Qüestiió

Quaderns d'Estadística i Investigació Operativa

Any 2001, volum 25, núm. 3
Segona època

Entitats patrocinadores:

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

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 2001, volum 25, núm. 3

SUMARI

EDITORIAL

El tercer número del volum 25, corresponent a l'any 2001, recull l'edició d'un total de vuit articles, repartits entre tres de les quatre seccions temàtiques de la revista, acompanyats de dues recensions a l'apartat «Comentaris de llibres» i de la ressenya habitual de novetats editorials en matèria estadística, publicades per l'Idescat i la resta de la Generalitat de Catalunya al llarg de l'any. Com ja s'ha fet en ocasions anteriors, *Qüestiió* promou la publicació de ponències que s'han donat a conèixer en fòrums i seminaris en el nostre país, com fou el cas del «Tercer Congrés Europeu de Matemàtiques» (10-14 juliol 2000). En aquest número, a més de continuar l'edició d'alguns d'aquests treballs, s'hi publiquen quatre ponències convidades a les *Jornades internacionals sobre Data Mining* (Barcelona, 14-15 desembre 2000) organitzades per la Xarxa Temàtica «Enquestes i Qualitat de la Informació Estadística», els quals es distribueixen entre la secció d'«Estadística» i la secció d'«Estadística Oficial».

D'altra banda, *Qüestiió* es complau a anunciar als seus lectors i subscriptors que, coincidint amb els 25 volums editats des de 1977, en el decurs de l'any 2002 s'iniciarà una tercera època de la revista. Els objectius bàsics d'aquesta nova etapa se centren en una millora qualitativa dels continguts a fi que esdevingui una revista més competitiva com per situar-la en els estàndards internacionals més exigents, tot augmentant el seu reconeixement acadèmic i l'impacte entre la comunitat científica aplegada en l'estadística aplicada i la investigació operativa. Per tal d'avançar en aquesta nova orientació, les normes d'acceptació d'articles sotmesos a *Qüestiió* es modifiquen en els termes que ja recull aquest número, amb la publicació dels originals acceptats exclusivament en anglès, per bé que s'acompanyaran d'un breu resum dels mateixos en català. Des del punt de vista organitzatiu, el lector també advertirà un canvi en la direcció executiva de la revista, en virtut del qual el Dr. Michael Greenacre, catedràtic de la Universitat Pompeu Fabra, substitueix el Dr. Tomàs Aluja, de la Universitat Politècnica de Catalunya, en qualitat d'editor executiu de *Qüestiió*.

Per últim, com ja és habitual en el darrer número de cada volum, es presenta a continuació l'evolució d'articles en el decurs de l'interval gener-desembre del 2001, el qual ha enregistrat el moviment següent:

- articles sotmesos: 27
- articles en procés d'avaluació: 20
- articles acceptats: 27 (20 publicats i 7 en espera de publicació)
- articles rebutjats: 9

En relació a les dades anteriors, és interessant constatar que el nombre d'articles sotmesos en aquest darrer any se situa justament en la mitjana del període 1992-2001 (2a època), la qual cosa permet apreciar una certa consolidació del flux anual d'originals rebuts a *Qüestiió*, a l'entorn dels 25-30 articles. També val la pena remarcar que, amb

les xifres del 2001, el temps d'acceptació d'un article en els darrers anys se situa en una mitjana de 9,2 mesos, reduint desviacions enregistrades en els anys 1997 i 1998, mentre que el percentatge de rebuig d'articles sotmesos es manté a l'entorn del 25-35% dels darrers exercicis. Finalment, respecte de la difusió electrònica dels continguts de *Qüestiió*, tot indica que en el decurs del 2001 continua accelerant-se el ritme de consultes al seu website, de manera que el nombre d'accessos han superat les 135.000 peticions http, gairebé triplicant les 56.000 consultes al web en el decurs de l'any 2000.

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

En aquest tercer número del volum 25 s'hi publiquen vuit articles, sis d'Estadística, un d'Investigació Operativa i un d'Estadística Oficial. Dins la secció «Estadística», el primer article titulat *Una generalización de los procesos estocásticos Log-Normal y de Gompertz como procesos de Itô*, de J. Gómez i F. Buendía, és un estudi d'un procés de difusió que estén el model logarítmic-normal i el model de creixement de Gompertz, plantejant el problema com una equació diferencial d'Itô, la qual cosa permet els autors obtenir més propietats que utilitzant una metodologia clàssica. En el segon article, *Detección de M señales gaussianas utilizando el desarrollo modificado de un proceso estocástico*, de J. Navarro i J. C. Ruíz, es proposa una metodologia per a la detecció de senyals que utilitza un desenvolupament modificat del procés estocàstic, que convergeix al procés, però diferent del desenvolupament de Karhunen-Loève i que, per tant, no necessita el dificultós càlcul d'autovalors i d'autofuncions de la funció de covariància, amb interessants conseqüències pràctiques. Pel que fa al tercer article, *Estimación no paramétrica de la función de riesgo: aplicaciones a sismología*, de A. Quintela i G. Estévez, s'obté una estimació de la funció de risc sense fer cap hipòtesi sobre la distribució de la variable (és a dir, una estimació no paramètrica), ni sobre la independència dels valors de la mostra. L'estimador que proposen els autors és tipus nucli, amb una adequada selecció de la finestra, la qual cosa presenta alguns avantatges sobre altres estimadors, com ara el basat en la funció de distribució empírica. Els autors justifiquen el mètode amb algunes simulacions i una aplicació a dades de sismologia. A continuació, i dins la mateixa secció, es publiquen tres articles més provinents d'una selecció de ponències presentades a les *Jornades internacionals sobre Data Mining*. En aquest context, el primer article, *La Minería de Datos, entre la estadística y la inteligencia artificial*, de T. Aluja, exposa el problema de com buscar i resumir els aspectes interessants d'un banc de dades, analitzant els fonaments estadístics d'aquesta nova tècnica, comparant nomenclatures, fent una panoràmica de l'abast actual i donant una visió de futur, a més d'un exemple de la seva aplicabilitat força il·lustratiu. A continuació, l'article

de L. Garrido i J. I. Latorre, *Aplicaciones empresariales de Data Mining*, presenta les idees bàsiques de la mineria de dades insistint, sobretot, des de l'àmbit de les xarxes neuronals i comentant dos exemples de predicció borsària i de propagació del foc en cables elèctrics. En tercer lloc, l'article *Practical Data Mining in a large utility company*, a càrrec de G. Hébrail, il·lustra la utilitat del Data Mining per a l'estudi de corbes de consum d'electricitat, presentant la transformació simbòlica de corbes, i la predicció de la demanda, que permet tractar grans bases de dades i problemes de dades faltants.

L'únic article a la secció «Investigació Operativa» es titula *Secuenciación dinámica de fabricación flexible mediante aprendizaje automático: análisis de los principales sistemas de secuenciación existentes*, de P. Priore, D. de la Fuente, J. Puente i A. Gómez, on s'hi comenta la necessitat d'obtenir un mètode que permeti seleccionar a cada moment, i mitjançant aprenentatge automàtic, la regla més adequada de seqüenciació dels treballs en els sistemes de fabricació. Els seus autors fan una revisió quasi exhaustiva dels sistemes de seqüenciació existents i proposen algunes idees per millorar aquests mètodes.

A continuació, la secció «Estadística Oficial» prossegueix amb la publicació selectiva de ponències presentades a les esmentades jornades sobre mineria de dades de l'any 2000, amb l'article *El cens de població: un assaig d'interpretació mitjançant Data Mining*, de C. Guisande i F. Subirada. En aquest treball, els autors exposen l'experiència conjunta d'Idescat, Cesca i IBM en l'aplicació de tècniques de segmentació i d'associació a les extenses bases de dades censals sobre individus i llars catalanes; entre d'altres conclusions, en destaca l'efecte que pot tenir la simultaneïtat en el tractament de les variables censals en l'estratègia convencional d'anàlisis temporals i territorials al si de l'estadística oficial.

Comentari d'altres seccions i apartats

La «Secció docent i problemes» inclou la presentació successiva de nous enunciats i la resolució dels problemes publicats en el número immediatament anterior. Seguidament, la secció «Comentaris de llibres» acull una detallada ressenya de F. Udina sobre el voluminós llibre *Statisticians of the Centuries*, de C. C. Heyde i E. Seneta, basat en la recopilació biogràfica de 103 estadístics nascuts abans del segle XX, a partir d'una iniciativa de l'ISI, i on s'hi destaca la transversalitat de l'estadística en tots els àmbits del coneixement científic. En segon lloc, E. Bonet comenta el manual *Probabilitats*, de la Dra. Marta Sanz, en la qual es posa de manifest la solidesa de la seva exposició sobre l'estadística inferencial i el càlcul probabilístic, així com l'encert en l'exposició i rendiment dels exemples que s'hi inclouen.

El darrer apartat, dedicat a «Ressenyes d'activitats institucionals», inclou, com ja és habitual, una revisió actualitzada d'activitats de la Sociedad Española de Biometría, amb l'anunci dels cursos monogràfics organitzats en col·laboració amb altres entitats. En segon lloc, s'ofereix una nova recensió del «Training for European Statisticians Institute» que *Qüestiió* redifon sistemàticament des del 1997, amb la relació dels cursos del *Core Programme 2002* que s'impartiran de gener a novembre del 2002, adreçats principalment als membres dels òrgans d'estadística oficial en l'àmbit comunitari. D'entre els cursos previstos s'inclou un anunci específic al curs sobre les tècniques d'estimació de petites àrees, *Estimation for small areas*, que s'impartirà a la seu de l'Idescat el mes de març del 2002, i un primer anunci sobre el Workshop que, sota el títol de *Research and Training in Failure Time Methods in the New Millennium* (Barcelona, 12-14 juny 2002), la Universitat Politècnica de Catalunya organitza amb la col·laboració, entre d'altres, de l'Idescat i patrocinadors de *Qüestiió*. El número conclou, com és habitual, amb la relació de novetats editorials de l'Institut d'Estadística de Catalunya i de la resta de la Generalitat de Catalunya que han estat publicades en el decurs de l'any 2001 en l'àmbit de l'estadística.

Carles M. Cuadras, director executiu
Enric Ripoll, editor executiu

Estadística

QÜESTIIÓ, vol. 25, 3, p. 393-414, 2001

UNA GENERALIZACIÓN DE LOS PROCESOS ESTOCÁSTICOS LOG-NORMAL Y DE GOMPERTZ COMO PROCESOS DE ITÔ

JUAN GÓMEZ GARCÍA
FULGENCIO BUENDÍA MOYA*

Estudiamos una ecuación diferencial estocástica de Itô que es una generalización de los modelos estocásticos logarítmico-normal y de Gompert. Reducimos la ecuación mediante una transformación de cambio de estado a otra que resulta una generalización de la ecuación de Langevin, que rige el proceso de Uhlenbeck-Ornstein. A partir de la expresión analítica de las soluciones de ésta y de la original estudiamos las características estadísticas de ambos procesos solución, en particular los momentos de las distribuciones finito dimensionales, sus funciones de densidad de transición, las distribuciones límite y las condiciones de estacionariedad, obteniendo que la expresada generalización del proceso de U-O es el único proceso Gaussiano, Markoviano y estacionario no centrado en tiempo continuo. Por otra parte, se establece que las potencias del proceso lognormal-Gompertz generalizado satisfacen una E.D.E. del mismo tipo.

A generalization of the log-normal and Gompertz stochastic processes as Itô processes

Palabras clave: Ecuación diferencial estocástica, ecuaciones de Kolmogorov, proceso log-normal, proceso de Gompertz, proceso de Uhlenbeck-Ornstein, ecuación de Langevin

Clasificación AMS (MSC 2000): 60H10

* Departamento de Métodos Cuantitativos para la Economía. Facultad de Economía y Empresa. Universidad de Murcia. 30100 Murcia, España.

– Recibido en marzo de 2000.
– Aceptado en junio de 2001.

1. INTRODUCCIÓN

Los procesos estocásticos de difusión han sido profusamente empleados en las dos últimas décadas para analizar el comportamiento de fenómenos económicos, sociales, biológicos, médicos, etc. [véase, p.e., los libros de Bharucha-Reid (1960), Bartolomew (1973), Wong-Hajek (1985), Sobczyk (1991), Todorovic (1992), Ricciardi (1977), Malliaris and Brock (1982), Sengupta (1986), Gutierrez-Valderrama (1994), McShane (1994), Kijima (1997), etc.]. En particular, los procesos de difusión log-normales, unidimensionales y multidimensionales, se han aplicado en el estudio y modelización de variables económicas. Tintner and Sengupta (1972) consideran este tipo de procesos como gobernadores de aquellos fenómenos en que podemos suponer que la tendencia media de la evolución de la variable en periodos cortos de tiempo (media infinitesimal del proceso) y la desviación cuadrática media de la variable en periodos cortos de tiempo (varianza infinitesimal del proceso) son proporcionales al valor de la variable. Así, Tintner and Thomas (1963), Tintner and Patel (1966), Tintner y Bello (1968), Tintner and Sengupta (1972) y Moreno Bas (1974), por ejemplo, presentan el proceso definido por su función de densidad de transición o por sus coeficientes de tendencia y de difusión.

Por otra parte el modelo estocástico de Gompertz ha sido utilizado para el estudio del crecimiento de ciertas poblaciones o de la difusión de innovaciones tecnológicas y de nuevos productos en el mercado [ver Skiadas, Giovanis and Dimoticalis (1993)].

La presentación de los modelos estocásticos de difusión a partir de los coeficientes de tendencia y de difusión precisa de un análisis acerca de la existencia de un tal proceso y, sobre todo, de la confirmación de que la función de densidad de transición puede obtenerse como solución de alguna de las ecuaciones de Kolmogorov asociadas al proceso. Por otra parte, las ecuaciones de Kolmogorov, ecuaciones diferenciales en derivadas parciales de tipo parabólico, son en general de dificil resolución y se requieren a menudo técnicas especiales de aproximaciones numéricas [ver Bouleau (1988) o Todorovic (1992), p.e.]. En realidad estas ecuaciones han sido resueltas explícitamente sólo en unos pocos casos simples [ver Bharucha-Reid (1960), Arnold (1974), Bhattacharya-Waymire (1990) ó Sobczyk (1991), p.e.]. A este respecto hay que destacar el trabajo de Ricciardi (1976) sobre la posibilidad, en determinados casos muy particulares, de reducir las ecuaciones de Kolmogorov a las de un proceso Wiener, y el de Gutiérrez *et al.* (1997) acerca de la construcción de funciones de densidad de transición de ciertas extensiones no necesariamente homogéneas de un proceso de difusión. En cualquier caso, el conocer únicamente los momentos infinitesimales del proceso o la función de densidad de transición, limita las posibilidades de estudio del proceso.

En este trabajo consideramos un proceso de difusión que es una extensión de los modelos logarítmico-normal (o de crecimiento exponencial o Malthusiano) y de crecimiento de Gompertz, a través de la solución de una ecuación diferencial estocástica de Itô. De

esta manera, hacemos todo el estudio del proceso a partir exclusivamente de la expresión analítica de la solución de la ecuación. En particular, obtenemos las expresiones de los momentos de las distribuciones finito dimensionales del proceso, a diferencia de los trabajos vía ecuaciones de Kolmogorov, que no disponen de la expresión del proceso y sólo dan los de la distribución unidimensional o variable general del proceso. Una recopilación de estos modelos, para poblaciones, puede verse, entre otros muchos trabajos, en Ricciardi (1977) y (1986). Entendemos que este estudio de una generalización de los procesos lognormal y de Gompert vía ecuaciones de Itô mejora en cuanto a rigor de método y amplitud de resultados a los que se han realizado anteriormente vía ecuaciones de Kolmogorov. El trabajo se completa con un estudio de las potencias de los procesos log-normales, generalizados y estrictos, utilizando igualmente la E.D.E. que satisfacen, del que no tenemos conocimiento que existan tratamientos anteriores similares.

2. PLANTEAMIENTO DE LA E.D.E. EXISTENCIA Y UNICIDAD DE SOLUCIONES

Sea la ecuación diferencial estocástica (E.D.E.) de Itô

$$(2.1)\qquad \begin{aligned} &dX(t) = [a - b\log X(t)]X(t)\,dt + \sigma X(t)\,dW(t) \quad t \in [0,\infty) \\ &\text{con } X(0) = X_0 \in \mathbb{R}^+;\ a,b,\sigma \text{ constantes} \end{aligned}$$

donde $X(t)$ es un proceso con valores en $(0,\infty) = \mathbb{R}^+$ y, naturalmente, $W(t)$ un proceso de Wiener unidimensional estándar.

Se tiene aquí, pues, que las funciones coeficientes drift y de martingala de la ecuación son, respectivamente

$$m(t,x) = m(x) = (a - b\log x)x \qquad \sigma(t,x) = \sigma(x) = \sigma x$$

En la E.D.E. (2.1), para $b = 0$ aparece la E.D.E. Malthusiana, que rige el proceso logarítmico-normal, y para $a = 0$ se tiene el modelo estocástico de Gompertz [ver, p.e., Skiadas, Giovanis and Dimoticalis (1994)].

Los coeficientes de la E.D.E. (2.1) cumplen las condiciones que permiten plantear esta ecuación [ver, p.e., Wong and Hajek (1985)], ya que $\forall T \in [0,\infty)$ $\forall x \in (0,\infty)$ tenemos

$$\int_0^T |m(t,x)|dt = \int_0^T |a - b\log x|\,|x|\,dt = |a - b\log x|\,|x|\,T < \infty$$

$$\int_0^T |\sigma(t,x)|^2 dt = \int_0^T \sigma^2 x^2 dt = \sigma^2 x^2 T < \infty$$

En cuanto a probar la existencia de solución global única para (2.1), observaremos primero que esta ecuación es atípica en el sentido de que su coeficiente drift $m(t,x) = (a - b\log x)x$ sólo está definido para $x > 0$, y no sabemos si se le puede aplicar el resultado clásico del teorema de existencia y unicidad de soluciones, donde siempre se supone $m(t,x)$, $\sigma(t,x) : [t_0,T] \times \mathbb{R} \to \mathbb{R}$, $[t_0,T] \subseteq \mathbb{R}_0^+$, medibles, etc. Por tanto, en vez de aplicar directamente dicho teorema a (2.1), procederemos aplicando a esta ecuación una transformación de *cambio de estado* [ver, p.e. Bhattacharya-Waymire (1990), pg. 382 ó Ikeda-Watanabe (1989), pgs. 197 y siguientes]. Así, definimos la variable $Y(t) = \log X(t)$ y aplicando el lema de Itô obtenemos

$$dY = \frac{dX}{X} - \frac{1}{2}\left(\frac{dX}{X}\right)^2$$

que junto con (2.1) nos da

$$dY = \left(a - \frac{1}{2}\sigma^2 - bY\right)dt + \sigma dW$$

pues $\left(\frac{dX}{X}\right)^2 = [(a - b\log X)\,dt + \sigma dW]^2 = \sigma^2 dt$ (aplicando las reglas de cálculo estocástico). Entonces, para $Y(t)$ tenemos la E.D.E.

$$\text{(2.2)} \qquad \begin{aligned} &dY = (\mu - bY)\,dt + \sigma dW \text{ con } \mu = a - \frac{1}{2}\sigma^2 \\ &\text{con c.i. } Y(0) = \log X(0) = Y_0 \in \mathbb{R} \end{aligned}$$

Para esta ecuación (2.2) es sencillo comprobar la verificación de las condiciones del teorema de existencia y unicidad de soluciones de una E.D.E. autónoma o tiempo-independiente [ver, p.e. Arnold (1974), pgs. 152-153], así que existe una solución global única $Y(t)$ de (2.2) que es un proceso de difusión. Resulta de esto que $X(t) = e^{Y(t)}$ es un proceso de difusión (teorema de cambio de estado) que, evidentemente, verifica (2.1). Probamos así la existencia de procesos de difusión solución de la E.D.E. (2.1). La unicidad de las soluciones de (2.2) determina la de las soluciones de (2.1). (Vemos que la aplicación de la transformación $Y(t) = \log X(t)$ antes de saber si $X(t)$ existe debe entenderse en sentido condicional, es decir suponiendo que $X(t)$ existe, lo que será cierto si y sólo si $Y(t)$, solución de (2.2), existe).

Llegamos, pues, a que la E.D.E. (2.1) tiene una única solución global $\{X(t), t \geqslant 0\}$ que es un proceso de difusión con coeficientes de tendencia y de difusión

$$A_1(t,x) = m(t,x) = (a - b\log x)x \quad \text{y} \quad A_2(t,x) = \sigma^2(t,x) = \sigma^2 x^2$$

respectivamente, y cuya función de densidad de transición $p'(x,t/x_0,t_0)$ verifica, y es su única solución, la ecuación atrasada de Kolmogorov [ver Ikeda-Watanabe (1989), pg. 215]:

$$\frac{\partial p'(x,t/x_0,t_0)}{\partial t_0}+(a-b\log x_0)x_0\frac{\partial p'(x,t/x_0,t_0)}{\partial x_0}+\frac{1}{2}\sigma^2x_0^2\frac{\partial^2 p'(x,t/x_0,t_0)}{\partial x_0^2}=0$$

$$p'(x,t/x_0,t_0)=\delta(x-x_0),t>t_0 \tag{2.3}$$

3. ESTUDIO DEL PROCESO SOLUCIÓN DE LA E.D.E.

Para estudiar el proceso $\{X(t),t\geqslant 0\}$ solución de (2.1), resolveremos primero la ecuación (2.2)

$$dY=(\mu-bY)\,dt+\sigma dW \text{ con } \mu=a-\frac{1}{2}\sigma^2$$
$$\text{con c.i. } Y(0)=\log X(0)=Y_0\in\mathbb{R}$$

[Para $\mu=0$ y $b>0$ es la llamada ecuación de **Langevin** y el proceso solución el de **Uhlenbeck-Ornstein**: ver, p.e., Bouleau (1988)], cuya única solución $\{Y(t),t\geqslant 0\}$, como hemos visto antes, es una difusión con coeficientes de tendencia (o coeficiente drift) y de difusión, respectivamente

$$A_1'(t,y)=\mu-by \qquad A_2'(t,y)=\sigma^2$$

y cuya función de densidad de transición $p(y,t/y_0,t_0)$ verifica la ecuación atrasada de Kolmogorov

$$\frac{\partial p(y,t/y_0,t_0)}{\partial t_0}+(\mu-by_0)\frac{\partial p(y,t/y_0,t_0)}{\partial y_0}+\frac{1}{2}\sigma^2\frac{\partial^2 p(y,t/y_0,t_0)}{\partial y_0^2}=0$$

$$\text{con } p(y,t/y_0,t_0)=\delta(y-y_0),\ t>t_0 \tag{3.1}$$

Consideraremos ahora una ecuación similar a la (2.2) pero con condiciones iniciales más generales, tomando un punto $t_o\in[0,\infty)$ arbitrario:

$$dY=(\mu-bY)\,dt+\sigma dW \text{ con } \mu=a-\frac{1}{2}\sigma^2$$
$$\text{con c.i. } Y(t_0)=\log X(t_0)=C \tag{3.2}$$

donde $Y(t_0)=C$ es una variable aleatoria no degenerada e independiente de $W(t)-W(t_0)$, $t\geqslant t_0$. La solución de esta ecuación nos permitirá encontrar directamente la

función de densidad de transición del proceso $\{Y(t), t \geqslant 0\}$ solución de (2.2), y nos proporcionará también la solución de cualquier otra con condiciones iniciales más particulares, como la propia (2.2).

Establecemos la relación (para $b \neq 0$)

$$d\left\{e^{b(t-t_0)}\left[Y(t)-\frac{\mu}{b}\right]\right\} = e^{b(t-t_0)}b\left[Y(t)-\frac{\mu}{b}\right]dt + e^{b(t-t_0)}dY(t)$$

$$= e^{b(t-t_0)}\{dY(t)+[bY(t)-\mu]dt\} = e^{b(t-t_0)}\sigma dW(t) \tag{3.3}$$

que podemos expresar en su forma integral

$$e^{b(t-t_0)}\left[Y(t)-\frac{\mu}{b}\right] = Y(t_0)-\frac{\mu}{b}+\int_{t_0}^{t} e^{b(s-t_0)}\sigma dW(s) \tag{3.4}$$

o bien

$$Y(t) = \frac{\mu}{b}+e^{-b(t-t_0)}\left[Y(t_0)-\frac{\mu}{b}+\sigma\int_{t_0}^{t} e^{b(s-t_0)}dW(s)\right] \tag{3.5}$$

Si utilizamos ahora el resultado [ver, p.e. Ikeda-Watanabe (1989), pgs. 214-215, ó Arnold (1974), pgs. 146-148] que establece que la probabilidad de transición del proceso solución de (2.2) es la ley de probabilidad del proceso solución de la ecuación que se obtiene al sustituir en (2.2) la condición inicial por $Y(t_0) = y_0$, obtenemos de (3.5) que la densidad de transición $p(y,t/y_0,t_0)$ de $\{Y(t), t \geqslant 0\}$ es la función de densidad de la variable general del proceso

$$Y(t,t_0,y_0) = \tfrac{\mu}{b}+e^{-b(t-t_0)}\left[y_0-\tfrac{\mu}{b}+\sigma\int_{t_0}^{t} e^{b(s-t_0)}dW(s)\right] \tag{3.5$'$}$$

por lo que $Y(t,t_0,y_0)$ es la variable $Y(t)$ del proceso solución de (2.2) condicionada por $Y(t_0) = y_0$: $Y(t)/Y(t_0) = y_0$.

Puesto que la variable $\int_{t_0}^{t} e^{b(s-t_0)}dW(s)$ es $N\left(0, \int_{t_0}^{t} e^{2b(s-t_0)}ds\right)$ [ver propiedades de la integral estocástica, p.e. en Arnold (1974)], resulta de (3.5)$'$ que $Y(t,t_0,y_0)$, es una variable Gaussiana con

$$E[Y(t,t_0,y_0)] = E[Y(t)/Y(t_0) = y_0] = \frac{\mu}{b}+e^{-b(t-t_0)}\left(y_0-\frac{\mu}{b}\right) \tag{3.6}$$

$$Var[Y(t,t_0,y_0)] = Var[Y(t)/Y(t_0)=y_0] = e^{-2b(t-t_0)}\sigma^2\int_{t_0}^{t} e^{2b(s-t_0)}ds$$

$$= \frac{\sigma^2}{2b}e^{-2b(t-t_0)}\left[e^{2b(t-t_0)}-1\right] = \frac{\sigma^2}{2b}\left[1-e^{-2b(t-t_0)}\right] \tag{3.7}$$

Así que la función de densidad de transición $p(y,t/y_0,t_0)$ del proceso $\{Y(t), t \geqslant 0\}$, solución de (2.2) será

$$p(y,t/y_0,t_0) = f_{Y(t)/Y(t_0)}(y/y_0) =$$

$$= \frac{1}{\sqrt{\frac{\pi\sigma^2}{b}\left[1-e^{-2b(t-t_0)}\right]}} \exp\left\{\frac{-b\left[y-\frac{\mu}{b}-\left(y_0-\frac{\mu}{b}\right)e^{-b(t-t_0)}\right]^2}{\sigma^2\left[1-e^{-2b(t-t_0)}\right]}\right\}$$

con

$$\mu = a-\frac{1}{2}\sigma^2;\ y,y_0 \in \mathbb{R};\ t > t_0;\ f_{Y(t_0)}(y_0) > 0 \tag{3.8}$$

donde $f_{Y(t_0)}$ es la función de densidad de $Y(t_0)$.

Si en (3.5), (3.6) y (3.7) le damos a t_0 el valor 0 y tenemos en cuenta que $Y(0) = \log X(0) \in \mathbb{R}$, tenemos que la solución $\{Y(t), t \geqslant 0\}$ de la E.D.E. (2.2) es un proceso Gaussiano con

$$Y(t) = \frac{\mu}{b}+e^{-bt}\left[Y(0)-\frac{\mu}{b}+\sigma\int_0^t e^{bs}dW(s)\right] \tag{3.9}$$

$$E[Y(t)] = \frac{\mu}{b}+e^{-bt}\left(Y(0)-\frac{\mu}{b}\right) \tag{3.10}$$

(3.11)

$$Cov[Y(t),Y(s)] = e^{-b(t+s)}Cov\left[\sigma\int_0^t e^{bu}dW(u), \sigma\int_0^s e^{bu}dW(u)\right]$$

$$= e^{-b(t+s)}\left[\sigma^2 Cov\left(\int_0^t e^{bu}dW(u), \int_0^s e^{bu}dW(u)\right)\right]$$

$$= e^{-b(t+s)}\left[\sigma^2\int_0^{min(t,s)} e^{2bu}du\right] = e^{-b(t+s)}\left[\frac{\sigma^2}{2b}\left(e^{2bmin(t,s)}-1\right)\right]$$

(3.12) $$Var[Y(t)] = \frac{\sigma^2}{2b}\left[1 - e^{-2bt}\right]$$

y cuya función de densidad de transición está dada por (3.8).

La función generatriz de momentos de las distribuciones finito dimensionales del proceso $\{Y(t), t \geqslant 0\}$es, para cada colección finita $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_k}$ $(t_1 < t_2 < \cdots < t_k)$

(3.13)
$$F(u_1, u_2, \ldots, u_k) = E\left[\exp\left(\sum_{i=1}^{k} u_i Y_{t_i}\right)\right] = \exp\left[\sum_{i=1}^{k} u_i E(Y_{t_i}) + \tfrac{1}{2}\sum_{i,j=1}^{k} u_i u_j Cov\left(Y_{t_i}, Y_{t_j}\right)\right]$$

con $E(Y_{t_i})$ y $Cov\left(Y_{t_i}, Y_{t_j}\right)$ dadas por (3.10) y (3.11) respectivamente.

