la información precisa, en otros muchos casos, el experimentar en cada una de las condiciones de la matriz producto, exige un esfuerzo y una dedicación de recursos que podría disminuirse (Shoemaker *et al.*, 1991).

Por otra parte, la peculiaridad de los estadísticos que utiliza convierte su método en el seguimiento de unos procedimientos poco intuitivos que no facilitan el entender el "qué" se hace y "por qué" se hace. Además, se ha demostrado (Box *et al.*, 1986) que las técnicas estadísticas utilizadas para el análisis de los resultados obtenidos no son las más adecuadas.

A continuación se presenta una nueva metodología para el diseño de parámetros a través del ejemplo de un circuito con bobina y resistencia.

Tras el planteamiento del problema, se plantea su resolución utilizando la relación funcional entre la respuesta y los factores de diseño (en este caso es una fórmula conocida, aunque en general no se conoce). A continuación se aborda el problema a través de un modelo experimental y, finalmente, se presenta el gráfico Distancia-Variancia, que representa una alternativa al Método de Taguchi cuando se trata de

obtener una solución de compromiso entre la variabilidad de la respuesta y su distancia al óptimo.

2. PLANTEAMIENTO DEL PROBLEMA

Supongamos que deseamos montar circuitos del tipo:

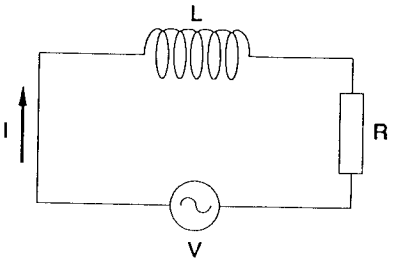


Figura 1.
Circuito con bobina y resistencia.

de forma que la intensidad máxima que circule por los mismos sea de 10 Amperios.

El valor de la intensidad máxima que circula por un circuito de este tipo viene dado por la fórmula:

$$I = \frac{V}{\sqrt{R^2 + (2\pi 50L)^2}} \tag{1}$$

Supongamos también que para el montaje de los circuitos disponemos de:

Fuentes de alimentación: Proporcionan un voltaje exacto igual a 100 voltios.

Resistencias: Existe una cierta variabilidad de unas unidades a otras (no en una misma unidad a los largo del tiempo). Dicha variabilidad sigue una distribución normal centrada en su valor nominal (R_n) y con una desviación tipo proporcional a dicho valor nominal.

Bobinas: Al igual que ocurre con las resistencias, disponemos de bobinas que presentan una cierta variabilidad de unos componentes a otros. Dicha variabilidad también sigue una distribución normal centrada en su valor nominal (L_n) y con desviación tipo proporcional a dicho valor.

El problema que nos planteamos resolver es: ¿Cuáles son los valores óptimos de R_n y L_n de forma que la intensidad máxima que circule por el circuito sea de 10 amperios, con la mínima variabilidad en torno a este valor?.

3. RESOLUCIÓN UTILIZANDO LA RELACIÓN FUNCIONAL CONOCIDA

3.1. Esperanza Matemática y Varianza de *I*

Al ser R y L variables aleatorias, la intensidad que circule por cada uno de los circuitos no será siempre la misma, sino que deberemos considerarla como una nueva variables aleatoria con una determinada distribución de probabilidad. De dicha distribución nos interesa conocer su esperanza matemática $\mathrm{E}(I)$ y su varianza $\mathrm{Var}(I)$.

La expresión de la esperanza matemática de I, dado que R y L son independientes, tiene la forma:

$$\mathrm{E}(I) = \frac{V}{\sqrt{R^2 + (2\pi 50L)^2}} f(R)f(L)\ \mathrm{d}R\ \mathrm{d}L$$

en la que $f(R)$ y $f(L)$ representan las funciones densidad de probabilidad de R y L respectivamente.

La integral planteada es —en principio— inabordable, y por ello recurriremos a la descomposición de la expresión de I en serie de Taylor en torno al punto (R_n, L_n), obteniendo:

$$\begin{aligned} I = I(R_nL_n) \ &+ \left.\frac{\partial I}{\partial R}\right]_{R_n\,L_n} (R-R_n) + \left.\frac{\partial I}{\partial L}\right]_{R_n\,L_n} (L-L_n) + \\ &+ \left.\frac{1}{2}\frac{\partial^2 I}{\partial R^2}\right]_{R_n\,L_n} (R-R_n)^2 + \left.\frac{1}{2}\frac{\partial^2 I}{\partial L^2}\right]_{R_n\,L_n} (L-L_n)^2 + \\ &+ \left.\frac{\partial^2 I}{\partial R \partial L}\right]_{R_n\,L_n} (R-R_n)(L-L_n) + \text{ Residuo} \end{aligned}$$

Por tanto:

$$\begin{aligned} \mathrm{E}(I) = I(R_nL_n) \ &+ \left.\frac{\partial I}{\partial R}\right]_{R_n\,L_n} \mathrm{E}(R-R_n) + \left.\frac{\partial I}{\partial L}\right]_{R_n\,L_n} \mathrm{E}(L-L_n) + \\ &+ \left.\frac{1}{2}\frac{\partial^2 I}{\partial R^2}\right]_{R_n\,L_n} \mathrm{E}(R-R_n)^2 + \left.\frac{1}{2}\frac{\partial^2 I}{\partial L^2}\right]_{R_n\,L_n} \mathrm{E}(L-L_n)^2 + \\ &+ \left.\frac{\partial^2 I}{\partial R \partial L}\right]_{R_n\,L_n} \mathrm{E}[(R-R_n)(L-L_n)] + \mathrm{E}(\text{Residuo}) \end{aligned}$$

Y tenemos que:

$$\begin{aligned} \mathrm{E}(R-R_n) &= \mathrm{E}(R)-\mathrm{E}(R_n)=0 \\ \mathrm{E}(L-L_n) &= \mathrm{E}(L)-\mathrm{E}(L_n)=0 \\ \mathrm{E}[(R-R_n)(L-L_n)] &= \mathrm{Cov}(R,L)=0 \\ \mathrm{E}(R-R_n)^2 &= \sigma_R^2 \\ \mathrm{E}(L-L_n)^2 &= \sigma_L^2 \end{aligned}$$

Así pues, podemos escribir:

$$\mathrm{E}(I) = I(R_n,L_n) + \frac{1}{2}\frac{\partial^2 I}{\partial R^2}\Big]_{R_n\,L_n}\sigma_R^2 + \frac{1}{2}\frac{\partial^2 I}{\partial L^2}\Big]_{R_n\,L_n}\sigma_L^2 + \mathrm{E(Residuo)}$$

A partir de este punto no se considera E(Residuo). Dicho valor es prácticamente despreciable en el rango de varianzas que se estudia. Nótese que el término de derivadas de tercer orden es nulo por ser $\mathrm{E}(R-R_n)^3 = \mathrm{E}(L-L_n)^3 = 0$ (el coeficiente de asimetría en la distribución normal es cero), así como las esperanzas de los otros términos de tercer orden (como $\mathrm{E}(R-R_n)^2(L-L_n)$). El término correspondiente a las derivadas de cuarto orden puede calcularse considerando que $\mathrm{E}(R-R_n)^4 = 3\sigma_R^4$ (análogamente para L). Sustituyendo los términos de derivadas por sus correspondientes expresiones se llega a:

$$\mathrm{E}(I) = 100\left[A^{-0.5} - \frac{1}{2}\left(A^{-1.5} - 3R_n^2A^{-2.5}\right)\sigma_R^2 - \frac{1}{2}\left(KA^{-1.5} - 3K^2L_n^2A^{-2.5}\right)\sigma_L^2\right]$$

Siendo:

$$\begin{aligned} A &= R_n^2 + (2\pi 50L_n)^2 \\ K &= (2\pi 50)^2 \end{aligned}$$

Para calcular la varianza de I puede aplicarse la fórmula:

$$\sigma_I^2 = \mathrm{E}(I)^2 - [\mathrm{E}(I)]^2 \tag{2}$$

$\mathrm{E}(I^2)$ puede calcularse utilizando el mismo procedimiento que para calcular $\mathrm{E}(I)$, descomponiendo la expresión de I^2 en serie de Taylor, obteniéndose:

$$\mathrm{E}(I^2) = 10.000\left[A^{-1} - \sigma_R^2\left(A^{-2} - 4R_n^2A^{-3}\right) - \sigma_L^2\left(KA^{-2} - 4K^2L_n^2A^{-3}\right)\right]$$

Tal como se ha indicado anteriormente, los valores de σ_R y σ_L se consideran proporcionales a R_n y L_n respectivamente, es decir:

$$\begin{aligned} \sigma_R &= k_R R_n \\ \sigma_L &= k_L L_n \end{aligned}$$

A partir de $\mathrm{E}(I)$ y $\mathrm{E}(I^2)$ es inmediato el cálculo de $\mathrm{Var}(I)$ mediante la expresión (2).

3.2. Valores óptimos de R_n y L_n.

La determinación de los valores óptimos de R_n y L_n puede realizarse cómodamente utilizando una hoja de cálculo para ordenador personal de la forma que a continuación se indica (ver tabla 1):

Tabla 1

Fragmento de la hoja de cálculo utilizada para estudiar las Var(y) *en función de* R_n *y* L_n *forzando* E(y) = 10.00.

	A	B	C	D	E	F	G
1	R_n	L_n	E(y)	V(y)		Kr:	0.1
2	0.00	0.03168	10.00	4.0357		Kl:	0.1
3	0.05	0.03168	10.00	4.0353			
4	0.10	0.03168	10.00	4.0344			
5	0.15	0.03167	10.00	4.0353			
6	0.20	0.03167	10.00	4.0332			
7	0.25	0.03167	10.00	4.0304			
8	0.30	0.03166	10.00	4.0295			
9	0.35	0.03166	10.00	4.0255			
10	0.40	0.03165	10.00	4.0234			
11	0.45	0.03165	10.00	4.0182			
12	0.50	0.03164	10.00	4.0149			
13	0.55	0.03163	10.00	4.0110			
14	0.60	0.03162	10.00	4.0064			
15	0.65	0.03161	10.00	4.0013			
	⋮	⋮	⋮	⋮			

En la 1ª columna se colocan los valores de R_n desde 0 hasta 10Ω (no tienen sentido valores fuera de este intervalo) con incrementos de 0.05Ω. En la 2ª columna se colocan los valores de L_n que junto con el correspondiente valor de R_n dan una E(I) = 10.

Los valores de L_n se han calculado en primera aproximación despejándolos de la expresión (1) y posteriormente se han ajustado por un procedimiento iterativo hasta conseguir que E(I) = 10.00. (El ajuste es necesario ya que $E[I(R,L)] \neq I(R_n,L_n)$).

De esta forma, es fácil descubrir cuales son los valores de R_n y L_n que dando E(I) = 10.00 producen asimismo un valor mínimo de Var(I) (cuarta columna).

A partir de sus correspondientes hojas de cálculo, se han construido los gráficos de las figuras 2 y 3, el primero de los cuales con $\sigma_R = 0.2R_n$, $\sigma_L = 0.01L$ y el segundo con $\sigma_R = 0.1R$, $\sigma_L = 0.01L$. En ellos se presenta la evolución de Var(I) en función de L (con E(I) = 10, por tanto a cada valor de L le corresponde un valor de R). A partir de la misma hoja de cálculo, modificando únicamente los coeficientes que figuran en el extremo superior derecho, se podrían obtener los gráficos para cualquier par de valores k_R, k_L.

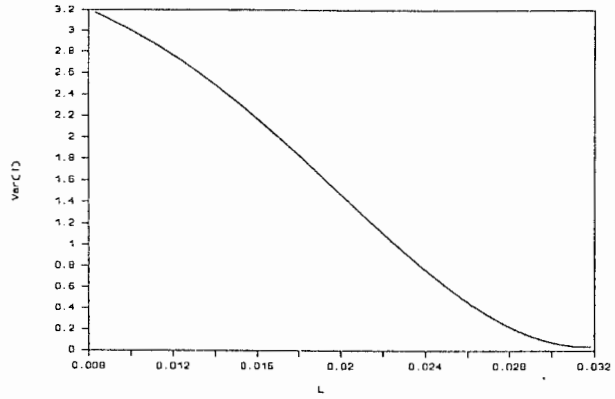


Figura 2.
Varianza de I en función de L_n con $k_R = 0.2$ y $k_L = 0.01$. (El valor de R_n es el obligado para obtener E(I) = 10).

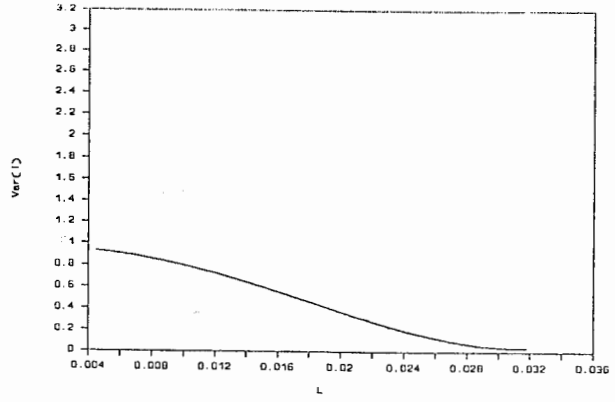


Figura 3.
Varianza de I en función de L_n con $k_R = 0.1$ y $k_L = 0.01$. (El valor de R_n es el obligado para obtener E(I) = 10).

4. METODOLOGÍA PROPUESTA CUANDO NO SE CONOCE LA RELACIÓN FUNCIONAL ENTRE LOS FACTORES Y LA RESPUESTA

La metodología que se propone se basa en el seguimiento de las siguientes etapas:

1. Establecer una hipótesis sobre el modelo de la respuesta.
2. Estimar los parámetros del modelo mediante el análisis de los resultados obtenidos en un plan de experimentación.
3. Utilizando el modelo obtenido, deducir la expresion de la varianza y, mediante un procedimiento análogo al utilizado en el análisis de la relación funcional real, determinar la combinación de valores de los factores que minimiza la variabilidad de la respuesta.

A continuación se describen cada una de estas etapas para el caso planteado.

4.1. Hipótesis sobre el modelo de la respuesta

Se trata de establecer una hipótesis sobre el tipo de relación existente entre la respuesta y los factores, que cumpla los siguientes requisitos:

a) Que explique suficientemente bien el comportamiento de la respuesta.

b) Que sea de fácil tratamiento, tanto en lo que se refiere a la estimación de los parámetros, como a su análisis posterior.

En general, en casos como el planteado se obtienen buenos resultados utilizando modelos cuadráticos (se trata de una aproximación similar a la que se obtendría mediante un desarrollo de Taylor de grado 2) y, por tanto, planteamos el siguiente modelo:

$$I = \beta_0 + \beta_1 R + \beta_2 L + \beta_{11} R^2 + \beta_{22} L^2 + \beta_{12} RL$$

4.2. Estimación de los parámetros del modelo

Los parámetros del modelo pueden estimarse mediante la realización de diseños factoriales fraccionales.

Cuando el modelo es cuadrático, no basta con un diseño a 2 niveles, sino que será necesario un diseño central compuesto (Myers, 1976) o bien un diseño a 3 niveles.

Si se tienen sólo 2 factores como en nuestro caso, el diseño central compuesto con longitud de brazo $\alpha = 1$ y el diseño a 3 niveles coinciden (resulta ser un 3^2) con la siguiente matriz de diseño:

$$\begin{bmatrix} -1 & -1 \\ -1 & 0 \\ -1 & 1 \\ 0 & -1 \\ 0 & 0 \\ 0 & 1 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

Fijando los niveles en los valores:

		Niveles		
		-1	0	$+1$
Factores	$R(\Omega)$	4	5	6
	$L(H)$	0.025	0.0275	0.030

se obtienen las condiciones de experimentación y los resultados que a continuación se indican:

R	L	I
4	0.0250	11.345
4	0.0275	10.503
4	0.0300	9.767
5	0.0250	10.740
5	0.0275	10.018
5	0.0300	9.373
6	0.0250	10.117
6	0.0275	9.507
6	0.0300	8.950

y estimando por regresión los parámetros del modelo se llega a la ecuación:

$$I = 30.5 - 1.64R - 825.77L + 6293.33L^2 + 41.1RL$$

Utilizando la aproximación lineal de transmisión de variabilidad a la respuesta se tiene que:

$$V(I) = \left[\frac{\partial I}{\partial R}|_{(R_n,L_n)}\right]^2 V(R_n) + \left[\frac{\partial I}{\partial L}|_{(R_n,L_n)}\right]^2 V(L_n)$$

y con el modelo obtenido se llega a:

$$(3) \quad V(I) = (-1.64 + 41.1L_n)^2 k_R^2 R_n^2 + (-825.77 + 12586.66L_n + 41.1R_n)^2 k_L^2 L_n^2$$

Siendo $k_R^2 R_n^2$ y $k_L^2 L_n^2$ las varianzas de R y L respectivamente (desviaciones tipo proporcionales al valor nominal).

4.3. Selección de los valores de los parámetros de diseño para minimizar la variabilidad de la respuesta

Utilizando un procedimiento idéntico al desarrollado anteriormente, se llega al gráfico de la figura 4 en el que sólo aparece el rango de valores de L_n en que se ha experimentado y se comparan los valores de la varianza de I en función de L_n según esta se calcule con el modelo real o el experimental.

Tal como puede observarse, las curvas son muy parecidas, y el valor óptimo de L es, tanto si se considera el modelo real como el experimental, el mayor posible. Por tanto, elegiremos los siguientes valores de $L = 0.030H$ y $R = 3.3\Omega$ (el que hace que $\mathrm{E}(I) = 10$ con $L = 0.030H$), obteniéndose una $\mathrm{Var}(I) = 0.14\,\mathrm{Amp}^2$.

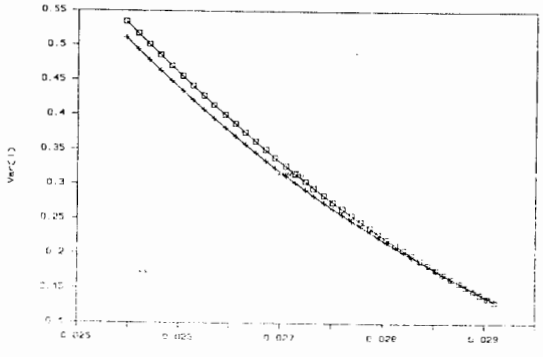


Figura 4.
Varianza de I en función de L_n según el modelo real (□) y el experimental (+).
$k_R = 0.2$ y $k_L = 0.01$.

5. NUEVO ENFOQUE DEL PROBLEMA. UTILIZACIÓN DEL GRÁFICO DISTANCIA-VARIANCIA

Otro enfoque del problema puede consistir en aceptar una cierta desviación del valor nominal de I si con esta desviación se consigue disminuir su variabilidad.

Suponiendo que los valores de R deben estar comprendidos entre 4 y 6 Ω y los de L entre 0.025 y 0.030 H, al igual que con el planteamiento anterior una hoja de cálculo sirve para resolver el problema. En la primera columna se colocan valores de R desde 4 hasta 6 Ω con incrementos de 0.2 Ω. Para cada valor de R, en la segunda columna se colocan valores de L desde 0.025 a 0.030 H con incrementos de 0.0005 H. De esta forma se obtiene un conjunto de combinaciones de valores de R y L (exactamente 121 combinaciones) y para cada una de ellas puede calcularse la varianza de I mediante la expresión (3) y también su esperanza matemática mediante la siguiente expresión, fácilmente deducible del modelo obtenido:

$$\mathrm{E}(I) = 30.5 - 1.64R_n - 825.77L_n + 6293.33(1 + k_L^2)L_n^2 + 41.1R_nL_n$$

A partir de la misma hoja de cálculo puede realizarse un diagrama bivariante de la varianza de la respuesta frente a la distancia al óptimo, que en nuestro caso será $10 - \mathrm{E}(I)$.

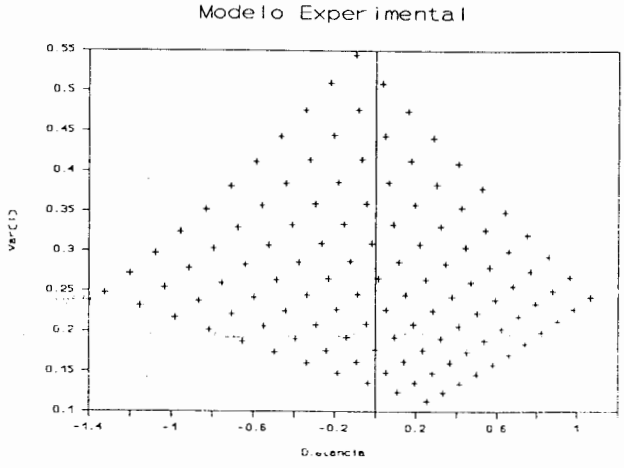


Figura 5.
Gráfico Distancia-Varianza según el modelo experimental. $k_R = 0.2$ y $k_L = 0.01$

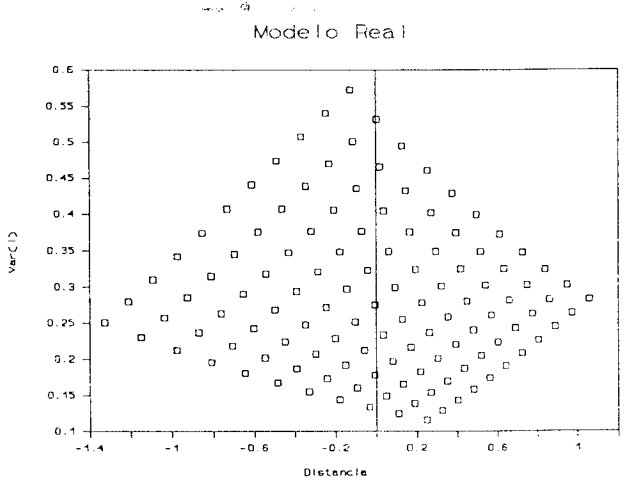


Figura 6.
Gráfico Distancia-Varianza a partir de la relación funcional real. $k_R = 0.2$ y $k_L = 0.01$

Cada punto del diagrama obtenido representa una combinación de valores de los factores de diseño, y a la vista del mismo podemos elegir el punto que mejor se adapte a nuestras necesidades. Del gráfico obtenido a partir del modelo experimental (figura 5) se deduce que la Var(I) puede disminuirse hasta 0.11 a costa de obtener una E(I) = 10.25. Utilizando la relación funcional real se obtiene el gráfico de la figura 6, prácticamente idéntico al anterior y del que pueden obtenerse las mismas conclusiones.

6. CONCLUSIONES

Dado que los componentes de un producto (ya sean componentes electrónicos, materias primas, ... etc) presentan variabilidad, las características de calidad del producto acabado también están afectadas por una cierta variación en torno a su valor objetivo.

La disminución de la variabilidad de una característica de calidad (lo que hemos denominado 'respuesta') puede abordarse de 2 formas distintas:

1) Estableciendo las toleracias que sean necesarias para las características de los componentes.

2) Determinar cuales deben ser los valores nominales de los componentes para que, aunque estos estén afectados de una cierta variabilidad, la respuesta se mantenga dentro de un rango de valores satisfactorio.

Naturalmente, el segundo método es mejor, ya que permite trabajar con materias primas más baratas (cuanto mayores sean las exigencias en las tolerancias más caro será el componente) y, además, asegura que el producto obtenido será correcto sin necesidad de inspección ni controles rigurosos que sí serían necesarios en el caso de estar obligados a trabajar con tolerancias reducidas.

En el presente artículo se ha presentado una metodología para determinar los valores de los parámetros de diseño de forma que la transmisión de variabilidad a la respuesta sea lo menor posible.

Se ha planteado también un procedimiento para obtener una aproximación a la relación funcional cuando ésta no se conoce y, finalmente, se ha presentado el gráfico Distancia-Variancia, que puede ser construido fácilmente con la ayuda de una hoja de cálculo para ordenador personal y que pone de manifiesto, de forma gráfica, toda la información relevante para que el diseñador elija los valores de los parámetros de diseño, atendiendo tanto a la variabilidad de la respuesta como a su distancia al valor óptimo.

BIBLIOGRAFÍA

[1] **G.E.P. Box** y **C.A. Fung** (1986). "Studies in Quality Improvement: Minimizing Transmitted Variation by Parametrer Design".*Center for Quality and Productivity Improvement.* **Report No. 8.** University of Wisconsin-Madison.

[2] **G.E.P. Box, R.N. Kackar, V.N. Nair, M. Phadke, A.C. Shoemaker** i **C.F.J. Wu** (1988). "Quality Practices in Japan". *Quality Progress.* Marzo, 1988.

[3] **G.E.P. Box** y **S. Jones** (1990). "Designing Products That Are Robust to the Environment". *Center for Quality and Productivity Improvement.* **Report No. 28.** University of Wisconsin-Madison.

[4] **D.M. Byrne** y **S. Taguchi** (1987). "The Taguchi Approach to Parameter Design". *Quality Progress.* Diciembre, 1987.

[5] **P. Grima** (1992). *Aportaciones Metodológicas al Diseño de Productos Robustos.* Tesis Doctoral. Universidad Politécnica de Cataluña.

[6] **S.P. Jones** (1990). *Designs for Minimizing the Effect of Envirinmental Variables.* Tesis Doctoral. University of Wisconsin-Madison.

[7] **R.N. Kackar** (1985). "Off-Line Quality Control, Parametrer Design, and the Taguchi Method". *Journal of Quality Technology*, **Vol. 17, 4**. October 1985.

[8] **P.K.H. Lin, L.P. Sullivan** y **G. Taguchi** (1990). "Using Taguchi Methods in Quality Engineering". *Quality Progress*, September 1990. **Volume XXIII, 9.**

[9] **R.H. Myers** (1976). *Response Surface Methodology.* Allyn and Bacon, Inc. Boston.

[10] **P.J. Ross** (1988). *Taguchi Techniques for Quality Engineering.* McGraw-Hill. New York.

[11] **T.P. Ryan** (1988). "Taguchi's Approach to Experimental Design: Some Concerns". *Quality Progress.* Mayo 1988.

[12] **A.C. Shoemaker, K.L. Tsui** y **C.F. Jeff Wu** (1991). "Economical Experimentation Methods for Robust Design". *Technometrics*, **Vol. 33, 4**.

[13] **G. Taguchi** y **M.S. Phadke** (1984). *Quality engineering through design optimization.* Conferencia pronunciada en IEEE GLOBECOM. Atlanta (USA).

[14] **G. Taguchi** (1986). *Introduction to Quality Engineering: Designing Quality Into Products and Processes.* Asian Productivity Organization. Tokyo.

[15] **G. Taguchi, E.A. Elsayed** y **T. Hsiang** (1989). *Quality Engineering in Production Systems.* McGraw-Hill. New York.

[16] **J. Tort-Martorell** (1987). "Comentarios sobre el 'Método Taguchi'". *Boletín de la Asociación Española para el Control de la Calidad.* Mayo 1987.

[17] **J. Tort-Martorell** (1989). *Interacciones y Medida del Ruido en el Método de Taguchi.* Ponencia en el IV Congreso Nacional de la Calidad. Madrid.

[18] **C.F.J. Wu, S.S. Mao** y **F.S. Ma** (1987). "An Investigation of OA-Based Methods for Parameter Design Optimitation". *Center for Quality and Productivity Improvement*, **Report No. 24.** University of Wisconsin-Madison.

ENGLISH SUMMARY:

DETERMINATION OF DESIGN FACTORS VALUES TO OBTAIN ROBUST PRODUCTS. APLICATION TO CIRCUITS WITH INDUCTANCE AND RESISTANCE

Albert Prat and Pere Grima

1. INTRODUCTION

Compared to older methods of controlling quality through final inspections, or the classic method of statistic control of the process, the form that we could currently designate as modern, and indeed the most economic way of assuring quality, consists in designing products so that their performance is maintained within the desired level even if they are not manufactured in the most suitable conditions, or if they are used in different conditions from those expected.

These products are considered to be "robust". The techniques for their design were introduced to the West by the Japanese engineer Genichi Taguchi, who gave his name to the most well known and often used method. In spite of general recognition of the contribution of Taguchi to the design of robust products, the method that he uses has some controversial aspects. Thus, although in some cases the number of experiments required is precisely the number necessary to obtain the specified information, in many other cases experience of each condition of the product demands an effort and a dedication of resources that could be reduced.

Furthermore, the peculiarity of the statistics that Taguchi uses means that his method follows certain non-intuitive procedures that do not ease the understanding of "what" is done and "why" it is done. Furthermore, it has been demonstrated (Box *et al*, 1986) that the statistical techniques used for the analysis of the results are not the most suitable.

This paper presents an alternative to the Taguchi method for the design of parameters using the example of a circuit with a coil and a resistance. These components present a random variability and the objective is for the current in the circuit to have minimal variability from its nominal value.

2. PRESENTATION OF THE PROBLEM

It is supposed that it is necessary to manufacture circuits made up of a power supply, a coil and a resistance, with the following characteristics:

Power Supply: An exact voltage equal to 100 volts.

Resistances: There is a certain variability from one unit to the other (not in the same unit in time). This variability follows a normal distribution centred on the nominal value (R) and has a standard deviation proportional to the nominal value.

Coils: Like the resistances, there are coils that present a certain variability from one component to another. This variability also follows a normal distribution centred on the nominal value (L_n) and a standard deviation proportional to the nominal value.

The problem that we are to solve is: what are the optimum values of R and L, so that the maximum current in the circuit will be 10 amperes, with minimal variability from this value?

3. RESOLUTION USING THE KNOWN FUNCTIONAL RELATIONSHIP

As R and L are random variables, the current flowing through each of the circuits will not always be the same, and we will have to consider it as a new random variable with a given probability distribution. We are interested in knowing the mathematical expectation E(I) and the variance Var(I) of this distribution.

These expressions are deduced by decomposition of the Taylor serial expression for I by assuming that the deviation of the values of R and L are proportional to their nominal values. This implies using a simple procedure based on the use of a spreadsheet to determine the nominal values of R and L to give the desired value for I with minimal variability from this value.

4. METHOD PROPOSED WHEN THE FUNCTIONAL RELATIONSHIP BETWEEN THE FACTORS AND RESPONSE IS NOT KNOWN

The method proposed is based on the following stages:

1. Establish a hypothesis for the response model.
2. Estimate the parameters of the model through analysis of the results obtained from experiments.
3. Use the resulting model to deduce the expression for the variance and, through an analogous procedure to that used in the analysis of the real functional relationship, determine the value combinations of the factors that minimize the variability of the response.

A quadratic equation model is suggested for response and its parameters are estimated through a compound central design. The conclusions drawn from the analysis of the model are identical to those obtained using the real functional relationship.

5. A NEW APPROACH TO THE PROBLEM. THE USE OF THE DISTANCE-VARIANCE GRAPH

Another approach to the problem could be to accept a certain deviation from the nominal value of I if this deviation reduces its variability.

This position is represented by the distance-variance graph.

Each point of the graph represents a value combination of the design factors, and we can choose the point that is best adapted to our needs. The conclusions obtained are practically identical whether the real function relation or the model obtained through experimentation are used.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 171-186, 1995

COMPARACIÓN DE CURVAS DE SUPERVIVENCIA GAMMA

EDUARDO BEAMONTE* y JOSÉ D. BERMÚDEZ‡

A Gamma Hierarchical Model is used for the Bayesian Analysis of Survival Data with covariates. We center our interest on the comparison of two groups of individuals using Monte-Carlo methods. Concretely, we use the Gibbs sampling to obtain a sample from the posterior distribution. The paper ends with the analysis of two data sets, one of them simulated and the other real.

Comparison of Gamma Survival Curves.

Key words: Censored data, Covariates, Gamma Hierarchical Model, Markov-Chain Monte-Carlo Methods, Prediction.

1. INTRODUCCIÓN

En un trabajo anterior (Bermúdez y Beamonte, 1993) propusimos un modelo Gamma con parámetros comunes a todos los individuos, para datos de supervivencia que podían estar progresivamente censurados por la derecha. Esto es, las observaciones realizadas en algunos individuos no muestran su tiempo total de supervivencia, tan sólo se sabe que ese tiempo es mayor que el observado, conocido como tiempo de censura, con un mecanismo de censura independiente del mecanismo de muerte.

*Eduardo Beamonte. Departamento de Economía Aplicada. Universitat de València. Avda. Blasco Ibáñez, 30. 46010 València.

‡José D. Bermúdez. Departamento de Estadística e Investigación Operativa. Universitat de València. C/ Dr. Moliner, 50. 46100 Burjasot. València.

Trabajo parcialmente subvencionado por la Generalitat Valenciana con cargo al proyecto GV-1081/93.

- Article rebut l'agost de 1994.
- Acceptat l'abril de 1995.

Analizábamos el modelo desde una perspectiva bayesiana, estudiando la distribución final mediante una muestra obtenida a partir de cadenas de Markov generadas por Monte-Carlo.

En este trabajo generalizamos el modelo anterior para disminuir su rigidez y poder incorporar covariables. Así, suponemos que el tiempo de supervivencia de cada individuo no tiene una distribución común a todos ellos, sino que esa distribución depende de ciertas características del individuo. Para ello proponemos el siguiente modelo jerárquico:

$$t \sim \mathrm{Ga}(t|\alpha,\beta)$$
$$(\log\alpha,\log\beta)' \sim \mathrm{NM}_2((\log\alpha,\log\beta)'|\mathrm{B}x,\mathrm{H}),$$

siendo t el tiempo de supervivencia y x el vector de covariables de dicho individuo, que habitualmente incorporará un primer elemento constante e igual a 1, como toda matriz de diseño en los modelos lineales generalizados. Es decir, cada tiempo de supervivencia es $\mathrm{Ga}(\alpha,\beta)$, pero los parámetros α y β son características propias no observables del individuo, que están relacionados con su vector de covariables a través de un mecanismo aleatorio que, a su vez, depende de ciertos hiperparámetros B y H comunes a todos los individuos. Así, $\log\alpha$ y $\log\beta$ siguen una distribución Normal bivariante con media $\mathrm{B}x$ y matriz de precisión H.

En este trabajo estamos especialmente interesados en la comparación de dos grupos de individuos. Para ello basta con incluir una única covariable dicotómica, aparte de la covariable constante; así, $x' = (1,1)$ si el individuo pertenece al primer grupo y $x' = (1,0)$ en otro caso. El objetivo primordial será hacer inferencias sobre los hiperparámetros relacionados con esa covariable dicotómica.

La notable complicación de la distribución final (independientemente de la inicial utilizada y acentuada por la presencia habitual de datos censurados) estimula a realizar su estudio a través de una muestra generada a partir de la misma. En esa línea fue pionero el trabajo de Gelfand y Smith (1990).

Uno de los métodos iterativos para la obtención de una muestra de la distribución final más utilizados en los últimos tiempos es el muestreo de Gibbs (ver, por ejemplo, Smith y Roberts, 1993, y referencias allí citadas). Básicamente, consiste en construir una función de transición que defina una cadena de Markov irreducible y para la que la distribución final sea estacionaria. Este proceso exige que sea fácil muestrear a partir de las distribuciones condicionales completas, pero eso suele ser trivial si el vector paramétrico se descompone de forma que tales distribuciones sean univariantes.

En el apartado 2 se analiza el modelo propuesto, detallando la aplicación del muestreo de Gibbs en el estudio de la distribución final. Este procedimento se utiliza en el apartado 3 para el análisis de dos bancos de datos que incluyen una covariable

dicotómica; uno de ellos simulados y el otro basado en datos reales. Por último, el apartado 4 recoge los comentarios y conclusiones finales.

2. ANÁLISIS DEL MODELO JERÁRQUICO GAMMA

Denotamos por $\{t_1,\ldots,t_r\}$ a los tiempos de supervivencia correspondientes a los datos no censurados y por $\{T_{r+1},\ldots,T_n\}$ a los tiempos de censura correspondientes a los datos censurados. El vector paramétrico completo objeto del muestreo de Gibbs es el formado por los parámetros del modelo, $(\alpha_1,\beta_1,\ldots,\alpha_n,\beta_n)$, los hiperparámetros, (B y H), y los tiempos de supervivencia no observados, $(t_{r+1},\ldots,t_n)$, sujetos a las restricciones $t_i > T_i, \quad i = r+1,\ldots,n$. La distribución final viene dada por:

$$f(\alpha_1,\beta_1,\ldots,\alpha_n,\beta_n,B,H,t_{r+1},\ldots,t_n|t_1,\ldots,t_r,T_{r+1},\ldots,T_n,X),$$

siendo $X_{n\times k}$ la matriz de covariables; esto es, una matriz cuya fila $i-$ésima coincide con el vector de covariables del $i-$ésimo individuo.

A partir de esa distribución podemos obtener la distribución marginal de interés:

$$f(B,H|t_1,\ldots,t_r,T_{r+1},\ldots,T_n,X).$$

Para utilizar el muestreo de Gibbs es necesario obtener las distribuciones condicionales completas. La correspondiente a un tiempo censurado no observado es:

$$\begin{aligned} f(t_i|t_1,\ldots,t_{i-1},\, t_{i+1},\ldots,t_n,\, T_{r+1},\ldots,T_n,X,\alpha_1,\beta_1,\ldots,\alpha_n,\beta_n,B,H) = \\ = f(t_i|T_i,\alpha_i,\beta_i), \quad i = r+1,\ldots,n, \end{aligned} \tag{1}$$

siendo $f(t_i|T_i,\alpha_i,\beta_i)$ una distribución Gamma truncada. Específicamente, Bermúdez y Beamonte (1993), $f(t_i|T_i,\alpha_i,\beta_i) \propto Ga(t_i|\alpha_i,\beta_i)$ si $t_i > T_i$.

La distribución condicional completa de los hiperparámetros es:

$$\begin{aligned} f(B,H|t_1,\ldots,t_n,T_{r+1},\ldots,T_n,X,\alpha_1,\beta_1,\ldots,\alpha_n,\beta_n) = f(B,H|X,\alpha_1,\beta_1,\ldots,\alpha_n,\beta_n) \propto \\ \propto f(B,H)f(\alpha_1,\beta_1,\ldots,\alpha_n,\beta_n|X,B,H), \end{aligned} \tag{2}$$

que se reduce a un problema habitual de regresión bivariante normal homocedástica: si se utiliza una distribución inicial $f(B,H)$ perteneciente a la familia Normal-Wishart, la distribución final también es Normal-Wishart cuya expresión puede consultarse, por ejemplo, en Broemeling (1985, capítulo 8, pp. 378-379).

La distribución condicional completa del parámetro α_i, $i = 1,\ldots,n$, es:

$$f(\alpha_i|\alpha_1,\ldots,\alpha_{i-1},\alpha_{i+1},\ldots,\alpha_n,\beta_1,\ldots,\beta_n,t_1,\ldots,t_n,T_{r+1},\ldots,T_n,X,B,H) =$$
$$= f(\alpha_i|t_i,\beta_i,x_i,B,H) \propto f(t_i|\alpha_i,\beta_i)f(\alpha_i|\beta_i,x_i,B,H) =$$
$$(3) \qquad = \mathrm{Ga}(t_i|\alpha_i,\beta_i)\mathrm{NM}(\log\alpha_i|\beta_i,x_i,B,H) \propto \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}t_i^{\alpha_i}\mathrm{NM}(\log\alpha_i|\mu_1',\sigma_1'^2),$$

con $\mu_1' = \mu_1 + \rho\dfrac{\sigma_1}{\sigma_2}(\ln\beta_i - \mu_2)$ y $\sigma_1'^2 = \dfrac{1}{h_{11}}$, los momentos de la distribución Normal condicionada obtenida a partir de la Normal bivariante con vector de medias $\mu = \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} = Bx_i$ y matriz de precisión $H = \begin{bmatrix}h_{11} & h_{12}\\ h_{12} & h_{22}\end{bmatrix} = \Sigma^{-1}$, con $\Sigma = \begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}$ y $\sigma_{12} = \rho\sigma_1\sigma_2$.

Y, finalmente, la distribución completa del parámetro β_i, $i = 1,\ldots,n$, es:

$$f(\beta_i|\beta_1,\ldots,\beta_{i-1},\beta_{i+1},\ldots,\beta_n,\alpha_1,\ldots,\alpha_n,t_1,\ldots,t_n,T_{r+1},\ldots,T_n,X,B,H) =$$
$$= f(\beta_i|t_i,\alpha_i,x_i,B,H) \propto f(t_i|\alpha_i,\beta_i)f(\beta_i|\alpha_i,x_i,B,H) =$$
$$(4) \quad = \mathrm{Ga}(t_i|\alpha_i,\beta_i)\mathrm{NM}(\log\beta_i|\alpha_i,x_i,B,H) \propto \beta_i^{\alpha_i}\exp(-\beta_i t_i)\mathrm{NM}(\log\beta_i|\mu_2',\sigma_2'^2),$$

siendo $\mu_2' = \mu_2 + \rho\dfrac{\sigma_2}{\sigma_1}(\ln\alpha_i - \mu_1)$ y $\sigma_2'^2 = \dfrac{1}{h_{22}}$, la media y varianza de la distribución Normal condicionada correspondiente.

Una vez construidas las distribuciones condicionales completas, y partiendo de un punto inicial $\alpha_1^{(0)},\beta_1^{(0)},\ldots,\alpha_n^{(0)},\beta_n^{(0)},B^{(0)},H^{(0)}$, el algoritmo de Gibbs puede implementarse de la siguiente forma:

Completamos los tiempos censurados, $T_{r+1},\ldots,T_n$, generando los tiempos de supervivencia no observados $(t_{r+1}^{(0)},\ldots,t_n^{(0)})$ a partir de (1) con $(\alpha_i,\beta_i)' = (\alpha_i^{(0)},\beta_i^{(0)})'$, $i = r+1,\ldots,n$. A continuación, generamos $\alpha_1^{(1)}$ a partir de (3), con $t_j = t_j^{(0)}$, $j = r+1,\ldots,n, \beta_1 = \beta_1^{(0)}$, $B = B^{(0)}$ y $H = H^{((0)}$, mediante el método de aceptación-rechazo (ver, por ejemplo, Devroye, 1986, p. 40), utilizando como función importante la densidad log-Student t $(\log\alpha|\mu_1',\sigma_1'^2)$. De forma similar, generamos $\beta_1^{(1)}$ a partir de (4), con $t_j = t_j^{(0)}$, $j = r+1,\ldots,n, \alpha_1 = \alpha_1^{(1)}, B = B^{(0)}$ y $H = H^{(0)}$, utilizando como función importante la densidad log-Student t $(\log\beta|\mu_2',\sigma_2'^2)$. Del mismo modo, generamos $\alpha_2^{(1)}$ a partir de (3), con $t_j = t_j^{(0)}$, $j = r+1,\ldots,n, \beta_2 = \beta_2^{(0)}, B = B^{(0)}$ y $H = H^{(0)}$, generamos $\beta_2^{(1)}$ a partir de (4), con $t_j = t_j^{(0)}$, $j = r+1,\ldots,n, \alpha_2 = \alpha_2^{(1)}, B = B^{(0)}$ y $H = H^{(0)}$, y así sucesivamente. Finalmente, generamos los hiperparámetros $B^{(1)}$ y $H^{(1)}$, a partir de la distribución Normal-Wishart (2), con $(\alpha_i,\beta_i) = (\alpha_i^{(1)},\beta_i^{(1)})$, $i = 1,\ldots,n$.

Repetimos el bucle hasta alcanzar estacionariedad en las cadenas de Markov y, posteriormente, hasta construir una muestra de la distribución final. A partir de esa muestra de la distribución final puede estudiarse cualquier característica deseada.

Comprobamos gráficamente la estacionariedad de las cadenas de Markov, monitorizando la evolución de las mismas. Actualmente sigue abierta la investigación de métodos que calculen de forma automática el número de pasos iniciales a desechar en la cadena de Markov, no habiéndose propuesto todavía ninguno de fácil utilización en aplicaciones concretas. También es un tema actual de investigación el número de cadenas a utilizar para la obtención de la muestra, e incluso si utilizar las cadenas enteras o solamente un submuestreo de las mismas que disminuya la autocorrelación, ver, por ejemplo, Gelman y Rubin (1992), Geyer (1992), MacEachern y Berliner (1994) y referencias allí citadas.

Este esquema teórico del algoritmo de Gibbs es válido para cualquier vector de covariables x. No obstante, a partir de ahora vamos a centrarnos en el caso particular de un vector de covariables dicotómico.

El tiempo medio de supervivencia de un individuo de la primera población es:

$$\begin{aligned} \mu_1 &= \mathrm{E}(t) = \mathrm{E}(\mathrm{E}(t|\alpha,\beta)) = \mathrm{E}(\alpha/\beta) = \mathrm{E}(\exp\{\log\alpha - \log\beta\}) = \\ &= \exp\{b_{11} + b_{12} - b_{21} - b_{22} + 1/2[\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2]^2\}, \end{aligned}$$

y podemos construir una muestra de la distribución final de μ_1 a partir de la muestra $\{\mathrm{B}^{(i)}, \mathrm{H}^{(i)}\}$. Similarmente, el cociente de los tiempos medios de supervivencia de las dos poblaciones es:

$$\tau = \frac{\mu_1}{\mu_2} = \exp\{b_{12} - b_{22}\}.$$

Si la distribución de τ está 'cerca' del 1, las dos poblaciones no serán diferentes en tiempos medios de supervivencia. De hecho, la hipótesis de igualdad de tiempos medios de supervivencia puede formularse como: $b_{12} = b_{22}$.

El estudio de la distribución bivariante $\{b_{12}, b_{22}\}$ permite contrastar la hipótesis de igualdad de poblaciones, pues las dos poblaciones serán iguales si y sólo si $b_{12} = b_{22} = 0$.

Por último, la distribución predictiva podemos aproximarla, por Monte-Carlo, de la siguiente forma:

$$\begin{aligned} f(t|x,t_1,\ldots,t_r,\mathrm{T}_{r+1},\ldots,\mathrm{T}_n,\mathrm{X}) &= \int_0^\infty \int_0^\infty \mathrm{Ga}(t|\alpha,\beta) \cdot \\ \cdot f(\alpha,\beta|x,t_1,\ldots,t_r,\mathrm{T}_{r+1},\ldots,\mathrm{T}_n,\mathrm{X}) d\alpha\, d\beta &\approx \frac{1}{m}\sum_{j=1}^{m} \mathrm{Ga}(t|\alpha_{(j)},\beta_{(j)}), \end{aligned}$$

con $\{(\alpha_{(j)},\beta_{(j)}),\quad j=1,\ldots,m\}$, una muestra obtenida a partir de:

$$\begin{aligned} f(\log\alpha,\log\beta|x,t_1,\ldots,t_r,T_{r+1},\ldots,T_n,X) &= \int \mathrm{NM}((\log\alpha,\log\beta)'|\mathrm{B}x,\mathrm{H})\cdot \\ \cdot f(\mathrm{B},\mathrm{H}|t_1,\ldots,t_r,T_{r+1},\ldots,T_n,X)d\mathrm{B}\,d\mathrm{H} &\sim \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \mathrm{NM}((\log\alpha,\log\beta)'|\mathrm{B}^{(i)}x,\mathrm{H}^{(i)}), \end{aligned}$$

siendo $\{(\mathrm{B}^{(i)},\mathrm{H}^{(i)}),\quad i=1,\ldots,\mathrm{N}\}$ la muestra obtenida mediante Gibbs sampling.

En todas las implementaciones de este algoritmo hemos construido el punto inicial de la siguiente manera: utilizamos la media y la varianza de los tiempos de supervivencia no censurados de cada uno de los grupos para obtener los pares paramétricos iniciales $(\alpha_i^{(0)},\beta_i^{(0)}),\quad i=1,\ldots,n$, (iguales para individuos del mismo grupo y que coinciden con los parámetros de una distribución Gamma cuya media y varianza igualan los correspondientes momentos muestrales de los tiempos de supervivencia no censurados) y una matriz $\mathrm{B}^{(0)}$ inicial, de modo que $\mathrm{B}^{(0)}x$ sea el vector columna formado por los logaritmos de los parámetros de la distribución Gamma anterior. Fijamos la matriz $\mathrm{H}^{(0)}=20\mathrm{I}_2$ (valor inicial que nos ha proporcionado buenos resultados en todas nues tras pruebas).

El algoritmo de Gibbs genera eventualmente una cadena de Markov estacionaria, con independencia del punto inicial elegido. Sin embargo, una elección razonable del mismo permite acele rar el proceso y conseguir estacionariedad en pocas etapas. El método aquí propuesto para la elección del punto inicial es sencillo, parece razonable y ha funcionado correctamente en todos los bancos de datos analizados.

3. EJEMPLOS NUMÉRICOS

El algoritmo expuesto en el apartado anterior lo em pleamos para analizar una batería de bancos simulados a partir del modelo jerárquico, que nos permitían comparar los resultados obtenidos con los verdaderos valores paramétricos utilizados para su simula ción. Todos esos estudios proporcionaron resultados similares. Aquí comentamos un banco de tamaño 200, que incluye 28 datos progresivamente censurados por la derecha.

También hemos utilizado este algoritmo en algún banco de datos reales. En este apartado incluimos el análisis de unos datos sobre remisión de leucemia pues son, posiblemente, los que se han utilizado con mayor frecuencia para com probar el funcionamiento de métodos estadísticos de comparación de curvas de supervivencia.

En los dos ejemplos aquí desarrollados hemos utilizado una distribución inicial Normal-Wishart poco informativa para los hiperparámetros, en concreto:

$$f(B,H) = NMW(B,H|M,A,T,\alpha) = NM(B^v|M^v, H \otimes A)W(H|T^{-1},\alpha),$$

con $M = 0_{2\times 2}, A = I_2, T = (0.01)I_2$ y $\alpha = 2$.

3.1. Análisis de un banco de datos simulados

Construimos un banco de 200 datos simulados a partir del modelo jerárquico Gamma con hiperparámetros $B = \begin{bmatrix} 4.6 & 0.2 \\ 2.3 & 0.0 \end{bmatrix}$ y $H = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$.

Tras realizar 20000 pasos en la cadena de Markov (19 minutos en una estación Sun SPARCclassic), partiendo de un punto inicial construido según el método propuesto en el apartado anterior, monitorizamos la evolución de las medias de los tiempos de supervivencia en cada uno de los grupos, representándolas cada 10 pasos en la figura 1. En dichas gráficas se observa estabilidad desde casi el principio, además de muy poca variabilidad en ambos grupos.

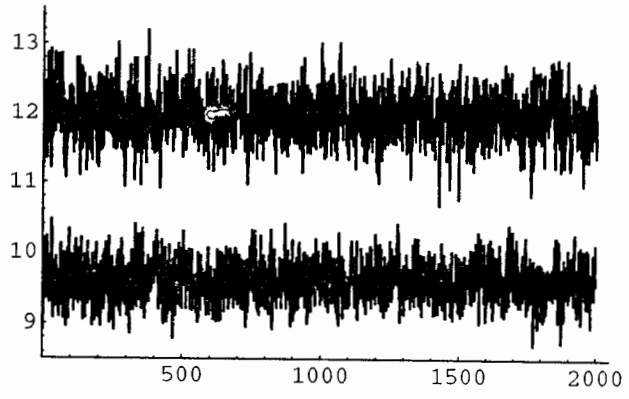


Figura 1.
Evolución de las medias de los tiempos de supervivencia en cada uno de los grupos del banco de datos simulados.

El estudio de la densidad predictiva lo realizamos desechando los 15000 primeros pasos de la cadena de Markov y, posteriormente, cogiendo uno de cada 10 hasta obtener una muestra de tamaño 500. Así, por Monte-Carlo obtuvimos las medias y

varianzas de las densidades predictivas para cada uno de los grupos. Se muestran en la tabla 1.

Tabla 1

Media y varianza de la predictiva en cada uno de los grupos del banco de datos simulados

	Grupo 1	Grupo 2
Media	12.3769	9.9653
Varianza	11.2328	7.7097

Las gráficas, para cada uno de los grupos, de las densidades predictivas (en trazo continuo) y de las densidades con las que realmente hemos simulado los datos (en trazo discontinuo), se recogen en la figura 2.

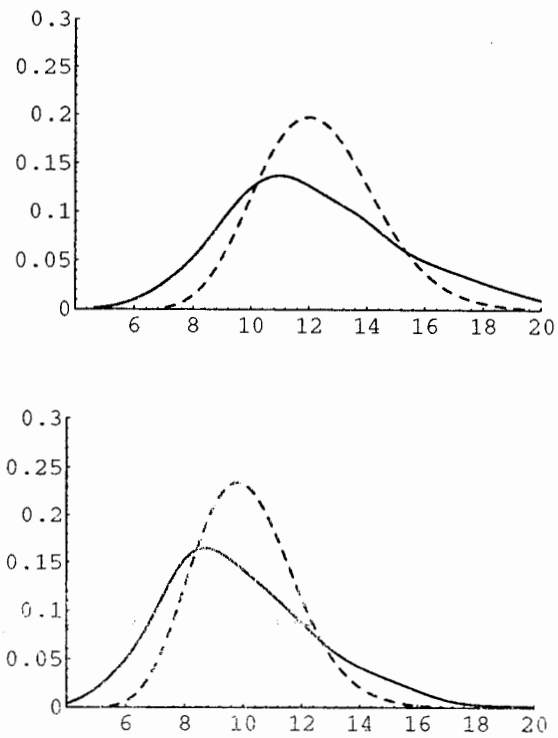


Figura 2.
Densidad predictiva y densidad correcta de cada uno de los grupos del banco de datos simulados.

Por último, hemos representado la nube de puntos de la distribución bivariante $\{b_{12}, b_{22}\}$ y el histograma de $\tau = \exp\{b_{12} - b_{22}\}$. Estas gráficas se muestran en la figura 3 y permiten comparar los dos grupos como ya comentamos en el apartado 2. Observamos que la nube de puntos se encuentra por debajo de la diagonal y no se acerca al cero, por lo que podemos deducir la diferencia entre los dos grupos.

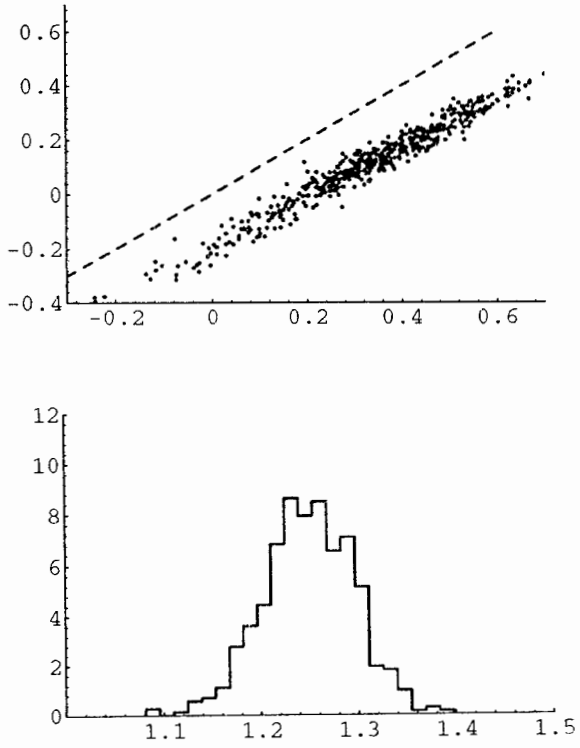


Figura 3.
Nube de puntos de $\{b_{12}, b_{22}\}$ e histograma de τ del banco de datos simulados.

Para comprobar la consistencia del estudio independientemente del punto inicial elegido, analizamos el mismo banco de datos partiendo de distintos puntos iniciales. Los resultados obtenidos fueron muy similares a los presentados. Por ejemplo, partiendo de un punto inicial que coincidía con los verdaderos valores paramétricos (esto es posible dado que los datos son simulados), observamos una evolución de las medias de los tiempos de supervivencia en cada uno de los grupos análoga a la de la figura 1 y realizando el estudio de la densidad predictiva siguiendo el mismo criterio, obtuvimos los siguientes estadísticos:

Tabla 2

Media y varianza de la predictiva en cada uno de los grupos del banco de datos simulados, utilizando puntos iniciales iguales a los verdaderos valores paramétricos

	Grupo 1	Grupo 2
Media	12.4865	10.0578
Varianza	10.5993	6.9834

3.2. Análisis de datos de leucemia

El banco de datos sometido a análisis consiste en 42 tiempos de remisión de la leucemia, medidos en semanas, en dos grupos con el mismo número de pacientes (Freireich, E.S. *et al.*, 1963). Al primero de ellos se le somete a tratamiento con 6-mercatopurina (6-MP), mientras que al segundo se le administra un placebo. Se recogen los datos en la tabla 3, donde el asterisco indica observación censurada.

Tabla 3

Tiempos de remisión (en semanas) de leucemia en dos grupos de pacientes

6-MP	6	6	6	6*	7	9*	10	10*	11*	13	16	17*	19*	20*	22	23	25*	32*	32*	34*	35*
Placebo	1	1	2	2	3	4	4	5	5	8	8	8	8	11	11	12	12	15	17	22	23

Estos datos han sido profusamente analizados por diferentes autores para estudiar nuevos métodos y tests de comparación de dos distribuciones de supervivencia. Así, Lawless (1982), previa constatación gráfica de que los tiempos de remisión en cada uno de los grupos podían provenir de distribuciones Weibull, obtuvo los estimadores máximo verosímiles de los parámetros y concluyó que ambos grupos eran distintos en tiempos medios de supervivencia utilizando intervalos de confianza. Asimismo, Lee (1992) utilizó el test F de Cox para distribuciones exponenciales para demostrar que los grupos no podían ser considerados iguales.

Al igual que en el subapartado anterior, construimos el punto de partida del algoritmo de Gibbs y monitorizamos la evolución de las medias de los tiempos de supervivencia en cada uno de los grupos, representándolas cada 20 pasos en la figura 4. También se observa estabilidad desde casi el principio, aunque mayor variabilidad en el primer grupo.

La evolución de los hiperparámetros B es bastante similar, mostrando estabilidad casi desde el principio de la cadena de Markov. De un modo totalmente análogo a la gráfica anterior, se muestra en la figura 5 la evolución de b_{12} y b_{22}, que son los dos parámetros que permiten diferenciar los grupos.

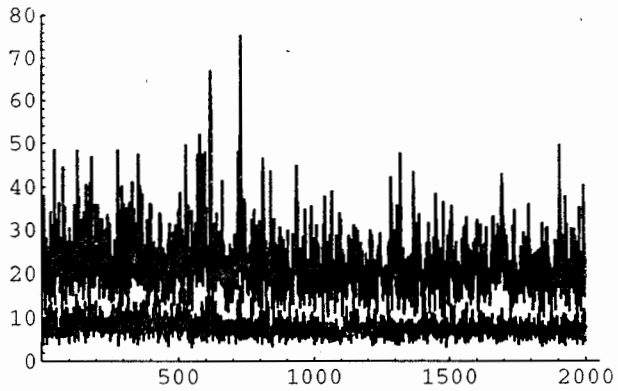


Figura 4.
Evolución de las medias de los tiempos de supervivencia en cada uno de los grupos del banco de datos de leucemia.

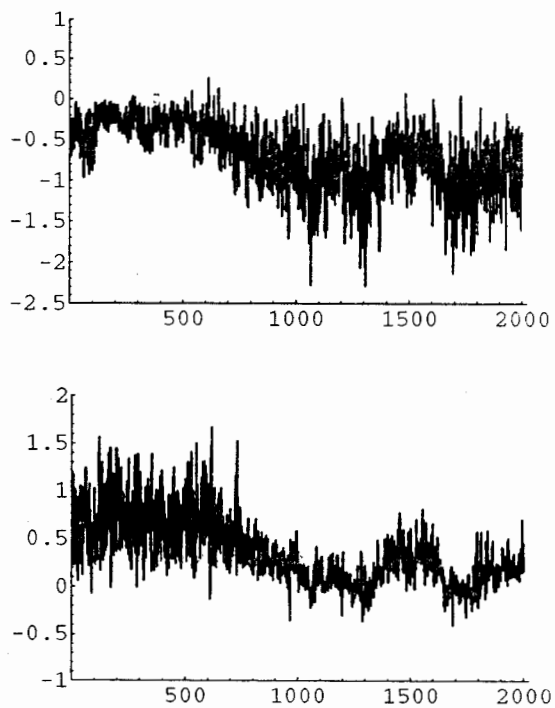


Figura 5.
Evolución de los hiperparámetros b_{12} y b_{22} en el banco de datos de leucemia.

El estudio de la densidad predictiva lo realizamos desechando los 20000 primeros pasos de la cadena de Markov y, posteriormente, cogiendo uno de cada 20 hasta obtener una muestra de tamaño 500. Así, por Monte-Carlo obtuvimos las medias y varianzas de las densidades predictivas para cada uno de los grupos. Se muestran en la tabla 4.

Tabla 4

Media y varianza de la predictiva en cada uno de los grupos en el banco de datos de leucemia

	Grupo 1	Grupo 2
Media	35.6264	13.3486
Varianza	2550.9143	692.7137

Representamos, en la figura 6, las gráficas de las densidades predictivas para cada uno de los dos grupos.

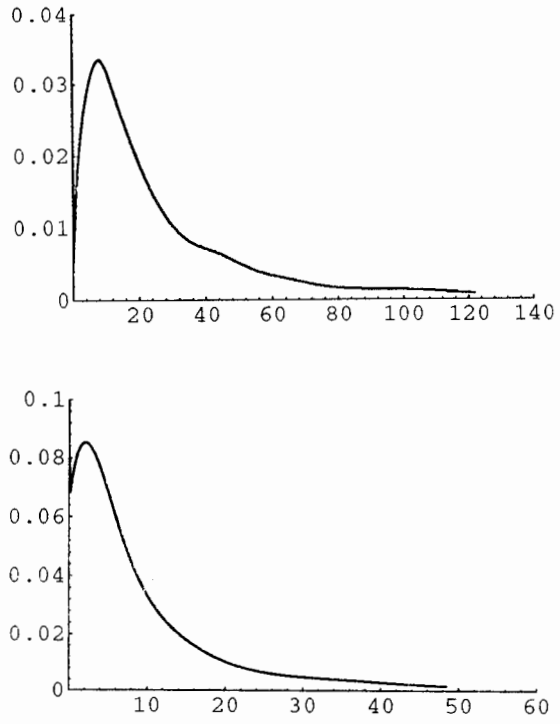


Figura 6.
Densidad predictiva de cada uno de los grupos en el banco de datos de leucemia.

Finalmente, en la figura 7 representamos la nube de puntos de la distribución bivariante $\{b_{12}, b_{22}\}$ y el histograma de $\tau = \exp\{b_{12} - b_{22}\}$ para ambos grupos. Observamos que la nube de puntos se encuentra por encima de la diagonal y no se acerca al cero, por lo que podemos deducir la diferencia entre los dos grupos.

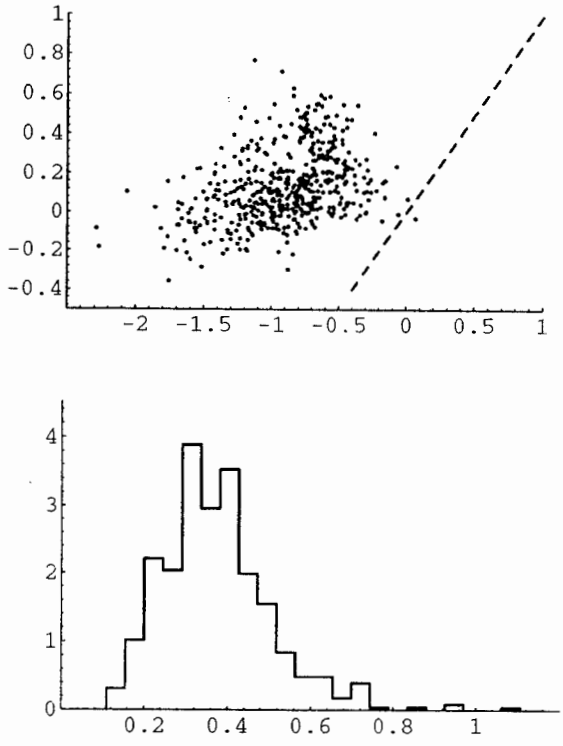


Figura 7.
Nube de puntos de $\{b_{12}, b_{22}\}$ e histograma de τ en el banco de datos de leucemia.