En particular, para la distribución unidimensional o variable general del proceso, $Y(t)$, se tiene

(3.14) $$F_{Y(t)}(u) = \exp\left\{u\left[\frac{\mu}{b} + e^{-bt}\left(Y(0) - \frac{\mu}{b}\right)\right] + \frac{1}{2}u^2\left[\frac{\sigma^2}{2b}\left(1 - e^{-2bt}\right)\right]\right\}$$

El proceso $\{X(t), t \geqslant 0\}$, solución única de la E.D.E. (2.1) está, después del estudio precedente, log-normalmente distribuido $\left(X(t) = X(0)\, e^{Y(t) - Y(0)}, X(0) = e^{Y(0)} \in \mathbb{R}\right)$y su función de densidad de transición $p'(x, t/x_0, t_0)$ es, si $x_0 = e^{y_0}$, aplicando la fórmula del cambio de variable para densidades condicionadas y después de (3.8):

$$p'(x, t/x_0, t_0) = f_{X(t)/X(t_0)}(x/x_0) = \left|\frac{1}{x}\right| f_{Y(t)/Y(t_0)}(\log x/\log x_0) =$$

$$= \frac{1}{x} p(\log x, t/\log x_0, t_0)$$

$$= \frac{1}{x\sqrt{\frac{\pi\sigma^2}{b}\left[1 - e^{-2b(t-t_0)}\right]}} \exp\left\{\frac{-b\left[\log x - \frac{\mu}{b} - \left(\log x_0 - \frac{\mu}{b}\right)e^{-b(t-t_0)}\right]^2}{\sigma^2\left[1 - e^{-2b(t-t_0)}\right]}\right\}$$

(3.15) $$\forall x, x_0 \in \mathbb{R}^+;\ t > t_0; f_{X(t_0)}(x_0) > 0, \text{donde } \mu = a - \frac{1}{2}\sigma^2$$

Los momentos de las distribuciones finito-dimensionales del proceso log-normal generalizado $\{X(t), t \geqslant 0\}$ se deducen ahora de las fórmulas (3.14) y (3.13) de la función

generatriz de momentos de las distribuciones unidimensional y k-dimensionales de su proceso Gaussiano soporte $\{Y(t), t \geqslant 0\}$. Son particularmente interesantes

$$E[X(t)] = E\left(e^{Y(t)}\right) = F_{Y(t)}(1) =$$

$$= \exp\left\{\left[\frac{\mu}{b} + e^{-bt}\left(\log X(0) - \frac{\mu}{b}\right)\right] + \frac{1}{2}\left[\frac{\sigma^2}{2b}\left(1 - e^{-2bt}\right)\right]\right\} \tag{3.16}$$

$$E\left[X(t)^2\right] = E\left(e^{2Y(t)}\right) = F_{Y(t)}(2) =$$

$$= \exp\left\{2\left[\frac{\mu}{b} + e^{-bt}\left(\log X(0) - \frac{\mu}{b}\right)\right] + 2\left[\frac{\sigma^2}{2b}\left(1 - e^{-2bt}\right)\right]\right\} \tag{3.17}$$

$$E(X_t X_s) = E\{\exp[Y(t) + Y(s)]\} = F(1,1) =$$

$$= \exp\left\{EY(t) + EY(s) + \frac{1}{2}VarY(t) + \frac{1}{2}VarY(s) + Cov[Y(t), Y(s)]\right\} =$$

$$= \frac{2\mu}{b} + \left[\log X(0) - \tfrac{\mu}{b}\right]\left(e^{-bt} + e^{-ts}\right) + \tfrac{\sigma^2}{4b}\left[2 - e^{-2bt} - e^{-2bs} + 2e^{-b(t+s)}\left(e^{2b\,min(t,s)} - 1\right)\right] \tag{3.18}$$

de donde se pueden obtener inmediatamente la varianza y la función covarianza del proceso.

(Observamos que si bien el proceso logarítmico-normal propiamente dicho es un caso particular, para $b = 0$, sus características estadísticas, y en particular su función de densidad de transición [ver, p.e., Capocelli and Ricciardi (1974) y para el caso no homogéneo Buendía Moya y Gómez García (1995)]

$$p(x,t/x_0,t_0) = \frac{1}{x\sqrt{2\pi\sigma^2(t-t_0)}}\exp\left\{-\frac{[\log x - \log x_0 - \mu(t-t_0)]^2}{2\sigma^2(t-t_0)}\right\} \quad \text{con } \mu = a - \frac{1}{2}\sigma^2 \tag{3.19}$$

no pueden obtenerse de las del proceso solución de la E.D.E. (2.1) haciendo $b = 0$, en cuyo caso la ecuación (2.2) es la E.D.E. del movimiento Browniano aritmético).

4. LAS DISTRIBUCIONES DE EQUILIBRIO Y LA ESTACIONARIEDAD

Como hemos indicado en el apartado anterior, para $b > 0$ el proceso $Y(t) = \log X(t)$ que, como sabemos, verifica (2.2):

$$dY = (\mu - bY)\,dt + \sigma dW \text{ con } \mu = a - \frac{1}{2}\sigma^2$$
$$Y(0) = Y_0 = \log X(0) \in \mathbb{R}$$

es una extensión del proceso de Uhlenbeck-Ornstein, que aparece como solución de la E.D.E. anterior para $\mu = 0$ y $b > 0$).

De (3.7) se obtiene:

Para $b < 0$, es $\lim_{t\to\infty} Var[Y(t)/Y(t_0) = y_0] = \infty$. (Sucede igual para $b = 0$, como puede verse en las referencias anteriores para ese caso). En contraste, para $b > 0$ aunque $Var[Y(t)/Y(t_0) = y_0]$ también crece con t, se tiene $\lim_{t\to\infty} Var[Y(t)/Y(t_0) = y_0] = \frac{\sigma^2}{2b}$. Similarmente, para $b > 0$, en (3.6) observamos que $\lim_{t\to\infty} E[Y(t)/Y(t_0) = y_0] = \frac{\mu}{b}$.

Si ahora aplicamos que una sucesión de variables aleatorias normales $Y_n \sim N(\mu_n, \sigma_n)$, con $\lim_{n\to\infty} \mu_n = \mu$ y $\lim_{n\to\infty} \sigma_n = \sigma$, converge en distribución a una variable normal $Y \sim N(\mu, \sigma)$ (basta considerar las funciones generatrices de momentos de las variables Y_n), obtenemos de (3.10) y (3.12) en particular, para $b > 0$, que $\{Y(t), t \geqslant 0\}$ y por tanto $\{X(t), t \geqslant 0\}$ (la función $X(t) = e^{Y(t)}$ conserva la convergencia en distribución) tienen una distribución de equilibrio no degenerada, que es independiente de la condición inicial. Esto es

(4.1) $$Y \sim N\left(\frac{\mu}{b}, \frac{\sigma^2}{2b}\right) \text{ independientemente de } Y(0).$$

(4.2) $$X \sim \text{Log-normal, con } E(X) = \exp\left(\frac{\mu}{b} + \frac{\sigma^2}{4b}\right)$$

que es, pues, independiente de $X(0)$.

Se tiene, por tanto, de lo precedente, como en (3.8) y (3.15), para las densidades de probabilidad de transición a las distribuciones límite

(4.3) $$p(y, y_0, t_0) = \frac{1}{\sqrt{\pi \frac{\sigma^2}{b}}} \exp\left[-\frac{b\left(y - \frac{\mu}{b}\right)^2}{\sigma^2}\right]$$

(4.4) $$p(x, x_0, t_0) = \frac{1}{x\sqrt{\pi \frac{\sigma^2}{b}}} \exp\left[-\frac{b\left(\log x - \frac{\mu}{b}\right)^2}{\sigma^2}\right]$$

que son independientes de y_0 (ó de x_0) y de t_0.

Podemos determinar las condiciones en las que el proceso $Y(t)$, solución de la E.D.E.

$$
\begin{aligned}
&dY = (\mu - bY)\,dt + \sigma dW \text{ con } \mu = a - \frac{1}{2}\sigma^2 \\
&Y(0) = C
\end{aligned}
\tag{4.5}
$$

donde C es una v.a. dada no degenerada e independiente de $W(t)$, $t \geqslant 0$, es estacionario.

Su solución, como en (3.9), es

$$
Y(t) = \frac{\mu}{b} + e^{-bt}\left[Y(0) - \frac{\mu}{b} + \sigma\int_0^t e^{bs} dW(s)\right] \tag{4.6}
$$

Entonces, si $Y(0)$ es Gaussiana, se tiene que $Y(t)$ es un proceso Gaussiano con

$$
E[Y(t)] = \frac{\mu}{b} + e^{-bt}\left\{E[Y(0)] - \frac{\mu}{b}\right\} \tag{4.7}
$$

(4.8)

$$
\begin{aligned}
Var[Y(t)] &= e^{-2bt}\left\{Var[Y(0)] + \sigma^2\int_0^t e^{2bs} ds\right\} = e^{-2bt}\left\{Var[Y(0)] + \frac{\sigma^2}{2b}\left(e^{2bt} - 1\right)\right\} \\
&= e^{-2bt}\left\{Var[Y(0)] + \frac{\sigma^2}{2b}\left(e^{2bt} - 1\right)\right\} = e^{-2bt}\left\{Var[Y(0)] - \frac{\sigma^2}{2b}\right\} + \frac{\sigma^2}{2b}
\end{aligned}
$$

(4.9)

$$
\begin{aligned}
Cov[Y(t), Y(s)] &= e^{-b(t+s)} Cov\left\{Y(0) + \sigma\int_0^t e^{bu} dW(u), Y(0) + \sigma\int_0^s e^{bu} dW(u)\right\} \\
&= e^{-b(t+s)}\left\{Var[Y(0)] + Cov\left[Y(0), \sigma\int_0^s e^{bu} dW(u)\right] + \right. \\
&\quad + Cov\left[Y(0), \sigma\int_0^t e^{bu} dW(u)\right] + \\
&\quad \left. + \sigma^2 Cov\left[\int_0^s e^{bu} dW(u), \int_0^t e^{bu} dW(u)\right]\right\} = \\
&= e^{-b(t+s)}\left\{Var[Y(0)] + \sigma^2\int_0^{min(t,s)} e^{2bu} du\right\} = \\
&= e^{-b(t+s)}\left\{Var[Y(0)] + \frac{\sigma^2}{2b}\left(e^{2b\,min(t,s)} - 1\right)\right\} = \\
&= e^{-b(t+s)}\left\{Var[Y(0)] - \frac{\sigma^2}{2b}\right\} + \frac{\sigma^2}{2b} e^{-b|t-s|}
\end{aligned}
$$

De modo que si

$$Var[Y(0)] = \frac{\sigma^2}{2b} \quad \text{y} \quad E[Y(0)] = \frac{\mu}{b} \tag{4.10}$$

se tiene

$$\begin{aligned} Cov[Y(t), Y(s)] &= \frac{\sigma^2}{2b} e^{-b|t-s|} \\ E[Y(t)] &= \frac{\mu}{b} = \text{cte} \end{aligned} \tag{4.11}$$

y el proceso $Y(t)$, solución de la E.D.E. (4.5), es *estacionario en sentido amplio* [ver, p.e., Wong (1971)], y puesto que es Gaussiano (si $Y(0)$ es una variable Gaussiana), es *estacionario*. Como es solución única de una E.D.E. es un proceso de Markov, pero independientemente de esto, las condiciones (4.10), que nos dan las fórmulas (4.11), y el carácter Gaussiano determinan la condición de Markov, pues la expresión de la covarianza verifica la condición necesaria y suficiente para que un proceso Gaussiano sea de Markov. Puesto que la función covarianza de un proceso Markoviano, Gaussiano y estacionario es necesariamente de la forma $R(t,s) = Ce^{-K|t-s|}$ [ver, p.e., Papoulis (1980)], y las funciones media y covarianza determinan completamente todas las distribuciones finito dimensionales de un proceso Gaussiano, concluimos que este proceso con la condiciones (4.10), y que además puede considerarse separable, es el único Gaussiano, Markoviano y estacionario no centrado en tiempo continuo. Es decir, dado un proceso Gaussiano, Markoviano y estacionario, con función de covarianza $R(t,s) = Ce^{-K|t-s|}$ y media m, su ley coincide con la del proceso Y solución de (4.5), tomando $b = K$, $\sigma^2 = 2KC$ y $a = (m+C)K$.

(Para $\mu = 0$ es centrado y se trata, entonces, del proceso de **U-O** propiamente dicho, como es bien sabido).

Obviamente, las condiciones (4.10), con $Y(0) = \log X(0)$, y el carácter Gaussiano de la distribución inicial $Y(0)$ determinan también la estacionariedad de nuestro proceso log-normal y de Gompertz generalizados, solución de la E.D.E. que resulta de modificar en (2.1) la condición inicial:

$$\begin{aligned} &dX(t) = [a - b\log X(t)]X(t)\,dt + \sigma X(t)\,dW(t) \quad t \in [0,\infty) \\ &X(0) = K \end{aligned} \tag{4.12}$$

donde ahora K es una v.a. no degenerada e independiente de $W(t), t \geqslant 0$.

5. POTENCIAS DE LOS PROCESOS LOG-NORMALES (generalizados y estrictos)

Consideramos de nuevo la ecuación diferencial de Itô (2.1):

$$dX = (a - b\log X)X\,dt + \sigma X\,dW \quad t \in [0,\infty)$$
$$X(0) = X_0 \in \mathbb{R}^+;\ a,b,\sigma \text{ ctes.}$$

donde hemos puesto X por $X(t)$ y W por $W(t)$, como haremos frecuentemente en adelante para simplificar la notación.

Sea ahora el proceso dado por

$$Q(t) = A(t)\,[X(t)]^{\gamma(t)} \tag{5.1}$$
$$A(t) \neq 0,\ \gamma(t) \neq 0$$

donde $A(t)$ y $\gamma(t)$ son funciones reales derivables para cada $t \in [0,\infty)$.

Aplicando el lema de Itô, de (2.1) y (5.1) obtenemos que la ecuación diferencial de Itô para Q es

$$\begin{aligned} dQ &= \left(\dot{A}X^\gamma + \dot{\gamma}AX^\gamma \log X\right)dt + A\gamma X^{\gamma-1}dX + \frac{1}{2}A\gamma(\gamma-1)X^{\gamma-2}(dX)^2 \\ &= \left(\dot{A}X^\gamma + \dot{\gamma}AX^\gamma \log X\right)dt + A\gamma X^{\gamma-1}\left[(a - b\log X)X\,dt + \sigma X\,dW\right] \\ &+ \frac{1}{2}A\gamma(\gamma-1)X^{\gamma-2}\sigma^2X^2dt = \\ &= \left[\frac{\dot{A}}{A} + \gamma\left(a + \frac{(\gamma-1)}{2}\sigma^2\right) + (\dot{\gamma} - \gamma b)\log X\right]AX^\gamma dt + \sigma\gamma AX^\gamma dW \\ &= \left[\frac{\dot{A}}{A} + \gamma\left(a + \frac{(\gamma-1)}{2}\sigma^2\right) + \left(\frac{\dot{\gamma}}{\gamma} - b\right)\gamma\log X\right]Q\,dt + \sigma\gamma Q\,dW \\ &= \left[\frac{\dot{A}}{A} + \gamma\left(a + \frac{(\gamma-1)}{2}\sigma^2\right) + \left(\frac{\dot{\gamma}}{\gamma} - b\right)(\log Q - \log A)\right]Q\,dt + \sigma\gamma Q\,dW \end{aligned} \tag{5.2}$$

con $\dot{A} = \frac{dA}{dt}$ y $\dot{\gamma} = \frac{d\gamma}{dt}$.

Si A y γ son constantes, de (5.2) con $\dot{A} = \dot{\gamma} = 0$, tenemos

$$\begin{aligned} dQ &= \left\{\gamma\left[a + \frac{(\gamma-1)}{2}\sigma^2\right] + b\log A - b\log Q\right\}Q\,dt + \sigma\gamma Q\,dW \\ &= (a' - b\log Q)\,Q\,dt + \sigma' Q\,dW \end{aligned} \tag{5.3}$$

donde $a' = \gamma\left[a + \frac{(\gamma-1)}{2}\sigma^2\right] + b\log A$ y $\sigma' = \sigma\gamma$ son constantes.

Comparando (2.1) con (5.3), vemos que la estructura de la E.D.E. para Q es la misma que para X, excepto para los valores de los parámetros:

Las variables aleatorias potencias de X (procesos log-normales generales) tienen la misma familia distribucional que X.

En el caso $b = 0$, que corresponde a un proceso $X(t)$ log-normal con E.D.E.

$$\begin{aligned} &dX = aX\,dt + \sigma X\,dW,\ t \in [0,\infty) \\ &X(0) = X_0 \in \mathbb{R}^+ \end{aligned} \tag{5.4}$$

la ecuación (5.3) puede escribirse

$$\begin{aligned} &dQ = a'Q\,dt + \sigma' Q\,dW,\ t \in [0,\infty) \\ &Q(0) = Q_0 \in \mathbb{R}^+ \end{aligned} \tag{5.5}$$

Q satisface la ecuación (5.4) cambiando a_0 por a' y σ por σ'. Por tanto, las potencias (exponente constante) de procesos log-normales son también log-normales.

Como la solución de (5.4) es [ver, p.e., Malliaris and Brock (1982)]

$$X(t) = X(0)\exp[\mu t + \sigma W(t)],\ t \geqslant 0 \quad \text{con } \mu = a - \frac{1}{2}\sigma^2 \tag{5.6}$$

de aquí y de (5.5), obtenemos

$$\begin{aligned} Q(t) &= Q(0)\exp[\mu' t + \sigma' W(t)] = Q(0)\exp\left[\left(a' - \frac{1}{2}\sigma'^2\right)t + \sigma' W(t)\right] \\ &= Q(0)\exp\left[\gamma\left(a - \frac{1}{2}\sigma^2\right)t + \gamma\sigma W(t)\right] \end{aligned} \tag{5.7}$$

y se tiene para la esperanza del proceso [ver, p.e., Skiadas, Giovanis and Dimoticalis (1994)]

$$E[Q(t)] = Q(0)\exp(a't) = Q(0)\exp\left[\gamma\left(a + \frac{\gamma-1}{2}\sigma^2\right)t\right] \tag{5.8}$$

que para $A = 1$ da

$$E\{[X(t)]^\gamma\} = [X(0)]^\gamma \exp\left[\gamma\left(a + \frac{\gamma-1}{2}\sigma^2\right)t\right] \tag{5.9}$$

y en el caso especial $\gamma = -1$ obtenemos

$$E\left[\frac{1}{X(t)}\right] = \frac{1}{X(0)} e^{(-a+\sigma^2)t} \tag{5.10}$$

OBSERVACIONES

1. Estos procesos que hemos estudiado como soluciones de ecuaciones de Itô tienen función de transición estacionaria, son homogéneos. Esto es natural porque están regidos por ecuaciones autónomas o tiempo-independientes. Entonces, la función de densidad de transición puede ponerse en la forma $p(t;x_0,x)$ [ó $p(t;y_0,y)$], pues

$$p(x,t/x_0,t_0) = p(x,t-t_0/x_0,0) = p(x,t'/x_0,0) = p(t';x_0,x)$$

   y cambiando t' por t podremos escribir $p(t;x_0,x)$. Pero en ese caso, en las ecuaciones atrasadas de difusión (2.3) habrá que sustituir el término $\dfrac{\partial p'(x,t/x_0,t_0)}{\partial t_0}$ por $-\dfrac{\partial p'(t;x_0,x)}{\partial t}$ y en la (3.1) el término $\dfrac{\partial p(y,t/y_0,t_0)}{\partial t_0}$ por $-\dfrac{\partial p(t;y_0,y)}{\partial t}$.

2. Si en la E.D.E. (2.1) que rige el procesos log-normal generalizado, sustituimos σ por $-\sigma$, el proceso solución de la nueva E.D.E. resultante tiene, evidentemente, el mismo coeficiente de tendencia $A_1(t,x) = (a - b\log x)x$ y el mismo coeficiente de difusión $A_2(t,x) = \sigma^2 x^2$ que el anterior, por lo que tiene el mismo generador infinitesimal, que determina unívocamente la función de transición, y la misma función de densidad de transición que la solución de (2.1). Como son procesos Markovianos con el mismo espacio de valores y la misma distribución inicial, los dos procesos tienen *la misma ley*. Y lo mismo puede decirse de la solución de la E.D.E. (2.2) si sustituimos σ por $-\sigma$. A esta misma conclusión se llega sin más que considerar que cambiar σ por - σ en las E.D.E. (2.1) ó (2.2) equivale a cambiar $W(t)$ por - $W(t)$, que también es un movimiento browniano estándar.

AGRADECIMIENTOS

Agradecemos a los evaluadores las precisas y minuciosas aclaraciones y sugerencias que nos han indicado, así como sus valiosos comentarios.

REFERENCIAS

Arnold, L. (1974). *Stochastic Differential Equations. Theory and Applications.* John Wiley.

Bartolomew, D. J. (1973). *Stochastic Model for Social Processes.* 2° edic. John Wiley.

Bharucha-Reid (1960). *Elements of the theory of Markov Processes. and their applications.* McGraw-Hill.

Bhattacharya, R. N. & Waymire, E. C. (1990). *Stochastic Processes with Applications.* John Wiley.

Bouleau, N. (1988). *Processus Stochastiques et Applications.* Hermann.

Buendía Moya, F. & Gómez García, J. (1995). «Estudio del proceso estocástico logarítmico-normal unidimensional con factores exógenos como solución de una ecuación de Itô». *IX Reunión Asepelt-España,* Santiago de Compostela.

Capocelli, R. M. & Ricciardi, L. M. (1974). «A diffusion model for population growth in random environment». *Theor. Pop. Biol.*, 5, 28-41.

Gihman and Skorohod (1972). *Stochastic Differential Equations.* Springer-Verlag.

Gutiérrez, R. & Valderrama, M., eds. (1994). *Selected Topics on Stochastic Modelling.* World Scientific.

Gutiérrez, R.; Ricciardi, L. M.; Román, P. & Torres, F. (1997). «First-Passage-Time Densities for Time non-Homogeneous Diffusion Processes». *Journal of Applied Probability*, 34, 623-631.

Ikeda, N-Watanabe, S. (1989). *Stochastic Differential Equations and Diffussion Processes.* North-Holland.

Kijima, M. (1997). *Markov Processes for Stochasting Modelling.* Chapman & Hall.

Malliaris, A. G. & Brock, W. A. (1982). *Stochastic Methods in Economics and Finance.* North-Holland.

Moreno Bas, E. (1974). «Una aplicación de los procesos markovianos de difusión logarítmico-normales a los ingresos y gastos de las Administraciones Públicas». *Hacienda Pública Española,* 0028.

McShane, E. J. (1994). *Stochastic Calculus and Stochastic Models.* Academic Press.

Papoulis, A. (1980). *Probabilidad, variables aleatorias y procesos estocásticos.* Eunibar.

Ricciardi, L. M. (1976). «On the transformation of diffusion processes in the Wiener process». *J. Math. Anal. Appl.*, 54, 185-199.

Ricciardi, L. M. (1977). «Diffusion Processes and Related Topics in Biology». *Lecture Notes in Biomathematics,* 14. Springer-Verlag.

Ricciardi, L. M. (1986). «Stochastic Population Theory: Diffussion Processes». In Hallam and Levin (eds.), *Biomathematics,* 17, *Mathematical Ecology.* Springer-Verlag.

Sengupta, J. K. (1986). *Stochastic Optimization and Economic Models.* D. Reidel Publishing Company.

Skiadas, C. H.; Giovanis, A. N. & Dimoticalis, I. (1993). «A sigmoid stochastic growth model derived from the revised exponential». In J. Janssen and C. H. Skiadas (eds.), *Applied Stochastic Models and Data Analysis.* World Scientific.

Skiadas, C. H.; Giovanis, A. N. & Dimoticalis, I. (1994). «Investigation of Stochastic. Differential Models: The Gompertzian Case». In Gutiérrez-Valderrama (eds.), *Selected Topics on Stochastic Modelling.* World Scientific).

Sobczyk, K. (1991). *Stochastic Differential Equations whit Applications to Physics and Engineering.* Kluwer Academic Publishers.

Tintner, G. & Bello (1968). «Aplicación de un proceso de difusión logarítmico-normal al crecimiento ecomómico». *Trabajos de Estadística,* 19, 83-97.

Tintner, G. & Patel, R. C. (1966). «A lognormal diffusion process applied to the development of Indian agriculture with some considerations on economic policy». *Journal of the Indian Society of Agricultural Statistic,* 18, 36-44.

Tintner, G. & Sengupta, J. K. (1972). *Stochastic Economics.* Academic Press.

Tintner, G. & Thomas, E. J. (1963). «Un modele stochastique de développement économique avec application a l'industrie Anglaise». *Revue d'Economie Politique,* 73, 143-47.

Todorovic, P. (1992). *An Introduction to Stochastic Processes and their Applications.* Springer-Verlag.

Wong, E. (1971). *Stochastic Processes in Information and Dynamical Systems.* McGraw-Hill Book Company.

Wong, E. & Hajek, B. K. (1985). *Stochastic Processes in Engineering System.* McGraw Hill.

ENGLISH SUMMARY

A GENERALIZATION OF THE LOG-NORMAL AND GOMPERTZ STOCHASTIC PROCESSES AS ITÔ PROCESSES

JUAN GÓMEZ GARCÍA
FULGENCIO BUENDÍA MOYA*

We study a stochastic process, of which Log-normal and Gompertz processes are particular cases, as solution of a stochastic differential equation (S.D.E.). The moments of finite dimensional distributions are studied along with the transition density function, the equilibrium distribution and the stationarity conditions. We deduce the that the powers of the process, and in particular those of the Log-normal one, verify an S.D.E. of the same type.

Keywords: Stochastic differential equation, Kolmogorov equations, log-normal process, Gompertz process, Uhlenbeck-Ornstein process, Langevin equation

AMS Classification (MSC 2000): 60H10

* Departamento de Métodos Cuantitativos para la Economía. Facultad de Economía y Empresa. Universidad de Murcia. 30100 Murcia, Spain.

- Received March 2000.
- Accepted June 2001.

1. INTRODUCTION

The log-normal stochastic diffusion processes have been extensively applied in the analysis of various economic variables. Their study, when taking as starting point the transition density function or that of the diffusion and trend coefficients (using Kolmogorov equations in this case) poses certain difficulties as well as huge limitations. We study herein a more general type of process and we start exclusively from the analytical expression of an S.D.E. solution.

2. S.D.E. APPROACH. EXISTENCE AND UNIQUENESS OF THE SOLUTIONS

Let the Itô S.D.E. be

$$\begin{aligned} &dX(t) = [a - b\log X(t)]X(t)\,dt + \sigma X(t)\,dW(t) \\ &\text{with } X(0) = X_0 \in \mathbb{R}^+;\ a,b,c \text{ constant} \end{aligned} \tag{2.1}$$

where $X(t)$ is a process with values in $(0,\infty) = \mathbb{R}^+$. In this S.D.E. (2.1), for $b = 0$ the Malthusian S.D.E., which governs the logarithmico-normal process appears, and for $a = 0$ we have the Gompertz stochastic model.

The equation makes sense and since its drift coefficient $m(t,x) = (a - b\log x)x$ is only defined for $x > 0$, it is atypical. Thus, in order to prove the existence of a unique global solution for (2.1), we will apply a *state change transformation* to this equation [see, e.g., Bhattacharya-Waymire (1990), pg. 382]. Thus we define the $Y(t) = \log X(t)$ variable, and from the Itô lemma, together with (2.1), we obtain

$$\begin{aligned} &dY = (\mu - bY)\,dt + \sigma dW \text{ with } \mu = a - \frac{1}{2}\sigma^2 \\ &\text{with } Y(0) = \log X(0) = Y_0 \in \mathbb{R} \end{aligned} \tag{2.2}$$

which has a unique solution $\{Y(t), t \geqslant 0\}$, which is a diffussion with drift and diffusion coefficients $A_1'(t,y) = \mu - by$ and $A_2'(t,y) = \sigma^2$, respectively [see, e.g., Arnold (1974), pgs. 152-153]. As a result of the above, $\left\{X(t) = e^{Y(t)}, t \geqslant 0\right\}$ is a diffusion process, the unique solution of (2.1), and with drift and diffusion coefficients $m(t,x) = (a - b\log x)x$ and $\sigma^2(t,x) = \sigma^2 x$, respectively.

3. STUDY OF THE S.D.E. SOLUTION PROCESS

First we will consider an equation similar to (2.2) but with more general initial conditions and taking any $t_o \in [0,\infty)$ point:

$$(3.1)\qquad \begin{aligned} &dY = (\mu - bY)\,dt + \sigma dW \text{ con } \mu = a - \frac{1}{2}\sigma^2 \\ &\text{con c.i. } Y(t_0) = \log X(t_0) = C \end{aligned}$$

where $Y(t_0) = C$ is a random variable which is non degenerated and independent of $W(t) - W(t_0)$, $t \geqslant t_0$. From the solution of this equation

$$(3.2)\qquad Y(t) = \frac{\mu}{b} + e^{-b(t-t_0)}\left[Y(t_0) - \frac{\mu}{b} + \sigma\int_{t_0}^{t} e^{b(s-t_0)}dW(s)\right]$$

we deduce that the transition density $p(y,t/y_0,t_0)$ of $\{Y(t), t \geqslant 0\}$, the solution of (2.2), is the density function of the general variable of the process [see, e.g. Ikeda-Watanabe (1989), pp. 214-215].

$$(3.3)\qquad Y(t,t_0,y_0) = \frac{\mu}{b} + e^{-b(t-t_0)}\left[y_0 - \frac{\mu}{b} + \sigma\int_{t_0}^{t} e^{b(s-t_0)}dW(s)\right]$$

Thus, we have

(3.4)

$$p(y,t/y_0,t_0) = \frac{1}{\sqrt{\frac{\pi\sigma^2}{b}\left[1 - e^{-2b(t-t_0)}\right]}} \exp\left\{\frac{-b\left[y - \frac{\mu}{b} - \left(y_0 - \frac{\mu}{b}\right)e^{-b(t-t_0)}\right]^2}{\sigma^2\left[1 - e^{-2b(t-t_0)}\right]}\right\}$$

Also from (3.2), if we give t_0 the value 0 and we take into account that $Y(0) = \log X(0) \in R$, we obtain that the solution $\{Y(t), t \geqslant 0\}$ of the S.D.E. (2.2) is a Gaussian process with

$$(3.5)\qquad Y(t) = \frac{\mu}{b} + e^{-bt}\left[Y(0) - \frac{\mu}{b} + \sigma\int_{0}^{t} e^{bs}dW(s)\right]$$

whose transition density function is given by (3.4). [For $\mu = 0$ and $b > 0$, (2.2) it is the Langevin equation and the solution process is the Uhlenbeck-Ornstein one: see, e.g., Bouleau (1988)]. From (3.5) we obtain the statistical characteristics of the process

$\{Y(t), t \geqslant 0\}$, in particular the generatrix function of moments of the finite-dimensional distributions. The $\{X(t), t \geqslant 0\}$ process, which is the solution of the S.D.E. (2.1) is, after the preceeding study, log-normally distruibuted $\left(X(t) = X(0)\, e^{Y(t)-Y(0)}, X(0) = \right.$ $\left. = e^{Y(0)} \in \mathbb{R}\right)$ and its transition density function $p'(x, t/x_0, t_0)$ is deduced immediately from (3.4). The moments of its finite-dimensional distributions are obtained from generatrix function of moments of the one-dimensional and k-dimensional distributions of the Gaussian process $\{Y(t), t \geqslant 0\}$.

4. THE EQUILIBRIUM DISTRIBUTIONS AND THE STATIONARITY

De (3.5), for $b < 0$, $\lim_{t\to\infty} Var[Y(t)] = \infty$ is obtained. (The same occurs for $b = 0$). For $b > 0$, $\lim_{t\to\infty} Var[Y(t)] = \frac{\sigma^2}{2b}$ and $\lim_{t\to\infty} E[Y(t)] = \frac{\mu}{b}$. It is concluded, for $b > 0$, that $\{Y(t), t \geqslant 0\}$ and, therefore, $\{X(t), t \geqslant 0\}$, have non degenerated equilibrium distributions (convergence in distribution) which are independent of the initial condition. That is $Y \sim N\left(\frac{\mu}{b}, \frac{\sigma^2}{2b}\right)$ and X~Log-normal.

Now let the $Y(t)$ process, the solution of the S.D.E., be

$$\begin{aligned} &dY = (\mu - bY)\, dt + \sigma dW \text{ with } \mu = a - \frac{1}{2}\sigma^2 \\ &Y(0) = C \end{aligned} \tag{4.1}$$

where C is a given non degenerated r.v. and independent of $W(t)$, $t \geqslant 0$. The solution de this S.D.E. is

$$Y(t) = \frac{\mu}{b} + e^{-bt}\left[Y(0) - \frac{\mu}{b} + \sigma \int_0^t e^{bs} dW(s)\right] \tag{4.2}$$

Then, if $Y(0)$ is Gaussian, it holds that $Y(t)$ is a Gaussian process. If, moreover, $Var[Y(0)] = \frac{\sigma^2}{2b}$ and $E[Y(0)] = \frac{\mu}{b}$, we obtain

$$Cov[Y(t), Y(s)] = \frac{\sigma^2}{2b} e^{-b|t-s|} \quad \text{and} \quad E[Y(t)] = \frac{\mu}{b} = cte \tag{4.3}$$

and the $Y(t)$ process, the solution of the S.D.E. (4.1), is *stationary* [see, e.g., Wong (1971)]. It is the only Gaussian, Markovian and stationary process which is not centred

on continuous time. (For $\mu = 0$ it is centred and we have the U-O process). The same conditions with $Y(0) = \log X(0)$, also determine the stationarity of our generalised log-normal and Gompertz processes.

5. POWERS OF THE LOG-NORMAL PROCESSES

If $\{X(t), t \geqslant 0\}$ is the solution of the S.D.E. (2.1), we consider the process given by

$$\text{(5.1)} \qquad \begin{cases} Q(t) = A(t)[X(t)]^{\gamma(t)} \\ A(t) \neq 0,\ \gamma(t) \neq 0 \end{cases}$$

with $A(t)$ and $\gamma(t)$ being real derivable functions for each $t \in [0,\infty)$. By applying the Itô lemma, of (2.1) y (5.1), we obtain the Itô differential equation for $Q(t)$, that for constant A and γ is

$$\text{(5.2)} \qquad dQ = (a' - b\log Q)\, Q dt + \sigma' Q dW$$

where $a' = \gamma\left[a + \dfrac{(\gamma-1)}{2}\sigma^2\right] + b\log A$ and $\sigma' = \sigma\gamma$ are constant.

Thus we have that the powers of log-normal, generalised or strict (case b=0) processes, are of the same type.

QÜESTIIÓ, vol. 25, 3, p. 415-436, 2001

DETECCIÓN DE *M* SEÑALES GAUSSIANAS UTILIZANDO EL DESARROLLO MODIFICADO DE UN PROCESO ESTOCÁSTICO*

JESÚS NAVARRO-MORENO
JUAN CARLOS RUIZ-MOLINA
Universidad de Jaén*

Utilizando el desarrollo modificado de un proceso estocástico se propone una nueva metodología, alternativa a la basada en el desarrollo de Karhunen-Loève, para el problema de detección de M señales Gaussianas en ruido Gaussiano blanco. Las soluciones proporcionadas no presentan el problema del cálculo de los autovalores y autofunciones asociados a la función de covarianza involucrada y son fácilmente implementables desde el punto de vista práctico.

Detection of *M* gaussian signals using the modified approximate Karhunen-Loève expansion of a stochastic process

Palabras clave: Desarrollo modificado de un proceso estocástico, detección de *M* señales gaussianas en ruido gaussiano blanco

Clasificación AMS (MSC 2000): 60G12

* Este trabajo ha sido financiado por el proyecto BFM2000-1103 del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I), Ministerio de Ciencia y Tecnología, España.
* Departamento de Estadística e I.O. Universidad de Jaén. Paraje Las Lagunillas, s/n. 23071 Jaén. E-mail: jnavarro,jcruiz@ujaen.es.
– Recibido en marzo de 2001.
– Aceptado en septiembre de 2001.

1. INTRODUCCIÓN

La detección de señales aleatorias en ruido Gaussiano blanco es un problema ampliamente estudiado en la literatura debido a su gran utilidad en la Teoría de la Comunicación Estadística (ver, por ejemplo, Van Trees, 1971 y Poor, 1994). En este problema se parte de un conjunto posible de señales aleatorias y el objetivo es decidir a partir de la medición o recepción de un proceso de observación que involucra a la señal y a un ruido perturbador, qué señal del conjunto ha sido emitida. El caso más sencillo que se puede presentar es decidir entre la presencia o la ausencia de una señal, denominado caso binario simple (Van Trees, 1971). Un ejemplo de este caso es el problema de detectar la presencia de un submarino usando un sistema de sonar. Sin embargo, existen otro tipo de situaciones prácticas, por ejemplo en astronomía, donde el número de señales a detectar es, en muchas ocasiones, superior a dos. Por esta razón, el estudio del problema de detección múltiple (M señales) es importante. Este problema puede ser modelizado por medio del siguiente conjunto de hipótesis:

$$H_i : Y(t) = S_i(t) + N(t), \qquad 0 \leq t \leq T, \qquad i = 1,\ldots,M,$$

donde las señales $\{S_i(t); t \in [0,T]\}$ son procesos Gaussianos, centrados y continuos en media cuadrática y $\{N(t); t \in [0,T]\}$ es un ruido Gaussiano blanco con parámetro de varianza $N_0/2$ y que es independiente de todas las señales.

Dado que el ruido blanco no puede considerarse rigurosamente un verdadero proceso aleatorio, el anterior problema puede ser transformado en otro equivalente a través de un proceso de integración de las observaciones (Poor, 1994). Así, obtenemos el modelo siguiente:

$$(1) \qquad H_i : X(t) = \int_0^t S_i(\tau)\, d\tau + W(t), \qquad 0 \leq t \leq T, \qquad i = 1,\ldots,M,$$

siendo $\{W(t); t \in [0,T]\}$ el proceso de Wiener con parámetro $N_0/2$ e independiente de las señales.

El objetivo de este problema consiste en determinar qué señal de entre las M posibles, $S_i(t)$, $i = 1,\ldots,M$, ha sido la realmente emitida a partir de la recepción del proceso de observación $X(t)$. Matemáticamente, pretendemos obtener una partición del espacio de las funciones muestrales $X(t)$, $\mathcal{S}_1,\ldots,\mathcal{S}_M$, de manera que si $X(t) \in \mathcal{S}_k$ entonces se elige la hipótesis H_k. Para determinar esta partición existen diversos criterios matemáticos como son el criterio de Neyman-Pearson, Bayes, etc. Debido a la sencillez que ofrece desde el punto de vista práctico, usualmente se utiliza el criterio de minimizar la pro-

babilidad total del error, $P(\varepsilon)$, dadas unas probabilidades a priori α_i, $i = 1,\ldots,M$[11]. Mediante esta regla de decisión obtenemos aquella partición que hace mínima la probabilidad del error asociada, es decir, que minimiza

$$P(\varepsilon) = 1 - \sum_{i=1}^{M} \alpha_i P_i(\mathcal{S}_i), \tag{2}$$

siendo P_i la probabilidad correspondiente a H_i. Puede demostrarse (Kadota, 1965) que bajo este último criterio, la partición óptima $(\bar{\mathcal{S}}_1,\ldots,\bar{\mathcal{S}}_M)$ para (1) viene determinada por

$$\bar{\mathcal{S}}_k = \left\{X(t); \alpha_k \frac{dP_k}{dP_0}(X) > \alpha_i \frac{dP_i}{dP_0}(X), i < k\right\} \bigcap \left\{X(t); \alpha_k \frac{dP_k}{dP_0}(X) \geq \alpha_i \frac{dP_i}{dP_0}(X), i > k\right\},$$

con P_0 la probabilidad inducida por $W(t)$ y $\frac{dP_i}{dP_0}(X)$ la derivada de Radon-Nikodym de P_i con respecto a P_0. Además, en Kadota (1965) puede encontrarse una nueva forma de expresar esta regla de decisión óptima que resulta más útil en la práctica. Esta nueva forma equivalente viene dada por

«elegir la hipótesis H_k si $I_k(X) = \max_{i=1,\ldots,M} I_i(X)$», (C1)

donde

$$I_i(X) = \log \frac{dP_i}{dP_0}(X) + \log \alpha_i, \qquad i = 1,\ldots,M. \tag{3}$$

Desde el punto de vista de la implementación, es necesario una forma de calcular las derivadas de Radon-Nikodym que aparecen en el sistema de detección anterior. Una de las herramientas más utilizadas para este propósito es el desarrollo de Karhunen-Loève de un proceso estocástico (Poor, 1994). Sin embargo, esta metodología presenta el inconveniente de que $\frac{dP_i}{dP_0}(X)$ y, por tanto, el estadístico de detección $I_i(X)$ dependen de los autovalores y autofunciones del operador integral involucrado y no existe un

[1]Estas probabilidades a priori α_i reflejan la información que se tiene acerca de cada hipótesis antes de realizar el experimento.

método estándar para calcularlos. Por ello, se utiliza una forma alternativa de representar el estadístico $I_i(X)$ denominada representación *estimadora-correladora* (Van Trees, 1971). Esta nueva forma está basada en el estimador lineal causal óptimo de la señal y es de gran utilidad cuando el estimador puede ser obtenido a través del filtro de Kalman. Sin embargo, para poder aplicar tal algoritmo de estimación es necesario que la señal verifique una ecuación de estados, hipótesis que en muchas ocasiones no se cumple.

El primer objetivo que nos planteamos en este trabajo será mostrar que el desarrollo modificado de un proceso estocástico (Ruiz-Molina y otros, 1999) permite obtener unas solución al problema de detección (1) evitando los problemas presentados por el desarrollo de Karhunen-Loève. El desarrollo modificado de un proceso estocástico es una representación aproximada tipo Karhunen-Loève que posee propiedades teóricas similares a éste pero que no presenta dificultades desde el punto de vista práctico. Así, se obtiene una solución para el problema de detección (1) que es fácilmente implementable y que no requiere el cálculo de los verdaderos autovalores y autofunciones.

Por otro lado, el segundo de los objetivos que nos planteamos será proporcionar una segunda forma alternativa para $I_i(X)$ que aproxima a la representación estimadora-correladora y que está basada en un estimador subóptimo derivado a partir del desarrollo modificado. La principal ventaja de este estimador subóptimo radica en su facilidad de cálculo ya que éste puede obtenerse recursivamente de forma similar al filtro de Kalman, no precisando que la señal verifique una ecuación de estados (Navarro-Moreno y otros, 2000).

1.1. Desarrollo Modificado

A continuación hacemos un resumen del desarrollo modificado introducido en Ruiz-Molina y otros (1999). Sea $\{X(t); t \in [0,T]\}$ un proceso de segundo orden definido sobre el espacio de probabilidad $(\Omega, \mathcal{B}, P)$, centrado, continuo en media cuadrática y con función de covarianza $R_X(t,s)$. Los autovalores y autofunciones asociados a $R_X(t,s)$ serán denotados por $\{\lambda_j\}_{j=1}^{\infty}$ y $\{\phi_j(t)\}_{j=1}^{\infty}$, respectivamente. Es bien conocido que el proceso $X(t)$ puede ser representado a través del desarrollo de Karhunen-Loève

$$X(t) = \sum_{j=1}^{\infty} X_j \phi_j(t), \qquad t \in [0,T],$$

donde la anterior serie converge en media cuadrática uniformemente en $t \in [0,T]$, y las v.a. X_j se definen de la forma

$$X_j = \int_0^T \phi_j(t) X(t)\, dt \qquad m.c., \qquad j = 1,2,\ldots \tag{4}$$

Desde un punto de vista práctico, el desarrollo de Karhunen-Loève presenta la dificultad de que no existe un método general para el cálculo de los autovalores y autofunciones de $R_X(t,s)$. Una alternativa consiste en utilizar procedimientos numéricos: el método de Rayleigh-Ritz (Baker, 1977). A partir de un conjunto de k funciones, $\{\varphi_1(t),\varphi_2(t),\ldots,\varphi_k(t)\}$, pertenecientes a un sistema completo de $L^2[0,T]$ arbitrario, este método permite obtener las autofunciones aproximadas

$$\tilde{\phi}_{jk}(t)=\sum_{i=1}^{k} a_{ji}\varphi_i(t), \qquad j=1,\ldots,k,$$

donde los coeficientes a_{ji} y los autovalores aproximados $\tilde{\lambda}_{jk}$ se obtienen a partir del problema de autovalores

$$\mathbf{A}\mathbf{a}_j=\tilde{\lambda}_{jk}\mathbf{B}\mathbf{a}_j, \qquad j=1,\ldots,k,$$

siendo los elementos de las matrices $\mathbf{A}=(A_{ij})$ y $\mathbf{B}=(B_{ij})$ de la forma

$$A_{ij}=\int_0^T\int_0^T R_X(t,s)\varphi_i(s)\varphi_j(t)\,dt\,ds, \qquad i,j=1,\ldots,k,$$
$$B_{ij}=\int_0^T \varphi_i(t)\varphi_j(t)\,dt, \qquad i,j=1,\ldots,k,$$

y los autovectores: $\mathbf{a}_j=(a_{j1},\ldots,a_{jk})'$, $j=1,\ldots,k$.

Los autovalores y autofunciones aproximados obtenidos tienen las siguientes propiedades: $0\leq\tilde{\lambda}_{jk}\leq\lambda_j$, $\tilde{\lambda}_{jk}\to\lambda_j$ y $\|\phi_j(t)-\tilde{\phi}_{jk}(t)\|_2\to 0$, cuando $k\to\infty$, donde $\|\cdot\|_2$ denota a la norma definida en $L^2[0,T]$. Además, el sistema de funciones $\{\tilde{\phi}_{jk}(t)\}_{j=1}^k$ es ortogonal en $L^2[0,T]$, por lo que a partir de ahora, consideraremos que también está ortonormalizado. Así, se verifica

$$\int_0^T\int_0^T R_X(t,s)\tilde{\phi}_{ik}(s)\tilde{\phi}_{jk}(t)\,dt\,ds=\tilde{\lambda}_{ik}\delta_{ij}.$$

Basadas en estas autofunciones aproximadas, en Ruiz-Molina y otros (1999) se propone un desarrollo aproximado tipo Karhunen-Loève denominado desarrollo modificado. Tal desarrollo se obtiene proyectando el proceso $X(t)$ sobre el subespacio de $L^2(\Omega,\mathcal{B},P)$ generado por el conjunto de variables aleatorias ortonormales $\{\tilde{\lambda}_{jk}^{-\frac{1}{2}}\tilde{X}_{jk}\}_{j=1}^n$, $n\leq k$, donde las variables $\tilde{X}_{jk}$ se definen de forma similar a (4) pero utilizando las autofunciones aproximadas en lugar de las verdaderas, es decir,

$$\tilde{X}_{jk}=\int_0^T\tilde{\phi}_{jk}(t)X(t)\,dt \qquad m.c., \qquad j=1,2,\ldots$$

El desarrollo modificado tiene la expresión

$$\check{X}_n(t) = \sum_{j=1}^{n} \tilde{X}_{jk}\hat{\phi}_{jk}(t), \qquad t \in [0,T], \tag{5}$$

donde

$$\hat{\phi}_{jk}(t) = \frac{1}{\tilde{\lambda}_{jk}} \int_0^T R_X(t,s)\tilde{\phi}_{jk}(s)\,ds, \qquad t \in [0,T].$$

Se demuestra en Ruiz-Molina y otros (1999) que las funciones $\hat{\phi}_{jk}(t)$, denominadas autofunciones modificadas, se aproximan mejor a la verdaderas que las aproximadas. De hecho, se tiene que $\left\|\phi_j(t) - \hat{\phi}_{jk}(t)\right\|_\infty \xrightarrow{k\uparrow\infty} 0$, donde $\|\cdot\|_\infty$ denota a la norma del supremo en $[0,T]$. Además, se verifica

$$\left\|X(t) - \check{X}_n(t)\right\|_H \xrightarrow{n\uparrow\infty} 0,$$

uniformemente en $t \in [0,T]$, donde $\|Z\|_H = (E[Z]^2)^{1/2}$, con $Z \in L^2(\Omega,\mathcal{B},P)$.

Nota 1. Aunque existen otros desarrollos en serie de un proceso estocástico diferentes al desarrollo de Karhunen-Loève, este último es óptimo en el sentido de que minimiza el error en media cuadrática con una representación finita del proceso. De esta forma, el desarrollo (5) permite aproximar con la precisión deseada al desarrollo truncado de Karhunen-Loève en n términos; es decir, para cada n fijo podemos elegir un k suficientemente grande de manera que se cumpla $\left\|X_n(t) - \check{X}_n(t)\right\|_H < \varepsilon$, con $\varepsilon > 0$ arbitrariamente pequeño y $X_n(t) = \sum_{j=1}^{n} X_j\phi_j(t)$. Así, a través del método propuesto se puede controlar en «cierta medida» la propiedad de optimalidad bajo truncamiento que posee el desarrollo de Karhunen-Loève.

Nota 2. Con objeto de abreviar la notación, a partir de ahora omitiremos el subíndice k en $\hat{\phi}_{jk}(t)$, $\tilde{\phi}_{jk}(t)$, $\tilde{\lambda}_{jk}$ y $\tilde{X}_{jk}$.

2. DETECCIÓN DE *M* SEÑALES GAUSSIANAS

En esta sección se aborda el problema de detección planteado en (1). Las funciones de covarianza de las M señales, $S_i(t)$, se denotarán por $R_{S_i}(t,s)$ y sus correspondientes autovalores y autofunciones se denotarán por $\{\lambda_{ij}\}_{j=1}^{\infty}$ y $\left\{\phi_{ij}(t)\right\}_{j=1}^{\infty}$, respectivamente. Se supondrá que el sistema $\left\{\phi_{ij}(t)\right\}_{j=1}^{\infty}$ es completo para $i = 1,\ldots,M$.

Como se ha indicado en la sección anterior, si se utiliza el criterio de minimizar la probabilidad total del error entonces la regla de decisión óptima viene determinada por (C1). El principal handicap de ésta es la implementación del estadístico $I_i(X)$ (3) debido a que es necesario calcular la derivada de Radon-Nikodym correspondiente. Esto puede ser solventado utilizando, por un lado, el desarrollo de Karhunen-Loève de cada señal

$$S_i(t) = \sum_{j=1}^{\infty} S_{ij}\phi_{ij}(t), \qquad t \in [0,T], \qquad i = 1,\ldots,M,$$

con $S_{ij} = \int_0^T \phi_{ij}(t)S_i(t)\,dt$ (m.c.) y por otro, que el proceso de Wiener admite la representación en serie (Poor, 1994)

$$W(t) = \sum_{j=1}^{\infty} \bar{W}_j \int_0^t \varphi_j(\tau)\,d\tau, \qquad t \in [0,T],$$

en el sentido de la media cuadrática, con $\{\varphi_j(t)\}_{j=1}^{\infty}$ un sistema ortonormal completo en $L^2[0,T]$ y $\bar{W}_j = \int_0^T \varphi_j(t)\,dW(t)$ (m.c.). Así, puede demostrarse (Poor, 1994) que el problema (1) es equivalente a

$$(6) \qquad H_i : X(t) = \sum_{j=1}^{\infty} \bar{X}_{ij} \int_0^t \phi_{ij}(\tau)\,d\tau, \qquad 0 \le t \le T, \qquad i = 1,\ldots,M,$$

donde $\bar{X}_{ij} = \int_0^T \phi_{ij}(t)\,dX(t)$ (m.c.).

Como consecuencia, se demuestra (Poor, 1994) que

$$\log \frac{dP_i}{dP_0}(X) = -\frac{1}{2}\sum_{j=1}^{\infty} \log\left(1+2\frac{\lambda_{ij}}{N_0}\right) + \frac{1}{N_0}\sum_{j=1}^{\infty} \frac{\lambda_{ij}}{\lambda_{ij}+\frac{N_0}{2}}\bar{X}_{ij}^2, \qquad i = 1,\ldots,M,$$

siendo la convergencia de la serie en el sentido de la media cuadrática. Por tanto, el estadístico $I_i(X)$ (3) toma la forma

$$(7) \quad I_i(X) = -\frac{1}{2}\sum_{j=1}^{\infty} \log\left(1+2\frac{\lambda_{ij}}{N_0}\right) + \frac{1}{N_0}\sum_{j=1}^{\infty} \frac{\lambda_{ij}}{\lambda_{ij}+\frac{N_0}{2}}\bar{X}_{ij}^2 + \log\alpha_i, \qquad i = 1,\ldots,M,$$

siendo la convergencia de la serie en el sentido de la media cuadrática.

Desde el punto de vista práctico, está solución presenta la dificultad de su dependencia explícita de los autovalores y autofunciones de $R_{S_i}(t,s)$. Para solucionar este problema se hará uso del desarrollo modificado de cada señal

$$\check{S}_{in}(t) = \sum_{j=1}^{n} \tilde{S}_{ij}\hat{\phi}_{ij}(t), \qquad t \in [0,T], \qquad i = 1,\ldots,M,$$

con $\tilde{S}_{ij} = \int_0^T \tilde{\phi}_{ij}(t) S_i(t)\, dt$ (m.c.).

Teorema 1. *Bajo cualquier hipótesis H_i, $i = 1,\ldots,M$, se tiene que*

$$\left\| X(t) - \sum_{j=1}^{n} \tilde{\tilde{X}}_{ij} \int_0^t \hat{\phi}_{ij}(\tau)\, d\tau \right\|_H \xrightarrow{n\uparrow\infty} 0,$$

donde $\tilde{\tilde{X}}_{ij} = \int_0^T \tilde{\phi}_{ij}(t)\, dX(t)$ *(m.c.).*

Demostración

En primer lugar, observar que

$$\bar{X}_{ij} = S_{ij} + \bar{W}_{ij},$$
$$\tilde{\bar{X}}_{ij} = \tilde{S}_{ij} + \tilde{\bar{W}}_{ij},$$

con $\bar{W}_{ij} = \int_0^T \phi_{ij}(t)\, dW(t)$ y $\tilde{\bar{W}}_{ij} = \int_0^T \tilde{\phi}_{ij}(t)\, dW(t)$.

Así, tenemos que

$$\begin{aligned} \left\| \bar{X}_{ij} - \tilde{\bar{X}}_{ij} \right\|_H &\le \left\| \tilde{S}_{ij} - \tilde{S}_{ij} \right\|_H + \left\| \bar{W}_{ij} - \tilde{\bar{W}}_{ij} \right\|_H \\ &\le \left(\sqrt{T} \left\| S_i(t_0) \right\|_H + \sqrt{\frac{N_0}{2}} \right) \left\| \phi_{ij}(t) - \tilde{\phi}_{ij}(t) \right\|_2 \xrightarrow{k\uparrow\infty} 0, \end{aligned} \tag{8}$$

donde se ha aplicado el Teorema del valor medio.

Por otro lado, se tiene que

$$\begin{aligned} &\left\| X(t) - \sum_{j=1}^{n} \tilde{\tilde{X}}_{ij} \int_0^t \hat{\phi}_{ij}(\tau)\, d\tau \right\|_H \\ &\le \left\| X(t) - \sum_{j=1}^{n} \bar{X}_{ij} \int_0^t \phi_{ij}(\tau)\, d\tau \right\|_H + \sum_{j=1}^{n} \left\| \bar{X}_{ij} \int_0^t \phi_{ij}(\tau)\, d\tau - \tilde{\tilde{X}}_{ij} \int_0^t \hat{\phi}_{ij}(\tau)\, d\tau \right\|_H. \end{aligned} \tag{9}$$

El primer término de (9) es convergente a cero por (6). Para demostrar que el segundo también tiende a cero, obsérvese que éste puede ser acotado de la siguiente manera:

$$(10)\quad \sum_{j=1}^{n}\left\|\bar{X}_{ij}\int_0^t\phi_{ij}(\tau)\,d\tau-\tilde{\bar{X}}_{ij}\int_0^t\hat{\phi}_{ij}(\tau)\,d\tau\right\|_H$$

$$\leq\sum_{i=1}^{n}\|\bar{X}_{ij}-\tilde{\bar{X}}_{ij}\|_H\left|\int_0^t\phi_{ij}(\tau)\,d\tau\right|+\sum_{i=1}^{n}\|\tilde{\bar{X}}_{ij}\|_H\left|\int_0^t\phi_{ij}(\tau)\,d\tau-\int_0^t\hat{\phi}_{ij}(\tau)\,d\tau\right|$$

$$\leq\sqrt{T}\sum_{i=1}^{n}\|\bar{X}_{ij}-\tilde{\bar{X}}_{ij}\|_H+K_1\sum_{i=1}^{n}\left|\int_0^t\phi_{ij}(\tau)\,d\tau-\int_0^t\hat{\phi}_{ij}(\tau)\,d\tau\right|,$$

donde la última desigualdad se ha obtenido teniendo en cuenta que

$$\left|\int_0^t\phi_{ij}(\tau)\,d\tau\right|\leq\left(\int_0^T\phi_{ij}^2(\tau)\,d\tau\right)^{\frac{1}{2}}\sqrt{T}=\sqrt{T}$$

$$\|\tilde{\bar{X}}_{ij}\|_H=\left(\tilde{\lambda}_{ij}+\frac{N_0}{2}\right)^{\frac{1}{2}}\leq\left(\int_0^T R_{S_i}(t,t)\,dt+\frac{N_0}{2}\right)^{\frac{1}{2}}=K_1<\infty.$$

Finalmente, por (8) y por las propiedades de convergencia de $\hat{\phi}_j(t)$, cuando $k\to\infty$, podemos elegir para cada n, un valor suficientemente grande de k, de manera que (10) sea arbitrariamente pequeño, con lo cual se demuestra el resultado. □

La consecuencia más importante que deducimos de este resultado es la posibilidad de aproximar el problema original (6) por el siguiente:

$$H_i: X(t)=\sum_{j=1}^{n}\tilde{\bar{X}}_{ij}\int_0^t\hat{\phi}_{ij}(\tau)\,d\tau,\qquad 0\leq t\leq T,\qquad i=1,\ldots,M,$$

para el cual el criterio óptimo es

$$\text{«elegir la hipótesis } H_k \text{ si } \tilde{I}_{kn}(X)=\max_{i=1,\ldots,M}\tilde{I}_{in}(X)\text{»},\qquad (\text{C1}a)$$

donde

$$\tilde{I}_{in}(X)=-\frac{1}{2}\sum_{j=1}^{n}\log\left(1+2\frac{\tilde{\lambda}_{ij}}{N_0}\right)+\frac{1}{N_0}\sum_{j=1}^{n}\frac{\tilde{\lambda}_{ij}}{\tilde{\lambda}_{ij}+\frac{N_0}{2}}\tilde{\bar{X}}_{ij}^2+\log\alpha_i,\qquad i=1,\ldots,M.$$

Teorema 2. *Bajo cualquier hipótesis H_i, $i=1,\ldots,M$, se tiene que*

$$\left\|I_i(X)-\tilde{I}_{in}(X)\right\|_H\xrightarrow{n\uparrow\infty}0.$$

Demostración

$$\left\| I_i(X) - \tilde{I}_{in}(X) \right\|_H \leq \mathcal{A}_1 + \mathcal{A}_2,$$

con

$$\mathcal{A}_1 = \left\| I_i(X) + \frac{1}{2}\sum_{j=1}^{n} \log\left(1+2\frac{\lambda_{ij}}{N_0}\right) - \frac{1}{N_0}\sum_{j=1}^{n} \frac{\lambda_{ij}}{\lambda_{ij}+\frac{N_0}{2}} \bar{X}_{ij}^2 - \log \alpha_i \right\|_H,$$

$$\mathcal{A}_2 = \frac{1}{2}\sum_{j=1}^{n} \left| \log\left(1+2\frac{\lambda_{ij}}{N_0}\right) - \log\left(1+2\frac{\tilde{\lambda}_{ij}}{N_0}\right) \right| + \frac{1}{N_0}\sum_{j=1}^{n} \left\| \frac{\lambda_{ij}}{\lambda_{ij}+\frac{N_0}{2}} \bar{X}_{ij}^2 - \frac{\tilde{\lambda}_{ij}}{\tilde{\lambda}_{ij}+\frac{N_0}{2}} \tilde{\bar{X}}_{ij}^2 \right\|_H.$$

El término $\mathcal{A}_1$ converge a cero por (7). Por otro lado, puesto que las variables aleatorias $\bar{X}_{ij} - \tilde{\bar{X}}_{ij}$ y $\bar{X}_{ij} + \tilde{\bar{X}}_{ij}$ son normales, aplicando la desigualdad de Cauchy-Schwarz obtenemos que

$$\begin{aligned} \left\| \bar{X}_{ij}^2 - \tilde{\bar{X}}_{ij}^2 \right\|_H &= \left(E\left[(\bar{X}_{ij} - \tilde{\bar{X}}_{ij})(\bar{X}_{ij} + \tilde{\bar{X}}_{ij}) \right]^2 \right)^{\frac{1}{2}} \\ &\leq \left(E\left[\bar{X}_{ij} - \tilde{\bar{X}}_{ij} \right]^4 E\left[\bar{X}_{ij} + \tilde{\bar{X}}_{ij} \right]^4 \right)^{\frac{1}{4}} \\ &= \sqrt{3} \left\| \bar{X}_{ij} - \tilde{\bar{X}}_{ij} \right\|_H \left\| \bar{X}_{ij} + \tilde{\bar{X}}_{ij} \right\|_H \\ &\leq K_2 \left\| \bar{X}_{ij} - \tilde{\bar{X}}_{ij} \right\|_H, \end{aligned} \tag{11}$$

donde $K_2 = 2\sqrt{3}K_1$.

Por tanto, utilizando (8) se tiene que

$$\left\| \bar{X}_{ij}^2 - \tilde{\bar{X}}_{ij}^2 \right\|_H \xrightarrow{k\uparrow\infty} 0.$$

Además, como se verifica que

$$\left| \frac{\lambda_{ij}}{\lambda_{ij}+\frac{N_0}{2}} - \frac{\tilde{\lambda}_{ij}}{\tilde{\lambda}_{ij}+\frac{N_0}{2}} \right| \xrightarrow{k\uparrow\infty} 0,$$

obtenemos que

$$\left\| \frac{\lambda_{ij}}{\lambda_{ij}+\frac{N_0}{2}} \bar{X}_{ij}^2 - \frac{\tilde{\lambda}_{ij}}{\tilde{\lambda}_{ij}+\frac{N_0}{2}} \tilde{\bar{X}}_{ij}^2 \right\|_H \xrightarrow{k\uparrow\infty} 0.$$

Así, $\mathcal{A}_2$ tiende a cero y el resultado se cumple. □

Por tanto, podemos concluir que se puede utilizar el criterio (C1a) como solución alternativa a la regla de decisión óptima (C1) en aquellos casos en los cuales no sea posible la obtención de los verdaderos autovalores y autofunciones.

Por otra parte, en Van Trees (1971) se demuestra, utilizando el estimador filtrado de la señal, que el estadístico $l_i(X)$ (3) puede escribirse de la forma

$$l_i(X) = \frac{2}{N_0}\int_0^T \hat{S}_i(t)\,dX(t) - \frac{1}{N_0}\int_0^T \hat{S}_i^{\,2}(t)\,dt + \log\alpha_i, \qquad i = 1,\ldots,M,$$

denominada forma estimadora-correladora, donde $\hat{S}_i(t) = \int_0^t h_i(t,\tau)\,dX(\tau)$ es el estimador causal de la señal $S_i(t)$ obtenido a partir de la señal perturbada por ruido blanco (bajo H_i). La función de respuesta a impulsos $h_i(t,\tau)$ satisface la ecuación integral

$$\int_0^t h_i(t,s)R_{S_i}(\tau,s)\,ds + \frac{N_0}{2}h_i(t,\tau) = R_{S_i}(\tau,t), \qquad 0 \le \tau \le t \le T.$$

La resolución de esta ecuación es difícil en general, lo que imposibilita la obtención del estimador óptimo. Por ello, a continuación se propone un nuevo estadístico $\hat{l}_{in}(X)$ que aproxima al verdadero, basado en el estimador subóptimo

$$\hat{S}_{in}(t) = \int_0^t \tilde{h}_{in}(t,\tau)\,dX(\tau),$$

donde $\tilde{h}_{in}(t,\tau)$ es la solución de la ecuación integral

$$\int_0^t \tilde{h}_{in}(t,s)R_{\check{S}_{in}}(\tau,s)\,ds + \frac{N_0}{2}\tilde{h}_{in}(t,\tau) = R_{\check{S}_{in}}(\tau,t), \qquad 0 \le \tau \le t \le T,$$

con $R_{\check{S}_{in}}(t,s)$ la función de covarianza de $\check{S}_{in}(t)$. Esta solución tiene la expresión

$$\tilde{h}_{in}(t,\tau) = \frac{2}{N_0}\hat{\Phi}'_{in}(t)\left[\tilde{\Lambda}_{in}^{-1} + \frac{2}{N_0}\int_0^t \hat{\Phi}_{in}(s)\hat{\Phi}'_{in}(s)ds\right]^{-1}\hat{\Phi}_{in}(\tau), \qquad 0 \le \tau \le t \le T,$$

siendo $\hat{\Phi}_{in}(t)$ un vector $n \times 1$ con entrada j-ésima la autofunción modificada $\hat{\phi}_{ij}(t)$ y $\tilde{\Lambda}_{in}$ una matriz diagonal con elementos los autovalores aproximados $\tilde{\lambda}_{ij}$, con $j = 1,\ldots,n$.