4. COMENTARIOS FINALES

El análisis de los dos ejemplos numéricos estudiados en el apartado anterior muestra claramente que existen diferencias entre los dos grupos de individuos. La respuesta bayesiana adecuada al problema de comparación de tiempos medios de supervivencia en las dos poblaciones vendría dada por el cálculo de algún intervalo de confianza

sobre la distribución final del parámetro $\tau = \exp\{b_{12} - b_{22}\}$. Como ya se comentó en el apartado 2, si existe igualdad de tiempos medios en las dos poblaciones $\tau = 1$. Similarmente, si existe igualdad entre las dos poblaciones, $b_{12} = b_{22} = 0$; por tanto, para responder a la pregunta de igualdad de poblaciones habría que construir alguna región de confianza de la distribución final conjunta de (b_{12}, b_{22}).

Como una primera solución, hemos obtenido un intervalo de confianza 0.95 para τ utilizando el histograma normalizado de la figura 7 como aproximación de su distribución final. Siguiendo las indicaciones de Wei y Tanner (1990), obtuvimos el intervalo (0.1525, 0.5637) que claramente no incluye el valor τ=1.

Por otro lado, en este trabajo hemos exigido que la distribución inicial sobre los hiperparámetros B y H pertenezca a la familia conjugada. Esto no supone una restricción importante: si se desea utilizar otra inicial el método de aceptación-rechazo permite obtener una muestra de la distribución final deseada a partir de una muestra construida según el procedimiento descrito en el apartado 2, empleando una distribución inicial Normal-Wishart que no sea muy distinta de la inicial que se desea utilizar.

Los comentarios de dos evaluadores anónimos sobre una versión anterior de este trabajo, han permitido mejorar notablemente la claridad en la exposición del mismo.

REFERENCIAS

[1] **Bermúdez, J.D.** y **Beamonte, E.** (1993). "Análisis bayesiano de datos de supervivencia Gamma utilizando muestreo de Gibbs". *Estadística Española*, **35**, 629–644.

[2] **Broemeling, L.D.** (1985). *Bayesian analysis of linear models*. New York: Marcel Dekker.

[3] **Devroye, L.** (1986). *Non-uniform random variate generation*. New York: Springer-Verlag.

[4] **Freireich, E.J., Gehan, E.A., Frei, E. et al.** (1963). "The effect of 6-Mercaptopurine on the duration of steroid-induced remissions in acute leukemia: a model for evaluation of other potential useful therapy". *Blood*, **21**, 699–716.

[5] **Gelfand, A.E.** and **Smith, A.F.M.** (1990). "Sampling-based approaches to calculating marginal densities". *J. Amer. Statist. Assoc.*, **85**, 398–409.

[6] **Gelman, A.** and **Rubin, D.B.** (1992). "Inference from iterative simulation (with discussion)". *Statistical Science*, **7**, 457–511.

[7] **Geyer, C.J.** (1992). "Practical Markov chain Monte Carlo (with discussion)". *Statistical Science*, **7**, 473–511.

[8] **Lawless, J.F.** (1982). *Statistical models and methods for lifetime data.* New York: John Wiley & Sons.

[9] **Lee, E.T.** (1992). *Statistical methods for survival data analysis.* 2nd. edition. New York: John Wiley & Sons.

[10] **MacEachern, S.N.** and **Berliner, L.M.** (1994). "Subsampling the Gibbs sampler". *The American Statistician*, **48**, 188–190.

[11] **Smith, A.F.M.** and **Roberts, G.O.** (1993). "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods". *J. Roy. Statist. Soc.* **B, 55**, 3–23.

[12] **Wei, G.C.G.** and **Tanner, M.A.** (1990). "Calculating the content and boundary of the highest posterior density region via data augmentation". *Biometrika*, **77**, 649–652.

ENGLISH SUMMARY:

COMPARISON OF GAMMA SURVIVAL CURVES

Eduardo Beamonte y José D. Bermúdez

In this paper we propose a hierarchical model for the bayesian analysis of survival data, progressively censored on the right, with covariates. This is a generalization of a previous work (Bermúdez and Beamonte, 1993) where each survival time is assumed to be distributed as a Ga (α, β) with $\log\alpha, \log\beta$ distributed as a bivariate Normal with mean Bx and precision matrix H, where x is the vector of covariates for each individual.

We are specially interested on the comparison of two groups of individuals. For that we considerate a dichotomous covariate, which identifies the group membership of each individual, together with a constant covariate equal to one. Our main objective is obtain inferences on the hiperparameters related with the dichotomous covariate.

The great complexity of the posterior distribution, independently of the prior distribution used and emphasised by the tipical presence of censored data, drive us to study it through Monte Carlo methods (in this line was pioneer the work of Gelfand and Smith, 1990). In order to do this, we use the Gibbs sampling, one of the most

important iterative methods of obtaining a sample from the posterior distribution (see Smith and Roberts, 1993, and references there). It requires that the sampling from all of the complete conditional distributions be easy, but that is our case.

If we denote the survival times corresponding to no censored data by $(t_1,\ldots,t_r)$ and the censorship times corresponding to censored data by $(T_{r+1},\ldots,T_n)$, then the parametric vector object of the Gibbs sampling is composed by the parameters of the model $(\alpha_1,\beta_1,\ldots,\alpha_n,\beta_n)$, the hiperparameters (B and H), and the nonobserved survival times $(t_{r+1},\ldots,t_n)$, subjets to the restrictions $t_i > T_i, \quad i = r+1,\ldots,n$. So the complete conditional distribution of a nonobserved survival time is the truncated Gamma (1) and the corresponding to the hiperparameters is the Normal-Wishart (2) (if we use a distribution belonging to that family for the prior f (B, H)). In both cases is relatively easy sampling from them. We solve the problem of sampling from the complete conditional distribution of α_i (3) and the corresponding to β_i (4), using the acceptance-rejection method with importance functions belonging to the log-t-student family.

We implement the algorithm of the Gibbs sampling with an initial point based on the first moments of the exact survival times and we check grafically the convergence of the Markov chains. Once a sample of the posterior distribution is obtained, we can calculate some of its characteristics by Monte Carlo methods and even to obtain the predictive distribution, of special interest in most applications.

The paper ends with the analysis of two data sets, one of them simulated and the other real. The first one, together with some others which we have analysed, permits to compare the results obtained with the true parameters. The real data consist in 42 remission times of leukemia (Freireich *et al.*, 1963), that has been analysed by many other authors. We apply our method and obtain the predictive distribution for each group. We conclude that the two groups are differents by obtaining a confidence interval of the parameter of interest, agreeing with previous works.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 187-215, 1995

ANÁLISIS COMPARATIVO DE CÁLCULO DE PREVISIONES UNIVARIABLES Y FUNCIÓN DE TRANSFERENCIA, MEDIANTE LAS METODOLOGÍAS DE BOX-JENKINS Y REDES NEURONALES

DAVID DE LA FUENTE GARCÍA* y RAÚL PINO DÍEZ*

En este trabajo se presenta una comparación de los resultados obtenidos en el cálculo de previsiones de series temporales univariables y de función de transferencia, mediante la metodología de Box-Jenkins y la utilización de Redes Neuronales Artificiales. Los resultados obtenidos indican que la aproximación de las redes neuronales, ha sido muy satisfactoria, tanto en lo referente a la bondad de las previsiones obtenidas, como, sobre todo, por la facilidad de su utilización. En las Redes Neuronales Artificiales, o "Redes de Conocimiento", todavía no son muy conocidos los mecanismos matemáticos internos que justifican su funcionamiento, sin embargo, esto no es óbice para comprobar que los resultados que dan son muy interesantes. El problema principal es que, al menos por el momento, el desarrollo de las redes neuronales es heurístico, no existen reglas fijas que determinen la estructura de la red neuronal apropiada a cada caso de estudio. En este trabajo se propone la utilización de la metodología de Box-Jenkins como un paso previo al diseño de la estructura de la red neuronal.

Comparative analysis of univariate and transfer-function forecasting using neural networks and Box-Jenkins techniques.

Key words: Forecasting, Multivariate Time-Series, Neural Network, Back Propagation, Training.

*David de la Fuente García y Raúl Pino Diez. Escuela Técnica Superior de Ingenieros Industriales e Ingenieros Informáticos de Gijón. Universidad de Oviedo. Carretera de Castiello, s/n. 33204. GIJÓN (Asturias).

E-mail: David @etsiig.uniovi.es

-Article rebut el juliol de 1993.

-Acceptat l'abril de 1995.

1. INTRODUCCIÓN

El análisis de series temporales, es una herramienta estadística muy importante en el estudio del comportamiento de datos dependientes del tiempo y la predicción de valores futuros. Una serie temporal, es una secuencia de valores medidos en unidades de tiempo discretas o continuas.

Una serie temporal multivariable, consiste en una secuencia de valores de varias variables contemporáneas cambiando con el tiempo. Un caso aparte, es cuando las variables medidas están significativamente correlacionadas, por ejemplo, cuando atributos similares están siendo medidos en diferentes localizaciones geográficas. En la previsión de un valor nuevo para cada variable, la mejor predicción, se conseguirá teniendo en cuenta las variaciones de las otras variables.

Históricamente, podemos decir que en la década de los ochenta, había muchas técnicas disponibles para el análisis de series temporales, las cuales asumían relaciones lineales entre las variables (Box- Jenkins, 1976). Pero en el mundo real los datos no presentan relaciones simples y es difícil hacer predicciones seguras. Las relaciones lineales y sus combinaciones a menudo son inadecuadas para hacer previsiones. Por ello, parece necesario que sean modelos no lineales, los que sirvan para analizar los datos temporales del mundo real.

Los modelos lineales de series temporales presentan además, ciertas desventajas, como la imposibilidad de explicar cambios repentinos de amplitud muy grande y en intervalos irregulares de tiempo (Tong, 1983).

Estudios posteriores (Tiao y Tsay, 1989), analizaron otros problemas del modelado lineal multivariable. Para resolverlos, se utilizaron modelos estadísticos no-lineales como "modelos de umbral" y "modelos bilineales" sugeridos por Tong (1990). Por otra parte, Granger y Newbold (1986), sugirieron utilizar transformaciones no lineales de los datos originales antes de realizar el modelado lineal tradicional. Farmer y Sidorowich (1987), mejoraron significativamente las predicciones sobre series temporales caóticas utilizando métodos de aproximaciones locales.

Desgraciadamente, a pesar de que en la década pasada se avanzó considerablemente, la formulación de modelos no lineales razonables es una tarea extremadamente difícil, debido a las simplificaciones hechas en la etapa de modelado, por ejemplo, omitir parámetros que son desconocidos o que no parece que afecten directamente a los datos observados etc.

Por todo ello, como el objetivo es hacer buenas previsiones, buscamos la alternativa de la utilización de redes neuronales artificiales para el cálculo de predicciones cuantitativas.

En este trabajo, hacemos una somera descripción de la metodología de Box-Jenkins para modelado temporal univariable, así como funciones de transferencia para el caso de una entrada y una salida. A continuación, describimos la aproximación de Redes Neuronales para finalizar comparando con varios ejemplos, los resultados obtenidos mediante ambos enfoques, en el modelado univariante y función de transferencia de una entrada y una salida.

2. MODELADO UNIVARIANTE Y DE FUNCION DE TRANSFERENCIA DE BOX-JENKINS

Se sabe que el procedimiento ARIMA (autorregresivo, integrado, media móvil), identifica, calcula estimaciones y previsiones de modelos, utilizando la metodología descrita por Box-Jenkins (1976). Los procedimientos ARIMA, permiten modelar una serie temporal discreta como una función de términos autorregresivos y términos media móvil pudiendo incluirse en el modelo factores estacionales y no estacionales de cada tipo, llamados habitualmente: MA, AR, SMA y SAR.

El proceso de modelado de series temporales según Box-Jenkins, consta de varias fases:

1- Identificación y propuesta de un modelo para los datos observados.

2- Estimación de los parámetros y contrastes.

3- Crítica y diagnosis del modelo.

4- Validación del modelo. (Si no es válido se propone otro modelo en el paso 1).

5- Previsiones.

También se modelan, en su caso, la serie diferenciada o la no diferenciada para eliminar tendencias, mediante la utilización de las diferencias estacionales y no estacionales.

Existen en el mercado bastantes paquetes comerciales que utilizan la metodología de Box-Jenkins univariable, STATGRAPHICS, BMDP, RATS, TSP, SPSS etc. Algunos de ellos, utilizan funciones de transferencia, como el BMDP, y muy pocos modelado multivariable (varias entradas y varias salidas), por ejemplo, SCA y RATS (solo modelo AR). Incluso, existen algunos paquetes de cálculo automático de previsiones univariantes como son AUTOBOX, FOCA etc. y últimamente han aparecido sistemas expertos en previsión univariante, como el SCA EXPERT. Para la realización de este trabajo, en el caso de modelado univariante hemos escogido el programa

STATGRAPHICS por su sencillez mientras que para el modelado de función de transferencia, hemos utilizado el BMDP.

2.1. Modelado Univariante

La forma básica del modelo que va a ser ajustado, es la siguiente:

$$W_t = \mu + \frac{\theta(B) \cdot \theta_s(B)}{\phi(B) \cdot \phi_s(B)} \cdot a_t$$

Este modelo, expresa los datos como una combinación de los valores de la serie y los valores de una entrada aleatoria, donde:

t	Índice de tiempo.
B	Operador de retardo.
W_t	Datos originales o una diferencia de estos datos.
μ	Término constante.
$\theta(B)$	Operador media móvil no estacional $(1 - \theta_1 B - \cdots - \theta_q B^q)$
$\phi(B)$	Operador autorregresivo no estacional $(1 - \phi_1 B - \cdots - \phi_p B^p)$
$\theta_s(B)$	Operador media móvil estacional $(1 - \theta_1 B^s - \cdots - \theta_Q B^{Qs})$
$\phi_s(B)$	Operador autorregresivo estacional $(1 - \phi_1 B^s - \cdots - \phi_P B^{Ps})$
a_t	Error aleatorio.

Este modelo se puede representar con la siguiente notación, $(\boldsymbol{p},\boldsymbol{d},\boldsymbol{q}) \times (\boldsymbol{P},\boldsymbol{D},\boldsymbol{Q})_s$, donde $\boldsymbol{p},\boldsymbol{d}$ y $\boldsymbol{q}$ representan los órdenes del término autorregresivo, de la diferencia y del término media móvil estacionales respectivamente, y $\boldsymbol{P},\boldsymbol{D}$ y $\boldsymbol{Q}$, los mismos órdenes pero estacionales. Por último, s representa la longitud de la estacionalidad.

La identificación y validación del modelo, se obtienen mediante el cálculo de las funciones de autocorrelación y autocorrelación parcial de la serie original y los residuos, respectivamente.

Para la estimación de los parámetros del modelo propuesto, se utiliza el algoritmo de Gauss- Marquardt, que es un método de optimización no lineal con restricciones y combina los algoritmos iterativos de Gauss-Newton y el de máxima pendiente.

El primer ejemplo que vamos a considerar para el modelado univariante, es la serie temporal que llamamos UN05.DAT, obtenida de la base de datos de Reilly, D.P. (1980), y cuyo gráfico representamos a continuación.

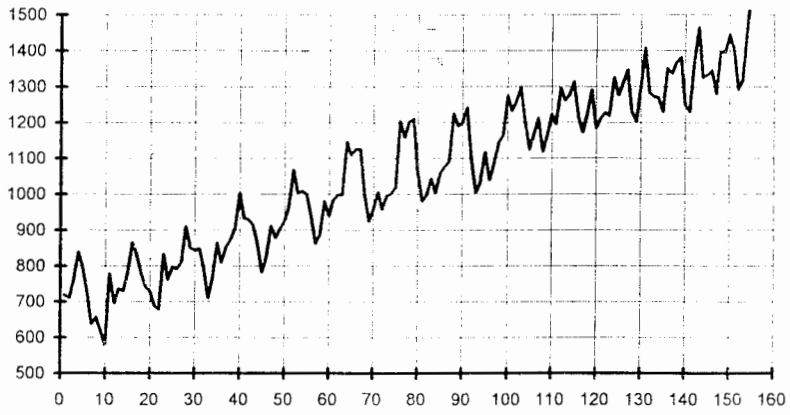


Figura 1.
Serie UN05.DAT.

Esta serie, consta de 155 valores, de los cuales hemos utilizado los primeros 134 como datos para al cálculo de las previsiones mediante la técnica de Box-Jenkins, el resto (21 valores), servirán para comprobar la bondad del método. Utilizando, como ya se ha comentado, el paquete STATGRAPHICS, se obtuvo el modelo:

$$(1-0.6B)(1-B^{12})Z = \text{CONS} + \text{NOISE}$$

Los resultados que se ven en la figura 2, son la representación gráfica de las previsiones calculadas, comparadas con los valores reales de la serie UN05.

Figura 2.
Valores reales/Previsiones Box-Jenkins.

2.2. Modelado de Funciones de Transferencia

Previo a este tipo de modelado, habría que considerar el llamado análisis de intervención que correspondería, a efectos puntuales en la evolución de la serie temporal, o a efectos que se producen a partir de un determinado instante en el desarrollo de la serie (un ejemplo del primer caso, podría ser un día de fiesta en el cálculo del consumo de electricidad, y un ejemplo del segundo podría ser la entrada en vigor de una ley a partir de una determinada fecha). No obstante, al ser casos particulares del modelado multivariante, no entraremos en ellos.

El modelado de funciones de transferencia, pretende encontrar un modelo estocástico que relacione varias variables explicativas $X_1, X_2, \ldots, X_n$ (input) y una variable explicada (output) Y (Figura 3). El procedimiento de cálculo del modelo, implica los mismos pasos que en el modelado univariante: identificación tentativa, estimación de parámetros eficientemente y verificación del mismo. Resultando en este caso, una regresión dinámica pero ahora con regresores las series $X_1, X_2, \ldots, X_n$ y el ruido del sistema.

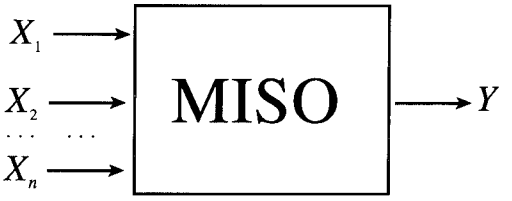


Figura 3.
Modelado de Función de Transferencia (Multiple Input Single Output).

El modelo de función de transferencia, se supone que es lineal, causal y discreto. Además, las entradas $X_1, X_2, \ldots, X_n$ son independientes y no existe realimentación entre ningún par de variables a relacionar. Por ello, el modelo propuesto tiene la siguiente expresión:

$$Y_t = \sum_{i=1}^{n} \frac{\omega_i(B)}{\delta_i(B)} \cdot X_{it} + \frac{\theta(B)}{\phi(B)} \cdot a_t$$

La herramienta básica para obtener el modelo adecuado, es la función de correlación cruzada; vamos a describir los pasos necesarios para el cálculo de una función de transferencia cuando tenemos una sola entrada, aunque es fácilmente generalizable y se encuentra en la literatura al uso.

Los pasos a realizar en la especificación, estimación y contraste de un modelo de función de transferencia son (Figura 4):

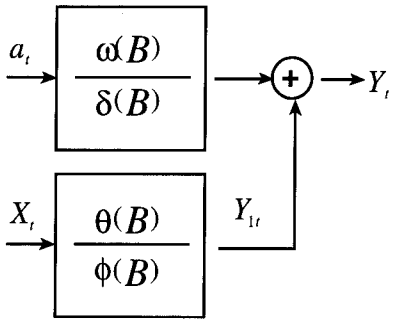


Figura 4.
Cálculo de la Función de Transferencia.

1) Se hace un análisis univariante de la serie $X(t)$.

2) Preblanqueado de la serie $Y(t)$ (esto es filtrar la serie $Y(t)$ en el modelo univariante de $X(t)$), y estimación inicial de los coeficientes de la respuesta $v(k)$ (o sea, cálculo de la función de correlación cruzada entre los residuos de la serie $X(t)$ y los obtenidos al filtrar la serie $Y(t)$).

3) Determinación de los órdenes de los polinomios de la función de transferencia, $w(B)$ y $\delta(B)$, a partir de la relación:

$$(1 - \delta_1 B^1 - \cdots - \delta_r B^r) \cdot (v_0 + v_1 B + \cdots) = (w_0 - w_1 B - \cdots - w_s B^s)$$

que nos da una orientación sobre los posibles valores de r y s. Un análisis detallado de los diferentes valores de r y s, nos permitirá obtener los más adecuados.

4) Determinación de los órdenes de los polinomios del término de error, $\phi(B)$ y $\theta(B)$, bien a través del análisis univariante de $Y(t)$ o bien estudiando el comportamiento de los residuos.

5) Etapa de estimación de los coeficientes del modelo: Mediante el procedimiento de mínimos cuadrados no lineales.

6) Etapa de validación, que consiste en comprobar que los residuos que se obtienen al relacionar $X(t)$ con $Y(t)$, son ruido blanco, si no ocurre esto, se debe de dar estructura al modelo de ruido que será, generalmente, muy semejante al modelo univariante de $Y(t)$. En esta última etapa de validación, se deben de utilizar, entre otros, los contrastes estadísticos **t** para la significación de los parámetros, matriz de correlación de los parámetros y contraste de ruido blanco y ausencia de correlación cruzada entre $X(t)$ y $a(t)$.

Como el objetivo del trabajo es comparar diferentes metodologías, vamos a utilizar un ejemplo muy conocido para el cálculo de la función de transferencia. A continuación, se muestran los resultados de la aplicación del método descrito, para el modelado de la función de transferencia en un ejemplo. En él, se trata de relacionar la concentración de CO_2 resultante, $Y(t)$, como salida de un horno de gas, en función de la tasa de gas que lo alimenta, $X(t)$. Las dos series $X(t)$ e $Y(t)$, constan de 296 datos, y aparecen representadas en la figura 5.

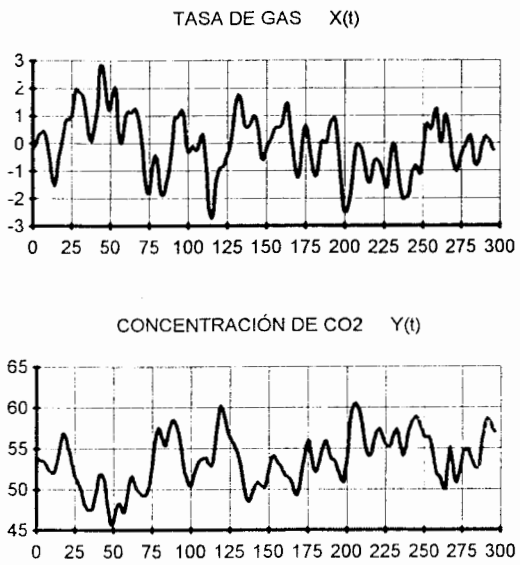


Figura 5.
Series Temporales $X(t)$ e $Y(t)$.

De la misma manera que se ha hecho en el caso univariante, utilizaremos los 250 primeros valores de las dos series temporales, $X(t)$ e $Y(t)$, para deducir el modelo y calcularemos 40 previsiones que comparadas con los valores reales nos darán una idea de la bondad del método. Utilizando el paquete comercial BMDP (programa P2T), se puede deducir el modelo:

$$(1 - 1.53B + 0.63B^2)Y_t = 53.4 + \frac{-0.53B^3 - 0.38B^4 - 0.52B^5}{1 - 0.55B}X_t + a_t$$

Los resultados de las previsiones obtenidas, junto con los valores reales de la serie $Y(t)$, se pueden observar en la figura 6.

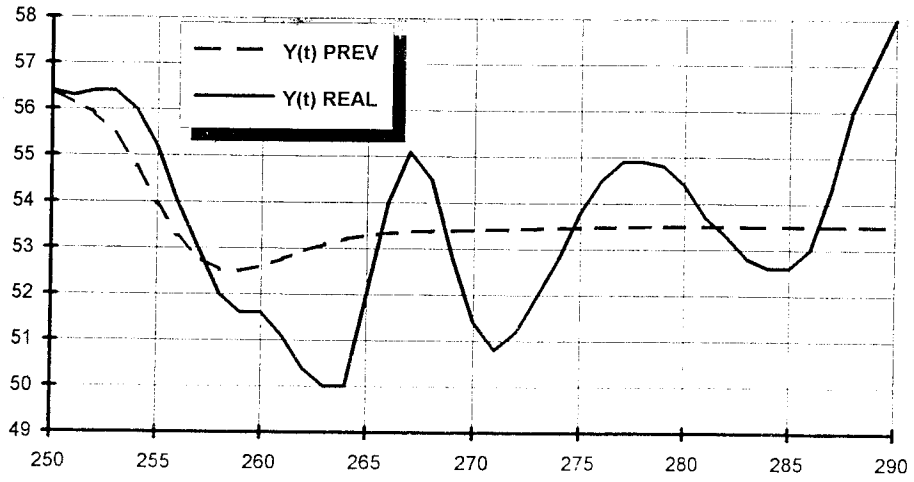


Figura 6.
Previsiones/Valores reales de $Y(t)$.

3. REDES NEURONALES

Las redes neuronales (o redes de neuronas artificiales), son modelos matemáticos simplificados de las redes de neuronas que constituyen el cerebro humano. Estos modelos, están compuestos por un conjunto de "neuronas artificiales" o conjunto de unidades que procesan e intercambian información. Las neuronas de una red, están estructuradas en distintas capas, de forma que una neurona de una capa está conectada con las de la capa siguiente, a las que puede enviar información.

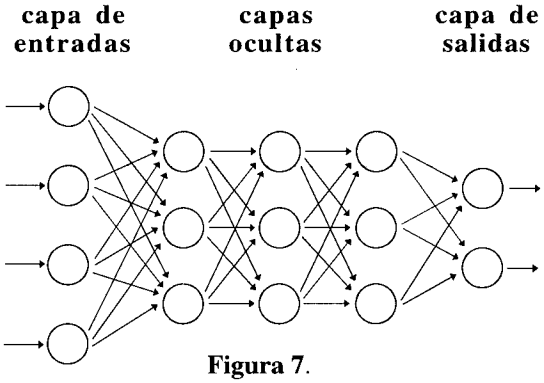


Figura 7.
Arquitectura de una Red Neuronal Artificial.

Tal como se refleja en la figura 7, la arquitectura más habitual consiste en una capa de neuronas de entrada que recibe la información "del exterior", una serie de capas intermedias (u "ocultas") y una capa de salidas, que proporciona "al exterior" el resultado del trabajo de la red.

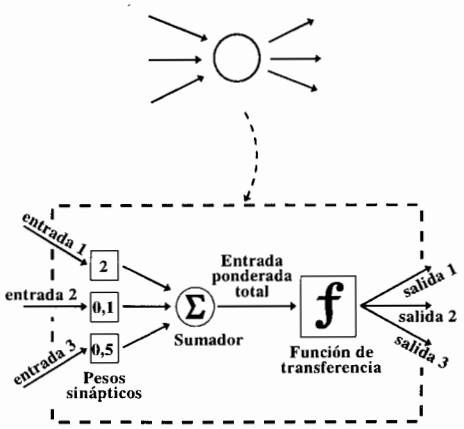


Figura 8.a.

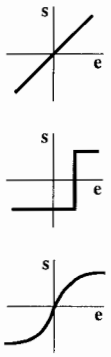


Figura 8.b.

Figura 8.
Funcionamiento de una neurona y tipos de función de transferencia.

Cada neurona, tal como se muestra en la figura 8.a, constituye una "unidad de procesamiento" de información, convierte un conjunto de señales de entrada en una salida que es difundida a las neuronas de la capa siguiente. Esta conversión se realiza en dos etapas: primero, cada una de las señales de entrada es multiplicada por un coeficiente de ponderación ("peso sináptico") atribuido a la conexión; todos los productos son sumados para obtener una cantidad denominada "entrada ponderada total". En una segunda fase, cada unidad utiliza una función de transferencia entrada-salida, o función de activación, que transforma la entrada ponderada total en una señal de salida que es la que se difunde a las neuronas de la capa siguiente. La función de transferencia puede ser de tres tipos, (Lippmann, 1987):

1. ***Lineal***. La actividad de salida es proporcional a la entrada ponderada total.
2. ***De umbral***. La salida queda fija a uno de dos niveles, dependiendo si la entrada ponderada total es mayor o menor que cierto valor crítico denominado "umbral".
3. ***Sigmoide***. La salida varía de forma continua dependiendo de la entrada ponderada total, pero esta dependencia no es lineal.

Habitualmente, se suele utilizar la sigmoide como función de transferencia cuando se trata de aplicar la tecnología de redes neuronales al procesado de señales no-lineales (Lapedes y Farber, 1987), aunque es necesario tener presente que las tres son aproximaciones bastante burdas de la actividad de las neuronas reales.

El proceso de aplicación de la tecnología de redes neuronales artificiales a un problema concreto, consta de tres etapas fundamentales:

1. **Diseño de la Red**. Es necesario decidir la arquitectura que va a tener la red , lo cual implica determinar el número de neuronas de la capa de entradas, el número de capas ocultas y las neuronas que contendrá cada una de ellas, y, por último, el número de neuronas de la capa de salidas. La arquitectura de la red dependerá, como es lógico, del problema concreto que se quiera resolver.

   En el caso que nos ocupa, queremos aplicar la tecnología de redes neuronales artificiales a la previsión de series temporales, basándonos en los datos históricos de la serie, por lo tanto, el número de neuronas de la capa de entrada coincidirá con el número de datos anteriores de la serie que son necesarios para calcular un valor concreto. La capa de salida estará compuesta por una sola neurona cuya salida será el valor de la previsión que queremos calcular. En cuanto al número de capas ocultas y las neuronas de estas capas, no existen reglas fijas que determinen los valores óptimos. Un número muy pequeño de capas ocultas puede hacer que el proceso de entrenamiento se alargue excesivamente, mientras que demasiadas capas ocultas llevan a que se produzca una memorización de los datos, en lugar de la deducción de los patrones que se derivan de los datos, con lo que se falsea la predicción (Richeson y Zimmermann, 1994).

   Por otra parte, el número de neuronas en cada capa oculta, dependerá de la complejidad del problema a estudiar, aunque algunos autores dan ciertas recomendaciones que van desde: la utilización de complicadísimas fórmulas para su cálculo (Zurada, 1991), o indicar que el número de neuronas de la capa oculta debe ser como mínimo el 75% del número de neuronas de entrada (Salchenberger, 1992), hasta aplicar la teoría matemática clásica de Kolmogorov para calcular el número de neuronas como $2k+1$, siendo "k" el número de neuronas de la capa de entrada (Zaremba, 1990).

   Hemos comprobado que el número de neuronas de la capa oculta, siempre que esté entre unos valores mínimo (por ejemplo el 75% de las neuronas de entrada) y máximo (5 veces ese mismo número), sólo influye en la velocidad de entrenamiento de la red, por lo que en nuestro caso hemos partido de un valor inicial $(2k+1)$, que posteriormente se ha ajustado, si durante el proceso de entrenamiento de la red se comprueba que mejora los resultados.

2. **Entrenamiento de la Red Neuronal Artificial**. El entrenamiento de una Red Neuronal consiste en la utilización de un algoritmo (generalmente se usa el al-

goritmo "Back Propagation") para ajustar los pesos sinápticos de las conexiones entre las neuronas.

El proceso consiste en presentar a la red inicial una batería de casos de entrenamiento, que se construyen utilizando los datos reales disponibles (el pasado de la serie temporal), tal como se comenta en el apartado siguiente. Cada uno de estos casos está compuesto por una serie de valores de entrada y el valor de salida correspondiente. Se asignan los valores de entrada a las neuronas de la capa de entradas, y se obtiene al final un valor de salida de la red neuronal. Esta respuesta se compara con la deseada u objetivo, mediante una función de error que da una medida de la eficacia de la configuración actual de pesos sinápticos de la red. El objetivo del aprendizaje es minimizar esta función de error.

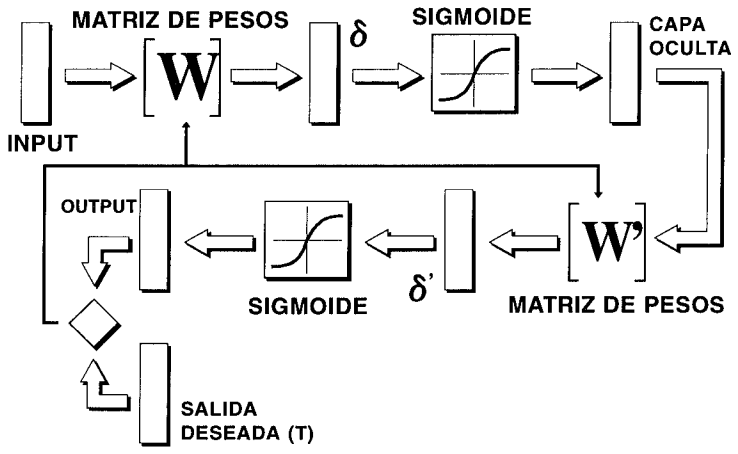


Figura 9.
Algoritmo "Back-Propagation".

Una vez calculado este error, se procede a realizar la fase "hacia atrás" ("backward") variando los pesos sinápticos de los enlaces en función de la magnitud del error cometido y de una constante ε llamada "coeficiente, tasa o ritmo de aprendizaje" que varía entre 0 y 1. La elección de ε es importante, ya que un valor bajo, da como resultado una lenta convergencia pues implicará variar los pesos muy poco, mientras que un valor excesivamente alto, puede conducir a oscilaciones en los pesos de la red, con lo que nunca se alcanzarían los pesos óptimos. Un valor usual de partida para ε es 0,3 aunque puede ser ajustado ligeramente durante el proceso de aprendizaje de la red.

Una vez ajustados los pesos sinápticos, se vuelven a presentar los casos a la red y se vuelve a calcular el error para de acuerdo con él, reajustar los pesos.

Este proceso continuará hasta que el error obtenido esté dentro de unos límites previamente fijados como aceptables, en este momento, el entrenamiento de la red habría finalizado. Una forma sencilla de acelerar el entrenamiento de la red, es añadir un término de momento "***m***" a la hora de ajustar los pesos, que recoja información sobre el último ajuste realizado (Chapman, 1994). Un valor habitual de partida suele ser $m = 0,7$, aunque también puede ser ajustado a cualquier valor entre 0 y 1 durante el proceso de entrenamiento de la red.

3. **Utilización de la Red Neuronal en "Modo Recuerdo"**. Una vez entrenada, la Red Neuronal está en condiciones de ser utilizada. Para ello, no hay más que presentar a la red un caso determinado (por ejemplo los últimos datos disponibles de una serie temporal) para que, utilizando los pesos sinápticos encontrados durante el proceso de entrenamiento, calcule la salida (la previsión del dato siguiente de la serie temporal).

La mayor parte de las aplicaciones de redes neuronales, se realizaron en lenguajes de programación convencionales como FORTRAN ó C. Sin embargo, de forma análoga a lo sucedido con los sistemas expertos, ante la expansión de la tecnología y las dificultades que presentan algunas de la etapas de diseño y entrenamiento de las redes, diversas instituciones y compañías, han empezado a comercializar ***entornos de desarrollo*** o "shells" que facilitan diversos tipos de arquitecturas y algoritmos de entrenamiento. Nosotros, hemos utilizado el paquete comercial denominado "NEUROSHELL" en aplicaciones sobre previsión, aunque existen otros como "ANSIM", "BACK PROPAGATION", etc, para ordenadores personales.

3.1. Modelado Univariante

En este primer caso, hemos tomado la serie UN05 (ya utilizada en este mismo trabajo) que consta de 155 datos, y hemos calculado previsiones utilizando la tecnología de redes neuronales. En el apartado siguiente de este artículo, se comparan estas previsiones con las obtenidas utilizando la metodología de Box-Jenkins.

Siguiendo los pasos indicados anteriormente, la primera fase consiste en el diseño de la red neuronal, es decir, número de neuronas de entrada, número de capas ocultas y número de neuronas en cada una de ellas, así como el número de neuronas de la capa de salidas. Como se ha comentado anteriormente, el número de capas ocultas influye en la velocidad del proceso de entrenamiento, en nuestro caso, teniendo en cuenta el problema que intentamos resolver, es suficiente con una sola capa oculta, ya que las pruebas que se han hecho con dos o tres no mejoraban significativamente los resultados obtenidos con una pero sí aumentaba de forma considerable el tiempo de entrenamiento. Por lo tanto, trabajaremos con redes neuronales del tipo "***d*−*n*−*s***",

llamando "d" al número de neuronas de la capa de entrada, "n" número de neuronas de la capa oculta, y "s" el número de neuronas de la capa de salida.

El valor de "s" es evidente, sólo queremos pronosticar un valor de la serie temporal utilizando para ello una serie de valores anteriores de la misma; por ello sólo habrá una neurona en la capa de salidas, por lo tanto, será una red del tipo "$\boldsymbol{d-n-1}$". El valor de "d", neuronas de la capa de entrada, dependerá del número de valores anteriores de la serie temporal que son necesarios para que la red deduzca un patrón o un modelo, de forma que pueda calcular el valor de salida correspondiente. El cálculo de "d" se realiza mediante pruebas sucesivas con varios diseños y escogiendo aquel con el que se obtienen los mejores resultados.

Se puede empezar con una red $5-11-1$. Cinco neuronas de entrada implica que cada valor de la serie temporal dependerá directamente de los cinco valores anteriores, ponemos 11 neuronas (resultado de aplicar la fórmula $2d+1$) en la capa oculta como número inicial, y una de salida. Esta arquitectura se muestra en la figura 10.

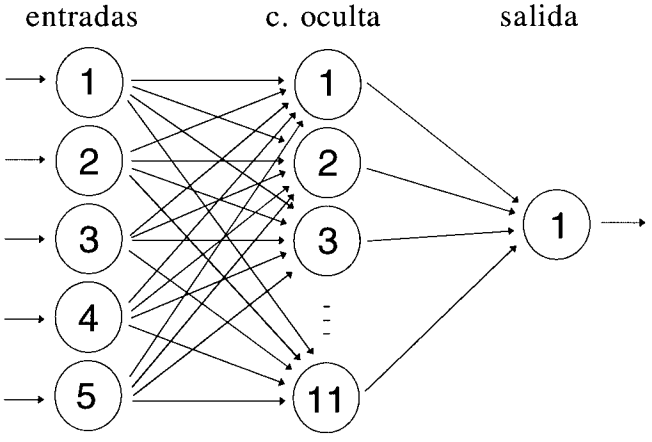


Figura 10.
Arquitectura de la red $5-11-1$.

Para el proceso de entrenamiento de esta red, disponemos de los 134 valores históricos de la serie temporal UN05. En algunas ocasiones, es útil hacer un procesado previo de los datos de entrada antes de presentarlos a la red para el entrenamiento, por ejemplo, para series que presentan grandes variaciones e irregularidades aleatorias, es conveniente trabajar con la serie en logaritmos de los datos de entrada, o efectuar un alisado exponencial en el caso de variaciones aleatorias menores (Kimoto *et al*, 1990). Además, es también útil normalizar los datos de entrada antes de presentarlos

a la red neuronal, esta normalización se utiliza para reducir el rango del conjunto de valores y que sean los apropiados para la función de transferencia o activación que se está usando. Generalmente se normalizan los datos entre 0 y 1 para prevenir el efecto de la saturación de la función de transferencia (sigmoide).

Teniendo en cuenta este diseño de la red, se construye el conjunto de casos de entrenamiento; cada uno de ellos se compone de cinco valores de entrada que determinan un valor de salida. Si llamamos $X1, X2, \ldots, X134$ a los valores históricos (ya normalizados entre 0 y 1) de la serie temporal, se podrían construir 129 casos:

$$
\begin{array}{cccccccc}
X1, & X2, & X3, & X4, & X5 & \longrightarrow & X6 \\
X2, & X3, & X4, & X5, & X6 & \longrightarrow & X7 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
X129, & X130, & X131, & X132, & X133 & \longrightarrow & X134
\end{array}
$$

Con estos 129 casos, comienza el proceso de entrenamiento de la red. Mediante las sucesivas pasadas de los casos por la red, y los ajustes que se van haciendo a los correspondientes pesos sinápticos de los enlaces entre las neuronas, se intenta reducir la diferencia entre la salida real y la salida que genera la red. El proceso termina cuando los errores entre la salidas real y generada por la red entran dentro de un intervalo previamente fijado como aceptable, o bien cuando, transcurrido un período de tiempo considerable, no se observan disminuciones en los errores, que quedan estabilizados en valores fuera del intervalo previamente fijado como aceptable.

Cuando el proceso de entrenamiento ha terminado, la red está preparada para ser utilizada en modo recuerdo, los enlaces entre las neuronas de las distintas capas tendrán unos pesos sinápticos que representan el modelo o patrones de la serie temporal, de forma que se puede pronosticar un valor de salida introduciendo los valores de entrada correspondientes. En nuestro caso, hemos introducido de nuevo los 129 casos para que la red calcule las 129 salidas que componen la serie UN05PR, que podemos comparar con la serie original UN05. Utilizamos para ello el Error Cuadrático Medio (MSE) calculado según la expresión:

$$MSE = \frac{\sum_{i \; 1}^{n} (X_i - S_i)^2}{n}$$

donde X_i son los valores de la serie UN05, S_i los datos pronosticados que pertenecen a la serie UN05PR, y n el total de casos de entrenamiento (en esta primera red, 129).

Ya hemos comentado que la obtención de la arquitectura óptima de la red en cuanto a número de neuronas de entrada, en la capa oculta y en la salida, en un principio sólo es posible mediante un proceso iterativo, entrenando sucesivamente una red tras otra. En nuestro caso hemos comenzado por la red 5-11-1, la red ha sido entrenada utilizando los 129 casos que se pueden construir con los valores históricos

de la serie, y, cuando el entrenamiento ha finalizado (en este caso, se comprobó que después de más de 1.000.000 de pasadas de los casos de entrenamiento por la red, los pesos sinápticos de los enlaces apenas variaban, de forma que la función de error que se intenta minimizar, quedaba estabilizada en un valor determinado), se introdujeron a la red ya entrenada, los mismos 129 casos para que la red dedujera la salida correspondiente para cada uno de ellos, y se calculó el error cuadrático medio (MSE), para comprobar la bondad de la red 5-11-1.

El mismo proceso se ha seguido para entrenar una segunda red, con la arquitectura 6-13-1 (cada valor de la serie UN05 depende de los seis valores anteriores), 7-15-1, 8-17-1, etc. (el número de neuronas de la capa oculta se ha calculado siempre utilizando la expresión $2d+1$, ya comentada), hasta la red 15-31-1, que ha sido la última que se ha entrenado. Todas ellas han utilizado un mínimo de 1.000.000 de pasadas con el programa "NeuroShell", hasta que se llegaba a un instante en el que la función de error se estabilizaba. En ese momento, se considera la red como suficientemente entrenada y apta para realizar previsiones. Después de calcular la serie UN05PR que pronosticaba cada red, se calculó el error cuadrático medio (MSE), obteniéndose los valores que se recogen en la tabla 1.

Tabla 1

Error Cuadrático Medio

Número de neuronas de entrada	**Arquitectura de la red entrenada**	**Número de casos de entrenamiento**	**Error Cuadrático Medio MSE**
5	5-11-1	129	0.00525371
6	6-13-1	128	0.00321710
7	7-15-1	127	0.01257860
8	8-17-1	126	0.00430616
9	9-19-1	125	0.00519821
10	10-21-1	124	0.00578450
11	11-23-1	123	0.01380814
12	**12-25-1**	**122**	**0.00047366**
13	13-27-1	121	0.00101259
14	14-29-1	120	0.00191643
15	15-31-1	119	0.00364024

Tabla 2

Pesos sinápticos finales de la Red 12-25-1

		NEURONAS DE LA CAPA DE ENTRADAS												SALIDA
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**1**
NEURONAS CAPA OCULTA	**1**	0.10	−0.18	0.25	0.14	0.67	0.55	0.55	0.32	0.15	−0.52	−0.42	0.78	0.54
	2	0.39	−0.25	−0.01	0.10	0.59	0.71	0.60	0.79	0.26	−0.16	−0.03	0.44	0.54
	3	0.20	−0.10	−0.10	0.20	0.72	0.31	0.27	0.00	−0.08	−0.64	0.24	0.83	0.42
	4	0.04	−0.55	−0.25	0.32	0.68	0.34	0.25	−0.22	0.16	−0.31	0.34	1.13	0.20
	5	0.46	0.33	0.23	0.34	0.50	0.18	0.40	−0.11	0.53	0.58	0.48	0.32	−0.62
	6	−0.23	−0.50	−0.86	0.41	0.60	0.81	−0.36	−0.51	0.07	−0.27	−0.25	−1.44	0.01
	7	0.46	0.32	0.20	0.20	0.45	−0.01	0.55	0.31	0.24	0.27	0.64	0.07	−0.18
	8	1.57	2.30	2.10	−1.18	−2.15	−1.71	−1.12	0.24	0.24	1.73	1.69	3.15	1.85
	9	0.54	1.99	0.78	−0.74	−1.49	−0.64	−0.51	0.92	−0.14	0.32	0.08	−3.93	−2.33
	10	0.12	0.81	0.67	−0.02	−0.13	−0.07	0.22	0.06	0.33	0.45	0.66	−1.44	−0.65
	11	0.29	0.61	0.40	0.03	−0.64	−0.14	0.07	0.62	0.33	0.87	0.67	−1.47	−0.58
	12	−0.32	−0.29	−0.18	−0.06	0.82	0.63	0.76	0.12	−0.57	−0.96	−0.07	1.18	0.91
	13	−0.06	−0.96	−0.26	0.25	0.57	0.40	0.63	−0.36	−0.53	−0.50	−0.14	1.84	0.60
	14	−0.30	−0.50	−0.93	0.30	0.75	0.46	−0.18	−0.62	0.11	0.05	−0.28	−1.27	−0.14
	15	0.48	0.68	0.19	0.46	−0.02	0.20	0.00	0.30	0.41	0.68	0.53	−0.88	−0.55
	16	0.38	0.44	0.31	0.39	0.27	0.46	0.39	0.61	0.06	0.27	0.41	−0.34	−0.02
	17	0.24	0.31	0.31	0.50	0.13	0.30	0.22	0.55	0.59	0.14	0.13	0.42	0.02
	18	0.35	−0.26	0.23	0.45	0.26	0.07	0.34	0.09	−0.27	−0.36	0.35	0.78	0.31
	19	−0.32	−0.49	−0.37	0.22	0.73	0.99	0.94	0.63	−0.51	−0.88	−0.52	0.82	1.26
	20	−0.08	0.49	−0.14	0.21	0.49	0.61	0.85	0.95	0.09	−0.29	−0.10	0.31	0.84
	21	0.32	0.30	0.42	0.01	0.04	0.39	0.39	0.32	0.37	0.63	0.52	−0.64	−0.29
	22	0.33	0.06	0.31	0.20	0.49	0.32	0.24	0.03	0.14	−0.21	0.40	0.53	0.07
	23	0.26	0.22	0.62	0.68	0.55	0.19	0.15	0.15	0.21	1.01	0.36	−0.16	−0.80
	24	0.36	0.06	0.45	0.48	0.33	0.55	0.56	0.51	0.35	0.29	0.39	−0.14	−0.04
	25	0.54	0.36	0.10	0.42	0.28	0.24	0.09	0.32	0.16	1.04	0.40	−0.32	−0.65

Tal como se puede comprobar en la tabla, el error cuadrático medio más bajo se obtiene para la red 12-25-1 (cada valor de la serie depende directamente de los doce valores anteriores). Este resultado parece lógico ya que la serie UN05 tiene una estacionalidad de orden doce, tal como se comprobó al hacer el estudio de la serie para calcular el modelo según Box-Jenkins. Una vez que se ha decidido que el valor óptimo de neuronas de entrada es 12, se puede intentar afinar más, variando ligeramente la tasa de aprendizaje, ε, y/o el momento, m. Se puede incluso intentar aumentar o disminuir ligeramente el número de neuronas de la capa oculta para comprobar si se reducen aún más los errores o se produce una convergencia más rápida de los pesos sinápticos de los enlaces. En este caso concreto se ha comprobado que se produce una ligerísima mejoría si se ajusta como tasa de aprendizaje el valor $\varepsilon = 0.3$ y el momento a $m = 0.2$, mientras que no se observa mejoría apreciable alguna variando el número de neuronas de la capa oculta. Los valores de los pesos sinápticos definitivos después de todo el proceso de entrenamiento de la red neuronal 12-25-1, son los que aparecen recogidos en la tabla 2, en la que se puede ver el peso sináptico del enlace entre cada una de las 25 neuronas de la capa oculta y cada una de las 12 neuronas de la capa de entradas, y la única neurona de la capa de salidas.

En la figura 11, se puede comprobar el ajuste de la serie UN05 deducida por la red neuronal 12-25-1, con la serie UN05 real.

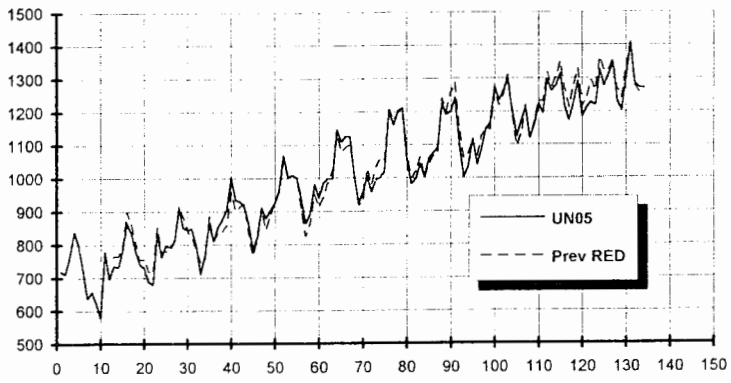


Figura 11.
Serie calculada por la Red 12-25-1 y serie UN05 real.

Una vez finalizado el proceso de entrenamiento de la red neuronal, se pasa al cálculo de previsiones, utilizando la red en "modo recuerdo", es decir, se introducen los nuevos casos a la red con los pesos sinápticos ya definitivos que se han ajustado durante el proceso de entrenamiento de la misma. En concreto, se introducen 20 casos para que la red calcule las previsiones $S135, S136, \ldots, S154$, y que representamos de la siguiente forma:

$X123\ X124\ X125\ X126\ X127\ X128\ X129\ X130\ X131\ X132\ X133\ X134$ ¿S135?
$X124\ X125\ X126\ X127\ X128\ X129\ X130\ X131\ X132\ X133\ X134\ X135$ ¿S136?
$X125\ X126\ X127\ X128\ X129\ X130\ X131\ X132\ X133\ X134\ X135\ X136$ ¿S137?

$X142\ X143\ X144\ X145\ X146\ X147\ X148\ X149\ X150\ X151\ X152\ X153$ ¿S154?

De esta manera de obtienen los 20 valores que aparecen en la figura 12 junto con los valores reales de la serie UN05.

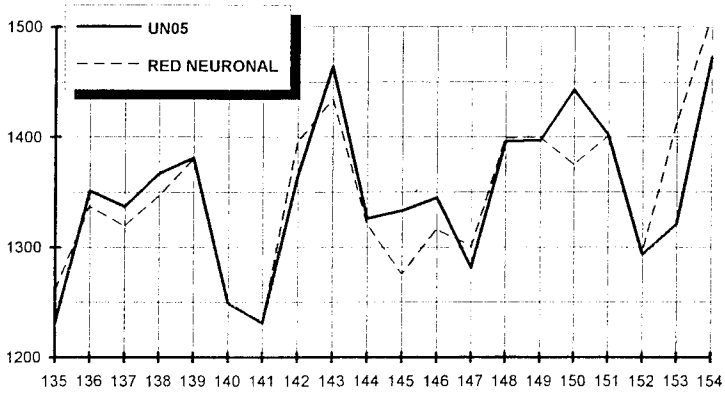


Figura 12.
Valores reales/Previsiones RED NEURONAL.

3.2. Modelado de Funciones de Transferencia

Nuestro objetivo es pronosticar los valores de una serie $Y(t)$ que dependen directamente de los de otra serie $X(t)$. Como vimos en el apartado 2.2, utilizamos el ejemplo del horno, en el cual tenemos como datos 296 valores históricos de las series $X(t)$ e $Y(t)$, y que están representados en la figura 5, de los cuales los primeros 250, nos servirán para entrenar a la red neuronal.

Al final del proceso, el modelo de red neuronal, hará las veces de función de transferencia entre $X(t)$ e $Y(t)$, de forma que para cada valor o serie de valores de $X(t)$, se tendrá un valor de $Y(t)$. Utilizaremos los últimos 46 datos de Y, para comprobar la bondad de las previsiones que se calcularon mediante la red neuronal, y las obtenidas utilizando la metodología de Box-Jenkins.

La red neuronal que se va a construir será de tipo $n-d-1$, es decir, será una red con "n" nodos de entrada, "d" nodos en la capa oculta y 1 nodo de salida. Los valores óptimos de n y d, se obtienen haciendo pruebas de la misma forma que en el caso univariante. El objetivo, es reducir el valor del error cuadrático medio (MSE).

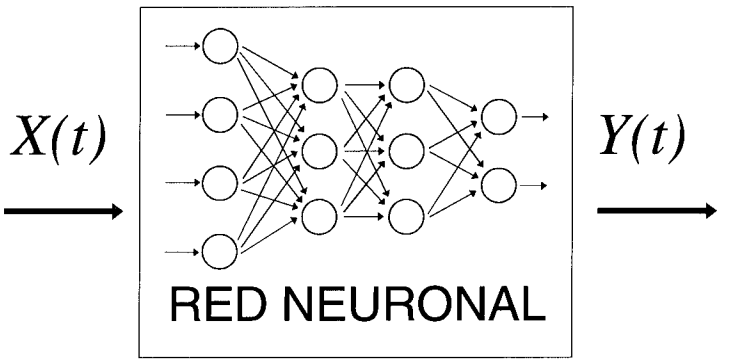


Figura 13.
Red Neuronal como función de transferencia.

La primera red que vamos a entrenar, es la que llamaremos RED 30, que indica que en cada caso utilizado para el entrenamiento, se utilizan 3 valores de la serie X, el del instante t, el de $t-1$ y el de $t-2$ (X_t,$X_{t\ 1}$ y $X_{t\ 2}$), para pronosticar el valor de Y retrasado cero períodos (Y_t). En este primer caso se dispone de un conjunto de 248 casos de entrenamiento:

$$
\begin{array}{llll}
X1, & X2, & X3 & \longrightarrow Y3 \\
X2, & X3, & X4 & \longrightarrow Y4 \\
\multicolumn{4}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\
\multicolumn{4}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\
X248, & X249, & X250 & \longrightarrow Y250
\end{array}
$$

Se seguirá un proceso similar al llevado a cabo en el modelado univariante para intentar llegar a la arquitectura óptima de la red neuronal, con una diferencia, además de ir aumentando el número de neuronas de la capa de entradas, se tendrá en cuenta también la posibilidad de que la respuesta pueda estar retrasada algún período respecto a la entrada. Según esto, probaremos con redes de 3, 4 o 5 neuronas en la capa de entrada, y para cada una de ellas, consideraremos que la salida puede estar retrasada

desde 0 a 4 períodos. Los resultados del proceso de entrenamiento de las diferentes redes neuronales aparecen recogidos en la tabla 3.

Tabla 3

Error Cuadrático Medio para diferentes configuraciones de la red neuronal.

		NÚMERO DE NEURONAS DE LA CAPA DE ENTRADAS		
		3	**4**	**5**
NÚMERO DE PERÍODOS DE RETRASO	**0**	0.02285230	0.01333500	0.00728100
	1	0.01524960	0.00556036	0.00234480
	2	0.00298400	0.00312900	0.00076860
	3	0.00324960	0.00117510	0.00074710
	4	0.00203520	0.00220270	0.00431380

En dicha tabla se observa que el valor mínimo del error cuadrático medio se obtiene para la RED53 (se trata de una red con 5 neuronas en la capa de entrada y en la que la salida se retrasa 3 períodos), el diseño de la red aparece en la figura 14. Una vez escogida la arquitectura idonea, se fue variando el número de neuronas de la capa oculta desde las 11 que se consideraron inicialmente (resultado de aplicar la expresión $2d+1$), hasta el 14, número de neuronas con el que se observó un menor valor del error, después de un tiempo considerable de entrenamiento. Los valores de los pesos sinápticos finales entre los enlaces de la red53, aparecen recogidos en la tabla 4.

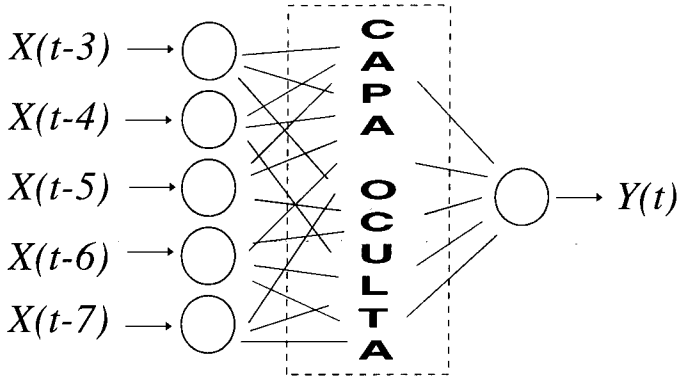


Figura 14.
Arquitectura de la RED53.

Tabla 4

Pesos sinápticos finales de la red 5-14-1.

		CAPA DE ENTRADAS					SALIDA
		1	**2**	**3**	**4**	**5**	**1**
		$X(t-3)$	$X(t-4)$	$X(t-5)$	$X(t-6)$	$X(t-7)$	$Y(t)$
	1	0.46	0.69	0.60	1.03	−0.46	−0.07
C	**2**	−6.13	−1.24	−0.12	0.46	1.34	2.53
A	**3**	0.57	1.10	0.24	1.37	−0.40	0.37
P	**4**	1.75	1.84	3.02	−0.27	0.16	−1.87
A	**5**	1.18	−0.73	0.51	−0.98	4.68	−1.39
	6	0.75	0.77	0.80	0.68	1.67	0.13
	7	0.78	0.55	0.29	0.89	−0.52	−0.05
O	**8**	0.64	0.55	0.90	0.95	0.90	0.18
C	**9**	0.41	0.60	0.69	1.29	0.10	0.32
U	**10**	0.85	1.19	0.31	0.87	−0.14	0.54
L	**11**	0.64	0.60	1.02	1.03	0.85	0.22
T	**12**	2.43	−2.12	−4.92	0.18	−0.86	2.69
A	**13**	−0.74	−0.95	−1.18	−0.17	−0.39	0.31
	14	−2.65	−2.05	−1.58	0.02	−0.17	1.27

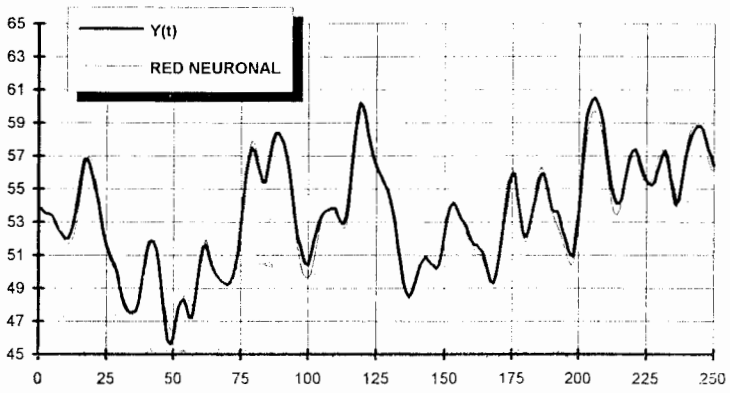


Figura 15.
Serie $Y(t)$ real y serie deducida por la red53.

En la figura 15, se puede comprobar el ajuste conseguido por la red neuronal con respecto a la serie $Y(t)$ original. Si ahora se calculan las previsiones introduciendo a la red53 entrenada, los 20 casos que proporcionan como salida las previsiones de $Y251$ a $Y290$:

$$
\begin{array}{lllllll}
X244, & X245, & X246, & X247, & X248 & \longrightarrow & ¿Y251? \\
X245, & X246, & X247, & X248, & X249 & \longrightarrow & ¿Y252? \\
X246, & X247, & X248, & X249, & X250 & \longrightarrow & ¿Y253? \\
\ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\
X283, & X284, & X285, & X286, & X287 & \longrightarrow & ¿Y290?
\end{array}
$$

El resultado de las previsiones aparece representado en la figura 16.

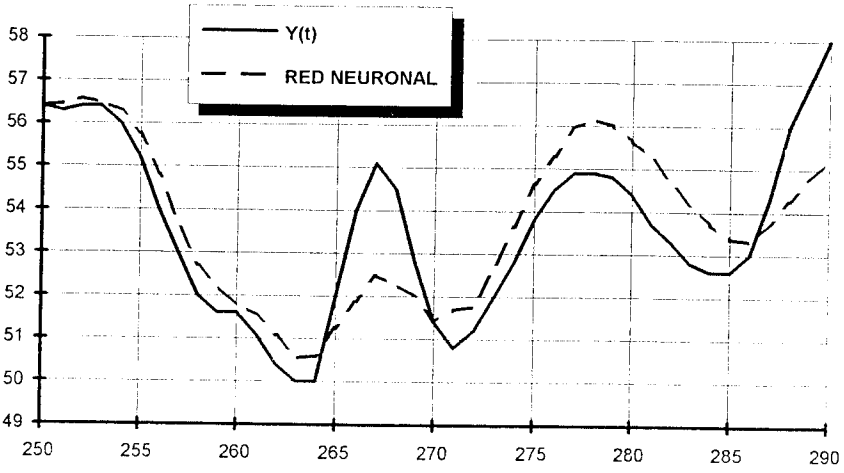


Figura 16.
Previsiones de la serie $Y(t)$ calculadas por la red53.

4. ANÁLISIS COMPARATIVO

En la figura 17, aparecen representadas las previsiones calculadas a partir de la red neuronal junto con las previsiones obtenidas utilizando la metodología de Box-Jenkins, comparadas con los valores reales de la serie UN05. Se puede comprobar que las previsiones calculadas por ambos métodos son bastante aceptables ya que parecen seguir a la serie real de una forma suficientemente correcta. En la tabla 5, están recogidos los valores del error cuadrático medio MSE, que han sido calculados comparando las previsiones con los valores reales de la serie UN05 para el modelado univariante.

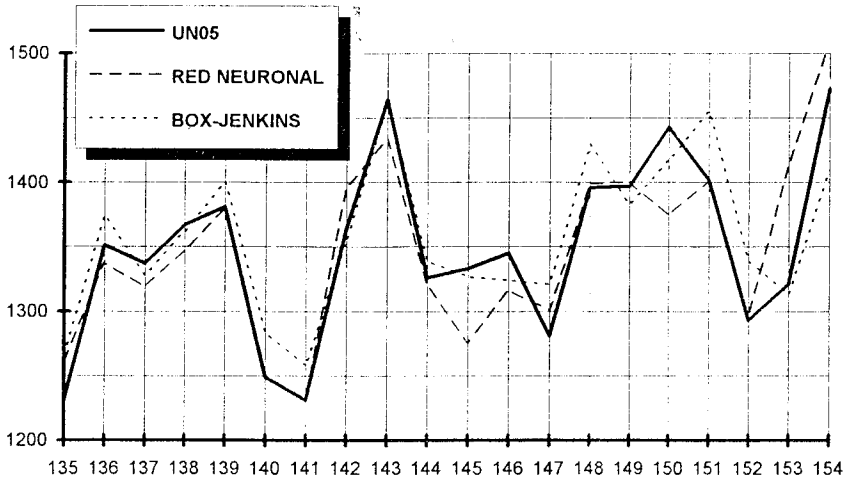


Figura 17.
Comparación de Previsiones en el modelado univariante.

Tabla 5
Valores del MSE de las previsiones.

	RED NEURONAL	BOX-JENKINS
MODELADO UNIVARIANTE	1.152,8852	876,8597
FUNCIÓN DE TRANSFERENCIA	1,3712	2,7615

Un aspecto destacable de las previsiones calculadas por ambos métodos, y que se puede comprobar en la figura 17, es que, en la mayoría de las ocasiones, ambos métodos aciertan en el cálculo de los llamados "turning points" de la serie, es decir, aquellos puntos en los que la serie cambia de tendencia. Este hecho es muy interesante puesto que en muchas ocasiones, más que obtener unas previsiones en las que una medida del error (como puede ser el error cuadrático medio que hemos calculado nosotros), sea muy pequeña en términos relativos, es mucho más importante averiguar con exactitud los puntos en los que la serie cambia de tendencia, aunque luego el error que se cometa sea mayor. En cuanto a este aspecto, se observa que tanto Box-Jenkins como la Red Neuronal aciertan bastante bien, sobre todo en la primera mitad de las previsiones calculadas.