En Navarro-Moreno y otros (2000) se demuestra que esta función de respuesta subóptima converge a la óptima en el sentido

$$\left\|h_i(t,\cdot) - \tilde{h}_{in}(t,\cdot)\right\|_2 \xrightarrow{n\uparrow\infty} 0,$$

uniformemente en $t \in [0,T]$. Utilizando este hecho, en Navarro-Moreno y otros (2000) se prueba también que

$$\left\|\hat{S}_i(t)-\hat{S}_{in}(t)\right\|_H \xrightarrow{n\uparrow\infty} 0,$$

uniformemente en $t\in[0,T]$. Tal estadístico, $\hat{I}_{in}(X)$, se define de la forma

$$\hat{I}_{in}(X)=\frac{2}{N_0}\int_0^T \hat{S}_{in}(t)\,dX(t)-\frac{1}{N_0}\int_0^T \hat{S}_{in}^2(t)\,dt+\log\alpha_i, \qquad i=1,\ldots,M.$$

Teorema 3. *Bajo cualquier hipótesis H_i, $i=1,\ldots,M$, se tiene que*

$$\left\|I_i(X)-\hat{I}_{in}(X)\right\|_H \xrightarrow{n\uparrow\infty} 0.$$

Demostración

$$\left\|I_i(X)-\hat{I}_{in}(X)\right\|_H$$

$$\leq \frac{2}{N_0}\left\|\int_0^T \hat{S}_i(t)\,dX(t)-\int_0^T \hat{S}_{in}(t)\,dX(t)\right\|_H \tag{12}$$

$$+\frac{1}{N_0}\left\|\int_0^T \hat{S}_i^2(t)\,dt-\int_0^T \hat{S}_{in}^2(t)\,dt\right\|_H. \tag{13}$$

La norma en (12) puede ser acotada de la forma

$$\left\|\int_0^T \hat{S}_i(t)\,dX(t)-\int_0^T \hat{S}_{in}(t)\,dX(t)\right\|_H$$

$$\leq \left\|\int_0^T \hat{S}_i(t)\,dW(t)-\int_0^T \hat{S}_{in}(t)\,dW(t)\right\|_H \tag{14}$$

$$+\left\|\int_0^T S_i(t)\left(\hat{S}_i(t)-\hat{S}_{in}(t)\right)dt\right\|_H. \tag{15}$$

Para (14) tenemos que

$$\left\|\int_0^T \hat{S}_i(t)\,dW(t)-\int_0^T \hat{S}_{in}(t)\,dW(t)\right\|_H=\sqrt{\frac{N_0}{2}}\left(\int_0^T \left\|\hat{S}_i(t)-\hat{S}_{in}(t)\right\|_H^2 dt\right)^{1/2} \xrightarrow{n\uparrow\infty} 0,$$

porque $\left\|\hat{S}_i(t)-\hat{S}_{in}(t)\right\|_H \xrightarrow{n\uparrow\infty} 0$ uniformemente en $t\in[0,T]$.

Puesto que $S_i(t)$ y $\hat{S}_i(t)-\hat{S}_{in}(t)$ son variables normales para cada t, haciendo un razonamiento similar al realizado en (11) y utilizando el Teorema del valor medio obtenemos para (15)

$$\left\|\int_0^T S_i(t)\left(\hat{S}_i(t)-\hat{S}_{in}(t)\right)dt\right\|_H \leq \int_0^T \left\|S_i(t)\left(\hat{S}_i(t)-\hat{S}_{in}(t)\right)\right\|_H dt$$
$$\leq \sqrt{3}\|S_i(t_0)\|_H \int_0^T \left\|\hat{S}_i(t)-\hat{S}_{in}(t)\right\|_H dt \xrightarrow{n\uparrow\infty} 0,$$

y entonces (12) tiende a cero.

Finalmente, también es inmediato comprobar que

$$\left\|\hat{S}_i^2(t)-\hat{S}_{in}^2(t)\right\|_H \xrightarrow{n\uparrow\infty} 0,$$

uniformemente en $t \in [0,T]$, con lo cual (13) tiende a cero, y así se sigue la afirmación. □

Como consecuencia obtenemos un nuevo criterio para resolver el problema (1),

«elegir la hipótesis H_k si $\hat{I}_{kn}(X) = \max_{i=1,\ldots,M} \hat{I}_{in}(X)$». (C1*b*)

Nota 3. Las soluciones (C1*a*) y (C1*b*) generalizan a las obtenidas en Ruiz-Molina y otros (2001) para el caso binario simple.

3. EJEMPLO

En la práctica, el cálculo de $P(\varepsilon)$ (2) que se comete con el criterio (C1) es difícil puesto que no existe un método general para determinarla. Normalmente se utilizan acotaciones o aproximaciones de ésta. Así, por ejemplo, si $\alpha_1 = \alpha_2 = 1/2$ en Van Trees (1971) puede obtenerse para el caso binario simple la siguiente aproximación:

$$P(\varepsilon) \simeq \frac{1}{2}\exp\left(\mu(s_m)+\frac{s_m^2}{2}\ddot{\mu}(s_m)\right)\left(1-\Phi\left(s_m\sqrt{\ddot{\mu}(s_m)}\right)\right)$$
(16)
$$+\frac{1}{2}\exp\left(\mu(s_m)+\frac{(1-s_m)^2}{2}\ddot{\mu}(s_m)\right)\left(1-\Phi\left((1-s_m)\sqrt{\ddot{\mu}(s_m)}\right)\right),$$

con Φ la función de distribución de una normal estándar y

$$\mu(s) = \frac{1}{2}\sum_{j=1}^{\infty}\log\left(\frac{(1+\frac{2\lambda_j}{N_0})^{(1-s)}}{1+\frac{2(1-s)\lambda_j}{N_0}}\right), \qquad 0 \leq s \leq 1, \tag{17}$$

siendo $\{\lambda_j\}_{j=1}^{\infty}$ el sistema de autovalores correspondiente a $R_S(t,s)$ y s_m es el punto tal que $\dot{\mu}(s_m) = \frac{d\mu(s)}{ds}\Big|_{s=s_m} = 0$.

En el caso de considerar el desarrollo modificado (5) de la señal $S(t)$ entonces la expresión (17) queda de la forma

$$\tilde{\mu}_n(s) = \frac{1}{2}\sum_{j=1}^{n} \log\left(\frac{(1+\frac{2\tilde{\lambda}_j}{N_0})^{(1-s)}}{1+\frac{2(1-s)\tilde{\lambda}_j}{N_0}}\right), \qquad 0 \le s \le 1,$$

Por tanto, si denotamos por $\tilde{P}_n(\varepsilon)$ a la probabilidad total del error que se comete con el criterio (C1a) obtenemos que

$$\begin{aligned}(18)\qquad \tilde{P}_n(\varepsilon) \simeq & \frac{1}{2}\exp\left(\tilde{\mu}_n(\tilde{s}_m) + \frac{\tilde{s}_m^2}{2}\ddot{\tilde{\mu}}_n(\tilde{s}_m)\right)\left(1-\Phi\left(\tilde{s}_m\sqrt{\ddot{\tilde{\mu}}_n(\tilde{s}_m)}\right)\right) \\ & + \frac{1}{2}\exp\left(\tilde{\mu}_n(\tilde{s}_m) + \frac{(1-\tilde{s}_m)^2}{2}\ddot{\tilde{\mu}}_n(\tilde{s}_m)\right)\left(1-\Phi\left((1-\tilde{s}_m)\sqrt{\ddot{\tilde{\mu}}_n(\tilde{s}_m)}\right)\right),\end{aligned}$$

con $\tilde{s}_m$ verificando que $\dot{\tilde{\mu}}_n(\tilde{s}_m) = 0$.

A continuación, se ilustra el funcionamiento del método propuesto comparando la aproximación (16) con (18) en un ejemplo práctico. Para ello, vamos a considerar que $\{S(t); t \in [0,1]\}$ es un proceso Gaussiano, centrado y con función de covarianza $R_S(t,s) = \min(t,s) - ts$. Además, supondremos que $W(t)$ es el proceso de Wiener estándar.

En Gutiérrez y otros (1992) puede encontrarse que los autovalores verdaderos para esta señal son de la forma $\lambda_j = \frac{1}{j^2\pi^2}, j = 1,2,\ldots$ Por tanto, aplicando (16) obtenemos que $P(\varepsilon) \simeq 0.48579$.

Por otro lado, en Gutiérrez y otros (1992) utilizando una base de funciones trigonométricas y otra de polinomios de Legendre con $k = 5$ se obtuvieron los siguientes autovalores aproximados:

Aut. aprox.	**Base trigon.**	**Base de polin.**
$\tilde{\lambda}_1$	0.10111	0.10132
$\tilde{\lambda}_2$	0.02533	0.01667
$\tilde{\lambda}_3$	0.01100	0.01111
$\tilde{\lambda}_4$	0.00633	0.00264
$\tilde{\lambda}_5$	0.00289	0.00135

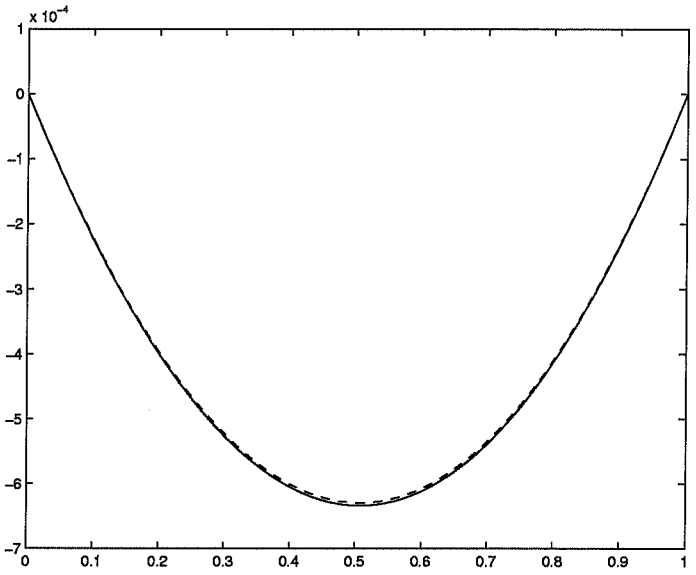


Figura 1. $\mu(s)$ (linea continua) y $\tilde{\mu}_5(s)$ correspondiente a la base trigonométrica.

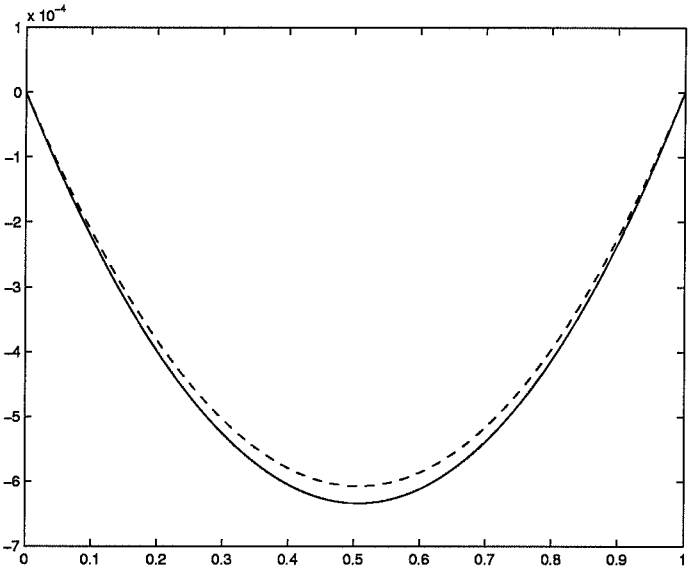


Figura 2. $\mu(s)$ (linea continua) y $\tilde{\mu}_5(s)$ correspondiente a la base de polinomios de Legendre.

De esta manera, aplicando (18) se obtiene que $\tilde{P}_5(\varepsilon) \simeq 0.48584$ para la base trigonométrica y $\tilde{P}_5(\varepsilon) \simeq 0.48609$ para la base de polinomios de Legendre.

Finalmente, en las Figuras 1 y 2 se representan $\mu(s)$ y $\tilde{\mu}_5(s)$ para las dos bases.

4. CONCLUSIONES

En este trabajo se ha propuesto una nueva solución para el problema de detección de M señales Gaussianas en ruido Gaussiano blanco. Haciendo uso del desarrollo modificado de un proceso estocástico se ha obtenido una regla de decisión subóptima que no presenta la dificultad del cálculo de autovalores y autofunciones.

Por otro lado, se ha propuesto una segunda solución aproximada basada en un estimador subóptimo obtenido mediante el desarrollo modificado. Este estimador subóptimo supone una alternativa al filtro de Kalman cuando no es posible obtener una ecuación de estados para la señal.

La principal ventaja de estas dos soluciones es su facilidad de implementación desde el punto de vista computacional y aunque ninguna proporcione la solución óptima, sin embargo, se puede controlar el error que se comete en la aproximación a la óptima sin mas que elegir adecuadamente el número de términos que conforman el desarrollo modificado.

Para finalizar, comentar que aunque el criterio utilizado haya sido el criterio de minimizar la probabilidad total del error, los resultados obtenidos pueden adaptarse sin dificultades a otro tipo de criterios, como es el criterio de Bayes. La importancia de este trabajo radica en las aproximaciones proporcionadas de la derivada de Radon-Nikodym, ya que ésta es la pieza fundamental a la hora de determinar las diversas reglas óptimas que proporcionan otros criterios. La razón de haber realizado el estudio de detección múltiple utilizando el criterio mencionado se debe a su facilidad de implementación, por lo que es usualmente utilizado en la literatura.

5. REFERENCIAS

Baker, C. T. H. (1977). *The Numerical Treatment of Integral Equations*. Oxford University Press, Oxford.

Gutiérrez, R., Ruiz, J. C. and Valderrama, M. J. (1992). «On the Numerical Expansion of a Second Order Stochastic Process». *Appl. Stochastic Models and Data Anal.*, 8(2), 67-77.

Kadota, T. T. (1965). «Optimum Reception of M-ary Gaussian Signals in Gaussian Noise». *Bell System Tech. J.*, 44, 2187-2197.

Navarro-Moreno, J., Ruiz-Molina, J. C. and Valderrama, M. J. (2000). «A Solution to Linear Estimation Problems Using Approximate Karhunen-Loève Expansions». *IEEE, Trans. Information Theory*, 46(4), 1677-1682.

Poor, H. V. (1994). *An Introduction to Signal Detection and Estimation*. Springer-Verlag, New York, 2nd edition.

Ruiz-Molina, J. C., Navarro, J. and Valderrama, M. J. (1999). «Differentiation of the Modified Approximative Karhunen-Loève Expansion of a Stochastic Process». *Statist. Prob. Lett.*, 42, 91-98.

Ruiz-Molina, J. C., Navarro-Moreno, J. and Oya, A. (2001). «Signal Detection Using Approximate Karhunen-Loève Expansions». *IEEE, Trans. Information Theory*, 47(4), 1672-1680.

Van Trees, H. L. (1971). *Detection, Estimation, and Modulation Theory. Part III*. Wiley, New York.

ENGLISH SUMMARY

DETECTION OF M GAUSSIAN SIGNALS USING THE MODIFIED APPROXIMATE KARHUNEN-LOÈVE EXPANSION OF A STOCHASTIC PROCESS*

JESÚS NAVARRO-MORENO
JUAN CARLOS RUIZ-MOLINA
Universidad de Jaén*

A new approach to the problem of detecting M Gaussian signals in white noise is proposed by using the modified approximate Karhunen-Loève expansion of a stochastic process. The solutions obtained do not require the calculation of true eigenvalues and eigenfunctions and they are easily implementable in real situations.

Keywords: Modified Approximate Karhunen-Loève Expansion of a Stochastic Process, Detection of M Gaussian Signals in White Gaussian Noise

AMS Classification (MSC 2000): 60G12

* This work was supported in part by Project BFM2000-1103, Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I), Ministerio de Ciencia y Tecnología, Spain.
* Departamento de Estadística e I.O. Universidad de Jaén. Paraje Las Lagunillas, s/n. 23071 Jaén. E-mail: jnavarro,jcruiz@ujaen.es.
–Received March 2001.
–Accepted September 2001.

1. INTRODUCTION

In this paper we treat the problem of detecting M Gaussian signals in additive white Gaussian noise. This problem can be formulated in the following way: suppose there are M stochastic signals, $S_i(t)$, $i=1,\ldots,M$, with *a priori* probabilities α_i, $0<\alpha_i<1$ and $\sum_{i=1}^{M}\alpha_i=1$, for transmission. The observation process $Y(t)$ consists of one of these M signals and an additive white Gaussian noise $N(t)$, i.e.

$$H_i: Y(t)=S_i(t)+N(t), \qquad 0\leq t\leq T, \qquad i=1,\ldots,M,$$

where the signals $\{S_i(t);t\in[0,T]\}$ are Gaussian stochastic processes with zero-mean and continuous in quadratic mean, and $\{N(t);t\in[0,T]\}$ is a white Gaussian noise with spectral height $N_0/2$ and independent of each signal. Denote by $R_{S_i}(t,s)$ the covariance function of $S_i(t)$.

Since the white noise does not exist as a mathematical random process in the ordinary sense, we convert the above model to one that is better posed by integrating the observation process (Poor, 1994). Thus, we obtain the following model:

$$(1) \qquad H_i: X(t)=\int_0^t S_i(\tau)\,d\tau+W(t), \qquad 0\leq t\leq T, \qquad i=1,\ldots,M,$$

where $\{W(t);t\in[0,T]\}$ is a Wiener process with spectral height $N_0/2$ and independent of each signal.

Our objective is to observe the observation process $X(t)$ and to decide which one of the M Gaussian signals, $S_i(t)$, $i=1,\ldots,M$, must have been received.

Kadota (1965) proposed the optimum receiver by using the minimum error-probability criterion in the following way:

$$\text{«to choose the hypothesis } H_k \text{ if } I_k(X)=\max_{i=1,\ldots,M} I_i(X)\text{»},$$

where

$$I_i(X)=\log\frac{dP_i}{dP_0}(X)+\log\alpha_i, \qquad i=1,\ldots,M,$$

with $\frac{dP_i}{dP_0}(X)$ the Radon-Nikodym derivative of P_i (probability measure described by H_i) with respect to P_0 (probability measure corresponding to $W(t)$).

From the practical standpoint, a method is necessary to compute the Radon-Nikodym derivative $\frac{dP_i}{dP_0}(X)$. One of the most widely used methodologies to solve this difficulty

is the Karhunen-Loève expansion of a stochastic process (Poor, 1994). Thus, by denoting the corresponding eigenvalues and eigenfunctions of $R_{S_i}(t,s)$ by $\{\lambda_{ij}\}_{j=1}^{\infty}$ and $\{\phi_{ij}(t)\}_{j=1}^{\infty}$, respectively, it is shown in Poor (1994) that

$$(2)\quad l_i(X) = -\frac{1}{2}\sum_{j=1}^{\infty}\log\left(1+2\frac{\lambda_{ij}}{N_0}\right)+\frac{1}{N_0}\sum_{j=1}^{\infty}\frac{\lambda_{ij}}{\lambda_{ij}+\frac{N_0}{2}}\bar{X}_{ij}^2+\log\alpha_i, \qquad i=1,\ldots,M,$$

where the convergence is in the mean-square sense and $\bar{X}_{ij}=\int_0^T\phi_{ij}(t)\,dX(t)$ (in q.m.).

The drawback of this methodology is that the detection statistic $l_i(X)$ depends explicitly on eigenvalues and eigenfunctions of the covariance kernel involved and there is no standard method to compute them.

Therefore, the statistic $l_i(X)$ is usually rewritten in the following form called the *estimator-correlator* representation (Van Trees, 1971):

$$(3)\quad l_i(X)=\frac{2}{N_0}\int_0^T\hat{S}_i(t)\,dX(t)-\frac{1}{N_0}\int_0^T\hat{S}_i^{\,2}(t)\,dt+\log\alpha_i, \qquad i=1,\ldots,M,$$

where $\hat{S}_i(t)=\int_0^t h_i(t,\tau)\,dX(\tau)$ is the causal estimator of $S_i(t)$ from signal plus white noise (under H_i). The impulse response function $h_i(t,\tau)$ satisfies the integral equation

$$\int_0^t h_i(t,s)R_{S_i}(\tau,s)\,ds+\frac{N_0}{2}h_i(t,\tau)=R_{S_i}(\tau,t), \qquad 0\le\tau\le t\le T.$$

However, there is not a general method to solve this integral equation and due to this, (3) is usually impractical to compute.

The first objective of this paper is to show that the modified approximate Karhunen-Loève expansion of a stochastic process (Ruiz-Molina *et al.*, 1999) which is based on the approximate eigenvalues and eigenfunctions calculated from the Rayleigh-Ritz method, allows us to obtain a solution to the detection problem (1) circumventing the above difficulties. Such a solution is easily implementable on a computer and does not require the calculation of true eigenvalues and eigenfunctions.

On the other hand, it is proposed an alternative form to the *estimator-correlator* representation which is based on a suboptimum estimator obtained from the modified approximate Karhunen-Loève expansion. The main advantage of this suboptimum estimator is that it can be derived through an efficient algorithm similar to the Kalman filter (Navarro-Moreno *et al.*, 2000).

For that, in the next subsection the modified approximate Karhunen-Loève expansion of a stochastic process introduced in Ruiz-Molina *et al.* (1999) is summarized.

1.1. Modified Approximate Karhunen-Loève Expansion

Let $\{X(t); t \in [0,T]\}$ be a stochastic process defined on the probability space $(\Omega, \mathcal{B}, P)$, with zero-mean, continuous in quadratic mean and covariance function $R_X(t,s)$. Let $\{\tilde{\lambda}_j\}_{j=1}^n$ and $\{\tilde{\phi}_j(t)\}_{j=1}^n$ be its corresponding approximate eigenvalues and eigenfunctions calculated from the Rayleigh-Ritz method (Baker, 1977). The modified approximate Karhunen-Loève expansion of the process $X(t)$ is defined in the following way (Ruiz-Molina *et al.*, 1999):

$$\breve{X}_n(t) = \sum_{j=1}^{n} \tilde{X}_j \hat{\phi}_j(t), \qquad t \in [0,T],$$

where

$$\hat{\phi}_j(t) = \frac{1}{\tilde{\lambda}_j} \int_0^T R_X(t,s)\tilde{\phi}_j(s)\,ds, \qquad t \in [0,T]$$

and $\tilde{X}_j = \int_0^T \tilde{\phi}_j(t)X(t)\,dt$ (in q.m.), with $E[\tilde{X}_j\tilde{X}_l] = \tilde{\lambda}_j\delta_{jl}$.

This expansion satisfies

$$\left\|X(t) - \breve{X}_n(t)\right\|_H \xrightarrow{n\uparrow\infty} 0,$$

uniformly in $t \in [0,T]$, where $\|Z\|_H = (E[Z]^2)^{1/2}$, with $Z \in L^2(\Omega, \mathcal{B}, P)$.

2. THE RESULTS

As we have already mentioned, the major drawback of (2) is its explicit dependence on eigenvalues and eigenfunctions of $R_{S_i}(t,s)$. This difficulty can be resolved by using the modified expansion of $S_i(t)$

$$\breve{S}_{in}(t) = \sum_{j=1}^{n} \tilde{S}_{ij}\hat{\phi}_{ij}(t), \qquad t \in [0,T], \qquad i = 1,\ldots,M,$$

with $\tilde{S}_{ij} = \int_0^T \tilde{\phi}_{ij}(t)S_i(t)\,dt$ (in q.m.).

Theorem 1. *Under any hypothesis H_i, $i = 1,\ldots,M$, we have*

$$\left\|I_i(X) - \tilde{I}_{in}(X)\right\|_H \xrightarrow{n\uparrow\infty} 0,$$

where

$$\tilde{I}_{in}(X) = -\frac{1}{2}\sum_{j=1}^{n}\log\left(1+2\frac{\tilde{\lambda}_{ij}}{N_0}\right)+\frac{1}{N_0}\sum_{j=1}^{n}\frac{\tilde{\lambda}_{ij}}{\tilde{\lambda}_{ij}+\frac{N_0}{2}}\tilde{\tilde{X}}_{ij}^2+\log\alpha_i, \qquad i=1,\ldots,M,$$

with $\tilde{\tilde{X}}_{ij} = \int_0^T \tilde{\phi}_{ij}(t)\,dX(t)$ *(in q.m.).*

Hence, we obtain the following approach to the problem (1):

«to choose the hypothesis H_k if $\tilde{I}_{kn}(X) = \max_{i=1,\ldots,M} \tilde{I}_{in}(X)$».

On the other hand, we propose here an alternative statistic to (3), $\hat{I}_{in}(X)$, which is based on the suboptimum estimator (Navarro *et al.*, 2000)

$$\hat{S}_{in}(t) = \int_0^t \tilde{h}_{in}(t,\tau)\,dX(\tau),$$

where $\tilde{h}_{in}(t,\tau)$ verifies the integral equation

$$\int_0^t \tilde{h}_{in}(t,s)R_{\breve{S}_{in}}(\tau,s)\,ds+\frac{N_0}{2}\tilde{h}_{in}(t,\tau) = R_{\breve{S}_{in}}(\tau,t), \qquad 0\le\tau\le t\le T,$$

where $R_{\breve{S}_{in}}(t,s)$ is the covariance function of $\breve{S}_{in}(t)$. Such a statistic $\hat{I}_{in}(X)$ is defined in the following way:

$$\hat{I}_{in}(X) = \frac{2}{N_0}\int_0^T \hat{S}_{in}(t)\,dX(t)-\frac{1}{N_0}\int_0^T \hat{S}_{in}^2(t)\,dt+\log\alpha_i, \qquad i=1,\ldots,M.$$

Theorem 2. *Under any hypothesis H_i, $i=1,\ldots,M$, we have*

$$\left\|I_i(X)-\hat{I}_{in}(X)\right\|_H \xrightarrow{n\uparrow\infty} 0.$$

Hence, it is obtained a new criterion to solve the problem (1):

«to choose the hypothesis H_k if $\hat{I}_{kn}(X) = \max_{i=1,\ldots,M} \hat{I}_{in}(X)$».

QÜESTIIÓ, vol. 25, 3, p. 437-477, 2001

ESTIMACIÓN NO PARAMÉTRICA DE LA FUNCIÓN DE RIESGO: APLICACIONES A SISMOLOGÍA

GRACIELA ESTÉVEZ PÉREZ
ALEJANDRO QUINTELA DEL RÍO
Universidad de A Coruña*

Se estudia la estimación de tipo no paramétrico de la función de riesgo o razón de fallo de una variable aleatoria real. A partir de una muestra $X_1, X_2, \ldots, X_n$ de datos no censurados y no necesariamente independientes, se considera un estimador cociente entre el estimador núcleo de la función de densidad y un estimador núcleo de la función de supervivencia, sobre el que se estudia el problema de selección del parámetro ventana.
Por medio de un estudio de simulación se observa la ventaja de utilizar este estimador frente al que estima la función de supervivencia a través de la función de distribución empírica. Finalmente, se realiza una aplicación práctica a datos de terremotos sucedidos en California y en Granada.

Palabras clave: Estimación no paramétrica, procesos fuertemente mixing, validación cruzada, razón de fallo

Clasificación AMS (MSC 2000): 62G05, 62G20, 62M99

*Departamento de Matemáticas-Facultad de Informática. Universidad de A Coruña. Campus de Elviña, s/n. 15071 A Coruña. Teléfono: 981167000, ext.: 2109.

–Recibido en julio de 1999.
–Aceptado en octubre de 2001.

1. INTRODUCCIÓN

En análisis de supervivencia, el interés se centra, por un lado, en una población homogénea en la que se considera un suceso puntual, normalmente llamado *fallo;* y, a la vez, en el tiempo de ocurrencia de dicho suceso, generalmente llamado *tiempo de fallo*. Para determinar el *tiempo de fallo* de forma precisa debe definirse sin ambigüedad el tiempo origen, la escala de medida de ese tiempo, y el significado del suceso puntual *fallo*. Tal suceso puede ser, por ejemplo, la muerte de pacientes tras algún tipo de transplante, la avería de algún componente de una máquina, la ruptura de algún material sometido a una fuerza o la ocurrencia de un terremoto. De esta forma, el tiempo de fallo será, respectivamente, el tiempo de vida después del transplante, el tiempo de duración de una máquina hasta su primera avería, el tiempo de ruptura del material y el tiempo entre terremotos sucesivos. Esto nos muestra como el ámbito de aplicación del análisis de supervivencia abarca disciplinas de índole tan diversa como la medicina, la química, estudios de fiabilidad o sismología.

Supongamos que X es una variable aleatoria real, continua, con función de densidad $f(\cdot)$ y función de distribución $F(\cdot)$. Además de estas formas matemáticamente equivalentes de representar la distribución de X, existen otras funciones particularmente útiles en análisis de supervivencia, como son la función de supervivencia o la función de razón de fallo.

La función de supervivencia se representa por $S(x)$ y se define como la probabilidad de que el tiempo de fallo sea superior a x, es decir,

$$S(x) = P(X > x).$$

De la definición de función de distribución $F(\cdot)$ y de su relación con la función de densidad $f(\cdot)$ se tiene que $f(x) = (-dS(x))/dx$. En aplicaciones médicas, la función de supervivencia, como su propio nombre indica, se interpreta como la probabilidad de que un individuo en estudio sobreviva al tiempo x, y ha sido ampliamente estudiada en contextos de censura y/o truncamiento (ver por ejemplo, Lee (1992) o Klein y Moeschberger (1997)).

La función de razón de fallo, que también es conocida como «función de azar», «fuerza de mortalidad», «función de intensidad» o «**función de riesgo**» se representa por $r(\cdot)$ y se define como

$$r(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x \mid X \geq x)}{\Delta x}. \tag{1}$$

Por la definición de probabilidad condicionada tenemos que

$$r(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}. \tag{2}$$

Si la variable X mide el tiempo de fallo, la función $r(\cdot)$ es una medida de la predisposición de fallo como una función del tiempo, en el sentido en que $r(x)\Delta x$ representa la proporción esperada de individuos que, sobreviviendo al tiempo x, fallan en el intervalo $(x, x+\Delta x)$. De esta forma, puede ser interpretada como el riesgo instantáneo de fallo en cada instante x sabiendo que el último fallo se ha producido en el instante 0.