En la tabla 5, se comprueba que, aunque el método de Box-Jenkins consigue unas previsiones mejores que las de la red neuronal en cuanto al error cuadrático medio, las previsiones de la red neuronal son igualmente aceptables. Sin embargo, hay que destacar el hecho de que cuando se estudió la arquitectura óptima de la red neuronal, se utilizó información conseguida durante la etapa de modelado de Box-Jenkins, por ejemplo, durante la etapa de modelado de Box-Jenkins se comprobó que se trataba de una serie con una estacionalidad de 12 períodos. Esta información se utilizó en la etapa de diseño de la red neuronal ya que se supuso que una red en la que cada valor depende de como mínimo de 12 valores anteriores, debería dar, en principio, mejores resultados que cualquier red con menos neuronas de entrada, como así ha resultado.

Este aspecto fue aún más importante en el caso de la red neuronal que tenía que funcionar como función de transferencia en el caso del horno. No existen de momento, reglas más o menos fijas que se puedan utilizar a la hora de diseñar la arquitectura de la red neuronal, por lo que la forma de llegar a una estructura que sea óptima, no consiste en otra cosa que ir probando red tras red hasta dar con la que pueda ser buena. Tal como se puede comprobar en la figura 18 y en la tabla 5, cuando se encuentra una red aceptable, las previsiones que se obtienen, son bastante mejores que las obtenidas por Box-Jenkins, el error cuadrático medio de las previsiones de la red es prácticamente la mitad del error cuadrático de las previsiones de Box-Jenkins.

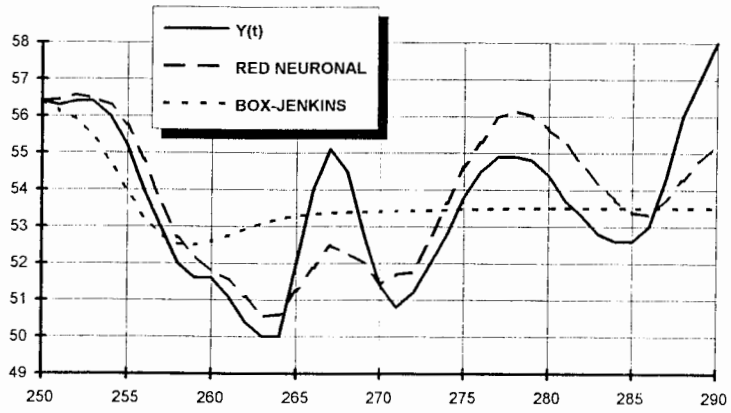


Figura 18.
Comparación de las previsiones en el caso de función de transferencia.

Sin embargo, en este segundo caso, fue todavía más difícil dar con la arquitectura óptima de la red. Gracias al estudio de las series que se ha de hacer para utilizar el método de Box-Jenkins, se pudo saber que la salida respondía a las variaciones de la entrada tres períodos más tarde. Luego, en la etapa de diseño de la red neuronal, se fueron probando redes en las que la salida aparecía retrasada cero, uno, dos, tres

y cuatro períodos con la esperanza de que la red en la que la salida estaba retrasada tres períodos fuera la que diera los mejores resultados, lo cual se pudo comprobar posteriormente con el cálculo del error cuadrático medio. Es decir, se llevó a cabo una "búsqueda dirigida" de la mejor red neuronal, o al menos de una buena red neuronal. De no haberlo hecho así, el proceso de búsqueda de la red podría llegar a ser ciertamente complicado porque podría eternizarse a no ser que, en un golpe de suerte, se diera con la red óptima.

5. CONCLUSIONES

Se ha hecho un estudio comparativo del cálculo de previsiones mediante Box-Jenkins y Redes Neuronales Artificiales. Aunque debería de comprobarse los resultados con muchos más casos, podemos intuir que se obtienen resultados aceptables con las dos metodologías para previsiones univariantes de series temporales.

En cambio, parece ser mejor el cálculo de previsiones de funciones de transferencia por Redes Neuronales, pero la dificultad de encontrar la red adecuada para cada caso es una limitación importante de esta aproximación. Hemos encontrado que este problema puede ser resuelto, al menos en parte, gracias a la información que se obtiene de la aplicación de la sistemática metodología de Box-Jenkins, que resulta ciertamente útil a la hora de hacer una "búsqueda dirigida" de la arquitectura adecuada de la red, y evita el tener que entrar en procesos de tipo "prueba y error" de forma indiscriminada.

REFERENCIAS

[1] **Alonso, G.** y **Becerril, J.L.** (1993). *Introducción a la Inteligencia Artificial.* Multimedia Ediciones S.A.

[2] **Box, G.E.P.** y **Jenkins, G.M.** (1976). *Time Series Analysis: Forecasting and Control.* 2nd. Ed. San Francisco: Holden Day.

[3] **Chapman, A.J.** (1994). "Stock market trading systems through neural networks: developing a model". *Int. Journal of Applied Expert Systems.* **12, 2.**

[4] **Eberhart, R.C.** y **Dobbins, R.W.** (1990). *Neural Network PC Tools.* Academic Press.

[5] **Farmer, J.D.** y **Sidorowich, J.J.** (1987). "Predicting chaotic time series". *Physical Review Letters*, **59**, 845–848.

[6] **Granger, C.W.J.** y **Newbold, P.** (1986). *Forecasting economic time series.* (2nd. ed.). Orlando, FL: Academic Press.

[7] **Kimoto, T., Asakawa, K., Yoda, M.** y **Takeoka, M.** (1990). "Stock market prediction system with modular neural networks". *IJCNN Int. Joint Conference on Neural Networks.* San Diego, CA.

[8] **Lapedes, A.** y **Farber, R.** (1987). "NonLinear signal processing using neural networks". *IEEE Conference on Neural Information Processing System - Natural and Synthetic.*

[9] **Lippmann, R.P.** (1987). "An Introduction to Computing with Neural Nets". *IEEE ASSP Magazine.* April 1987.

[10] **Neuroshell "Neural Network Shell Program".** (1989). *Ward Systems Group Inc.*

[11] **Priestley, M.B.** (1988). *Non-Linear and non-stationary time series analysis.* San Diego, CA: Academic Press.

[12] **Reilly, D.P.** (1980). "Experiences with an Automatic Box-Jenkins Modeling Algorithm". In *Time Series Analysis.* Ed. O.D. Anderson. (Amsterdam: North Holland), pp. 493–508.

[13] **Richeson, L.** y **Zimmermann, R.A.** (1994). "Predicting consumer credit performance: can neural networks outperform traditional statistical methods?". *Int. Journal of Applied Expert Systems.* **2, 2**, February, 1994.

[14] **Salchenberger, L.M., Cinar, E.M.** y **Lash, N.A.** (1992). "Neural Networks: a new tool for predicting thrift failures". *Decision Sciences.* July-August, 1992.

[15] **Tiao, G.C.** y **Box, G.E.P.** (1983). *An Introduction to Applied Multiple Time Series Analysis.* Illinois, USA: Scientific Computing Associated.

[16] **Tiao, G.C.** y **Tsay, R.S.** (1989). "Model specification in multivariate time series". *Journal of the Royal Statistical Society.* **B 51**, 157–213.

[17] **Tong, H.** (1983). "Threshold models in non-linear time series analysis". *Lecture Notes in Statisti.* **21**. New York: Springer-Verlag.

[18] **Tong, H.** (1990). *Non-Linear time series: A dynamical system approach.* Oxford: Oxford University Press.

[19] **Zaremba, T.** (1990). *Technology in Search of a Buck.* Neural Network PC Tools. Academic Press Inc.

[20] **Zurada, J.** (1991). *Using Nworks: An Extended Tutorial for Neural Works Proffessional II/Plus and NeuralWorks Explorer.* Pittsburg.

ENGLISH SUMMARY:

COMPARATIVE ANALYSIS OF UNIVARIATE AND TRANSFER-FUNCTION FORECASTING USING NEURAL NETWORKS AND BOX-JENKINS TECHNIQUES

David de la Fuente García and Raúl Pino Díez

Time series forecasting is a very important statistical tool in studying time dependent data behaviour and in predicting future values.

A time series is a set of data measured in discrete or continous time units. A multivariate time series consists of a set of data from several time dependent variables. A special case is when the measured variables are significantly correlated, for instance, when similar attributes are being measured at distinct geographical locations. When forecasting a new value for each variable, the best prediction is achieved by taking into account the variations of the other variables.

Traditionally speaking, and especially during the 1980's, there were many techniques available for time series analysis, which assumed linear relationships between variables (Box- Jenkins, 1976). But in real life situations, data do not have such simple relationships, and it is difficult to make good forecasts. This is the reason why it seems necessary to build non-linear models to analize time series data in real cases.

Linear time series models also present certain disadvantages, such as not explaining sudden changes over a very wide range during irregular time intervals (Tong, 1983).

Further studies (Tiao and Tsay, 1989) analyzed other problems involved in multivariable modelling. In order to solve them, non-linear statistical models were used, such as the "threshold" and "bilinear" models suggested by Tong (1990). Alternatively, Granger and Newbold (1986) suggested using non-linear transformations of the original data before traditional linear modelling. Farmer and Sidorowich (1987) significantly improved chaotic time series predictions by using local approximations methods.

Although in the past decade major advances have been made, non-linear modelling is, unfortunately, an extremely hard task, due to the simplifications made in the modelling step, such as omitting parameters which are unknown or which are expected not to have a direct effect on the data.

In order to achieve good forecasts, we have hence looked to artificial Neural Networks as an alternative for calculating quantitative predictions.

In this work, we present a overview of the Box-Jenkins methodology for univariate time series modelling and transfer functions with one input and one output. Subsequently, we describe the Neural Networks approach and then go on finally to compare both methods with a number of examples. From these examples, it can be seen that the forecasts from both methods are quite acceptable, because they follow the actual series in a sufficiently correct way.

A key aspect of forecasts obtained from both methods is that in most cases both of them succeed in calculating the series turning points, i.e., those observations at which the series trend changes. This is very important since it is often better to try to find out the exact observations at which the series trend changes, rather than to obtain a good forecast in terms of a relatively small error. With regards to this objective, it can be seen that both Box-Jenkins and Neural Networks techniques are fairly accurate, above all in the first half of the forecasts.

When the optimal architecture for the net was studied, information obtained from Box- Jenkins modelling step was used. For example, during the Box-Jenkins modelling step, it was proven that the series had an order 12 seasonality. This information was used in the design step of the nets because it was assumed that a 12-period related net would have better performance than any other net.

This aspect is even more important in the case of the net which has to work as a transfer function. Presently, there are no rules for designing the architecture of Neural Networks, so the best way to obtain an optimum structure is by testing net by net until we find the one which may be good. Thanks to the series analysis that has to be done in order to use the Box-Jenkins method, we were able to ascertain that the output followed the input variations with a 3-period delay. Later, in the design step of the Neural Net, we tested nets in which the output was delayed in zero, one, two, three and four periods. This was in the hope that the one with a 3-period delay was the best of all. We then proved this by observing the MSE. Thus, it was a "guided search" of the best net or, at least, of a good one.

In conclusion, we can state that, although the results have to be checked against many more experiences, good performances are obtained with both methods for univariate time series forecasting. On the other hand, Neural Networks seem to be better in transfer function forecasting, although a handicap for this method is the difficulty of finding the adequate net for each case. We found a partial solution to this problem in the application of the systematic Box- Jenkins method which is very useful when carrying out a "guided search" of the appropiate architecture for the net, thus avoiding indiscriminate "trial and error" processes.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 217-231, 1995

EL SISTEMA ADEST

J.M CARIDAD OCERIN* y R. ESPEJO MOHEDANO*

ADEST is an intelligent software enviroment, user friendly, oriented toward categorical data analysis. It's based on a generalization of Havranek procedure for log-linear model especification. It includes also a semi-automatic module to agregate categories whith low marginal expected frequencies, and several tools like a categorical data editor and a generator of BMDP programs.

ADEST System.

Key words: Categorical Data Analysis, Log-linear models, Multiple contingency tables.

1. INTRODUCCIÓN

La utilización de variables cualitativas es un hecho cada vez más frecuente en el ámbito de las Ciencias Sociales y Experimentales: análisis de encuestas, tratamiento de historias clínicas o epidemiológicas, y otras situaciones en análisis de datos.

Los investigadores y profesionales en estos campos no suelen ser expertos en Estadística, y en todo caso es más frecuente que no conozcan en profundidad las técnicas relativas a datos categorizados, por lo que se hace necesario la consulta a un estadístico profesional en las distintas fases del análisis de datos. Los desarrollos acaecidos en los últimos años en las aplicaciones software convencionales, sistemas potenciados y de inteligencia artificial en la construcción de sistemas expertos, permite, a veces, abordar la consultoria estadística de una forma sistemática hasta llegar a la elaboración de un logical de ayuda y apoyo al usuario final, disminuyendo la carga del consultor estadístico y facilitando la labor de éste.

* J.M. Caridad Ocerin y R. Espejo Mohedano. Departamento de Estadística. Escuela Superior de Ingenieros Agrónomos y Montes.
Avda. Menéndez Pidal s/n.
Apdo. 3048, 14080 Córdoba. Spain.

- Article rebut el febrer de 1994.
- Acceptat el setembre de 1994.

Tales herramientas son tremendamente poderosas para cualquier investigador ya que no sólo puede realizar un determinado análisis de forma más rápida, sino que pueden repetirlos tantas veces como se desee, quizás con una sola pequeña diferencia en el conjunto de parámetros. El ordenador ha cambiado la forma de pensar del estadístico en muchas áreas de la Estadística, en particular sobre ajuste de modelos. Según palabras de Hand (1992), estas ventajas también suponen un cierto peligro: el sobreajustar modelos es un peligro real, no en la forma convencional de sobreparametrizar un modelo, pero sí una forma más sutil e insidiosa de intentar ajustar muchos modelos. Existen métodos estadísticos para reconocer sobreparametrización, pero no para esta segunda clase de ajustes que quizás representa un problemas más serio. Por tanto, es necesario poner a disposición de usuarios de análisis de datos, un conjunto de herramientas inteligentes que eviten el uso inapropiado, y desgraciadamente muy frecuente, de los paquetes estadísticos, que dirijan su utilización e interpretación, lo que conlleva a la automatización de los procesos de decisión y selección de estrategias de análisis de datos. El disponer de tales herramientas supone un apoyo importante y cada vez más necesario en la utilización correcta de los métodos estadísticos y sus limitaciones.

Los modelos log-lineales constituyen un tipo de modelos utilizados en análisis de datos de tipo cualitativo. Además de representar la asociación entre los factores o variables en estudio, permiten cuantificar y determinar la influencia de los factores sobre las frecuencias de una tabla multidimencional, así como la frecuencia de efectos conjuntos de varios factores (interacciones) en cada celda de la tabla. En análisis exploratorio de datos, uno de los objetivos fundamentales es encontrar todos los modelos que describan de forma adecuada a los datos. Una primera aproximación para este problema, que puede utilizarse para tablas de contingencia pequeñas (menos de 3 o 4 factores), es ajustar todos los posibles modelos log-lineales jerárquicos y después utilizar alguna medida estadística, tipo AIC (Akaike 1973,1987) o BIC (Raftery 1986), para seleccionar aquellos modelos log-lineales que representen de forma óptima a los datos. Para tablas de más de 4 dimensiones, el problema empieza a dificultarse por la complejidad computacional que conlleva. Con 4, 5 y 6 variables, el número de modelos log-lineales jerárquicos posibles es 168, 7.581 y 7.828.354 respectivamente. De forma natural, mientras mayores términos de interacción contenga un modelo, mayor será la complejidad en la interpretación de éste; incluso cuando se consideran sólo modelos de orden 2 o menor, 5 y 6 variables producen 1.451 y 40.070 modelos log-lineales respectivamente. Esta complejidad computacional estriba no sólo en la generación y mantenimiento de estos modelos sino también a la hora de ajustarlos.

Muchos procedimientos de selección de modelos son capaces de seleccionar un único modelo a través de una serie de tests de significación sobre parámetros o grupos de parámetros de modelos. Una revisión de la mayoría de estos procedimientos puede consultarse en Wrigley (1985). Los modelos seleccionados mediante este tipo de técnicas difieren a menudo dependiendo del procedimiento elegido, incluso uti-

lizando el mismo test estadístico. Según palabras de Tukey (1985) "La Ciencia es el resultado de trabajar con múltiples hipótesis", lo que incita a adoptar métodos que proporcionen respuestas múltiples. Las aproximaciones de optimización, con un criterio bien definido de optimabilidad de modelos, son más consistentes puesto que el criterio clasifica sin reparar en el método de búsqueda.

Este tipo de aproximaciones son computacionalmente más complejas que las basadas en test de significación sobre parámetros de modelos. Edwards y Havranek (1984, 1987) han desarrollado un procedimiento de búsqueda de modelos log-lineales jerárquicos basado en un criterio de optimización, y que en este trabajo se generaliza introduciendo modificaciones necesarias en procesos de modelización, selección de un punto inicial de partida en el procedimiento, además de dotarlo de elementos de decisión necesarios para su completa automatización.

Otro problema frecuente en el análisis estadístico en general, y en tablas de contingencia y modelos log-lineales en particular, radica en un tamaño muestral pequeño y que puede provocar obtener tablas dispersas, es decir, tablas en donde aparecen ciertas combinaciones de celdas con frecuencias bajas. Este tipo de tablas presentan graves inconvenientes tanto a nivel teórico, al no poder garantizar la convergencia de algunos estadísticos usuales, como computacionales, al tener que reproducir algunas frecuencias esperadas cero.

Poco puede hacerse en este tipo de problemas. La solución obvia sería buscar más datos, pero, desgraciadamente, en la mayoría de los casos esto es inviable. Tradicionalmente se han añadido constantes a valores muestrales, pero estas prácticas no son recomendables para todos los modelos y todas las circunstancias. También se han propuesto tests estadísticos alternativos (Cressie y Read 1984), pero la falta de potencia de éstos aún no ha sido resuelta. Colapsar (refundir) sobre categorías o variables, de tal forma que se garantice un tamaño mínimo en las frecuencias esperadas, es el camino por el que optan la mayoría de los investigadores, y aún cuando este tipo de estrategias generalmente afecta a la estimación de los parámetros del modelo log-lineal, mientras se indique que las relaciones encontradas pertenecen a la tabla actualmente utilizada, esta técnica puede ser la más prudente y en algunos casos la única viable (Agresti 1990).

En esta línea, el trabajo que nos ocupa presenta un mecanismo de refundición de categorías de variables en tablas de contingencia en las que ocurren los problemas aludidos. Dicho mecanismo es semiautomático en el sentido de que es el investigador el que tiene la última decisión sobre adoptar o no las recomendaciones ofertadas como las más óptimas, atendiendo a las características propias de las variables. Además se implementa un entorno inteligente de desarrollo adecuado y cómodo para el tratamiento de un conjunto de datos de tipo cualitativo, desde una primera fase de descripción de variables implicadas en el estudio y manipulación de datos, hasta la obtención de mecanismos y herramientas de control necesarias en procesos de modelización para

llegar a conclusiones y resultados que se ajusten a los requerimientos marcados. Esto se aborda mediante la elaboración de un lógical integrado que incluye cinco módulos principales: un módulo de creación de descripciones, otro de introducción controlada de datos, un tercer módulo de verificación, análisis y refundición semiautomática de categorías en tablas de contingencia dispersas, el módulo de selección automática de modelos log-lineales, y un último de generación de ficheros de instrucciones BMDP. El enfoque elegido es interactivo, de uso simple para no expertos con amplios sistemas de ayuda, y sobre sistemas estándares.

La fase de cálculo numérico está resuelta de forma satisfactoria con los paquetes estadísticos convencionales, por lo que se ha considerado más adecuado enlazar las técnicas previas de modelización con la generación automática de uno o varios ficheros de instrucciones BMDP, así como la importación de los ficheros de datos preparados con un logical de tratamiento de datos categóricos adecuado al usuario final, aunque en los procesos de decisión incorporados en el análisis de datos se necesitan procesos numéricos de estimación de modelos log-lineales, que se han desarrollado e incorporado al sistema.

El sistema ADEST está programado en Turbo-Pascal 6.0, bajo sistema operativo DOS en ordenadores compatibles IBM que dispongan de tarjeta gráfica VGA, coprocesador matemático y ratón compatible Microsoft.

2. SELECCIÓN DE MODELOS

Para describir el procedimiento se necesita introducir la siguiente notación: la inclusión de modelos log-lineales se expresa en la forma habitual, así pues, $m_1 \leq m_2$ significa que m_1 es un submodelo de m_2, es decir en m_2 están los términos de m_1 y alguno más. Lógicamente, $m_1 < m_2$ se usa para decir que $m_1 \leq m_2$ y $m_1 \neq m_2$. A partir de este momento, y por razones de simplificación, las referencias al término "modelo" han de entenderse como "modelo log-lineal". Para cualquier conjunto de modelos $\mathcal{S}$ en $\mathcal{F}$ se definen $\max(\mathcal{S})$ y $\min(\mathcal{S})$ como:

$$\begin{aligned} \max(\mathcal{S}) &= \{s \in \mathcal{S} \ / \ \text{si} \quad s < t \Rightarrow t \notin \mathcal{S}\} \\ \min(\mathcal{S}) &= \{s \in \mathcal{S} \ / \ \text{si} \quad t < s \Rightarrow t \notin \mathcal{S}\} \end{aligned}$$

Los principios básicos en los que se basa el método de selección de modelos hacen referencia al principio de coherencia (Gabriel, 1969) y la utilización de dos reglas de uso frecuente en el ámbito del análisis multidimensional. El principio de coherencia implica que un procedimiento que involucre contrastar un conjunto de modelos, no debe de aceptar un modelo mientras se rechace un modelo más general, entendiéndose éste como un modelo que incluye al primero. Claramente, si se considera un modelo

compatible con los datos, es absurdo estimar un modelo más general incompatible con ellos. Las dos reglas a las que se ha hecho referencia son las que siguen:

1. Si un modelo se acepta, entonces todos los modelos que lo incluyan se consideran aceptados.
2. Si un modelo es rechazado, entonces todos sus submodelos se consideran rechazados.

En lo que sigue, se dice que un modelo es $d-$aceptado (débilmente aceptado) si incluye un modelo más simple aceptado, es decir deducimos que puede ser aceptado, y $d-$rechazado si está incluido en un modelo rechazado.

Estas reglas verifican el principio de coherencia y son prácticas puesto que reducen de forma considerable el número de modelos a ser ajustados. Así pues, los modelos $d-$aceptados o $d-$rechazados no necesitan ser considerados como especificaciones alternativas.

El procedimiento propuesto busca un conjunto $\mathcal{A}$ de modelos aceptados y un conjunto $\mathcal{R}$ de modelos rechazados, de tal forma que cualquier otro modelo de $\mathcal{F}$ (familia de todos los modelos log-lineales jerárquicos posibles que se pueden formar a partir de un conjunto de variables dadas) o contiene un modelo de $\mathcal{A}$, y por tanto es $d-$aceptado, o está contenido en un modelo de $\mathcal{R}$, y por tanto es $d-$rechazado.

Para cualquier modelo $m \in \mathcal{F}$, dado un nivel de significación α, un test de significación adecuado permitirá aceptar o rechazar su ajuste a los datos. Por tanto, en cualquier estado se pueden clasificar los modelos en $\mathcal{F}$ en tres subconjuntos:

$$\begin{aligned} \text{A} &= \{m \in \mathcal{F}/m \text{ se acepta}\} \\ \text{R} &= \{m \in \mathcal{F}/m \text{ se rechaza}\} \\ \text{I} &= \{m \in \mathcal{F}/m \text{ no ha sido contrastado}\} \end{aligned}$$

Si $m \in \mathcal{F}$ se ajusta a los datos, puede ser incluido en el conjunto A; si $m \in \mathcal{F}$ se rechaza pertenecerá a R, y si m no ha sido contrastado estará incluido en I. El propósito del procedimiento de selección de modelos es reducir el número de modelos de I hasta conseguir que $\text{I} = \emptyset$ con el menor coste computacional y de tiempo posibles.

Las construcciones básicas usadas para determinar qué modelo hay que contrastar son los conceptos de $a-$dual y $r-$dual de un conjunto dado. El $a-$dual de un conjunto de modelos $\mathcal{S}$, escrito $D_a(\mathcal{S})$ se define como el conjunto de modelos minimales en $\mathcal{F}$ que no están contenidos en ningún modelo de $\mathcal{S}$. Este concepto es de interés cuando los modelos en $\mathcal{S}$ han sido rechazados.

$$D_a(\mathcal{S}) = \min\{m \in \mathcal{F}/m \not\leq m^*, \forall m^* \in \mathcal{S}\}$$

De forma similar, el r-dual de $\mathcal{S}$, escrito $D_r(\mathcal{S})$, se define como el conjunto de modelos en $\mathcal{F}$ que no contienen ningún modelo en $\mathcal{S}$, siendo de utilidad cuando los modelos

en $\mathcal{S}$ han sido aceptados.

$$D_r(\mathcal{S}) = \max\{m \in \mathcal{F} / m^* \not\leq m, \quad \forall m^* \in \mathcal{S}\}$$

Si $\mathcal{S}$ está vacío, se define: $D_a(\mathcal{S}) = \min(\mathcal{F})$ y $D_r(\mathcal{S}) = \max(\mathcal{F})$

El primer paso, como ya se ha dicho, consiste en contrastar un conjunto inicial de modelos $\mathcal{M}I$ y construir los conjuntos $\mathcal{A}$ y $\mathcal{R}$.

$$\begin{aligned} \mathcal{A} &= \min\{m \in \mathcal{M}I / m \text{ se acepta}\} \\ \mathcal{R} &= \max\{m \in \mathcal{M}I / m \text{ se rechaza}\} \end{aligned}$$

El siguiente paso consiste en contrastar o $D_a(\mathcal{R}) \setminus \mathcal{A}$ (modelos en $D_a(\mathcal{R})$ que no pertenecen a $\mathcal{A}$) o bien $D_r(\mathcal{A}) \setminus \mathcal{R}$, y así sucesivamente actualizando los conjuntos $\mathcal{A}$ y $\mathcal{R}$. Si suponemos que en cualquier paso los modelos en $D_r(\mathcal{A}) \setminus \mathcal{R}$ son contrastados y rechazados, entonces después de haber actualizado $\mathcal{R}$, se tiene que $D_r(\mathcal{A}) \setminus \mathcal{R}$ está vacío y por tanto en este punto puede parar el procedimiento. A las mismas conclusiones se llega si los modelos en $D_a(\mathcal{R}) \setminus \mathcal{A}$ se contrastan y todos han sido aceptados.

Cuando se aplica el procedimiento a un problema particular hay que considerar cuidadosamente (a) qué test de ajuste se va a usar, (b) qué conjunto de modelos inicial se va a contrastar y (c) si contrastar $D_r(\mathcal{A}) \setminus \mathcal{R}$ o $D_a(\mathcal{R}) \setminus \mathcal{A}$ en cada paso.

Una modificación al procedimiento anterior es la que sigue. En muchas aplicaciones podrían considerarse modelos que sólo contengan un modelo dado m_0, es decir un submodelo, o bien que una determinada interacción aparezca en el modelo final. Esto puede ser posible en el contexto del procedimiento, al asignar inicialmente los modelos en $D_r(m_0)$ a $\mathcal{R}$, sin contrastar. Por tanto, para cualquier modelo $m \in \mathcal{F}$, m_0 es un submodelo de m si y sólo si m no está contenido en algún modelo de $D_r(m_0)$, con lo que se obtiene el resultado deseado, es decir, si se asigna $D_r(m_0) = \mathcal{R}$, se rechazan a priori todos aquellos modelos que no contengan a m_0.

El test estadístico de ajuste que se ha utilizado es el clásico test chi-cuadrado que garantiza el principio de coherencia, uno de los pilares básicos del procedimiento.

El punto de partida, es decir el conjunto de modelos inicial, es un punto crucial en el desarrollo del procedimiento no en cuanto a la solución final que, obviamente, ha de ser la misma a un mismo nivel de significación independientemente también del conjunto $D_a(\mathcal{R})$ o $D_r(\mathcal{A})$ seleccionado en cada paso (Edwards, 1987), sino al tiempo de cómputo necesario para llegar a dicha solución. No es lo mismo partir de un conjunto de modelos próximo a la solución que otro mucho más alejado; el número de pasos necesarios para converger es mucho mayor en el segundo caso. Por tanto, si el investigador conoce o sospecha algún tipo de relación o interacción entre variables, puede comenzar el procedimiento con esta información, pero si no es éste

el caso, el proceso automático de selección ha de construir un conjunto de modelos inicial que de una forma no muy compleja intente reducir el tiempo de proceso y a la vez tenga en consideración la naturaleza de las variables en estudio. Los contrastes STP (Simultaneous Test Procedure) sobre la existencia de efectos de un orden k o superior, pueden también ayudar en la selección del conjunto inicial.

Partiendo de la hipótesis no restrictiva y lógica de que al menos existe una variable con mayor peso respecto a las demás, entre el conjunto de variables implicadas, el primer objetivo es construir un conjunto de modelos anidados (un conjunto m_2 se dice anidado respecto de m_1 cuando los parámetros de m_2 constituyen un subconjunto de los parámetros de m_1) que tenga en consideración tales variables y sus interrelaciones. Es norma usual en análisis exploratorios de datos incluir el término de mayor interacción entre variables con mayor peso y concentrar el proceso de modelización sobre términos que relacionen variables con menor peso.

Así pues, si $X = \{x_1, x_2, \ldots, x_n\}$ representa el conjunto de variables sobre las que se va a operar, e $Y = \{y_1, y_2, \ldots, y_m\}$ $(m < n)$ el subconjunto de X formado por las variables de mayor peso, se puede considerar el conjunto de modelos iniciales $\mathcal{M}I$ como:

$$\mathcal{M}I = \{\{G \cup C_i\} \cup \{\acute{C}_j\}\}_{i=1,\ldots,k}^{j=k+1,\ldots,n-1}$$

donde:

- k es el orden de la mayor interacción G posible entre las variables en Y
- $\{\{G \cup C_i\}\}$ es el conjunto de modelos de la forma $\{G \cup C_i\}$ $i=1,\ldots,k$ y $C_i = \bigcup_x c_i$ con c_i interacción de orden i en X
- $\{\{\acute{C}_j\}\}$ es el conjunto de modelos de la forma $\{\acute{C}_j\}$ $j=k+1,\ldots,n-1$ y $\acute{C}_j = \bigcup_x \acute{c}_j$ con $\acute{c}_j$ interacción de orden j en X.

Otro posible conjunto inicial de modelos podría ser considerar el modelo no saturado de mayor orden, es decir aquel que contiene todas las interacciones de orden $n-1$, garantizando, de igual forma, que las interrelaciones máximas entre variables de mayor peso están incluidas. Análisis sobre tiempos de cómputo necesarios para obtener la solución final mediante procesos de simulación de tablas, confirman que iniciando el proceso de selección partiendo de modelos anidados, el tiempo empleado se reduce de forma considerable con una ganancia aproximada de 4 a 1.

Otro estado a considerar dentro del procedimiento de selección supone la elección entre la construcción del $D_a(\mathcal{R})$ o el $D_r(\mathcal{A})$. Dada la naturaleza lógica y algebraica del método para la construcción de estos conjuntos (Havranek,1987), el número de combinaciones entre generadores de modelos de $\mathcal{A}$ o $\mathcal{R}$, y por tanto el número de modelos intermedios para tales construcciones, puede llegar a ser excesivo y supone otro punto crítico en el proceso de control, tanto en la velocidad de proceso como

en las limitaciones de la memoria disponible. Lo ideal en cuanto a tiempo de proceso sería construir mediante el algoritmo oportuno todos los modelos intermedios y minimizar éstos para obtener el $D_a(\mathcal{R})$ o el $D_r(\mathcal{A})$, pero tal número de modelos intermedios, aunque se utilizan estructuras de datos dinámicas para almacenarlos, supone un desborde de la memoria disponible relativamente rápido en equipos con pocos recursos.

La solución dada al problema planteado consiste en pasar un proceso de construcción-minimización de $D_a(\mathcal{R})$ o $D_r(\mathcal{A})$ después de la generación de cada modelo y no esperar a que se generen todos. De esta forma el proceso consume más tiempo pero las necesidades de memoria disminuyen de forma considerable ya que se evita manejar muchos modelos repetidos o incluidos en otros.

Por último, se han implementado dos mecanismos de control y decisión para automatizar de forma completa el método de selección de modelos. El primero para decidir si construir el $D_a(\mathcal{R})$, el $D_r(\mathcal{A})$ o ambos, analizando el número de combinaciones entre generadores de modelos necesarios para tales construcciones. Si el número de estas combinaciones es relativamente pequeño se considera indiferente uno u otro, en caso contrario se opta por aquel que requiere menor número de modelos intermedios para su construcción. El segundo mecanismo de control está implementado para decidir si contrastar los modelos del $D_a(\mathcal{R})$ o el $D_r(\mathcal{A})$, ateniéndose a la siguientes normas y orden: se selecciona el conjunto con un menor número de modelos y dentro de éstos, el que tenga un menor número total de generadores, y por último, si coinciden ambos criterios, el de menor varianza de generadores.

El hecho de adoptar estas normas o reglas viene dado, principalmente por el tiempo de proceso consumido en el procedimiento de cálculo de frecuencias esperadas estimadas para cada modelo (Bishop, Fienberg y Holland, 1975). Aunque no es posible conocer cuanto tiempo puede consumir este proceso (al ser un proceso iterativo, el número de iteraciones dependerá de la naturaleza de los datos), si se conoce el número de pasos necesarios en cada iteración: tantos como generadores tenga el modelo. Por otra parte, también se ha seguido otro proceso lógico y tradicional en la elección de modelos: el criterio de parsimonia, lo que supone elegir los modelos o conjunto de modelos más simples.

3. REFUNDICIÓN DE CATEGORÍAS

Un problema frecuente en el análisis estadístico, en particular en análisis de tablas de contingencia y modelos log-lineales, radica en un tamaño muestral pequeño. Si muchas frecuencias esperadas de celdas de una tabla de contingencia multidimensional

tiene valores bajos ($< 5 o < 1$), no puede esperarse que los tests estadísticos G^2 de razón de verosimilitudes y X^2 de Pearson se aproximen a la distribución teórica chi-cuadrado (Haberman, 1988). Es más, los efectos estandarizados (Goodman, 1971) y los errores ajustados no se aproximan a la distribución normal estándar. Cochran estudió en una serie de artículos la aproximación chi-cuadrado para X^2: en 1954 sugirió que para contrastar la independencia con más de un grado de libertad, un valor mínimo de 1 era permisible siempre y cuando no existan más de un 20% de valores esperados de celdas por debajo de 5. Larntz (1978), Koehler y Larntz (1980) y Koehler (1986) muestran que el estadístico X^2 se comporta mejor que G^2 para tamaños muestrales pequeños y tablas dispersas. Muestran que la distribución muestral de G^2, generalmente, se aproxima en menor grado a una chi-cuadrado que X^2 cuando n/N es menor que 5, siendo n el tamaño muestral y N el número de celdas.

Agresti (1990) aborda este problema y realiza una recopilación de técnicas alternativas, entre las que cita a Cressie y Read (1984) con los tests de potencia divergente basados en la familia de estadísticos:

$$\frac{2}{\lambda(\lambda+1)} \sum n_i \left[\left(\frac{n_i}{\hat{m}_i} \right)^{\lambda} - 1 \right] \quad ; \quad -\infty < \lambda < \infty$$

Cita también que otras adaptaciones de test chi-cuadrado han sido propuestas por Berry y Mielke (1988) y Zelterman (1987) y comenta que las ventajas de estas adaptaciones sobre los tests estadísticos convencionales no son claras en muchas circunstancias y además, la falta de fuerza de estos tests aún no ha sido resuelta.

En el sentido referenciado en la introducción sobre refundición de categorías o variables, de tal forma que se garantice un tamaño mínimo en las frecuencias esperadas, y considerando tablas multidimensionales, se ha trabajado en criterios que realicen la refundición semiautomática de categorías. Para ello se han definido los cuatro criterios de decisión y que se exponen a continuación, refiriéndose todos ellos a la tabla de frecuencias esperadas bajo un determinado modelo log-lineal. Los criterios de decisión adoptados son los siguientes:

- Se considera que una tabla no necesita una refundición cuando al menos el 80% de sus valores son mayores que 5 y todos mayores que 1, aunque este criterio se puede alterar en cada caso concreto.

- Se construye el vector ordenado *IRV* (índice de refundición de variables) cuyos elementos son de la forma (V^{X_i}); $1 \leq i \leq n_v$ (n_v=número de variables) y cada elemento V^{X_i} se corresponde con el número de líneas (filas o columnas en el caso bidimensional) "defectuosas" en relación al número total de categorías la variable X_i, siempre y cuando el número de categorías de la variable sea mayor o igual a 3. Si en algunas variables estos valores coinciden, la ordenación

se realiza conforme a aquella que tenga más categorías (esta última decisión parte de la lógica de que al refundir se pierde una categoría). La variable correspondiente al primer elemento de *IRV* será aquella en la que se ha de producir una refundición entre dos de sus categorías. Para saber si una línea es defectuosa se procederá de la siguiente forma:

- Se busca la línea con más valores menores que 5 calculándose dicho número de valores.
- Se le calcula el "tope" (número entero entre 0 y 100 dado como información a priori) por ciento a ese número de valores encontrado.
- Si una línea tiene más valores menores que 5 que el módulo del porcentaje calculado anteriormente, se considera defectuosa.

De cualquier forma es el investigador el que decide si acepta refundir alguna categoría de la variable propuesta por el programa según el criterio anterior, o por el contrario prefiere otra.

- La categoría "peor" dentro de una variable (y por tanto la que necesita ser refundida con otra) será aquélla correspondiente al primer elemento del vector ordenado IRC^{X_k}. IRC^{X_k} es un vector cuyos elementos son de la forma $(C_j^{X_k})\ 1 \leq j \leq n_{ck}$ (n_{ck}=número de categorías de X_k) para la variable X_k seleccionada en el paso anterior. Cada elemento $C_j^{X_k}$ es el número de valores menores que 5 que contiene la categoría j de X_k. En el caso de existir más de una categoría con el mismo valor, la ordenación se realiza conforme a aquella que tenga menor valor medio.

- Para decidir con qué categoría ha de refundirse la encontrada con anterioridad, hay que distinguir entre que la escala de medida de la variable en cuestión sea nominal u ordinal, es decir, si sus categorías están ordenadas o no:

  - *Ordinal:* Sólo puede refundirse con una categoría adyacente. Si la peor categoría es la primera o la última, sólo hay una posibilidad: La segunda y la penúltima, respectivamente. Si no es ése el caso, existen dos y se tomará la peor. TABCONT encuentra ésa peor buscando en IRC^{X_k} los valores correspondientes a las categorías adyacentes y seleccionando aquella que figure primero dentro de IRC^{X_k}, pero permite al usuario seleccionar la otra posibilidad.
  - *Nominal:* La peor categoría puede refundirse con cualquier otra. En este caso, TABCONT presenta al usuario las categorías correspondientes a los restantes elementos de IRC^{X_k}. En este punto es el usuario el que ha de decidir la primera refundición de la lista que le parezca lógica (de esta forma nos aseguramos haber realizado la mejor refundición posible).

Tanto si las opciones presentadas son aceptadas o se introducen otras distintas, el proceso continua adicionando las frecuencias de las categorías obtenidas o seleccionadas, reajustando tablas y actualizando la información sobre variables y categorías, hasta que se cumplan las condiciones del test.

4. GENERADOR DE FICHEROS DE INSTRUCCIONES BMDP

Como un módulo más integrado en el sistema ADEST, GEN-BMDP tiene como objetivo la construcción de un fichero de instrucciones BMDP. Parte de la idea de algunos autores de que los sistemas estadísticos pudieran imitar de manera automática los pasos seguidos por un estadístico experimentado y generar igualmente de forma automática los correspondientes interfaces con los paquetes estadísticos, en este caso BMDP.

Está enfocado hacia el investigador que ya familiarizado con alguna técnica estadística y conozca bien sus objetivos (en este apartado el sistema puede dejarlos claros al seleccionar un conjunto de modelos log-lineales óptimo, para posteriormente ser tratados y analizados con el paquete estadístico), pueda disponer de una herramienta de fácil manejo que le descargue de la tarea de programar en el lenguaje propio del paquete estadístico BMDP.

Naturalmente, se han seguido unos pasos que se corresponden con la lógica seguida en la codificación de instrucciones BMDP:

- generación automática de párrafos generales de descripción de datos,
- generación automática de los párrafos particulares correspondientes a los objetivos específicos del estudio.

A partir de un fichero de descripción sobre las características de las variables, se obtiene la información necesaria para concluir el primer paso, con lo que se generarían los párrafos:

/INPUT	CASES	FILES
	FORMAT	TITLE
	VARIABLES	
/VARIABLES	NAMES	MAX
	USE	MIN
	MISSING	GROUP
/GROUP	CODES	NAMES
	CUTPOINT	

Para el segundo, una vez leído el fichero de descripción, se presenta al investigador un protocolo de preguntas distinto según el programa BMDP seleccionado y las características descritas para las variables en uso. Dicho protocolo de preguntas está basado en las posibilidades que permite el programa seleccionado. Si éste es muy amplio, como podría ser 4F, habrá que realizar más preguntas al usuario para concretar qué tratamiento se desea exactamente. Las preguntas en cada caso son distintas según las respuestas precedentes y la naturaleza de las variables utilizadas. Puede decirse, en este sentido, que GEN-BMDP es seudointeligente y además, ofrece la posibilidad de ayuda en cada momento al usuario, con referencia a las preguntas formuladas y posibilidades presentadas a éste. Como ya se ha comentado con anterioridad, el sistema ADEST está orientado hacia el tratamiento de datos de tipo categórico, con lo que el generador de ficheros de instrucciones cubre los siguientes programas BMDP:

- 1D, 2D, 5D — descripción y tabulación de datos,
- 4F — tablas de contingencia y modelos log-lineales,
- CA, CAM — análisis de correspondencias simple y múltiple,
- LR, PR — regresión logística dicotómica y policotómica paso a paso.

Una vez seleccionado el programa BMDP y finalizado el protocolo de preguntas correspondiente, GEN-BMDP dispone de la información necesaria para poder generar el fichero de instrucciones.

REFERENCIAS

[1] **Agresti, Alan** (1984). *Analysis of ordinal categorical data.* Wiley Interscience.

[2] **Agresti, Alan** (1990). *Categorical data analysis.* Wiley interscience.

[3] **Akaike, H.** (1973). "Information theory and an extension of the maximun likelihood principle". In B.N. Petrov and B.F. Csaki (Eds), *Second International Symposium on Information Theory*, 267–281. Budapest: Academiai Kiado.

[4] **Akaike, H.** (1987). "Factor Analysis and AIC". *Psychometrika,* **52**, 317–332.

[5] **Baker, R.J.** and **Nelder, J.A.** (1978). *The GLIM system. Release 3. Generalized linear interactive modelling manual.* Oxford: N.A.G.

[6] **Berry, K.J.** and **Mielke, P.W.** (1988). "Montecarlo comparisons of the asymptotic chi-square and likelihood ratio test with the nonasymptotic chi-square test for sparse rxc tables". *Psychol. Bull.,* **103**, 256–264.

[7] **Bishop, Y.S., Fienberg, P.** and **Holland, P.** (1975). *Discrete multivariate analysis: Theory and practice*. MIT Press, Cambridge.

[8] **Caridad, J.M.** and **Espejo, R.** (1991). *Sistema experto en análisis de datos categorizados.* I Seminario Internacional de sistemas expertos en la agricultura mediterránea. Córdoba.

[9] **Caridad, J.M.** and **Espejo, R.** (1991). *Inteligencia artificial y estadística aplicada.* I Seminario Internacional de sistemas expertos en la agricultura mediterránea. Córdoba.

[10] **Clogg, C.C.** and **Becker, M.** (1986). "Log-linear model with SPSS". *Computer science and statistics.*, 263–269. North Holland. Amsterdam.

[11] **Clogg, C.C.** and **Eliason, S.R.** (1987). "Some common problems in log-linear analysis". *Sociological Methods and Research,* **16**, 8–14.

[12] **Cochran, W.G.** (1954). "Some method of strengthening the common chi-square test". *Biometrics*, **10**, 417–451.

[13] **Cressie, N.** and **Read, T.R.** (1984). "Multinomial goodness of fit". *J. Roy. Statist. Soc.,* **46**, 440–464.

[14] **Cressie, N.** and **Read, T.R.** (1989). "Pearson X^2 and the loglikelihood ratio statistic G^2: A comparative review". *Internat. Statist. Rev.,* **57**, 19–43.

[15] **Dixon, W.J.** (1990). *Ed's Statistical software manual. Vol 1 y 2.* University of California Press.

[16] **Edwards, D.** and **Havránek, T.** (1985). "A fast procedure for model search in multidimensional contingency tables". *Biometrika*, **72**, 339–351.

[17] **Edwards, D.** and **Havránek, T.** (1987). "A fast model selection procedure for large families of model". *J.A.S.A.*, **82, 397**, 205–213.

[18] **Gabriel, K.R.** (1969). "Simultaneous test procedures — some theory of multiple comparisons". *Annals of Mathematical Statistics*, **40**, 224–250.

[19] **Goodman, L.A.** (1971). "The analysis of multiple contingency tables: stepwise procedures and direct estimation methods for buildings models for multiple clasifications". *Technometrics,* **13**, 33–61.

[20] **Goodman, L.A.** (1984). *The analysis of cross-classified data having ordered categories.* Harvard University Press.

[21] **Haberman, S.J.** (1973). "The analisys of residuals in cross-classification tables". *Biometrics*, **29**, 205–220.

[22] **Haberman, S.J.** (1988). "A warning on the use of chi-square statistics with frequency tables with small expected cell counts". *J.A.S.A.*, **83**, 555–560.

[23] **Havránek, T.** (1984). "A procedure for model search in multidimensional contingency tables". *Biometrics*, **40**, 95–100.

[24] **Koehler, K.** and **Larntz, K.** (1980). "An empirical investigation of goodness of fit statistics for sparse multidimensiona tables". *J.A.S.A.*, **75**, 336–344.

[25] **Koehler, K.** (1986). "Goodness of fit test for log-linear models in sparse contingency tables". *J.A.S.A.*, **81**, 483–493.

[26] **Larntz, K.** (1987). "Small-sample comparison of exact levels for chi-squared goodness of fit statistics". *J.A.S.A.*, **73**, 253–263.

[27] **Nelder, J.A.** and **Baker, R.J.** (1985). "Statitstical software: progress and prospects". *Computer Science and Statistics. Proc. of de 16th symposium on the interface*, 33–37. Amsterdam.

[28] **Raftery, A. E.** (1986). "Choosing models for cross-classifications". *Amer. Socciol. Rev.,* **51**, 145–146.

[29] **Tukey, J.W.** (1985). "Comment on more inteligence statistical software and statistical expert systems: Future directions". *The American Statistician,* **39**, 12–14.

[30] **Wrigley, D.** (1985). *Categorical data analysis for geographers and environmental scientists.* Longman. London.

[31] **Zelterman, D.** (1987). "Goodness of fit test for large sparse multidimensional distributions". *J.A.S.A.,* **82**, 624–629.

ENGLISH SUMMARY:

ADEST SYSTEM

J.M. Caridad Ocerin and R. Espejo Mohedano

ADEST is an intelligent software environment for categorical data analysis. It is oriented to end users of log-linear models and contingency tables, who belong to medical or social sciences backgrounds, that is, to professionals who are familiar with the main concepts of multivariate categorical data, but who are not experts in this field. It is an enhanced package, which helps and guides in log-linear model building, and solves some of the practical problems associated with these statistical techniques, like dealing with low expected frecuencies.

In log-linear model specifications, the computational problems become really serious when the number of variables exceeds four; for example there are 7.581 five variable hierarchical models and 7.828.354 with six variables.

Some stepwise procedures for model building lead to one final model after a sequence of goodness of fit test. This can be misleading, as in many situationes it is not possible to specify an optimun model, but a set of alternative models. The stepwise method could also produce errors in the specification depending on the first variables included or excluded.

One alternative is the Edwards and Havranek (1984, 1987) selection procedure, and a generalization of it is implemented in ADEST. From an initial set of possible models $\mathcal{F}$ we use a sequence of acceptable models $\mathcal{A}$, non-acceptable models $\mathcal{R}$, and non-tested models I. This last set is reduced, in a finite number of steps, to the empty set.

A model m is weakly accepted if there is a submodel $m \leq m$ that belongs to the set $\mathcal{A}$, and weakly rejected if $m \in \mathcal{R}$. It is not necessary, then, to test the weakly accepted or rejected models. The set:

$$D_a(\mathcal{R}) = \min\{m \in \mathcal{F} / m \not\leq m \ , \quad \forall m \ \in \mathcal{R}\}$$

includes the minimal log-linear models not included in rejected models, and $D_a(\mathcal{R}) \setminus \mathcal{A}$ $= D_a(\mathcal{R}) \cap \mathcal{A}^c$ (where $\mathcal{A}^c$ is the set of models in $\mathcal{F}$-$\mathcal{A}$) is the set of models to be tested; if one of these models is rejected, it must be included in $\mathcal{R}$ and the set $D_a(\mathcal{R}) \setminus \mathcal{A}$ is then reduced. The procedure stop when $D_a(\mathcal{R}) \setminus \mathcal{A} = \emptyset$. Of course it is possible to define the set:

$$D_r(\mathcal{A}) = \max\{m \in \mathcal{F} / m \ \not\leq m, \quad \forall m \ \in \mathcal{A}\}$$

and use $D_r(\mathcal{A}) \setminus \mathcal{R} = D_r(\mathcal{A}) \cap \mathcal{R}^c$ (where $\mathcal{R}^c$ is the set of models in $\mathcal{F}$-$\mathcal{R}$); if one of these models is accepted, $\mathcal{A}$ is should be updated and the new $D_r(\mathcal{A}) \setminus \mathcal{R}$ has fewer elements; again this procedure is repeated until this set is empty.

To increase the computational efficiency ADEST uses both sets $D_a(\mathcal{R}) \setminus \mathcal{A}$ and $D_r(\mathcal{A}) \setminus \mathcal{R}$, and the sequential procedure begins with an initial set of models so that the number of steps is minimized. In addition the highest level of interaction to be included in the final set of acceptable (non weakly) models can be stated beforehand, and again, this accelerates the selection procedure.

The final set of log-linear models selected, with a fixed significance level is independent of the initial set of models used in the sequential procedure, but the computational time is dependent of this initial set, so it is important a good selection as starting values.

Also ADEST has a module to produce automatic aggregation of categories to avoid the problem of low expected frequencies in the chi-square test, taking account of the measurement scale of the variables, the expected frequencies associated with each test, and the information provided by the user to guide the collapsability of the categories at different stages of the model selection.

The ADEST environment is user friendly, avalaible for use on DOS based computers (although there is also a Unix version) and it includes an interface with BMDP categorical data analysis programs.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 233-257, 1995

IMPLEMENTACIÓ D'UN ALGORISME PRIMAL-DUAL DE PUNT INTERIOR AMB FITES SUPERIORS A LES VARIABLES

J. CASTRO*

Universitat Politècnica de Catalunya

Aquest document presenta una implementació d'un algorisme primal-dual de punt interior, que permet considerar variables afitades superiorment. A diferència d'altres codis que obtenen la direcció del mètode de Newton a cada iteració usant tècniques de matrius simètriques indefinides, la implementació feta descompon el procés fins arribar a un sistema simètric i definit positiu. Això permet que es pugui fer una factorització simbòlica prèvia per tal de conèixer la topologia de la descomposició de Cholesky que s'efectuarà a cada iteració, agilitzant així el procés de càlcul. El codi desenvolupat es compara amb dos paquets de programació lineal capdavanters: Minos 5.3 (una implementació eficient de l'algorisme del símplex) i LoQo (codi de punt interior basat també en un algorisme primal-dual). Els problemes resolts en la comparació pertanyen al conjunt de problemes de la Netlib (una bateria estàndard de problemes de programació lineal).

An implementation of a primal-dual interior point algorithm with upper-bounded variables.

Key words: Algorisme Primal-Dual, Mètodes de Barrera Logarísmica, Mètodes de Punt Interior, Programació Lineal, Resultats Computacionals.

*J. Castro. Dept. d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya. c. Pau Gargallo 5, 08028 Barcelona, Spain.

- Article rebut el desembre de 1994.
- Acceptat el juny de 1995.

1. INTRODUCCIÓ

Des de l'aparició el 1984 d'un algorisme basat en transformacions projectives a Karmarkar(1984) per resoldre problemes de programació lineal (anomenat també algorisme de Karmarkar), es va iniciar una forta activitat de recerca dins del camp dels anomenats mètodes de punt interior. Aquests, a diferència del fins llavors majoritàriament emprat algorisme del símplex, arriben al punt òptim a través de l'interior de la zona factible delimitada pel conjunt de constriccions, en comptes de bellugar-se per la frontera com ho feien els iterats obtinguts amb el símplex. L'algorisme de les transformacions projectives, però, va ser precedit per l'algorisme de l'el.lipsoid (Khachiyan(1979)), també de complexitat polinòmica, encara que mai va ser un competidor del símplex des del punt de vista pràctic. Des d'aquell article inicial d'en Karmarkar nous mètodes van aparèixer, entre ells els anomenats d'escalat afí, on la transformació projectiva de l'algorisme de Karmarkar era substituïda per una transformació afí (Barnes(1986), Monma i Morton(1987), Vanderbei *et al*(1986)). Posteriorment van aparèixer els anomenats mètodes primal-dual (Monteiro i Adler(1989)), basats en la substitució de les constriccions de no-negativitat de les variables per una barrera logarísmica (Wright(1991)). Els mètodes de punt interior han guanyat un ampli reconeixement com a procediment d'optimització per a problemes lineals en els darrers anys, i s'han mostrat força més eficients que l'algorisme del símplex.

El codi presentat en aquest document implementa un mètode primal-dual de punt interior. Malgrat ser, des d'un punt de vista formal, més complexos que els mètodes d'escalat afí, els mètodes primal-dual s'han mostrat computacionalment més eficients que aquells. La implementació feta tracta les fites superiors a les variables de forma explícita, sense afegir constriccions addicionals. Això fa que no incrementem la dimensió del sistema simètric i definit positiu a solucionar a cada iteració (donat que aquesta dimensió és exactament el nombre de constriccions del nostre problema). A més, el fet de que la topologia d'aquest sistema sigui sempre la mateixa a cada iteració, permet usar rutines de factorització simbòlica per tal de determinar a priori el nombre i la situació dels elements nous no-zero creats en fer la descomposició de Cholesky. Aquesta factorització simbòlica caldrà fer-la només una vegada al principi de l'execució, i serà aprofitada posteriorment a cada iteració. En aquest punt el codi difereix clarament respecte altres implementacions capdavanteres d'algorismes primal-dual (com ara Vanderbei(1992), Vanderbei(1994)), que solucionen un sistema simètric i indefinit mitjançant una factorització de Bunch-Parlett (Duff *et al*(1986), Vanderbei i Carpenter(1993)), que, en teoria, ha de ser lleugerament menys eficient que una factorització de Cholesky coneixent el patró d'esparsitat.

El present document s'estructurarà de la següent forma. En primer lloc es detallaran les bases de l'algorisme primal-dual amb fites superiors a algunes variables. A continuació es comentaran certs aspectes referents a detalls de la implementació

feta. Finalment el codi desenvolupat es comparará amb dos codis capdavanters d'optimització lineal, usant una bateria estàndard de 80 problemes de programació lineal.

2. L'ALGORISME PRIMAL-DUAL CONSIDERANT FITES SUPERIORS A LES VARIABLES

En aquesta secció es farà un breu esment del mètode primal-dual usat, una especialització del mètode general per a problemes amb fites superiors en algunes variables. No es pretén, però, aprofondir en els aspectes de convergència ni justificar el mètode. Per més detalls sobre els mètodes primal-dual hom pot consultar Arbel(1993) i Monteiro i Adler(1989).

2.1. Formulació dels problemes primal i dual

Considerem el següent problema de minimització amb fites superiors a algunes variables

$$(1)\qquad \begin{aligned} \min\ & c^t x \\ \text{subj. a}\ & Ax = b \\ & 0 \le x_u \le \bar{x}_u \\ & 0 \le x_l \end{aligned}$$

on $x_u \in \mathbb{R}^{n_u}$, $x_l \in \mathbb{R}^{n_l}$, $x = (x_u^t, x_l^t)^t$, $x \in \mathbb{R}^n$ (llavors $n = n_u + n_l$), $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ i $A \in \mathbb{R}^{m \times n}$. Considerant un particionament escaient de c i A, i afegint folgues $f \in \mathbb{R}^{n_u}$ per als límits superiors, el problema anterior pot ser escrit com:

$$(2)\qquad \begin{aligned} \min\ & c_u^t x_u + c_l^t x_l \\ \text{subj. a}\ & A_u x_u + A_l x_l = b \\ & x_u + f = \bar{x}_u \\ & x \ge 0 \\ & f \ge 0 \end{aligned}$$

on $c_u \in \mathbb{R}^{n_u}$, $c_l \in \mathbb{R}^{n_l}$, $A_u \in \mathbb{R}^{m \times n_u}$ i $A_l \in \mathbb{R}^{m \times n_l}$.

Considerant les variables duals $y \in \mathbb{R}^m$, $u \in \mathbb{R}^{n_u}$, i les folgues $z = (z_u^t, z_l^t)^t$ ($z_u \in \mathbb{R}^{n_u}$, $z_l \in \mathbb{R}^{n_l}$, per tant $z \in \mathbb{R}^n$) i $w \in \mathbb{R}^{n_u}$, el problema dual associat a (2) pot ser escrit com:

$$
\begin{aligned}
&\max\ b^t y + \vec{x}_u^t u \\
&\text{subj. a } A_u^t y + u + z_u = c_u \\
&\qquad\qquad A_l^t y + z_l = c_l \\
&\qquad\qquad u + w = 0 \\
&\qquad y \in \mathbb{R}^m \quad u \in \mathbb{R}^{n_u} \\
&\qquad z_l \geq 0 \quad z_u \geq 0 \quad w \geq 0
\end{aligned} \tag{3}
$$

De la tercera constricció es té que $u = -w$ i, aleshores, eliminant u el problema dual pot ser formulat com:

$$
\begin{aligned}
&\max\ b^t y - \vec{x}_u^t w \\
&\text{subj. a } A_u^t y + z_u - w = c_u \\
&\qquad\qquad A_l^t y + z_l = c_l \\
&\qquad y \in \mathbb{R}^m \\
&\qquad z_l \geq 0 \quad z_u \geq 0 \quad w \geq 0
\end{aligned} \tag{4}
$$

2.2. Obtenció dels Lagrangians a través una barrera logarísmica

Als problemes primal i dual prèviament definits les restriccions de no-negativitat de les variables poden ser substituïdes per una penalització de barrera logarísmica (tal i com descriu Wright(1991)). Llavors el problema primal (2) pot ser expressat com:

$$
\begin{aligned}
&\min\ c_u^t x_u + c_l^t x_l - \mu \textstyle\sum_{i=1}^{n} \ln x_i - \mu \sum_{i=1}^{n_u} \ln(\bar{x}_{u_i} - x_{u_i}) \\
&\text{subj. a } A_u x_u + A_l x_l = b \\
&\qquad \mu \geq 0
\end{aligned} \tag{5}
$$

mentre que el dual (4) pot formular-se tal i com segueix:

$$
\begin{aligned}
&\max\ b^t y - \vec{x}_u^t w + \mu \textstyle\sum_{i=1}^{n} \ln z_i + \mu \sum_{i=1}^{n_u} \ln w_i \\
&\text{subj. a } A_u^t y + z_u - w = c_u \\
&\qquad\qquad A_l^t y + z_l = c_l \\
&\qquad y \in \mathbb{R}^m \quad \mu \geq 0
\end{aligned} \tag{6}
$$

A continuació, associant els multiplicadors de Lagrange $y \in \mathbb{R}^m$ a les constriccions d'igualtat de (5) construim el Lagrangià L_p del problema primal:

$$(7)\quad L_p(x,y,\mu) = c_u^t x_u + c_l^t x_l - \mu \sum_{i=1}^{n} \ln x_i - \mu \sum_{i=1}^{n_u} \ln(\bar{x}_{u_i} - x_{u_i}) - y^t(A_u x_u + A_l x_l - b)$$

Anàlogament, associant els multiplicadors $x_u \in \mathbb{R}^{n_u}$ i $x_l \in \mathbb{R}^{n_l}$ a les constriccions d'igualtat de (6) obtenim L_d, el Lagrangià del dual:

$$\begin{aligned}(8)\quad L_d(x,y,z,w,\mu) &= b^t y - \bar{x}_u^t w + \mu \sum_{i=1}^{n} \ln z_i + \\ &+ \mu \sum_{i=1}^{n_u} \ln w_i - x_u^t(A_u^t y + z_u - w - c_u) - x_l^t(A_l^t y + z_l - c_l)\end{aligned}$$

2.3. Condicions d'optimalitat de Kuhn-Tucker de primer ordre

Si denotem per e_l el vector l-dimensional de 1's i considerem X_u, X_l, X, Z_u, Z_l, Z, W, F matrius diagonals, on:

$$(9)\quad \begin{aligned} e_l &= (1_1,\ldots,1_l)^t \\ X_u &= \operatorname{diag}(x_{u_1},\ldots,x_{u_{n_u}}) \\ X_l &= \operatorname{diag}(x_{l_1},\ldots,x_{l_{n_l}}) \\ X &= \begin{pmatrix} X_u & 0 \\ 0 & X_l \end{pmatrix} \\ Z_u &= \operatorname{diag}(z_{u_1},\ldots,z_{u_{n_u}}) \\ Z_l &= \operatorname{diag}(z_{l_1},\ldots,z_{l_{n_l}}) \\ Z &= \begin{pmatrix} Z_u & 0 \\ 0 & Z_l \end{pmatrix} \\ W &= \operatorname{diag}(w_1,\ldots,w_{n_u}) \\ F &= \operatorname{diag}(\bar{x}_{u_1} - x_{u_1},\ldots,\bar{x}_{u_{n_u}} - x_{u_{n_u}}) \end{aligned}$$

llavors les condicions d'optimalitat necessàries de primer ordre de L_p poden ser escrites com:

$$(10)\quad \frac{\partial L_p}{\partial x_u} = c_u - \mu X_u^{-1} e_{n_u} + \mu F^{-1} e_{n_u} - A_u^t y := 0$$

$$(11)\quad \frac{\partial L_p}{\partial x_l} = c_l - \mu X_l^{-1} e_{n_l} - A_l^t y := 0$$

$$(12)\quad \frac{\partial L_p}{\partial y} = -(A_u x_u + A_l x_l - b) := 0$$

Per la seva banda, les condicions d'optimalitat de primer ordre del Lagrangià dual (8) són:

$$\frac{\partial L_d}{\partial x_u} = c_u - A_u^t y - z_u + w := 0 \tag{13}$$

$$\frac{\partial L_d}{\partial x_l} = c_l - A_l^t y - z_l := 0 \tag{14}$$

$$\frac{\partial L_d}{\partial y} = b - A_u x_u - A_l x_l := 0 \tag{15}$$

$$\frac{\partial L_d}{\partial z} = \mu Z^{-1} e_n - X e_n := 0 \tag{16}$$

$$\frac{\partial L_d}{\partial w} = \mu W^{-1} e_{n_u} - F e_{n_u} := 0 \tag{17}$$

Les condicions (12) i (15) són la mateixa, i imposen la *factibilitat primal* a l'òptim. Les condicions (13) i (14) imposen la *factibilitat dual* de la solució. Les condicions (16) i (17) (usant ()) poden ser reescrites com:

$$\left.\begin{aligned} Z_u X_u e_{n_u} &= \mu e_{n_u} \\ Z_l X_l e_{n_l} &= \mu e_{n_l} \end{aligned}\right\} \tag{18}$$

$$F W e_{n_u} = \mu e_{n_u} \tag{19}$$

i les anomenarem condicions de *complementarietat* (això degut a que, quan $\mu \to 0$ tenim que (18) i (19) són exactament les condicions del teorema de la folga complementària, les quals han de ser garantides per la solució del problema primal i dual). Finalment, es pot veure fàcilment que, verificant-se la factibilitat primal, la factibilitat dual i la complementarietat, automàticament queden satisfetes les dues condicions restants (10) i (11):

• pel que fa a (10) veiem que:

$$\begin{aligned} 0 &\stackrel{?}{=} c_u - \mu X_u^{-1} e_{n_u} + \mu F^{-1} e_{n_u} - A_u^t y \\ &= A_u^t y + z_u - w - \mu X_u^{-1} e_{n_u} + \mu F^{-1} e_{n_u} - A_u^t y \quad \text{[usant (13)]} \\ &= Z_u e_{n_u} - W e_{n_u} - \mu X_u^{-1} e_{n_u} + \mu F^{-1} e_{n_u} \quad \text{[usant (9)]} \\ &= (Z_u e_{n_u} - \mu X_u^{-1} e_{n_u}) + (\mu F^{-1} e_{n_u} - W e_{n_u}) \\ &= 0 \quad \text{[usant (18) i (19)]} \end{aligned}$$

• per la seva banda, per (11) observem que:

$$\begin{aligned} 0 &\stackrel{?}{=} c_l - \mu X_l^{-1} e_{n_l} - A_l^t \\ &= A_l^t y + z_l - \mu X_l^{-1} e_{n_l} - A_l^t y \quad \text{[usant (14)]} \\ &= Z_l e_{n_l} - \mu X_l^{-1} e_{n_l} \quad \text{[usant (9)]} \\ &= 0 \quad \text{[usant (8)]} \end{aligned}$$

Per tant, les condicions finals de Kuhn-Tucker de primer ordre que ha de verificar un punt per ser considerat òptim del nostre problema seran les sis següents:

$$X_u Z_u e_{n_u} = \mu e_{n_u} \tag{20}$$

$$X_l Z_l e_{n_l} = \mu e_{n_l} \tag{21}$$

$$A_u x_u + A_l x_l = b \tag{22}$$

$$A_l^t y + z_l = c_l \tag{23}$$

$$FW e_{n_u} = \mu e_{n_u} \tag{24}$$

$$A_u^t y + z_u - w = c_u \tag{25}$$

Clarament, quan $n_u = 0$ ($n = n_l$, és a dir, no hi ha cap variable amb fita superior) només cal considerar les equacions (21, 22 i 23), les quals són, efectivament, les condicions d'optimalitat del mètode primal-dual estàndard sense fites a les variables (com es pot veure a Arbel(1993)).