En muchas situaciones prácticas es evidente que el conocimiento de la función de razón de fallo de una variable en estudio supondría grandes ventajas, de ahí el interés de estimar dicha función. Para tal propósito, el primer planteamiento que surge es considerar una familia paramétrica de distribuciones (Weibull, Gamma o Exponencial son algunos de los modelos más utilizados en casos prácticos) y, utilizando un conjunto de observaciones $X_1, \ldots, X_n$ de la variable X, estimar el vector de parámetros desconocido. En Cox y Oakes (1984) aparece una descripción de algunas de las familias paramétricas de distribuciones más usadas, así como criterios para la elección adecuada de tales familias. También, en su capítulo 3 puede verse una revisión de métodos de inferencia paramétrica basados en la función de verosimilitud. Otras obras de consulta sobre este tema son los anteriormente citados Lee (1992) y Klein y Moeschberger (1997).

Sin embargo, en muchos casos no se dispone de información suficiente para precisar la familia de distribuciones a la que pertenece la variable en estudio. En tales situaciones es deseable disponer de métodos no paramétricos para estimar la razón de fallo a partir de la información que proporciona un simple conjunto de observaciones $X_1, \ldots, X_n$.

El problema de estimación no paramétrica de la función de razón de fallo ha sido tratado con bastante profundidad en la literatura, bajo la hipótesis de datos independientes. Algunas referencias al respecto son Watson y Leadbetter (1964*a* y *b*), Ahmad (1976), Hollander y Proschan (1984), Prakasa Rao y Van Ryzin (1985), Singpurwalla y Wong (1983) y Hassani, Sarda y Vieu (1986).

En ciertas ocasiones, sin embargo, la hipótesis de independencia entre las observaciones $X_1, \ldots, X_n$ no tiene por qué ser adecuada. En estudios sismológicos, por ejemplo, es mucho más realista suponer que los tiempos entre terremotos son dependientes, aunque esta dependencia se vaya debilitando para movimientos separados temporalmente. Esto nos llevará a considerar que las observaciones pertenecen a una serie de tiempo —que supondremos estrictamente estacionaria— y que cumplirá alguna condición de dependencia que sea razonable desde un punto de vista práctico, y que permita una manipulación matemática sencilla.

En este trabajo consideraremos la estimación no paramétrica de la función de razón de fallo (2) bajo condiciones de dependencia entre los datos y no censura. En la sección 2 definimos el estimador utilizado y las condiciones matemáticas de dependencia. Detallamos también algunas propiedades matemáticas del mismo, junto con una manera de elegir el parámetro de suavización en los estimadores no paramétricos. En la sección 3 se exponen algunos ejemplos de simulación que evidencian la mejora de utilizar el estimador definido en la sección 1 frente a un estimador utilizado en Sarda y Vieu (1989). En la sección 4 se analiza el procedimiento de selección del parámetro ventana desde un punto de vista local. Finalmente, en la sección 5 se exponen dos ejemplos de utilización práctica de la estimación no paramétrica del riesgo instantáneo de ocurrencia de terremotos, en dos regiones geográficas y periodos de tiempo diferentes.

2. ESTIMACIÓN NÚCLEO DE LA FUNCIÓN DE RIESGO BAJO CONDICIONES DE DEPENDENCIA

Muchos de los estimadores de la función de riesgo o razón de fallo y, en particular, el que consideraremos en este trabajo, se definen como un cociente entre un estimador no paramétrico de la función de densidad $f(\cdot)$ y un estimador no paramétrico de la función de supervivencia $1-F(\cdot)$ (Watson y Leadbetter (1964*a* y *b*)). Nosotros consideraremos el estimador obtenido a partir de una muestra $X_1,\ldots,X_n$ dado por

$$r_h(x)=\frac{f_h(x)}{1-F_h(x)}, \tag{3}$$

donde $f_h(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)$ es el conocido estimador núcleo de Parzen-Rosenblatt, y $F_h(\cdot)$ es el estimador núcleo de $F(\cdot)$, definido por

$$F_h(x)=\frac{1}{n}\sum_{i=1}^{n}H\left(\frac{x-X_i}{h}\right)=\int_{-\infty}^{x}f_h(t)dt. \tag{4}$$

La función núcleo (kernel) $K:\mathbb{R}\longrightarrow\mathbb{R}$ verifica $\int K(u)\,du=1$ y $H(\cdot)$ se obtiene a partir de $K(\cdot)$ mediante $H(x)=\int_{-\infty}^{x}K(u)du$. El parámetro $h=h(n)\in\mathbb{R}^+$ es el llamado *parámetro de suavización* o *ventana,* que controla el grado de suavización inherente en todo proceso de estimación de tipo no paramétrico.

Este estimador presenta varias ventajas con respecto al estimador que utiliza la función de distribución empírica para estimar la función $F(\cdot)$:

$$r_{h,1}(x) = \frac{f_h(x)}{1 - F_n(x)}, \text{ con } F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leq x\}}. \tag{5}$$

En primer lugar, este último estimador no es continuo, puesto que $F_n(\cdot)$ tampoco lo es. Además, $F_n(\cdot)$ presenta una deficiencia relativa con respecto a $F_h(\cdot)$ en términos de error cuadrático medio (Reiss (1981)). Un ejemplo de esta deficiencia se pondrá de manifiesto en la sección 3 en simulaciones. Este estimador ha sido estudiado por Sarda y Vieu (1989) en distintos contextos de dependencia (entre ellos el de fuertemente mixing, definido más adelante). Dichos autores demuestran propiedades de consistencia fuerte casi segura para el estimador, y estudian las propiedades de optimalidad de los procedimientos de validación cruzada global y local para la selección del parámetro ventana.

El estimador dado en (3) ha sido estudiado por Youndjé, Sarda y Vieu (1996) en un contexto de independencia. Para poder trabajar bajo la hipótesis de que los datos de la muestra en estudio sean dependientes, consideramos la siguiente definición:

Definición 2.1. *Sea* $\{X_t; t \in \mathbb{N}\}$ *una sucesión de variables aleatorias, y sea* $\mathcal{F}_k^{k+n}$ *la* σ*-álgebra generada por las variables* $X_k, X_{k+1}, \ldots, X_{k+n}$. *Definamos*

$$\alpha(n) = \sup_{k \in \mathbb{N}} \left\{ |P(A \cap B) - P(A)P(B)| ;\ A \in \mathcal{F}_0^k,\ B \in \mathcal{F}_{k+n}^{\infty} \right\}$$

Entonces, se dice que $\{X_t : t \in \mathbb{N}\}$ *es fuertemente mixing o* α*-mixing si* $\lim_{n \to \infty} \alpha(n) = 0$.

Esta condición es de las más débiles entre las utilizadas en estudios con datos dependientes. La monografía de Doukhan (1994) es un buen referente para el estudio de éstas y otras condiciones similares. En dicha monografía se demuestra que un gran número de procesos dependientes cumplen condiciones de tipo mixing. Por ejemplo, los procesos ARMA estacionarios con ruido absolutamente continuo son fuertemente mixing. Para más referencias sobre estimación de tipo núcleo bajo hipótesis de datos mixing consultar, por ejemplo, Györfi *et al.* (1990).

Para poder trabajar en la práctica con $r_h(\cdot)$ es necesario elegir una función núcleo $K(\cdot)$, así como la ventana $h = h(n)$. Mientras la elección de $K(\cdot)$ no resulta demasiado importante, la elección de la ventana h juega un papel preponderante en el comportamiento del estimador. En Estévez y Quintela (1999) se estudian propiedades de consistencia puntual y uniforme de tipo casi seguro para el estimador $r_h(\cdot)$ bajo la suposición de que la muestra $X_1, \ldots, X_n$ es fuertemente mixing. Se prueba también la normalidad

asintótica del mismo y se describe un método de validación cruzada para la selección del parámetro ventana h, del que se prueba su optimalidad asintótica.

Una ventana es considerada óptima si minimiza algún criterio de error para el estimador $r_h(\cdot)$, como por ejemplo el *Error cuadrático promedio:*

$$ASE(h) = \frac{1}{n}\sum_{i=1}^{n}(r_h(X_i) - r(X_i))^2 W(X_i), \tag{6}$$

donde $W(\cdot)$ es alguna función de pesos no negativa, introducida para evitar el llamado «efecto frontera», propio de los estimadores no paramétricos de curvas (Gasser y Müller (1984)). También puede considerarse el *Error cuadrático integrado:*

$$ISE(h) = \int (r_h(x) - r(x))^2 W(x)\, dx. \tag{7}$$

o el *Error cuadrático medio integrado*, definido como la esperanza de $ISE(h)$. Sin embargo, debido a que r_h es un estimador definido como un cociente, este último error puede no existir, y se debe trabajar entonces con el error

$$MISE^*(h) = E\int \left[\left(\frac{f_h(x)}{1-F_h(x)} - \frac{f(x)}{1-F(x)}\right)\frac{1-F_h(x)}{1-F(x)}\right]^2 W(x)\, dx. \tag{8}$$

Estas tres medidas de error son asintóticamente equivalentes (Vieu (1991*a*)), por lo que es indiferente utilizar una u otra. Este último autor obtiene la siguiente expresión que especifica el comportamiento asintótico del error $MISE^*(h)$:

$$MISE^*(h) = C_1 h^{2k} + C_2 \frac{1}{nh} + o(MISE^*(h)), \tag{9}$$

donde

$$C_1 = \int \left[\left(f^{(k)}(x) + F^{(k)}(x) r(x)\right)\frac{1}{k!}\int u^k K(u)\, du\right]^2 \frac{W(x) f(x)}{(1-F(x))^2} dx$$

y

$$C_2 = \int \left(f(x)\int K^2(u)\, du\right)\frac{W(x) f(x)}{(1-F(x))^2} dx$$

son constantes reales positivas y $k \geq 2$ es el número de derivadas que se asumen para la función de densidad $f(\cdot)$. El primer término de (9) es debido al sesgo de $r_h(\cdot)$ y el segundo término a su varianza. Podemos derivar dicha expresión y obtener una ventana óptima asintótica, dada por:

$$h_0 = \left(\frac{C_2}{2kC_1}\right)^{1/(2k+1)} n^{-1/(2k+1)}.$$

Debido a que C_1 y C_2 incluyen términos desconocidos, esta ventana es imposible de calcular en la práctica a través de una muestra de datos reales. Una forma de obtener una estimación de tal ventana óptima es utilizar las ideas de validación cruzada de Bowman (1984) y Rudemo (1982) que se basan en calcular h minimizando alguna estimación de

$$ISE(h) = \int r_h^2(x) W(x)\, dx + \int r^2(x) W(x)\, dx - 2E\left(\frac{r_h(X)}{1-F(X)} W(X)\right).$$

En Estévez y Quintela (1999) se propone elegir el h que minimice

$$(10) \qquad CV_{l_n}(h) = \int r_h^2(x) W(x)\, dx - \frac{2}{n}\sum_{i=1}^{n} \frac{f_h^{-i}(X_i)}{\left(1-F_h^{-i}(X_i)\right)\left(1-F_n(X_i)\right)} W(X_i),$$

donde $f_h^{-i}(\cdot)$ y $F_h^{-i}(\cdot)$ se construyen como

$$(11) \qquad \begin{aligned} f_h^{-i}(x) &= n_{l_n}^{-1} \sum_{|j-i|>l_n} K_h(x, X_j), \\ F_h^{-i}(x) &= n_{l_n}^{-1} \sum_{|j-i|>l_n} H_h(x, X_j), \end{aligned}$$

con l_n un entero positivo, n_{l_n} verificando $nn_{l_n} = \#\{(i,j) / |i-j| > l_n\}$, y las funciones $K_h(\cdot)$ y $H_h(\cdot)$ dadas por $K_h(x,y) = \frac{1}{h} K\left(\frac{x-y}{h}\right)$ y $H_h(x,y) = H\left(\frac{x-y}{h}\right)$. $f_h^{-i}(x)$ y $F_h^{-i}(x)$ son, de este modo, los estimadores de $f(x)$ y $F(x)$, respectivamente, pero sin utilizar los datos próximos a X_i en el tiempo: $X_{i-l_n}, X_{i-l_n+1}, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{i+l_n}$. Este tipo de validación cruzada se conoce como *validación cruzada modificada*, y ha sido utilizada en otros contextos de estimación no paramétrica, como la densidad (Hart y Vieu (1990)) o la regresión (Härdle y Vieu (1992), Chu y Marron (1991), Quintela (1994)).

Estévez y Quintela (1999) demuestran que, bajo condiciones habituales en estimación no paramétrica con datos α-mixing, la ventana $\widehat{h}(l_n)$ que minimiza $CV_{l_n}(h)$ es asintóticamente óptima, en el sentido de que

$$\frac{d_I(\widehat{h}(l_n))}{\inf_h d_I(h)} \longrightarrow 1 \text{ casi seguro.} \tag{12}$$

donde $d_I(h) = ISE(h)$, $ASE(h)$ o $MISE^*(h)$.

Sarda y Vieu (1989) demuestran una propiedad similar a (12), considerando el estimador $r_{h,1}(\cdot)$ en vez de $r_h(\cdot)$. En este caso, la función de validación cruzada a minimizar es prácticamente similar a la de (10), y toma la forma

$$CV_{l_n,1}(h) = \int r_{h,1}^2(x) W(x)\, dx - \frac{2}{n}\sum_{i=1}^{n} \frac{f_h^{-i}(X_i)}{(1-F_n(X_i))^2} W(X_i). \tag{13}$$

3. SIMULACIONES

En esta sección realizamos un análisis comparativo, por medio de un estudio de simulación, del comportamiento de los dos estimadores de la función de razón de fallo tratados hasta el momento, $r_h(\cdot)$ y $r_{h,1}(\cdot)$. El estudio realizado es el siguiente:

Hemos generado M muestras, de N observaciones cada una, de un modelo dependiente con distribución Normal y otro con distribución Gamma, del siguiente modo:

Distribución normal

Cada muestra $\{X_t\}_t$ está formada por observaciones de un modelo $AR(1)$ con distribución $N(0,1)$ obtenidas de la forma:

$$X_t = \rho X_{t-1} + \left(1-\rho^2\right)^{1/2} e_t, \tag{14}$$

donde $\{e_t\}_t$ son variables aleatorias independientes e idénticamente distribuidas según una $N(0,1)$. El valor de la autocorrelación ρ se ha variado en el conjunto $\{0, \pm 0.3, \pm 0.8\}$, con el fin de ver cómo afecta la dependencia de los datos al procedimiento de validación cruzada.

Distribución gamma

Las observaciones X_t siguen una ley Gamma con parámetros de escala y forma iguales a 1 y 1.5, respectivamente. Una variable aleatoria sigue una distribución Gamma con parámetro de escala β y parámetro de forma α si su función de densidad es:

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \text{ para } x > 0. \tag{15}$$

Las observaciones se obtienen de la forma

$$X_t = \frac{1}{2} \sum_{j=t}^{t+2} Y_j^2, \tag{16}$$

donde $Y_1, Y_2, \ldots$, son variables aleatorias independientes e idénticamente distribuidas según una $N(0,1)$. De esta manera se obtiene una muestra de datos $X_1, \ldots, X_n$ que sigue un modelo 3-dependiente (es decir, los grupos de variables $(X_i, X_{i+1}, \ldots, X_j)$ y $(X_{j+3}, \ldots, X_k)$ son independientes). Esta condición de dependencia es, obviamente, más restrictiva que la condición fuertemente mixing.

Para cada una de estas distribuciones se ha considerado como función de pesos $W(x) = I_{[-2,2]}(x)$ y $W(x) = I_{[0.5,4.5]}(x)$, respectivamente, y se ha calculado, para cada muestra:

- La ventana que minimiza el error cuadrático promedio (ASE) de $r_h(\cdot)$, que denotaremos por h_{ASE}.
- La ventana que minimiza el error cuadrático promedio (ASE) de $r_{h,1}(\cdot)$, que denotaremos por $h_{ASE,1}$.
- La ventana de validación cruzada, $\widehat{h}(l_n)$ para el estimador núcleo $r_h(\cdot)$.
- La ventana de validación cruzada, $\widehat{h}_1(l_n)$ para el estimador $r_{h,1}(\cdot)$.

En los dos últimos casos, el valor del parámetro l_n se ha variado desde 0 hasta 10. Para cada una de estas 4 ventanas se ha calculado el error cuadrático ASE asociado. En todos los casos se ha usado como función kernel el núcleo de Epanechnikov:

$$K(u) = \frac{3}{4}\left(1 - u^2\right) I_{[-1,1]}(u).$$

Finalmente, hemos calculado las medias y desviaciones típicas de los resultados (amplitudes de banda y errores ASE) sobre las M muestras. Esta última medida aparecerá entre paréntesis.

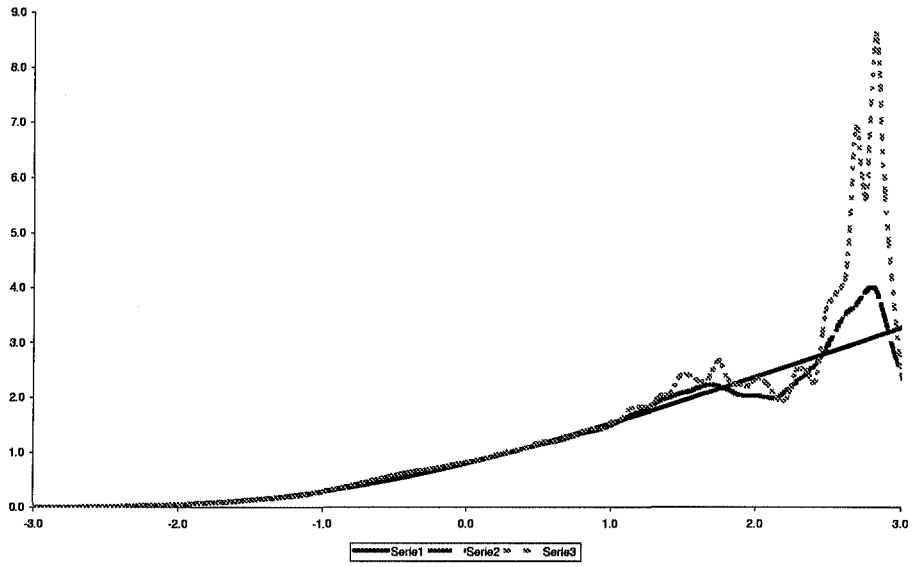


Figura 1. Razón de fallo teórica y estimada. Serie 1: curva teórica. Serie 2: estimación núcleo en el denominador. Serie 3: distribución empírica en el denominador.

En primer lugar, y para hacernos una idea del comportamiento de los estimadores $r_{h,1}(\cdot)$ y $r_h(\cdot)$, presentamos una gráfica, correspondiente a una única muestra de tamaño 1000 de una distribución $N(0,1)$, generada según (14) con el valor de $\rho = 0.8$. En la figura 1 comparamos los estimadores antes citados con el gráfico de la función de razón de fallo teórica de una distribución normal estándar. Las ventanas, elegidas según los métodos de validación cruzada (10) y (13) son, respectivamente, 0.45 y 0.43 (en ambos casos hemos tomado $l_n = 0$).

En la figura 1 puede apreciarse que, en los valores más altos de la variable, el estimador que utiliza en su denominador la función de distribución empírica se aproxima peor a la función teórica. Esta peor aproximación se aprecia con claridad en las tablas siguientes, donde se ha elegido $M = 200$ y $N = 1000$. Para cada uno de los estimadores y distribuciones, se han calculado las ventanas de validación cruzada y *ASE* junto con los errores cuadráticos promedio para las distintas estimaciones. Los resultados obtenidos para los valores de $\rho = 0$, 0.3 y 0.8 se muestran en las siguientes tablas:

Distribución normal

Tabla I. $\rho = 0$.

$h_{ASE} = 0.5107$	(0.1323)	$ASE(h_{ASE}) = 0.0060$	(0.0043)
$h_{ASE,1} = 0.5721$	(0.1827)	$ASE(h_{ASE,1}) = 0.0104$	(0.0124)

l_n	$\widehat{h}(l_n)$	$ASE\left(\widehat{h}(l_n)\right)$	$\widehat{h}_1(l_n)$	$ASE\left(\widehat{h}_1(l_n)\right)$
0	0.4158 (0.1997)	0.0208 (0.0349)	0.6676 (0.3605)	0.0239 (0.0362)
1	0.4147 (0.1985)	0.0206 (0.0343)	0.6569 (0.3616)	0.0247 (0.0383)
2	0.4166 (0.1989)	0.0207 (0.0346)	0.6535 (0.3618)	0.0247 (0.0383)
3	0.4052 (0.1999)	0.0214 (0.0349)	0.6392 (0.3638)	0.0246 (0.0380)
4	0.4059 (0.2018)	0.0216 (0.0352)	0.6466 (0.3649)	0.0251 (0.0386)
5	0.3986 (0.1968)	0.0216 (0.0354)	0.6499 (0.3646)	0.0252 (0.0386)
6	0.3990 (0.1989)	0.0221 (0.0357)	0.6478 (0.3640)	0.0252 (0.0385)
7	0.3945 (0.1980)	0.0224 (0.0359)	0.6415 (0.3648)	0.0251 (0.0382)
8	0.3912 (0.1941)	0.0223 (0.0359)	0.6395 (0.3642)	0.0247 (0.0362)
9	0.3915 (0.1951)	0.0218 (0.0353)	0.6399 (0.3687)	0.0255 (0.0382)
10	0.4007 (0.1927)	0.0207 (0.0342)	0.6652 (0.3566)	0.0226 (0.0330)

Tabla II. $\rho = 0.3$.

$h_{ASE} = 0.4940$	(0.1607)	$ASE(h_{ASE}) = 0.0056$	(0.0046)
$h_{ASE,1} = 0.5355$	(0.1386)	$ASE(h_{ASE,1}) = 0.0107$	(0.0101)

l_n	$\widehat{h}(l_n)$	$ASE\left(\widehat{h}(l_n)\right)$	$\widehat{h}_1(l_n)$	$ASE\left(\widehat{h}_1(l_n)\right)$
0	0.4972 (0.1967)	0.0174 (0.0244)	0.6524 (0.3340)	0.0244 (0.0290)
1	0.4952 (0.1987)	0.0175 (0.0244)	0.6592 (0.3283)	0.0238 (0.0283)
2	0.4937 (0.1967)	0.0175 (0.0245)	0.6550 (0.3297)	0.0239 (0.0283)
3	0.4905 (0.1978)	0.0177 (0.0249)	0.6608 (0.3378)	0.0247 (0.0293)
4	0.4937 (0.1992)	0.0176 (0.0246)	0.6646 (0.3374)	0.0246 (0.0294)
5	0.4930 (0.2015)	0.0176 (0.0246)	0.6593 (0.3399)	0.0250 (0.0299)
6	0.4902 (0.2031)	0.0177 (0.0246)	0.6550 (0.3410)	0.0247 (0.0294)
7	0.4963 (0.2027)	0.0177 (0.0246)	0.6547 (0.3408)	0.0244 (0.0293)
8	0.4950 (0.2001)	0.0175 (0.0246)	0.6565 (0.3375)	0.0243 (0.0293)
9	0.4993 (0.2001)	0.0174 (0.0245)	0.6566 (0.3367)	0.0236 (0.0282)
10	0.4921 (0.2003)	0.0175 (0.0245)	0.6578 (0.3397)	0.0238 (0.0282)

Tabla III. $\rho = 0.8$.

$h_{ASE} = 0.5894$	(0.2673)	$ASE(h_{ASE}) = 0.0100$	(0.0102)
$h_{ASE,1} = 0.5497$	(0.2345)	$ASE(h_{ASE,1}) = 0.0250$	(0.0271)

l_n	$\widehat{h}(l_n)$	$ASE\left(\widehat{h}(l_n)\right)$	$\widehat{h}_1(l_n)$	$ASE\left(\widehat{h}_1(l_n)\right)$
0	0.4657 (0.1790)	0.0255 (0.0241)	0.6423 (0.3464)	0.0379 (0.0382)
1	0.4666 (0.1975)	0.0273 (0.0260)	0.6672 (0.3711)	0.0396 (0.0421)
2	0.4688 (0.1910)	0.0265 (0.0252)	0.6734 (0.3924)	0.0406 (0.0428)
3	0.4629 (0.1947)	0.0269 (0.0251)	0.6830 (0.4016)	0.0406 (0.0427)
4	0.4554 (0.1967)	0.0279 (0.0284)	0.6826 (0.4071)	0.0410 (0.0427)
5	0.4531 (0.1907)	0.0278 (0.0284)	0.6853 (0.4115)	0.0412 (0.0426)
6	0.4455 (0.1953)	0.0284 (0.0286)	0.6881 (0.4125)	0.0413 (0.0432)
7	0.4510 (0.1986)	0.0287 (0.0283)	0.6871 (0.4182)	0.0419 (0.0442)
8	0.4516 (0.1923)	0.0281 (0.0281)	0.6892 (0.4155)	0.0418 (0.0442)
9	0.4497 (0.1973)	0.0284 (0.0283)	0.6875 (0.4177)	0.0419 (0.0442)
10	0.4438 (0.1932)	0.0277 (0.0269)	0.6948 (0.4081)	0.0411 (0.0436)

Los valores de $\rho = -0.3$ y $\rho = -0.8$ en (14) producen resultados similares. Pasamos ahora a exponer los resultados de la distribución Gamma (16):

Tabla IV.

$h_{ASE} = 0.21101$	(0.2190)	$ASE(h_{ASE}) = 0.0018$	(0.0013)
$h_{ASE,1} = 1.3670$	(0.2048)	$ASE(h_{ASE,1}) = 0.0263$	(0.0153)

l_n	$\widehat{h}(l_n)$	$ASE\left(\widehat{h}(l_n)\right)$	$\widehat{h}_1(l_n)$	$ASE\left(\widehat{h}_1(l_n)\right)$
0	0.6925 (0.3605)	0.0309 (0.0220)	0.6255 (0.3053)	0.0384 (0.0210)
1	0.8108 (0.3916)	0.0273 (0.0219)	0.8022 (0.3196)	0.0337 (0.0185)
2	0.8373 (0.3835)	0.0260 (0.0214)	0.8115 (0.3225)	0.0336 (0.0191)
3	0.8375 (0.3786)	0.0259 (0.0212)	0.8212 (0.3122)	0.0333 (0.0186)
4	0.8596 (0.3683)	0.0248 (0.0205)	0.8247 (0.3151)	0.0331 (0.0181)
5	0.8560 (0.3724)	0.0250 (0.0207)	0.8244 (0.3178)	0.0331 (0.0181)
6	0.8495 (0.3771)	0.0253 (0.0209)	0.8234 (0.3158)	0.0331 (0.0181)
7	0.8504 (0.3768)	0.0253 (0.0209)	0.8135 (0.3214)	0.0335 (0.0185)
8	0.8308 (0.3716)	0.0258 (0.0210)	0.8115 (0.3206)	0.0335 (0.0185)
9	0.8415 (0.3725)	0.0255 (0.0209)	0.8131 (0.3228)	0.0335 (0.0185)
10	0.8457 (0.3709)	0.0253 (0.0208)	0.8095 (0.3246)	0.0335 (0.0184)

Desde el punto de vista comparativo de los dos estimadores, observamos que el método de validación cruzada tiende a producir, en el caso de la distribución Normal, una infrasuavización de la ventana óptima h_{ASE} para $r_h(\cdot)$ y, sin embargo, una sobresuavización de $h_{ASE,1}$ para $r_{h,1}(\cdot)$. Para la distribución Gamma, no obstante, dicho método infrasuaviza las ventanas para ambos estimadores. Observamos también que, para cada valor de ρ y cada l_n, los errores $ASE\left(\widehat{h}(l_n)\right)$ son menores, en media, que los correspondientes $ASE\left(\widehat{h}_1(l_n)\right)$ y con dispersiones más pequeñas. En resumen, parece que el uso del estimador núcleo va a proporcionar buenas estimaciones, no sólo mejores que las de $r_{h,1}(\cdot)$, sino también en comparación con la estimación óptima, $r_{h_{ASE}}(\cdot)$.

Si observamos los resultados para cada estimador por separado, nos damos cuenta de que el valor del parámetro l_n en el procedimiento de validación cruzada tiene un efecto muy débil, al contrario de lo que ocurre en otros problemas de estimación. En contextos similares, como en estimación de densidad (Hart y Vieu (1990)) o regresión (Quintela (1994)), cuando trabajamos, por ejemplo, con datos que presentan autocorrelación positiva, interesa aumentar el valor de l_n a medida que ésta crece. Aquí, bajo el modelo $AR(1)$ con distribución marginal $N(0,1)$, se observa que un incremento de l_n prácticamente no modifica el valor de la amplitud de banda obtenida para $l_n = 0$. Para $\rho = 0.8$ por ejemplo, las ventanas seleccionadas por validación cruzada se mueven en el intervalo $[0.44, 0.47]$, sin observarse una tendencia clara al aumentar l_n (más detalles sobre esta afirmación pueden consultarse en Estévez y Quintela (1999)). Sin embargo, bajo la distribución Gamma, sí se logran ventanas ligeramente distintas al utilizar $l_n = 0$ o $l_n > 0$, aunque tal diferencia tampoco conlleva grandes mejorías en lo que a errores de estimación se refiere. La explicación de este hecho está en la forma (simétrica a la J) de la función de validación cruzada.

4. SELECCIÓN LOCAL DEL PARÁMETRO VENTANA

En la estimación no paramétrica de curvas es un fenómeno relativamente frecuente que la utilización de un único parámetro ventana, para la estimación de una curva en concreto, no presente buenos resultados. Así, en regiones de baja densidad de datos, es preferible elegir un parámetro ventana grande a fin de basar la estimación en un número suficientemente grande de puntos. Mientras, en aquellas áreas en las que existen un mayor número de datos, es preferible elegir parámetros ventana pequeños, pues evidenciarán más claramente las características locales de la función a estimar. Con esta filosofía se construyen los estimadores de ventana local, que utilizan una ventana diferente h_x dependiendo del punto concreto x donde se realice la estimación. Básicamente, se pretende seleccionar una ventana h_x que minimice algún estimador del *Error cuadrático medio puntual:*

(17) $$MSE_x(h) = E(r_h(x) - r(x))^2.$$

En esta sección definimos un método de selección local de la ventana que se demuestra asintóticamente óptimo en el sentido de (12), con (12) adaptado a un método de validación cruzada de tipo local. Concretamente, para cada punto x de un intervalo $[a,b] \subset \text{Soporte}(X)$ ($\text{Soporte}(X) = \{x \in \mathbb{R}/f(x) > 0\}$), elegiremos una ventana h_x minimizando la función

(18) $$CV_{x,l_n}(h) = \int_a^b r_h^2(y) W_{n,x}(y)\, dy - 2n^{-1} \sum_{i=1}^{n} \frac{f_h^{-i}(X_i)}{(1 - F_h^{-i}(X_i))(1 - F_n(X_i))} W_{n,x}(X_i),$$

donde $f_h^{-i}(\cdot)$ y $F_h^{-i}(\cdot)$ se definen igual que en (11), y $W_{n,x}(\cdot)$ es una sucesión de funciones de peso centradas en el punto x. Es habitual elegir funciones del tipo

(19) $$W_{n,x}(y) = \begin{cases} 1 & \text{si} \quad y \in [x - t_{0,n}, x + t_{0,n}] \\ 0 & \text{si} \quad y \notin [x - t_{0,n}, x + t_{0,n}] \end{cases}$$

siendo $t_{0,n}$ un número real que tiende a 0 cuando n tiende a infinito, y dependiente de x (Vieu (1991*b*), Quintela y Vilar (1992)). Como puede verse en estos trabajos, la elección de esta función $W_{n,x}(\cdot)$ no influye de manera especial en el procedimiento de selección de la ventana.