2.4. Solució del sistema no-lineal

El sistema no-lineal d'equacions (20–25) resultant es resoldrà usant el mètode de Newton, on el sistema lineal d'equacions $J^i d^i = -f^i$ —essent J^i el jacobià del sistema, d^i la direcció de Newton i f^i l'avaluació del sistema al punt actual— a ser resolt a cada iteració i és:

(26)

$$\begin{array}{|cccccc|}
\hline
Z_n & & & X_u & & \\
\hline
 & Z_l & & & X_l & \\
\hline
A_u & A_l & & & & \\
\hline
 & & A_l^t & & \mathbb{I}_{n_l} & \\
\hline
-W & & & & & F \\
\hline
 & & A_u^t & \mathbb{I}_{n_u} & & -\mathbb{I}_{n_u} \\
\hline
\end{array}
\begin{array}{|c|}
\hline
dx_u \\
\hline
dx_l \\
\hline
dy \\
\hline
dz_u \\
\hline
dz_l \\
\hline
dw \\
\hline
\end{array}
=
\begin{array}{|c|}
\hline
\mu e_{n_u} - X_u Z_u e_{n_u} \\
\hline
\mu e_{n_l} - X_l Z_l e_{n_l} \\
\hline
b - Ax \\
\hline
c_l - A_l^t y - z_l \\
\hline
\mu e_{n_u} - WFe_{n_u} \\
\hline
c_u - A_u^t y - z_u + w \\
\hline
\end{array}$$

Definint els nous vectors (que representen els termes independents del sistema lineal anterior):

$$b_{1_l} = \mu e_{n_l} - X_l Z_l e_{n_l} \tag{27}$$

$$b_{1_u} = \mu e_{n_u} - X_u Z_u e_{n_u} \tag{28}$$

$$b_2 = \mu e_{n_u} - WFe_{n_u} \tag{29}$$

$$b_3 = b - Ax \tag{30}$$

$$b_{4_l} = c_l - A_l^t y - z_l \tag{31}$$

$$b_{4_u} = c_u - A_u^t y - z_u + w \tag{32}$$

s'arriba a que la solució de (26) ve donada per:

$$(ASA^t)dy = b_3 + ASr \tag{33}$$

$$dx = S(A^t dy - r) \tag{34}$$

$$dw = F^{-1}(b_2 + W dx_u) \tag{35}$$

$$dz_u = b_{4_u} + dw - A_u^t dy \tag{36}$$

$$dz_l = b_{4_l} - A_l^t dy \tag{37}$$

essent

$$\begin{aligned} r &= (r_u^t, r_l^t)^t \quad r \in \mathbb{R}^n \quad r_u \in \mathbb{R}^{n_u} \quad r_l \in \mathbb{R}^{n_l} \\ r_u &= F^{-1}b_2 + b_{4_u} - X_u^{-1}b_{1_u} \quad r_l = b_{4_l} - X_l^{-1}b_{1_l} \end{aligned} \tag{38}$$

i

$$\begin{aligned} S &= \begin{pmatrix} S_u & 0 \\ 0 & S_l \end{pmatrix} \quad S \in \mathbb{R}^{n\times n}, \quad S_u \in \mathbb{R}^{n_u \times n_u}, \quad S_l \in \mathbb{R}^{n_l \times n_l u} \\ S_u &= FX_u(Z_u F + X_u W)^{-1} \qquad S_l = Z_l^{-1} X_l \end{aligned} \tag{39}$$

on S_u i S_l poden ser calculades directament, donat que no són més que productes i sumes de matrius diagonals. A l'apèndix 1 es pot trobar el procés de reducció de (26) a la seqüència (33–37).

El codi presentat soluciona a cada iteració el sistema (26) a través de (33–37), a diferència d'altres —com Vanderbei(1992)— on (26) es soluciona usant tècniques per matrius simètriques indefinides basades en la factorització de Bunch-Parlett (Duff *et al*(1986), Vanderbei i Carpenter(1993), Vanderbei(1994)).

També cal fer notar que en la solució del nostre sistema cal invertir les matrius diagonals X, Z, W, F definides a (9). Naturalment això només és possible si cap element diagonal és igual a 0. Per tant, per poder garantir l'obtenció de (33–37) cal que les variables primals i duals no arribin mai a les seves fites (inferiors o superiors), és a dir, hem d'obligar a tenir sempre punts estrictament interiors respecte les seves fites.

2.5. Actualització del nou punt i del paràmetre μ de penalització

Un cop s'han calculat les direccions (33–37), cal actualitzar les noves variables primals i duals per a la següent iteració. És a dir, calcularem:

$$
\begin{aligned}
x^{(i+1)} &= x^{(i)} + \alpha_p dx \\
y^{(i+1)} &= y^{(i)} + \alpha_d dy \\
z^{(i+1)} &= z^{(i)} + \alpha_d dz \\
w^{(i+1)} &= w^{(i)} + \alpha_d dw
\end{aligned}
\tag{40}
$$

on α_p i α_d són tals que preserven el fet de que cada iterat sigui un punt interior (és a dir, α_p manté que $0<x_l^{(i+1)}$ i $0<x_u^{(i+1)}<\bar{x}_u$, mentre que α_d garanteix que $0<z$ i $0<w$).

Només queda fer l'actualització del paràmetre μ de la barrera logarísmica abans de tornar a calcular les direccions (33–37) pel nou iterat calculat prèviament. Com és habitual als mètodes primal-dual, el paràmetre μ s'actualitza en funció de la distància que hi ha al punt actual entre la funció primal i la funció dual (que anomenarem *gap dual*). Al nostre cas concret, tenim que:

$$
\begin{aligned}
\text{[usant (2) i (4)]} \qquad \textit{gap dual} &= c^t x - (b^t y - \bar{x}_u^t w) \\
\text{[usant (2)]} \qquad &= c^t x - x^t A^t y + \bar{x}_u^t w \\
&= (c^t - y^t A)x + \bar{x}_u^t w \\
\text{[usant (4)]} \qquad &= (z_u^t - w^t \quad z_l^t) \begin{pmatrix} x_u \\ x_l \end{pmatrix} + \bar{x}_u^t w \\
&= z_u^t x_u + z_l^t x_l - x_u^t w + \bar{x}_u^t w \\
\text{[usant (2)]} \qquad &= z^t x + f^t w \\
\text{[usant (20), (21) i (24)]} \qquad &= \mu n + \mu n_u \\
&= (n + n_u)\mu \\
&\Downarrow \\
\mu &= \frac{\textit{gap dual}}{(n + n_u)}
\end{aligned}
$$

Per tal de que a la següent iteració es millori el gap dual (garantint la convergència de l'algorisme), es prendrà com nou paràmetre $\mu^{(i+1)}$ una fracció de l'anterior. És a dir:

$$
\mu^{(i+1)} = \sigma \frac{\textit{gap dual}}{(n + n_u)} \quad \text{on} \quad 0 < \sigma < 1 \tag{41}
$$

3. IMPLEMENTACIÓ DE L'ALGORISME

En aquesta secció es detallaran alguns aspectes de la implementació de l'algorisme realitzada. Els cinc punts concrets a tractar seran la forma de solucionar el sistema (33) a cada iteració, com escollir el punt inicial $(x^{(0)}, y^{(0)}, z^{(0)}, w^{(0)})$, com calcular

les passes α_p i α_d de (40), com fer l'actualització del paràmetre μ de la barrera logarísmica i quines són les condicions d'aturada del procés iteratiu de minimització.

3.1. Obtenció de *dy* a cada iteració

El pas més costós, des del punt de vista computacional, de tot l'algorisme és l'obtenció de *dy* a través del sistema (33). Donat que es garanteix que en tot moment el punt actual és interior respecte les fites simples de les variables, tenim que la matriu ASA^t serà simètrica i definida positiva, sempre i quan A sigui de rang complet (és a dir, sempre i quan rang(A)= m). D'igual forma, per a la majoria de problemes la matriu A acostuma a ser força esparsa, i sovint aquesta propietat es manté en construir ASA^t. Per tant sembla raonable aplicar tècniques de resolució de sistemes especialment dissenyades per a matrius simètriques definides positives i esparses (és a dir, s'emprarà una factorització de Cholesky per a matrius esparses i de gran dimensió). Concretament, la implementació realitzada utilitza el paquet SPARSPAK (George i Liu(1981)) dissenyat per aquest tipus de sistemes.

En fer la factorització de Cholesky de ASA^t cal tenir molt present l'ordenació de les seves files (o columnes), donat que una mala ordenació pot donar lloc a que es creïn un gran nombre d'elements no-zero, tot degradant l'esparsitat original de la matriu. Per tant, abans de fer la factorització, cal reordenar les files de ASA^t de forma que garantim la creació del menor nombre possible d'elements. Hi ha dos possibles mètodes a aplicar per tal de trobar aquesta ordenació millor: l'ordenació del *mínim grau* ("mínimum-degree ordering"), i la del *mínim ompliment* ("mínimum-local-fill-in ordering") (Duff *et al*(1986)). Al codi desenvolupat s'ha usat la rutina GENQMD del paquet SPARSPAK, la qual implementa l'algorisme del mínim grau.

Cal també observar que la topologia de la matriu ASA^t no varia durant tota l'execució del programa. Tan sols hi ha una variació en els coeficients de la matriu diagonal S definida a (39). Aquest fet és clau, donat que implica que només cal fer una única reordenació de files mitjançant l'algorisme del mínim grau al principi, en comptes d'una vegada a cada iteració. D'igual forma també es pot explotar el fet de tenir una topologia constant fent una factorització simbòlica prèvia del sistema. Això ens proporciona el patró d'esparsitat del sistema un cop ja factoritzat, i agilitza clarament el posterior procés de càlcul.

Per tal de calcular de forma eficient ASA^t (només la part sub i diagonal, però, ja que és una matriu simètrica) el codi usa una variació de la idea proposada a Monma i Morton(1987). Considerem que la part subdiagonal de ASA^t s'emmagatzema per columnes al vector `lnz(maxlnz)`, on `maxlnz` ès el nombre d'elements no nuls de la part subdiagonal de ASA^t un cop realitzada la factorització simbòlica, mentre que els elements diagonals es troben al vector `diag(m)`. El nombre d'elements que ori-

ginalment es tenen a ASA^t (abans de factoritzar) correspon al nombre de parells de files de A $(A_{i_1,.}, A_{i_2,.})$ tal que alguna variable x_j apareix amb un coeficient no nul en ambdues constriccions i_1 i i_2 (és a dir $\exists\ j : a_{i_1,j} \neq 0$ i $a_{i_2,j} \neq 0$). Considerem que el nombre de parells de files que donen lloc a un element no nul de la part sub i diagonal de ASA^t és `np`. Llavors tindrem un vector `ilnz(np)` que ens dirà per a cada parell de files amb elements coincidents quina posició ocupa dins els vectors `lnz(.)` o `diag(.)`, tenint en compte la informació de la factorització simbòlica prèviament feta. Per tal de que `ilnz(.)` mantingui una única numeració, els vectors `lnz(.)` i `diag(.)` s'emmagatzemen a memòria de forma consecutiva, dins un vector que podem anomenar `asa(maxlnz+m)`, i llavors aquelles posicions i tals que $1 \leq$ `ilnz(i)` $\leq$ `maxlnz` correspondran a la part subdiagonal, mentre que aquelles on `maxlnz` $<$ `ilnz(i)` $\leq$ `maxlnz+m` estaran associades a la part diagonal. També cal un vector `ifillin(nfill)`, on `nfill` és el nombre d'elements no-zero creats a la part subdiagonal en factoritzar (és a dir, `nfill` = `maxlnz` $-$ (`np` $-$ `m`)). Aquest vector ens dóna les posicions dins `lnz(.)` dels nous elements creats, les quals han de ser inicialitzades a zero a cada iteració. A més, disposarem de tres vectors més anomenats `ka(np+1)`, `la(.)` i `va(.)`. El vector `ka(.)` per cada parell de files (i_1, i_2) de A que donen lloc a un element no-zero té un punter als vectors `la(.)` i `va(.)`. El primer d'aquests, `la(.)`, ens diu quines són les columnes j amb elements no-zero a les dues files i_1 i i_2 que estem considerant, mentre que `va(.)` ens dóna directament el producte $a_{i_1,j} \cdot a_{i_2,j}$. El nombre total de columnes amb elements no-zero d'un parell de files es troba fent `ka(i+1)` $-$ `ka(i)`, essent i l'entrada associada a aquesta parella de files dins el vector `ka(.)`. Per exemple, si considerem la matriu de constriccions A següent:

$$A = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix} \\ \begin{pmatrix} & 3 & 2 & 1 & & \\ 3 & & 5 & & 1 & \\ 2 & 3 & -1 & & & 1 \end{pmatrix} \end{array}$$

on `m` $= 3$, `np` $= 6$, `maxlnz` $= 3$ i `nfill` $= 0$, tenim que els vectors anteriorment definits són:

`(i)`	=	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
`ka(i)`	=	1	4	5	7	10	12	16								
`la(i)`	=	2	3	4	3	2	3	1	3	5	1	3	1	2	3	6
`va(i)`	=	9	4	1	10	9	−2	9	25	1	6	−5	4	9	1	1
`ilnz(i)`	=	4	1	2	5	3	6									

Amb aquesta estructura de dades, considerant que la matriu diagonal S es troba al vector `s(n)` de dimensió nombre de columnes de A, la creació de la part sub i

diagonal de ASA^t es redueix simplement a fer:

```
per i  = 1 fins nfill fer
       asa(ifillin(i)) = 0.0
fi-per
per i  = 1 fins np fer
       acum = 0.0
       per j  = ka(i) fins ka(i+1) − 1 fer
              acum = acum + va(j)· s(la(j))
       fi-per
asa(ilnz(i)) = acum
fi-per
```

Tot i que la matriu ASA^t es suposa definida positiva, podria ser que, si A no és de rang complet, s'arribés a tenir pivots diagonals amb valor nul en fer la descomposició de Cholesky. Aquest fet podria ser evitat si s'usés pivotació parcial en escollir el pivot, però això ens modificaria l'esparsitat de la descomposició desaprofitant la factorització simbòlica feta prèviament. El que realment s'ha fet, seguint una recomanació de Monma i Morton(1987), ha estat, en comptes de solucionar el sistema $(ASA^t)x = b$, solucionar el sistema escalat $(A(S/\gamma)A^t)x = b/\gamma$, on $\gamma = \max\{S_{ii}, i = 1,\ldots,m\}$. Si en solucionar el nou sistema escalat trobem un pivot nul, automàticament li assignem un valor petit (p.e, 10^{-12}), i continuem la factorització. Això és equivalent a introduir una petita perturbació dins la matriu A de forma que eliminem el fet de tenir alguna fila combinació lineal d'altres. Els resultats obtinguts amb aquesta regla han estat força satisfactoris. Tanmateix, a més d'aquesta simple estratègia, s'han provat d'altres de més acurades, com ara la triangularització de Gill-Murray (Gill, Murray i Wright(1981)), sense haver obtingut un millor resultat.

3.2. Elecció del punt inicial

Un dels punts clau dins de l'algorisme és el càlcul del punt inicial d'iteració. Aquest punt, que ha de ser estrictament interior, s'inicialitza atenent a dos criteris principals. El primer és el de no fixar variables a un valor que estigui a prop de les seves fites, donat que llavors tindríem un punt "poc interior". El segon és que, donat que el punt òptim ha de satisfer les condicions (20–25), interessa que el punt inicial satisfaci ja d'entrada el màxim nombre de condicions d'optimalitat possible. Dit això, el criteri seguit a l'hora de trobar el punt inicial és el següent:

a) Variables x: $x_l = 100$ i $x_u = \min\left\{100, \dfrac{\bar{x}_u}{2}\right\}$.

b) Variables y: $y = 0$ (y són variables lliures i poden tenir qualsevol valor).

c) Variables z i w: donat que hem fixat $y = 0$, llavors les condicions de factibilitat dual (23) i (25) queden directament com

$$\begin{aligned} z_l &= c_l \\ z_u - w &= c_u \end{aligned}$$

La primera condició es podrà satisfer si $\forall i \; c_{l_i} > 0$, ja que sinó tindríem una component de z_l no-interior. A més, no només interessa que totes les components de c_l siguin positives, sinó que també cal que estiguin lluny de zero per evitar tenir un punt "poc interior". Per tal de garantir això, z_l s'inicialitza com $z_{l_i} = \max\{c_{l_i}, 100\}$. La segona condició sempre la satisfarem inicialitzant w i z_u com segueix:

$$w_i, z_{u_i} = \begin{cases} w_i = 100, z_{u_i} = c_{u_i} + 100 & \text{si } c_{u_i} \geq 0 \\ z_{u_i} = 100, w_i = 100 - c_{u_i} & \text{si } c_{u_i} < 0 \end{cases}$$

3.3. Càlcul de les passes α_p i α_d

Les passes α_p i α_d usades a (40) es calculen de forma que en tot moment es mantinguin els valors de les variables entre fites. A més, sempre que sigui possible fer-ho sense violar les fites de les variables, s'intenta que prenguin el valor de 1, ja que així estaríem resolent el sistema lineal (26) usant la direcció exacta del mètode de Newton. Per tal de garantir ambdós objectius α_p i α_d són calculades com:

$$\alpha_p = \min\left\{\rho \cdot \min\{\frac{-x_i}{dx_i} \; \forall \; i : dx_i < 0\}, \rho \cdot \min\{\frac{\overline{x}_u - x_{u_i}}{dx_{u_i}} \; \forall \; i : dx_{u_i} > 0\}, 1\right\}$$

(42)

$$\alpha_d = \min\left\{\rho \cdot \min\{\frac{-z_i}{dz_i} \; \forall \; i : dz_i < 0\}, \rho \cdot \min\{\frac{-w_i}{dw_i} \; \forall \; i : dw_i < 0\}, 1\right\}$$

on ρ és un valor proper a 1, usat per treure fora de la seva fita la variable que ens proporciona la passa màxima. A la implementació feta ρ=0.99.

3.4. Actualització del paràmetre μ

L'expressió de l'actualització del paràmetre μ a la nova iteració i+1 venia donada per l'equació (41) i era $\mu^{(i+1)} = \sigma \cdot gap\ dual/(n + n_u)$, on σ havia de ser un valor tal que $0 < \sigma < 1$. En general un valor de σ=0.1 acostuma a ser prou satisfactori. Tanmateix

s'ha usat una actualització dinàmica de μ, tal i com es descriu a Vanderbei(1994). Sigui $\alpha_{pd} = \min\{\alpha_p, \alpha_d\}$, α_p i α_d definits a (42), llavors el paràmetre σ es troba a cada iteració com:

$$\sigma = \left(\frac{1 - \alpha_{pd}}{10\alpha_{pd} + 1} \right)^2 \tag{43}$$

Això fa que, quan α_{pd} és proper a 1 (és a dir, no ha estat necessari escurçar o bé α_p o bé α_d a causa de les fites) tenim que σ tendeix a 0, amb el qual disminuirem $\mu^{(i+1)}$. Per l'altra banda, quan α_{pd} s'apropa a 0 (alguna variable està a prop de la seva fita) augmentem $\mu^{(i+1)}$. Donat que el paràmetre μ controla la barrera logarísmica, actualitzant σ d'aquesta forma estem fent que, quan les variables estan lluny de les fites, disminuim la barrera logarísmica, mentre que quan alguna està a prop incrementem el seu pes dins el Lagrangià.

3.5. Condicions d'aturada

Les condicions d'aturada implementades són les habituals dels mètodes primal-dual: un punt es considera òptim quan supera un test de factibilitat primal, un de factibilitat dual i un sobre el *gap dual* entre les funcions objectiu primal i dual. Concretament les tres condicions a complir són:

$$\begin{aligned}
\frac{\|Ax - b\|}{1 + \|b\|_2} &< \varepsilon_p \qquad \text{Factibilitat Primal} \\
\frac{\left\| \begin{pmatrix} A_u^t y + z_u - w \\ A_l^t y + z_l \end{pmatrix} - c \right\|_2}{1 + \|c\|_2} &< \varepsilon_d \qquad \text{Factibilitat Dual} \\
\frac{|c^t x - (b^t y - x_u^t w)|}{1 + c^t x} &< \varepsilon_g \qquad \text{Gap Dual}
\end{aligned} \tag{44}$$

A la implementació realitzada els valors usats per ε_p, ε_d i ε_g són $\varepsilon_p = \varepsilon_d = 10^{-6}$ i $\varepsilon_g = 10^{-8}$.

3.6. Resultats computacionals

Presentem en aquesta secció els resultats computacionals obtinguts amb la implementació feta de l'algorisme presentat en apartats anteriors. S'han usat un total de 80 problemes de programació lineal pertanyents a la bateria de tests de la Netlib (Gay(1985)). Tots aquests tests han estat agafats a través d'un ftp anònim al compte

Taula I

Característiques dels problemes test

Nom	Constr.	Vars.	No-zeros A
25fv47	822	1571	11127
80bau3b	2263	9799	29063
adlittle	57	97	465
afiro	28	32	88
agg	489	163	2541
agg2	517	302	4515
agg3	517	302	4531
bandm	306	472	2659
beaconfd	174	262	3476
blend	75	83	521
bnl1	644	1175	6129
bnl2	2325	3489	16124
boeing1	351	384	3865
boeing2	167	143	1339
bore3d	234	315	1525
brandy	221	249	2150
czprob	930	3523	14173
d2q06c	2172	5167	35674
d6cube	416	6184	43888
degen2	445	534	4449
degen3	1504	1818	26230
e226	224	282	2767
etamacro	401	688	2489
fffff800	525	854	6235
finnis	498	614	2714
fit1d	25	1026	14430
fit1p	628	1677	10894
fit2d	26	10500	138018
fit2p	3001	13525	60784
ganges	1310	1681	7021
gfrd-pnc	617	1092	3467
greenbea	2393	5405	31499
grow15	301	645	5665
grow22	441	946	8318
grow7	141	301	2633
israel	175	142	2358
kb2	44	41	291
lotfi	154	308	1086
maros	847	1443	10006
maros-r7	3137	9408	151120

Nom	Constr.	Vars.	No-zeros A
nesm	663	2923	13988
pilot	1442	3652	43220
pilot87	2031	4883	73804
pilotnov	976	2172	13129
recipe	92	180	752
sc105	106	103	281
sc205	206	203	552
sc50a	51	48	131
sc50b	51	48	119
scagr25	472	500	2029
scagr7	130	140	553
scfxm1	331	457	2612
scfxm2	661	914	5229
scfxm3	991	1371	7846
scorpion	389	358	1708
scrs8	491	1169	4029
scsd1	78	760	3148
scsd6	148	1350	5666
scsd8	398	2750	11334
sctap1	301	480	2052
sctap2	1091	1880	8124
sctap3	1481	2480	10734
seba	516	1028	4874
share1b	118	225	1182
share2b	97	79	730
shell	537	1775	4900
ship04l	403	2118	8450
ship04s	403	1458	5810
ship08l	779	4283	17085
ship08s	779	2387	9501
ship12l	1152	5427	21597
ship12s	1152	2763	10941
sierra	1228	2036	9252
standata	360	1075	3038
standgub	362	1184	3147
standmps	468	1075	3686
stocfor1	118	111	474
stocfor2	2158	2031	9492
woodlp	245	2594	70216
woodw	1099	8405	37478

netlib.att.com, i es troben dins del directory */netlib/lp/data*. Només s'han usat problemes sense variables lliures a la seva formulació, donat que la implementació feta no contempla aquesta possibilitat. La Taula I mostra les principals característiques dels problemes test usats. La taula presenta per a cada problema el seu nom (columna

"Nom"), el nombre de constriccions del problema (columna "Constr."), el nombre de variables (columna "Vars."), i el nombre d'elements no-zero de la matriu de constriccions (columna "No-zeros A").

Taula II

Resultats obtinguts amb els problemes test

	IP			LoQo			Minos 5.3		
Nom	Valor òptim	Iter.	CPU	Valor òptim	Iter.	CPU	Valor òptim	Iter.	CPU
25fv47	5501.8	45	23.9	5501.8	29	19.8	5501.8	6477	75.8
80bau3b	987224.2	71	61.3	987224.2	44	52.4	987229.0	11158	239.2
adlittle	225495.0	20	0.1	225495.0	14	0.1	225495.0	108	0.3
afiro	464.8	13	0.0	464.8	13	0.0	464.8	14	0.1
agg	35991767.3	43	6.8	35991767.3	26	1.9	35991767.3	158	1.6
agg2	20239252.3	37	9.5	20239252.3	22	4.4	20239252.4	223	2.2
agg3	10312115.9	39	9.8	10312115.9	22	4.4	10312115.9	255	2.2
bandm	158.6	34	1.5	158.6	20	1.6	158.6	474	2.7
beaconfd	33592.5	18	1.3	33592.5	14	1.2	33592.5	95	1.1
blend	30.8	20	0.1	30.8	14	0.2	30.8	115	0.4
bnl1	1977.6	51	7.5	1977.6	35	7.1	1977.6	1166	9.5
bnl2	1811.2	61	124.7	1811.2	40	122.4	1811.2	5441	125.0
boeing1	335.2	44	4.0	335.2	28	3.1	335.2	579	3.0
boeing2	315.0	32	0.8	315.0	28	1.1	315.0	159	0.7
bore3d	1373.1	29	0.9	1373.1	17	0.9	1373.1	118	1.1
brandy	1518.5	29	1.2	1518.5	22	1.5	1518.5	310	1.7
czprob	2185196.7	56	12.8	2185196.7	38	10.9	2185196.7	1506	17.8
d2q06c	122784.2	58	263.6	122784.2	36	247.0	122784.2	43934	1246.4
d6cube	315.5	51	80.8	315.5	23	75.2	315.5	96780	1049.1
degen2	1435.2	26	6.7	1435.2	14	4.7	1435.2	974	6.5
degen3	987.3	38	165.6	987.3	17	73.4	987.3	7109	158.4
e226	18.8	39	1.7	18.8	22	1.4	18.8	455	1.9
etamacro	755.7	49	5.6	755.7	30	8.9	755.7	761	4.1
fffff800	555679.6	78	17.8	555679.6	36	10.6	555679.6	272	3.3
finnis	172791.1	42	2.7	172791.1	26	2.2	172791.0	547	3.4
fit1d	9146.4	56	4.8	9146.4	21	5.3	9146.4	2626	6.4
fit1p	9146.4	26	381.6	9146.4	26	4.8	9146.4	894	13.6
fit2d	68464.3	48	44.7	68464.3	24	195.0	68464.3	32235	291.2
fit2p	*(a)*			68464.3	24	41.1	68464.3	14240	1128.3
ganges	109585.7	41	15.2	109585.7	23	10.7	109585.8	663	9.5
gfrd pnc	6902236.0	55	1.9	6902236.0	19	1.6	6902236.0	686	5.7
greenbea	*(b)*			72555247.5	47	86.9	72462405.9	24537	645.5
grow15	106870941.1	24	2.0	106870941.3	23	3.0	106870941.3	426	4.3
grow22	160834336.2	25	3.3	160834336.5	28	5.2	160834336.5	676	8.0
grow7	47787811.8	23	0.8	47787811.8	20	1.2	47787811.8	175	1.4
israel	896644.8	71	12.3	896644.8	28	1.2	896644.8	250	1.2
kb2	1749.9	28	0.1	1749.9	21	0.2	1749.9	50	0.2
lotfi	25.3	29	0.5	25.3	25	0.8	25.3	205	0.8
maros	58063.7	65	22.2	58063.7	28	14.0	58063.7	1910	20.0
maros r7	*(a)*			1497185.2	22	3603.3	*(c)*		

(Continua)

Taula II (cont.)

Resultats obtinguts amb els problemes test

Nom	IP			LoQo			Minos 5.3		
	Valor òptim	Iter.	CPU	Valor òptim	Iter.	CPU	Valor òptim	Iter.	CPU
nesm	14076036.6	65	19.4	14076036.5	38	22.2	14076073.1	3061	26.4
pilot	557.5	72	520.5	557.5	44	463.3	557.4	15996	738.5
pilot87	301.7	79	2312.7	301.6	46	1923.7	301.7	24284	3041.2
pilotnov	4497.3	40	40.1	4497.3	29	39.7	4497.3	2525	34.9
recipe	266.6	18	0.1	266.6	13	0.3	266.6	27	0.4
sc105	52.2	15	0.1	52.2	14	0.1	52.2	44	0.3
sc205	52.2	17	0.2	52.2	17	0.4	52.2	77	0.6
sc50a	64.6	15	0.0	64.6	14	0.1	64.6	28	0.2
sc50b	70.0	12	0.0	70.0	13	0.1	70.0	15	0.2
scagr25	14753433.1	47	1.4	14753433.1	18	1.0	14753433.1	319	2.4
scagr7	2331389.8	23	0.2	2331389.8	15	0.2	2331389.8	106	0.5
scfxm1	18416.8	38	1.7	18416.8	26	1.9	18416.8	361	2.1
scfxm2	36660.3	46	4.8	36660.3	28	4.2	36660.3	753	6.5
scfxm3	54901.3	46	8.4	54901.3	28	6.4	54901.3	1097	13.4
scorpion	1878.1	23	0.7	1878.1	16	0.7	1878.1	134	1.3
scrs8	904.3	37	2.5	904.3	23	2.9	904.3	682	5.3
scsd1	8.7	14	0.3	8.7	15	0.9	8.7	359	1.8
scsd6	50.5	17	0.8	50.5	17	1.6	50.5	1106	5.2
scsd8	905.0	17	1.8	905.0	17	3.4	905.0	3026	25.6
sctap1	1412.3	30	0.8	1412.3	18	0.9	1412.2	257	1.6
sctap2	1724.8	33	7.7	1724.8	16	4.9	1724.8	661	9.0
sctap3	1424.0	36	11.9	1424.0	16	6.2	1424.0	953	20.2
seba	15711.6	40	68.4	15711.6	19	2.1	15711.6	366	3.5
share1b	76589.3	89	1.0	76589.3	26	0.7	76589.3	250	0.8
share2b	415.7	20	0.2	415.7	14	0.2	415.7	111	0.4
shell	1208825347.3	32	1.6	1208825346.9	25	3.0	1208825346.0	321	4.5
ship04l	1793324.5	21	2.4	1793324.6	19	2.9	1793324.5	290	4.3
ship04s	1798714.7	24	1.6	1798714.7	19	2.1	1798714.7	157	2.9
ship08l	1909055.2	27	6.9	1909055.2	20	6.5	1909055.2	473	10.0
ship08s	1920098.2	24	3.3	1920098.2	20	3.7	1920098.2	256	5.3
ship12l	1470187.9	25	10.0	1470187.9	24	9.7	1470187.9	961	18.6
ship12s	1489236.1	25	4.8	1489236.1	22	4.8	1489236.1	437	9.0
sierra	15394362.2	25	6.6	15394362.2	21	5.9	15394362.2	864	12.2
standata	1257.7	29	1.4	1257.7	23	1.9	1257.7	131	2.3
standgub	1257.7	29	1.4	1257.7	23	2.0	1257.7	105	2.2
standmps	1406.0	42	3.7	1406.0	32	3.4	1406.0	383	3.0
stocfor1	41132.0	22	0.2	41132.0	17	0.2	41132.0	92	0.4
stocfor2	39024.4	45	19.6	39024.4	31	11.7	39024.4	3092	69.8
woodlp	1.4	38	57.1	1.4	28	53.8	1.4	870	21.1
woodw	1.3	51	49.1	1.3	27	41.1	1.3	3572	76.5
PROMIG		37	58.1		23	46.0		3747	97.5

(a) No hi ha suficient memòria per executar aquest problema.

(b) Problemes de convergència.

(c) Error numèric. No es poden satisfer les constriccions de forma acurada.

La implementació feta (a la qual ens referirem com IP) ha esta codificada en ANSI−C i s'ha comparat amb dos codis capdavanters de programació lineal. El primer d'ells és el paquet Minos 5.3 (Murtagh i Saunders(1983)), el qual està basat en el mètode del símplex per a programació lineal. El segon és el paquet LoQo (Vanderbei(1992)), una implementació eficient d'un mètode primal-dual per a programació lineal i quadràtica. Totes les execucions s'han realitzat sobre una SunSparc 10/41 sota UNIX, amb un únic processador RISC de 40Mhz i aproximadament 20Mflops, i amb 64 Mbytes de memòria (32Mbytes reals i 32Mbytes mapejats a disc). La Taula II mostra els resultats obtinguts amb cada paquet. Per a cada un dels tres codis (IP, LoQo i Minos 5.3) es presenta el valor òptim assolit (columna "Valor òptim"), el nombre d'iteracions realitzat (columna "Iter.") i el temps de CPU en segons requerit (columna "CPU").

Es pot observar a la Taula II com els tres codis en tot moment van assolir els mateixos valors òptims (hi ha petites diferències, però són menyspreables). També es pot veure que IP i LoQo realitzen moltes menys iteracions que Minos, degut a que estan basats en un mètode de punt interior en comptes de en el símplex. I comparant només aquests dos, en general IP realitza més iteracions que LoQo. El codi desenvolupat (IP) no va poder executar dos problemes ("fit2p" i "maros-r7") degut a problemes de memòria. LoQo no va tenir aquests problemes perquè no resol el sistema (26) a través de (33–37), sinó directament aplicant tècniques per a matrius simètriques i indefinides. Aquest fet explica també el comportament tan diferent que tenen ambdós codis en alguns exemples (per exemple, als problemes "fit1p" i "fit2d"). IP va tenir problemes de convergència al problema "greenbea", mentre que Minos no va poder satisfer les constriccions de forma acurada al test "maros-r7". La darrera fila de la Taula II mostra el valor promig per a cada codi pel que fa al nombre d'iteracions i temps de CPU. Pot observar-se com els dos codis de punt interior es mostren més eficients que Minos. També es veu que LoQo té un millor rendiment que IP. Tanmateix IP realitza moltes més iteracions que LoQo, el qual indica que efectivament IP té un cost per iteració menor que LoQo, com s'havia previst inicialment per la forma especial de solucionar el sistema (26).

El dit anteriorment sobre el rendiment de cada codi pot observar-se clarament a les Fig. 1 i 2. La Fig. 1 mostra el temps de CPU (eix y) pels tres codis segons la mida del problema (eix x). Com a mida del problema s'ha agafat el nombre de constriccions de cada un (tot i que es podria haver seguit un altre criteri com el nombre de variables, d'elements no-zero de la matriu o una expressió funció d'aquests valors). Només es presenten els problemes amb una mida relativament gran (per ex., més de 500 constriccions), per simplificar la gràfica. El temps de CPU s'ha escalat logarísmicament. Pot observar-se que, en general, Minos es troba per sobre dels dos codis de punt interior, mentre que LoQo dóna un millor rendiment. IP es troba en una situació intermèdia. Per la seva banda la Fig. 2 ens mostra, per a IP i LoQo, el temps per iteració (a l'eix y com abans) segons la mida del problema (eix x). Només

es mostren, com a la figura anterior, els problemes amb més de 500 constriccions. De nou també el temps per iteració s'ha escalat logarísmicament. Pot observar-se com IP té tendència a donar millors temps per iteració. Malgrat això, el fet de que requereixi força més iteracions que LoQo per convergir a l'òptim fa que no tingui un comportament global tan eficient com aquell.

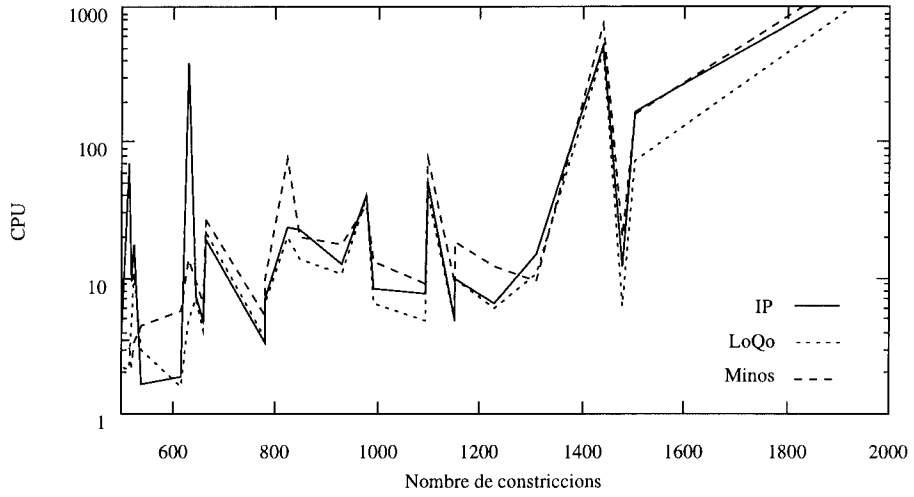


Figura 1. Temps de CPU per IP, LoQo i Minos segons la mida del problema.

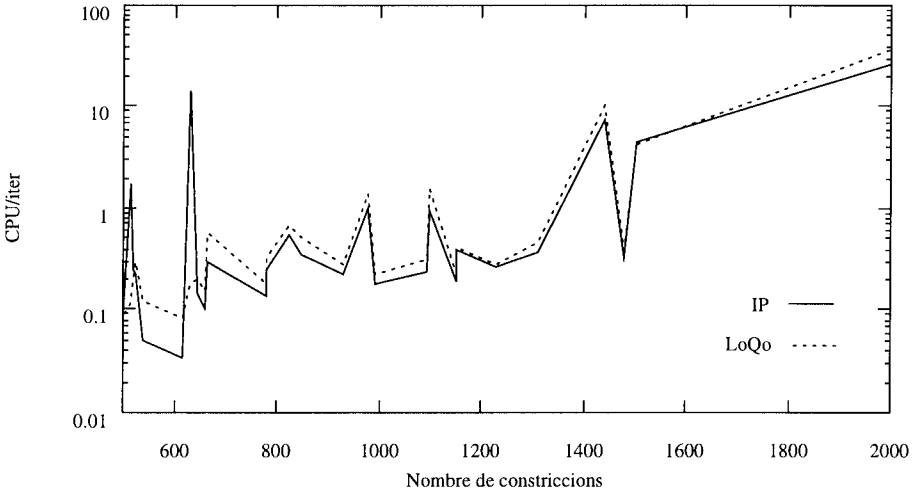


Figura 2. Temps per iteració per IP i LoQo segons la mida del problema.

5. CONCLUSIONS

Aquest document ha presentat un algorisme primal-dual de punt interior especialitzat per tractar les fites superiors a les variables de forma explícita. Una implementació de l'algorisme ha estat comparada amb dos paquets capdavanters de programació lineal. La implementació feta s'ha mostrat més eficient que Minos5.3, una implementació eficient del mètode del símplex, però menys que LoQo, basat també en un mètode primal-dual. Tanmateix, tot i que el codi desenvolupat té un major cost que LoQo, també realitza moltes més iteracions. S'ha pogut comprovar que el seu cost per iteració és menor, degut a la forma especial de calcular la direcció de Newton resolent un sistema simètric i definit positiu (LoQo resol un sistema simètric i indefinit). Això fa que es pugui explotar el fet de tenir la mateixa topologia durant tota l'execució i poder realitzar una factorització simbòlica prèvia. Queda obert el problema de millorar la convergència del codi desenvolupat pel que fa al nombre d'iteracions realitzades, amb el qual els temps de càlcul s'atansarien — i potser millorarien— els del codi LoQo.

AGRAÏMENTS

Voldria agrair la tasca dels revisors anònims, tant per la lectura tan acurada que han fet de l'article, la qual m'ha fet adonar de petits detalls que havia obviat, com per tots el seus comentaris i suggeriments.

APÈNDIX 1. SOLUCIÓ DEL SISTEMA LINEAL D'EQUACIONS PLANTEJAT A (26)

El sistema lineal d'equacions a resoldre a cada iteració a l'algorisme presentat és:

$$Z_u dx_u + X_u dz_u = b_{1_u} \tag{45}$$

$$Z_l dx_l + X_l dz_l = b_{1_l} \tag{46}$$

$$-W dx_u + F dw = b_2 \tag{47}$$

$$A_u dx_u + A_l dx_l = b_3 \tag{48}$$

$$A_u^t dy + dz_u - dw = b_{4_u} \tag{49}$$

$$A_l^t dy + dz_l = b_{4_l} \tag{50}$$

Aïllant dz_u de (49), dz_l de (50), dw de (47), dx_u de (45) i dx_l de (46) obtenim:

$$dz_u = b_{4_u} - A_u^t dy + dw \tag{51}$$

$$dz_l = b_{4_l} - A_l^t dy \tag{52}$$

$$dw = F^{-1}(b_2 + W dx_u) \tag{53}$$

$$dx_u = Z_u^{-1}(b_{1_u} - X_u dz_u) \tag{54}$$

$$dx_l = Z_l^{-1}(b_{1_l} - X_l dz_l) \tag{55}$$

Substituïnt (51) i (53) a (54) tenim que:

$$\begin{aligned} dx_u &= Z_u^{-1}(b_{1_u} - X_u[b_{4_u} - A_u^t dy + dw]) \\ &= Z_u^{-1}(b_{1_u} - X_u b_{4_u} - X_u dw + X_u A_u^t dy) \\ &= Z_u^{-1}(b_{1_u} - X_u b_{4_u} - X_u[F^{-1}(b_2 + W dx_u)] + X_u A_u^t dy) \\ &= Z_u^{-1}(b_{1_u} - X_u b_{4_u} - X_u F^{-1} b_2 - X_u F^{-1} W dx_u + X_u A_u^t dy) \\ &= Z_u^{-1} b_{1_u} - Z_u^{-1} X_u b_{4_u} - Z_u^{-1} X_u F^{-1} b_2 - Z_u^{-1} X_u F^{-1} W dx_u + Z_u^{-1} X_u A_u^t dy \end{aligned}$$

Agrupant termes s'arriba a

$$(\mathbb{1} + Z_u^{-1} X_u F^{-1} W) dx_u = Z_u^{-1} b_{1_u} - Z_u^{-1} X_u b_{4_u} - Z_u^{-1} X_u F^{-1} b_2 + Z_u^{-1} X_u A_u^t dy$$

Definint T_u com

$$T_u = \mathbb{1} + Z_u^{-1} X_u F^{-1} W$$

la seva inversa ve donada directament com

$$T_u^{-1} = F Z_u (F Z_u + X_u W)^{-1} \tag{56}$$

Llavors l'expressió anterior de dx_u pot ser escrita com

$$dx_u = T_u^{-1}(Z_u^{-1} b_{1_u} - Z_u^{-1} X_u b_{4_u} - Z_u^{-1} X_u F^{-1} b_2 + Z_u^{-1} X_u A_u^t dy) \tag{57}$$

Anàlogament, substituïnt (52) a (55) s'obté directament que

$$\begin{aligned} dx_l &= Z_l^{-1}(b_{1_l} - X_l[b_{4_l} - A_l^t dy]) \\ dx_l &= Z_l^{-1} b_{1_l} - Z_l^{-1} X_l b_{4_l} + Z_l^{-1} X_l A_l^t dy \end{aligned} \tag{58}$$

Ara, usant (58) i (59), l'equació (48) queda com segueix:

$$\begin{aligned} A_u[T_u^{-1}(Z_u^{-1} b_{1_u} - Z_u^{-1} X_u b_{4_u} - Z_u^{-1} X_u F^{-1} b_2 + Z_u^{-1} X_u A_u^t dy)] + \\ + A_l[Z_l^{-1} b_{1_l} - Z_l^{-1} X_l b_{4_l} + Z_l^{-1} X_l A_l^t dy] = b_3 \\ A_u T_u^{-1} Z_u^{-1}(b_{1_u} - X_u b_{4_u} - X_u F^{-1} b_2) + A_u T_u^{-1} Z_u^{-1} X_u A_u^t dy + \\ + A_l Z_l^{-1}(b_{1_l} - X_l b_{4_l}) + A_l Z_l^{-1} X_l A_l^t dy = b_3 \end{aligned} \tag{59}$$

Definint S_u i S_l com

$$S_u = T_u\ {}^1Z_u\ {}^1X_u = FX_u(FZ_u + X_uW)\ {}^1 \tag{60}$$

$$S_l = Z_l\ {}^1X_l \tag{61}$$

i agrupant termes, l'expressió (60) queda

$$(A_uS_uA_u^t + A_lS_lA_l^t)dy = b_3 + A_uT_u\ {}^1Z_u\ {}^1(X_uF\ {}^1b_2 + X_ub_{4_u} - b_{1_u}) + A_lZ_l\ {}^1(X_lb_{4_l} - b_{1_l})$$

$$(A_uS_uA_u^t + A_lS_lA_l^t)dy = b_3 + A_uS_u(F\ {}^1b_2 + b_{4_u} - X_u\ {}^1b_{1_u}) + A_lS_l(b_{4_l} - X_l\ {}^1b_{1_l})$$

L'expressió final del càlcul dy és:

$$(ASA^t)dy = b_3 + ASr \tag{62}$$

on r i S són respectivament:

$$r = (r_u^t, r_l^t)^t \qquad r_u = F\ {}^1b_2 + b_{4_u} - X_u\ {}^1b_{1_u} \qquad r_l = b_{4_l} - X_l\ {}^1b_{1_l} \tag{63}$$

$$S = \begin{pmatrix} S_u & 0 \\ 0 & S_l \end{pmatrix} \tag{64}$$

Un cop hem calculat dy, substituint el seu valor a (58) i (59) dóna:

$$\begin{aligned} dx_u &= T_u\ {}^1(Z_u\ {}^1b_{1_u} - Z_u\ {}^1X_ub_{4_u} - Z_u\ {}^1X_uF\ {}^1b_2 + Z_u\ {}^1X_uA_u^tdy) \\ &= T_u\ {}^1Z_u\ {}^1X_u(X_u\ {}^1b_{1_u} - b_{4_u} - F\ {}^1b_2 + A_u^tdy) \\ dx_u &= S_u(A_u^tdy - r_u) \end{aligned} \tag{65}$$

$$\begin{aligned} dx_l &= Z_l\ {}^1b_{1_l} - Z_l\ {}^1X_lb_{4_l} + Z_l\ {}^1X_lA_l^tdy \\ &= Z_l\ {}^1X_l(X_l\ {}^1b_{1_l} - b_{4_l} + A_l^tdy) \\ dx_l &= S_l(A_l^tdy - r_l) \end{aligned} \tag{66}$$

Agrupant (66) i (67) tenim que l'expressió final del calcul de dx és:

$$dx = S(A^tdy - r) \tag{67}$$

Un cop dy i dx són coneguts, dz i dw es calculen directament a través de les expressions inicials (51), (52) i (53).

BIBLIOGRAFIA

[1] **Arbel, A.** (1993). *Exploring Interior-Point Linear Programming. Algorithms and Software*. The MIT Press, Cambridge, Massachusetts.

[2] **Barnes, E.R.** (1986). "A variation on Karmarkar's algorithm for solving linear programming problems". *Mathematical Programming*, **36**, 174–182.

[3] **Duff, I.S., A.M. Erisman** i **J.K. Reid.** (1986). *Direct Methods for Sparse Matrices*. Oxford University Press, New York.

[4] **Gay, D.M.** (1985) "Electronic mail distribution of linear programming test problems". *Mathematical Programming Society COAL Newsletter*, **13**, 10–12.

[5] **George, J.A.** i **J.W.H. Liu.** (1981). *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall, Englewood Cliffs, NJ, USA.

[6] **Gill, P.E., W. Murray** i **M.H. Wright.** (1981). *Practical Optimization*. Academic Press, London, UK.

[7] **Karmarkar, N.K.** (1984). "A new polynomial time algorithm for linear programming". *Combinatorica*, **4**, 373–395.

[8] **Khachiyan, G.** (1979). "A polynomial algorithm in linear programming". *Doklady Akademii Nauk SSSR*, **244(S)**, 1093–1096, traduït en *Soviet Mathematics Doklady*, **20(1)**, 191–194.

[9] **Monma, C.L** i **A.J Morton** (1987). "Computational experience with a dual affine variant of Karmarkar's method for linear programming". *Operations Research Letters*, **6**, 261–267.

[10] **Monteiro, R.D.C** i **I. Adler** (1989). "Interior path following primal-dual algorithms. Part I: linear programming". *Mathematical Programming*, **44**, 27–41.

[11] **Murtagh, B.A.** i **M.A. Saunders** (1983). "MINOS 5.0. User's guide". Dept. of Operations Research, Stanford University, CA, USA.

[12] **Vanderbei, R.J.** (1992). "LOQO User's Manual". Princeton University, Princeton, NJ, USA.

[13] **Vanderbei, R.J.** (1994). "An interior point code for quadratic programming". Princeton University, Princeton, NJ, USA.

[14] **Vanderbei, R.J.** i **T.J. Carpenter** (1993). "Symmetric indefinite systems for interior point methods". *Mathematical Programming*, **58**, 1-32.

[15] **Vanderbei, R.J., M.S. Meketon** i **B.A. Freedman** (1986). "A modification of Karmarkar's linear programming algorithm". *Algorithmica*, **1**, 395–407.

[16] **Wright, M.H.** (1991). "Interior methods for constrained optimization". *Acta Numerica*, 341–407.

ENGLISH SUMMARY:

AN IMPLEMENTATION OF A PRIMAL-DUAL INTERIOR POINT ALGORITHM WITH UPPER-BOUNDED VARIABLES

J. Castro

This paper presents an implementation of a primal-dual interior point algorithm considering upper-bounded variables without adding additional constraints. The algorithm is described and formulated, following the traditional way of the primal-dual methods, which attempt to solve the Khun-Tucker first order optimality conditions of the primal and dual Lagrangian functions including a logarithmic barrier function. Imposing the first order optimality conditions means solving a nonlinear system of equations. This is done through Newton's method of solving sets of nonlinear equations. In the algorithm developed, Newton's direction is computed by solving a symmetric and positive definite system with matrix ASA^t, A being the constraints matrix and S a diagonal matrix where $S_{ii}>0$. In this point the code differs from other implementations which find out Newton's direction by solving a symmetric indefinite system. In this last case the sparsity pattern of the Cholesky decomposition is not known a priori, since partial pivoting must be applied. However, in the approach of the algorithm presented, since the system to be solved is symmetric and positive definite, it can be assumed that it is sufficiently well-conditioned (thus avoiding partial pivoting), and a previous symbolic factorization can be made in order to know the sparsity pattern of the Cholesky decomposition. In theory, this fact should improve the performance of the process of computing Newton's direction.

Once a brief outline of the algorithm has been made, the paper focuses on some implementation details. The data structure employed to store the information for managing and creating the sub and diagonal part of system matrix ASA^t is instrumental to ensure a good performance. Another key point is the choice of the initial iteration point of Newton's method. When using Newton's method, the step size usually employed is one. In this case, due to the variables bounds, sometimes this step will have to be shorter. The way of computing this step is also detailed. The logarithmic barrier function considered in the formulation of the primal and dual Lagrangian functions has a parameter (denoted as μ) which must be updated at each iteration, and an account is given of how to update it. Finally, a brief exposition of the optimality conditions required to consider a point as optimal is made.

The implementation developed is compared with two state of the art linear programming codes: Minos 5.3 (an implementation of the simplex method) and LoQo

(a primal-dual interior point code for linear and quadratic programming). To test the codes 80 linear programming problems taken from the *Netlib* standard collection have been employed. The results obtained show that the performance of the code developed is between those of LoQo and Minos (LoQo is the most efficient of the three). However, the code developed is faster in performing one single iteration than LoQo (as it was theoretically expected), but it has a slower convergence since it requires much more iterations. It remains as an open problem to improve its convergence to reach (even improve) the efficiency of the LoQo package.

Estadística Oficial

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 261-320, 1995

INTRODUCCIÓ ALS SISTEMES DE METAINFORMACIÓ ESTADÍSTICA

ISIDRE CANALS CABIRÓ*

D'una banda, es defineix el concepte i la tipologia de la metainformació estadística i es descriuen els problemes plantejats als instituts d'estadística per a la seva organització en sistemes integrats, de cara a la satisfacció de les necessitats dels usuaris d'informació estadística, tant externs com interns.

D'altra banda, es passa revista als diferents projectes i experiències d'altres instituts agrupats segons els enfocaments tècnics possibles, incloent-hi els plantejaments i recomanacions dels grups de treball específics d'institucions internacionals.

Finalment, es treuen algunes conclusions i criteris pràctics tant per definir una estratègia com per dissenyar un sistema global de metainformació estadística.

Aquest article s'ha de considerar introductori a la problemàtica i complementari de l'article de Sundgren, també publicat en aquest número de ***Qüestiió.***

Introduction to Statistical Metainformation Systems.

Key words: Metainformació estadística, Metadades, Sistema de metainformació estadística, Sistema d'informació estadística.

* Isidre Canals Cabiró. Institut d'Estadística de Catalunya. Dept. d'Economia i Finances. Generalitat de Catalunya. Via Laietana, 58. 08003 Barcelona.

Les opinions expressades en aquest article no reflecteixen necessàriament les de l'Institut.

- Article rebut el maig de 1995.
- Acceptat el setembre de 1995.

1. INTRODUCCIÓ A LA METAINFORMACIÓ ESTADÍSTICA

La manera més simple de definir la metainformació estadística és la d'identificar-la amb la *informació sobre la informació estadística.*

Resulta també la més interessant des de la perspectiva d'un institut d'estadística. En tot cas, més general que el significat corrent del terme metadades, quan es refereix només a la informació relativa a dades numèriques concretes (tant si es tracta de *microdades*, és a dir, dades individualitzades, com de *macrodades*, és a dir, dades agregades en forma de taules)[1].

En efecte, per a un institut d'estadística, el significat del terme "metainformació estadística" inclou també tot el que es refereix a informació de l'estadística en general (concretat, però, en la pràctica, al país de referència i a l'àmbit de les competències de l'institut de què es tracti). Plantejada així la qüestió, resoldre el problema de l'organització de la metainformació estadística resulta vital, no solament pel que fa a l'organització de la informació estadística al servei de les necessitats internes d'un institut d'estadística, sinó també pel que fa a l'articulació d'unes relacions eficients entre l'institut i la societat a la que serveix.

Així, no semblarà estrany que quan Gillman, del *U.S. Bureau of the Census*, presentava la seva visió del problema de la metainformació, l'encapçalava amb aquestes paraules, amb les que s'havia expressat oficialment quin hauria de ser l'objectiu global a perseguir a l'era de la revolució de les tecnologies de la informació i la comunicació, pel que fa a l'estadística: *"un sistema de recollida, producció i difusió de dades electròniques sense discontinuïtats entre les seves fases i íntimament entrelligat amb la societat a la que mesura"*. (Gillman 1994).

Des d'una perspectiva més operativa, dos són els punts de vista amb els que, al si d'un institut d'estadística, pot ser plantejada la metainformació estadística, i que corresponen *grosso modo*: el primer, al punt de vista dels usuaris (i de l'*ús*) de la informació estadística, reprès pels responsables i tècnics (temàtics, informàtics i especialistes d'informació) encarregats de la difusió d'estadístiques, i el segon, al punt de vista de la *producció*, encarnat pels responsables i tècnics (estadístics, temàtics i informàtics) al càrrec de la producció d'informació estadística.

Aquests dos punts de vista indueixen no solament plantejaments diferents respecte de la problemàtica de la metainformació estadística, sinó que responen a necessitats

[1]Cal advertir, però, que la tasca de recerca terminològica i d'estandardització empresa pel grup de treball METIS (de la Conferència d'Estadístics Europeus, dedicat als problemes de la metainformació estadística) distingeix entre *"statistical metainformation"* (= *An information on statistical metadata or statistical metainformation*) i *"metadata"* (= *Data representing statistical metainformation*). Vg [Prazenka(1994)]. Vg també l'apartat 2.4.2).

diferents, segons les tasques i coneixements dels usuaris (interns i externs), que requeriran informacions diferenciades sobre la mateixa qüestió. Aquests plantejaments, però, estan destinats a convergir en el futur, en la mesura que responen a funcions complementàries en tot institut d'estadística i, d'altra banda, l'organització de la informació haurà de tendir a presentar-se als usuaris de tota mena amb un únic accès i una conceptualització global consistent.

La convergència provindrà també de l'ampliació del terme "usuaris" per comprendre també (a més dels usuaris finals) tota mena d'usuaris interns. D'aquesta manera, la metainformació estadística ja no serà només aquella informació complementària, necessària per a un *ús final* correcte de les dades, sinó també (i, des del punt de vista de la gestió d'un institut d'estadística, sobretot) aquella informació necessària per a la *producció* eficient de les dades. Podríem parlar, així, d'una metainformació estadística final i d'una metainformació estadística instrumental.

Afegim que, al llarg d'aquest article, el terme "usuaris" manté una certa ambigüetat, donat que, segons el context d'on provingui la informació descrita, el seu significat pot anar del més restrictiu al més ampli. Confiem que el bon sentit del lector sàpiga discernir en cada part el significat més adequat.

1.1. La metainformació estadística des del punt de vista de l'ús de la informació estadística

Una primera consideració ens porta als orígens de la metainformació estadística, als que fa referència el terme anglès *metadata*, utilitzat per primera vegada fa una vintena d'anys per Sundgren[2] (Institut Suec d'estadística) i que identifica aquella part de la metainformació més propera a les dades pròpiament dites.

En efecte, la necessitat de les metadades hauria nascut per la preocupació (que s'ha manifestat tostemps, amb més o menys força) de que les dades estadístiques vagin sempre acompanyades de les informacions necessàries per a interpretar-les correctament. Recordem que les dades estadístiques, fins i tot les més simples, són sempre estimacions, i que només es poden determinar els límits entre els que es troba el valor real o veritable d'una variable. I, així, les notes que apareixen al peu d'una taula, o annexes a ella, constitueixen la manifestació més elemental de la metainformació estadística: definicions, característiques de la mostra d'una enquesta, diferències en la data o període cobert, mètode emprat per a la recollida o l'estimació de valors, incidències en el treball de camp, etc.[3]

[2]Bo Sundgren és reconegut com un pioner en l'estudi i plantejament conceptual de la meta-informació i els sistemes de metainformació estadística.

[3]Vg. Silver (1993), on fa un plantejament sistemàtic de les notes al peu de taules estadístiques, i on es

Del que es tracta, en definitiva, és de proporcionar a l'usuari, en una publicació estadística, a l'ensems de les dades el seu context, condicionant decisiu de la fidelitat amb la que reflecteixen el fenòmen de la realitat que pretenen medir. Només així els usuaris de l'estadística podran estar en disposició d'utilitzar- les correctament.

Aquesta materialització de la metainformació estadística en notes acompanyants de les taules estadístiques en publicacions impreses continúa essent molt important; en l'actualitat, però, cal tenir en compte dos nous mitjans de difusió de la informació estadística: l'*edició electrònica* i l'*accés interactiu a bases de dades* públiques, que hem de considerar també des de la perspectiva de la metainformació.

Comencem per constatar que ha començat amb força creixent i irreversible l'edició d'informació estadística en forma digitalitzada i en suports magnètics (cintes i disquets) i òptics (CD-ROM). I aquest fet obliga a plantejar la problemàtica de les metadades amb característiques noves. D'una banda, pel problema dels formats sota els quals hauran de ser distribuïdes la informació estadística i la metainformació associada (en el que no entrem ara, però és decisiu a llarg termini). De l'altra (sobretot), per les enormes possibilitats obertes en l'organització conjunta d'ambdues menes d'informacions en sistemes interactius, provistos de les funcionalitats vehiculades per les interfícies gràfiques d'usuari[4].

L'altre mitjà de difusió de la informació estadística consisteix a oferir al públic l'accés interactiu a bases de dades[5]. Aquests sistemes, de les seves primeres manifestacions als anys 70 ençà, han passat per vàries generacions, pel que fa a les seves funcionalitats. Avui dia, es troben en plena crisi de transformació, coexistint inharmònicament diverses formes, en conflicte i al mateix temps convergents: bases de dades documentals de recuperació d'informació clàssiques (*retrieval systems*), bases de dades relacionals (ambdues eventualment dotades d'interfícies gràfiques d'usuari), servidors *World-Wide-Web (WWW)* al si de la xarxa Internet (que utilitzen tècniques hipertext simples), etc.

Així, doncs, ens cal preguntar-nos com s'ha d'integrar la metainformació en aquests mitjans digitalitzats. La resposta haurà de tenir en compte, d'una banda, les més grans potencialitats dels sistemes informatitzats, però també, de l'altra, el fet que els usuaris de la informació estadística també han evolucionat i les seves neces-

proposa una taxonomia específica en 22 menes de notes, codificades i agrupades en 5 grups principals.

[4]Les interfícies gràfiques d'usuari (*GUI = Graphical User Interfaces*) corresponen a tot un paradigma informàtic, altrament identificat amb les lletres *MIWM*, que corresponen als seus quatre components essencials: *Menu-Icon-Window-Mouse*), també anomenats *sistemes de manipulació directa* (Shneiderman).

[5]El *Consorci d'Informació i Documentació de Catalunya*, precedent històric i base de l'actual *Institut d'Estadística de Catalunya*, fou pioner a Espanya, tant en l'accés a les bases de dades en línia científiques i tècniques del primer servidor europeu (*European Space Agency*), com en la producció i posada en servei públic de la primera base de dades estadística (ESPAN) sobre les fonts estadístiques espanyoles, i, en general, en l'aplicació de la informàtica als problemes documentals.

sitats són més diversificades i complexes. De manera que no n'hi haurà prou amb preveure simples notes acompanyant les dades. D'una banda, les peces de metainformació hauran d'incloure no solament textos llargs (legislació, p. ex.) sino també documents composats (projectes tècnics, incloent gràfiques, taules, cronogrames, etc.) i, fins i tot, bases de dades estructurades (com en el cas de les Classificacions multinivells). Tindrem així un conjunt voluminós i heterogeni de peces d'informació, l'organització de les interrelacions entre les quals no és gens simple. D'afegitó, els usuaris ja no s'acontentaran amb la simple consulta de la informació; voldran (i és raonable que ho esperin així) importar-la al seu ordinador i integrar-la al seu propi entorn de treball de la manera més simple possible.