Para el establecimiento de un resultado de optimalidad asintótica para el procedimiento de validación cruzada local utilizaremos las siguientes hipótesis:

(H.1) $\{X_n\}_{n \in \mathbb{N}}$ es una sucesión estrictamente estacionaria de variables α–mixing.

(H.2) La sucesión de pesos locales $W_{n,x}(\cdot)$ verifica:

$$W_{n,x}(u) \geq 0, \int W_{n,x}(u)\, du = 1,$$

$$\sup_{x,u} W_{n,x}(u) = O\left(n^{\beta}\right) \quad \text{con} \quad \beta < \frac{1 - 2ku}{2}.$$

(H.3) Para toda función $g(\cdot)$ continua se verifica:

$$\int W_{n,x}(u)\, g(u)\, du \longrightarrow g(x) \text{ uniformemente para } x \in [a,b].$$

(H.4) Existe $\gamma > 0$ tal que $\forall x \in [a,b]$, $1 - F(x) > \gamma$ y $f(x) \geq \gamma$.

(H.5) Existe $M_1, 0 < M_1 < \infty$, tal que $f(u) \leq M_1$ $\forall u \in [a,b]$.

(H.6) La densidad $f(\cdot)$ tiene $k(\geq 2)$ derivadas continuas.

(H.7) Para todo $j \geq 2$ existe la función de densidad conjunta $f_{1,j}(\cdot,\cdot)$ y satisface:

$$\sup_{u,v\in[a,b]} \int |f_{1,j}(u,v) - f(u)f(v)|du\,dv \leq \alpha(j-1).$$

(H.8) La función núcleo $K(\cdot)$ es una función de soporte compacto que encierra área 1, es decir, $\int K(u)\,du = 1$.

(H.9) $K(\cdot)$ es de orden k, esto es

$$\int u^m K(u)du = 0, \text{ para } m = 1,\ldots,k-1 \text{ y } 0 < \int u^k K(u)du < \infty.$$

(H.10) $K(\cdot)$ es Lipschitz continua, es decir, $\exists C_K$ $0 < C_K < \infty$ tal que

$$|K(x) - K(y)| \leq C_K |x-y| \quad \forall x,y \in \mathbb{R}.$$

(H.11) La transformada de Fourier de $K(\cdot)$ es absolutamente integrable.

(H.12) $h \in H_n = [A_1 n^{-u}, A_2 n^{-v}]$, $0 < A_1 < A_2 < \infty$, $0 < v \leq 1/(2k+1) \leq u < 2/(4k+1)$.

(H.13) $0 \leq l_n \leq C n^{r_1}$, con $0 < r_1 < 1 - \beta - u(4k+1)/2$.

(H.14) Para

$$r_2 = U + V + (2\beta + 2u + 4ku)(2 + \frac{U}{V}),$$

$$U = 1 + 2u + 2ku + \beta - v, \qquad V = 2 - 2\beta - u(1+4k) - 2r_1,$$

se tiene que

$$\tilde{\alpha}(n^{r_1}) = \sup_{t > n^{r_1}} \alpha(t) = o\left(n^{-r_2}\right).$$

Teorema 4.1. *Bajo las hipótesis* **(H.1)-(H.14)**, *se tiene que*

$$\sup_{x\in[a,b]} \frac{d_x(\widehat{h}_x(l_n))}{\inf_{h\in H_n} d_x(h)} \longrightarrow 1 \quad \textit{casi seguro},$$

donde $\widehat{h}_x(l_n)$ *es la ventana de validación cruzada local y* $d_x(h)$ *es* $MSE_x(h)$ *(17) o cualquiera de las versiones locales de los errores cuadráticos (6), (7) y (8):*

$$ASE_x(h) = \frac{1}{n}\sum_{i=1}^{n}(r_h(X_i) - r(X_i))^2 W_{n,x}(X_i),$$

$$ISE_x(h) = \int_a^b (r_h(y) - r(y))^2 W_{n,x}(y)dy$$

y

$$MISE_x^*(h) = E\int\left[\left(\frac{f_h(y)}{1-F_h(y)} - \frac{f(y)}{1-F(y)}\right)\frac{1-F_h(y)}{1-F(y)}\right]^2 W_{n,x}(y)\,dy.$$

Comentario 4.1. *Las hipótesis bajo las que se establece este teorema son del mismo tipo de las utilizadas en los trabajos de Hart y Vieu (1990) y Quintela y Vilar (1992), por lo que sus comentarios sobre las mismas siguen siendo válidos en nuestro contexto. Estas hipótesis son, a su vez, las mismas que las utilizadas en el teorema 6 de Estévez y Quintela (1999), a las que se han añadido las correspondientes a la utilización de las funciones de peso locales* $W_{n,x}(\cdot)$.

La demostración de este teorema sigue los mismos pasos que el teorema 6 de Estévez y Quintela (1999), pero teniendo en cuenta las hipótesis **(H.2)** y **(H.3)** sobre los pesos locales. La demostración hace uso, también, de los siguientes resultados sobre equivalencia asintótica de los errores cuadráticos locales y de la descomposición asintótica de los mismos:

Teorema 4.2. *Bajo las hipótesis del Teorema 4.1, se tiene que*

$$\sup_{h\in H_n}\frac{\left|d_x(h) - d_x'(h)\right|}{d_x'(h)} \longrightarrow 0 \quad \textit{casi seguro,}$$

y

$$d_x(h) = C_1 h^{2k} + C_2\frac{1}{nh} + o\left(h^{2k} + \frac{1}{nh}\right),$$

para $d_x(h)$ *y* $d_x'(h)$ *cualesquiera de las funciones* $MSE_x(h)$, $ASE_x(h)$, $ISE_x(h)$ *o* $MISE_x^*(h)$.

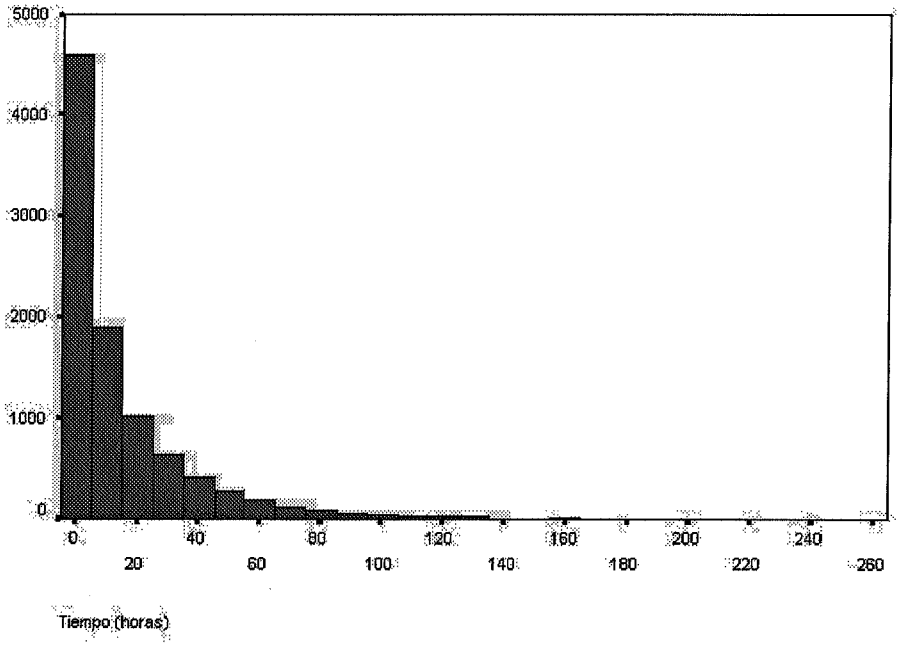


Figura 2. Histograma de la serie de datos de California.

La demostración de los teoremas 4.1 y 4.2 se ha resumido en el Apéndice del presente trabajo.

5. APLICACIONES A DATOS SÍSMICOS

El modelo estadístico más simple para ajustar una serie de tiempo de ocurrencia entre terremotos es el proceso de Poisson (Vere-Jones (1970)). Bajo este modelo, los intervalos de tiempo entre eventos consecutivos están distribuidos exponencialmente, de modo que la función de riesgo es una línea recta de altura igual al inverso del valor medio de la variable.

Sin embargo, este modelo presupone independencia entre los eventos, de modo que la ocurrencia de un terremoto no estaría influenciada por la de los terremotos previos, lo que obviamente resulta una hipótesis bastante alejada de la realidad.

La desviación de las observaciones de este modelo se atribuye, principalmente, a la existencia de un número muy grande de intervalos de tiempo pequeños, debido a que los terremotos no ocurren como sucesos aislados, sino que tienden a agruparse en «clusters», es decir, un movimiento concreto viene seguido y/o precedido, habitualmente, por otra serie de «replicas» y/o «precursores» (Lomnitz y Hax (1967), Vere-Jones y Davies (1966)).

Un modelo estadístico de tipo paramétrico que ajusta más convenientemente esta tendencia al agrupamiento o «clustering» de los movimientos sísmicos es la distribución Gamma ((Vere-Jones (1970), Udias y Rice (1975)).

Los datos de nuestro primer ejemplo corresponden a un fichero con la información de los terremotos ocurridos en California (Estados Unidos), desde el 1 de enero de 1980 hasta el 12 de octubre de 1996 (inclusive). De ellos, hemos seleccionado aquellos que tienen magnitud mayor o igual que 3 en la escala de Richter, y profundidad menor o igual que 50 km. De esta forma, estamos considerando aquellos terremotos cuya localización goza de mayor precisión. Considerando el tiempo de ocurrencia de cada uno de estos movimientos sísmicos, construimos la serie de tiempo $\{\Delta t_i\}_{i=1}^n$ de intervalos de tiempo (en horas) entre movimientos consecutivos. Se obtiene una muestra de tamaño $n = 9471$ datos.

En primer lugar mostramos, en la figura 2, un histograma de los datos $\{\Delta t_i\}_{i=1}^n$. En éste puede observarse el efecto de agrupamiento antes citado, al concentrarse una gran cantidad de datos en valores próximos a cero. Un estudio descriptivo de los mismos arroja, como valores significativos: media muestral = 15.53, mediana = 5.48, desviación típica = 24.204, mínimo de la variable = 0.00027 y máximo de la variable = 256.2886.

Un ajuste de tipo paramétrico de la densidad de estos datos a una distribución Gamma como en (15) proporciona unos valores para α y β de 0.4117 y 0.02651, respectivamente. Sin embargo, tanto el test Chi-cuadrado como el de Kolmogorov-Smirnov de bondad de ajuste arrojan valores para el p-valor del test iguales o muy próximos a cero, indicando que debe rechazarse la hipótesis de pertenencia de los datos a la distribución especificada.

Para realizar una estimación no paramétrica de la función de riesgo asociada a la muestra $\{\Delta t_i\}_{i=1}^n$, aplicamos el algoritmo de validación cruzada (10) a dichos datos. Debido a la escasa influencia del valor l_n, lo hemos tomado igual a cero. El valor de la ventana así obtenida es $\widehat{h}(0) = 0.25$ (horas), y la gráfica de la función de razón de fallo $r_h(\cdot)$ estimada con dicho parámetro aparece en la figura 3. En la misma, hemos incluido la gráfica de la función de razón de fallo que correspondería a una distribución Gamma de parámetros los estimados a partir de la muestra ($\alpha = 0.4117$ y $\beta = 0.02651$).

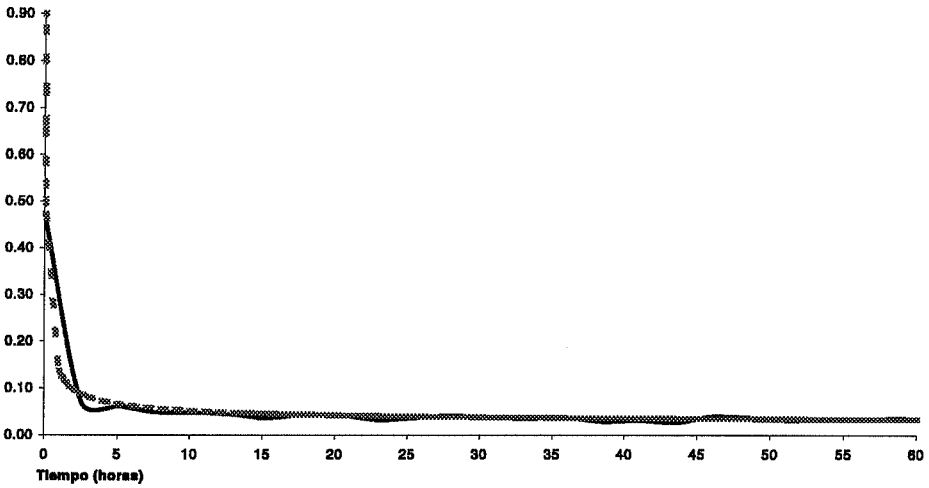


Figura 3. Estimaciones paramétrica y no paramétrica con los datos de California. Línea sólida: estimación no paramétrica. Línea punteada: estimación paramétrica.

Atendiendo a la estimación no paramétrica, podemos observar el elevado riesgo que existe en los primeros momentos, y como la función tiende a ser constante a partir de las 6 horas, aproximadamente. En la gráfica, el eje de abscisas sólo alcanza hasta el valor 60 para una mejor visualización, puesto que a partir de dicho valor sólo existen un cinco por ciento del total de los datos. De hecho, el cuantil de orden 90 ya es el valor 44.29 horas. Un efecto decreciente como el de esta figura puede observarse en las figuras 8 y 9 de Rice y Rosenblatt (1976), donde se analiza un conjunto de terremotos ocurridos en la Falla de San Andreas, una región comprendida en nuestro espacio geográfico de análisis.

La comparación con la curva ajustada a una distribución Gamma revela la potente capacidad de la estimación no paramétrica para una estimación más correcta del riesgo instantáneo de ocurrencia de un terremoto. En efecto, la curva que correspondería a una distribución Gamma indicaría unos valores del riesgo no aceptables en los valores del tiempo más próximos a cero (el eje de ordenadas es una asíntota de la curva ajustada paramétricamente), mientras que ambas estimaciones son muy similares a partir de las 5 o 6 horas, aproximadamente.

El segundo ejemplo que hemos considerado corresponde a la serie de tiempo formada por los intervalos de tiempo (en horas) entre la ocurrencia de terremotos consecutivos en Granada, durante el periodo de tiempo comprendido desde enero de 1993 hasta diciembre de 1995 (datos obtenidos del Instituto Geográfico Nacional). Esta serie de tiempo $\{\Delta t_i\}_{i=1}^{n}$ se compone de $n = 172$ datos.

Aplicando el procedimiento de validación cruzada global a esta serie de tiempo, obtenemos como valores para las ventanas $\widehat{h}(0) = 38.2$, $\widehat{h}(1) = 38.0$, $\widehat{h}(2) = 31.0$, $\widehat{h}(3) = 30.2$, $\widehat{h}(4) = 30.4$ y $\widehat{h}(5) = 29.8$ que proporcionan, prácticamente, las mismas estimaciones. En la figura 5 (comentada más adelante) se muestra una gráfica de la curva estimada con el parámetro $\widehat{h}(0)$.

Para comprobar los resultados prácticos del procedimiento de selección de la ventana mediante validación cruzada local, aplicamos el método definido en (18) a este conjunto de datos. Para ello, hemos realizado una partición del intervalo muestral en 33 puntos equiespaciados y, para cada uno de ellos, hemos calculado una ventana local. Como funciones de peso hemos tomado (19), con $t_{0,n}$ la mitad de la desviación típica de los datos. El valor de l_n se eligió igual a cero. Una representación gráfica de las diferentes ventanas locales obtenidas $\widehat{h}_x(0)$ se presenta en la figura 4.

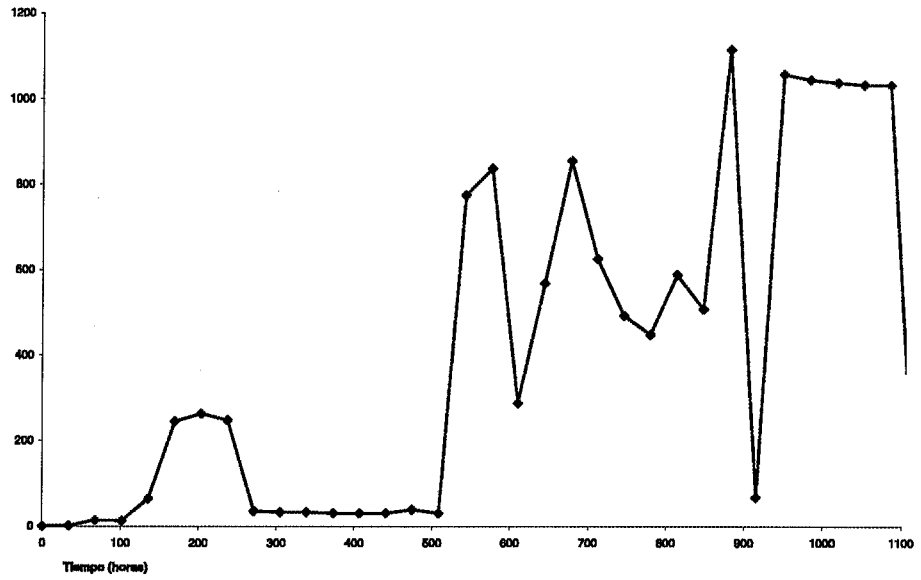


Figura 4. Ventanas locales.

En la misma podemos observar la tendencia a obtenerse ventanas más grandes a medida que el tiempo crece, puesto que la dispersión de los datos es también mayor. La figura 5 muestra las estimaciones locales y globales de la función de riesgo para los datos en estudio:

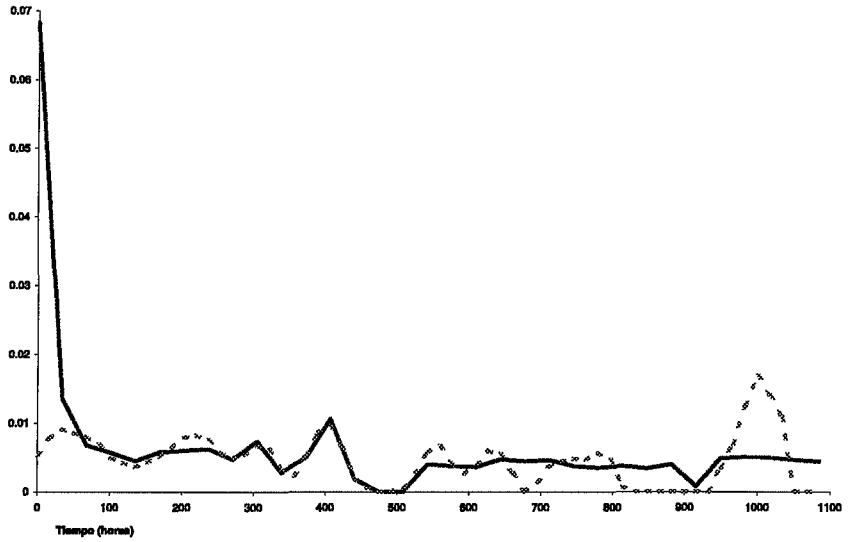


Figura 5. Estimaciones global y local de la razón de fallo con los datos de Granda. Línea sólida: estimación local. Línea punteada: Estimación global.

Puede observarse que los estimadores son similares lejos de las fronteras. Para intervalos de tiempo grandes, el estimador local corrige los valores excesivos que muestra la estimación global. Igualmente, en los intervalos próximos a cero también se corrige la mala estimación causada por una ventana global que resulta excesivamente grande en dicho extremo. Este hecho pone también de manifiesto el enorme interés que puede tener una estimación del riesgo de tipo local cuando se maneja una muestra con no demasiados datos. En el primer caso analizado, el de los terremotos de California, la estimación global produjo resultados muy aceptables gracias a que la muestra contenía un número de datos muy grande. En el segundo ejemplo, el tamaño de muestra pequeño produjo una estimación global del riesgo muy alejada de la realidad en los valores próximos a cero, y este problema pudo evitarse mediante una estimación de tipo local.

6. APÉNDICE: DEMOSTRACIONES DE LOS RESULTADOS RELATIVOS AL USO DE VENTANAS LOCALES

En la presente sección se dan las demostraciones resumidas de los resultados de la sección 4, relativos al uso de ventanas locales en estimación núcleo de la función de razón de fallo. Las pruebas detalladas pueden ser vistas en Estévez (2001).

Demostración del Teorema 4.1. La demostración es muy similar a la correspondiente del caso global (Estévez y Quintela (1999) prueba de su Teorema 6), por lo que nos ceñiremos a una revisión superficial de la misma. La diferencia entre ambas pruebas estriba en la función de pesos; en tanto en el Teorema 6 de Estévez y Quintela (1999) se hacía uso de una función acotada, ahora se dispone de una sucesión de funciones $W_{n,x}(\cdot)$ verificando: $\sup_{x,u} W_{n,x}(u) = O(n^{\beta})$ con $\beta < \frac{1-2ku}{2}$, es decir, tienen la posibilidad de tender a infinito con el tamaño muestral aunque de un modo muy lento.

Al igual que en el caso global, la equivalencia asintótica entre los errores locales $ASE_x(h)$, $ISE_x(h)$, $MISE_x^*(h)$ y $MSE_x^*(h)$ (Teorema 4.2) permite demostrar el presente resultado probando, para cada $x \in [a,b]$, la existencia de una v.a. T, independiente de h, que satisfaga:

$$\sup_{h \in H_n} \frac{|ISE_x(h) - CV_{x,l_n}(h) - T|}{ISE_x(h)} \longrightarrow 0 \quad c.s.. \tag{20}$$

Para ello, se tratará de localizar la v.a. T y, simultáneamente, lograr una descomposición de fácil manejo para $ISE_x(r_h) - CV_{l_n,x}(h)$. En primer lugar, de la definición de las funciones $ISE_x(h)$ y $CV_{x,l_n}(h)$ se obtiene que

$$\begin{aligned} ISE_x(r_h) - CV_{x,l_n}(h) = & \int_a^b \frac{f^2(y)}{(1-F(y))^2} W_{n,x}(y)\,dy - \\ & -2\int_a^b \frac{f_h(y)}{(1-F_h(y))(1-F(y))} W_{n,x}(y) f(y) dy + \\ & +\frac{2}{n}\sum_{i=1}^{n} \frac{f_h^{-i}(X_i)}{(1-F_h^{-i}(X_i))(1-F_n(X_i))} W_{n,x}(X_i). \end{aligned} \tag{21}$$

Ahora bien, por analogía con el caso global, utilizando el Teorema 4.2 junto con el hecho de que $f(\cdot)$ está acotada en el intervalo $[a,b]$ y las hipótesis **[H.2]** y **[H.4]**, se obtiene la siguiente descomposición para el segundo sumando de (21):

$$\begin{aligned} & \int_a^b \frac{f_h(y)}{(1-F_h(y))(1-F(y))} W_{n,x}(y) f(y) dy = \\ & = \int_a^b \frac{f_h(y)}{(1-F(y))^2} W_{n,x}(y) f(y) dy - \int_a^b \frac{f(y)(F(y)-F_h(y))}{(1-F(y))^3} W_{n,x}(y) f(y) dy - \end{aligned}$$

$$(22)\qquad -\int_a^b \frac{(f_h(y)-f(y))(F(y)-EF_h(y))}{(1-F(y))^3} W_{n,x}(y) f(y)dy -$$

$$-\int_a^b \frac{(f_h(y)-f(y))(EF_h(y)-F_h(y))}{(1-F(y))^3} W_{n,x}(y) f(y)dy +$$

$$+\int_a^b \frac{f(y)(F(y)-F_h(y))^2}{(1-F(y))^4} W_{n,x}(y) f(y)dy + o(ISE_x(h)) \quad c.s..$$

De forma análoga, el tercer sumando de (21) se puede escribir como

$$\frac{2}{n}\sum_{i=1}^n \frac{f_h^{-i}(X_i)}{(1-F_h^{-i}(X_i))(1-F_n(X_i))} W_{n,x}(X_i) =$$

$$= \frac{2}{n}\sum_{i=1}^n \frac{f_h^{-i}(X_i)}{(1-F(X_i))^2} W_{n,x}(X_i) + T' - \frac{2}{n}\sum_{i=1}^n \frac{f(X_i)(F(X_i)-F_h^{-i}(X_i))}{(1-F(X_i))^3} W_{n,x}(X_i) -$$

$$(23)\quad -\frac{2}{n}\sum_{i=1}^n \frac{(f_h^{-i}(X_i)-f(X_i))(F(X_i)-\overline{E}[F_h(X_i)/X_i])}{(1-F(X_i))^3} W_{n,x}(X_i) -$$

$$-\frac{2}{n}\sum_{i=1}^n \frac{(f_h^{-i}(X_i)-f(X_i))(\overline{E}[F_h(X_i)/X_i]-F_h^{-i}(X_i))}{(1-F(X_i))^3} W_{n,x}(X_i) +$$

$$+\frac{2}{n}\sum_{i=1}^n \frac{f(X_i)(F(X_i)-F_h^{-i}(X_i))^2}{(1-F(X_i))^4} W_{n,x}(X_i) + o(ISE_x(h)) \quad c.s.,$$

donde

$$T' = \frac{2}{n}\sum_{i=1}^n \frac{f(X_i)(F_n(X_i)-F(X_i))}{(1-F(X_i))^3} W_{n,x}(X_i)$$

y $\overline{E}[F_h(X_i)/X_i]$ es la media condicional si los datos fuesen independientes. Es importante reseñar que si bien la prueba de (22) se obtiene, en virtud de la condición $\int W_{n,x}(u)\,du = 1$, trabajando como si hubiese una única función de pesos $W(\cdot)$, la demostración de (23) requiere más cautela y el uso de $\sup_{x,u} W_{n,x}(u) = O(n^{\beta})$ con $\beta <$

$\frac{1-2ku}{2}$. Así pues, recopilando (21)-(23), se obtiene la siguiente factorización para el numerador de (20):

$$(24)\qquad ISE_x(h) - CV_{x,l_n}(h) - T = 2CT^{(1)}_{l_n,x}(h) - 2CT^{(2)}_{l_n,x}(h) - 2CT^{(3)}_{l_n,x}(h) - \\ - 2CT^{(4)}_{l_n,x}(h) + 2CT^{(5)}_{l_n,x}(h) + o(ISE_x(h)) \quad c.s.,$$

donde la v.a. T está dada por

$$T = \frac{2}{n}\sum_{i=1}^{n} \frac{f(X_i)}{(1-F(X_i))^2} W_{n,x}(X_i) - \int_a^b \frac{f^2(y)}{(1-F(y))^2} W_{n,x}(y)\,dy + \\ + \frac{2}{n}\sum_{i=1}^{n} \frac{f(X_i)\,(F_n(X_i) - F(X_i))}{(1-F(X_i))^3} W_{n,x}(X_i)$$

y los términos $CT^{(i)}_{l_n,x}(h)$ son:

$$CT^{(1)}_{l_n,x}(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{f_h^{-i}(X_i)\,W_{n,x}(X_i)}{(1-F(X_i))^2} - \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)\,W_{n,x}(X_i)}{(1-F(X_i))^2} + \\ + \int_a^b \frac{f^2(y)}{(1-F(y))^2} W_{n,x}(y)\,dy - \int_a^b \frac{f_h(y)}{(1-F(y))^2} f(y)\,W_{n,x}(y)\;dy;$$

$$CT^{(2)}_{l_n,x}(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(f_h^{-i}(X_i) - f(X_i)\right)\left(F(X_i) - \overline{E}[F_h(X_i)/X_i]\right)}{(1-F(X_i))^3} W_{n,x}(X_i) - \\ - \int_a^b \frac{(f_h(y) - f(y))\,(F(y) - EF_h(y))}{(1-F(y))^3} W_{n,x}(y)\,f(y)\;dy;$$

$$CT^{(3)}_{l_n,x}(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(f_h^{-i}(X_i) - f(X_i)\right)\left(\overline{E}[F_h(X_i)/X_i] - F_h^{-i}(X_i)\right)}{(1-F(X_i))^3} W_{n,x}(X_i) - \\ - \int_a^b \frac{(f_h(y) - f(y))\,(EF_h(y) - F_h(y))}{(1-F(y))^3} W_{n,x}(y)\,f(y)\;dy;$$

$$CT_{l_n,x}^{(4)}(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)\left(F(X_i) - F_h^{-i}(X_i)\right)}{(1-F(X_i))^3} W_{n,x}(X_i) -$$

$$- \int_a^b \frac{f(y)(F(y) - F_h(y))}{(1-F(y))^3} W_{n,x}(y) f(y)\, dy$$

y

$$CT_{l_n,x}^{(5)}(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)\left(F(X_i) - F_h^{-i}(X_i)\right)^2}{(1-F(X_i))^4} W_{n,x}(X_i) -$$

$$- \int_a^b \frac{f(y)(F(y) - F_h(y))^2}{(1-F(y))^4} W_{n,x}(y) f(y)\, dy.$$

Por consiguiente, la demostración de (20) es consecuencia directa de (24) y

$$\sup_{h\in H_n} \frac{\left|CT_{l_n,x}^{(i)}(h)\right|}{ISE_x(h)} \longrightarrow 0 \quad c.s., \text{ para } i = 1,\ldots,5. \tag{25}$$

Para probar (25) seguiremos, como hasta el momento, esencialmente los mismos pasos que en el procedimiento de validación cruzada global, notando que también, en virtud del Teorema 4.2, $ISE_x(h) \geq Ch^{2k}\left(C\frac{1}{nh}\right)$ *c.s.*.

Prueba de (25) para $i = 1$ $(i = 2)$. Comparando $CT_{l_n,x}^{(1)}(h)$ $\left(\text{respectivamente } CT_{l_n,x}^{(2)}(h)\right)$ con el término $T_x(h)$ de la demostración del Teorema 3.1.2 de Quintela (1992), se observa que son idénticos salvo un pequeño matiz: la sucesión de funciones de ponderación. Mientras el autor utiliza una sucesión de funciones $W_{n,x}(\cdot)$ que cumple **[H.2]** y **[H.3]**, en nuestro caso aparece

$$\frac{W_{n,x}(\cdot)}{(1-F(\cdot))^2} \left(\text{respectivamente } \frac{B(\cdot)W_{n,x}(\cdot)}{(1-F(\cdot))^3}, \text{ con } B(y) = \right.$$

$$\left. = F(y) - \int H\left(\frac{y-u}{h}\right) f(u)\, du\right)$$

que, por otra parte, también satisface las condiciones anteriores. Así pues, notando que los órdenes de convergencia de los errores $MISE_x(\cdot)$ para densidad y razón de fallo coinciden (ver Teoremas 4.2.4 y 4.2.6 en Estévez (2001)), su prueba se adapta de forma trivial a nuestro contexto.

Prueba de (25) para $i = 5$. Así como en el caso global, el término $CT^{(5)}_{l_n,x}(h)$ puede ser acotado tal y como sigue:

$$\begin{aligned}\left|CT^{(5)}_{l_n,x}(h)\right| \leq & \left|\frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)\left(F(X_i)-F_h(X_i)\right)^2}{\left(1-F(X_i)\right)^4}W_{n,x}(X_i)-\right.\\ & \left.-\int_a^b \frac{f(y)\left(F(y)-F_h(y)\right)^2}{\left(1-F(y)\right)^4}W_{n,x}(y)\,f(y)\,dy\right|+\\ & +\left|\frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)\left(F_h(X_i)-F_h^{-i}(X_i)\right)^2}{\left(1-F(X_i)\right)^4}W_{n,x}(X_i)\right|+\\ & +\left|\frac{2}{n}\sum_{i=1}^{n}\left(F(X_i)-F_h(X_i)\right)\left(F_h(X_i)-F_h^{-i}(X_i)\right)\frac{f(X_i)}{\left(1-F(X_i)\right)^4}W_{n,x}(X_i)\right|,\end{aligned}$$

de lo que se deriva (25) sin más que utilizar la equivalencia asintótica entre los errores locales $ASE_x(F_h)$ e $ISE_x(F_h)$ (ver Teorema 4.2.5 en Estévez (2001)), el Lema 2 de Estévez y Quintela (1999), la condición **[H.13]** y el hecho de que $F_h(y) - F_h^{-i}(y) = O\left(n^{r_1-1}\right)$ *c.s.*.

Prueba de (25) para $i = 3$. Con un modo de proceder análogo al del término precedente, se obtiene que

$$\begin{aligned}\left|CT^{(3)}_{l_n,x}(h)\right| \leq & \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\left(f_h^{-i}(X_i)-f(X_i)\right)\left(\overline{E}\left[F_h(X_i)/X_i\right]-F_h(X_i)\right)}{\left(1-F(X_i)\right)^3}W_{n,x}(X_i)\right|+\\ & +\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\left(f_h^{-i}(X_i)-f(X_i)\right)\left(F_h(X_i)-F_h^{-i}(X_i)\right)}{\left(1-F(X_i)\right)^3}W_{n,x}(X_i)\right|+\\ & +\left|\int_a^b\frac{\left(f_h(y)-f(y)\right)\left(EF_h(y)-F_h(y)\right)}{\left(1-F(y)\right)^3}W_{n,x}(y)\,f(y)\,dy\right|,\end{aligned}$$

que da lugar, trabajando como en el caso global, a (25) sin más que aplicar los Lemas 2, 3 y 12 de Estévez y Quintela (1999) y las hipótesis **[H.12]** y **[H.13]**.

Prueba de (25) para $i=4$. Llegado a este punto de la demostración, interesa considerar H_n' un subconjunto finito de H_n formado por puntos equiespaciados y tal que **[H.15]** $\#H_n' = n^{\rho}$, siendo ρ un número real verificando $2u(1+k)+\beta-v<\rho<2u(1+k)+\beta-v+V$.

Además, para cada $h \in H_n$ denotaremos por h^* el punto de H_n' más próximo a h. Pues bien, con esta notación la prueba de (25) para $i=4$ es consecuencia inmediata de los siguientes lemas:

Lema 6.1. *Bajo las hipótesis del Teorema 4.1, se verifica que:*

$$\sup_{h\in H_n'} \frac{\left|CT^{(4)}_{l_n^*,x}(h)\right|}{ISE_x(h)} \longrightarrow 0 \quad c.s..$$

Lema 6.2. *Bajo las hipótesis del Teorema 4.1, se verifica que:*

$$\sup_{h\in H_n} \frac{\left|CT^{(4)}_{l_n^*,x}(h) - CT^{(4)}_{l_n^*,x}(h^*)\right|}{ISE_x(h)} \longrightarrow 0 \quad c.s..$$

Lema 6.3. *Bajo las hipótesis del Teorema 4.1, se verifica que:*

$$\sup_{h\in H_n} \frac{\left|CT^{(4)}_{l_n,x}(h) - CT^{(4)}_{l_n^*,x}(h)\right|}{ISE_x(h)} \longrightarrow 0 \quad c.s..$$

Demostración del Lema 6.3. No es más que una réplica de la prueba del Lema 14 de Estévez y Quintela (1999) con la salvedad de que en lugar de su condición **[H.13]** se utiliza $\sup_{x,u} W_{n,x}(u) = O\left(n^{\beta}\right)$. De modo que,

$$\left|CT^{(4)}_{l_n,x}(h) - CT^{(4)}_{l_n^*,x}(h)\right| = O\left(n^{\beta+r_1-1}\right) \quad c.s.,$$

que junto con **[H.12]** y **[H.13]** concluye la demostración.

Demostración del Lema 6.2. Al igual que en el Lema 6.3, la prueba de este resultado se hace exactamente igual que en el caso global, con cuidado de acotar la función de pesos por n^{β} en lugar de hacerlo por una constante.

Demostración del Lema 6.1. Una vez más, la prueba es una réplica de la correspondiente en la situación global, en este caso del Lema 16 de Estévez y Quintela (1999). Un esquema de la misma se muestra a continuación.

El término $CT^{(4)}_{l^*_n,x}(h)$ puede escribirse como

$$-CT^{(4)}_{l^*_n,x}(h) = \frac{1}{nn_{l^*_n}} \sum\sum_{|i-j|>l^*_n} \Gamma_x(i,j),$$

donde

$$\Gamma_x(i,j) = (H_h(X_j,X_i) - F(X_j))\, D_{n,x}(X_j) - \int_a^b (H_h(y,X_i) - F(y))\, D_{n,x}(y)\, f(y)dy$$

y $D_{n,x}(y) = \dfrac{W_{n,x}(y)f(y)}{(1-F(y))^3}$. Definiendo

$$\Gamma^*_x(j) = \int (H_h(X_j,u) - F(X_j))\, D_{n,x}(X_j)\, f(u)du -$$

$$- \int_a^b \int (H_h(y,u) - F(y))\, D_{n,x}(y)\, f(y) f(u)dydu,$$

$$\Psi_x(i,j) = \Gamma_x(i,j) - \Gamma^*_x(j) \quad \text{y} \quad T_x(h) = \sum\sum_{|i-j|>l^*_n} \Psi_x(i,j),$$

y observando que $n_{l^*_n} \backsim n$, todo lo que tenemos que probar es

$$\sup_{h\in H'_n} n^{-2}h^{-2k}|T_x(h)| \longrightarrow 0 \quad c.s.$$

y

$$\sup_{h\in H'_n} h \left| \sum_{j=1}^{n} \Gamma^*_x(j) \right| \longrightarrow 0 \quad c.s..$$

La demostración de las convergencias anteriores sigue los mismos pasos que las correspondientes en el caso global (ver (22) y (23) en Estévez y Quintela (1999)), con la precaución, una vez más, de acotar las funciones de peso por n^{β} y posteriormente hacer uso de **[H.2]**.

Demostración del Teorema 4.2. Se observa que

$$r_h(y) - r(y) = \left(\frac{f_h(y)}{1-F_h(y)} - \frac{f(y)}{1-F(y)} \right) \frac{1-F_h(y)}{1-F(y)} + \tag{26}$$
$$+ \frac{\left(f_h(y) - r(y)\overline{F}_h(y)\right)\left(\overline{F}(y) - \overline{F}_h(y)\right)}{\overline{F}_h(y)\overline{F}(y)},$$

y puesto que $F_h(x)$ converge uniformemente a $F(x)$ sobre $x \in [a,b]$ y $h \in H_n$,

$$\left| \frac{\left(f_h(y) - r(y)\overline{F}_h(y)\right)\left(\overline{F}(y) - \overline{F}_h(y)\right)}{\overline{F}_h(y)\overline{F}(y)} \right| \leq c_n \left| \left(\frac{f_h(y)}{1-F_h(y)} - \frac{f(y)}{1-F(y)} \right) \frac{1-F_h(y)}{1-F(y)} \right|, \tag{27}$$

siendo c_n una sucesión de reales que tiende a cero. Por lo tanto, se obtiene que

$$ISE_x(h) = ISE_x^*(h) + o\left(ISE_x^*(h)\right),$$

y

$$ASE_x(r_h) = ASE_x^*(r_h) + o\left(ASE_x^*(r_h)\right),$$

donde

$$ISE_x^*(h) = \int_a^b \left[\left(\frac{f_h(y)}{1-F_h(y)} - \frac{f(y)}{1-F(y)} \right) \frac{1-F_h(y)}{1-F(y)} \right]^2 W_{n,x}(y)\, dy \tag{28}$$

y

$$ASE_x^*(h) = \sum_{i=1}^{n} \left[\left(\frac{f_h(X_i)}{1-F_h(X_i)} - \frac{f(X_i)}{1-F(X_i)} \right) \frac{1-F_h(X_i)}{1-F(X_i)} \right]^2 W_{n,x}(X_i) \tag{29}$$

son las versiones modificadas de los errores $ISE_x(h)$ y $ASE_x(h)$, respectivamente.

Por otra parte, cualquiera de las medidas locales modificadas puede ser vista como el error correspondiente, sin modificar, de $r_h'(y) = \frac{f_h(y)}{\overline{F}(y)} - \frac{r(y)\overline{F}_h(y)}{\overline{F}(y)}$ considerado como

estimador de la función 0. En otras palabras, $MISE_x^*(h) = MISE_x(r_h')$, $MSE_x^*(h) = MSE_x(r_h')$, $ISE_x^*(h) = ISE_x(r_h')$ y $ASE_x^*(h) = ASE_x(r_h')$. Si a todo ello añadimos el hecho de que $r_h'(\cdot)$ puede escribirse como un estimador tipo delta, es decir,

$$r_h'(y) = n^{-1}\sum_{i=1}^{n}\frac{K_h(y,X_i) - r(y)(1-H_h(y,X_i))}{1-F(y)} = n^{-1}\sum_{i=1}^{n}K_h^*(y,X_i),$$

la demostración de la primera expresión en el Teorema 4.2 será análoga a la del Teorema 3.1.1 de Quintela (1992) remplazando $K_h(\cdot)$ por $K_h^*(\cdot)$.

A continuación se da un esquema de la prueba de la segunda expresión en el Teorema 4.2. Tal y como se ha apuntado previamente, el $MISE_x^*(h)$ puede ser visto como el Error Cuadrático Medio Integrado de un estimador tipo delta de la función 0, esto es,

$$MISE_x^*(h) = E\int_a^b (r_h'(y))^2 W_{n,x}(y)\,dy.$$

Nótese además, que

$$r_h'(y) = \frac{f_h(y) - f(y)}{1-F(y)} + \frac{(F_h(y) - F(y))\,r(y)}{1-F(y)}. \tag{30}$$

Ahora bien, dado que es posible escribir

$$MISE_x^*(h) = E\int_a^b (r_h'(y) - Er_h'(y))^2 W_{n,x}(y)\,dy + \int_a^b (Er_h'(y))^2 W_{n,x}(y)\,dy,$$

una aplicación del Lema 4.2.4 de Estévez (2001) conduce a una descomposición del mismo en suma de dos términos, salvo una $o\left(\frac{1}{nh}\right)$. Tales términos,

$$S_x(r_h') = \int_a^b (Er_h'(y))^2 W_{n,x}(y)\,dy$$

y

$$\overline{R}_x(r_h') = \overline{E}\int_a^b (r_h'(y) - Er_h'(y))^2 W_{n,x}(y)\,dy$$

coinciden con sus homólogos en el caso de independencia, resultando por consiguiente de fácil manejo.

Así pues, en referencia con el término asociado al sesgo de $r'_h(\cdot)\left(S_x\left(r'_h\right)\right)$, es suficiente aplicar integración por partes, hacer ciertos cambios de varible y utilizar **[H.9]** para obtener:

$$E f_h(y) - f(y) = \int K(z) f(y - hz)\, dz - f(y) = \tag{31}$$
$$= h^k \left(\frac{1}{k!} \int t^k K(t)\, dt\right) f^{(k)}(y) + o\left(h^k\right)$$

y

$$EF_h(y) - F(y) = \int K(z) F(y - hz)\, dz - F(y) = \tag{32}$$
$$= h^k \left(\frac{1}{k!} \int t^k K(t)\, dt\right) F^{(k)}(y) + o\left(h^k\right),$$

y así poder concluir que

$$S_x\left(r'_h\right) = h^{2k} \int_a^b (B_r(y))^2 W_{n,x}(y)\, dy + o\left(h^{2k}\right), \tag{33}$$

con

$$B_r(y) = \left(\frac{1}{k!} \int t^k K(t)\, dt\right) \frac{\left(f^{(k)}(y) + r(y) F^{(k)}(y)\right)}{(1 - F(y))}.$$

En cuanto al término de la varianza $\left(\overline{R}_x(r_h)\right)$, en virtud de (30) se tiene que:

$$\overline{R}_x\left(r'_h\right) = \int_a^b \overline{E}\left[(f_h(y) - E f_h(y))^2\right] \frac{W_{n,x}(y)}{(1 - F(y))^2} dy +$$
$$+ \int_a^b \overline{E}\left[(F_h(y) - EF_h(y))^2\right] \frac{W_{n,x}(y) r^2(y)}{(1 - F(y))^2} dy +$$
$$+ 2 \int_a^b \overline{E}\left[(f_h(y) - E f_h(y))(F_h(y) - EF_h(y))\right] \frac{W_{n,x}(y) r(y)}{(1 - F(y))^2} dy,$$

reduciendo el problema de calcular varianzas y covarianzas de estimadores núcleo en el supuesto de dependencia al cálculo, notablemente más simple, de las mismas cantidades bajo independencia. Para el primer sumando es suficiente efectuar un cambio de variable y aplicar los desarrollos de Taylor de orden 0 y k a la función $f(\cdot)$ para obtener que

$$\text{(34)} \qquad \overline{E}\left[(f_h(y)-Ef_h(y))^2\right] = n^{-1}E\left[(K_h(y,X))^2\right] - n^{-1}\left[EK_h(y,X)\right]^2 = \\ = \frac{1}{nh}f(y)\int K^2(t)\,dt + o\left(\frac{1}{nh}\right).$$

Análogamente, para el sumando que involucra a la función de distribución y su suavizador, se consigue:

(35)

$$\overline{E}\left[(F_h(y)-EF_h(y))^2\right] = n^{-1}E\left[(H_h(y,X))^2\right] - n^{-1}\left[EH_h(y,X)\right]^2 = \\ = \frac{F(y)(1-F(y))}{n} - \frac{h}{n}\left(\int 2zK(z)H(z)\,dz\right)f(y) + o\left(\frac{h}{n}\right) = \\ = o\left(\frac{1}{nh}\right).$$

Finalmente, el término de la covarianza se puede acotar mediante:

$$\overline{E}\left[(f_h(y)-Ef_h(y))(F_h(y)-EF_h(y))\right] = \\ = \frac{1}{n}E\left[K_h(y,X)H_h(y,X)\right] - \frac{1}{n}EK_h(y,X)EH_h(y,X) \le C\frac{1}{n} = o\left(\frac{1}{nh}\right).$$

Por lo tanto, se obtiene que

$$\text{(36)} \qquad \overline{R}_x(r_h) = \frac{1}{nh}\int_a^b V_r^2(y)W_{n,x}(y)\,dy + o\left(\frac{1}{nh}\right),$$

con

$$V_r^2(y) = \left(\int K^2(t)\, dt \right) \frac{f(y)}{(1-F(y))^2}.$$

En definitiva, combinando (33) y (36) se consigue el resultado deseado.

7. AGRADECIMIENTOS

Deseamos expresar nuestra gratitud al profesor Philippe Vieu (Universidad Paul Sabatier, Toulouse, Francia) por sus comentarios, que nos han ayudado a realizar una mejor presentación del trabajo. También deseamos agradecer la gentileza del profesor Angel Venedikov (Academia de Ciencias de Bulgaria), que nos cedió un amplio fichero de datos con los terremotos sucedidos en California durante el periodo de tiempo estudiado en este artículo.

Este trabajo ha sido financiado por la Xunta de Galicia mediante el proyecto de investigación XUGA 10503A98 y por la DGES bajo el proyecto de investigación PB 98-0182-C02-01.

8. REFERENCIAS

Ahmad, I. A. (1976). «Uniform strong convergence of the generalized failure rate estimate». *Bull. Math. Statist.*, 17, 77-84.

Bowman, A. (1984). «An alternative method of cross-validation for the smoothing of density estimates». *Biometrika*, 71, 353-360.

Cox, D. R. & Oakes, D. (1984). *Analysis of survival data.* Chapman and Hall.

Doukhan, P. (1994). «Mixing: properties and examples». *Lecture Notes in Statistics*, 85, Springer-Verlag.

Chu, C. K. & Marron, J. S. (1991). «Comparison of two bandwidth selectors with dependent errors». *The Annals of Statistics*, 4, 1906-1918.

Estévez, G. (2001). *Estimación tipo núcleo de la función de razón de fallo bajo condiciones de dependencia.* Tesis Doctoral. Universidad de Santiago de Compostela.

Estévez, G. & Quintela, A. (1999). «Nonparametric estimation of the hazard function under dependent conditions». *Communications in Statistics-Theory and Methods*, 28, 10, 2297-2331.

Gasser, T. & Müller, H. G. (1984). «Estimating regression functions and their derivatives by the kernel method». *Scandinavian Journal of Statistics*, 11, 171-185.

Györfi, L., Härdle, W., Sarda, P. & Vieu, P. (1990). «Nonparametric curve estimation from time series», *Lecture Notes in Statistics*, 60, Springer-Verlag.

Härdle, W. & Vieu, P. (1992). «Kernel regression smoothing of time series», *Journal of Time Series Analysis*, 13, 209-232.

Hart, J. & Vieu, P. (1990). «Data-driven bandwidth choice for density estimation based on dependent data». *The Annals of Statistics*, 18, 873-890.

Hassani, S., Sarda, P. & Vieu, P. (1986). «Approche non paramétrique en théorie de la fiabilité». *Revue de Statistiques Appliquées*, vol. XXXV, n.° 4.

Hollander, M. & Proschan, F. (1984). «Nonparametric concepts and methods in reliability». *Handbook of Statistics*, vol. 4. P. R. Krishnaiah and P. K. Sen, eds., pp. 613-655, North-Holland, Amsterdam.

Klein, J. P. & Moeschberger, M. L. (1997). «Survival Analysis». *Techniques for censored and truncated data*. Springer-Verlag.

Lee, E. T. (1992). *Statistical methods for survival data analysis*. Wiley Series in Probability and Mathematical Statistics.

Lomnitz, C. & Hax, A. (1967). «Clustering in aftershock sequences». *The earth beneath the continents. Geophys. Monograph. 10, Am. Geophys. Union.*

Prakasa Rao, B. L. S. & Van Ryzin, J. (1985). «Asymptotic theory for two estimators of the generalized failure rate». *Statistical Theory and Data Analysis*. K. Matusita, ed., North-Holland, Amsterdam.

Quintela, A. (1992). *Cálculo del parámetro de suavización en la estimación no paramétrica de curvas con datos dependientes*. Tesis Doctoral, Universidad de Santiago de Compostela.

Quintela, A. (1994). «A plug-in technique in nonparametric regression with dependence». *Communications in Statistics-Theory and Methods*, 23, 2581-2603.

Quintela, A. & Vilar, J. M. (1992). «A local cross-validation algorithm for dependent data». *TEST*, 1, 123-153.

Reiss, R. D. (1981). «Nonparametric estimation of smooth distribution functions». *Scandinavian Journal of Statistics*, 8, 116-119.

Rice, J. & Rosenblatt, M. (1976). «Estimation of the log survivor function and hazard function». *Sankhyã*, A, 38, 60-78.

Rudemo, M. (1982). «Empirical choice of histograms and kernel density estimates». *Scandinavian Journal of Statistics*, 9, 65-78.

Sarda, P. & Vieu, P. (1989). «Estimation non paramétrique de la fonction de hasard». *Cahiers du C.E.R.O.*, 31, 241-265.

Singpurwalla, N. D. & Wong, M. (1983). «Estimation of the failure rate-A survey of nonparametric methods. Part I: Non-Bayesian methods». *Communications in Statistics-Theory and Methods*, 12, 559-588.

Udias, A. & Rice, J. (1975). «Statistical analysis of microearthquakes activity near San Andreas Geophysical Observatory. Hollister, California». *Bulletin of the Seismological Society of America*, 65, 809-828.

Vere-Jones, D. (1970). «Stochastic models for earthquake ocurrence». *Journal of the Royal Statistical Society*, Ser. B, 32, 1-62.

Vere-Jones, D. & Davies, D. (1966). «A statistical survey of earthquakes in the main seismic region of New Zealand, II. Time series analysis». *New Zealand J. Geol. Geophys*, 9, 251-284.

Vieu, P. (1991*a*). «Quadratic errors for nonparametric estimates under dependence». *Journal of multivariate analysis*, 39, 324-347.

Vieu, P. (1991*b*). «Nonparametric regression: optimal local bandwidth choice». *Journal of the Royal Statistical Society*, Ser. B, 53, 453-464.

Watson, G. S. & Leadbetter, M. R. (1964*a*). «Hazard Analysis I». *Biometrika*, 51, 175-184.

Watson, G. S. & Leadbetter, M. R. (1964*b*). «Hazard Analysis II». *Sankhyã*, A, 26, 110-116.

Youndjé, É., Sarda, P. & Vieu, P. (1996). «Optimal smooth hazard estimates». *TEST*, 5, 379-394.

ENGLISH SUMMARY

NONPARAMETRIC ESTIMATION OF THE HAZARD FUNCTION: APPLICATIONS TO SEISMOLOGY

GRACIELA ESTÉVEZ PÉREZ
ALEJANDRO QUINTELA DEL RÍO
Universidad de A Coruña*

We study the nonparametric estimation of the risk or hazard function of a real random variable. Through an uncensored random sample of data $X_1, X_2, \ldots, X_n$, not necessarily independent, we consider an estimator built as a ratio of a kernel estimator of the density function and a kernel estimator of the survival function. We consider the bandwidth selection problem in this estimator.
We show, by means of a simulation example, the improvement in the estimates if we use this estimator instead of the one that uses in the denominator the empirical distribution function. Finally, we present a practical application to data of earthquakes in California (United States) and Granada (Spain).

Keywords: Nonparametric estimation, strong mixing processes, cross-validation, hazard function

AMS Classification (MSC 2000): 62G05, 62G20, 62M99

* Departamento de Matemáticas-Facultad de Informática. Universidad de A Coruña. Campus de Elviña, s/n. 15071 A Coruña. Teléfono: 981167000, ext.: 2109.

- Received July 1999.
- Accepted October 2001.

1. INTRODUCTION

Let X be a real random variable with probability density function $f(\cdot)$ and distribution function $F(\cdot)$ with respect to some $\sigma-$finite measure on $\mathbb{R}$. The **failure rate** or **hazard function** is defined by $r(x) = \lim_{\Delta_x \longrightarrow 0} \dfrac{P(x \leq X < x+\Delta_x / X \geq x)}{\Delta_x}$.

By the definition of conditional probability, we have that

$$r(x) = \frac{f(x)}{1-F(x)} = \frac{f(x)}{\bar{F}(x)}, \tag{1}$$

considered when $\bar{F}(x) > 0$, that is, $r(\cdot)$ is defined in the set $S = \{x \in \mathbb{R} / \bar{F}(x) > 0\}$.

If we consider that the random variable X measures the «failure time», or the time of occurrence of a particular event in an homogeneous population, $r(x)\Delta_x$ can be interpreted as the approximate probability of one subject «fails» in the time interval $[x, x+\Delta_x)$, given the subject has survived to time x, i.e. the instantaneous probability of failure at x, given the last fail happened in the moment 0.

Through this point of view, the estimation of the hazard function is a problem of considerable interest, specially to medical researchers, logistics planners, reliability engineers, and seismologists. In medical research, there exists a lot of applications of the hazard estimation (Hassani, Sarda and Vieu (1986)). In seismology, the hazard rate function might be thought as the instantaneous risk of the occurrence of an earthquake at time x, known that at time 0 the last earthquake has happened (Rice and Rosenblatt (1976)).

A practical question is the estimation of $r(\cdot)$ based on a random sample $X_1, \ldots, X_n$ of X. This problem has been extensively discussed in the literature in the case of i.i.d. random variables. See, for example, Watson and Leadbetter (1964*a*, *b*), Ahmad (1976), Hollander and Proschan (1984) and Hassani, Sarda and Vieu (1986).

However, in several fields of applications the independence assumption is far from being realistic. This is for instance the case of microearthquake studies (Rice and Rosenblatt (1976)). Therefore, we will consider that the random data pertain to a strictly stationary sequence of strong mixing variables (Rosenblatt (1956)).

In this paper, we approach the nonparametric estimation of hazard function (1) under dependence conditions and no censored. In the next section we define the method of estimation and recall some of its properties. Section 3 shows some simulations that exhibit such properties. In Section 4 we study the local cross-validation method for the calculation of the bandwidths. Finally, in Section 5 we present an application to a set of microearthquake data.

2. KERNEL ESTIMATION OF HAZARD FUNCTION UNDER DEPENDENCE CONDITIONS

Several estimators proposed, and in particular the one which will be object of study in this work, are built as the ratio of a nonparametric estimator of $f(\cdot)$ and a nonparametric estimator of $\bar{F}(\cdot) = 1 - F(\cdot)$. (For a survey on different types of estimators of the hazard function see Rice and Rosenblatt (1976)). We will consider a smooth estimator of $r(\cdot)$, introduced by Watson and Leadbetter (1964*a*, *b*), but we will work in a context of α–mixing data (Rosenblatt (1956)):

$$r_h(x) = \frac{f_h(x)}{1 - F_h(x)}, \tag{2}$$

where $f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$ is the known Parzen-Rosenblatt estimator of $f(\cdot)$, and $F_h(\cdot)$ is the kernel estimator of $F(\cdot)$ defined by

$$F_h(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - X_i}{h}\right) = \int_{-\infty}^{x} f_h(t)dt. \tag{3}$$

with $K(\cdot)$ a kernel function, $H(x) = \int_{-\infty}^{x} K(u)du$ and $h = h(n) \in \mathbb{R}^+$ is the smoothing parameter, or bandwidth.

This estimator presents some advantages as compared to estimators that use the empirical distribution function to estimate the function $F(\cdot)$ $\left(r_{h,1}(x) = \frac{f_h(x)}{1 - F_n(x)}\right)$, such as its continuity, and presents improvements considered in terms of quadratic errors (see e.g. Youndjé, Sarda and Vieu (1996) and references therein). Consistency properties and asymptotic normality were studied by Estévez and Quintela (1999). The authors have also developed a smoothing parameter selection method for this estimator, through the least-squares cross-validation, showing that it is asymptotically optimal, in the sense that

$$\frac{d_I(\widehat{h}(l_n))}{\inf_h d_I(h)} \longrightarrow 1 \ a.s.$$

where $d_I(\cdot)$ is any of quadratic errors $ISE(\cdot)$, $ASE(\cdot)$ or $MISE^*(\cdot)$ and $\widehat{h}(l_n)$ is the cross-validation bandwidth with l_n a parameter depending of the amount of dependence.

3. SIMULATIONS

In order to compare the estimators $r_h(\cdot)$ and $r_{h,1}(\cdot)$ we carried out the following study of simulation: we generated M samples with N data from an autoregressive process of order one with distribution $N(0,1)$ or from a m-dependent model with distribution $\Gamma(1,1.5)$. For each one of the samples we calculated the cross-validation and the optimal bandwidth and the corresponding errors. The obtained results assure the kernel estimator $(r_h(\cdot))$ is behaved better than $r_{h,1}(\cdot)$. With regard to bandwidth selection, the results do not agree with what happened in density estimation (Hart and Vieu (1990)) or regression estimation (Quintela (1994)). Here, we can observe that the cross-validation bandwidth practically does not change with l_n. The horizontal convergence rates of cross-validation procedure maybe could explain this small influence of l_n.

4. LOCAL BANDWIDTH CHOICE

In this section we investigate a data-driven local selection method pertaining to the minimization, in each point x of an interval $[a,b]$, of the pointwise Mean Squared Error

$$MSE_x(h) = E(r_h(x) - r(x))^2. \tag{4}$$

To select, in each point x, a data-driven local bandwidth h_x we propose to minimize some estimate of MSE_x. The criterion used to estimate MSE_x is a local adaptation of the usual cross-validation rule. In effect, for each point x in the interval $[a,b]$, we will chose the bandwidth $\widehat{h}_x(l_n)$ minimizing the function:

$$CV_{x,l_n}(h) = \int_a^b r_h^2(y)\, W_{n,x}(y)\, dy - 2n^{-1} \sum_{i=1}^{n} \frac{f_h^{-i}(X_i)}{(1 - F_h^{-i}(X_i))\,(1 - F_n(X_i))} W_{n,x}(X_i), \tag{5}$$

where $f_h^{-i}(\cdot)$ and $F_h^{-i}(\cdot)$ are defined as in the global case and $W_{n,x}(\cdot)$ is a sequence of weight functions concentrated around x.

The local bandwidth also is asymptotically optimal, in the sense that

$$\sup_{x \in [a,b]} \frac{d_x(\widehat{h}_x(l_n))}{\inf_{h \in H_n} d_x(h)} \longrightarrow 1 \quad a.s.,$$

where $d_x(\cdot)$ is any of local quadratic errors $ASE_x(\cdot)$, $ISE_x(\cdot)$ or $MISE_x^*(\cdot)$. It is important to note that the previous result uses the asymptotic equivalence between such errors, that is,

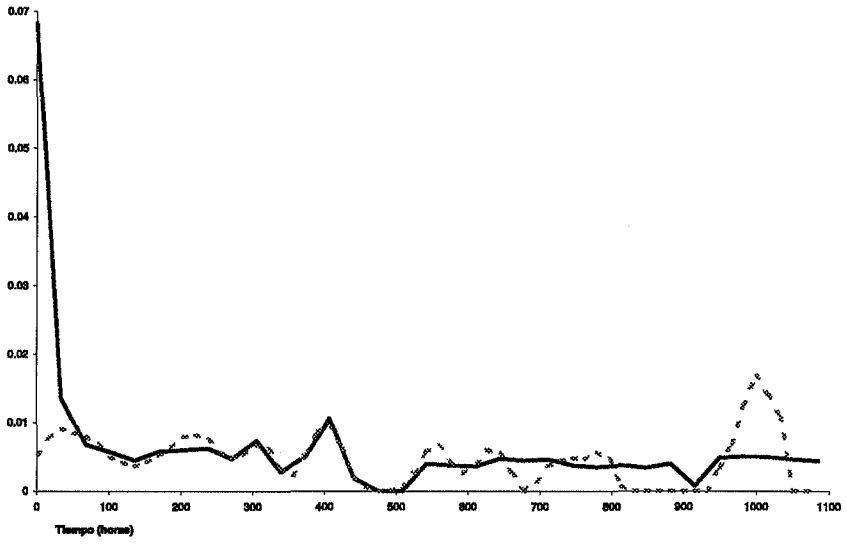
$$\sup_{h\in H_n} \frac{\left|d_x(h) - d'_x(h)\right|}{d'_x(h)} \longrightarrow 0 \quad a.s..$$

These results are collected in the Theorems 4.1 and 4.2 of the Spanish version of the paper.

5. APPLICATIONS TO SEISMOLOGY

Our applications deal with the seismicity analysis. We have applied the cross-validation procedure and the hazard kernel estimation to a geological set of data. Our time series are obtained of the global set of earthquakes observations recorded in two well different regions: California (USA) and Granada (Spain). From the whole data registered, we select our data sets and we consider the time series formed by the time intervals (in hours) between the occurrence of each two earthquakes. These samples are composed of dependent data (effecting several tests of independence, in no case it could be admitted the independence hypothesis between the data). Applying the cross-validation procedure to these time series, we obtain, for the region of Granada, the hazard kernel estimate using the cross-validation bandwidth.

We have also obtained the local bandwidth and the corresponding estimations. The resulting curves have been represented in the following figure:



We can observe that the local method allows us to estimate more efficiently the hazard function over the full sample interval than using a global estimate.