El resultat final és el de que un institut d'estadística ha de preveure l'emmagatzemament (o elaboració ad hoc) d'una munió de peces de metainformació heterogènies i organitzar-les en un veritable sistema de metainformació estadística, provist de les funcionalitats adequades als fins a què es destina.

La natura tècnica d'aquests sistemes d'informació pot correspondre a la dels sistemes clàssics de recuperació d'informació, però també pot adoptar avantatjosament la dels sistemes hipertext.

Efectivament, els sistemes hipertext, altrament dits navegacionals per la seva capacitat de permetre l'usuari *navegar* al seu gust per entre les peces d'informació (*nodes*), atorguen una gran llibertat al dissenyador a l'hora d'interrelacionar-les mitjançant *lligams* (*links*). I aquesta és una característica especialment adequada quan es tracta d'interrelacionar la informació amb la metainformació estadística. Justament, però, per aquella llibertat, és molt important estudiar amb cura els principis a què haurien de respondre l'estructura dels lligams i els mapes conceptuals constitutius del *browser*, com s'anomena l'instrument bàsic de navegació dels usuaris[6].

En resum, la perspectiva dels usuaris de la informació estadística ens porta a organitzar sistemes de recuperació i difusió com a eines molt adequades per a satisfer les necessitats modernes d'informació sobre la informació estadística. És més, *la perspectiva de l'ús pot ser utilitzada fins i tot a l'hora d'organitzar el propi sistema d'informació estadística d'un institut.* Si així es fa, aquest és anomenat per Sundgren (1993) *sistema d'informació estadística orientat a l'usuari* (*user-oriented*) o bé *induït per l'usuari* (*user-driven*).

Pel que fa a la metainformació estadística, aquella generalització i diversitat de mitjans a disposició del públic planteja la possibilitat i l'exigència addicional, per als dissenyadors dels sistemes, de considerar les necessitats específiques de grups

[6]Vg Canals (1990), on es tracta aquest problema específic, amb caràcter general, per a qualsevol mena d'informació en els sistemes hipertext.

d'usuaris diferenciats, i en especial la del grup d'usuaris sense coneixements tècnics amplis; la qual cosa es tradueix en la incorporació de peces de metainformació estadística de tipus explicatiu i pedagògic, a més de les de caire fonamentalment tècnic.

Finalment, no cal oblidar la necessitat més general i evident dels usuaris potencials d'informació estadística: la de conèixer les fonts estadístiques disponibles sobre un tema determinat. Així, la descripció de les fonts d'un país hauria de constituir una peça bàsica en l'articulació del sistema general de metainformació estadística corresponent al seu Sistema estadístic.

1.2. La metainformació estadística des del punt de vista de la producció d'informació estadística

Si ara ens situem en el lloc dels responsables i tècnics de tota mena involucrats en la *producció* d'informació estadística que és, evidentment, la funció bàsica de tot institut d'estadística, la metainformació estadística que ens interessarà serà tota aquella informació addicional necessària per produir una informació estadística en particular, o un conjunt d'informacions estadístiques en general (a banda, naturalment, de la informació numèrica mateixa). Així, totes aquelles informacions relatives a totes i cadascuna de les fases de la cadena de producció de la informació estadística (disseny, obtenció de les dades primàries, constitució dels arxius de base, estimació, explotació[7]) constituïran peces de metainformació pertinents, independentment de la seva forma i suport del sistema d'informació en el que puguin estar integrades.

Així, per començar, tota la *documentació tècnica* relativa a les operacions estadístiques constituirà una part substancial de la metainformació estadística, com també la informació legislativa i l'administrativa associada o relacionada. Un altre conjunt típic és el de la *informació instrumental*, entre la que podem distingir la relacionada amb el tractament informàtic (documentació de sistemes, programes, codis, etc.) i la normativa (classificacions, codis territorials, etc.).

Amb aquesta simple enumeració es comprèn a la vegada la complexitat i la trascendència per a un institut d'estadística d'organitzar eficientment aquesta munió de peces d'informació heterogènia, de manera que el seu accés, consulta i eventual reutilització siguin no solament factibles sinó facilitats al màxim.

Un primer desenvolupament informàtic (*Data dictionaries*) ha vingut a oferir-se com a una solució tecnològica específica, al menys per a les necessitats d'informació instrumental comuna a un conjunt de bases de dades al si d'un mateix sistema. Que puguin constituir el fonament tecnològic d'un sistema complet de metainformació es-

[7] Als que caldria afegir la *difusió* i, fins i tot, l'eficàcia social en l'*ús* de la informació estadística.

tadística, però, està per veure, donada l'heterogeneïtat de les peces d'informació que aquest ha de tractar, la diversitat de les funcionalitats reclamades i la dispersió dels nuclis d'arxiu existents, que necessàriament s'han d'integrar en aquell sistema, i que ultrapassen clarament els objectius dels diccionaris de dades.

Des de la perspectiva del seu ús, és evident que les diferents menes de metainformació de què ara parlem estan concebudes per estar bàsicament al servei del personal relacionat amb la producció en un institut d'estadística. Per això, els sistemes que l'organitzen són anomenats per Sundgren *sistemes de metainformació estadística orientats o induïts per la producció* (*production-driven*), marcant així una diferència important amb l'orientació dels sistemes de què parlàvem a l'apartat anterior, al servei prioritari dels usuaris (externs sobretot) de la informació estadística.

Ara bé, dit això, convé afegir que les noves possibilitats de la telemàtica, entre d'altres coses, obren perspectives inèdites, que tenen com a conseqüència que aquesta metainformació estadística, induïda per les tasques de la producció, s'apropi i arribi a confondre's (al menys en part) amb la reclamada per les necessitats dels usuaris finals d'informació estadística.

En efecte, en primer lloc, si considerem, com és lògic i cada vegada més estès, que la feina d'un institut d'estadística no s'exhaureix amb l'explotació, sinó que inclou també la difusió (i sobretot la difusió activa, més enllà d'unes quantes publicacions impreses), tot el que hem dit dels sistemes *induïts pels usuaris* a l'apartat 1.1 pot ser reprès per ser integrat ara amb la perspectiva interna d'un institut d'estadística.

En segon lloc, avui dia, entre els usuaris externs, abans poc versats en els tecnicismes estadístics, hi ha cada vegada més usuaris amb els coneixements tècnics i la capacitat i els instruments necessaris no solament per interpretar correctament les dades, i eventualment reutilitzar mètodes i peces d'informació instrumentals, sinó també per explotar pel seu compte arxius de microdades (en la mesura que puguin ser difosos en forma no contradictòria amb les restriccions del secret estadístic).

En tercer lloc, les exigències de coordinació (i fins i tot de cooperació) entre instituts porten a permetre l'accés dels tècnics d'uns instituts a la informació tècnica dels altres. I això, tant al si del sistema estadístic d'un país[8], com a nivell de la coordinació europea o internacional entre instituts d'estadística[9].

---

[8]Un exemple n'és el *Sistema Estadístic de Catalunya*, estructurat entorn del *Pla Estadístic de Catalunya* (*PEC*), com a eina legislativa bàsica de planificació quadriennal, i al si del qual l'*Institut d'Estadística de Catalunya* exerceix un paper fonamental de coordinació tècnica entre els diferents organismes que hi participen, essent-ne garantia de la seva correcta execució.

[9]Un exemple n'és la coordinació a nivell de la Unió Europea, que ha de rebre en els propers anys un impuls decisiu amb la implementació del projecte *DSIS* (*Distributed Statistical Information System*) d'EUROSTAT (1992).

1.3. Perspectiva històrica

Arribats aquí, pot ser instructiu seguir l'evolució històrica que ha seguit la metainformació estadística.

Reformulant les fases definides per Nordbotten (1993), que descriuen el desenvolupament històric de mètodes de tractament de la metainformació estadística, podem definir sis fases relativament diferenciades (*Vg. Figura 1*).

En primer lloc, recordem que s'ha reconegut sempre, des dels inicis de la producció d'estadística oficial, la necessitat d'acompanyar les dades (extretes, per exemple, dels censos primitius) amb textos descrivint l'abast i els procediments i mètodes utilitzats per a la recollida de les dades. Les metadades més comunes eren, i probablement encara són avui dia, les notes al peu de les taules estadístiques.

En una segona fase, cap als anys 50, amb les primeres aplicacions de sistemes electro-mecànics de tractament de dades, certs detalls metodològics (si bé molt codificats i críptics) podien ser inclosos a col·leccions de targes perforades, les quals podien així anomenar-se arxius autoexplicatius (*self-descripting files*), encara que només amb un abús del llenguatge podríem avui dia admetre aquesta terminologia.

Només en una tercera fase, als anys 60, amb la nova generació d'ordinadors i la generalització de la cinta magnètica, és possible parlar d'arxius de dades sistemàtics (*data archives*), que podien ser objecte d'intercanvi. Aquests arxius incloïen codis i notes de metainformació, si bé tan succints, que obligaven a acompanyar-los d'una documentació explicativa a part, en paper.

Els reptes que representaven les noves possibilitats informàtiques (tot i restringides per la gairebé exclusiva consideració de camps de longitud fixa) varen fer l'objecte de grans debats al si de la comunitat estadística, encapçalats pels treballs metodològics de la Conferència d'Estadístics Europeus, a través del "*Working Group for EDP (Electronic Data Processing)*".

Ara bé, com fa notar Sundgren amb força, cal recordar que la informatització de la gestió de les operacions estadístiques va implicar en la pràctica, i des dels seus inicis, la desafortunada conseqüència de la desintegració (separació) de la relació natural que havia existit entre dades i metadades als sistemes manuals. En efecte, històricament, les preguntes, instruccions i respostes d'un qüestionari no es separaven en cap dels processos de la cadena manual; la mecanització, en canvi, que es limitava als processos de recompte (numèrics), només recollia les dades nues[10].

[10] Aquesta conseqüència, negativa, de la gairebé exclusiva dedicació al càlcul numèric de la informàtica tradicional, es retroba en molts camps (p. ex. en la informàtica documental), malgrat algunes veus aïllades, com la de Ted Nelson, pioner de l'hipertext, que als anys 60 propugnava les *literary machines* enfront dels *computers*.

abans de 1950	**Descripcions textuals.**
1950	**Autodescripcions per interpretació de targes perforades.**
1960	**Arxius de dades estadístiques (*data archives*). Només codis.**
1970	**Sistemes d'informació estadística. Concepte de metadades.**
1980	**Sistemes de metainformació estadística.**
1990	**Experiències diverses (producció *vs* ús). Consciència de complexitat i globalitat.**

Figura 1
Fases històriques de la metainformació estadística.

Només va ser en una quarta fase, als anys 70, quan s'obre pas el concepte de sistemes d'informació estadística, superant el d'arxius individuals, i la metainformació estadística comença a ser estudiada com un problema específic pels instituts d'estadística, sobretot per Sundgren (1973) en els seus treballs seminals.

En els anys 80, en una cinquena fase, comença a estar clar que la problemàtica de la metainformació estadística era més complexa del que havia semblat, i calia desenvolupar noves eines conceptuals i tècniques. D'altra banda, en una època de turbulència tecnològica, en la que coexistien grans sistemes centralitzats en temps real i xarxes locals de PCs, els esforços i les inversions es concentren en els instituts en el desenvolupament de sistemes d'informació estadística, deixant de banda la metainformació estadística. Tot i així, el debat sobre les metadades no s'atura i, fins i tot, comencen a desenvolupar-se alguns sistemes de metainformació estadística, si bé parcials.

Finalment, en una sisena fase, en la que ens trobem, en ple canvi de paradigma informàtic, sembla, d'una banda, que la tecnologia necessària per a la integració i la navegació entre informacions i sistemes heterogenis comença a estar disponible (orientació a l'objecte, hipertext/hipermèdia, interfícies gràfiques d'usuari), però les aplicacions divergeixen entre sistemes orientats a la producció o als usuaris, tant en els conceptes com en les plataformes utilitzades (Sadreddini 1993).

En contraposició a la *desintegració* entre les dades i les metadades patida fins ara com a resultat de la informatització, Sundgren subratlla que "una característica essencial de la gestió moderna de les meta dades és que està[11] *reintegrada* amb la gestió de les dades objecte".

D'altra banda, s'ha reafirmat la consciència de la manca d'un model conceptual adequat per al tractament i la integració de la metainformació estadística. I, sobretot, s'admet l'absoluta necessitat que qualsevol pas endavant haurà de ser fet en perfecta coordinació internacional. La globalització de l'economia no permet fer-ho altrament. I això implica haver resolt molts problemes d'homogeneïtzació de formats, conceptes, sistemes i xarxes, més enllà dels purament lingüístics.

En aquest sentit, si al febrer de 1993 se celebrava el 1r *Statistical MetaInformation Systems Workshop*, organitzat per EUROSTAT[12], i al 1994 s'aprovaven les grans línies del *DSIS (Distributed Statistical Information Systems)*, el 1995 ha vist el desplegament al gran públic de l'estratègia de les *"superautopistes de la informació"*, que plantegen noves i fortes exigències als vells i no resolts problemes de la metainformació, però també ofereixen enormes potencialitats.

1.4. Tipologia de la metainformació estadística segons la seva natura

1.4.1. Diversitat de la metainformació estadística

Si centrem ara la nostra atenció en les diferents menes de metainformació estadística que caldria organitzar hipotèticament en un sistema, ens trobem que una característica important del conjunt que podríem llistar és la seva diversitat, característica que constitueix un dels problemes bàsics a resoldre.

Diversitat tant en el que fa a la natura de la pròpia informació com en el que fa als formats en els que es presenta, com a les característiques dels suports que la vehiculen.

Respecte de la natura, i en una primera anàlisi, un criteri decisiu és el que separa, d'una banda, la metainformació que es refereix a *microdades* (bé a les dades mateixes, emmagatzemades en un arxiu individualitzat, bé a qualsevol de les fases de l'operació estadística que les ha creat) i, de l'altra, la que es refereix a *macrodades*, és a dir, les taules resultat de l'explotació de les microdades.

[11] No sempre, però seria desitjable que així fos, afegim nosaltres.

[12] La lectura de les actes d'aquest *Workshop* constitueix una magnífica oportunitat d'apropar-se a la problemàtica de la metainformació estadística (EUROSTAT (1993).

La metainformació sobre microdades[13] serà, més aviat, d'interès per als tècnics d'un institut d'estadística, mentre la referent a les macrodades anirá més aviat destinada als usuaris de les dades, normalment externs a l'institut d'estadística.

1.4.2. Metainformació sobre les microdades

Ara bé, dintre del primer grup, podem identificar menes d'informació molt diferents segons la seva natura, des de codis i detalls técnics molt concrets utilitzats pels informàtics per al tractament dels arxius en el transcurs d'una operació estadística, fins a la documentació tècnica relativa al disseny d'un qüestionari o d'una operació de camp (recollida), com també al detall de les incidències ocorregudes durant la seva implementació, passant per les consideracions (des dels punts de vista temàtic o bé estadístic) relatives a les variables i característiques de la població objecte de quantificació o mesura.

Dintre encara d'aquest primer grup, és a dir, la metainformació sobre les microdades, trobem les menes de metainformació que es refereixen o s'utilitzen en un marc més ampli al d'una operació estadística concreta (o d'una sèrie d'elles): ens referim, per exemple, al conjunt de programes informàtics, tant aplicatius com de base (llenguatges o bases de dades), i també a una mena de metainformació extraordinàriament important per a l'estadística: les Classificacions i Nomenclatures.

Aquestes darreres, per la seva pròpia natura, pels contextos internacional i multilingüístic de la seva creació i la complexitat del seu tractament multinivells jeràrquics, plantegen problemes específics, que justifiquen per a molts instituts d'estadística el seu tractament autònom (i, fins i tot, prioritari dins la metainformació estadística).

Altres menes de metainformació general cauen també dintre d'aquest primer grup, com poden ser la legislació i la reglamentació administrativa, quan es refereixen a una o vàries operacions estadístiques, o bé a tot el sistema estadístic o a alguna de les seves parts.

1.4.3. Metainformació sobre les macrodades

En el que fa al segon grup, és a dir, la metainformació sobre les macrodades, ha estat objecte d'anàlisi exhaustiva per Sundgren (1991), qui ha elaborat una pro-

[13] Vegeu Sundgren (1994) (reproduït en aquest mateix número de **Qüestiió**) i altres papers del mateix autor, per una classificació sistemàtica i exhaustiva de la metainformació referent a les micro-dades. Aquí només pretenem il·lustrar la problemàtica amb alguns exemples.

posta ambiciosa i maximalista, basada en la consideració de les macrodades a la llum del procés d'agregació de microdades, del que són resultat. La pregunta que tracta de respondre Sundgren és la següent: Quines descripcions d'unes macrodades determinades cal que acompanyin aquestes per tal que els usuaris potencials d'una taula (aïllada o formant part d'una base de dades) estiguin en situació de jutjar la seva utilitat i fiabilitat en un context d'ús determinat?

La resposta és, aparentment, molt simple. Cal donar informació, tant sobre la població objecte d'interès i les seves característiques (variables) com sobre els mètodes i processos que poden condicionar la fiabilitat dels resultats finals; és a dir, els processos de recollida de les microdades i els de la seva agregació per obtenir les macrodades.

D'aquí es dedueix (i Sundgren ho recorda expressament) que la metainformació a donar sobre unes macrodades ha d'incloure també la informació pertinent sobre les microdades de les que provenen i sobre el procés d'agregació que les relliga a ambdues.

En la seva proposta, Sundgren especifica que cal proporcionar, per a unes macrodades determinades, les descripcions següents:

a) Els paràmetres dels quals, per a cada població i domini d'interès, les macrodades es suposa que són estimacions.

b) Les variables, els objectes unitaris de les quals han estat investigats (observats) al nivell micro de cara a l'estimació dels paràmetres del nivell macro.

c) Complementàriament, la funció "ideal" d'estimació, per deduir els paràmetres a partir dels valors observats de les variables.

d) La manera concreta com han estat portades a terme les observacions, i totes les incidències susceptibles d'haver afectat la qualitat dels resultats.

e) La comparabilitat en el temps i en l'espai, en general i, específicament, pel que té a veure amb les classificacions i nomenclatures (nacionals i internacionals).

Des d'una altra perspectiva, i en un paper posterior, Sundgren (1993) afegeix consideracions particulars per a dues menes de macrodades diferents:

- les taules estadístiques publicades.
- els conjunts de macrodades emmagatzemades en suports electrònics o digitals (disquets, CD-ROMs, bases de dades interrogables en línia).

En aquest sentit, si bé la informació a subministrar és en essència la mateixa, el diferent nivell d'agrupament i forma de presentació exigeix/permet la preparació de peces de metainformació estadística molt diferents, per atendre/aprofitar les molt

diverses necessitats/potencialitats dels sistemes en qüestió (multiplicades pels nivells diferents d'exigència dels diferents usuaris: des del públic a l'especialista temàtic o estadístic)[14].

1.5. Heterogeneïtat de les peces de metainformació estadística

A l'extrema heterogeneïtat de les peces de metainformació estadística segons la seva *natura*, evidenciada en la descripció tipològica anterior, cal que hi afegim ara les heterogeneïtats derivades del *format* en què es presenten les peces de metainformació estadística, del suport en que es vehiculen, del context dels *sistemes* en què s'integren i de l'*objectiu* a què es destinen.

El problema del *format* n'és un de temible donats, d'una banda, la diversitat de sistemes informàtics existents, amb exigències/potencialitats diferents respecte del format de descripció de dades requerit/possible, i de l'altra les necessitats de coordinació internacional en un moment, com l'actual, en que convergeixen les potencialitats tecnològiques (xarxes telemàtiques, p. ex.) amb les necessitats polítiques (el fet de la Unió Europea, i la mundialització de l'economia, en general). Una part important de les preocupacions del *grup de treball METIS*[15] és justament la d'integrar les exigències de models (*IRDS = Information Resource Dictionary System o IRM = Information Resources Management*, d'ISO, etc.) i formats (EDIFACT, ODA, SGML, etc.) internacionals a les necessitats específiques de la metainformació estadística.

Un aspecte no banal del format és el referent a les alternatives de presentació de la informació digitalitzada: en imatge (mapa de bits), en ASCII, en format de processador de textos (Word, WordPerfect), general de consulta (PDF, d'Acrobat, p. ex.), en document composat d'alguna mena (OpenDoc, OLE, p. ex.), o bé en forma d'enregistrament estructurat d'una base de dades (o DBF, en general).

Heterogeneïtat també respecte del *suport*, una vegada més amb la doble visió de les exigències/potencialitats ofertes per cadascun d'ells, i dintre de la tendència a la generalització de l'edició electrònica en totes les seves formes.

Heterogeneïtat, igualment, derivada del context del/dels *sistema/es* en el/els que s'integren les peces de metainformació estadística, de manera especial en la seva in-

[14]L'expressió dual "necessitat/potencialitat" és emprada per tal de ressaltar que es tracta de dues cares de la mateixa moneda, segons que es miri l'aspecte restrictiu d'un sistema (les seves limitacions) o l'aspecte positiu (les seves potencialitats). En la mesura que els sistemes informàtics evolucionen històricament en el sentit de disminuir les limitacions i augmentar les potencialitats, és possible plantejar-se objectius (d'integració i de coordinació) més forts cada vegada, utilitzant sistemes més moderns i avançats. Tenint en compte que queda sempre el problema de la transició dels vells sistemes als nous.

[15]Ja esmentat a la nota 1). *Vegeu també l'apartat 2.4*)

terrelació. Així, per posar un exemple del que volem dir, una peça de metainformació estadística pot consistir en un text presentat en una finestra en la que diversos descriptors o "botons" (*hot spots*) són lligams (*links*) associats a d'altres peces d'informació, al si d'un sistema hipertext, de manera que l'usuari, com és típic d'aquests sistemes, pot navegar lliurement entre aquelles peces d'informació interrelacionades.

Aquesta situació és molt diferent de la que es produeix en sistemes de bases de dades relacionals, en les que les peces d'informació han estat, *ab initio*, formalment estructurades segons la lògica conceptual de la informació, traduïda en un determinat esquema d'entitats-relacions.

Heterogeneïtat, finalment, derivada de l'*objectiu* a què es destinen les peces de metainformació estadística. Un parell d' exemples il·lustraran a què ens referim:

a) En primer lloc, suposem que es tracti d'una *nota al peu d'una taula* o associada amb ella. Doncs bé, serà molt diferent si va adreçada a un informàtic (en el qual cas serà breu i molt codificada), a un temàtic (text tècnic rigorós i complet des del punt de vista del significat de les xifres), o al públic no tècnic (text explicatiu més llarg, i fins i tot pedagògic, però menys rigorós i sense detalls tècnics). De fet, una nota associada a una taula pot fer referència a les seves condicions de producció, a les de la seva reutilització en un context diferent, o simplement, a les del seu "bon ús" (per tal d'evitar males interpretacions del seu significat).

b) En segon lloc, suposem que es tracti d'una *tipologia de tabulació*, és a dir, el conjunt d'ítems que classifiquen una variable en una taula. En aquest cas, hem de recordar que tipologies molt diverses poden haver estat originades per la mateixa Classificació de base, segons ordres d'agregació diferents. Per exemple, l'activitat econòmica dels establiments pot classificar-se segons una tipologia a 4 sectors, o 24 branques, que són agregacions diverses dels ítems de la *Classificació Nacional d'Activitats Econòmiques*. Doncs bé, un problema típic amb el que es troba l'usuari d'informació estadística tabulada és el d'identificar l'abast dels ítems a què corresponen els rètols de files i columnes de la taula (molt abreujats normalment per raó de l'espai disponible) cosa que requereix notes explicatives a part. Però l'usuari pot voler a més reutilitzar les etiquetes normals de la tipologia en el seu context de treball, i això requerirà que el sistema de metainformació estadística ho prevegi. En un institut d'estadística és essencial que un únic sistema de Classificacions sigui la font d'autoritat de totes les tipologies que s'utilitzin, que aquestes li estiguin integrades, i així mateix, que siguin automàticament accessibles des de qualsevol taula que les utilitzi en algun dels múltiples sistemes difosos per l'institut (publicació electrònica, base de dades en línia). D'altra banda, l'usuari haurà de poder, a més, triar la llengua en què li interessi la tipologia en qüestió.

1.6. Escenaris d'usos i usuaris

La consideració de l'objectiu en vistes del qual els usuaris necessitaran accedir a peces de metainformació estadística ens porta a plantejar el tema dels diferents tipus d'usuaris d'un sistema de metainformació estadística.

Des d'un punt de vista general, poden ser usuaris de metainformació estadística tant els que utilitzen informació estadística per ella mateixa com els que la produeixen, retrobant així el doble punt de vista amb el que podem enfocar la metainformació estadística, i al que al·ludíem a l'inici d'aquest article. Així, usuaris i productors d'informació estadística són les dues categories bàsiques d'usuaris de metainformació estadística. Tal com ja hem apuntat, "productors" són els diferents tècnics i especialistes (estadístics, temàtics, informàtics) que treballen al si d'un institut d'estadística en tasques relacionades amb la producció d'informació estadística, i "usuaris" seran les persones i institucions que la utilitzen (tant microdades com macrodades), normalment externs a l'institut. Uns i altres necessitaran accedir a certes peces de metainformació estadística relacionades amb la informació estadística amb la que treballen o que volen utilitzar. I ja hem fet notar que la metainformació estadística necessària vindrà determinada pel tipus específic d'objectiu o tasca en el context del qual l'usuari de metainformació estadística la necessita. És tot el problema dels *usos*, que s'afegeix a la categorització dels *usuaris*, i que fa que la identificació de les peces de metainformació estadística necessàries per a cadascun d'ells sigui a la vegada més complexa però també més pragmàtica, en la mesura en què, si es posa en relació amb el sistema de metainformació estadística l'eficàcia d'aquest, en la pràctica la seva utilitat serà més gran.

En aquest punt, voldríem fer tres observacions:

a) La *primera* té a veure amb la funció de difusió de la informació dels instituts d'estadística, cada vegada més important, a la que ja hem al·ludit a l'apartat 1.2), i que exerceixen diferents menes de tècnics (als estadístics, temàtics i informàtics s'afegeixen els especialistes en sistemes d'informació i documentació). Els *difusors* d'informació estadística (a no confondre amb els tècnics de la unitat orgànica de Difusió, quan aquesta existeix) han passat de les seves tasques tradicionals d'editar les publicacions estadístiques i prestar serveis de biblioteca i documentació al disseny i implementació de tota una panòplia de sistemes d'informació, des de bases de dades bibliogràfiques o factuals interrogables en línia a diferents formes d'edició electrònica). Així, doncs, en la mesura en què conjunts de peces de metainformació estadística han de ser elements constitutius importants d'aquests sistemes, en relació amb la informació estadística que difonguin, els *difusors* dels instituts d'estadística es constituiran en una nova categoria d'usuaris de metainformació estadística.

b) La *segona* està relacionada amb l'anterior. En efecte, en la mesura en què els sistemes de difusió d'informació estadística incorporen peces de metainformació estadística, els tècnics que participen en la seva implementació podran integrar-hi les peces de metainformació estadística que ja existeixin i estiguin disponibles; d'altra banda, però, serà necessari que en produeixin moltes altres, bé perquè les existents no s'adaptin al context dels usuaris dels sistemes, bé perquè es tracti de peces de metainformació estadística no necessàries per a la producció, bé perquè estiguin adreçades a categories específiques d'usuaris o a necessitats molt concretes.

c) La *tercera* consisteix simplement a esmentar unes situacions concretes en les que es poden trobar diferents tècnics d'un institut d'estadística i que constitueixen il·lustracions de la varietat d'usos de la metainformació estadística en diferents contextos. Els exemples han estat suggerits per les funcions de persones reals de l'*Institut d'Estadística de Catalunya,* si bé hem substituït els seus noms per lletres, i estereotipat en certa manera les seves funcions:

> El tècnic ***A***, en la seva feina de satisfer demandes del públic de taules estadístiques específiques per ser transmeses en disquet, necessita recollir la metainformació estadística necessària, en forma de definicions i notes (que poden trobar-se a la base de dades BEMCAT, però també en d'altres llocs i en formats diversos).
>
> El tècnic ***B***, en la seva feina de constituir arxius de microdades (prèviament preparats per no infringir el secret estadístic) per a la seva difusió, necessita integrar també al suport electrònic escollit (cinta o CD-ROM) la metainformació estadística necessària: definicions de variables, segur, però també diferents aspectes de la producció, de cara a proporcionar elements per a jutjar de la fiabilitat i comparabilitat de les dades en diferents contextos, i també afegirà descripcions del mètode utilitzat per obtenir l'arxiu susceptible de ser difós com a subconjunt de l'arxiu individualitzat de base. L'arxiu de Documentació Tècnica (definit a la *Norma 0*, però no implementat) li podria proporcionar la informació necessària, però l'única manera de que no hagi de recopiar-la (o, pitjor, reredactar-la) seria que estés prevista aquesta necessitat i es redactessin i s'emmagatzemessin les peces corresponents en el moment de constituir l'arxiu (digital) de Documentació Tècnica de l'operació estadística en qüestió.
>
> ***C***, especialista temàtic/a en la preparació dels Padrons Municipals d'Habitants 1996, necessita consultar informació metodològica general (internacional i pròpia) relativa als Censos de Població anteriors.
>
> ***D***, informàtic/a en la seva feina de produir, a partir dels arxius de base, noves taules per a formar part de publicacions o per a integrar-les a la base

de dades BEMCAT, necessita, d'una banda, codis i paràmetres relatius als arxius per a l'ús correcte dels programes informàtics i, de l'altra, els literals per a etiquetes de files i columnes i notes textuals (definicions, comentaris...). Es barregen, doncs, necessitats internes molt tècniques i les externes, (textos per als usuaris), en el context de la base de dades.

E, estadístic/a en el transcurs del disseny operatiu d'una nova operació estadística, necessita incorporar-hi una tipologia pròpia, obtinguda (s'entén automàticament) per agregacions de classes i/o ítems d'una classificació oficial.

F, estadístic/a en el transcurs del disseny operatiu d'una nova operació estadística, necessita informació diversa (metodològica, definicional, tipològica i estadística) per tal de comparar l'eficiència de mostres diferents del mateix univers utilitzades en diferents operacions del passat.

G, per a preparar els originals de les publicacions estadístiques en el sistema Macintosh d'edició electrònica (*desk top publishing*) utilitzat, necessita manipular els paquets d'informació estadística i textual associats que rep, i en els formats adequats per minimitzar la feina.

H, en la seva feina de construir un CD-ROM per difondre informació censal, té moltes necessitats de metainformació estadística com, per exemple:

- recollir tota mena de definicions, notes d'ús, qüestionaris, etc.
- integrar etiquetes de tipologies de variables als rètols de files i columnes

Això implica, en general, que les necessitats derivades dels sistemes, directes o indirectes, de difusió (BEMCAT, CD-ROMs, publicacions, videotex, etc.) han d'estar previstes ja a les fases de la producció i articulació de sistemes de producció, so pena d'haver de reescriue i adaptar les mateixes informacions en diferents contextos.

Finalment, ***I***, un administrador de l'institut, necessita recuperar informació de calendaris, personal, costos, etc. d'una operació anterior per a comparar-los amb una altra anàloga en curs.

Etcètera.

Des de la perspectiva diferent d'identificar els usos a què un sistema de metainformació estadística hauria d'atendre de cara al seu disseny, Sundgren (1992) defineix 5 escenaris d'ús:

- Escenari 1: Perspectiva orientada a l'usuari final.
- Escenari 2: Perspectiva orientada a la producció.

- Escenari 3: Perspectiva orientada al disseny (per fases).
- Escenari 4: Perspectiva gerencial.
- Escenari 5: Perspectiva tècnica.

Per a cadascun d'aquests escenaris, Sundgren, analitza les tasques i subtasques relacionades, les necessitats de cada fase o les dels subsistemes implicats, aportant una sèrie de consideracions molt suggestives.

1.7. Cap a sistemes integrats de metainformació estadística

Arribats a aquest punt, creiem necessari explicitar un parell d'aspectes importants, que poden haver passat desapercebuts i que, en tot cas, no hem justificat a bastament.

1.7.1. La generació de les peces de metainformació estadística

El primer d'aquests aspectes es refereix a la responsabilitat de fabricació de les peces de metainformació estadística al si d'un institut d'estadística.

En efecte, hem parlat en algun moment donant per suposat que les peces de metainformació estadística existien prèviament, o es fabricaven com a part dels processos normals de producció d'informació estadística. La realitat és, però, que en gran part caldrà fabricar-les expressament, atenent als usos a què vagin destinades i (molt important) amb el format i les característiques formals requerides pel sistema o subsistema en el què hagin de ser incorporades.

Ara bé, un problema especialment delicat és el de a quina unitat orgànica o funcional atribuir aquesta responsabilitat, al si d'un institut d'estadística. D'una banda, són els tècnics de producció els que tenen la informació i els coneixements necessaris per produir les peces de metainformació estadística tècniques i d'ús relacionades amb la creació i manipulació de la informació estadística en general, sobretot les microdades, però també, en part, les macrodades. El problema és que això desborda el seu objectiu funcional i el seu interès i, en la pràctica, la documentació que estan ben disposats a generar és solament la que necessiten per a l'eficàcia de la seva feina. D'altra banda, els tècnics relacionats amb els mitjans i sistemes de difusió (en sentit ampli) són els que són sensibles per la seva funció a les necessitats dels usuaris i, per tant, a ells pertoca fer els esforços que calgui per crear les diverses i nombroses peces de metainformació estadística necessàries. Per fer-ho, però, els caldrà disposar de la informació, bruta per dir-ho d'alguna manera, a partir de la qual,

mitjançant manipulacions i tractaments diversos puguin crear-les; i això de la manera més automàtica que sigui possible.

A més a més, hi ha conjunts de peces de metainformació estadística que, essent necessària la seva disponibilitat o organització en sistemes de recuperació d'informació de cara als usuaris externs, no provenen de les tasques de producció. Per exemple, la legislació i la documentació tècnica metodològica i internacional.

Altres, encara, tot i estar relacionades amb la feina de producció, són de caire no tècnic com, per exemple, les que informen sobre les fonts (impreses o digitals) a través de les quals es difonen els resultats de les operacions estadístiques.

Altres peces de metainformació estadística, finalment, tot i ser eines de producció, tenen un caràcter d'ús general i polivalent, que obliga a articular-les en sistemes autònoms i complexos. Aquest és el cas de les Classificacions, Nomenclatures i Codis territorials.

Tots aquests conjunts addicionals de metainformació estadística poden ser també responsabilitat funcional dels tècnics de difusió (que, com hem dit, van més enllà de la unitat de Difusió, quan existeix).

1.7.2. L'organització d'un sistema de metainformació estadística

El segon aspecte es refereix a l'organització dels conjunts de peces de metainformació estadística en sistemes d'informació.

Fins ara, hem parlat de metainformació estadística o de sistemes de metainformació estadística indistintament o, en tot cas, sense justificar la diferència.

Ara bé, en introduir el concepte de "sistema" i parlar de "sistema de metainformació estadística", entenem que els diferents conjunts i les peces de metainformació estadística mateixes s'articulen en una unitat d'ordre superior, en la que les interrelacions entre elles i amb els outputs i prestacions oferts als usuaris prenen un nou significat a la llum de la nova globalitat. Segons la proposta METIS, elaborada per Prazenka (1994):

> *A Statistical Metainformation System (SMS) is the information system that uses and stores statistical metadata and produces statistical metainformation for purpose of supporting decision making concerning an information system. The object of the Statistical Metainformation System is the statistical information system.*

De fet, l'organització de la metainformació estadística en un o diversos sistemes, i la natura d'aquests, com també el seu entorn informàtic, és un tema subjecte a

debat. Si l'especificitat d'alguns conjunts de metainformació estadística (bé sigui per la seva natura, bé per la mena d'usuaris o usos a qui vagin destinats) reclama un tractament autònom (com és el cas de les Classificacions, com exemple paradigmàtic), les interrelacions entre els diferents sistemes resultants reclamen una coordinació global forta (de cara a minimitzar duplicitats i a maximitzar-ne l'ús eficient per part dels usuaris).

Avui dia, amb el rerafons de la revolució tecnològica en curs derivada de les transformacions de les tecnologies de la informació i la comunicació, la *integració* (paraula-clau del moment) de sistemes a múltiples nivells d'informació és a la vegada un objectiu ambiciós, però a l'abast.

Dit això, ¿com compaginar el llast que la història ha deixat en els instituts d'estadística en forma de configuracions informàtiques determinades (si no obsoletes, sí corresponents a concepcions diverses) amb la necessitat de caminar cap a sistemes veritablement integrats entre les diferents menes de metainformació estadística, les necessitats dels usuaris, i entre la metainformació estadística i la informació estadística associada?

La resposta no existeix encara, almenys suficientment contrastada i assumida. Però el que sí podem fer és passar revista a les diferents formes com, en la pràctica, els instituts d'estadística d'altres països aborden el problema, i també quins són els seus plantejaments i els dels organismes internacionals sobre la metainformació estadística. Aquest és el significat de la segona part d'aquest article.

2. LA DIVERSITAT DE L'EXPERIÈNCIA INTERNACIONAL

2.1. Introducció al panorama internacional

El mètode que utilitzarem per passar revista a l'experiència internacional sobre la metainformació estadística no és el de fer-ho tractant d'esbrinar com s'ho ha plantejat cada institut d'estadística, sinó el de veure, per a cadascuna de les possibles formes que pot adoptar el sistema global de metainformació estadística o cadascun dels components o mitjans autònoms de metainformació estadística, quins són els problemes específics que es plantegen i quines han estat les solucions que, en la pràctica, han estat proposades, mitjançant projectes o experiències.

D'aquesta manera, el que pretenem és, prenent les experiències esmentades com a simples il·lustracions de solucions o plantejaments, donar una visió més detallada de

la problemàtica implicada en els sistemes de metainformació estadística, en la mesura que els instituts d'estadística se la plantegen.

No cal buscar, doncs, el plantejament específic de cada institut d'estadística; només són esmentats aquells que aporten algun projecte o experiència en un tema concret, que hem conegut i ens ha semblat il·lustratiu.

D'altra banda, tampoc ens ha preocupat massa si els exemples aportats podien estar desfassats (de fet, gran part de la informació prové del *Statistical Metainformation Systems Workshop* de 1993, i també de les reunions del grup de treball METIS), entenent que el que ens interessa aquí és més exemplificar tipus de solucions que donar notícia de solucions "al dia".

Així mateix, reconeixem un cert desequilibri material en el tractament dels diferents punts, inevitablement supeditat a la informació disponible.

En definitiva, passarem revista als punt següents:

- Sistemes generals de metainformació estadística, orientats a la producció.
- Paquets de metainformació estadística especialitzats.
- Sistemes generals de metainformació estadística, orientats als usuaris.
- Sistemes documentals i metainformació estadística.
- La metainformació estadística com informació associada a conjunts d'informació estadística, en suports digitals de difusió.
- El sistema de metainformació estadística com instrument per a la coordinació.
- La metainformació estadística integrada en un sistema expert.

2.2. Projectes i experiències en curs

2.2.1. Sistemes generals de metainformació estadística, orientats a la producció

La primera opció a considerar consisteix a aplicar a la metainformació estadística la mateixa tècnica de construir bases de dades relacionals sobre grans ordinadors (*mainframes*), utilitzada generalment als instituts d'estadística per mantenir centralment algunes o totes les tasques de la producció estadística. Es tracta, doncs, de dissenyar sistemes generals orientats a la producció, en els que la metainformació estadística considerada és la d'interès per als tècnics involucrats en la producció, l'objectiu dels quals, implícitament o expressa, és la d'arribar a constituir sistemes

complets i *actius* (anomenant així els sistemes que pretenen no solament referenciar sinó també integrar la metainformació estadística amb la informació estadística associada al si del mateix sistema).

En la pràctica, les dificultats tècniques derivades de la mateixa ambició de l'objectiu, i també les resistències humanes a les fortes exigències de coordinació implicades (a l'haver de documentar amb més amplitud de la que cadascú requeriria per a les seves pròpies necessitats), fan que les aplicacions de metainformació estadística acabin essent en realitat parcials, d'una banda, i aïllades del sistema d'informació estadística, de l'altra.

2.2.1.1. L'òptica "Dictionnaires de Données Statistiques (DDS)" de l'INSEE

L'enfocament "diccionari de dades" (*data dictionary*), en el context dels sistemes informàtics en general, fou impulsat per Shoshani, i va donar lloc a la incorporació de programari especialitzat als entorns de bases de dades, com és el cas del *CDD* (*Common Data Dictionary*) de Digital, sistema orientat a l'objecte de repositori de dades que centralitza les metadades corresponents a les dades dels diferents sistemes corrent en l'entorn de referència (VMS, p. ex.). És utilitzat a l'*Institut d'Estadística de Catalunya.*

L'INSEE (1990) va generalitzar el concepte per donar-l'hi un significat general dintre del Sistema Estadístic de França. La seva experiència, descrita per Lazarou (1993) és paradigmàtica, havent realitzat ja al 1983 un primer prototipus del sistema *DDS* (*Dictionnaires de Données Statistiques*). El seu objectiu final era molt ambiciós, puix que es tractava d'arribar a constituir un conjunt de sistemes articulats, que oferissin globalment l'accés al públic al conjunt del Sistema Estadístic de França, tot i que l'INSEE no tenia (ni té encara) competències sobre ell.

En la pràctica, la implementació s'ha reduït al mòdul "Production" (només de l'INSEE), composat dels submòduls "DDS-Enquêtes" i "DDS-Nomenclatures", que en la seva versió actual es configuren sota l'arquitectura client-servidor com a bases de dades relacionals sota ORACLE sobre un servidor UNIX, i accessibles des de clients PC/Windows en xarxa local, existint interfícies amb SAS. Les entitats (en un esquema d'entitats-relacions) considerades per a les Enquestes són: *Questionnaire, Question, Spécification, Programme, Variable, Nomenclature, Poste, Fichier, Tableau, Incident, Concept, Archive, Journal.*

L'institut d'estadística de Rumania està implementant aquesta versió del *DDS* per al control de la producció d'enquestes, com un component d'un sistema més general, en desenvolupament (Vg. Marina 1994).

D'altra banda, l'òptica "diccionaris de dades" fou el model conceptual de referència sobre el que es basava la nostra proposta d'estratègia *DIDAC* (*Diccionaris de Dades de Catalunya*) per a l'*Institut d'Estadística de Catalunya*, presentada al 1990 en un informe intern.

Altres instituts d'estadística han pres opcions semblants, menys ambicioses, com a simple ampliació de la seva experiència en bases de dades relacionals aplicades a la metainformació estadística, com, per exemple, l'*Instituto Nacional de Estadística* espanyol.

En efecte, l'INE (1994) (*Vegeu també Villar 1993*) que havia implementat al 1986 una petita versió d'un sistema de diccionaris de dades, ha desenvolupat al 1994 el *S.I.D. (Sistema de Información Documental)*, com l'estructuració informàtica mitjançant una base de dades relacional del conjunt d'informacions i documents tècnics relacionats amb cadascuna de les etapes de les operacions estadístiques a través del "Centro de Proceso de Datos", i d'acord amb la "Norma de Desarrollo de Aplicaciones", que paral·lelament s'ha implementat. Malgrat que es tracta, doncs, d'un sistema marcat pel seu entorn informàtic, en la mesura que en el seu disseny (via la *Norma* esmentada) s'han tingut en compte les informacions necessàries al conjunt de tècnics de l'INE, té un significat volgudament ampli, reduït tanmateix per l'orientació de producció inherent al plantejament.

2.2.1.2. L'enfocament "information management"

El plantejament "information management" que fa Ronald Graves (1994), en el que es basa l'actual estratègia de "Statistics Canada", respon a una consideració conceptualment simple i que té l'avantatge de permetre integrar els sistemes relacionals existents amb els nous tractaments documentals: la de considerar que en l'estadística, tenim dades i metadades. Mentre les dades (xifres, taules, arxius numèrics) són abundants i han estat tractades amb detall, les metadades (textos i documents de tota mena) estan disperses i han estat descurades. Tenim, doncs, dos paquets d'informació, que han de ser tractats amb mètodes diferents, constituint subsistemes d'un sistema superior que els integra i interrelaciona. Graves, d'una banda, estableix un esquema conceptual del sistema global, a través d'una definició operativa d'*informació* (que entén englobant *dades* i *documents*) i, de l'altra, preveu, al costat de les bases de dades relacionals que tracten les dades, la utilització de sistemes capaços de tractar informació heterogènia (referències, textos, documents composats, etc.), que no són únics (sistemes documentals tradicionals, sistemes de gestió d'informació, com BASIS/Plus, sistemes hipertext, etc).

L'objectiu és el d'anar integrant sota l'esquema global diversos mòduls (com *IBOSS = Internal Bank of Statistics System i EBOSS = External Bank of Statistics Canada*) i també productes (el catàleg de fonts, CD-ROMs específics, i d'altres).

Aquest plantejament apareix com a resultat de l'anàlisi de l'experiència de "Statistics Canada" en l'elaboració de diversos productes relacionats amb la metainformació estadística, com el CD-ROM *CANSIM* (semestral): conté bàsicament 500.000 sèries temporals (actualitzades semestralment) i està provist d'una interfície de consulta i explicació en hipertext, articulada amb el *SDDS (Statistical Data Directory System)*, veritable sistema de metainformació estadística estructurat sota BASIS. D'altres CD-ROMs publicats fins ara són el LMAS (mercat de treball) i els censals.

Afegim que hi ha relacions evidents entre aquest plantejament i alguns dels que esmentarem a l'apartat 2.2.4) i al 2.2.6).

2.2.2. Paquets de metainformació especialitzats

Aquesta opció suposa la renúncia (expressa o, més sovint, implícita) a plantejar-se un sistema general de metainformació estadística, per a concentrar-se en el plantejament de solucions per a conjunts particulars de metainformació estadística.

És una opció molt estesa entre els instituts d'estadística, si bé amb característiques especials segons l'experiència de cadascun, i també segons la sensibilitat relativa als problemes de la difusió al públic respecte dels de la producció.

En general, són les Classificacions i Nomenclatures els tipus de metainformació estadística tractats prioritàriament, fins i tot en absència de la seva consideració explícita com a metainformació, simplement per la seva importància intrínseca, tant per a la producció com per a la difusió d'informació estadística.

Així s'està fent, per exemple, a l'*Institut d'Estadística de Catalunya*, que té en curs d'elaboració una base de dades (*SICONO = Sistema d'informació de Codis i Nomenclatures*), interrogable en línia sobre l'ordinador central VAX.

EUROSTAT, per la seva banda, que edita les Classificacions en disquets, entre d'altres suports, prepara el projecte EDINOMEN, independent però integrable amb DSIS, l'objectiu del qual és el d'establir un sistema d'intercanvi electrònic de Classificacions i Nomenclatures entre els instituts d'estadística (Lebaube 1993).

Esmentem, així mateix, l'existència de programari comercial, com p. ex. BLAISE, desenvolupat per al disseny, recollida i control de la informació d'enquestes assistits per ordinador. En la seva versió BLAISE III, el tractament de les dades i les metadades corresponents està integrat (Schuerhoff 1993). El *Centre for Educational Sociology* n'ha fet una aplicació (Lamb 1993), en el context del projecte EISI, del programa DOSES (EUROSTAT) (enfocat a sistemes experts en estadística).

2.2.3. Sistemes generals de metainformació estadística, orientats als usuaris

Aquesta opció és la resposta lògica quan la preocupació prioritària (no única) és la de proporcionar als usuaris l'accés a la metainformació estadística i les prestacions que responen a les necessitats, generals i les particulars de grups d'usuaris específics; essent l'objectiu del sistema el de permetre consultes intel·ligents, una recuperació d'informació eficient i un suport efectiu a l'ús de la informació estadística en els diversos contextos de treball dels usuaris.

En la pràctica, un sistema amb aquestes funcionalitats pot adoptar diverses formes, i consistir, bé simplement en una interfície d'usuari per accedir a bases de dades centrals o distribuïdes, bé en una veritable arquitectura client-servidor, bé en un sistema autosuficient, bé en alguna de les variants intermitges. Veiem-ne alguns exemples.

2.2.3.1. Interfície d'usuari integradora

Existeixen múltiples variants d'interfícies, més o menys mediatitzades per les característiques de les bases de dades a les que han de donar accés. En la mesura, però, en què es dona la importància que veritablement té al terme "interfície d'usuari", com a sistema que "tradueix" els conceptes en què els usuaris tenen plantejades les seves necessitats i usos potencials a l'estructura (conceptual, organitzativa, formal i material) amb què les peces de metainformació estadística estan emmagatzemades a les diferents bases d'informació disponibles, es van decantant elements comuns de solució, encara en estat embrionari.

1) **Tesaures**

   Si es tracta de "traduir" conceptes, una eina tradicional és un *tesaure* estructurat, que organitza els termes constitutius d'un lèxic, juntament amb les seves interrelacions, incloent tota mena de sinònims i termes utilitzats en diferents contextos. Kopp (1993) descriu el Tesaure utilitzat per l'institut estadístic de Berlín i explica el seu rol integrador com a peça bàsica del sistema cooperatiu (interciutats) de metainformació estadística desenvolupat a Alemanya, sobre la base dels conceptes teòrics elaborats per Appel (1993). La raó d'utilitzar el tesaure com a la interfície d'usuari bàsica "*rau en el reconeixement de que només la integració del llenguatge corrent en un sistema de metainformació estadística pot reduir la complexitat, que seria massa gran en qualsevol altra circumstància d'assolir un projecte tan ambiciós*" (Appel 1994).

2) **Sistemes hipertext**

   Una altra manera (no contradictòria amb l'anterior) és la d'enfocar la interfície com un *browser* d'un sistema hipertext; en aquest cas, el conjunt de mapes conceptuals o "cartes de navegar" constitutives del browser (juntament amb les funcionalitats

de navegació pertinents) poden limitar-se a donar accés a les peces de metainformació estadística emmagatzemades a les bases d'informació tradicionals, o bé (al menys en part) contenir les pròpies peces de metainformació estadística. Aquesta darrera opció és particularment interessant, quan es tracta d'informació no estructurada o sota formes gràfiques o emmagatzemades en mapes de bits (imatge).

De fet, el programari EMMA (*Vegeu més avall*) presenta un mapa conceptual en hipertext com una de les formes de consulta i ja hem citat el CD-ROM CANSIM, provist d'una interfície en hipertext. En realitat, si bé no hi ha encara massa exemples operatius, és una opinió generalitzada que els sistemes hipertext jugaran un paper important en les interfícies dels sistemes d'informació i metainformació estadística.

A l'*Institut d'Estadística de Catalunya* es va elaborar l'any 1989 *munCAT*, "Un hiperdocument d'estadística demogràfica de les comarques i municipis de Catalunya". Desenvolupat sota HyperCard per a Macintosh, es tractava d'un prototipus que implementava experimentalment una sèrie de conceptes originals constituint una veritable proposta d'interfície especialitzada: d'una banda, relligant una sèrie de taules dels Padrons Municipals d'Habitants 1986 amb conjunts molt diversos de metainformació estadística (notes, legislació, gràfiques, mapes automàtics de coropletes, bibliografia, etc), "navegable" mitjançant mapes conceptuals a diversos nivells; de l'altra, oferint una sèrie de funcionalitats, estructurades en un conjunt sistemàtic i coherent (*munCAT*, juntament amb d'altres realitzacions, està descrit per Canals 1992). Alguns d'aquests conceptes han estat aplicats a la interfície de consulta inclosa al CD-ROM *Cens de Població de Catalunya 1991*, en curs d'elaboració (desenvolupat sota FoxPro per a PC/Windows) (*Vegeu l'apartat 2.2.5.2*).

2.2.3.2. El sistema PC-DOK de "Statistics Sweden", orientat al document

Tal com explica Malmborg (1993), el plantejament del sistema PC-DOK respon, a la vegada, als plantejaments conceptuals i teòrics generats al si del mateix institut suec d'estadística, on treballa Sundgren, i a l'experiència dels sistemes ja existents:

- DOK, un sistema documental sobre mainframe, contenint la documentació tècnica sobre els sistemes de producció, relacionats amb el tractament informàtic.
- ARK-DOC, utilitzant el mateix sistema, però contenint descripcions d'arxius bruts d'enquestes (típicament en cinta magnètica).
- DAISY, sistema de metainformació estadística per difondre informació sobre micro i macrodades sobre el mercat de treball. Es tracta d'un sistema sobre PC i l'usuari selecciona i s'emporta jocs de disquets amb programari i una (meta)- base de dades.

PC-DOK ha de substituir DOK i ARK-DOC tot suplementant i ampliant DAISY.

Des del punt de vista conceptual, PC-DOK reposa sobre eines ja experimentades, com SCBDOK, normativa de descripció de la documentació tècnica d'operacions estadístiques de producció[16]. Des del punt de vista de disseny, PC-DOK, es basa en l'orientació a l'objecte, accepta documents composats, i conté una interfície gràfica d'usuari sota Windows, amb connexions via SQL amb les bases de dades centrals; un model gràfic de presentació de la informació estadística (Malmborg 1988) és part important de cara a la seva facilitat d'ús, i adopta una forma semblant a la d'un tesaure gràfic, interpretable també com un mapa conceptual navegable (hipertext).

Si bé el plantejament de PC-DOK és general (i és per això que l'hem inclòs en aquest apartat), combina tècniques tractades a d'altres apartats. Cal així mateix fer notar que el programari PC-AXIS, inclòs a l'apartat 2.2.5, constitueix un avenç de PC-DOK.

És interessant d'afegir que (segons un text intern de Per Cronholm, tramès per Malmborg via e-mail), al 1995 s'ha iniciat l'anomenat *Database project*, que està constituït per tres components: *Metadata, Micro/macro* i *Output*, que complementen i/o substitueixen els sistemes descrits suara.

Metadata tindrà 2 parts: una base de dades contenint textos i una altra contenint metainformació emmagatzemada en SQL. Explícitament, el seu objectiu és doble: el de ser la porta d'entrada per recuperar informació a les bases de dades i publicacions, i el de ser instrument de coordinació de la producció d'informació estadística.

El component en format SQL contindrà també les classificacions, nomenclatures i codis territorials. També contindrà les descripcions dels fitxers individualitzats (*observation registers*) i de les macrodades (*Micro/macro database*).

La base de dades *Output* organitzarà l'accéss a tota aquella informació (micro, macro i meta).

2.2.3.3. *El sistema EMMA (Enhanced Metainformation Management Architecture), de World Systems (Europe)*

EMMA és el primer producte comercial de programari concebut específicament per suportar totes les necessitats d'un sistema global de metainformació estadística. Ha estat elaborat per World Systems (Europe) i dissenyat per De Vaney (1994c),

[16]Normativa que Sundgren reprén i re-elabora en les seves plantilles de les "Guidelines"; d'altra banda, l'Institut d'Estadística de Catalunya l'ha utilitzat com a referència per a la seva pròpia normativa de l'Arxiu de Documentació Tècnica.

mantenint un estret contacte tècnic amb el grup METIS; d'altra banda, el mateix De Vaney participa en diversos projectes relacionats amb la metainformació estadística (com els ja esmentats SMILE i EISI). EMMA, que té una arquitectura modular, està basat en els principis client-servidor i API (Application Program Interfaces) per als lligams amb d'altres aplicacions, està provist d'una rica interfície gràfica d'usuari, on utilitza tècniques hipertext, i està desenvolupat per a treballar sota PC/Windows.

La filosofia d'EMMA és la de construir un sistema de metainformació complet separat de les dades, que se suposa que resideixen prèviament en els sistemes d'un institut, dedicant, això sí, un dels seus mòduls (*Data Repository*) a la comunicació i accés amb les bases de dades, amb les que acaba configurant un entorn integrat. La seva primera versió, que funciona encara com a monousuari, està essent avaluada oficialment per diversos instituts d'estadística europeus.

Els mòduls bàsics són els següents:

1 *Subject-Matter Knowledge-Base Module (SMKB)*
2 *Metainformation System Module (MIS)*
3 *Nomenclature Management System (NMS)*
4 *Data Repository Module (DR)*

Als que s'afegeixen les funcionalitats necessàries per a l'administració i manteniment del sistema (*Vegeu les Figures 2 i 3.*).

2.2.4. Sistemes documentals i metainformació estadística

En el context de la metainformació estadística, en la mesura en què es tracta d'organitzar i integrar en un mateix sistema conjunts heterogenis de peces de metainformació estadística, l'opció d'utilitzar l'enfocament dels sistemes documentals és pertinent, tant si el que es fa *és* interpretar que el SME[17] és un sistema documental de ple dret, com si el que es vol és aplicar les tècniques documentals a conjunts aïllats de peces de metainformació estadística.

Prendre aquesta opció suposa al mateix temps que el SME és un sistema *passiu*, és a dir, que es limita a referenciar les dades estadístiques, i es tradueix a aplicar les tècniques dels sistemes de recuperació d'informació (RI) a les peces de metainformació estadística, considerades com a documents. Aquelles consisteixen tradicionalment a emmagatzemar substituts dels documents (en forma de fitxes descriptives, amb referències i descriptors o resums, indicatius de la informació continguda) i construir mecanismes de recuperació i consulta d'informació, més o menys sofisticats.

[17] D'ara endavant, utilitzarem les sigles SME per a referir-nos a un Sistema de Metainformació Estadística.

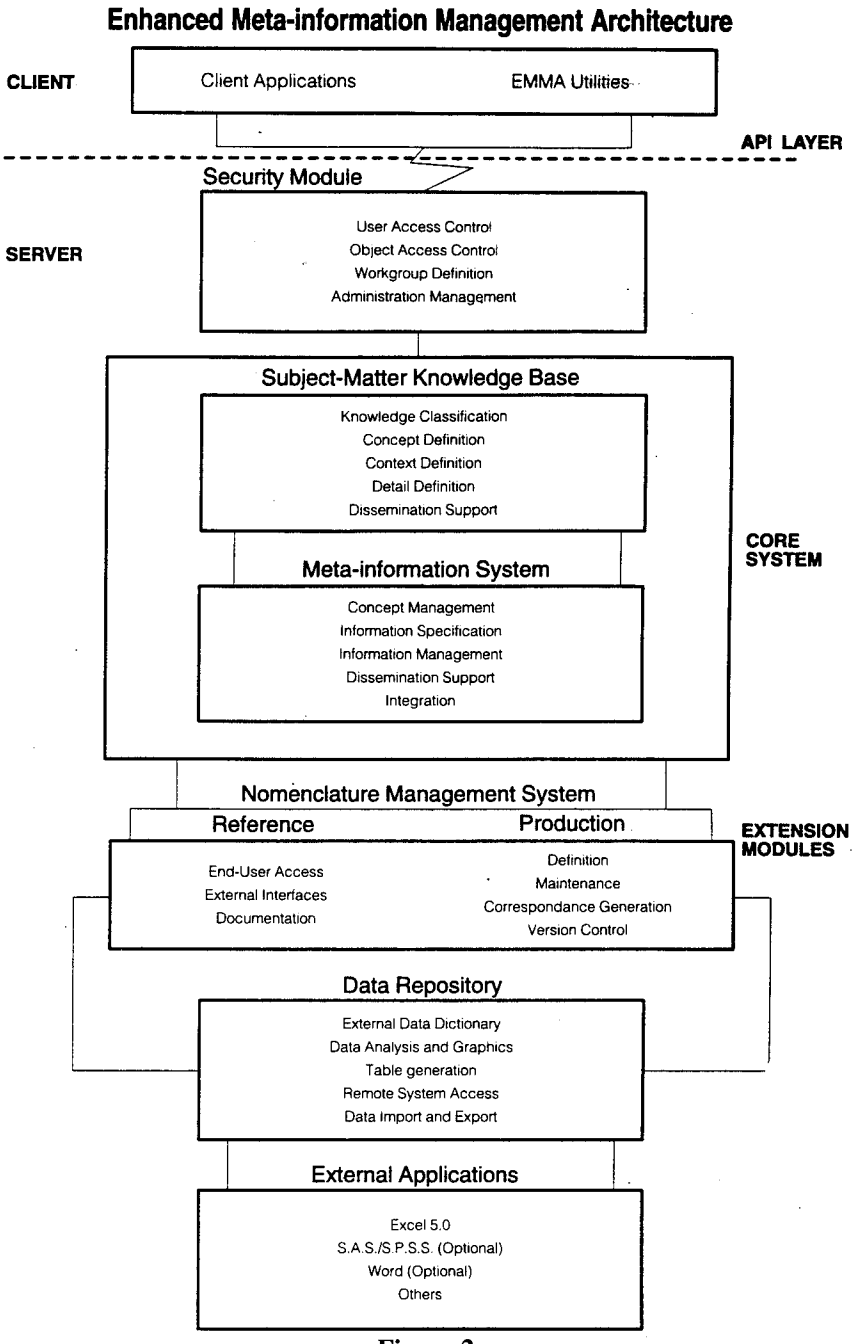


Figura 2

EMMA. Estructura modular.

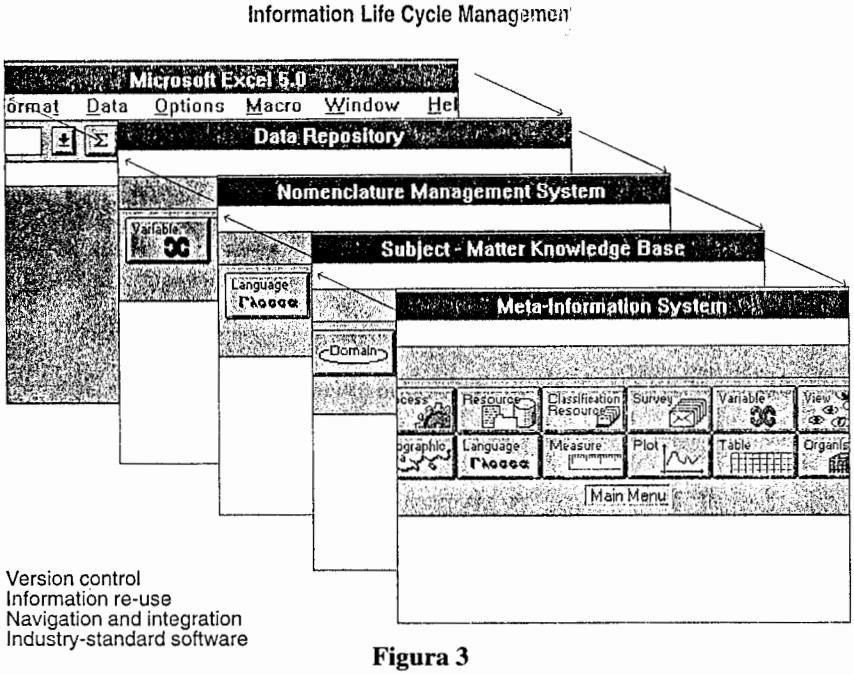


Figura 3

EMMA. Il·lustració de pantalles de mòduls.

Avui dia, però, els nous sistemes documentals es troben en plena evolució, en la mesura que els enregistraments poden contenir el text íntegre (en ASCII o PDF d'Acrobat, p. ex.) o, fins i tot, documents composats, en algun dels formats disponibles: CDA (DEC), Bento (OpenDoc,) OLE (Microsoft), etc. D'altra banda, les tècniques de recuperació s'allunyen de la mecànica de l'àlgebra booleana per adoptar algorismes més sofisticats (probabilístics, de proximitat semàntica, etc.)

Amb aquesta evolució es produeix una convergència dels sistemes documentals amb les bases de dades relacionals, però mantenint la més gran llibertat d'aquells en el tractament d'informació no estructurada.

Una alternativa radical a les tècniques de RI, cada vegada més sovint integrable a qualsevol sistema d'informació, és la de les tècniques *hipertext/hipermèdia*, que ofereixen la navegabilitat lliure entre les peces d'informació (il·limitadament heterogènies) com a forma de consulta específica, a base de seguir les associacions establertes (bé per l'autor, bé per l'usuari) mitjançant lligams (*links*) entre aquelles.

L'opció documental per als SME va normalment associada a un enfocament de la metainformació estadística orientada a l'usuari, en el que es prioritzen sobretot les necessitats dels usuaris externs. D'altra banda, la no-exigència d'informació estruc-

turada per part dels sistemes documentals els dota d'una gran capacitat integradora de sistemes de tota mena, especialment bases de dades heterogènies.