6. ACKNOWLEDGMENTS

We would like to express our gratitude to the professor Philippe Vieu (University Paul Sabatier, Toulouse, France) for its comments, which have helped us to prepare a better version of the paper. Thanks are also due to the professor Angel Venedikov (Academic of Sciences of Bulgaria), that gave us a large data file with the earthquakes of California. This research was financed by the Xunta de Galicia under project XUGA 10503A98 and by DGES under research project PB 98-0182-C02-01.

7. REFERENCES

Ahmad, I. A. (1976). «Uniform strong convergence of the generalized failure rate estimate». *Bull. Math. Statist.*, 17, 77-84.

Estévez, G. & Quintela, A. (1999). «Nonparametric estimation of the hazard function under dependent conditions». *Communications in Statistics-Theory and Methods*, 28, 10, 2297-2331.

Hart, J. & Vieu, P. (1990). «Data-driven bandwidth choice for density estimation based on dependent data». *The Annals of Statistics*, 18, 873-890.

Hassani, S., Sarda, P. & Vieu, P. (1986). «Approche non paramétrique en théorie de la fiabilité». *Revue de Statistiques Appliquées*, vol. XXXV, n.° 4.

Hollander, M. & Proschan, F. (1984). «Nonparametric concepts and methods in reliability». *Handbook of Statistics*, vol. 4. P. R. Krishnaiah and P. K. Sen, eds., pp. 613-655, North-Holland, Amsterdam.

Quintela, A. (1994). «A plug-in technique in nonparametric regression with dependence». *Communications in Statistics-Theory and Methods*, 23 , 2581-2603.

Rice, J. & Rosenblatt, M. (1976). «Estimation of the log survivor function and hazard function». *Sankhyã*, A, 38, 60-78.

Rosenblatt, M. (1956). «A central limit theorem and a strong mixing condition». *Proccedings of the National Academic Science*, 42, 43-47.

Watson, G. S. & Leadbetter, M. R. (1964*a*). «Hazard Analysis I». *Biometrika*, 51, 175-184.

Watson, G. S. & Leadbetter, M. R. (1964*b*). «Hazard Analysis II». *Sankhyã*, A, 26, 110-116.

Youndjé, É., Sarda, P. & Vieu, P. (1996). «Optimal smooth hazard estimates». *TEST*, 5, 379-394.

QÜESTIIÓ, vol. 25, 3, p. 479-498, 2001

LA MINERÍA DE DATOS, ENTRE LA ESTADÍSTICA Y LA INTELIGENCIA ARTIFICIAL

TOMÀS ALUJA
Universitat Politècnica de Catalunya*

En la pasada década hemos asistido a la irrupción de un nuevo concepto en el mundo empresarial: el «data mining» (minería de datos). Algunas empresas han implementado unidades de minería de datos estrechamente vinculados a la dirección de la empresa y en los foros empresariales las sesiones dedicadas a la minería de datos han sido las protagonistas. La minería de datos se presenta como una disciplina nueva, ligada a la Inteligencia Artificial y diferenciada de la Estadística. Por otro lado, en el mundo estadístico más académico, la minería de datos ha sido considerada en su inicio como una moda más, aparecida después de los sistemas expertos, conocida desde hacía tiempo bajo el nombre de «data fishing».

¿Es esto realmente así? En este artículo abordaremos las raíces estadísticas de la minería de datos, los problemas que trata, haremos una panorámica sobre el alcance actual de la minería de datos, presentaremos un ejemplo de su aplicación en el mundo de la audiencia de televisión y, por último, daremos una visión de futuro.

Data mining, between statistics and artificial intelligence

Palabras clave: Data mining, análisis de datos, modelización, inteligencia artificial, KDD, redes neuronales, árboles de decisión

Clasificación AMS (MSC 2000): 62-07, 68T10, 62P30

* Departamento de Estadística e Investigación Operativa. Universitat Politècnica de Catalunya (UPC). E-mail: tomas.aluja@upc.es

- Recibido en abril de 2001.
- Aceptado en noviembre de 2001.

1. INTRODUCCIÓN

El almacenamiento de datos se ha convertido en una tarea rutinaria de los sistemas de información de las organizaciones. Esto es aún más evidente en las empresas de la nueva economía, el e-comercio, la telefonía, el marketing directo, etc. Los datos almacenados son un tesoro para las organizaciones, es donde se guardan las interacciones pasadas con los clientes, la contabilidad de sus procesos internos, representan la memoria de la organización. Pero con tener memoria no es suficiente, hay que pasar a la acción inteligente sobre los datos para extraer la información que almacenan. Este es el objetivo de la Minería de Datos.

En primer lugar situemos la minería de datos a partir de algunas definiciones que se ha dado sobre la misma:

> Data Mining: «the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners» (Hand, 1998)
>
> «Iterative process of extracting hidden predictive patterns from large databases, using AI technologies as well as statistics techniques» (Mena, 1999)

La última, sin ser la definición más popular, enfatiza, sin embargo, cuáles son las raíces de la minería de datos: la Inteligencia Artificial (en particular *Machine learning)* y la Estadística.

Si buscamos a su vez definiciones de estas dos disciplinas:

> *Machine learning*: «a branch of AI that deals with the design and application of learning algorithms» (Mena, 1999)
>
> Estadística: «a branch of Applied Mathematics, and may be regarded as mathematics applied to observational data ... Statistics may be regarded as
>
> 1. the study of populations
> 2. the study of variation
> 3. the study of methods of the reduction of data»
>
> (Fisher, 1925)
>
> «methodology for extracting information from data and expressing the amount of uncertainity in decisions we make» (C. R. Rao, 1989).

Observemos que ya en 1925, Sir R. Fisher consideró la estadística bajo tres ópticas diferentes, como el estudio de poblaciones, lo cual está en el propio origen de la disciplina, como el estudio de la variabilidad que permite la modelización de los fenómenos

teniendo en cuenta la aleatoriedad presente en la naturaleza y como métodos de síntesis de la información contenida en los datos. La definición más moderna de C. R. Rao permite resaltar las coincidencias con la definición de minería de datos presentada más arriba.

Es claro que para cualquier persona vinculada con la estadística puede hablarse de dos tipos de estadística, una que podemos denominar Estadística Exploratoria (≅ Data Analysis) y otra que podemos denominar Estadística Inferencial (≅ modelling).

Si bien las fronteras entre ambos tipos de estadística no siempre es fácil de establecer, y a menudo la primera se presenta como la fase previa de la segunda (Cox & Snell, 1982) (Rao, 1989), existe una diferencia conceptual importante entre ambos tipos de Estadística. La Estadística Inferencial se refiere al paradigma central del quehacer estadístico, esto es, decidir entre varias hipótesis a partir de las consecuencias observadas. Consiste en incorporar la aleatoriedad dentro de la decisión (ya sea en su forma paramétrica mediante el razonamiento deductivo para delucidar la verosimilitud de cada hipótesis o por métodos computacionales).

«Data analysis» tiene un sentido muy general de estadística aplicada (con una connotación de aproximación pragmática e informatizada). Jean Paul Benzecri expresaba perfectamente el espíritu de este enfoque cuando afirmaba en sus cursos de 1965 que «le modèle doit suivre les données et non l'inverse» (si bien «Data Analysis» es más amplio que el equivalente francés de «analyse des données», el cual queda circunscrito a Análisis Multivariante Exploratorio). En este enfoque, no se trata de no tener en cuenta la naturaleza aleatoria de los datos (es obvio, por ejemplo, que cuando se seleccionan las componentes principales «significativas» para realizar una clasificación se está tratando de eliminar la parte aleatoria de los datos), sino que primero son los datos y es a partir de estos que se busca manifestar la información relevante para los problemas planteados.

Se puede constatar, sin embargo, que muchos de los problemas abordados en Análisis de Datos son comunes con la Inteligencia Artificial. Estas dos disciplinas, como a menudo sucede en el entorno académico, se han desarrollado la una a espaldas de la otra, dando lugar a nomenclaturas totalmente diferentes para problemas iguales. La Tabla 1, elaborada por el profesor L. Lebart, muestra las equivalencias para el problema de la predicción con redes neuronales.

Resumiendo mucho, podemos decir que la Inteligencia Artificial ha estado más preocupada en ofrecer soluciones algorítmicas con un coste computacional aceptable, mientras que la Estadística se ha preocupado más del poder de generalización de los resultados obtenidos, esto es, poder inferir los resultados a situaciones más generales que la estudiada.

Tabla 1. Equivalencias de nomenclatura entre la Estadística y la Inteligencia Artificial para el problema de predicción por redes neuronales.

Inteligencia Artificial	*Estadística*
red (network)	modelo
ejemplos (patterns)	observaciones, individuos
features, inputs, outputs	variables
inputs	variables explicativas
outputs, targets	variables de respuesta
errores	residuos
training, learning	estimación
función de error, coste	criterio de ajuste
pesos, coef. sinápticos	parámetros
aprendizaje supervisado	regresión, discriminación
aprenendizaje no supervisado	clasificación

Como mera ilustración de las aportaciones de ambas disciplinas al problema de la predicción, señalemos los hitos históricos de la regresión para la predicción de una variable continua (Galton, 1890), el análisis discriminante para la predicción de una variable nominal (Fisher, 1937), el AID para la construcción de árboles de decisión (Sonquist y Morgan, 1964), MARS (Friedman, 1991)... en Estadística, mientras que en el campo de la Inteligencia Artificial podemos citar el perceptrón, antecedente de las modernas redes neuronales (Rosemblat, 1958), los sistemas expertos, secuencias de reglas «if - then – else», para la toma de decisiones en los años setenta, los algoritmos genéticos (Holland, 1970), también los árboles de decisión (Quinlan, 1986)...

1.1. Nuevos problemas

La progresiva utilización de los avances tecnológicos por las empresas e instituciones hace aparecer nuevas colectas de datos y nuevos problemas. «Development in hardware have contributed to statistics by giving us many new and interestings sorts of data to analyse. Data have been able to be captured and stored quickly and cheaply by spectrometers, telescopes, process measuring devices, ... From these instruments have come new research problems. New applications have not arisen in science alone. Hardware changes have led to sophisticated point-of-sales terminals, bar-code readers and the ability to store and recall the huge volumes of data that are constantly being collected in warehouses, retail stores, government departments and financial institutions. Attempts to use such data to improve business performance have led to the field of data mining» (Cameron, 1997).

Un campo privilegiado de aplicación de las técnicas de minería de datos es el marketing, concretamente todo aquello que se agrupa bajo el nombre de CRM (*costumer relationship management*), donde el objetivo es conocer lo mejor posible los clientes para poder satisfacerlos mejor y asegurar así la rentabilidad de las empresas. Problemas tales como estimar el potencial económico de los clientes, modelizar la probabilidad de baja, medir la satisfacción por el servicio, descubrir nuevos segmentos de clientes potenciales, etc., son problemas que los responsables de la acción comercial de las empresas deben afrontar.

Pero no sólo las empresas o las instituciones son generadoras de nuevos problemas que afrontar, otros campos científicos también generan nuevos problemas donde la minería de datos se convierte en imprescindible, tales como las investigaciones originadas a raíz del proyecto Genoma, ¿qué secuencias de genes motivan la aparición de enfermedades?, ¿lo hacen de forma determinista o en probabilidad? También la información transmitida por satélite puede proporcionar avances a fenómenos hasta hoy difíciles de explicar, tales como la vulcanología, los terremotos o el clima, etc. La Tierra está dejando de ser el marco de referencia único para serlo cada vez más el sistema solar, como lo prueba la influencia que tienen las erupciones solares en las telecomunicaciones vía satélite. Otros campos de gran actualidad son encontrar métodos de predicción fiables, rápidos y baratos sobre la composición de los alimentos a partir del análisis del espectro infrarrojo de estos alimentos u otros análisis químicos.

Todo esto comporta la necesidad de tratar tablas de datos complejos y de tamaño inimaginable hasta ahora. Esta situación es nueva para el estadístico y bastante alejada de la clásica muestra aleatoria de observaciones independientes formada por algunas decenas de variables y unos cuantos millares de individuos. Tal como señala D. Hand (1998), ahora los datos son «secondary, messy, with many missings, noisy and not representative». Esto supone un reto para la estadística que obligará a repensar los esquemas clásicos de la inferencia estadística y de significación de los resultados observados. Si bien el nivel de significación fisheriano continua siendo válido para detectar la discrepancia entre los datos observados y la hipótesis formulada, es obvio que el orden de magnitud de los «p-value» es ahora muy inferior al acostumbrado. También es claro que en este contexto cobran renovada importancia los métodos de inducción computacionales, de simulación por Monte Carlo, bootstrap, etc.

2. ¿QUÉ PROBLEMAS ABORDA LA MINERÍA DE DATOS?

Cualquier problema para el que existan datos históricos almacenados es un problema susceptible de ser tratado mediante técnicas de Minería de Datos. Sin pretender ser exhaustivos la siguiente es una lista ilustrativa:

Búsqueda de lo inesperado por descripción de la realidad multivariante. Un principio clásico de la Estadística, el principio de la parsimonia, ya no es ahora válido (si bien siempre serán preferibles los modelos simples). Para describir un fenómeno cuantas más variables tengamos mejor, más ricas, más globales y más coherentes serán las descripciones y más fácil será detectar lo inesperado, esto es, aquello que no habíamos previsto y que resulta valioso para entender mejor el comportamiento de algún grupo de individuos, lo cual se ve favorecido por el hecho de trabajar con muestras grandes. Las muestras aleatorias son suficientes para describir la regularidad estadística global, pero no para detectar comportamientos particulares de subgrupos.

Búsqueda de asociaciones. Un cierto suceso, ¿está asociado a otro suceso?, ¿podemos inferir que determinados sucesos ocurren simultaneamente más de lo que seria esperable si fuesen independientes?, ¿es posible sugerir un producto, sabiendo que otro ha sido adquirido?

Definición de tipologías. Los consumidores son, a efectos prácticos, infinitos, pero los tipos de consumidores distintos son un número mucho más pequeño. Detectar estos tipos distintos, su perfil de compra y proyectarlos sobre toda la población, es una operación imprescindible a la hora de programar una política de marketing. Por otro lado, las tipologías no tienen que ser necesariamente de consumo, pueden ser de opiniones, valores, condiciones de vida, etc.

Detección de ciclos temporales. Todo consumidor sigue un ciclo de necesidades que ocasionan actos de compra distintos a lo largo de su vida. Detectar los diferentes ciclos y la fase donde se sitúa cada consumidor ayudará a crear complicidades y adecuar la oferta de productos a las necesidades y crear fidelización.

Predicción. A menudo deberemos efectuar predicciones: ¿cuál es la probabilidad de baja de un cliente?, ¿cuál es el precio de una vivienda concreta?, ¿lloverá mañana? Estas y muchas más son preguntas que deberemos responder, para ello construiremos un modelo a partir de los datos históricos. Si la variable de respuesta es continua (p. e. la rentabilidad de un cliente) diremos que se trata de un problema de regresión, mientras que si la variable de respuesta es categórica (p. e. la compra o no de un producto) diremos que se trata de un problema de clasificación.

3. LAS TÉCNICAS

En general, cualquiera que sea el problema a resolver, no existe una única técnica para solucionarlo, sino que puede ser abordado siguiendo aproximaciones distintas. El número de técnicas es muy grande y sólo puede crecer en el futuro. También aquí, sin pretender ser exhaustivos, la siguiente es una lista de técnicas con una breve reseña.

Análisis Factoriales Descriptivos. Permiten hacer visualizaciones de realidades multivariantes complejas y, por ende, manifestar las regularidades estadísticas, así como eventuales discrepancias respecto de aquella y sugerir hipótesis de explicación.

«Market Basket Analysis» o análisis de la cesta de la compra. Permite detectar qué productos se adquieren conjuntamente, permite incorporar variables técnicas que ayudan en la interpretación, como el día de la semana, localización, forma de pago. También puede aplicarse en contextos diferentes del de las grandes superficies, en particular el e-comercio, e incorporar el factor temporal.

Técnicas de «clustering». Son técnicas que parten de una medida de proximidad entre individuos y a partir de ahí, buscar los grupos de individuos más parecidos entre sí, según una serie de variables mesuradas.

Series Temporales. A partir de la serie de comportamiento histórica, permite modelizar las componentes básicas de la serie, tendencia, ciclo y estacionalidad y así poder hacer predicciones para el futuro, tales como cifra de ventas, previsión de consumo de un producto o servicio, etc.

Redes bayesianas. Consiste en representar todos los posibles sucesos en que estamos interesados mediante un grafo de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones.

Modelos Lineales Generalizados. Son modelos que permiten tratar diferentes tipos de variables de respuesta, por ejemplo la preferencia entre productos concurrentes en el mercado. Al mismo tiempo, los modelos estadísticos se enriquecen cada vez más y se hacen más flexibles y adaptativos, permitiendo abordar problemas cada vez más complejos: (GAM, Projection Pursuit, PLS, MARS, ...).

Previsión local. La idea de base es que individuos parecidos tendrán comportamientos similares respecto de una cierta variable de respuesta. La técnica consiste en situar los individuos en un espacio euclídeo y hacer predicciones de su comportamiento a partir del comportamiento observado en sus vecinos.

Redes neuronales. Inspiradas en el modelo biológico, son generalizaciones de modelos estadísticos clásicos. Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo. Permite aprender en contextos difíciles, sin precisar la formulación de un modelo concreto. Su principal inconveniente es que para el usuario son una caja negra.

Árboles de decisión. Permiten obtener de forma visual las reglas de decisión bajo las cuales operan los consumidores, a partir de datos históricos almacenados. Su principal ventaja es la facilidad de interpretación.

Algoritmos genéticos. También aquí se simula el modelo biológico de la evolución de las especies, sólo que a una velocidad infinitamente mayor. Es una técnica muy prometedora. En principio cualquier problema que se plantee, como la optimización de una combinación entre distintas componentes, estando estas componentes sujetas a restricciones, puede resolverse mediante algoritmos genéticos.

Un enriquecimiento de las posibilidades de análisis son los sistemas híbridos, esto es, la combinación de dos o más técnicas para mejorar la eficiencia en la resolución de un problema, como por ejemplo, utilizar un algoritmo genético para inicializar una red neuronal, o bien utilizar un árbol decisión como variable de entrada en una regresión logística.

En el futuro, el campo de actuación de la minería de datos no puede sino crecer. En particular debemos mencionar en estos momentos el análisis de datos recibidos por internet y «on line», dando lugar al *web mining*, donde las técnicas de *data mining* se utilizan para optimizar las interacciones a través de la *web*. ¿Cuáles son las secuencias de páginas más visitadas?, ¿qué páginas visitan los que compran?, ¿los que compran, vuelven a conectarse?, ¿cuales son las «killer pages»?, ¿una vez efectuada una adquisición, qué productos puedo sugerir?, son algunas de las preguntas que los responsables de comercio electrónico de las empresas se están formulando en estos momentos.

También los datos objeto de análisis pueden ser textos, dando lugar al *text mining*. Esto es particularmente útil en el análisis de las encuestas de satisfacción percibida por los usuarios. La utilización de las frases realmente escritas supone un enriquecimiento de los análisis realizados sólo con información numérica. También la utilización del *text mining* para la síntesis y la presentación de la información encontrada en la web es un campo actual de investigación. Más a largo plazo podrán utilizarse la voz o las imágenes.

Otra de las nuevas vías de investigación es el *fuzzy mining*, esto es, la utilización de las técnicas de minería de datos con objetos simbólicos, que representen más fidedignamente la incertidumbre que se tiene de los objetos que se estudian.

La tendencia actual más prometedora sería la de integrar los dos puntos de vista, provinientes de la estadística y de la Inteligencia Artificial, en las soluciones algorítmicas propuestas, de forma de aprovechar los puntos fuertes de ambas disciplinas. En consecuencia los algoritmos deberían contemplar las dos siguientes propiedades básicas:

Poder de generalización a poblaciones diferentes de la observada. Lo cual implica implementar técnicas eficientes de validación de resultados, ya sea a partir del conocimiento de la distribución muestral de los estadísticos del modelo o por métodos computacionales como la validación cruzada, etc.

Escalabilidad. Dado el volumen de datos a tratar, el coste de los algoritmos ha de ser todo lo lineal que sea posible respecto de los parámetros que definen el coste, en particular respecto del número de individuos.

4. COMPARACIÓN DE TÉCNICAS

Una pregunta que nos podemos formular es cuál es el mejor método para resolver un problema. La experiencia nos muestra que excepto ciertos problemas específicos y difíciles, la mayoría de problemas abordados en minería de datos dan resultados comparables cualquiera que sea la técnica utilizada. Hemos realizado una prueba con un fichero de 4000 individuos y 15 variables para explicar dos variables de respuesta sobre la adquisición de un cierto producto, el primero es un producto que podríamos calificar de relativamente «fácil» de predecir, mientras que el segundo es claramente más difícil. Hemos efectuado la predicción de ambas variables mediante 4 técnicas alternativas:

- Análisis Discriminante
- Redes neuronales
- Árboles de decisión
- Regresión Logística

Para medir la calidad de la predicción por cada método, hemos seleccionado al azar 4 muestras de 1000 individuos cada una como muestras de aprendizaje, utilizando los 3000 restantes como muestra de validación. Para cada método hemos realizado 3 ejecuciones cambiando ligeramente los parámetros del modelo. Por tanto, en total disponemos de 12 ejecuciones por método. Tomando el promedio de la probabilidad de acierto en las muestras de aprendizaje y en las de validación obtenemos los resultados que se muestran en la Tabla 2:

Tabla 2. Comparación de la probabilidad de acierto según 4 métodos de predicción.

Problema 1	*Apren.*	*Test*
Análisis Discriminante	71.13%	**69.71%**
Redes Neuronales	71.63%	**69.12%**
Árboles de Clasificación	72.94%	**70.31%**
Regresión logística	74.18%	**71.33%**
Problema 2	*Apren.*	*Test*
Análisis Discriminante	62.18%	**61.39%**
Redes Neuronales	62.29%	**60.19%**
Árboles de Clasificación	62.70%	**61.03%**
Regresión logística	65.28%	**59.36%**

Observando los resultados vemos que las probabilidades de acierto en la muestra de validación son bastante parecidas para los cuatro tipos de modelos utilizados.

5. EJEMPLO DE APLICACIÓN. DEFINICIÓN DE TARGETS COMPORTAMENTALES DE CONSUMO TELEVISIVO

Las innovaciones tecnológicas en el mundo audiovisual, producen el almacenamiento de una cantidad ingente de datos. El análisis de estos datos permite una mejora en la toma de decisiones por parte de las organizaciones implicadas.

En audiometría se dispone de información minuto a minuto de la audiencia realizada por un panel de familias. Estas observaciones pasan por un proceso de validación y enriquecimiento a partir de los datos sociodemográficos disponibles sobre los panelistas y por el minutado de programas y spots. Posteriormente, la muestra obtenida se afecta con un factor de elevación para obtener datos a nivel poblacional.

El problema planteado es el de definir targets compuestos explicativos del consumo de programas del género «Revistas del corazón», el cual ha experimentado un notable aumento en los últimos años en la programación televisiva.

Los datos analizados han sido todos los programas de este género emitidos durante el año 1997. La variable de respuesta ha sido los minutos semanales de visión de los programas de tipo *rosa*, en las cadenas estatales y para todos los individuos mayores de 3 años. Las variables explicativas son todas las sociodemográficas.

En el año 1997 se realizaron un total de 707 emisiones para este tipo de programas con una audiencia promedio de 32 minutos semanales por individuo.

El interés del problema planteado es claro, tanto para las propias Televisiones y Productoras, como técnica alternativa para definir targets afines a cualquier programa, como para las empresas de publicidad, anunciantes, agencias o centrales, al poder efectuar un «matching» entre los targets de consumo televisivo con el target consumidor del producto anunciado y así poder ser introducido en un sistema de «media-planing» para la compra de publicidad.

Clásicamente, este problema se soluciona mediante técnicas estadísticas simples, como es la distribución por variable o análisis de perfiles simples. La simplicidad de esta técnica es su principal ventaja. Así, por ejemplo, en el histograma de la Figura 1, mostrando el perfil de la audiencia de dibujos animados respecto de la edad, es clara la preferencia del segmento de niños (4 a 15 años) de este tipo de programas, pero también se observa una cierta afinidad con el segmento de personas jubiladas (más de 65 años). El histograma no revela si esta afinidad es propia del segmento o es debida a la presencia de niños en el hogar.

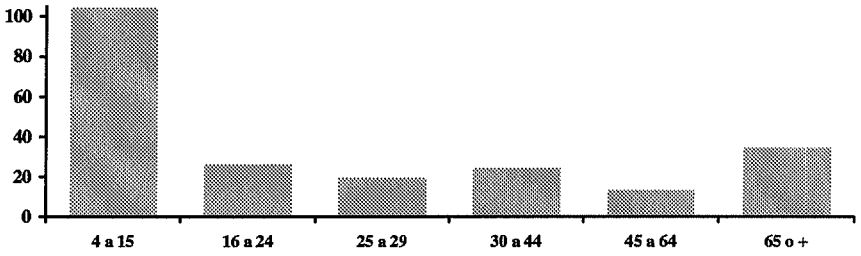


Figura 1. Perfil de audiencia de dibujos animados según la edad.

Una forma de obtención directa de targets compuestos del consumo televisivo de los programas *rosa*, es la utilización de árboles de decisión para explicar la audiencia de este tipo de programas. Escogemos esta metodología por su aplicabilidad inmediata de los resultados obtenidos. Estos resultados se obtienen de forma visual. La Figura 2 esquematiza el proceso de Minería de Datos en audiometría.

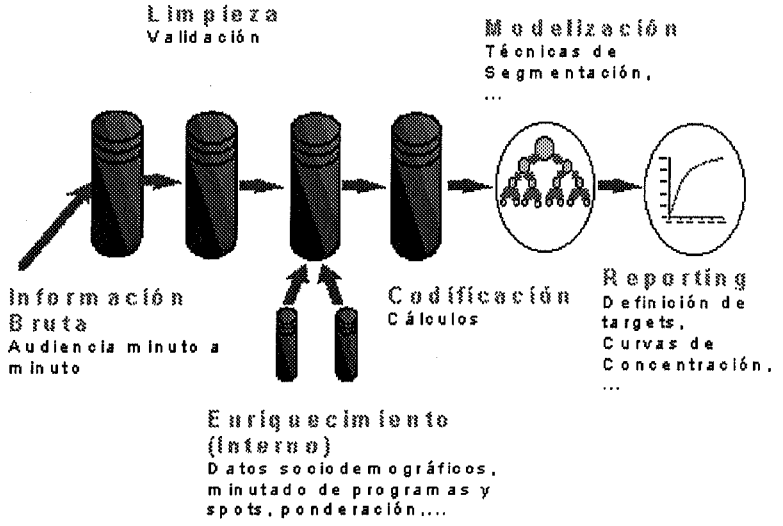


Figura 2. Sistema de KDD en audiometría.

El proceso de generación de un árbol es un proceso iterativo. Empieza situando toda la muestra disponible en el nodo raíz, a partir del cual, por sucesivas particiones, se obtienen las ramas del árbol hasta los nodos terminales u hojas, formadas por conjuntos de individuos que han visto un número similar de minutos los programas *rosa*.

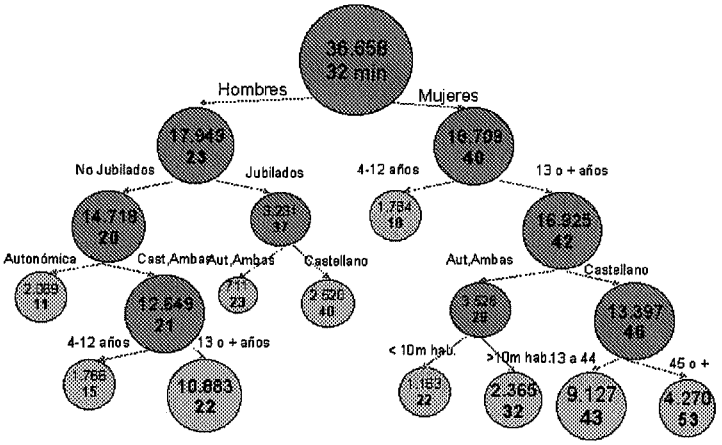


Figura 3. Ejemplo de árbol de decisión.

El algoritmo de construcción de un árbol de decisión implementa el siguiente bucle:

Hacer para cada nodo:

1. Verificar el criterio de parada del proceso en el nodo.
2. Definir la lista de todas las particiones posibles del nodo.
3. Seleccionar la partición óptima.
4. Generar la partición seleccionada.

La Figura 3 ilustra las primeras particiones del árbol generado.

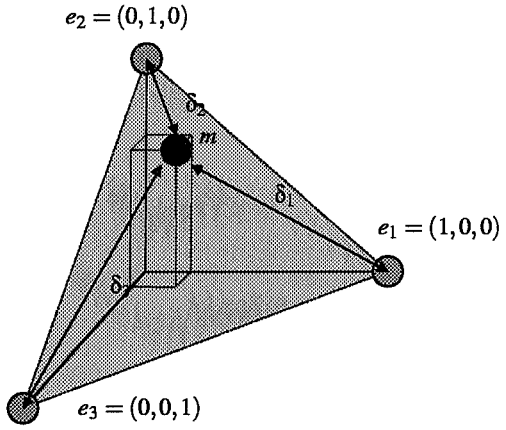
El árbol obtenido ilustra bien el proceso seguido, cada nodo da el número de individuos que contiene y el promedio de audiencia de estos individuos en los programas *rosa*. La obtención visual de los resultados permite a su vez su crítica, en efecto, la programación ofrecida por cada cadena condiciona la visión que puede hacerse de sus programas.

Existen varios algoritmos para la construcción de árboles de decisión: AID, CHAID, CART, C4.5. La diferencia básica es, aparte del hecho de que los árboles generados sean binarios o n-arios, la definición de la partición óptima de un nodo. Ciertamente seleccionar la partición óptima implica previamente definir un criterio de optimalidad.

Nosotros expresamos el criterio a optimizar en función de la pureza del nodo $i(t)$, definida por la siguiente fórmula:

$$i(t) = \frac{\sum_{i \in t} w_{it}\, \delta(i, m_t)}{\sum_{i \in t} w_{it}}$$

Función de los pesos de los individuos w_{it} del nodo y de las distancias de estos individuos al representante del nodo m_t.



Para el caso de variables de respuesta continuas y utilizando la métrica euclídea (norma L_2), la fórmula anterior se reduce a la conocida fórmula de la variancia de la variable de respuesta (y_i) en el nodo:

$$i(t) = \frac{\sum_{i \in t} (y_i - \bar{y}_t)^2}{n_t}$$

En este caso es obvio que efectuar una partición de un nodo implica descomponer la variancia total (V_T) del nodo original en dos componentes, una es la variancia *intra* (V_w) y la otra es la variancia *inter* (V_b):

$$V_T = V_b + V_w$$

Por tanto, maximizar la pureza de los nodos hijos implica minimizar V_w y por consiguiente maximizar V_b, esto es, encontrar dos nodos con la diferencia de medias lo más significativa posible (teorema de Huyghens).

El otro criterio a verificar en cada nodo es el criterio de parada, ya que en caso contrario podemos hacer crecer un árbol hasta que todos los nodos sean puros o contengan un solo individuo. Es evidente que entonces habríamos sobreparametrizado el árbol. Cuanto más avanzamos en la construcción del árbol, menos fiables son las particiones que se obtienen. Una manera de evitar esto es utilizar una técnica de validación como criterio de parada. Cuando se produzcan diferencias significativas entre la muestra de aprendizaje y la de validación, significa que las particiones no son estables.

La figura 5 muestra la calidad del árbol en función de su tamaño. En la muestra de aprendizaje esta medida es siempre monótona creciente, mientras que en la muestra de