Ja hem pogut veure l'aplicació de tècniques documentals en alguns dels sistemes esmentats fins ara (l'esquema de Graves per al Canadà, PC-DOK, munCAT), i també la veurem en d'altres dels que ens ocuparem en els apartats següents.

Per acabar, només volem esmentar una aplicació específicament documental: la dels sistemes que donen informació sobre les fonts estadístiques rellevants, que hem identificat a l'apartat 1.1) com una de les menes de metainformació estadística més importants per als usuaris finals. No debades Allin (1993) identifica aquesta necessitat d'informació com la primera per a aquests, d'entre les 7 capes en les que agrupa les necessitats de metainformació estadística dels usuaris finals.

Malgrat això, són encara pocs els països que disposen d'una "guia de fonts estadístiques", la necessitat de la qual és cada vegada més reconeguda. El mateix Allin (1993), per exemple, en reclama una per al Regne Unit. I, en l'esquema d'"Statistics Canada", està prevista la conversió del "Statistics Canada Catalogue" en una base documental informatitzada.

Fou l'INSEE el primer que va publicar el "Répertoire de Sources Statistiques" (als anys 60s) i el Servei d'Estudis de Barcelona del Banc Urquijo qui publicava al 1969 la primera "*Guía de Fuentes Estadísticas de España*", continuada i ampliada pel *Consorci d'Informació i Documentació de Catalunya* amb l'*Inventario de Estadísticas de España*, complementat després amb les dues bases de dades públiques (sota BASIS) interrogables en línia: ESPAN i ESCAT (aquesta sobre les fonts estadístiques de Catalunya). Actualment, estan sotmeses a revisió sobre la base de les competències que la llei atribueix a l'*Institut d'Estadística de Catalunya* per a coordinar els organismes productors d'estadística.

2.2.5. La metainformació estadística com informació associada a conjunts d'informació estadística, en suports digitals de difusió

Un problema especial de tractament de la metainformació estadística és el que es planteja quan un institut d'estadística, dintre de la seva estratègia de difusió general, es proposa incorporar la metainformació estadística associada als conjunts d'informació estadística que difon al públic. En realitat, no es tracta d'un sol problema sinó de tants com mètodes i mitjans de difusió utilitzi, i la veritat és que, avui dia, a més del mitjà clàssic de les publicacions impreses (encara, naturalment, indipensables, si bé no sabem per quant de temps), l'oferta tecnològica i metodològica és variada i en curs de transformació, basada, naturalment, en la digitalització prèvia de la informació. És el que s'ha anomenat l'edició electrònica.

Una classificació operativa d'aquests mitjans pot ser la següent:

1) Disquets.
2) CD-ROM i altres suports de memòria massiva.
3) Bases de dades interactives públiques.
4) Servidors *WWW (World Wide Web)* per Internet.

2.2.5.1. Disquets

Disquets contenint taules preformatades és una manera simple i barata de distribució d'informació estadística, oferint una gran varietat de possibilitats: des de taules en ASCII (amb delimitador de columnes pel tabulador) fins a sèries de taules o arxius de microdades provistos de programari especial per a l'extracció, amb possibilitats de certa manipulació de taules.

Aquesta darrera forma és la més útil a l'usuari, però exigeix haver definit algun format que permeti la seva reutilització per a diferents informacions, i que faciliti els intercanvis d'informació entre un institut i els seus usuaris. Exemples interessants de formats incorporats a programari: *STATVIEW* (Netherlands Statistical Institute), *EXTRACT* (US Bureau of the Census) (Zeisset 1993) i *PC-AXIS* (Statistics Sweden) (Nordbäck 1992).

L'INE, per la seva banda, ha desenvolupat el programari *INEDAT*, per a PC/MS-DOS, que permet visualitzar i extreure taules i documents a partir de les seves publicacions estadístiques en disquet. Una nova versió gràfica permet visualitzar informacions en finestres simultànies, i accedir des d'una taula a la informació metodològica associada. També distribueix en disquets algunes Classificacions, provistes d'un sistema de consulta per a PC/MS-DOS (*SIN=Sistema Informatizado de Nomenclaturas*).

El problema general és com incorporar la metainformació estadística, donada l'escassa capacitat dels disquets quan hi volem incorporar informació en formats rics (la qual cosa obliga sovint a comprimir- la). Una possibilitat consisteix a difondre disquets on les taules i la documentació annexa només poden ser consultades, sense possibilitats de manipulació per l'usuari. S'utilitzen aleshores llenguatges orientats als objectes com C++ o programari de desenvolupament hipertext (HyperCard, ToolBook, Folio, Smartext, etc.). Exemples recents en són:

Bizkaiko Zentsuak (Censos de Bizkaia). Diputació Foral de Bizkaia.

Informació de Base. Pla Territorial Metropolità de Barcelona (Generalitat de Catalunya)

Catalunya en Xifres. Institut d'Estadística de Catalunya

2.2.5.2. *CD-ROM com a suport de memòria massiva*

La gran virtut del CD-ROM és la seva gran capacitat de memòria, que permet eixamplar enormement les possibilitats d'emmagatzemament, gestió, manipulació, presentació i extracció, tant de la informació estadística com de la metainformació estadística associada, per ser utilitzat mitjançant aparells lectors de CD-ROM en entorns PC o Macintosh, en general.

No tot, però, és possible alhora i, en l'estat actual de les coses, un compromís és encara necessari entre la sofisticació i multiplicitat de prestacions i la simplicitat d'ús per a l'usuari, en un context de demanda potencial real encara reduïda (tot i que creix ràpidament).

No podem aquí detallar la problemàtica general de la informació estadística distribuïda en CD-ROM. Deixem constància només de la diversitat de plantejaments dels instituts d'estadística, que utilitzen criteris, metodologies i programari diferents, la qual cosa representa un problema afegit per als usuaris.

Respecte de la metainformació estadística, comencem per dir que en el CD-ROM ja no tenim els problemes de capacitat a que al·ludíem al parlar dels disquets. Així podrem, en principi, pensar en incorporar informació textual en formats tipogràfics rics, i també gràfiques, o mapes, per exemple, i utilitzar-los no simplement com a peces d'informació passives, per a ser desplegades a la pantalla, sinó també per ser manipulades per l'usuari o bé formant part d'interfícies gràfiques d'usuari completes.

El problema n'és un de metodològic i tècnic en el context dels sistemes d'informació. D'una banda ¿com associar cadascuna de les peces de metainformació estadística a la informació estadística corresponent (i recordem l'heterogeneïtat d'aquella)? De l'altra, com construir al mateix temps i, complementàriament, un sistema de consulta de la metainformació estadística prou sofisticat per ser útil a usuaris molt diversos (en les seves necessitats i en els seus coneixements)?

La solució passa per fer compatibles i coherents en un mateix sistema dos sistemes de filosofia diferent: l'enfocament *base de dades*, adequat per a tot el que fa a la informació estadística, i l'enfocament hipertext, adequat pel que fa a les peces de metainformació estadística i a la interfície general d'usuari. Així ho van descobrint amb la pròpia experiència els instituts que han elaborat CD-ROMs, tot i que l'estat de la tecnologia no permet desenvolupar fàcilment aquestes solucions híbrides, més fàcils i potents ara com ara en l'entorn Macintosh que sota PC/Windows. De tota manera, l'evolució general cap a sistemes orientats a l'objecte promou la convergència tècnica entre aquells dos enfocaments; a més a més, en la pràctica actual, cada vegada hi ha més casos d'incorporació de tècniques hipertext als sistemes clàssics de bases de dades. El *Help Compiler* i el *Multimedia Viewer* de Microsoft en són exem-

ples. Esmentem també *LinkWorks*, de Digital, que permet desenvolupar interfícies en Macintosh i PC/Windows associades a bases de dades corrent sobre VAX (en entorn VMS).

De fet, retrobem, encapsulats en un suport material, tots els problemes de la metainformació estadística, en general. L'avantatge que tenim ara és el de que, al tractar-se d'un suport autònom d'informació, el CD-ROM permet als instituts d'estadística experimentar en l'organització de sistemes de metainformació estadística a petita escala, sense interferir amb la producció, tenint a més la possibilitat de l'efecte demostració sobre l'eficàcia de certes tècniques.

Tanmateix, no cal oblidar que la informació i metainformació estadística a encapsular en un CD-ROM segueix essent tributària dels sistemes d'informació de l'institut i, per tant, caldrà preveure que subministrin les peces d'informació i de metainformació estadística *amb les característiques i en la forma i format específics al sistema utilitzat al CD-ROM*, sota pena d'haver de repetir molta feina inútilment. D'aquí se'n dedueix que, si bé el CD-ROM permet construir sistemes autònoms, el seu plantejament no pot portar-se a terme totalment a banda dels sistemes de producció; cal un plantejament global, en el que el CD-ROM sigui una peça del conjunt. És més, partint de la base que un nombre a priori indeterminat, però important, de peces de metainformació estadística caldrà que siguin definides i elaborades expressament per al CD-ROM, donat que no corresponen a necessitats del sistema de producció, serà convenient que l'esmentada definició sigui feta al si del mateix plantejament global, preveient que, en el futur, estigui compresa i sigui coherent amb l'estratègia de l'institut en qüestió.

Com a exemples de CD-ROM amb plantejament específic de metainformació estadística, podem reprendre, en primer lloc, els ja esmentats *LMAS* i *CANSIM*, del Canadà, descrits per Podehl (1993). El CD-ROM *LMAS (Labour Market Activity Survey)* conté informació provinent de l'enquesta sobre el mercat de treball a diferents nivells. La documentació sobre l'enquesta a donar als usuaris ocupa varis volums i és molt complexa, de manera que s'ha organitzat en un sistema hipertext, desenvolupat sota *Folio Views* per a PC/Windows, que complementa i s'articula amb un altre sistema de gestió de dades, desenvolupat amb un programari de la *Ohio State University*. El conjunt ocupa 2 gigabytes d'informació distribuïts en 3 discos compactes.

El CD-ROM *CANSIM*, que es publica semestralment des de fa uns anys com a actualitzacions de la base de dades de sèries temporals del Canadà no contenia fins al 1993 altra metainformació estadística que unes definicions bàsiques, moment en el que fou reelaborat per tal de, a partir de l'experiència de LMAS, incorporar-hi un sistema hipertext. Aquest consistia (segons el projecte de Podehl 1993) en articular una interfície entre el banc de sèries, el directori de dades (*SDDS=Statistical Data*

Directory System) i el catàleg "explicat" de publicacions, de manera que l'usuari hauria de poder anar de l'un a l'altre seguint el fil d'un tema determinat.

Un altre exemple d'articulació de la metainformació estadística en un CD-ROM és el *Cens de Població de Catalunya 1991* en curs d'elaboració a l'*Institut d'Estadística de Catalunya*, per ser consultat sota Windows. La informació estadística consisteix en un conjunt de taules bàsiques censals per a tots els municipis de Catalunya i tots els nivells territorials superiors. La interfície gràfica de consulta, desenvolupada sota *FoxPro/Windows*, permet seleccionar, manipular i exportar taules a diferents nivells, visualitzant-les en finestres simultànies, i obtenir automàticament piràmides de població, gràfiques i mapes de coropletes, mitjançant barres de menús i paletes de navegació (Vegeu Figura 4).

En el que fa a la metainformació estadística, s'han previst dos entorns diferents, utilitzant tècniques hipertext. Al primer, consistent en finestres especials d'*Info*, s'accedeix quan l'usuari prem la icona "i", i es presenten les definicions i tipologies de tabulació corresponents a la taula que està activa en aquell moment. Aquestes finestres d'*Info* contenen a més "botons" que permeten accedir a l'altre entorn, el mòdul META, anant a parar directament a peces de metainformació estadística relacionades amb aquella taula activa. Aquest mòdul és així un entorn connectat amb el de les taules, però autònom, i ha estat desenvolupat utilitzant el *Help Compiler*, integrable amb *FoxPro/Windows*, i que disposa de la seva pròpia interfície de consulta i navegació. La metainformació estadística continguda és la documentació metodològica bàsica del Cens de Població 91. Podria, eventualment, utilitzar-se el *Multimedia Viewer*, enlloc del Help Compiler, ja que es tracta del mateix entorn, però amb moltes més prestacions, i permetria construir un sistema de metainformació estadística complet a diferents nivells sobre les múltiples i heterogènies peces d'informació del *Sistema Estadístic de Catalunya*, com ja es va experimentar en el prototipus *munCAT*, sobre Macintosh.

Citem també el CD-ROM "Anuari estadístic de la ciutat de Barcelona. 1993", que és un reflex fidel de la versió en paper; es tracta, doncs, d'un sistema provist d'una interfície de presentació de textos i taules, sense possibilitat de tractament, desenvolupat sota Microsoft Viewer, i editat per l'Ajuntament de Barcelona.

Un altre ús del CD-ROM per part dels instituts d'estadística és el d'emprar-lo com a vehicle de difusió d'arxius de microdades. En aquest cas, es prioritza simplement la capacitat de memòria massiva d'aquell suport sobre la sofisticació de la interfície, i no s'hi inclou normalment cap mòdul de metainformació estadística, entenent-se que els usuaris d'aquesta mena d'informació són experts que coneixen a bastament amb quina informació se les han d'haver. Aquest ús del CD-ROM, suport de memòria òptica, l'assimila a d'altres suports de memòria magnètica massiva.

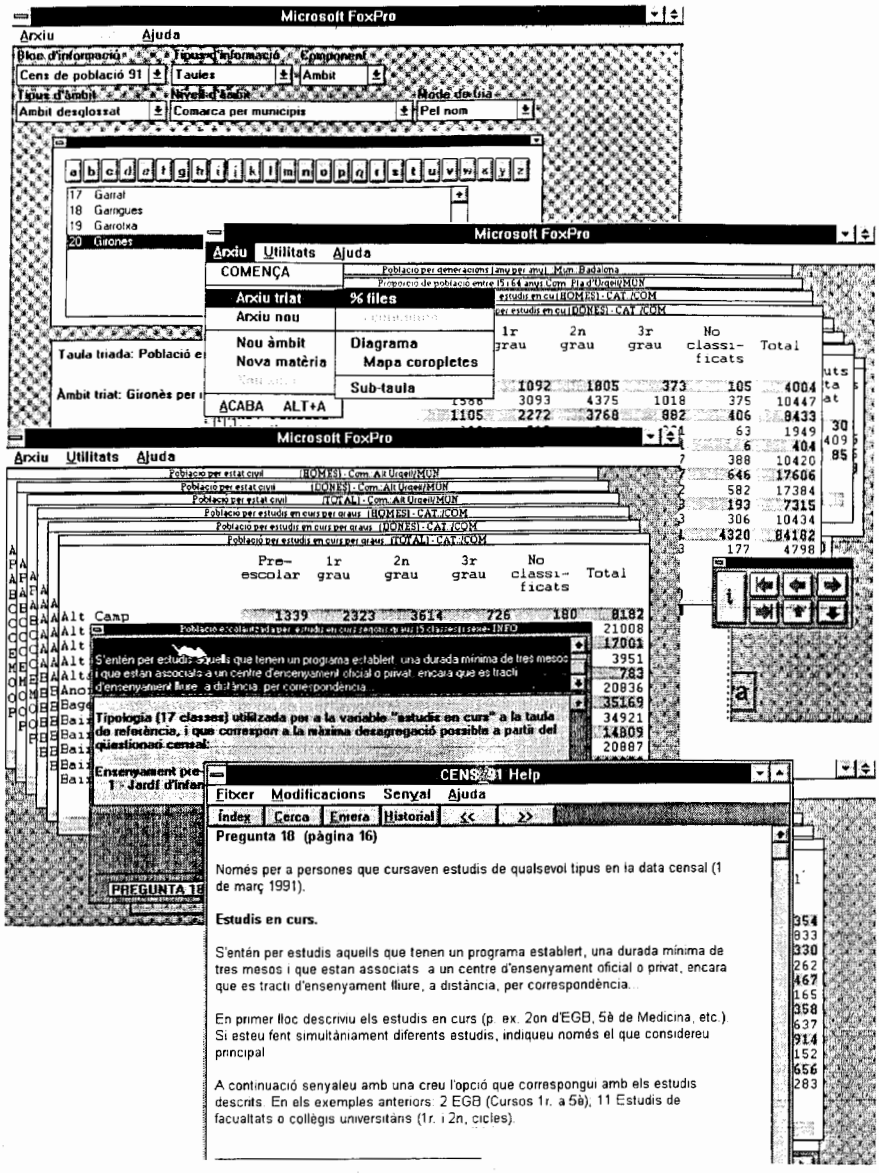


Figura 4

CD-ROM "Cens de Població de Catalunya 1991". Il·lustració de pantalles. Menús, finestra Info, Finestra Meta...

2.2.5.3. *Bases de dades interactives públiques*

En les bases de dades interactives que els instituts d'estadística han anat constituïnt i posant al servei del públic (directament, o a través de servidors comercials), tant si eren bases de dades bibliogràfiques (tipus "guies de fonts estadístiques") com bases de dades numèriques (tipus "sèries temporals" o "arxius estadístics de base territorial"), no es preveia en general cap mòdul ni tractament específic de metainformació estadística.

La raó és doble: d'una banda, el propi plantejament metodològic de la metainformació estadística és recent; de l'altra, hi ha dificultats tècniques per a engarçar i articul·lar la metainformació estadística amb el mode de tractament de la informació numèrica utilitzat per oferir la interrogació pública, que normalment correspon a desenvolupaments molt específics de l'enfocament bases de dades relacionals sobre grans ordinadors (mainframes), els quals només recentment han començat a incorporar lentament l'orientació a l'objecte. Justament, aquestes dificultats han fet que l'opció CD-ROM s'utilitzi com a terreny d'experimentació de noves solucions, tal com hem apuntat.

A més, l'òptica tradicional d'interrogació de bases de dades numèriques obligava els usuaris a aprendre i manegar llenguatges complexos de definició, selecció i extracció de taules, en un entorn relativament opac per als no experts, que ja ni intentaven aclarir-s'hi.

Deixem constància de que s'ha realitzat per part dels desenvolupadors un cert esforç per fer més còmode l'accés a les dades (tant si són arxius de sèries temporals com taules), a través de noves interfícies d'usuari on, malgrat tot, la documentació metodològica, quan hi és, està disponible en mòduls separats, si bé en certs casos s'hi pot accedir directament des d'una taula.

a) *BEMCAT i SAETA*

La base de dades *BEMCAT (Base d'Estadístiques Municipals i Comarcals de Catalunya)*, desenvolupada per l'*Institut d'Estadística de Catalunya* n'és un exemple. Conté taules amb estadístiques de tota mena relatives als municipis, comarques i altres nivells territorials de Catalunya, i dóna accés opcionalment a unes "descripcions temàtiques" associades a una taula o als nivells temàtics superiors. *BEMCAT*, base de dades relacional RDB residint en un ordinador VAX, és accessible (de moment, internament) bé per una interfície de menús (per emulació de terminal VT200), bé per una interfície gràfica en OSF/MOTIF.

Una altra forma de procedir consisteix a proveir els usuaris amb un programari de selecció i extracció de taules. Un exemple n'és el programari *SAETA (Sistema de Archivo y Extracción de Tablas)*, desenvolupat per l'INE per a visualitzar i

extreure taules i documents metodològics associats amb elles a partir de les seves bases de dades. També el PC-AXIS, de l'institut suec, ja citat, fa aquesta mateixa funció.

b) *MDB (U.S. Bureau of the Census)*

El plantejament de futur que fa l'*US Bureau of the Census (BOC)* respecte de la metainformació estadística confirma les dificultats per oferir solucions combinant en un sol sistema la informació estadística i els paquets de metainformació estadística necessaris. Tal com explica Gillman (1994), el BOC s'ha concentrat en treballar sobre la metainformació estadística en paral·lel a les bases de dades numèriques dissenyant, per encàrrec del Departament de Comerç, un prototipus de base de dades pública anomenada *MetaDataBase (MDB)*, l'objectiu de la qual seria el d'assistir als usuaris d'informació estadística en la localització dels conjunts específics de dades pertinents a les seves necessitats concretes, ocasionals o sistemàtiques. La informació estadística de referència prevista per a la base de dades final no es limita a la del BOC; al contrari, es pretén incorporar totes les fonts de dades de l'administració federal, si bé el prototipus inicial es limita a un subconjunt de la informació del BOC. D'altra banda, quedaria per a un projecte posterior l'articulació directa amb les dades estadístiques mateixes.

Als efectes de *MDB*, la metainformació estadística d'interès es divideix en tres categories: *de sistema, d'aplicació* i *administrativa*, per tal de satisfer les diverses necessitats dels usuaris. La *de sistema* es refereix, bàsicament, a informació sobre la lògica i el tractament informàtic utilitzat en la creació de les dades; la *d'aplicació* comprèn els aspectes estadístics necessaris per interpretar correctament les dades, i l'*administrativa* correspon a aspectes institucionals i administratius necessaris per accedir a les dades i a les persones de contacte pertinents. Com es veu, l'objectiu és a la vegada comprensiu i pràctic en el que fa a la metainformació estadística, si bé, recordem-ho, parcial, en la mesura que *només* intenta crear un SME deslligat de les dades mateixes.

El model de referència del sistema MDB en ell mateix és l'estàndar *IRDS (Vegeu l'apartat 2.3.3)* i es preveu connexions amb els sistemes de recuperació ja implementats i públicament disponibles per a conjunts particulars de metainformació estadística, com són, p. ex.: *ARRK (Automated Reference Rack)*[18], *DES=Data Extraction System* (Survey-on-Call) i *CENDATA* (sobre dades censals, accessible per Dialog i Compuserve), etc. D'altra banda, també s'utilitzarà el format "BOX file" ja existent com a llenguatge de descripció de taules, que permet incloure camps específics de metadades.

[18]ARRK és un sistema hipertext, desenvolupat amb el programari de desenvolupament *Smartext*, de Lotus.

Un dels requeriments de *MDB*, (amb exigències de llarg abast en el que fa a la coordinació entre organismes productors) que té el seu paral·lel en una de les preocupacions de METIS, és el d'elaborar un lèxic comú terminològic i un model de dades coherent amb la nova filosofia i consistent amb les heterogènies pràctiques dels diferents productors.

Des del punt de vista informàtic, per a la primera realització del prototipus, encara a petita escala (enfocat a la verificació de la consistència lògica i pràctica del model), s'utilitza *Infospan Open Repository*, l'únic (fins a la data) sistema de desenvolupament existent complint les especificacions *IRDS*.

c) *Videotex*

Apart de les bases de dades interactives clàssiques, com a sistemes de recuperació d'informació, hem de citar el sistema Videotex que, per les seves característiques i limitacions, no permet utilitzar-lo fins ara més que com a sistema de presentació d'informació i metainformació estadística molt sintètiques per al públic en general; o bé com a sistema de publicació electrònica de dades molt freqüents i poc detallades (exemple típic, l'IPC = Índexs de Preus al Consum). Entre d'altres, l'INE manté a Espanya una base per Ibertex i l'*Institut d'Estadística de Catalunya* posa ara al públic una primera base d'informació estadística, anomenada XIFRES, també per la xarxa Ibertex.

2.2.5.4. Servidors WWW (World Wide Web) per a Internet

Un dels desenvolupaments dels darrers anys conduents a fer més fàcil l'accès i la interrogació de les bases de dades interactives ha consistit en desenvolupar interfícies d'usuari per a entorns PC/Windows i Macintosh, i que actuaven bé com a interfícies de comunicació simples, bé com a clients en un entorn client-servidor.

Sense invalidar aquest desenvolupament, els instituts d'estadística més avançats no han deixat d'explorar les noves possibilitats d'accés pràcticament universal ofertes per la xarxa Internet, de primer constituint servidors d'arxius "ftp anonymous" i "gophers" i, darrerament, creant servidors de pàgines *World-Wide-Web (WWW)*, en el que està esdevenint una moda abassegadora o, potser, un nou servei indispensable per a distribuir no solament informació estadística (tant micro com macrodades) sinó també i, pel que a nosaltres interessa, sobretot, metainformació estadística. La raó és senzilla i se sustenta en una doble constatació: d'una banda, permet continuar posant a disposició del públic les bases de dades estadístiques ja existents, però amb la possibilitat de construir interfícies gràfiques d'usuari relativament independents dels sistemes del servidor i, al mateix temps, d'una gran simplicitat, amb l'avantatge addicional que, definint un sol entorn, es posa a disposició pràcticament de tot el món;

d'altra banda, la filosofia de base de l'entorn *WWW* és hipertext, i les pàgines web s'escriuen amb un llenguatge d'orientació "document composat" (*HTML=HyperText Management Language*, un subconjunt de l'estàndar *SGML*), que poden anar provistes de "botons" (*links*) que les associen no solament amb d'altres parts del mateix servidor sinó també amb les de qualsevol altre servidor WEB d'Internet. Les possibilitats que aquesta tecnologia ofereix per a la metainformació estadística són, així, impresionants, tot i recordant els problemes de tipus conceptual i d'estandardització als que hem fet referència més amunt.

Com explica Cathryn Dippo (1994), l'*US Bureau of Labour Statistics* aprofita l'entorn *WWW* per a distribuir la informació estadística (bàsicament, sèries temporals territorialitzades i enquestes especials) ja disponible a les seves bases de dades *BLS*, a través d'una interfície gràfica Web, que simplifica la selecció de taules i el format d'exportació desitjat. A més, el *BLS* s'ha pres seriosament l'oportunitat d'aprofitar les prestacions hipertext i la presentació de documents en format ric (PDF d'*Acrobat*) per oferir diversos conjunts de metainformació estadística: d'una banda, la documentació metodològica associada als arxius estadístics, cartogràfics i gràfics i, de l'altra, documentació metodològica general, programari especialitzat i informació institucional (l'actualitat estadística, organigrames i persones a contactar, calendaris d'operacions estadístiques, etc.).

2.2.6. El sistema de metainformació estadística, com a instrument per a la coordinació

Ja hem esmentat en algun moment que la metainformació estadística pot jugar un paper integrador de primer ordre, en la mesura que se situa en el punt més proper a l'usuari i, per tant, més global (*Vegeu, a més, l'apartat 2.4*). D'aquí a fer-li jugar el paper d'instrument per a la coordinació de diverses institucions no hi ha més que un pas.

Així ho ha vist, per exemple, l'institut d'estadística de Berlín que, tal com defensa Appel (1994), entén que un SME és una precondició per assolir una estandardització bàsica en els processos de producció i, així, estén el seu sistema *DUVA* a les 50 grans ciutats alemanyes, l'objectiu del qual és el de coordinar els registres municipals administratius de població. Ja hem esmentat a l'apartat 2.2.3.1 el paper fonamental que el Tesaure jugava al si del seu SME com un dels sis mòduls (Tesaure, Textos, microdades, macrodades, Directori i Confirmacions de qualitat). En l'articulació progressiva del SME de DUVA es van establint normes consensuades de procediments i definicions, entorn d'un model de dades, que incorpora expressament els estàndars internacionals, com els requisits GESMES (EDIFACT) per a l'intercanvi electrònic d'estadístiques.

Un altre exemple n'és *GENIE* (Walker 1993), un macroprojecte liderat per la Universitat de Loughborough que consistirà en el sistema de base del *Global Environmental Change Data Network*, que utilitza al seu torn la xarxa JANET.

L'esmentem aquí, encara que l'objectiu del sistema es redueix a recollir i proporcionar informació al servei dels centres de recerca en una temàtica específica (el canvi climàtic del món) i els productors d'estadístiques oficials només hi participen com a proveïdors de dades, per la raó que el nucli integrador de les bases de dades a constituir serà un sistema de metainformació estadística, la definició i alimentació dels elements del qual serà l'arma principal en la coordinació de les interrelacions amb els diversos centres.

Recordem, finalment, com el projecte DSIS d'EUROSTAT atorgava un paper integrador al *common reference environment*, del que la metainformació estadística és part fonamental.

2.2.7. La metainformació estadística, integrada en un sistema expert

Tal com explica Hand (1992), una de les àrees més prometedores de l'aplicació de les tècniques de la Intel·ligència Artificial (IA) a l'estadística és la dels sistemes experts, en la mesura que es proposen ajudar els usuaris no-experts en l'anàlisi de la informació estadística, suggerint, p. ex., hipòtesis, mètodes alternatius d'exploració i d'interpretació, etc., per a conjunts de dades concretes.

D'Angiolini (1993) va més lluny defensant que el propi sistema d'informació estadística pot enfocar-se amb l'òptica "coneixement", definint-lo aleshores com:

"el corpus de coneixements d'un agent (personal o institucional) l'objectiu del qual és augmentar el seu propi coneixement sobre l'entorn, mitjançant activitats estadístiques com:

- recollir informació (estadística) a base d'observar parts determinades del món real,
- aplicar tècniques d'anàlisi estadística per a augmentar el seu coneixement".

En aquest context, és evident que la metainformació estadística ha de jogar un paper fonamental en la configuració i contingut de la base de coneixement del sistema. El programa *DOSES* d'EUROSTAT així ho ha entès al seleccionar dos dels projectes especialment dedicats a la metainformació estadística: *MMD* i *EISI*.

MMD (Modelling Metadata), descrit per Darius (1993) consisteix en un model formalitzat de l'ús de la metainformació estadística en relació amb les diferents tasques dels usuaris d'un sistema d'informació estadística. Una anàlisi de les tasques des

del punt de vista dels coneixements implicats ha portat a la definició d'un conjunt d'operadors, articulables en cadenes, uns dels quals signifiquen manipulacions de les estructures de dades/metadades que donen com a resultat altres estructures de dades/metadades (p. ex, selecció d'una columna, fusió de conjunts, etc.); mentre d'altres generen resultats finals (generació de taules, generació de gràfics). Aquests operadors tenen com a característica especial que no es limiten a tractaments mecànics, sinó que tracten les dades i les metadades en funció de les circumstàncies. Un prototipus ha estat elaborat com una aplicació en l'entorn SAS, per validar conceptes del model.

En el segon projecte sobre metainformació estadística, *EISI (Expert Interface to Statistical Information)* (De Vaney 1992), s'ha elaborat un prototipus, l'arquitectura del qual consisteix en la Interfície d'Usuari i la Base de Coneixement. La interfície és doble: la primera (*Domain Browser*) proporciona mecanismes de navegació i consulta de la base de coneixement, utilitzant un motor hipertext; la segona (*Domain Assistant*) suporta l'accés directe a la metainformació a través de raonaments en el context dels problemes específics del domini d'aplicació, i també mitjançant l'expansió de les especificacions de solucions parcials als problemes, en termes de la metainformació disponible. Pel que fa a la Base de Coneixement, es tracta d'una Base estructurada en tres components: *The usage scenario library, The metainformation database* i *The domain encyclopædia.* La base de dades de metainformació gestiona representacions dels elements, estructures i relacions de metainformació del domini d'aplicació. L'arquitectura d'aquesta base de dades reposa sobre el model elaborat al si del grup de treball METIS: *User's Guide to metaInformation Systems in Statistical Offices.* (UN/ECE 1990).

Un altre dels projectes seleccionats, *CASIP (Complete Automated System for Information Processing)*, descrit per Saris (1992), consisteix en realitat en l'elaboració d'un conjunt de sistemes experts concebuts per suportar cadascuna de les fases de treball d'un camp concret: les enquestes de pressupostos familiars: *EDC (Expert System in Data Collection), MDBS (Micro Data Database System), EDA (Expert System for Data Analysis)* i *EPSS (Expert System for Presentation of Summary Statistics).*

EDA (Expert System for Data Analysis), descrit per Gibert (1992) i Aluja (1993) i elaborat al si del Departament d'Estadística i Investigació Operativa de la UPC, és un entorn integrat per a la producció i l'anàlisi de taules estadístiques i està composat així per dos mòduls: *STG (Statistical Table Generator)* i *STA (Statistical Table Analyzer).* El primer, *STG*, està preparat per a la producció de taules complexes de síntesi a partir de les dades i metadades emmagatzemades pel sistema precedent (*MIDAS*) o bé directament a partir de fitxers ASCII. Al seu torn les taules generades poden ser emmagatzemades a la base de dades estadística del sistema (que inclou també les metadades oportunes) o bé exportades en formats ASCII o LOTUS 1-2-3. El segon mòdul, *STA*, ofereix un conjunt d'eines generals per emmagatzemar, organitzar, tractar, i analitzar gràficament la informació continguda a la base de dades estadística. Els resultats

(dades, gràfics, noves taules...) poden així mateix ser exportats a processadors de textos corrents o a sistemes d'autoedició per a la producció (i eventual publicació) d'informes estadístics.

Com es veu per aquest petit resum, l'activitat en aquest camp és notable. Resta per veure la convergència en la pràctica dels conceptes i funcionalitats involucrats en el desenvolupament de sistemes experts amb els que s'utilitzen en els altres diversos enfocaments, tampoc homogenis entre sí.

2.3. Conclusions de l'experiència internacional

Efectivament, hem pogut constatar la diversitat de plantejaments dels diferents instituts d'estadística, que provenen, tant de la diversitat de situacions de partida com (sobretot) del fet de plantejar-se diversament la problemàtica de la metainformació estadística: bé com un conjunt de problemes particulars o bé com un problema global. Els instituts que prenen aquesta darrera opció resulten ser també els que són més conscients de l'evolució en curs cap a una obertura dràstica dels seus actius d'informació a uns usuaris cada vegada més diversificats i il·lustrats, i de la necessitat de dissenyar i implementar serveis més sofisticats i integrables amb els entorns de consulta i treball dels usuaris. Al mateix temps, són també els més confiats en què la revolució en curs de les tecnologies de la informació i la comunicació i la disponibilitat de sistemes cada vegada més sofisticats fan concebibles solucions pràctiques a aquells objectius.

Ara bé, en la mesura en què s'intenta un plantejament global de la metainformació estadística es manifesta clarament la necessitat d'un model conceptual potent, en el que les diferents tasques, funcions i peces d'informació d'un institut d'estadística tenen el seu lloc, i en el que les múltiples, variades i complexes necessitats dels diferents usuaris estan previstes de manera pràctica i eficient. I això en un context de creixent interdependència i coordinació a diferents nivells, entre instituts, empreses i administracions.

D'aquí la importància de la tasca de reflexió endegada de fa vint anys ençà pels pioners de la metainformació estadística, Bo Sundgren al capdavant.

La conclusió comuna, integrada explícitament per Sundgren als seus treballs, i expressada col·lectivament pels participants al grup de treball METIS, és d'una lògica aclaparadora.

En poques paraules, és la següent:

> *Només si comprenem en totes les seves virtualitats les interrelacions entre les diferents tasques d'informació d'un institut d'estadística i som capaços*

de la seva conceptualització per la via d'un esquema globalitzador, podrem interpretar correctament les interrelacions entre la metainformació estadística i la informació estadística i, per tant, podrem dissenyar sistemes de metainformació estadística integrats.

És clar que si recordem que la necessitat de sistemes de metainformació estadística vé determinada per objectius d'eficiència en l'articulació de sistemes d'informació estadística, ens pot semblar que estem en un cercle viciós si ara diem que el disseny correcte de sistemes de metainformació estadística depèn de la correcta comprensió dels sistemes d'informació estadística en un institut d'estadística. De fet, com fa notar Sundgren (que parla de "simbiosi entre les dades i les metadades"), no ha d'estranyar l'íntima interrelació entre ambdues menes de sistemes perquè és la que existeix entre un sistema d'informació i el sistema objecte sobre el que aquell pretén informar.

Aquesta necessitat no exclou la dificultat de l'empresa, que al seu torn explica les dificultats i vacil·lacions amb què actuen els instituts d'estadística.

Per acabar aquest apartat, ens agradaria fer-ho resumint l'essencial de les observacions que fa Sundgren (1994b) en les seves *Guidelines*[19]. Són les següents:

Molts sistemes de metainformació estadística han fracassat en el passat degut a que:

a) La recollida i manteniment de metainformació estadística és una feina avorrida, cara i llarga en el temps.

b) Desafortunadament, existeix una desconnexió entre els usuaris i els productors de metainformació estadística al si d'un institut. Els que necessiten la metainformació estadística no poden produir-la ells mateixos. D'altra banda, els que "posseeixen" els coneixements sobre les dades estadístiques no troben que en puguin treure massa avantatges de sistemes formalitzats i automatitzats de metainformació estadística.

Així, doncs, des d'una perspectiva positiva, per evitar les males experiències del passat, Sundgren al·ludeix a les següents característiques desitjables del futur SME (Sistema de Metainformació Estadística):

a) Les activitats de recollida de metainformació estadística haurien de ser minimitzades en el sentit de que cap peça de metainformació estadística no hauria de ser entrada més d'una vegada, i les peces susceptibles de ser deduïdes ho haurien de ser de manera informatitzada, i no manualment.

[19] A l'apartat 5.1 (*Experiències del passat i implicacions pel futur*). Aquest resum no pretén substituir el text original, més detallat.

b) Les activitats de recollida retrospectiva massiva de metainformació estadística haurien de ser evitades. En canvi, el major nombre possible de peces de metainformació estadística haurien de ser generades com a subproductes d'altres activitats.

c) S'hauria d'introduir alguna mena d'anàlisi cost/benefici en l'arquitectura d'un SME amb l'objectiu de relacionar usuaris i productors de metainformació estadística d'una manera constructiva.

En definitiva, acaba dient:

> *"Els sistemes d'informació estadística són actius valuosos. Ara bé, sense sistemes de metainformació estadística adequadament integrats, la vàlua d'aquells sistemes es redueix dràsticament. Donat que avui dia els sistemes d'informació estadística estan d'una manera general formalitzats i informatitzats, els sistemes de metainformació estadística han de ser també formalitzats i informatitzats, si es vol que els fluxos de metainformació segueixin el pas de la seva informació objecte, la informació estadística."*

2.4. El grup de treball METIS (UN/ECE)

2.4.1. Programa de treball

El grup de treball METIS, organitzat al si de la *Conferència d'Estadístics Europeus* (al seu torn dintre de la *Comisió Econòmica per a Europa de les Nacions Unides*), està treballant amb força d'uns anys ençà, per tal de consensuar entre els tècnics representants dels instituts d'estadística dels països europeus (més enllà de la Unió Europea, i als que s'afegeixen regularment a més els de Canadà i dels Estats Units), una sèrie d'aspectes trascendentals sobre la metainformació estadística i el desenvolupament de sistemes de metainformació estadística.

El seu programa de treball, a més de mantenir-se al corrent del d'altres institucions internacionals (en particular, el programa DOSES i, a partir d'ara, el nou programa DOSIS d'EUROSTAT), s'ha concretat especialment en les línies següents:

a) Intercanvi d'informació entre els instituts sobre les línies en les que es basen per a desenvolupar sistemes, globals o específics, de metainformació estadística, i els ensenyaments de les seves respectives experiències.
b) Elaboració d'una terminologia específica comuna per a la metainformació estadística.
c) Inventari i eventual elaboració d'estàndars per ser proposats a les instàncies pertinents.

d) Metodologia per al disseny i desenvolupament de sistemes de metainformació estadística.
e) Criteris d'avaluació de sistemes de metainformació estadística.
f) Mètodes formals per a la descripció de les peces de metainformació estadística.
g) Problemes específics dels països de l'Europa de l'Est per a la transició dels seus sistemes estadístics.

La darrera reunió de treball va tenir lloc a Ginebra, al novembre 1994, mentre la propera no tindrà lloc fins al 1996/97.

2.4.2. METIS i la terminologia

Es tracta d'establir un lèxic de termes relacionats amb la metainformació estadística, i també les seves definicions, com a eina comuna de treball. La darrera versió de l'esborrany, preparada per l'eslovac Prazenka (1994) fou distribuïda a la reunió de novembre 1994, i serà actualitzada i esmenada en el transcurs de 1995.

2.4.3. METIS i els estàndars

Una línia de treball molt important del grup METIS és la relativa a l'estandarització de diversos aspectes de la metainformació estadística.

Una primera tasca consistí en l'elaboració d'un primer element de referència, ja publicat per UN/ECE (1990): *User's guide to metainformation systems in statistical offices*.

Després es va definir la relació entre la metainformació estadística i la informació estadística, al si d'un model de referència general, el *Statistical Reference Model*, proposat per Olenski (1991) en el que es varen identificar tres capes:

- capa de la metainformació estadística (*metadata layer*)
- capa de la informació estadística (*data layer*)
- capa del processament (*computing layer*)

Per a la capa del processament, el punt de partida és considerar un sistema de metainformació estadística, bé com un *Information Resource Dictionary System (IRDS)*, bé com un *Information Resources Management System (IRMS)*, de manera complementària, estàndars en curs de desenvolupament per part de l'ISO[20].

[20]S'està treballant en un estàndar més general, el *PCTE = Portable Common Tool Environment*, del que l'IRDS seria un sub-conjunt, però passaran anys abans que se'n pugui disposar.

D'altra banda, reconeixent que les altres dues capes estan molt relacionades i són ambdues importants per a la metainformació estadística, METIS està treballant en la constitució d'un *Inventory on international standards applicable in statistics* (1a versió, 1993) i en l'anàlisi de l'enfocament adequat per a cadascuna d'elles. Així, l'enfocament basat en el *missatge* s'ha considerat el més idoni per a l'estandardització de la capa de la informació estadística i, per tant, s'ha reconegut la importància de la tasca que ha portat a terme el grup MD6 al si del UN/EDIFACT (Western European Board), que ha elaborat l'estàndar *GESMES (Generic Statistical Message)* per a la definició del missatge estadístic en l'intercanvi electrònic de dades (EDI). Tècnicament, GESMES és un missatge autodefinit, composat de tres parts, de les que les dues primeres són menes de metainformació (administrativa i tècnica, respectivament) referides a les dades, contingudes en la tercera part[21].

En canvi, per a la capa de metainformació estadística, s'ha considerat adequat l'enfocament basat en l'*indicador estadístic* i, conseqüentment, Olenski va elaborar al 1993 una proposta de "model genèric formal per a l'estandardització de la metainformació i els indicadors estadístics", reelaborada i aprovada finalment al 1994. Aquest model, basat en un treball previ en el que es definia un *model semàntic estructural d'indicador estadístic*, accepta l'estàndar GESMES i l'introdueix com un subconjunt. Descriu el seu contingut dient que cobreix els principis de representació de conjunts (holdings) específics de metainformació estadística i defineix regles generals per a l'organització dels indicadors segons diferents formes de missatges: qüestionaris, taules, arxius, gràfiques, etc.

2.4.4. Metodologia per al disseny de sistemes de metainformació estadística

L'objectiu d'aquesta línia de treball ha variat en el temps. Si l'any 1991 s'encarregava a Sundgren que definís "un sistema pilot de metainformació estadística", amb la intenció d'implementar-lo com a banc de proves, al 1993, sota la proposta del mateix Sundgren, l'objectiu es concentrava en els aspectes més genèrics de definir un model de sistema i d'establir unes recomanacions per al disseny de sistemes de metainformació estadística, abandonant justament tota pretensió d'incidir en els aspectes informàtics. A la reunió de novembre 1994, Sundgren (1994) presentava el text definitiu de les recomanacions, sota el títol: *Guidelines for the Modelling of Statistical Data and Metadata*[22], que era aprovat oficialment.

[21] Vegeu-ne una descripció per Maurer (1992). Durant el 1995 s'està procedint a la difusió de GESMES, per a documents (com "*Message Implementation Guide for GESMES/ECOSER*") i programari d'implementació (com "*Windows Help File for GESMES*".

[22] Aquest document és reproduït íntegrament en aquest mateix número de **Qüestiió**.

El raonament de Sundgren, explicitat en diverses ocasions i sobre el que reposen les seves *Guidelines*, comença preguntant-se quin pot ser el concepte de "sistema de metainformació estadística", en el sentit fort de "sistema", més enllà de simples conjunts de metadades, més o menys interrelacionades. El problema consisteix així en trobar els criteris més adequats i eficients que ens permetin articular relacions conceptualment significatives entre les diferents unitats i components d'un sistema de metainformació estadística.

Ara bé, tot sistema d'informació, per definició, es refereix i s'articula entorn d'un sistema objecte, que és el que el justifica i en determina les característiques. Així, *un sistema de* **metainformació** *estadística seria un sistema d'informació l'objecte del qual és un sistema d'***informació** *estadística.* I en la mesura que existeixen diferents menes de sistemes d'informació estadística, també hi haurà diferents menes de sistemes de metainformació estadística.

Per exemple, tradicionalment, els sistemes d'informació dels instituts d'estadística s'han organitzat en general amb una *orientació a la producció (input-oriented)*; és a dir, s'articulen entorn de les operacions de producció estadística i de la sèrie de processos associats a les seves fases.

Ara bé, des del punt de vista de l'ús i dels usuaris de la informació estadística, com subratlla Sundgren, seria "*més adequat organitzar els sistemes d'informació estadística com a sistemes de recuperació i difusió, sobre la base de les necessitats potencials d'informació dels usuaris*". Un sistema d'informació estadística així *orientat a l'usuari (output-oriented)* estaria en les millors condicions per presentar als seus usuaris (tant els tècnics interns com externs) una imatge global ben integrada i conceptualment coherent.

Justament, *organitzar conceptualment i sistemàtica aquesta imatge d'un sistema d'informació estadística és una precondició del sistema de metainformació estadística corresponent, esdevenint així un dels seus objectius inesquivables.*

Així, doncs, per la seva pròpia natura, un sistema de metainformació estadística es presenta al seus usuaris potencials amb la pretensió d'ajudar-los a fer-se una representació mental adequada del sistema d'informació estadística, que és el seu sistema objecte. El resultat esperat és que les accions dels usuaris respecte d'aquest darrer sistema seran així més ajustades, comprehensives i eficaces.

El mètode seguit per Sundgren, en l'elaboració de les seves *Guidelines*, consistirà, doncs, en primer lloc, en definir les *funcions* d'un sistema d'informació estadística i dels diferents *agents* que realitzen les tasques associades en aquelles, en la mesura en què són considerats els usuaris potencials "nats" d'un sistema de metainformació estadística, a afegir als usuaris externs; en segon lloc, tractarà de passar revista a les *necessitats d'informació* (usos) de totes aquestes menes d'*usuaris*, materialitzades,

d'una banda, en certs conjunts o paquets de *peces d'informació* i, de l'altra, en certes funcionalitats o *requeriments del sistema*, que haurien de fer possibles els usos considerats necessaris.

Cal subratllar aquí la transcendència metodològica de considerar, com fa Sundgren, que la funció de *difusió* constitueix una de les funcions normals d'un institut d'estadística, d'on es dedueix que les necessitats de metainformació estadística derivades de les tasques relacionades amb aquella funció són necessitats internes normals de l'institut i no, simplement, necessitats marginals o addicionals.

Efectivament, en la mesura que els agents difusors d'informació estadística d'un institut incorporen en els plantejaments dels sistemes de difusió les necessitats reals dels usuaris externs (i faran justament la seva feina en la mesura en què ho aconsegueixin) podrem analitzar i integrar les necessitats d'aquells sistemes en el plantejament global de les necessitats de metainformació estadística a satisfer pel sistema global de metainformació estadística (SME).

Una altra innovació metodològica de Sundgren en la tipificació d'usuaris d'informació estadística (i, en conseqüència, de metainformació estadística), és la de considerar les pròpies eines de programari (*software tools*) com a usuaris especials, la qual cosa permet analitzar i identificar les seves necessitats sistemàtiques per a integrar-les també en el plantejament global del SME.

En definitiva, a través de la rigorosa consideració dels conceptes implicats en cadascuna d'aquestes fases, Sundgren proposa una sèrie de diagrames de fluxos, que representen models diferents i complementaris d'un sistema d'informació estadística, dels que en dedueix una sèrie de classes de peces de metainformació estadística, per a cadascuna de les quals és aleshores factible definir la informació necessària a recollir en forma de plantilles de documentació esquemàtiques.

Defineix, així, les següents plantilles:

- Declaració de qualitat per a la informació estadística (*micro/macrodata*).
- Documentació d'un arxiu individualitzat de base (*observation register*).
- Documentació d'una enquesta i el seu sistema de producció (*statistical survey*).

Defineix finalment un llenguatge formal de descripció d'objectes de metainformació estadística (en tres capes) i una sèrie de diagrames, que sintetitzen els fluxos d'un sistema de metainformació estadística i n'organitzen l'arquitectura.

Abans, però, fa una sèrie de consideracions sobre les lliçons a treure de les experiències i fracassos del passat, que hem cregut oportú de reprendre a l'apartat 2.3).

2.4.5. Criteris d'avaluació de sistemes de metainformació estadística

Considerant important, malgrat la manca de materials metodològics suficients, l'establiment d'una panòplia de criteris que permetin avaluar l'eficiència i idoneïtat dels sistemes de metainformació estadística, ha estat elaborat un primer esborrany de proposta per De Vaney (1994a) (de World Systems), que defineix l'abast de la tasca i les fases d'avaluació següents: determinació dels requeriments, establiment dels criteris, desenvolupament del model de cost/benefici, eficàcia del procés d'avaluació i anàlisi dels resultats.

Un aspecte polèmic és el de la mesura en què s'han d'aplicar criteris de confidencialitat a la metainformació estadística, en relació amb la informació estadística associada. Hi ha, però, consens en el principi de que "tota la informació i la metainformació estadístiques hauria de ser, en principi, pública; les restriccions només haurien de ser aplicades quan es pugui deduir informació sobre persones individuals".

2.4.6. Mètodes formals de descripció de metainformació estadística

Christofer De Vaney (1994b) ha defensat l'aplicació de les tècniques *FM (Formal Methods)* al desenvolupament de components de metainformació estadística per al seu ús en el context de sistemes de metainformació estadística. El seu raonament és el següent:

En primer lloc, els objectes (o classes d'objectes) de metainformació estadística és probable que siguin utilitzats en la implementació de múltiples sistemes, en diferents entorns. L'ús de tècniques FM en la fase de disseny pot permetre garantir que la semàntica de l'ús és consistent a través de les implementacions.

En segon lloc, en la mesura que els objectes de metainformació estadística i les operacions associades són complexes, les notacions FM poden ser utilitzades com a base per a noves notacions específiques de la metainformació estadística.

En tercer lloc, les tècniques FM poden ser utilitzades com a base per a la definició dels estàndars de metainformació estadística i dels sistemes de metainformació estadística.

En la pràctica, les tècniques FM han estat efectivament aplicades en dos projectes específics relacionats amb la metainformació estadística, en el marc del Programa DOSES d'EUROSTAT: el llenguatge *SMILE, (Statistical Metainformation Language Environment)* desenvolupat, amb l'ajuda d'un sistema expert, per a representar objectes de metainformació estadística i mantenir una base de dades d'aquests objectes, i l'*EISI Domain Assistant*, (*EISI = Expert Interface to Statistical Information Systems*), una extensió de SMILE en el context d'una shell d'inferència per resoldre problemes estadístics (*Vegeu l'apartat 2.2.7*).

En el mateix informe citat, De Vaney donava notícia d'un nou projecte que reprenia el llenguatge SMILE, però basat en els objectes descrits al "Model genèric d'indicadors estadístics" d'Olenski. (1994)

2.5. La metainformació estadística i EUROSTAT

El grup de treball *Metadata Task Force* és tributari, en la definició de la seva feina, de l'activitat estadística d'EUROSTAT que, d'una banda, es concentra en macrodades i, de l'altra, en taules comparables al nivell dels països de la Unió Europea. Així, doncs, els problemes de coordinació i integració d'informació heterogènia, típics de la metainformació estadística, s'agreugen en aquest cas, on l'harmonització de definicions i nomenclatures es troba en el primer pla de les preocupacions, complicades per la necessitat del multilingüisme. (Byfuglien 1994)

L'organització de l'encontre *Statistical Metainformation Systems Workshop*, al 1993, per part d'EUROSTAT (1993), va representar una magnífica ocasió de debatre els problemes de la metainformació estadística en un context obert i tècnic. Ja hem esmentat també (apartat 2.2.7) el programa DOSES de sistemes experts aplicats a la informació estadística.

Tanmateix, als efectes institucionals a escala europea, és el projecte *DSIS (Distributed Statistical Information Systems)* d'EUROSTAT (1992) el que pot tenir efectes importants, en la mesura en que es proposa establir un sistema d'estadística distribuïda entre els països de la Unió. En l'estudi de viabilitat del *DSIS*, es defineix l'anomenat *common reference environment* com a base tècnica que garantiria l'accés a les diferents informacions nacionals sota uns conceptes comuns. Als nostres efectes, cal assenyalar que el *DSIS* reconeix la importància decisiva de la metainformació estadística al si d'aquest entorn comú, a la que fa jugar el paper d'esquema conceptual global, bàsic en la interfície d'usuari del futur sistema. El projecte no ha passat encara de la fase d'estudi, però en tot cas, aquest reconeixement de la metainformació estadística és significatiu. D'altra banda, ja hem esmentat el projecte *EDINOMEN*, independent però integrable amb DSIS, l'objectiu del qual és d'establir un sistema d'intercanvi electrònic de Classificacions i Nomenclatures. (Lebaube 1993)

3. OBSERVACIONS FINALS

3.1. L'estratègia de Sundgren

En una primera versió de les "*Guidelines*", Sundgren (1993b) desenvolupava una sèrie de criteris que configuren una estratègia recomanable per als instituts d'estadís-

tica per al desenvolupament d'un sistema de metainformació estadística. Tot i que aquest text ha desaparegut a la versió definitiva, hem volgut recuperar-lo pels elements pràctics que conté.

El text comença establint les següents premisses:

a) És irrealista pretendre l'immediat desenvolupament d'un sistema de metainformació estadística "complet" que satisfagui totes les necessitats relatives a la metainformació estadística, manifestades en les seves diferents formes. Això sembla evident. Però és que, a més, aquesta afirmació vé confirmada per certes dissortades experiències sofertes pels "impacients".

b) D'altra banda, és igualment arriscat planejar el desenvolupament d'un sistema de metainformació estadística, focalitzat sobre algunes necessitats importants, però sense parar atenció en les exigències i prerequisits relacionats amb elles. Per exemple, "mentre és natural i altament recomanable per a un institut d'estadística modern atorgar prioritat a les necessitats de metainformació estadística dels usuaris d'estadística, fóra estúpid negligir les interdependències entre la tasca subjacent i d'altres funcions d'infrastructura d'un sistema de metainformació estadística més complet". I Sundgren acaba comparant irònicament aquesta manera d'actuar amb la d'aquell que "critica les despeses realitzades en la producció rigorosa d'estadístiques d'atur quan tothom pot llegir les xifres d'atur als diaris cada dia".

Així, la hipòtesi de partida per a Sundgren és que

> "*Cal donar prioritat a les necessitats dels usuaris d'estadístiques, tot i tenint en compte que la metainformació que els usuaris necessiten ha de ser produïda amb criteris econòmics, i també de manera tal que la metainformació i les funcionalitats a obtenir hauran de ser de qualitat*"

En aquest context, cal considerar usuaris tant els interns de l'institut com els externs, i respecte de la metainformació estadística, cal incloure tant les necessitats del seu ús instrumental al si de les diferents funcions (producció, difusió) com les d'ús final dels usuaris externs.

Sundgren preconitza que un institut d'estadística que es proposi desenvolupar un sistema de metainformació estadística segueixi unes fases amb els objectius següents:

1) Definir l'arquitectura global horitzó per al conjunt dels sistemes de metainformació, tal com haurien de ser a llarg termini: el que en podem dir *infrastructura de metainformació.*

2) Implementar aquesta infrastructura de metainformació de manera progressiva, pas a pas, i començant amb aquells subsistemes que:

a) es necessitin amb més urgència per part dels usuaris, o bé

b) es necessitin per a un funcionament eficient dels subsistemes a).

Amb aquests criteris, respecte de les fases del cicle de vida dels sistemes d'informació estadística (d'aquells que es considerin prioritaris), la tasca s'haurà de centrar naturalment en la fase operativa (tant en la perspectiva d'ús com de producció), deixant de banda les fases de disseny i de gestió.

Òbviament, d'acord amb el que hem dit suara, la prioritat aniria als subsistemes més enfocats als usuaris i, en segon lloc, a aquells subsistemes orientats a la producció, però que siguin essencials per al bon funcionament d'aquells.

3.2. Conclusions

Sigui quina sigui la solució adoptada per a la implementació d'un SME, que no podrà ser d'altra banda sinó progressiva, ens permetem subratllar les següents condicions, que entenem com a necessàries, si bé no suficients, per garantir l'eficiència i el reeiximent de l'operació.

1) Que el disseny del SME a llarg termini sigui global, respongui a un esquema conceptual rigorós i es defineixin les interrelacions entre els subsistemes existents, o a crear progressivament, i també entre les diferents peces de metainformació estadística, a la llum d'aquell model global, que podem considerar, d'altra banda, simplement com un sistema virtual, que pot no arribar a implementar-se mai.
2) Que en la selecció, disseny i implementació dels subsistemes reals, components del SME, l'orientació a l'usuari sigui prioritària, tot i garantint, en la mesura necessària, l'eficàcia de la producció d'informació estadística.
3) L'esmentada orientació a l'usuari hauria de partir de la consideració a llarg termini del conjunt de les necessitats del Sistema Estadístic del país o territori de l'institut d'estadística, independentment de l'abast de les seves competències. Com a conseqüència, les necessitats de coordinació en cadascuna de les fases de les activitats dels diferents organismes productors d'estadística oficial del país formaran part del SME global.
4) Donades aquestes característiques del SME global, en el seu disseny haurien de participar activament els tècnics més propers i sensibles a les necessitats de l'ús de la informació i la metainformació estadístiques en les diferents fases i funcions de l'institut. D'altra banda, la responsabilitat del disseny hauria d'estar situada orgànicament sota la màxima autoritat de direcció.
5) Donades les incerteses de partida, pot ser prudent triar els components a implementar primer amb el criteri de relativa autonomia respecte dels sistemes

centrals, per tal de no interferir amb les tasques bàsiques d'un institut d'estadística. Això pot portar a experimentar sistemes d'edició electrònica, bé sigui en CD-ROM, bé en interfícies gràfiques d'usuari a bases de dades (eventualment en entorn WWW).

6) Paral·lelament, s'haurien d'implementar els subsistemes relacionats amb l'eficiència conjunta de la producció, sobretot els necessaris per a la coordinació tècnica interna. Caldrà fer-ho, però, dintre de l'esquema global i conceptual de referència, per tal de garantir la coherència lògica entre els diversos components.

7) Finalment, en l'estat actual de les coses, en el que el tema de la integració de la metainformació estadística en els sistemes d'informació estadística està a les beceroles, caldrà fer un seguiment atent dels desenvolupaments que es produeixin a nivell internacional, tant en el terreny dels estàndars com en el de les aplicacions pràctiques. És a dir, fer el que se'n diu una "vetlla tecnològica" atenta sobre el tema.

REFERÈNCIES

[1] **Aluja T, Balaguer J.M., Martí-Recober M., Nafria E.** (1993). "The EDA/System. A system for the production and analysis of summary statistics". *Statistical Metainformation Systems Workshop. Proceedings.* (EUROSTAT) Luxembourg, February 2–4, 1993.

[2] **Appel, Günther** (1993). (Statistiches Landesamt Berlin). "A metada driven statistical information system". *Statistical Metainformation Systems Workshop. Proceedings.*

[3] **Appel, Günther** (1994). (Statistiches Landesamt Berlin). "Management aspects of a statistical metainformation system (SMS)". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 6. Geneva : UN/ECE, 1994.

[4] **Byfuglien, Jan** (1994). (Eurostat). "Some plans and issues developing metadata at Eurostat". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 12 Geneva : UN/ECE, 1994.

[5] **Canals, Isidre** (1992). (Institut d'Estadística de Catalunya). "Los sistemas hipertexto e hipermedia en el contexto de los futuros libros electrónicos. Reflexión conceptual ilustrada con casos prácticos". *4as Jornadas de Información y Documentación en Ciencias de la Salud.* Bilbao, Junio 1992.

[6] **Darius P., Boucneau M., De Greef P., De Feber E., Froeschl K.** "Modelling metadata". *Statistical Metainformation Systems Workshop. Proceedings.*

[7] **De Vaney, Christofer** (1994a). (World Systems). "Evaluation criteria of statistical metainformation systems". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 20. Geneva : UN/ECE, 1994.

[8] **De Vaney, Christofer** (1994b). (World Systems). "Formal Methods and the design of statistical metadata". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 22. Geneva: UN/ECE, 1994.

[9] **De Vaney, Christofer** (1994c). (World Systems). "EMMA. Enhanced metainformation Management Architectura". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 14. Geneva: UN/ECE, 1994.

[10] **Dippo, Cathryn S.** (1994) (US Bureau of Labor Statistics). "A customer focus on metadata". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 11. Geneva: UN/ECE, 1994.

[11] **D'Angiolini, Giovanna** (1993). "Data and metadata: The need for a knowledge approach". *Statistical Metainformation Systems Workshop. Proceedings.*

[12] **EUROSTAT** (1992). *DSIS — Distributed Statistical Information System. Feasibility Study.* Version 2. Luxembourg : Logica, 1992 77p. Annexos.

[13] **EUROSTAT** (1993). *Statistical Metainformation Systems Workshop. Proceedings.* Luxembourg, February 2–4, 1993.

[14] **Gibert C., Martí-Recober M., Aluja T.** (1992). *EDA: An Expert System for Data Analysis.* Barcelona : UPC, Dept. d'Estadística i Investigació Operativa, Maig 1992. Document de recerca.

[15] **Gillman, Daniel W. & Appel, Martin V.** (1994). (U.S. Bureau of the Census). "Metadata Database Development at the Census Bureau". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 10. Geneva : UN/ECE, 1994.

[16] **Graves, Ronald** (1994). (Statistics Canada). "Information holdings within Statistics Canada. A framework". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 18. Geneva: UN/ECE, 1994.

[17] **Hand, D.J.** (1992). "Artificial Intelligence in statistics". *New Technologies and Techniques for Statistics. Proceedings.* EUROSTAT. Bonn, 24-26 Feb. 1992. p. 133–140.

[18] **INE** (1994). *S.I.D. Manual de referencia y uso. Documentación del subsistema de gestión.* Madrid : Instituto Nacional de estadística, 1994. (Doc. interno).

[19] **INSEE** (1990). *Les Dictionnaires de Données Statistiques (DDS). Présentation et Manuel de Référence.* Paris: INSEE, 1990

[20] **Kopp, Norbert** (1993). (Statistiches Landesamt Berlin). "The Thesaurus as a user interface for the metainformation system". *Statistical Metainformation Systems Workshop. Proceedings.*

[21] **Lamb, Joanne** (1993). "Metada in survey processing". *Statistical Metainformation Systems Workshop. Proceedings.*

[22] **Lazarou, Georges** (1993). (INSEE). "Les Dictionnaires de Données Statistiques (DDS)". *UN/ECE. METIS Meeting.* Dec. 1993. CRP 8. Geneva: UN/ECE, 1993.

[23] **Lebaube, Philippe** (1993). (EUROSTAT). "Exchange systems for classifications and standardization issues". *Statistical Metainformation Systems Workshop. Proceedings.*

[24] **Malmborg Eric** (1988). (Statistics Sweden). "Design of the user interface for an object- oriented statistical database". *Statistical and Scientific Database Management, Fourth International Working Conference SSDBM*, Rome, June1988. Available: Statistics Sweden R & D Report 1988:11. (cit. Malmborg 1993).

[25] **Malmborg E. & Lisagor L.** (1993). (Statistics Sweden). "Implementing a statistical metainformation system". *Statistical Metainformation Systems Workshop. Proceedings.*

[26] **Marina, Liana** (1994). "Using Statistical Data Dictionary (DDS) client/server towards a Statistical Metainformation System in NCS, Romania". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 23. Geneva: UN/ECE, 1994.

[27] **Maurer A., Kubler J.E.** (1992). "EDI and statistics. The need for a Generic Statistical Message". *New Technologies and Techniques for Statistics. Proceedings.* EUROSTAT. Bonn, 24-26 Feb. 1992. p. 258–267.

[28] **Nordbäck, Lars** (1990). (Statistics Sweden). *An illustrated booklet on the dissemination of statistics in the 1990s.* Stockholm : Statistics Sweden, 1990.

[29] **Nordbäck, Lars** (1992). (Statistics Sweden). "The PC-AXIS vision, the liberation of official statistics". *New Technologies and Techniques for Statistics. Proceedings.* EUROSTAT. Bonn, 24-26 Feb. 1992. p. 218–225.

[30] **Nordbotten, Svein** (1993). (Univ. Bergen). "Statistical meta-knowledge and -data". *Statistical Metainformation Systems Workshop. Proceedings.* Luxembourg, February 2-4, 1993.

[31] **Olenski, Jozef** (1991). (CSO Poland). "Reference Model for Standardization in Statistics". *UN/ECE. METIS Meeting.* Dec. 91 Working Paper 3. Geneva: UN/ECE, 1991.

[32] **Olenski Jozef** (1994). (CSO Poland). "Standardization of Statistical Indicators and Metadata ". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 17 Geneva: UN/ECE, 1994.

[33] **Prazenka Dusan** (1994). "Common Terminology of METIS". (2nd version). *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 16. Geneva : UN/ECE, 1994.

[34] **Sadreddini M.A., Bell D.A.** (1993). "Integrating distributed and heterogeneous statistical data bases". *Statistical Metainformation Systems Workshop. Proceedings.*

[35] **Saris Willem E., Prastacos Moulicos, Martí Recober Manuel** (1992). "CASIP - A complete automated system for information Processing in family budget research". *New Technologies and Techniques for Statistics. Proceedings.* EUROSTAT. Bonn, 24-26 Feb. 1992. p. 80–87.

[36] **Schuerhoff M. & Bethlehem J.G.** (1993) (Netherlands Central Bureau of Statistics). "Handling data and metadata in BLAISE III". *Statistical Metainformation Systems Workshop. Proceedings.*

[37] **Shoshani A., Wong H.K.T.** (1985). "Statistical and scientific database issues". *IEEE Transactions on Software Engineering*, SE 11 (10), 1985. (cit. D'Angiolini).

[38] **Silver, Mick** (1993) (Cardiff Business School). "The role of footnotes in a statistical metainformation system". *Statistical Metainformation Systems Workshop. Proceedings.*

[39] **Sundgren, Bo** (1973). (Statistics Sweden). "An infological approach to data bases". *Urval 7* (cit. Nordbotten).

[40] **Sundgren, Bo** (1990). (Statistics Sweden). *Conceptual modelling and related methods and tools for computed-aided design of information systems* Stockholm : SCB. R & D Report, 1990.

[41] **Sundgren, Bo** (1991a). (Statistics Sweden). *What metainformation should accompany statistical macrodata*? Stockholm : SCB. R & D Report, 1991:9.

[42] **Sundgren, Bo** (1991b). (Statistics Sweden). *Statistical metainformation and metainformation systems.* Stockholm : SCB. R & D Report, 1991:11.

[43] **Sundgren, Bo** (1992a). (Statistics Sweden). *Organizing the metainformation systems of a statistical office.* Stockholm : SCB. R & D Report, 1992:10.

[44] **Sundgren, Bo** (1992b). (Statistics Sweden). *Some properties of statistical information: Pragmatics, Semantics and Syntactics.* Stockholm: SCB. R & D Report, 1992:16.

[45] **Sundgren, Bo** (1993a). (Statistics Sweden). "Modelling Metainformation Systems". *Statistical Metainformation Systems. Workshop. Proceedings.* Luxembourg, February 2-4, 1993.

[46] **Sundgren, Bo** (1993b). (Statistics Sweden). *Guidelines on the design and implementation of statistical metainformation systems.* Stockholm : SCB. R & D Report, 1993:4.

[47] **Sundgren, Bo** (1994a). (Statistics Sweden). "Statistical metadata and metainformation systems". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 15. Geneva: UN/ECE, 1994.

[48] **Sundgren, Bo** (1994b). (Statistics Sweden). *Guidelines for the Modelling of Statistical Data and Metadata.* Stockholm : SCB. R & D Reports, 1994. (Reproduït íntegrament en aquest número de **Qüestiió**).

[49] **UN/ECE** (1990). *User's Guide to Metainformation Systems in Statistical Offices. (METIS).* Geneva: UN/ECE, 1990. (cit. De Vaney).

[50] **Villán, I.** (1991) (INE). "Metadata management at the Spanish National Institute of Statistics: SID project". *UN/ECE. METIS Meeting.* Dec. 91 Working Paper 11. Geneva: UN/ECE, 1991.

[51] **Walker D., Newman I., Mather P., Ruggles C, Medyckyj-Scott D.** (1993). "Federal network based metainformation and central metadata management. Some initial proposals related to the GENIE approach". *Statistical Metainformation Systems Workshop. Proceedings.*

[52] **Zeisset, P.T.** (1993). (US Bureau of the census). "Metainformation for summary statistics: The EXTRACT experience". *Statistical Metainformation Systems Workshop. Proceedings.*

ENGLISH SUMMARY:

INTRODUCTION TO STATISTICAL METAINFORMATION SYSTEMS

Isidre Canals Cabiró

This paper aims to provide an introduction to the concept and problems related to statistical metainformation, defined broadly as *information on statistical information.*

The first part ("Introduction to statistical metainformation systems") takes the perspective of an official statistical body faced with the problem of organizing statistical metainformation (metadata) in order to provide its users with all additional information needed to interpret and use statistical information (data). These complementary elements include the methods used for producing that statistical information, and also definitions, classifications, territorial codes etc.

It is stated that statistical metainformation (or metadata) can be analyzed from two points of view: from the user and the producer perspectives, which induces two types of statistical metainformation system: user-driven and production-driven. In fact, both internal and external users should be considered as potential users of a statistical metainformation system.

From the viewpoint of the use of statistical information the paper reviews the new means of disseminating information (electronic books on CD-ROM, online databases...) and raises questions on how to integrate statistical metainformation with the data. It is also stated that the user perspective on metainformation can be fruitfully used to design the statistical information system of a statistical body, thus creating a *user-oriented statistical information system.* The complexity of this task can be recognized if we take into account the fact that potential users are heterogeneous in their needs and that now, more than ever, they possess the skills and the technical means to manage statistical information by themselves.

From the viewpoint of the production of statistical information, this being the basic function of every statistical body, the statistical metainformation of interest to us will be all the additional information necessary to produce statistical information correctly and efficiently. So, all the information needed at every stage of the production chain (design, data collection, primary registers, estimation, tabulation) and also dissemination stage will be considered items of metainformation. Technical documents relating to statistical operations, and information related to different types of tools (computer systems, classifications, codes...) are typical examples of the elements of *production-driven statistical metainformation systems.*

Both viewpoints, user and producer, not only imply different statements of the problem, but they also convey different needs, according to different functions and tasks of the users (both internal and external), who require different items of information about the same subject. Nevertheless, both approaches may converge in the future, as they are related to complementary functions in a statistical body. Moreover, the organization of the information concerned will have a tendency to be conveyed to the users through a consistent global conceptual scheme.

This convergence will also be driven by the fact that users will be both, internal and external. Broadly speaking, external users will be concerned basically with the correct end use of data, while internal users will be concerned with the efficient production of data.

The paper proceeds to analyze and classify the different types of statistical metainformation, according to their nature, format, support and intended goal. The resulting strong heterogeneity of this information constitutes an added difficulty to the problem of organising a statistical metainformation system. A series of examples in different contexts of use of statistical metainformation serves the purpose of illustrating the variety of uses and users of statistical metainformation.

The first part ends with the following definition by METIS (the working group of the Conference of European Statisticians):

> *A Statistical Metainformation System (SMS) is the information system that uses and stores statistical metadata and produces statistical metainformation for purpose of supporting decision making concerning an information system. The object of the Statistical Metainformation System is the statistical information system.*

The second part of the paper ("The diversity of international experience") is devoted to a survey of some projects and experiments carried out by different statistical bodies in the field of organising statistical metainformation systems, and also of proposals made by international bodies.

The projects and experiments reviewed have been grouped, according to the method or approach adopted, as follows:

1) General production-oriented statistical metainformation systems. (The "Dictionnaires de Données Statistiques (DDS)" and information management approaches)

2) Specialized statistical metainformation systems.

3) General user-oriented statistical metainformation systems. (Integrated user interface, PC-DOK system and EMMA, Enhanced Meta- information Management Architecture)

4) Document systems and statistical metainformation systems.

5) Statistical metainformation as associated with sets of statistical information on disseminating digital supports. (Diskettes, CD-ROM, Public interactive databases and Internet World Wide Web servers)

6) Statistical metainformation system as a coordination tool

7) Statistical metainformation, integrated in an expert system

The second part also describes the workings of and the opinions issued by specialised working groups of international bodies. Work programs of METIS are described briefly, especially those relating to the methodology of design and implementation of statistical metainformation systems. The proposal written by Sundgren ("Statistics Sweden"), officially adopted by METIS as guidelines is reproduced in this volume of **Qüestiió**.

The third and last part of this paper repeats some recommendations made by Sundgren to the statistical bodies on their strategy for the development of statistical metainformation systems, and summarizes some general conclusions, emphasizing the need for a long-term global design of the system, even in the (advisable) case of its partial and progressive implementation.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 321-357, 1995

GUIDELINES FOR THE MODELLING OF STATISTICAL DATA AND METADATA

BO SUNDGREN*

The paper consists of five chapters. In chapter 1 a conceptual foundation is proposed, introducing the concepts of statistical metadata, statistical data, and statistical information systems. Chapter 2 discusses which subjects (users and producers of statistical data) and objects (software tools) have needs for statistical metadata, and for which purposes. In chapter 3 the contents of these metadata needs are analysed more systematically, and in more detail; particular emphasis is given to the metadata needs of users of statistical data. Chapter 4 investigates possible sources for the statistical metadata needed. Finally, chapter 5 discusses how the metadata infrastructure of a statistical service could best be organised.

Key words: Statistical metainformation, Statistical metadata, Metadata infrastructure, Statistical metainformation system.

1. CONCEPTUAL FOUNDATION

1.1. Statistical metadata —what are they, and why are they needed?

We shall base our definition of statistical metadata on the following short statements about metadata in general:

- Metadata are physical representations of metainformation (metaknowledge) —similarly as data are representations of information (knowledge).

*Bo Sundgren. Statistics Sweden. S-11581 STOCKHOLM

-Article rebut el novembre de 1994.

-Acceptat el febrer de 1995.

- Metadata inform about data —and about processes producing and using data.
- Metadata are data which are needed for proper production and usage of the data they inform about.
- Like data inform about objects in a "real-world" object system, metadata inform about information objects in an information system —metaobjects. Like "real-world" objects, metaobjects have properties and are related to each other by object relations.
- Like data are managed by information systems, metadata are managed by metainformation systems. Thus a metainformation system uses and produces metadata, informing about data, and it fulfills its tasks by means of functions like "metadata collection", "metadata processing", "metadata storage", and "metadata dissemination".
- A metainformation system may be active or passive. An active metainformation system is physically integrated with the information system containing the data that the metadata in the metainformation system informs about. A passive metainformation system contains only references to data, not the data themselves.

Note that the first three statements above may be regarded as three "projections" of a three-dimensional definition of metadata:

- a **syntactical projection** (representation-oriented)
- a **semantical projection** (contents-oriented)
- a **pragmatical projection** (purpose-oriented)

The fourth statement above can be regarded as a further explication of the semantics of metainformation, and the fifth and sixth statements can be regarded as further explications of the syntactical aspects of metainformation.

In the first statement "knowledge" was introduced as an alternative term for "information". In general, "knowledge" should be interpreted as a concept, which is both "wider" and "deeper" than "information". The concept of information often focuses on factual knowledge (as opposed to knowledge having the character of definitions, rules, laws, algorithms, etc). Like "knowledge" the term "information" implies interpretation by means of a human brain, but "knowledge" may imply further "digestion" or "understanding" as well.

In practice, one does not always maintain a strict distinction between "data", "information", and "knowledge". In fact it is often possible to see the same phenomenon alternatively as, say, an immediate aspect of information or as a mediate aspect of data —or vice versa. For many practitioners "data" seem to be more concrete and understandable than "information". Thus in a discussion with practitioners it

may sometimes be advisable to present concepts and arguments in terms of metadata (rather than metainformation). The advice will —to some extent—be followed in this paper.

Now we shall apply the general statements above to our main topic —the topic of statistical metadata and statistical metainformation systems.

Statistical metadata are data which are needed for proper production and usage of statistical data. They describe statistical data and —to some extent— processes and tools involved in the production and usage of statistical data. Expressed briefly, statistical metadata are data about statistical data.

A **statistical metainformation system** is a system, which uses and produces statististical metadata, informing about statistical data, and which fulfills its tasks by means of functions like "statistical metadata collection", "statistical metadata processing", "statistical metadata storage", and "statistical metadata dissemination".

Like other metainformation systems, a statistical metainformation system may be active or passive, as defined above. A user of an active metainformation system, who has identified some potentially interesting data, can immediately proceed to retrieve the data from the same system. Such a system is an integrated information/metainformation system. In contrast, a user of a passive metainformation system, who has identified some potentially interesting data, will have to retrieve these data from another system.

1.2. Statistical data

Statistical data are the primary objects of the descriptions provided by statistical metadata. Thus in order to understand the meaning and contents of statistical metadata, we must have some understanding of what statistical data are, and what it is about them that may have to be described.

Figure 1.1 attempts to highlight the essence of statistical data and of the processes resulting in statistical data.

Statistical data may be microdata or macrodata, defined as follows.

1.2.1. Microdata

Microdata, sometimes called **observation data** or **measurement data** are the result of observations or measurements of a set of **object characteristics** (states and events). An object characteristic can be formalized as an ordered pair

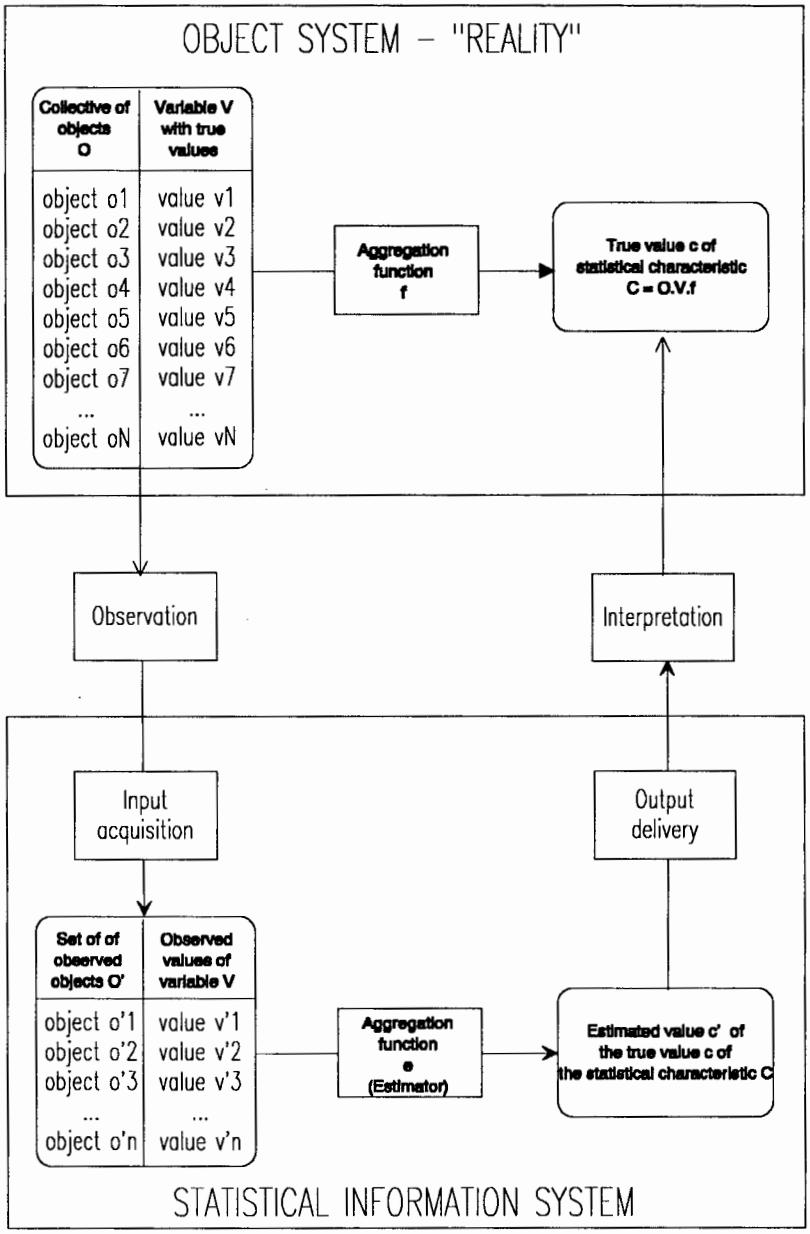


Figure 1.1
Illustration of some fundamental concepts in statistical information processing.

(1) $C_o =< O(t), V(t) >$ or, with dot notation, $C_o = O(t) \cdot V(t)$

where

(i) O is an object type;

(ii) V is a variable;

(iii) t is a time parameter.

Sometimes O will rather be a *vector* of object types, in which case V will be a **relation** or a variable that is based upon a relation, e g "quantity (of a commodity) exported (from one country to another country)".

$O(t)$ is a **population of objects** existing at time t, and $V(t)$ is the status at time t of a variable V, which is relevant for the objects in $O(t)$. In the most general case the time parameter t may not have the same value in $V(t)$ as in $O(t)$.

The basic building-blocks of information about observations and measurements of object characteristics are so-called **elementary messages** (e-messages) with the semantical structure

(2)

$m_o =< lo_i, p, t >$ or, with an alternative notation, $m_o = [o_i \cdot V(t) = v_j]$

where

(i) o_i is an object instance belonging to the object type O;

(ii) p is a property, typically expressed as a value a_j of a variable V;

(iii) t is an instance of time (point or interval) at/during which the object is supposed to have (had) the property p.

Alternatively o_i could be a vector of objects, p being a relation (like "married") or a $< V, vj >$ pair, where V is based upon a relation (like in the "export" example above).

In a typical interaction between a statistical information system and an input provider, the latter receives a set of questions, often hierarchically structured by respondent. The respondent is sometimes identical with (one of) the object(s) observed. The questions are accompanied by some metadata in the form of explanations, instructions, etc. In some systems additional metadata may be requested interactively by the respondent, if and when they are needed. When observation messages are returned to the statistical information system, they may be accompanied by other types of

metadata, informing about, say, some exceptional circumstances noted in connection with the observation process.

When the hierarchically structured sets of observation messages enter the statistical information system, they are often —sooner or later— transformed into **flat files** or **relational tables** in accordance with relatively well standardized procedures, supported by many commercial software products (cf form handling tools of relational database management systems). The accompanying metainformation should ideally be systematically taken care of by a parallel process, but this process has not yet reached any degree of standardization.

Observation registers, containing observed and/or derived microdata, are —beside collections of statistics/macrodata (cf next section)— the most important type of data output from statistical surveys. More and more competent users of statistics demand access to microdata, for their own analyses, in their own computer environments. Statistical services may respond to such demands by preparing files of **anonymized microdata**, for example so-called **public files**.

An external (re)user of an observation register may not be in a position where he or she has access to the staff, who once (maybe years earlier) produced the data. Thus the observation register will have to be accompanied by an appropriate set of metadata.

1.2.2. Macrodata: statistics

Macrodata, in daily talk simply referred to as "**statistics**", are the result of estimations of a **set of statistical characteristics** (statistical concepts). The estimations are made on the basis of a set of microdata, that is, a set of observations of a set of object characteristics. A statistical characteristic can be formalized as a triple

(3) $< O(t), V(t), f >$ or, with dot notation, $C_s = O(t) \cdot V(t) \cdot f$

where

(i) $O(t) \cdot V(t)$ is an object characteristic;

(ii) f is a statistical measure, that is, an aggregation function (count, sum, average, correlation, etc) summarizing the true values of $V(t)$ for the objects in $O(t)$.

In the most general case, V is a vector of variables, each one of which may be qualified by a separate time parameter, that is, t is a vector, too.

The basic building-blocks of the aggregated statistical information contained in statistical tables are **statistical e-messages** with the semantical structure

(4) $$m_s = \left[e\left(O(t_1) \cdot V(t_2) \cdot f\right) = a' \right]$$

where

(i) $O(t_1)$ is a population of objects existing at/during time t_1;

(ii) $V(t_2)$ is the status of a (vector of) variable(s) at/during time(s) t_2;

(iii) f is a statistical measure;

(iv) e is an **estimator**, that is, a function providing estimates of the true values of a statistical characteristic C_s on the basis of observed values of one or more object characteristics.

Statistical macrodata are organized in certain typical structures. For example, statistics users are often interested to obtain estimated values of "the same" statistical characteristic for

- a series of time periods (rather than a signle one) —"**time series data**"; and/or
- a sturctured set of object populations (rather than a single one) —"**cross-sectional data**".

The following format is general enough to cover most structures of statistical meta-information that are demanded by statistics users:

$$O(t_a)(\text{with } p_a)\,(\text{by } V_g(t_g)) \cdot V_b(t_b) \cdot f \tag{5}$$

where

(i) $O(t_a)$ is a (series of) population(s) of objects existing at/during t_a, which, in the case of time series data, is a parameter varying over a certain range of times;

(ii) p_a is a property, the **alfa property**, selecting a subset of $O(t_a)$;

(iii) $V_g(t_g)$ is a vector of variables, the **gamma variables**, which are cross-classifying the population(s) $O(t_a)$, and t_g is a vector of time parameters corresponding to the vector of variables; in the case of time series data, each time parameter will vary over a range of values matching the range of values of the time parameter t_a;

(iv) $V(t_b)$ is a vector of variables, the **beta variables**, the summarized values of which are estimated, and t_b is a vector of time parameters corresponding to the vector of variables; in the case of time series data, each time parameter will vary over a range of values matching the range of values of the time parameter t_a;

(v) f is an aggregation function.

The general structure (1.5) for statistical macrodata is sometimes referred to as a **box structure** or an **alfa-beta-gamma-tau structure**, where "alfa" refers to the selection property, "beta" to the summarized variables, "gamma" to the crossclassifying variables, and "tau" to the time parameters, as just discussed. This structure is very useful for systematical analyses and descriptions of statistical macrodata. Figure 1.2 shows an example of an alfa-beta-gamma-tau analysis of the statistical macrodata published in a statistical publication.

Traditional forms of macrodata, like statistical tables, are typically accompanied by meta-data in the form of headings, column and row labels, footnotes, comments, etc. Today electronical equivalents of statistical tables are at least equally important, and such outputs are often the result of user-initiated interactions, which involve the processing of metadata provided alternately by the user (search questions etc) and by the statistical information system.

The GESMES format is a proposed standard for representation of statistical macrodata and accompanying metadata. "GESMES" stands for "Generic Statistical Message", and the standard proposal is developed by the UN/EDIFACT Message Development Group 6.1.

1.3. Statistical information systems

Beside the statistical data themselves, the **processes** of the information systems associated with the statistical data are important description objects of statistical metadata. As a basis for later discussion of metadata describing these processes, we shall here give a brief analysis of the concept of an information system in general, and of a statistical information system in particular.

With a very general formulation an **information system** has the purpose to help its users

- to establish **mind models** of a certain **object system**, a certain "piece of reality"; thereby helping them
- to *understand* the object system; and
- to *plan, implement, monitor, and evaluate actions* vis-à-vis the object system.

A **statistical information system** fulfills these tasks by

- providing **statistical information**, that is, information about collectives of objects (rather than individual objects) in the object system;
- supporting so-called **directive actions**, like general-level planning, decision-making, and evaluation.

ALFA COMPONENTS	GAMMA COMPONENTS	BETA COMPONENTS	TAU COMPONENTS
Table 1 (page 1): Estimated resident population ('000).			
Persons resident in Australia at a certain point of time; subset of "persons"	State of residence	Count/1000	Yearly: 1985-06-30--1990-06-30 Quarterly: 1989-09-30--1990-12-31
Table 2 (page 1): Components of resident population growth, year ended 30 June 1990.			
Person events during a certain time interval, causing an increase or decrease of the number of residents in an Australian state; subset of "person events"	1. State of person event. 2. Event classification: birth/death, migration (overseas, interstate).	1. Population growth = sum of population growth contribution caused by event (+1 or -1). 2. Rate of growth = (1)/(number of resident persons at the ... of the time interval; from table 1).	The year 1989-07-01--1990-06-30.
Table 3 (page 1): Mean resident population ('000).			
Persons resident in Australia some time during a certain time interval; subset of "persons"	State of residence.	Mean resident population ('000) computed on the basis of counts for ... successive time periods according to the formula ...	One year periods: 1985, 1986, ..., 1990. Two year periods: 1984-85, 1985-86, ..., 1989-1990.
Table 4 (page 2): Live births registered.			
Live births registered during a certain time period; subset of "person events"	State of registration.	Count	Quarter years ending 1989-09-30--1990-12-31.
Table 5 (page 2): Deaths registered.			
Deaths registered during a certain time period; subset of "person events"	State of registration.	Count	Quarter years ending 1989-09-30--1990-12-31.
Table 6 (page 2): Marriages registered.			
Marriages registered during a certain time period; subset of "person events"	State of registration.	Count	Quarter years ending 1989-09-30--1990-12-d31.
Table 7 (page 2): Divorces granted.			
Divorces registered during a certain time period; subset of "person events"	State of registration.	Count	Years ending 1985-12-31--1990-12-31.

Figure 1.2

Part of an alfa-beta-gamma-tau analysis of the tables in "Monthly Summary of Statistics Australia". Cf Sundgren (1991d).

A statistical information system accomplishes its tasks by performing three major functions:

(F1) an **input acquisition function**, which directly and/or indirectly observes (measures) certain object system characteristics, and which prepares and stores the observation data obtained as microdata in an observation register;

(F2) an **aggregation function**, which transforms the microdata produced by the input acquisition function into macrodata, or "statistics", which are estimated values of statistical characteristics;

(F3) an **output delivery function**, which makes macrodata (statistics) available to the users, and which assists the users to interpret and analyze the data further.

Figure 1.3 illustrates a break-down of the three major functions into more concrete subfunctions and tasks. In a traditional survey-processing system the functions, subfunctions, and tasks have been carried out more or less serially, in the order indicated by figure 1.3: from top to bottom, and from left to right.

Modern technology permits a much more flexible organization of the processes for producing and disseminating statistics. The production system in figure 1.4 is assumed to be:

- **database-oriented:** the microdata and macrodata, which are stored and processed, are communicated within and between the functions and subfunctions via a database;
- **self-describing:** the microdata and macrodata are described by means of accompanying **metadata**, which are stored in the database, and which are consistently transformed, whenever the described data are transformed.

The architecture visualized in figure 1.4 covers many different types of statistical information systems: survey processing systems, register management systems, user-driven retrieval systems, to mention the most important ones.

A **survey processing system** focuses on a data collection process, resulting in a collection of microdata, which are aggregated into estimated values of certain statistical characteristics.

A **user-driven retrieval system** focuses on the needs of a particular category of statistics users, and aims at making available macrodata and microdata from different surveys (and other sources), which may be relevant for the particular category of users.

Register management is an important auxiliary process for statistics production. There are two kinds of registers, which are particularly important for statistical

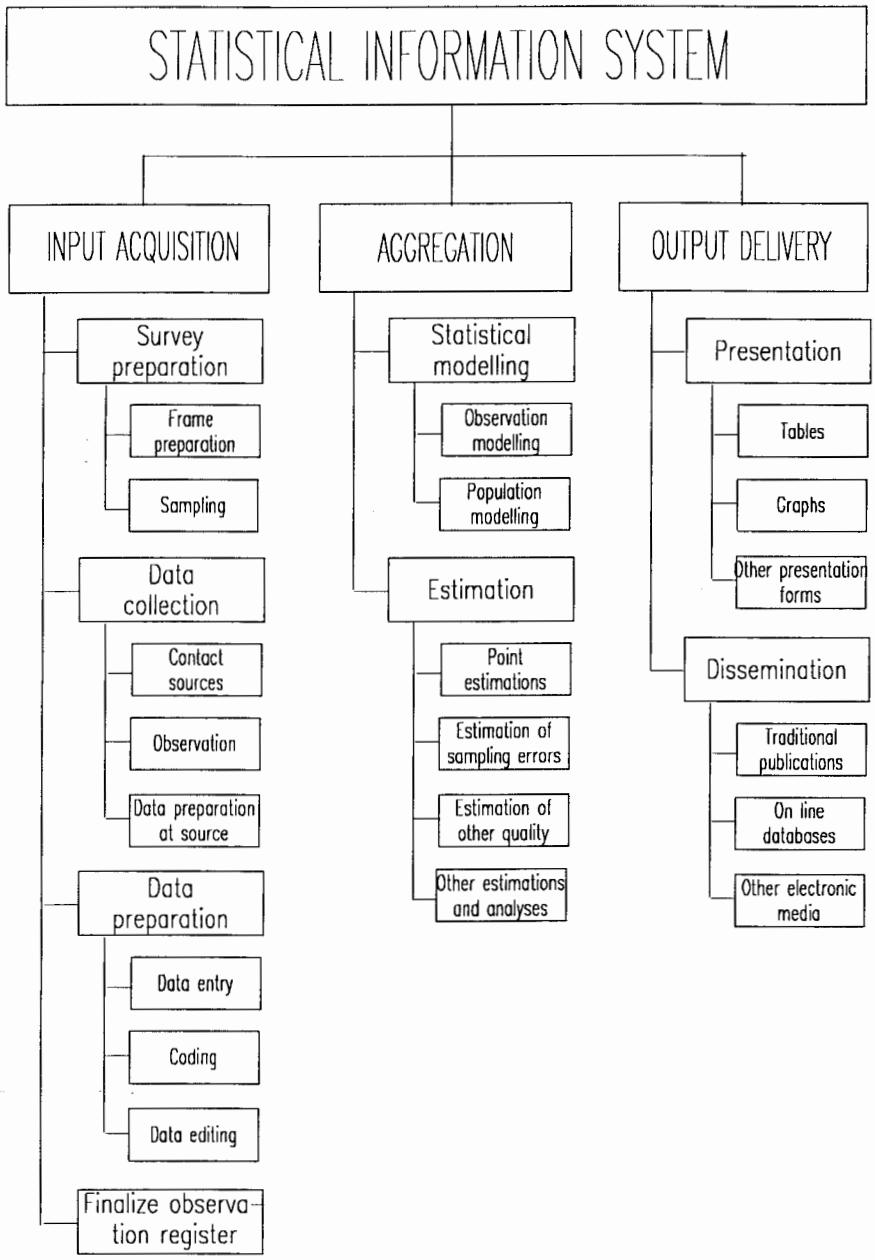


Figure 1.3
A functionally oriented model of a statistical information system.

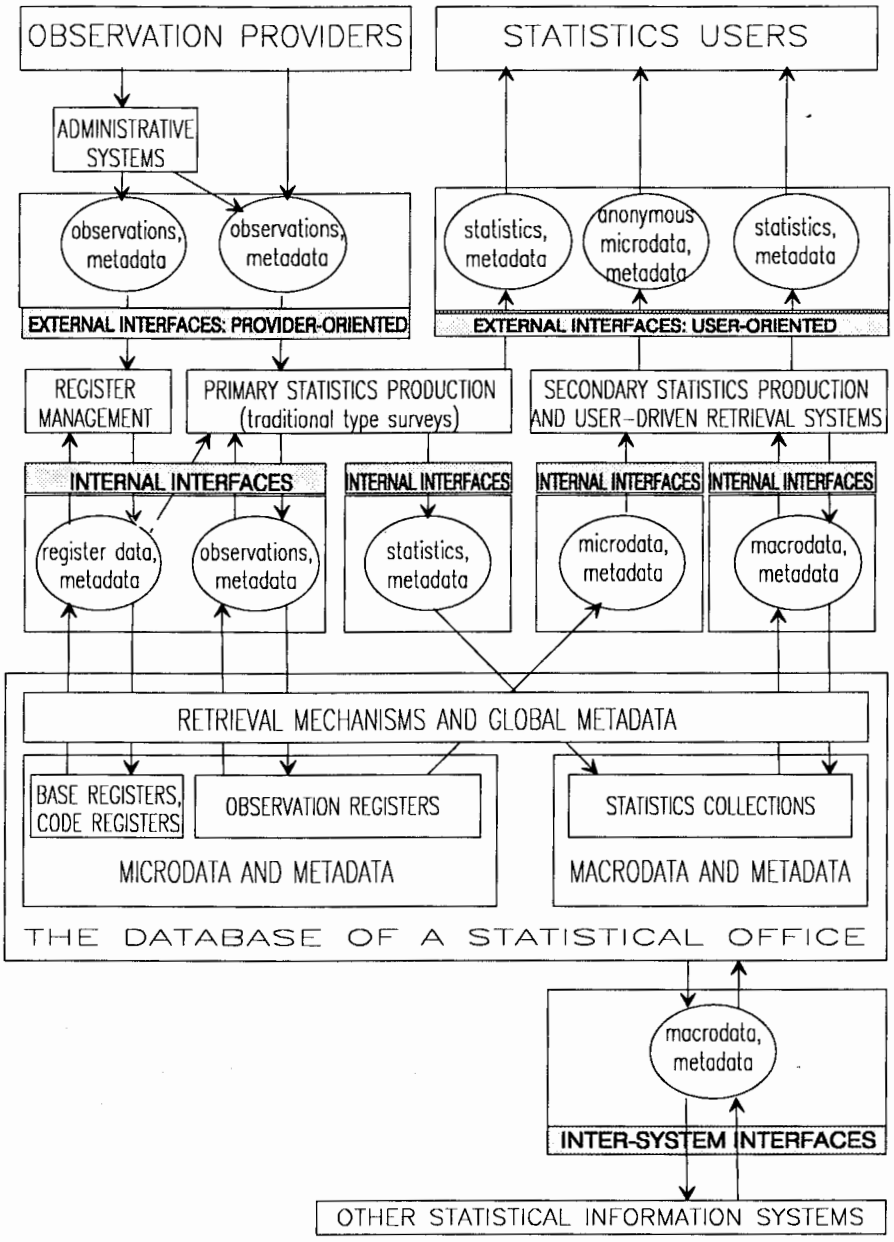


Figure 1.4
A model of a self-describing database-oriented statistical information system.

information systems: base registers and code registers. A **base register** establishes and maintains an authorized list of the objects belonging to a certain population. A **code register** establishes and maintains an authorized list of the values belonging to the value set of a certain variable or classification.

A complex statistical information system may contain many statistical surveys, retrieval systems, and registers as subsystems. "The statistical information system of a country" is an example of such a complex statistical information system.

2. WHO NEEDS STATISTICAL METADATA —AND FOR WHAT PURPOSES?

Users and producers of statistical data are obvious users of statistical metadata. In this chapter we shall identify the major purposes for which people belonging to these two categories actually need statistical metadata. Then, in the next chapter, we shall make a more detailed analysis of which kinds of the metadata that they need. In addition we shall treat separately the needs for statistical metadata that are associated with the software tools that are used in the production and usage of statistical data. Naturally these needs for statistical metadata can directly or indirectly be derived from the needs of users and producers of statistical data. However, software tools have such a special and important role that we find it appropriate to pay special attention to these needs.

2.1. Users of statistical data

Schematically the usage process proceeds as follows. First, a would-be user with some kind of question or problem is looking for statistical data of potential relevance for his/her problem. Second, the user identifies some statistical data of potential interest and decides to retrieve these data. Third, the user analyses and interprets the statistical data thus retrieved, and then possibly reiterates parts of the search, retrieval, and analysis procedure again.

In each one of the steps of the outlined usage procedure the user of statistical data will need some metadata. The width and depth of the statistical metadata needed will among other things depend on the pre-knowledge of the user. Different categories of users have different prerequisites and requirements.

Figure 2.1 gives an overview of some important categories of users and usages of statistical data and statistical information systems.

WHO?	WHY?
"Government"	Plan, monitor, and evaluate actions
"Companies"	Business decisions
"Organisations"	Negotiations, lobbying
"Researchers"	Analyse, understand, explain real-world phenomena
"General public"	Participate in democratical processes
"Politicians"	
"Journalists"	

Figure 2.1
Users and usages of statistical data and statistical information systems.

Different users may have very different needs, and as a consequence a modern statistical service will have to offer a wide range of different products and services. Figures 2.2a and 2.2b illustrate two ways of analysing this situation. Figure 2.2a shows how the characteristics of a number of "typical" user categories can be described and analyzed, and figure 2.2b shows a way of structuring the range of products and services provided by a statistical service.

2.2. Producers of statistical data

"Production of statistical data" can be given a broad or a narrow interpretation. In a narrow sense the production process includes the operative production steps as outlined in figure 1.3 earlier in this paper. In a broad sense "production" covers the whole life cycle of a statistical survey or a statistical information system, including design, implementation, operation, monitoring, maintenance, and evaluation.

Here we shall use the broader interpretation of "production", and "producers of statistical data" will include

- **designers** of statistical surveys and statistical information systems: subject matter statisticians, statistical methodologists, and information system specialists;
- **input data providers:** respondents, contact persons etc;
- **production statisticians** (here "production" is given a narrower interpretation).

USER CATEGORY BY CHARACTERISTIC	"Ministry of finance"	"Researcher/ scientist"	"Analyst - public sector"	"Analyst - private sector"	"Actor on the finance market"	"International organization"	"Journalist"	"Politician"	"Teacher/student in school"	"Interested citizen"
Competence: · subject matter · statistical · EDP										
Knowledge about relevant data sources: · broad · deep										
Quality requirements: · contents · accuracy · availability										
Needs for search systems, documentation, and metainformation										
Resources: · hardware · software · expertise · money · "trading objects"										

Figure 2.2a

A scheme for analyzing the profiles of different categories of statistics users.

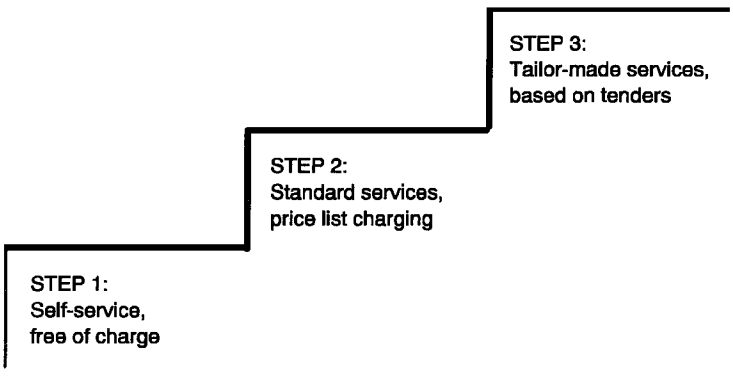


Figure 2.2b
"The stairs of services".

All these categories of producers of statistical data have their typical metadata needs. For example, a designer of a statistical survey will need to know about user needs, how similar surveys have been designed in the past and/or by other statistical services. A data provider to a statistical survey will be interested to know about the purposes of the survey and about the costs and benefits of his/her participation. A production statistician will need check-lists and other production system documentation in order to remember how to carry out the production steps properly, and in order to be able to train new staff members whenever necessary. An auditor evaluating a statistical information system will need metadata concerning the functioning of the system, including feed-back information from the users.

2.3. Software tools

A software tool needs certain metadata for its proper functioning. For example, it needs a formal description of the data that it is requested to process, that is, something like a file and record description. It also needs textual metadata in order to be able to present output data and communicate properly with the software tool users.

Thus a software tool needs metadata. On the other hand its users need metadata informing about the software tool as such. For example, a user of a software tool for some kind of statistical analysis will need information about the tool in order to be able to handle the tool properly, and in order to be able to interpret the results from the analysis in a responsible way.

3. WHAT STATISTICAL METADATA ARE NEEDED?

In this chapter we shall analyse which statistical metadata are needed by users and producers of statistical data and by software tools. We shall put particular emphasis on the metadata needed by users of statistical data.

3.1. Metadata needed by users of statistical data

3.1.1. Declarative metadata concerning specific sets of statistical data

In order to judge the usefulness of some specific statistical data (macrodata and/or microdata) for his/her purpose, a potential user needs to know something about

- the **contents** (meaning) of the statistical data, making it possible for the user to judge the **relevance** of the data with regard to his/her question or problem;
- the **accuracy** (precision, reliability) of the statistical data, that is, how well the measurements/estimations actually measure/estimate what was intended (by the designers) to be measured/estimated;
- the **availability** of the statistical data, that is, how the user can get access to them.

These three aspects of statistical data are often regarded as three major dimensions of the **quality of statistical data**, and a description of the contents, accuracy, and availability of a set of statistical data may be termed a **quality declaration** of the data.

The term "quality" has actually two basically different meanings. According to one interpretation, "quality" is identical with "**good quality**". This interpretation assumes that there is a general agreement on what "good quality" means in terms of **properties** of the entity whose quality is measured. For statistical data it is not easy to establish *absolute* quality criteria. The same statistical data (based on certain definitions and measurement procedures) may be quite adequate for one usage and completely inadequate for another, depending on how relevant the definitions and measurement procedures are for the respective purposes. The most fundamental quality requirement on statistical data is that the quality-relevant properties of the statistical data should be *known*, that is, the statistics should have **known quality**, and this quality should be well documented in some kind of quality declaration. If (and only if) these conditions are satisfied, a potential user of certain statistical data can (independently of the producer) asses the quality of the data *relative* to his/her particular needs.

Recalling figure 1.1 we may identify three conceptual levels, the relations between which are fundamental for understanding the quality concept that we are discussing here (cf figure 3.1):

(L1) the level of a **statistical characteristic of interest to a user**, that is, the "ideal" statistical characteristic to be estimated, if there were no practical restrictions or contradictory needs of different users;

(L2) the level of the **statistical characteristic decided to be estimated**, after practical restrictions and contradictory needs have been taken into account by the designers;

(L3) the level of the **actually achieved estimated value of the statistical characteristic**, as produced by the implemented production system, affected by different kinds of errors and uncertainties.

The discrepancy between levels L1 and L2 may be called the **relevance discrepancy.** It can be finally judged only by the user. What the producer can do in order to make it easier for the user to reach his/her judgement is to provide certain contents-oriented metadata as will be further discussed below.

The discrepancy between levels L2 and L3 may be called the **accuracy discrepancy.** It can be described by the producer of the statistical data.

Note. The conceptual discussion that we have just carried out for statistical characteristics (and thus for macrodata) can be carried out analogously for object characteristics (and thus for microdata); cf section 1.2.

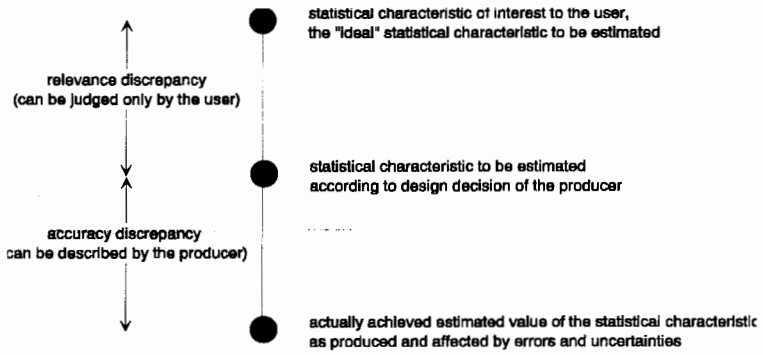


Figure 3.1
The quality of statistical data from a user's point of view.

The **quality declaration templet** in figure 3.2 provides a possible structuring of the metadata needed by a user who wants to judge the potential usefulness of a certain set of statistical data for a certain question or problem. With some minor modifications the templet can be used for microdata as well as for macrodata.

3.1.2. Process-oriented metadata concerning specific sets of statistical data

Declarative metadata are not always sufficient to give a user of a set of statistical data an adequate understanding of the data. Sometimes it may be necessary for the user to get more detailed information about how the production processes behind the statistical data were actually carried out.

The documentation templet in figure 3.3 illustrates what kinds of process-oriented metadata that a user of a set of statistical microdata (a so-called observation register) could need, and how these metadata could be organised.

QUALITY DECLARATION OF STATISTICAL DATA

0 Administrative information	**1 Contents**
0.0 Documentation templet 0.1 Source of the data (survey etc) 0.2 Name and identity of the data set 0.3 Responsible organisation/person	1.1 Overviews 1.1.1 Verbal description 1.1.2 Object graph 1.1.3 Structured lists of characteristics 1.2 Measurement instruments 1.2.1 Questionnaires 1.2.2 Other instruments 1.3 Object aspects 1.3.1 Object types 1.3.2 Populations 1.3.3 Samples (if applicable) 1.3.4 Domains of interest (if applicable) 1.4 Property aspects 1.4.1 Variables (by population) 1.4.2 Statistical measures (if applicable) 1.5 Time aspects 1.5.1 Reference time(s) 1.5.2 Frequency (if applicable) 1.5.3 Production time 1.5.4 Comparability over time (if applicable) 1.6 Comparability with other data
2 Accuracy	**3 Availability**
2.1 Overall accuracy 2.2 Error sources (sources of uncertainty) 2.2.1 Coverage 2.2.2 Sampling 2.2.3 Measurement 2.2.4 Non-response 2.2.5 Processing 2.2.6 Model assumptions	3.1 Physical storage 3.2 Presentation and dissemination 3.2.1 Traditional publications 3.2.2 On line databases 3.2.3 Other media 3.3 Further documentation and contact points 3.4 Related data sets

Figure 3.2
Quality declaration templet for statistical data (microdata or macrodata).

3.1.3. Global metadata and general knowledge

A user looking for statistical data which are possibly relevant for his/her question or problem will need **global metadata** spanning over many statistical surveys in order to be able to identify and locate the possibly relevant statistical data. Examples of such global metadata are:

- **descriptions** —more or less formalized— of available statistics and observation registers;
- well-structured and informative **tables of contents** —as global as possible— specifying available statistics and observation registers and giving references to the more detailed descriptions;
- **indexes** to the tables of contents
- **thesauri** for supporting the process of specifying search questions by providing broader terms, narrower terms, and related terms

When analysing statistical data resulting from a retrieval operation, a user will need metadata belonging to the category of **general knowledge** like **handbooks** describing different methods of statistical analysis.

3.2. Metadata needed by producers of statistical data

A producer of statistical data needs

- the same metadata as the users (in order to be able to help them)
- detailed information about the production process (in order to remember how to carry out the production steps correctly, and in order to train new staff members)
- feed-back information from the users and input data providers (respondents etc) in order to tune and improve the production processes
- general knowledge and global metadata (in order to design/redesign production processes)

3.2.1. Survey-specific metadata concerning the production process

Figure 3.4 suggests a templet for organising the survey-specific metadata needed by those who are responsible for monitoring, operating, maintaining, and redesigning a survey production system.

OBSERVATION REGISTER DOCUMENTATION

0 Administrative information 0.0 Documentation templet 0.1 Survey name and identification, organisation and persons responsible 0.2 Documentation modules and subsystems 0.3 Archived data sets and published statistics 0.4 References to other relevant documentation	**1 Survey contents** 1.1 Domain of interest and target domain, verbal description 1.2 Target domain, formal description 1.2.1 Target objects, description and object graph 1.2.2 Target populations 1.2.3 Target variables 1.3 Survey outputs 1.3.1 Structured overview of the tabulation plan 1.3.2 Publications in printed form 1.3.3 Electronical distribution 1.3.4 Database storage
2 Survey plan 2.1 Frame procedure and observation objects 2.1.1 Overview 2.1.2 Frame and its links to objects 2.1.3 Frame production 2.1.4 Overcoverage and undercoverage 2.2 Sampling procedure (if applicable) 2.3 Data collection procedure 2.3.1 Observation objects, description and object graph 2.3.2 Data sources, including contact procedures 2.3.3 Observation variables and measurement instruments 2.3.4 Interruptions (including actions at overcoverage) 2.3.5 Non-response actions 2.4 Planned data preparation (coding, data entry, editing and correction) 2.5 Planned observation register 2.5.1 Overview 2.5.2 Object types, including derived object types 2.5.3 Object graph 2.5.4 Object/variable-matrixes, including derived variables 2.5.5 Data set descriptions 2.5.6 Derivation procedures (in complicated cases)	**3 Completed data collection** 3.1 Frame production 3.2 Sampling 3.3 Data collection 3.3.1 Communication with the data providers 3.3.2 Measurements, experiences of instruments 3.3.3 Interruptions/overcoverage, actions taken 3.3.4 Non-response, causes and actions taken 3.3.5 Editing and correction at data collection time 3.4 Data preparation (coding, data entry, editing and correction) 3.5 Production of final observation register 3.5.1 Treatment of interruption/overcoverage objects 3.5.2 Treatment of non-response objects 3.5.3 Treatment of partial non-response 3.5.4 Frequency counts of overcoverage, responses, non-responses etc 3.5.5 Completed derivations of derived objects and variables
4 Statistical processing and presentation 4.1 Observation models 4.1.1 Sampling 4.1.2 Non-response 4.1.3 Measurement/observation 4.1.4 Frame coverage 4.1.5 Total model 4.2 Population models 4.3 Computation formulae for estimations 4.3.1 Point estimations 4.3.2 Estimations of sampling errors (variance estimations) 4.3.3 Estimation/judgment of other quality characteristics 4.4 Analyses 4.5 Presentation and dissemination procedures	**5 -**
6 Log-book	

Figure 3.3
Documentation templet for an observation register and the survey behind.

DOCUMENTATION TEMPLET FOR A STATISTICAL SURVEY

0 Administrative information	**1 Survey contents**
0.0 Documentation templet 0.1 Survey name and identification, organisation and persons responsible 0.2 Documentation modules and subsystems 0.3 Archived data sets and published statistics 0.4 References to other relevant documentation	1.1 Domain of interest and target domain, verbal description 1.2 Target domain, formal description 1.2.1 Target objects, description and object graph 1.2.2 Target populations 1.2.3 Target variables 1.3 Survey outputs 1.3.1 Structured overview of the tabulation plan 1.3.2 Publications in printed form 1.3.3 Electronical distribution 1.3.4 Database storage
2 Survey plan 2.1 Frame procedure and observation objects 2.1.1 Overview 2.1.2 Frame and its links to objects 2.1.3 Frame production 2.1.4 Overcoverage and undercoverage 2.2 Sampling procedure (if applicable) 2.3 Data collection procedure 2.3.1 Observation objects, description and object graph 2.3.2 Data sources, including contact procedures 2.3.3 Observation variables and measurement instruments 2.3.4 Interruptions (including actions at overcoverage) 2.3.5 Non-response actions 2.4 Planned data preparation (coding, data entry, editing and correction) 2.5 Planned observation register 2.5.1 Overview 2.5.2 Object types, including derived object types 2.5.3 Object graph 2.5.4 Object/variable-matrixes, including derived variables 2.5.5 Data set descriptions 2.5.6 Derivation procedures (in complicated cases)	**3 -**
4 Statistical processing and presentation 4.1 Observation models 4.1.1 Sampling 4.1.2 Non-response 4.1.3 Measurement/observation 4.1.4 Frame coverage 4.1.5 Total model 4.2 Population models 4.3 Computation formulae for estimations 4.3.1 Point estimations 4.3.2 Estimations of sampling errors (variance estimations) 4.3.3 Estimation/judgment of other quality characteristics 4.4 Analyses 4.5 Presentation and dissemination procedures	**5 Data processing system** 5.0 System overview 5.0.1 Verbal description 5.0.2 System flow 5.1* Subsystem description 5.1.1 Overview 5.1.1.1 Verbal description 5.1.1.2 System flow 5.1.2 Component descriptions 5.1.2.1 Data sets 5.1.2.2 Processes 5.1.2.3 Other components
6 Log-book	

Figure 3.4
Documentation templet for a statistical survey and its production system.

3.2.2. Global metadata and general knowledge

Producers of statistical data need global metadata and metadata of "general knowledge" type when they design, maintain, and evaluate surveys and production system. For example, they need information about how similar tasks are solved in other surveys and production system. Such information may be available in a global metadatabase. They also need general methodological knowledge, which may be available in handbooks, encyclopedia, and expert systems.

3.2.3. Feed-back information from the users

An important type of metadata for producers is feed-back from the users indicating user satisfaction (or dissatisfaction) with the statistical data and services provided. This type of feed-back metadata should be broken down by user categories and service types (cf chapter 2, in particular figure 2.2 a and b). Requests for statistical data, which cannot be satisfied by the present survey design, may be logged and used as a basis for future improvements of the survey.

3.3. Metadata related to software tools

This category includes

- metadata *needed by* software tools
- metadata *about* software tools, needed by the users of the tools

Both kinds of metadata needs are directly or indirectly related to user/producer needs.

3.3.1. Factual metadata

Semi-structured, ranging from highly formalized metadata concerning representation formats (file descriptions, record layouts, code descriptions etc) to text labels (names etc) and free-text descriptions. Important global metadata of this kind: information about classifications.

3.3.2. Algorithmic metadata

Algortithmic metadata include

- the algorithms as such behind statistical procedures, including procedures for statistical analysis;
- descriptions of the algorithms.

4. SOURCES OF STATISTICAL METADATA

A good principle for data management in general, and metadata management in particular, is to capture data/metadata as early as possible and only once. Ideally data/metadata should be captured with as little human effort as possible, and as a direct result of events that imply the "birth" of the data/metadata. For example, a decision to design a statistical survey in a certain way may imply the birth of (some of) the metadata describing the definition of a statistical characteristic.

In the previous chapter we identified a large number of metadata items, which are needed to satisfy important metadata needs of users and producers of statistical data. A systematical analysis of when these metadata items could best be captured could start from an analysis of the typical life histories of statistical characteristics, statistical surveys, and statistical information systems. Major phases of these life histories are

- design processes
- operation and evaluation processes
- maintenance and redesign processes

4.1. Design processes

We have already discussed how designers of statistical surveys and statistical information systems can make use of metadata originating from other surveys and information systems. However, design processes do not only make use of metadata; they also (as a side-effect) give rise to new metadata, which can be most useful for many different purposes, if they are properly taken care of.

A design process produces a number of design decisions, and if these decisions are properly documented —as they should be anyhow— the decision documents (which will typically be computer-stored today) will be excellent sources of statistical metadata.

Schematically, a design process for a statistical survey will consist of a series of design steps, corresponding to the boxes in figure 1.3 earlier in this paper. Each one of these design steps results in a number of design decisions, and these design decisions consist of metadata concerning the planned survey, which have their natural "slots" in the structures of documentation templets such as those proposed in figures 3.2, 3.3, and 3.4 above.

The natural way of taking care of the metadata flow from a design process may be to incrementally update a growing production system documentation of the survey, starting from an empty documentation templet and ending with a more or less completely filled one (cf figure 3.4).

4.2. Operation and evaluation processes

Similarly as the design process for a statistical survey produces a natural flow of metadata that can be used for updating an incrementally growing production system documentation, the operation of the survey (cf figure 1.3 again) can be so organised as to produce a flow of metadata incrementally updating documentations of the observation registers and statistics that are produced by the survey.

For example, an observation register documentation based on the templet in figure 3.3 can be incrementally filled as follows. First, the slots of chapters 0, 1, and 2 are filled by copying the corresponding metadata from the production system documentation completed during the design process. Then, chapter 3 is filled, slot by slot, as the different production steps are carried out. Some of the original design decisions may have to be modified already during the first operation of the survey, and then some slots of the already completed parts of the production system documentation will have to be updated.

A survey production system should be designed to give some automatical feedback in order to give the producers a basis for evaluating the information contents and functionality of the system. Another way to evaluate a survey production system is to compare it with similar systems elsewhere as well as with more or less established "current best methods". In order to facilitate such comparisons, it should be a routine task for every survey production to contribute some metadata to a global metadatabase as a side-effect of its own operation and evaluation processes.

4.3. Maintenance and redesign processes

An evaluation process may lead to radical redesign or to less radical maintenance. Maintenance also includes actions which typically have to be undertaken in order to adapt a survey production system to changes in the environment, e.g. changes in systems providing input data to the system under consideration.

Maintenance and redesign processes will typically generate metadata of the same nature as the original design process.

5. ORGANISING FLOWS AND HOLDINGS OF STATISTICAL METADATA

5.1. Experiences from the past and implications for the future

Many statistical metainformation systems in the past have failed for various reasons:

- Metadata collection is dull, expensive, and time-consuming.
- The natural metadata providers, those who know something about the object data to which the metadata refer, are difficult to motivate. They do not themselves need the metadata. At least they do not need the metadata as a formalised part of a system, since they have the metadata in their own heads.
- The metadata users, on the other hand, will not find a metainformation system to be of much value, until the metadatabase covers data from many surveys. This is so, because users of statistical data are usually interested in several (related) collections of data, which have been collected by different surveys.
- There is an unfortunate disjunction between users and producers of metadata. Those who need the metadata cannot themselves produce the metadata. On the other hand, those who "own" the knowledge about statistical data, do not benefit very much from formalized and automated availability of the metadata.
- In comparison with a well-functioning market, a typical metainformation system from the past has been lacking some important features: (a) relations between the supply side and the demand side have been virtually non-existing; (b) the supply side has had most of the costs but little benefit, whereas the demand side, which has been intended to get the benefits, have not had to pay for them, or, if they have had to pay, the payments may not have been routed back to the metadata suppliers, and have thus failed to have the "self-regulating" effect that such payments usually have in a market.

From the past experiences we can learn some lessons concerning the desirable architecture of future metainformation systems:

- Metadata collection activities should be minimized in the sense that no metadata should have to be entered more than once, and derivable metadata should be automatically derived rather than manually entered.
- Huge retrospective metadata collection activities should be avoided. Instead as much as possible of the metadata input flow should be generated as a side-effect of other activities. For example, the more or less formalized descriptions that are typically generated by systems analysis and design activities should be

automatically captured and organised as potential metadata for the information system under development;

- Some type of cost/benefit mechanism needs to be introduced into the architecture of a metainformation system in order to relate users and producers of metadata in a constructive way. The mechanism needs to be relatively sophisticated, since there is a many-to-many relation between users and producers of metadata: the same metadata may be used by many different users, and the same user may need metadata concerning many several data collections from several producers.

Statistical information systems are valuable assets. However, without properly integrated metainformation systems, the value of the information systems is drastically reduced. Since today's statistical information systems are by and large formalized and computerized, the metainformation systems must also be formalized and automated, if the pace of the metadata flow is to keep up with the pace of the object data flow.

There may be exceptional cases, where a single survey (or a small set of closely related surveys) has a statical and well-known set of users, who essentially need data from this survey (or set of surveys) alone, and who can rely on person-to-person contacts with the survey staff, together with oral traditions between themselves, for satisfying their needs of metadata. However, most survey data will be used by many different users, and most users of statistical data need to interpret and combine data from different sources.

5.2. Metadata holdings

The statistical metadata managed by a statistical service need to be stored in **metadata holdings**, organised into one or more **statistical metadatabases**. A statistical metadatabase may be **active** (integrated with a statistical database) or **passive** (separate from a statistical database). It may contain **local metadata** concerning individual surveys and/or **global metadata** concerning a wide range of surveys and data collections. It may be physically stored and maintained **locally** or **centrally** in the organisation.

As we have seen earlier in this paper, there is a need to store a lot of different kinds of metadata in a statistical metainformation system. The metadata can be categorized in several different dimensions, for example:

- by **metaobject type**;
- by being **microdata-oriented** or **macrodata-oriented**;
- by **data type** (quantitative, qualitative, textual);

- by **type of formalism** (fixed-format facts, logical expressions, mathematical expressions, algorithms, graphs, free text);
- by being **data-oriented** or **process-oriented**;
- by being **procedural** or **declarative**;
- by representing **specific facts** or **general knowledge**.

Since the metadata will be in many different forms, relatively advanced **database management software** will be needed for handling metadata holdings properly.

In many respects a statistical metadatabase can be designed in the same way as any other kind of database. For example, it is advisable to use a so-called **object graph** (or Entity-Relationship graph) for modelling the contents of a statistical metadatabase. Figure 5.1 shows an example of such a **metaobject graph**, containing **metaobjects**, **metavariables**, etc. The metaobject types should be interpreted as follows:

BOX a structured set of statistical characteristics (cf section 1.2), a socalled box;
POP population of objects (statistical units);
SAM sample of objects from a population;
XCL crossclassification of the population into (sub)domains of interest;
PAR parameter, statistical characteristic;
VAR variable;
VAS value set of one or more variables;
VAL value in value set;
SUR survey;

In figure 5.1 most (meta)object types occur in three versions: an **occurrence version** (*occ*), a **series version** (*ser*), and a **type version** (*typ*); corresponding to three layers of the conceptual model: an **occurrence layer**, a **series layer**, and a **type layer**. The division into three layers has the following background.

A typical production pattern is that "the same" survey is repeated at regular time intervals, for example monthly, quarterly, or yearly. In such cases it is appropriate to speak about a **survey series**. Surveys producing indexes and other indicators (like unemployment rates) are typical examples of such time series of "similar" surveys.

The individual survey repetitions within a survey series are never identical. Some component or aspect of the survey design is often changed, if only marginally. For example, a new data item may be added, another one may be redefined, etc. Even if the survey design should be the same between survey repetitions, the conditions under which the survey is carried out will change, which will result in changes in response rates and other aspects of the quality of the survey data.

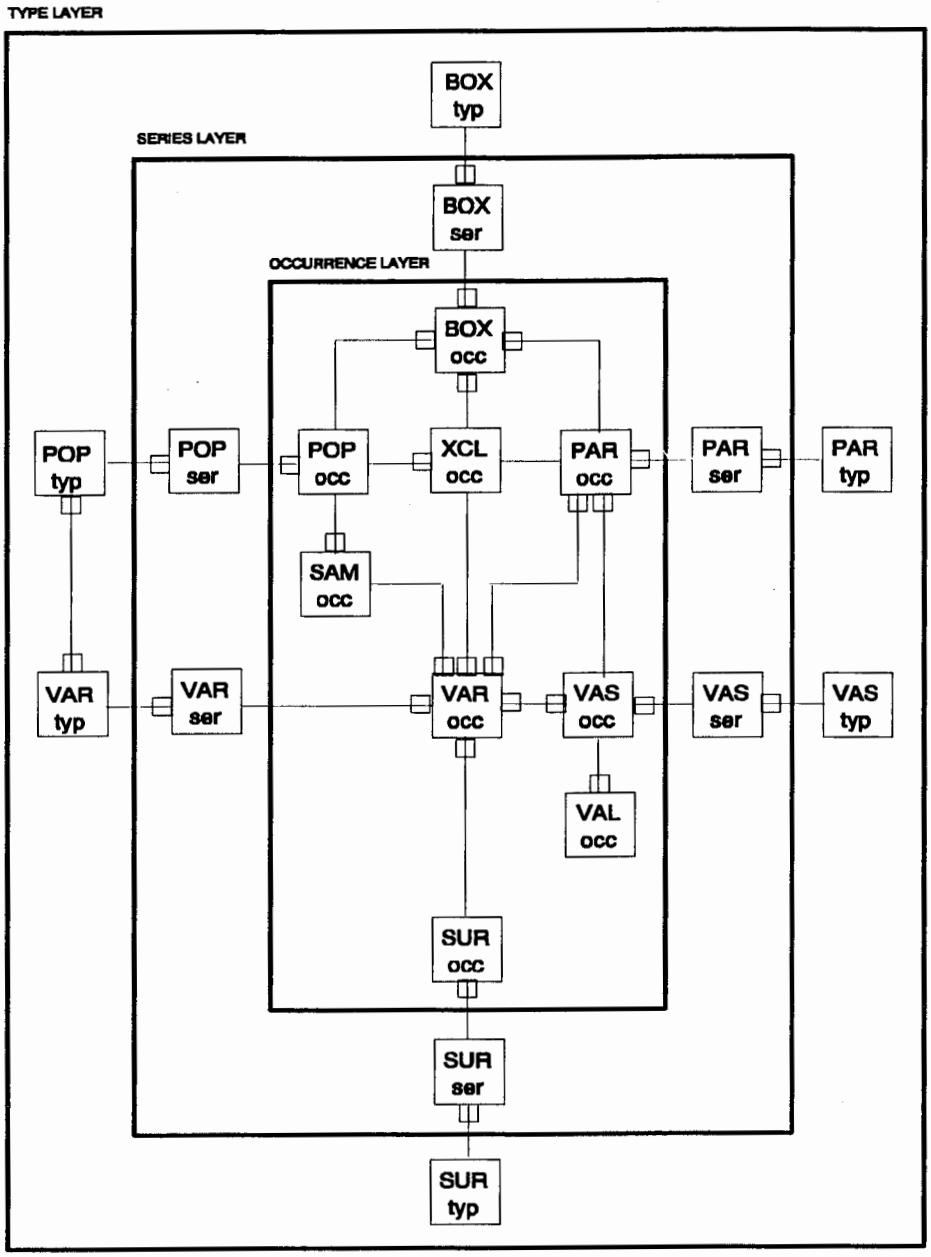


Figure 5.1
A metaobject graph modelling the contents of a statistical metadatabase.

Thus the metadata for different survey repetitions within a survey series will be different to a certain extent. *Both* the metadata generated by survey design decisions *and* the metadata generated by the survey process itself will change over time.

In principle, every metadata item *may* change from one survey repetition to the next one. On the other hand, many relevant metadata items *will not* change between survey repetitions. A failure to recognize properly *both* the similarities and the dissimilarities between different survey repetitions will negatively affect the **comparability in time**, an extremely important quality component for many users of statistics.

A similar problem concerns **comparability in space**, where "space" is a generic concept, covering not only geographical subdivisions, but also many other forms of classifications, where it is meaningful to recognize some kind of proximity and/or (fuzzy) similarity between different instances (occurrences) of one and the same type. For example, populations and variables with "similar" definitions may be good **substitutes** for each other with respect to certain needs.

The three-layer model in fiugre 5.1 is one way of taking care of the user needs for comparability in time and space. The **type layer** should contain metadata, which are "usually" the same, or at least "similar" for different members of the same type. The type level metadata have the character of "**general rules**" or "**typical descriptions**"; **exceptions** to the rules can be given for subtypes and/or occurrences of the types.

Analogously, the **series layer** should contain metadata, which are "more or less" the same for different repetitions within a time series. Once again exceptions to the typical descriptions can be given on the occurrence level.

The **occurrence layer** should primarily contain metadata, which are known to be different between different occurrences within the same series, or the same type, respectively. High variability in this sense is typical for most **operation-based metavariables**, like "measurement problems" and "non-response rate". **Design-based metavariables** will not change their values between repetitions of "the same" survey to the same extent.

To summarize, many metavariables will have to be recorded on the occurrence level. However, if a metavariable is known to be relatively stable over time, it could be recorded on the series level, provided that there is an option to record **occurrence level exceptions** from the **series level rule**. The exceptions could result in footnotes in appropriate places, when the data are presented.

For example, if the measurement procedure for a variable is usually the same from survey repetition to survey repetition, the information about the measurement procedure could be given for the "VAR series" metaobject. If something unusal should occur with the measurement procedure during some particular repetition of the

survey, this could be noted as an exception from the general rule, and the exceptional information would be recorded for the appropriate "VAR occurrence" metaobject.

If a metavariable is less stable, but still does not vary too much over time, it may be better to make the primary recordings on the occurrence level, but complement this information with some "**overview information**", which is given on such a level of abstraction that it becomes stable over time.

For example, if response rates vary rather modestly over time, one could give information about the "**normal**" response rate span on the series level and give an "alarm signal" on the occurrence level, whenever the response rate falls outside the "normal span".

One could apply similar principles for determining the distribution of metadata between the type layer and the series layer of the metadatabase. "Normal" values of metavariables could be given on the type level, and exceptions from what is regarded as "normal" could be signalled on the series and occurrence levels.

5.3. Metadata flows

Every statistical service function, which somehow manages data, should also manage the metadata, which is associated with the data.

In fact automation and computerization of survey management has up to recently implied **disintegration** of the natural relationships between statistical data and metadata, which existed in earlier manual systems. For example, consider a questionnaire. When it has been completed, it contains both data (answers to questions) and the associated metadata (the questions themselves and accompanying instructions for answering the questions). As long as the forms were processed manually, the data and metadata continued to go "hand in hand" throughout all the processing steps, until the final tables had been produced. Automation primarily aimed at rationalizing the counting process, a process which deals with the object data only. Thus the object data became separated from the metadata. When a programmer, in a later production step, was to compose readable tables, he or she would have to (re)introduce metadata, explaining the meaning of the data in the tables, but at that stage the original metadata (questions, instructions, etc) might very well have been lost track of. Thus the metadata in the presented tables would not normally be the result of systematical, formalized transformations of the metadata in the questionnaires.

An essential feature of modern metadata management is that it is **reintegrated** with object data management, so that for example the metadata describing the figures in presented tables would in fact be the result of a chain of systematical, formally well-

defined, and automated transformation processes, starting with the metadata in the questionnaire, or maybe even earlier, with the metadata generated by design decisions preceding the (computer-aided) construction of the questionnaire.

During all activities of all phases of the life-cycle of a statistical system, the different actors produce decisions, documents, etc, which contain metadata. If the metadata are properly captured and organized, they may become very useful, when the same statistical information system, or other ones, require metadata input.

It should be a challenge for every statistical service to organize its metadata flows in such a way that

- as many metadata as possible can be obtained from existing metadata holdings, whenever they are needed by a certain actor in a certain statistical system;
- as few metadata as possible have to be produced for its own sake, rather than as a side- effect of other (necessary) activities of the statistical systems monitored by the statistical office.

Sharing of metadata (as well as sharing of object data) **within and between systems** is essential for any statistical service aiming at rational, computer-supported planning and operation of its statistics production. Systematical, automated exchange of metadata between different activities promotes two good causes at the same time:

- it **decreases the burden** on metadata providers;
- it **increases the benefits** gained from metadata, which are already available.

International standards (like GESMES) for the storage and exchange of statistical data and metadata should significantly facilitate the efforts of statistical services to systematize and automate exchange of data and metadata, both internally and externally.

5.4. Interdependencies between metadata flows

There are important interdependencies in the metadata flows between

- local and global metadata holdings
- different statistical information systems
- different phases of the life cycle of each statistical information system..

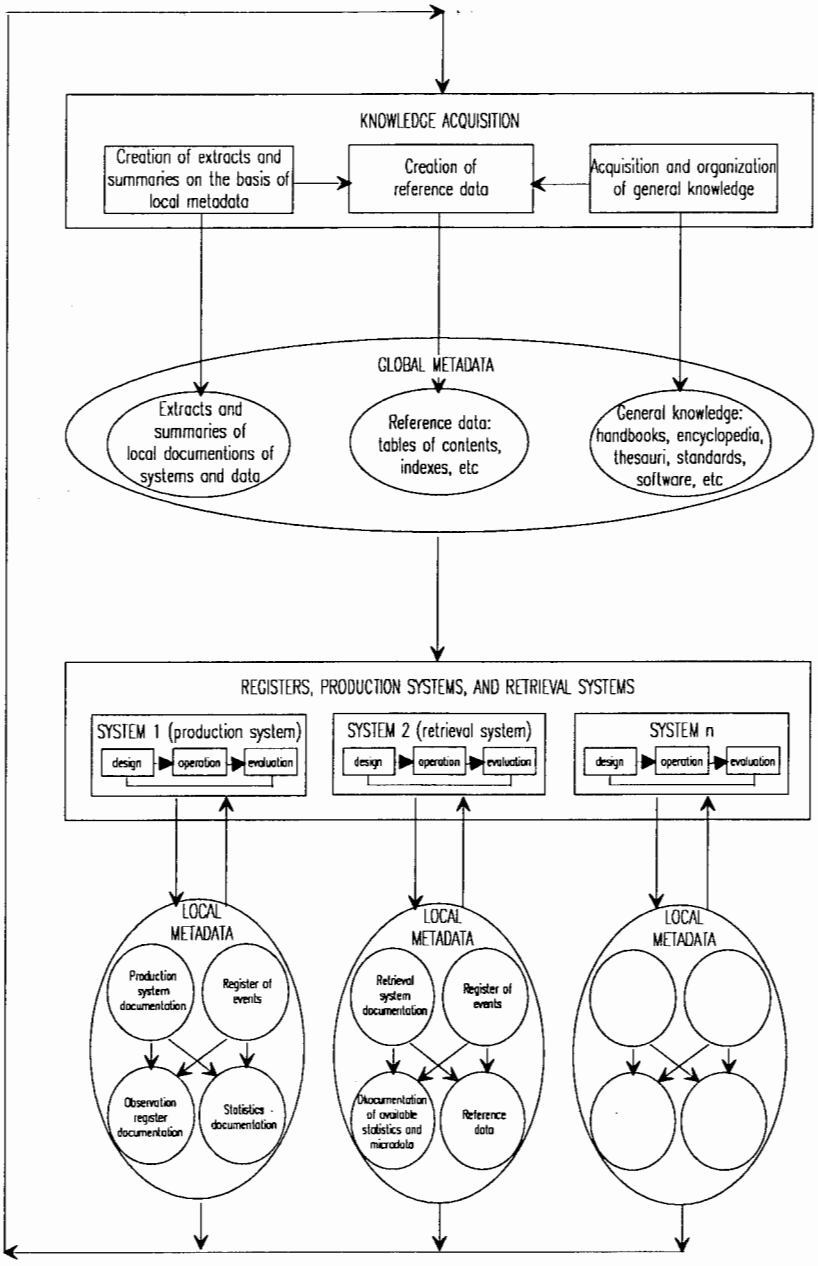


Figure 5.2
Metadata flows for a system of different kinds of statistical information systems.

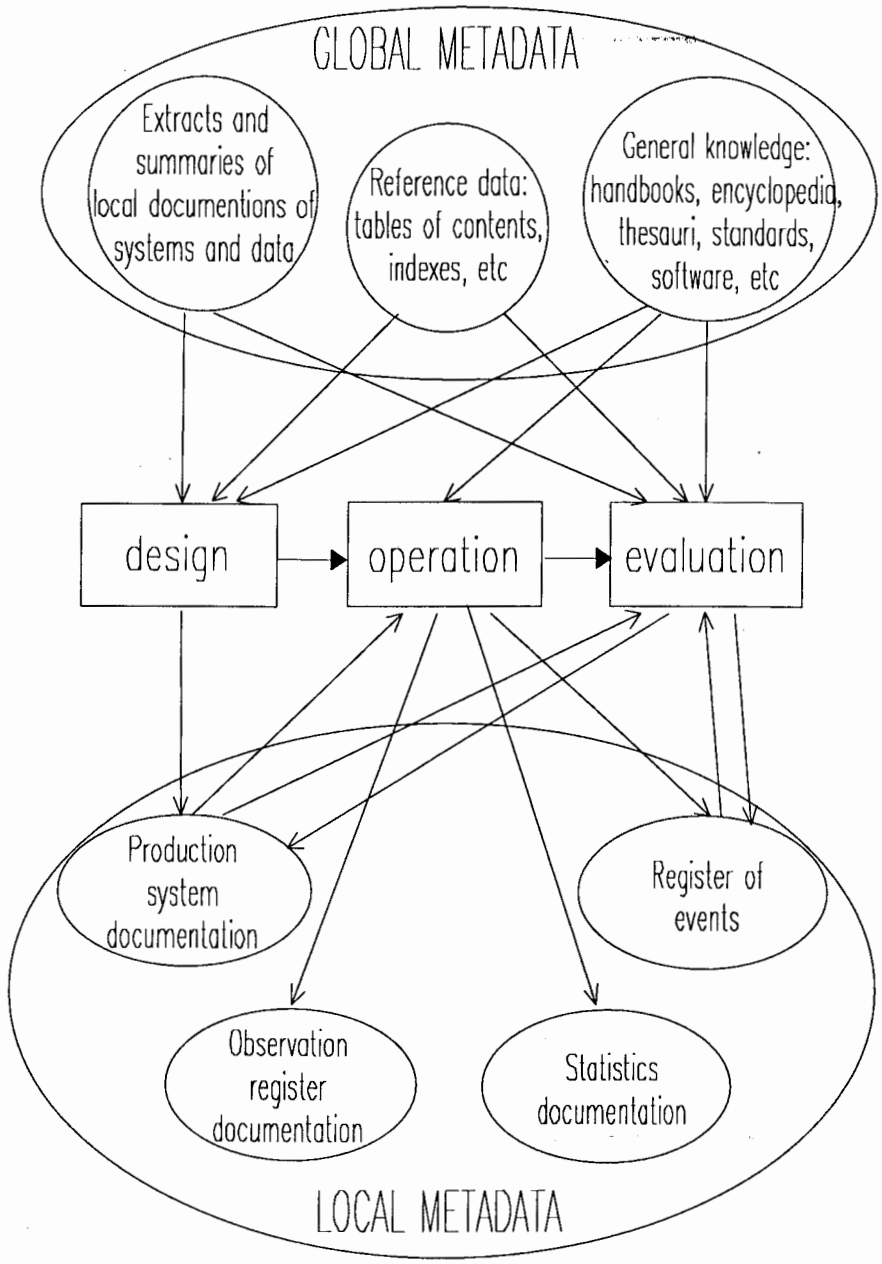


Figure 5.3
Metadata flows for the three major phases of the life cycle of a survey production system.

For example, many metadata are naturally generated and captured as the result of design processes and design decisions, and many of these metadata (for example names, definitions, and storage formats of different statistical characteristics and their components) are later needed (at operation time) by producers and users of statistical data, as well as by software products supporting the tasks of users and producers. Figure 5.2 shows an important feed-back loop from the local metadatabases of a number of (different types of) statistical information systems to a common global metadatabase. The local databases contain detailed knowledge concerning specific systems and their data (observation registers and statistics). The local metadata have to be processed (as automatically as possible) in order to create extracts and summaries, which can be managed in the global metadatabase, from which metadata can be retrieved by local as well as global and external systems. In order to make the retrieval of global metadata as efficient and user- friendly as possible, the extracts and summaries have to be further processed (once again as automatically as possible) in order to create and maintain reference data like tables of contents and indexes.

The generation of extracts, summaries, and reference data is one part of the knowledge acquisition process for the global metadatabase. Another part is the acquisition of general knowledge: handbooks, encyclopedia, thesauri, standards, software, etc. The acquisition of general knowledge can be performed rather independently of the feed-back loop just described. However, there is a potential for creating an "intelligent" **inductive learning loop** from the local and global specific knowledge to the global general knowledge. At the present state of the art this inductive learning loop will be highly dependent on human efforts, but artificial intelligence may contribute increasingly in the future.

Figure 5.3 focuses on the metadata flows for one statistical information system only, a survey production system. However, for this single system, figure 5.3 gives a more precise description of the metadata flows for each one of three major phases of the life cycle of the system: design, operation, and evaluation.

REFERENCES

[1] **Appel, G.** (1993). "Metadata Driven Statistical Information Systems". *Proceedings of the Statistical Metainformation Systems Workshop in Luxemburg*, February 1993.

[2] **Fröschl, K.A.** (1993). "Towards an Operative View of Semantic Metadata". *Proceedings of the Statistical Metainformation Systems Workshop in Luxemburg*, February 1993.

[3] **Klas, A.** (1985). "The Metainformation System: Its Structure and Role in the Statistical Information System". *Journal of Official Statistics*, Vol. **1**, *n°* **4**, pp 413–426.

[4] **Lamb, J.** (1993). "Metadata in Survey Processing". *Proceedings of the Statistical Metainformation Systems Workshop in Luxemburg*, February 1993.

[5] **Langefors B.** (1966). *Theoretical Analysis of Information Systems.* Studentlitteratur, Lund.

[6] **Malmborg, E.** and **Lisagor, L.** (1993). "Implementing a Statistical Meta-Information System". *Proceedings of the Statistical Metainformation Systems Workshop in Luxemburg*, February 1993. Also in *Statistical Journal of the United Nations UN/ECE* 2/1993. Also available from *Statistics Sweden.*

[7] **Malmborg, E.** and **Sundgren, B.** (1994). " Integration of statistical information systems — theory and practice". *Proceedings of the Seventh International Conference on Scientific and Statistical Database Management*, University of Virginia, USA, September 1994, IEEE Computer Society Press.

[8] **Nordbotten, S.** (1993). *Statistical Meta-Knowledge and -Data.* Invited opening lecture for the Statistical Metainformation Systems Workshop in Luxemburg, February 1993.

[9] **Rosén, B.** and **Sundgren, B.** (1991). "Documentation for reuse of microdata from the surveys carried out by Statistics Sweden". *Statistics Sweden.* Original report in Swedish. English translation available.

[10] **Silver, M.S.** (1993). "The Role of Footnotes in a Statistical Meta-Information System". *Proceedings of the Statistical Metainformation Systems Workshop in Luxemburg*, February 1993.

[11] **Statistical Computing Project** (1984). *Users Guide to Metainformation Systems in Statistical Offices.* United Nations, Economic Commission for Europe, Conference of European Statisticians.

[12] **Sundgren B.** (1973). "An Infological Approach to Data Bases". *Statistics Sweden.*

[13] **Sundgren, B.** (1980). "Meta-Information in Statistical Agencies". *Statistics Sweden.*

[14] **Sundgren B.** (1984). "Conceptual Design of Data Bases and Information Systems". *Statistics Sweden.*

[15] **Sundgren, B.** (1989). *Conceptual Modelling as an Instrument for Formal Specification of Statistical Information Systems.* ISI 47th Session, Paris.

[16] **Sundgren, B.** (1991a). "What metainformation should accompany statistical macrodata?". *Report for the June 1991 Meeting of Working Party 9 of the OECD Industrial Committee* as a basis for a discussion on the topic of Standards for Metadata in International Databases. Also available from *Statistics Sweden.*

[17] **Sundgren, B.** (1991b). "Statistical Metainformation and Metainformation Systems". *Report for the UN/ECE METIS Group*, established within the programme of work of the Conference of European Statisticians. Also available from *Statistics Sweden.*

[18] **Sundgren, B.** (1991c). "Towards a Unified Data and Metadata System at the Australian Bureau of Statistics". *Consultancy report for the Australian Bureau of Statistics (ABS).* By permission of the ABS, the report is also available from the author.

[19] **Sundgren, B.** (1991d). "Some Properties of Statistical Information: Pragmatics, Semantics, and Syntactics". *Statistics Sweden*, 1991.

[20] **Sundgren, B.** (1992a). "Organizing the Metainformation Systems of a Statistical Office". *Report for the UN/ECE METIS Group*, established within the programme of work of the Conference of European Statisticians. Also available from *Statistics Sweden.*

[21] **Sundgren, B.** (1992b). "Statistical Metainformation Systems —Pragmatics, Semantics, and Syntactics". Invited paper for the *Statistical Metainformation Systems Workshop* in Luxemburg, February 1993. Also available from *Statistics Sweden.*

[22] **Sundgren, B.** (1993). "Guidelines on the Design and Implementation of Statistical Metainformation Systems". *Report for the UN/ECE METIS Group*, established within the programme of work of the Conference of European Statisticians. Also available from *Statistics Sweden.*

[23] **Sundgren, B.** (1994a). "Statistical information systems in a modern society: roles, functions, and system designs". Invited paper for the *Baltic Workshop on National Infrastructure Databases*, Vilnius, Lithuania, May 1994. Also available from *Statistics Sweden.*

[24] **Sundgren, B.** (1994b). "Statistics production in the 1990's". Invited paper for *Eurostat's High-Level Seminar on Strategy for Statistical Computing*, in Luxemburg, July 1994. Also available from *Statistics Sweden.*

[25] **UN Western European EDIFACT Borard, Message Development Group 6 Statistics** (1993). *GESMES 93 For the exchange of multidimensional statistical arrays and time-series data: "Guidance to users" and "Reference Guide".* Published by Eurostat, Luxemburg.

SECCIÓ DOCENT I PROBLEMES

La introducció de la nova "SECCIÓ DOCENT I PROBLEMES" a la revista QÜESTIIÓ es fa amb l'objectiu d'incloure una secció on es publiquen articles de caire docent, difícilment publicables en revistes de recerca. Alhora es continua amb l'antiga secció de problemes. A cada número de QÜESTIIÓ s'inclourà d'un a tres problemes i les solucions es donaran en el número següent.

Els lectors poden, si ho volen, proposar problemes amb les solucions pertinents i enviar-los a QÜESTIIÓ, que farà una selecció i en publicarà els més adequats, fent la corresponent referència a l'autor.

També seran ben rebudes solucions alternatives a les propostes fetes per l'autor dels problemes; l'editorial es reservarà, però, el dret a publicar-les.

QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 361-380, 1995

LA ESTADÍSTICA DE DOMINIO PÚBLICO

C. RADHAKRISHNA RAO*

Pennsylvania State University

La vida es el arte de obtener suficientes conclusiones a partir de evidencias insuficientes.

Samuel Butler

Para entender los pensamientos de Dios debemos estudiar estadísticas, dado que éstas son las medidas de su voluntad.

Francis Nightingale

6.1. Ciencia para todos

En su libro sobre *La función Social de la Ciencia,* publicado en 1939, J.D. Bernal escribió:

> De nada sirve que mejoremos la comunicación que los científicos tienen entre sí acerca de su propio trabajo, si al mismo tiempo no logramos que un conocimiento real de la ciencia llegue a ser, en nuestra época, parte de la vida diaria.

Tan sólo medio siglo más tarde se ha reconocido la importancia de lo que dijo Bernal, y se han realizado serios esfuerzos para difundir el conocimiento científico entre el público. Las Academias Nacionales de Ciencias de los paises avanzados han nombrado equipos para examinar el problema y sugerir caminos para su resolución. Hace 5 años la Real Sociedad del Reino Unido empezó a publicar una revista, llamada *Science and Public Affairs,* con el propósito de fomentar el conocimiento, por parte del público, de las publicaciones científicas y aclarar las implicaciones en la vida

*C. Radhakrishna Rao. Department of Statistics. Pennsylvania State University. University Park, PA 16802.

Aquest article correspon al capítol 6 del llibre *Estadística y Verdad* de C. Radhakrishna Rao. Publicat amb l'autorització de l'Editorial P.P.U.

diaria de los descubrimientos científicos y tecnológicos. El nuevo eslogan puesto en circulación por la Real Sociedad es

"La Ciencia es para todos"

Sin lugar a dudas, la Ciencia impregna casi todo lo que hacemos en sociedad, y la importancia del conocimiento de la ciencia para el hombre de la calle no precisa ser remarcada. El público debe conocer cómo la nueva tecnología puede serle útil para mejorar su nivel de vida. Deben conocer las consecuencias de la explotación de nuevos descubrimientos para su propio beneficio sin hacer caso de los efectos perjudiciales para la sociedad y el medio ambiente. Deberían ser conscientes que una política gubernamental, como establecer plantas de energía nuclear por todo el país, afectará sus vidas y las de sus hijos.

Cuando Bernal escribió el libro, la Estadística no era conocida como una disciplina independiente. Creció en importancia en el segundo cuarto del presente siglo, como un método para extraer información de datos observados y como el camino lógico de tomar decisiones en casos de incertidumbre. Como tal, el conocimiento de la estadística es de gran valía para la humanidad en todos los sentidos de la vida. Si Bernal estuviese vivo hoy día para sacar una nueva edición de *La Función Social de la Ciencia* podría haber añadido, impresionado por la ubicuidad de la Estadística, que el conocimiento público de la ciencia estadística es mucho más importante que cualquier otro campo científico.

6.2. Datos, información y conocimiento

El único problema de una cosa segura es la incertidumbre.

¿Qué es la Estadística? ¿Es ciencia, tecnología, lógica, o arte? ¿Es una disciplina independiente como las matemáticas, la medicina, la química y la biología, con una temática a estudiar bien definida? ¿Qué fenómenos estudiamos con la Estadística?

La Estadística es una disciplina peculiar que no tiene por objeto ninguna parte concreta de la realidad por si misma. Parece que existe y se desarrolla para resolver problemas de otras áreas. En palabras de L.J. Savage

> "La Estadística es básicamente parasitaria: vive del trabajo de otros. No es un demérito de dicha disciplina. Actualmente se reconoce que muchos huéspedes morirían de no ser por los parásitos que albergan: algunos animales no pueden digerir su alimento. Así ocurre también en muchos campos del esfuerzo humano, tal vez no desaparezcan pero ciertamente se debilitarían sin la existencia de la Estadística."

La Estadística no se ha instaurado en el currículum académico de las universidades hasta el presente siglo. Ni siquiera ahora el papel de la Estadística en la ciencia y en la sociedad ha sido bien entendido por el hombre de la calle y los profesionales de diversas áreas.

No hace mucho tiempo, habían ideas falsas y escepticismo acerca de la Estadística expresadas en afirmaciones como las siguientes:

* Mentiras, condenadas mentiras y estadísticas.

* La Estadística no dispensa la necesidad de razonar.

* Conozco la respuesta, dadme una estadística para justificarla.

* Puedes probar cualquier cosa con estadísticas.

Era también el objeto de chistes tales como:

* La Estadística es como un bikini. Pone de relieve lo obvio pero oculta lo esencial.

Actualmente la estadística ha llegado a ser una palabra mágica que da apariencia de realidad a los enunciados que formulamos:

* *La Estadística demuestra* que fumar cigarrillos es malo.

* *De acuerdo con las estadísticas,* los hombres que permanecen solteros mueren diez años antes.

* *Estadísticamente hablando* los padres altos tienen hijos altos.

* Una encuesta por *muestreo estadístico* ha revelado que tomar una aspirina a días alternos reduce el riesgo de un segundo ataque cardíaco.

* Hay una *evidencia estadística* de que los hijos nacidos en segundo lugar son más inteligentes que los nacidos en el primero.

* La *Estadística confirma* que una toma de 500 mg. de vitamina C cada día prolonga la vida 6 años.

* Un *estudio estadístico* ha revelado que los maridos dominados por su mujer tienen mayor probabilidad de sufrir un ataque cardíaco.

La Estadística como disciplina científica tiene una corta historia, pero como *información numérica* tiene una larga antigüedad. Hay varios documentos antiguos que contienen información numérica acerca de paises (estados), sus recursos y composición de la población. Esto explica el origen de la palabra *estadística* como una descripción objetiva de un *estado.* Referencias a censos de población y agrícolas

sabemos hoy en día que se pueden encontrar en el libro chino *Kuan Tzu* (1000 a.C.), *Viejo Testamento* (1500 a.C.) y *Arthaśāstra of Kautilya* (300 a.C.).

Un ejemplo de antiguos registros estadísticos, recientemente descubiertos, son las cifras encontradas en una pirámide de un faraón egipcio que vivió hace unos 50 siglos (3000 a.C.). Éstas se referían a la captura de:

120.000	prisioneros de guerra
400.000	bueyes
1.422.000	cabras

después de una guerra y gracias al ejército de un victorioso faraón. ¿Cómo llegaron a ser estas cifras tan armoniosamente redondeadas? ¿Eran cifras efectivas hechas por los reales guardianes contables o cifras ficticias concebidas para la activa imaginación del victorioso faraón? ¿Era el drástico redondeo de las cifras un intento de subrayar la magnitud del botín?

Samuel Johnson creía:

"*Los números redondos son siempre falsos*"

lo que ya había sido anticipado por Weirus, un físico alemán del siglo XVI, una época en que la mayor parte de Europa estaba sometida al temor de las enfermedades y la brujería. Weirus calculó que exactamente

7.405.926

fantasmas habitaban la Tierra! La mayoría de la gente creyó que tal cifra debía ser el recuento real ya que Weirus era un hombre erudito.

Recuerdo lo que se recomendaba en una *Guía de Impuestos* mientras rellenaba mi declaración de impuestos en U.S.A.

> "Un detallado examen del informe G.A.O. confirma una forma importante de reducir las auditorías. Evitar redondear a dólares cuando se detallan ganancias y gastos. Cifras tales como $100, $250, $400, $600 inducen sospechas al inspector, mientras que cifras tales como $171, $313, $496 disminuyen la probabilidad de inspección. Si tiene que tasar algunos gastos, hágalo en cantidades raras".

La definición etimológica de la palabra estadística significa *datos* obtenidos por diversos medios. ¿Qué transmiten los datos y cómo debemos usarlos para lograr un objetivo determinado? Para ello, debemos saber qué clase y cuánta *información* hay en los *datos observados* utilizable para resolver un *problema dado.* ¿Qué es la información? Quizás, la definición más lógica, como la dada por Claude Shannon, un experto en teoría de la información, es "hacer disipar la incertidumbre", que es la piedra angular de la resolución de un problema. Los datos, por si mismos, no

son una respuesta a un determinado problema. Pero es el material básico, a partir del cual podemos evaluar lo bien que podemos resolver un problema, cuan dudosa es una respuesta particular o bien qué confianza podemos poner en ella. Los datos observados necesitan ser procesados para averiguar hasta qué grado la incertidumbre puede disiparse. El conocer la cantidad de incertidumbre asociada a los datos es la llave para tomar la decisión apropiada. Ello nos permite sospesar las consecuencias de diferentes opciones y escoger una que sea la menos perjudicial. La Estadística tal como es entendida actualmente es la lógica a través de la cual podemos subir un peldaño en la escalera que nos lleva de los *datos* a la *información.*

A medida que la información aumenta gradualmente, reduciendo la incertidumbre a un nivel mínimo aceptable, vamos subiendo varios peldaños más en la escalera del estado del *conocimiento,* lo que nos da seguridad en las decisiones que tomamos (sujetas naturalmente a un inevitable aunque pequeño riesgo). Tal nivel de conocimiento puede no ser alcanzable en todas las áreas y en todas las situaciones. Esto crea la necesidad de la estadística, como la metodología de la toma de decisiones bajo un nivel de incertidumbre asociada a los datos obtenidos.

De acuerdo con el distinguido científico, Rustum Roy, el conocimiento que encaja en un cuerpo determinado del saber, lo amplía, constituyendo nuevo saber, lo que supone un peldaño más en la escalera del conocimiento.

No es más que un antiguo proverbio:

¿La ruta de la sabiduría?
Bien, es llano y simple de expresar
Errar
y errar
y otra vez errar
Pero cada vez menos
y menos
y menos.

6.3. La revolución de la Información y la comprensión de la Estadística

> *Llegará un tiempo tal vez no muy lejano que se comprenderá que para una formación completa como ciudadano eficiente ..., es tan necesario saber calcular, pensar en términos de promedios, máximos y mínimos, como lo es ahora saber leer y escribir.*
>
> H.G. Wells

La prosperidad de la humanidad dependió en el pasado de la revolución agrícola y luego de la revolución industrial. Pero ello no nos ha alejado de aliviar la miseria de la

gente en aspectos como hambre y enfermedades. El principal obstáculo para progresar ha sido nuestra incapacidad para prever el futuro y tomar decisiones políticas sabias. La política sana se basa en una buena información. Así hay necesidad de ampliar la base de datos para reducir la incertidumbre y tomar mejores decisiones.

La importancia de la información como ingrediente clave en la planificación y ejecución de un proyecto más que la pericia tecnológica, es ahora ampliamente reconocido, y somos testigos de la revolución informática, ya que empresas tanto públicas como privadas están haciendo grandes inversiones en la adquisición y procesamiento de la información. Se dice que en los EE.UU. alrededor del 40 ó 50% de los empleados en el sector público y en el privado están ocupados en estas actividades.

Que hay demanda pública de estadística se demuestra por el hecho de que los periódicos dedican considerable espacio en dar toda clase de información. Tenemos la predicción detallada del tiempo por un periodo que se prolonga a alrededor de una semana, hecho que permite planificar nuestras actividades al aire libre. Están las cotizaciones de la Bolsa, que nos indican qué inversiones pueden sernos más provechosas. Una sección especial está dedicada a los deportes con el fin de mantenernos informados de los acontecimientos deportivos de todas las partes del mundo. Un periódico diario de Edmonton, Canadá, publica lo que se denomina el índice diario de mosquitos, con el fin de convencer al público de que las autoridades municipales están haciendo los máximos esfuerzos para controlar el nivel de mosquitos en las ciudades. El New York Times dedica casi el 30% de su espacio para toda clase de estadísticas así como reportajes basados en ellas.

Hay revistas con estudios de consumidores que informan al público acerca de los precios de artículos de consumo y los resultados comparativos de varios productos del mercado.

La comprensión de la estadística resulta importante a varios niveles. El primero es a nivel individual. La necesidad de conocer las tres reglas (leer, escribir y contar) es bien conocida. Pero esto no es suficiente para hacer frente a las incertidumbres que encuentra un individuo en cada momento de su vida. Tendrá que tomar decisiones para matricularse en un colegio, casarse, hacer inversiones y resolver los problemas del trabajo diario. Esto requiere diferentes tipos de habilidades, que nosotros podemos llamar la cuarta regla: el razonamiento estadístico, comprensión de las incertidumbres de la naturaleza y del comportamiento humano y minimización del riesgo en la toma de decisiones, haciendo servir la propia experiencia y la colectiva. Además, el conocimiento estadístico para un individuo será una ventaja para su propia protección y la de su familia contra infecciones, contra la propaganda de los políticos y de los anuncios poco escrupulosos, de los negociantes, eliminando la superstición que es peor que la enfermedad, aprovechándose de las predicciones del tiempo, enterándose

de desastres como la radiación que se escapa de las plantas nucleares y muchas otras cosas que afectan a su vida y sobre las que no posee control.

¿Necesita el hombre de la calle estudiar estadística para adquirir lo que denominamos la cuarta regla? La respuesta es no. Una cierta educación estadística en la enseñanza media, junto con las matemáticas, sería suficiente. Nuestro actual sistema educativo está más orientado a estimular a los estudiantes a creer en la palabra escrita y les previene contra la toma de decisiones con riesgo simbolizado en frases como "No cuentes los pollos antes de que salgan del cascarón", en lugar de prepararlos para vivir en un mundo incierto y aprender a hacer frente a situaciones límite de la vida moderna sin precipitación.

Debemos aprender cómo enfrentarnos a un riesgo calculado. Recientemente, hubo un reportaje en la prensa, que decía que entre los nombres grabados en el "Vietnam Veterans Memorial" en Washington, hay por lo menos 38 que erróneamente han sido dados por muertos. Cuando la persona responsable fue preguntado por ello, dijo: "No era posible en el momento de la construcción saber si estos soldados estaban muertos, porque los datos eran incompletos. Yo no sabía si sería posible añadir nombres una vez construido el Memorial. Tuve la creencia de que sus nombres podían perderse para la historia si no los hubiera incluido".

En el siguiente nivel tenemos a políticos y artífices de la política, para quienes el conocimiento estadístico es importante. Los gobiernos tienen una descomunal maquinaria administrativa para recopilar datos. Son estos los medios que se usan para tomar decisiones políticas correctas en la administración cotidiana y formular planes de largo alcance para fines sociales. Los políticos intentan buscar consejos técnicos para tomar decisiones. No obstante, es importante que por ellos mismos adquieran algún conocimiento técnico para comprender e interpretar la información. Las siguientes anécdotas ilustran este punto.

> Los estadísticos del Gobierno y de la industria a menudo se hallan frente a barreras lingüísticas con sus jefes. El jefe de una oficina estadística, un oficial administrativo, se entrevistaba con un grupo de estadísticos que se quejaban de que en un informe recibido de otra organización, algunas estimaciones no detallaban el error estándar.[1] El jefe, al ser informado, inmediatamente señaló: "¿Hay estándares para los errores también?"

> Un informe sometido a Tea Board por un asesor estadístico, contenía una tabla con el título: Número estimado de gente que toma el té con error estándar. Pronto fue enviada una carta al estadístico preguntando qué clase de error estándar es el que la gente toma con el té.

[1]El error estándar es usualmente un número adjunto a una estimación, para dar una idea aproximada de la magnitud del error en la misma.

Una comisión regia, revisando un informe estadístico en el que se decía que las familias de clase media tenían una media de 2.2 hijos, comentó:

> La cifra de 2.2 hijos por mujer adulta es en algunos aspectos absurda. Se sugiere que se ayude a la clase media pagándoles dinero con el fin de incrementar la media hasta una cifra más redondeada y conveniente.
>
> Punch

Un ministro de sanidad quedó intrigado por la afirmación formulada por un estadístico de que 3.2 personas de cada 1000 enfermaron y murieron durante el último año. Preguntó a su secretario privado, un administrativo, ¿cómo puede ser que mueran 3.2 personas? El secretario respondió:

> "Señor, cuando un estadístico dice que 3.2 personas han muerto, significa que 3 personas realmente han fallecido y que 2 están a punto de morir".

Las decisiones en la política gubernamental son importantes por su efecto sobre millones de personas. Necesitan una sólida información e igualmente sólida metodología para procesarla.

Finalmente, existen profesionales en medicina, economía, ciencia y tecnología para quienes la interpretación y análisis de datos es hasta cierto punto una parte necesaria de su trabajo.

6.4. Números lúgubres

No me lo cuentes, en lúgubres números
La vida no es más que un sueño vacio.

H.W. Longfellow

Continuamente se nos previene, a través de los periódicos, revistas y otros medios de comunicación, de los buenos y de los perjudiciales efectos de nuestros hábitos dietéticos, hacer ejercicio, sobre la costumbre de fumar y beber, tensiones en nuestra profesión y otras actividades diarias. La información viene expresada en números que representan pérdidas o ganancias en determinadas unidades. He aquí algunos números lúgubres reproducidos de Cohen y Lee (1979). (Ver Tabla 1).

¿Cómo hemos de interpretar estas cifras? ¿Qué mensaje nos transmiten? ¿De qué utilidad son para un individuo en la modelación de su estilo de vida, para incrementar su felicidad?

Tabla 6.1

Disminución en la Esperanza de vida debida a varias causas*

Causa	Días	Causa	Días	
Soltero	3500	Alcohol	130	
Zurdo	3285	Accidentes con armas de fuego	11	
Soltera	1600	Radiación natural	8	
Con 30% de sobrepeso	1300	Rayos X (origen médico)	6	
Con 20% de sobrepeso	900	Café	6	
Fumar cigarrillos (varón)	2250	Contraceptivos orales	5	
Fumar cigarrillos (hembras)	800	Bebidas dietéticas	2	
Fumar cigarros	330	Test de Papanicolau	−4	**
Fumar en pipa	220	Alarma anti humo en casa	−10	
Trabajos peligrosos, accidentes	300	"airbag"	−50	
Accidentes de trabajo	74	Unidad móvil-Cuidados coronarios	−125	

* Datos relativos a otras causas pueden hallarse en Cohen y Lee.

** Los números negativos indican ganancia en esperanza de vida.

Consideremos la primera cifra de la Tabla 6.1, la pérdida en esperanza de vida de un hombre si permanece soltero. Ésta puede ser obtenida de información asequible usualmente en los registros de fallecimientos por sexos, estado civil y edad al fallecer. De los registros de varones, simplemente debe computarse separadamente, el promedio de edad al fallecer para los casados y para los solteros. La diferencia en estos promedios es la cifra: 3500 dias. Esto probablemente facilita una amplia evidencia del riesgo de quedarse soltero, habla favorablemente de la institución del matrimonio

y da un fuerte respaldo al consejo de casarse lo antes posible y *ahorrar* alrededor de *10 años* de vida! No obstante, ello no implica una causa [casarse] y un efecto [vivir 10 años más] aplicable *a cada individuo.* Puede ser bastante probable que para un individuo determinado, casarse sea suicida! Sin duda, una detallada tabulación de los registros de fallecimientos clasificados en subgrupos de varones de acuerdo con varias características personales sería más informativa. Diferentes subgrupos pueden tener comportamientos diferentes en pérdida o ganancia de esperanza de vida. Un determinado individuo puede analizar su propia personalidad y comparar su caso con la cifra indicada para el subgrupo de personas con características similares a las suyas.

Se aprecia en la Tabla 6.1 que los zurdos fallecen alrededor de 9 años antes que los diestros. ¿Implica ello que hay algo genéticamente anómalo en los zurdos? Quizás no: la diferencia puede ser debida a la desventaja que los zurdos tienen viviendo en un mundo donde la mayor parte de facilidades se han hecho a medida para los diestros. Sin embargo, la información estadística puede servir a los zurdos para protegerse contra posibles peligros.

Un promedio, en general, facilita una amplia indicación de las características de un grupo de individuos (población) como un todo. Sirve provechosamente para comparar poblaciones. De este modo podemos decir que una población de individuos con una media de ingresos de 1000$ al mes tiene mejor nivel de vida que otra con 500$ al mes. Un promedio no nos dice nada sobre las disparidades de los ingresos individuales. Por ejemplo, éstos pueden variar de 20$ a 100000$ y el promedio ser 1000$. Las diferencias entre los ingresos individuales dentro de una población, la llamada *variabilidad,* es también apropiada para comparar poblaciones. En muchos casos, una cifra promedio y alguna medida de variabilidad (como la gama de ingresos) facilita información de cierto valor práctico. Un promedio por sí mismo puede ser decepcionante y no se puede utilizar en todos los casos para hacer afirmaciones relativas a un individuo. Imaginemos que se informa a una persona, que no sabe nadar, que puede cruzar un río porque su estatura es superior a la profundidad media del río!

6.5. La predicción del tiempo

> *Un pronosticador digno de confianza es aquel cuyo micrófono está lo bastante próximo a una ventana para que así pueda decidir si usar la predicción oficial o hacer una propia.*

Hace algunos años los pronósticos del tiempo acostumbraban a usar expresiones como estas: lloverá mañana, probablemente lloverá mañana, no se esperan precipitaciones para mañana, etc. Los pronósticos eran frecuentemente equivocados. Pero hoy en día las predicciones tienen diferente lectura: existe un 60% de probabilidades de que

llueva mañana. ¿Qué significa este 60%? ¿Contiene esta afirmación más información que las anteriores predicciones? Quizás, para quienes no saben qué significa la palabra "probabilidad", las predicciones diarias pueden ser algo confusas y dar la impresión de que no son tan precisas como acostumbraban a ser.

Hay un elemento de incertidumbre en la predicción sea cual sea su base. Así, hablando en pura lógica, una predicción sin ninguna indicación de su precisión no es prácticamente útil para poder tomar decisiones. La cifra del 60% en la predicción del tiempo nos da una medida de la exactitud de la predicción. Implica que en las ocasiones en que se hace un pronóstico parecido, lloverá alrededor del 60% de las veces y no lloverá el 40% de las veces. Naturalmente que no es posible decir en qué ocasión particular lloverá. En este sentido, la predicción "hay un 60% de probabilidad de que mañana llueva" es más informativa y lógica de formular que decir categóricamente "mañana lloverá". ¿En qué sentido es esta predicción útil?

Supongamos que hay que decidir si coger un paraguas o no en base al pronóstico del tiempo que dice "hay un 60% de probabilidad de que mañana llueva". Supongamos que la inconveniencia que puede causarle el coger el paraguas cualquier día pueda ser medida en términos monetarios como m doláres y la pérdida que significa mojarse a causa de la lluvia por no llevar el paraguas sea r dólares. Entonces la esperanza de pérdida en dólares bajo las dos posibles decisiones que se pueden tomar cuando la probabilidad de lluvia es del 60%, es como sigue:

Decisión	Pérdida esperada
Llevar paraguas	m
No llevar paraguas	$.6(r) + .4(0) = 6r/10$

Se puede minimizar la pérdida decidiendo llevar un paraguas si $m \leq 6r/10$ y no llevarlo si $m > 6r/10$.

Esta es una simple demostración de cómo la medida de la exactitud o inexactitud de una predicción puede ser usada para sospesar las consecuencias de las posibles diferentes decisiones y escoger la mejor de ellas. No existe una base para tomar una decisión si no se ha cuantificado la incertidumbre asociada a la predicción.

6.6. Sondeos de opinión pública

Tan pronto como me pongo a pensar, me asaltan las dudas.

Oscar Levant

En el pasado, los reyes trataban de conocer la opinión pública usando una red de espías. Probablemente, la información así recogida les ayudaba a configurar la política

pública, decretando leyes y obligando su cumplimiento. La historia de los sondeos para conocer la opinión pública, empezó con la primera publicación de las encuestas Gallup. Actualmente estas consultas han llegado a ser rutinarias en periódicos y otros medios de comunicación jugando un importante papel en los mismos. Recogen la opinión del público en diferentes asuntos políticos, sociales y económicos, publicando resúmenes de los resultados. Estas encuestas de opinión son muy valiosas en los sistemas políticos democráticos. Indican a los líderes políticos y a la burocracia cuáles son las aspiraciones, deseos y necesidades de la sociedad. También son noticia informando a los ciudadanos sobre lo que piensan los demás. Esto puede ayudar a cristalizar la opinión pública en asuntos importantes.

Los resultados de las encuestas de opinión pública se anuncian en un determinado estilo estadístico, que necesita una aclaración. Por ejemplo, las noticias radiadas pueden ser:

> "El porcentaje de población que aprueba la política exterior del Presidente es de 42 con un margen de error de más menos 4 puntos".

En lugar de dar una cifra única como respuesta, se da un intervalo $(42-4, 42+4) = (38,46)$. ¿Cómo se obtiene y cómo se interpreta?

Supongamos que el verdadero porcentaje de adultos en América que aprueba la política exterior del Presidente es un número determinado que llamamos T. Para conocer el número T, es necesario contactar con todos los americanos adultos y lograr sus respuestas a la pregunta, ¿Aprueba la política exterior del Presidente? Esta es una labor imposible si tiene que hallarse una respuesta rápida. Lo mejor que podemos hacer es lograr una estimación, que sea una buena aproximación a T. La información media se logra telefoneando a un determinado número de personas "elegidas al azar" y anotando sus respuestas. Si r de cada p personas contactadas responden diciendo sí, entonces la estimación de T sería $100(r/p)$. Naturalmente, hay algún error en la estimación porque hemos tomado solamente una muestra de la población (una pequeña fracción del número de adultos de EE.UU.). Si se contacta con otro grupo de p individuos, podemos obtener una estimación diferente. ¿Cuál es el error en una estimación determinada? Basándonos en una teoría desarrollada por dos estadísticos, J. Neyman y E.S. Pearson, es posible calcular un número e tal que el verdadero valor de T se encuentre en el intervalo

$$100(r/p) - e, 100(r/p) + e$$

con una alta "probabilidad", normalmente elegida como el 95% (o bien 99%). Esto significa que el suceso tal que el intervalo no cubra el valor real es tan raro como sacar una bola blanca en una extracción al azar de una bolsa que contiene 5 (ó 1) bolas blancas y 95 (99) bolas negras.

La validez de los resultados de las encuestas públicas depende de lo "representativa" que sea la muestra de individuos escogida. Es evidente que el resultado dependerá de la composición de la afiliación política de los individuos escogidos (Republicanos o Demócratas). Aún suponiendo que no haya habido un determinado sesgo al escoger los individuos con respecto a sus afiliaciones políticas, los resultados pueden ser viciados si algunos individuos no responden pero ocurre que pertenecen a un partido político determinado. En cualquier informe, es seguro que habrá un cierto grado de respuestas nulas, y el error que introduce esta circunstancia es difícil de valorar a menos que se disponga de información adicional.

6.7. Superstición y procesos psicosomáticos

Cuando se le preguntó por qué no creía en la astrología, el lógico Raymond Smullyan respondió que era Géminis y que los Géminis nunca creen en la astrología.

Un amigo mío, un buen cristiano, dio a la Iglesia la totalidad del primer mes de su salario en su primer trabajo. Cuando le pregunté si creía en Dios, contestó: "yo no sé si Dios existe o no, pero por si acaso es más seguro creer que Dios existe y actuar de acuerdo con ello". Quizás creencias y supersticiones ocupan un lugar en la vida de cada uno, pero es un peligro cuando son las únicas guías de nuestras actividades.

¿Tienen los procesos psicosomáticos algún efecto en el funcionamiento biológico de nuestro cuerpo? No hay evidencia experimental ni en un sentido ni en otro. No obstante, de vez en cuando, aparecen estudios en apoyo de anécdotas referentes a los efectos de la "mente sobre la materia". En un reciente trabajo, David Phillips de la Universidad de California, San Diego, examinó la proporción de muertes en un periodo de 25 años entre las mujeres chino-americanas mayores de edad alrededor de una fiesta clave, el Festival de la Cosecha. Encontró que los fallecimientos bajaron un 35.1% sobre lo que sería normal una semana antes de la fiesta y subieron un 34.6% una semana después, lo que parece indicar que uno puede ejercer cierto poder para retrasar su muerte hasta después del acontecimiento esperado.

En un estudio previo Phillips (1977) obtuvo datos referidos a los meses de nacimiento y defunción de 1251 americanos famosos y demostró similares resultados. La siguiente Tabla 6.2 facilita los datos conseguidos por Phillips junto con los datos sobre los académicos indios de la Royal Society.

Se aprecia en la Tabla 6.2 que el número de muertes en los meses anteriores es inferior a las ocurridas en los meses durante y después del mes de nacimiento. Este fenómeno es más pronunciado en el caso de las personalidades más famosas. Los datos globalmente considerados parecen indicar que hay una tendencia a retrasar la propia muerte hasta después del cumpleaños.

Tabla 6.2

Número de muertos antes, durante y después del mes de nacimiento

	meses antes						mes de nacimiento	meses después					Total	
	6	5	4	3	2	1		1	2	3	4	5		*p*
Muestra 1	24	31	20	23	34	16	26	36	37	41	26	34	348	.575
Muestra 2	66	69	67	73	67	70	93	82	84	73	87	72	903	.544
Muestra 3	0	2	1	9	2	2	3	2	0	1	3	2	18	.611

p = Proporción de fallecimientos durante y después del mes de nacimiento.

Muestra 1. Famosos citados en "Cuatrocientos Notables Americanos".

Muestra 2. Las personas citadas bajo esta categoría forman parte de las más importantes familias que publican los 3 volúmes de "Who is Who"[2] de los años 1951–60, 1943–50 y 1897–1942.

Muestra 3. Académicos indios de la Royal Society fallecidos.

¿Indican estos estudios que algunas personas pueden emplear su fuerza de voluntad para retardar la fecha de su muerte hasta que ocurre un acontecimiento importante, tal como un nacimiento, un festival o un aniversario? Un famoso ejemplo señalado en este sentido es el de Thomas Jefferson; se sabe que demoró su muerte hasta el 4 de julio de 1826 —exactamente 50 años después de la firma de la Declaración de Independencia— sólo después de preguntar a su doctor, "¿Es el 4?"

Estudios aislados como el publicado de David Phillips no necesariamente cuentan toda la historia. En el trabajo de investigación, es normal que el mismo problema sea estudiado por un gran número de investigadores y sólo aquellos resultados positivos, quizás por azar, son publicados y conocidos. Los que indicasen resultados negativos no son generalmente publicados y permanecen archivados, una situación a la que nos referimos como el "problema del cajón de los archivos". Por tanto, hace falta mucha precaución en aceptar los resultados publicados y sacar conclusiones de los mismos.

6.8. La Estadística y la Ley

> *Las leyes no son generalmente entendidas por tres clases de personas: aquellas que las hacen, aquellas que las ejecutan y aquellas que las sufren si las transgreden.*
>
> Halifax

Es importante no sólo que se haga justicia si no que además lo parezca.

[2] "Quién es Quién". N. del T.

Durante la última década, los conceptos y métodos estadísticos han jugado un importante cometido en resolver complejas situaciones referidas a casos de derechos civiles. Ejemplos típicos son las disputas de paternidad, alegaciones de discriminación contra grupos minoritarios a la hora de obtener empleo y oportunidades de encontrar vivienda, regulación del entorno y seguridad, y protección del consumidor contra anuncios engañosos. En todos estos casos, los argumentos se han basado en datos estadísticos y su interpretación. Un juez tiene que determinar la credibilidad de la evidencia cuantitativa que se le haya presentado y decidir sobre la responsabilidad legal en cada caso, así como la apropiada compensación. Este proceso precisa que todas las partes interesadas, todos los implicados en el pleito, abogados de cada parte, y, quizás lo más importante, los jueces que tienen que decidir, posean conocimientos de estadística y de los peligros más habituales en su uso y su interpretación.

Consideremos el caso de *Eison contra la ciudad de Knoxville,* en el que una mujer candidata a la Academia de Policia de Knoxville, reclamó que un test de fuerza y resistencia usado por la Academia era discriminatorio contra el sexo femenino. Como evidencia, Eison facilitó los resultados del test en su curso.

Tabla 6.3

Proporción de aprobados entre los aspirantes de un curso

	Aprobados	Suspendidos	porcentaje de aprobados
Mujeres	6	3	.666
Hombres	34	3	.919
Total	40	6	.870

Tabla 6.4

Proporción de aprobados entre los aspirantes de toda la Academia

	Aprobados	Suspendidos	% de aprobados
Mujeres	16	3	.842
Hombres	64	3	.955
Total	80	6	.930

Ella argumentó que la regla de los 4/5 de la EEOC (Comisión para la igualdad de oportunidades de empleo) se violaba ya que la relación .666/.919 = 0.725 era bastante

menor que los 4/5 = .8. El juez pidió los resultados globales de la Academia, que fueron los siguientes[3]:

En este caso, la relación es (.842)/(.955) = .882 > .8. El juez dijo sensatamente que lo que era relevante era la "totalidad de las personas" al hacer un test y no un subconjunto particular de las mismas. Este es un ejemplo típico donde las partes interesadas tratan de seleccionar una parte de los datos que parecen diferir de la totalidad de los mismos, aplicándola a su caso específico.

A menudo, la evidencia cuantitativa es expresada en forma de un promedio o una proporción, basado en una encuesta sobre una pequeña parte de los individuos de una población, acerca de una medición particular u opinión. ¿Representaba la cifra indicada la característica particular de la totalidad de la población? Depende mucho de que el número de individuos relacionados sea el adecuado y de la ausencia de desviaciones en su selección.

El dar por buenas las estimaciones muestrales de una población precisa de un cuidadoso examen del procedimiento seguido al llevar a cabo la encuesta, de cómo asegurar la representatividad de la muestra y usar un tamaño muestral adecuado para asegurar un determinado grado de precisión en las estimaciones resultantes. La justicia estaría mejor servida si los jueces tuviesen algún conocimiento de la metodología aplicable en las encuestas, con el fin de que pudiesen facilitarles la decisión, en cada caso individual, de si aceptar o rechazar unas estimaciones muestrales. No es que estemos sugiriendo que un juez deba ser un estadístico cualificado, pero sería ventajoso para un juez que tuviera algún conocimiento de la inferencia estadística y de la incertidumbre que conlleva tomar una decisión, para así poder formarse una opinión independiente basada en los argumentos estadísticos que le sean presentados.

Cualquier juicio envuelve la evaluación de la probabilidad de que un determinado suceso sea verdadero, dadas todas las evidencias, tomar una decisión y considerar las consecuencias de condenar a una persona inocente y de no llegar a condenar a un culpable. Los frases habituales para expresar verbalmente varios grados de probabilidad son como las siguientes:

(1) el predominio de las evidencias;

(2) evidencia clara y convincente;

(3) evidencia clara, inequívoca y convincente;

(4) prueba más allá de una duda razonable;

[3]Las diferencias entre las proporciones de hombres y mujeres, respecto a una cierta propiedad, al considerar un resultado parcial y el resultado globlal, se conoce como paradoja de Simpson. N. del T.

Con el fin de determinar cómo interpretan los jueces generalmente estos criterios para calificar las pruebas, el juez Weinstein estudió a sus compañeros en los tribunales del distrito, cuyas probabilidades expresadas en porcentajes se detallan en la siguiente tabla.

Tabla 6.5

Probabilidades asociadas con varios estandares de calificación de las pruebas según los jueces del distrito Este de Nueva York

Juez	Preponderancia (%)	Claro y convincente (%)	Claro, inequívoco y convincente (%)	Prueba más allá de una duda razonable (%)
1	50+	60-70	65-75	80
2	50+	67	70	76
3	50+	60	70	85
4	51	65	67	90
5	50+	Criterio esquivo y poco útil		90
6	50+	70+	70+	85
7	50+	70+	80+	95
8	50.1	75	75	85
9	50+	60	90	85
10	51	No puede estimarse numéricamente		

Fuente: U.S.v. Fatico 458 F. Supp. 388 (1978) pág. 410.

Se aprecia que existe unanimidad en el orden creciente de las probabilidades otorgadas, para los cuatro criterios (1)-(4) descritos. No obstante, hay alguna variación entre los jueces en las probabilidades asignadas al más alto grado de certeza.

De hecho, existe una sofisticada técnica estadística, el método de Bayes, según el cual la *probabilidad a priori* de que un individuo sea culpable, según el juez, puede ser actualizada usando las evidencias disponibles con un grado determinado de credibilidad. Esta probabilidad condicionada a las evidencias disponibles se llama *probabilidad a posteriori* y constituye el principal dato al tomar decisiones. Parece que

la teoría Bayesiana para la toma de decisiones, tal como se desarrolla en estadística, proporciona una base objetiva para administrar justicia.

6.9. Percepción extrasensorial y coincidencias asombrosas

> *El universo está más gobernado por probabilidades estadísticas que por lógica. Pero esto lo hace todavía más maravilloso. Si la vida es como obtener seiscientas veces seguidas el mismo resultado en un juego de azar, sabemos que no es probable que ésto suceda más que una sola vez en muchos siglos, pero también sabemos que ello puede ocurrir en esta habitación, esta noche, sin perturbar el frágil orden cósmico. Ello resulta tranquilizador.*
>
> G.K. Chesterton

De vez en cuando nos llegan estudios sobre individuos que poseen percepción extrasensorial (PES) con la habilidad de leer la mente de otros, astrólogos que hacen predicciones exactas y coincidencias asombrosas, como que alguien gane a la lotería dos veces en cuatro meses. Estos acontecimientos son noticia y quizás resultan interesantes de leer. ¿Sugieren la existencia de poderes ocultos que los causan?

Es quizás poco prudente descartar completamente la posibilidad de que existan ciertos individuos con extraordinarias habilidades (como PES), o que las posiciones de los planetas en el momento del nacimiento determinen el curso de los acontecimientos de la vida de un individuo. No obstante, el anuncio de historias afortunadas, a menudo sobre una base selectiva, no nos proporciona una gran evidencia a favor de tales posibilidades.

Basta considerar, por ejemplo, un experimento típicamente extrasensorial, donde se pide a una persona que adivine cuál de los dos posibles objetos con los que se experimenta, ha sido escogido y puesto debajo de una carpeta. La posibilidad de que un individuo acierte con todas las respuestas correctas en cuatro pruebas repetidas, por puro azar, es 1/16. Esto significa que si 64 individuos de una población arbitraria son puestos a prueba, hay una alta probabilidad de que haya entre 3 y 4 individuos que contesten correctamente. Este experimento no sugiere que estos 3 ó 4 individuos tengan PES. No obstante, si sólo se publicasen tales logros, atraerían nuestra atención.

Consideremos otro ejemplo. Si se está en una fiesta con al menos 23 personas y les preguntamos sobre sus fechas de nacimiento, podemos encontrar a 2 de ellos con la misma fecha. Esto podría parecer una coincidencia asombrosa, pero los cálculos probabilísticos demuestran que esta circunstancia ocurre con una probabilidad del 50%.

En un artículo publicado en el *Journal of the American Statistical Association* (Vol. 84, pp. 853-880), dos profesores de la Universidad de Harvard, Diaconis y Mosteller, demuestran que la mayor parte de las coincidencias, hechos que pueden parecer asombrosos, son sucesos que tienen una probabilidad razonable de ocurrir de vez en cuando.

Existe una ley estadística que indica que con un tamaño muestral suficientemente grande, cualquier suceso, aunque sea pequeña la probabilidad de que suceda en un ensayo aislado, acabará ocurriendo. Puede ocurrir en cualquier momento sin podérsele atribuir ninguna causa especial.

6.10. Difundamos la terminología estadística

Deseo que él quisiera aclarar su explicación.

Lord Byron.

Estudiamos las 3 reglas de lectura, escritura y aritmética en la escuela. Todo esto no es suficiente. Hay una gran necesidad de saber cómo manejar las situaciones de incerditumbre. ¿Cómo tomaremos una decisión cuando no tenemos suficiente información? Esto ha desconcertado a los filósofos durante los siglos pasados. Ahora tenemos un camino lógico para permitir la incertidumbre en la toma de decisiones. Lo podemos llamar la cuarta regla, razonamiento (de tipo inductivo) a partir de insuficientes premisas. Se deberían hacer intentos para introducir la cuarta regla en una etapa temprana del curriculum escolar. Esto puede llevarse a cabo a través de ejemplos de sucesos impredecibles de la naturaleza, variabilidad entre los individuos y errores de medición, y explicando lo que podemos aprender a partir de los datos observados o de la información obtenida en dichas situaciones.

También deberíamos intentar explorar la posibilidad de usar los medios de comunicación, los periódicos, la radio y la televisión, para educar continuamente al público sobre las consecuencias de acciones tomadas por el Gobierno y los hallazgos de los científicos. Esto precisa de periodistas con la habilidad necesaria para interpretar informaciones estadísticas y así poder dar noticia sobre ellas en un sentido imparcial. Sin lugar a dudas, los nuevos periodistas tienen algunas limitaciones. Tienen que escribir historias en el sentido de que no ofendan a la clase dirigente y que a la par sean lo bastante sensacionalistas para ser aceptadas por los editores para su publicación. Pueden no tener la experiencia para formar un juicio independiente y prefieren resumir lo que los expertos desean fomentar. Quizás, existe la necesidad de formar a periodistas capacitándolos para escribir trabajos sobre temas estadísticos. Tengo entendido que el profesor F. Mosteller de la Universidad de Harvard imparte periódicamente cursos de estadística para formar periodistas que puedan escribir imparcialmente sobre materias estadísticas y en un sentido comprensible para el público.

Esta es una loable tentativa y deberían llevarse a cabo esfuerzos para introducir, en las universidades, cursos ordinarios de estadística orientados a escritores sobre temas científicos.

6.11. La estadística como una tecnología clave

En el pasado, la economía de un país dependía de lo bien preparado que se encontrara para la guerra. Estamos presenciando hoy en día una transformación de las amenazas y la confrontación hacia la conciliación y negociación. El mayor problema de las próximas décadas para cualquier país no será el desafío de la guerra sino el de la paz. El campo de batalla del futuro será el económico y el bienestar social. En donde deberemos luchar va a ser contra el hambre y las privaciones que afligen la sociedad. Parece que no estamos plenamente preparados para tal cometido. Nuestro éxito dependerá de que se consiga y procese la información necesaria para tomar una decisión óptima gracias a la cual los recursos disponibles, tanto en el terreno personal como en el material, sean explotados al máximo para mejorar el nivel de vida de los ciudadanos. Esto debe ser llevado a cabo de un modo cuidadoso con el fin de asegurar lo siguiente:

- El progreso sea equitativo y sostenible.
- No deben causarse daños irreversibles a la biosfera.
- No haya contaminación moral (o degradación de los valores humanos).

La Estadística podría ser la clave tecnológica para lograr esta revolución, una tecnología para dar forma a un nuevo mundo a través de la paz.

Referencias

Cohen, B. y Lee, I.S. (1979). "A catalog of risks". *Health Physics,* **36**, 707-722.

Diaconis, P. y Mosteller, F. (1989). "Methods for studying coincidences". *J. Amer. Statist. Assoc.,* **84**, 853-880.

Phillips, D.P. (1977). "Deathday and birthday: An unexpected connection". En *Statistics: A Guide to Biological and Health Sciences* (Eds. J.M. Tanur *et al.*) pp. 111-125, Holden Hay Inc., San Francisco.

Autorització de domiciliació bancària per al pagament de les subscripcions anuals de la revista **Qüestiió**

El sotasignat _____

autoritza el Banc/Caixa _____

Agència núm. _____ adreça _____

Codi postal _____ ciutat _____

a abonar a la revista **Qüestiió** amb càrrec al meu compte número _____

_____, les subscripcions a **Qüestiió**.

_____ , a _____ d _____ de 19 _____

(Signatura)

El sotasignat _____

autoritza la revista **Qüestiió** a carregar al meu compte número _____

_____ , al Banc/Caixa _____

_____

Agència núm. _____ adreça _____

Codi postal _____ ciutat _____

l'import de les tarifes vigents de les subscripcions a la revista **Qüestiió.**

_____ , a _____ d _____ de 19 _____

(Signatura)

Butlleta de subscripció a la revista **Qüestiió**

Nom i cognoms _____

_____

Empresa/Institució _____

_____

Adreça _____

_____

Codi postal _____ ciutat _____

Tel. _____ Fax _____

Data d'expedició _____

Signatura: DNI o NIF _____

Desitjo subscriure'm a **Qüestiió** per a l'any 1995.
El preu de la subscripció és de 2.700 PTA.

Forma de pagament

☐ Transferència al compte de la Caixa de Catalunya número 6985/77, Agència 100, Comte d'Urgell 162, 08036 Barcelona

☐ Domiciliació bancària

☐ Taló nominatiu a l'Institut d'Estadística de Catalunya

☐ Gir postal

☐ En efectiu

Retornar aquesta butlleta (o una fotocòpia) a:

Qüestiió:
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona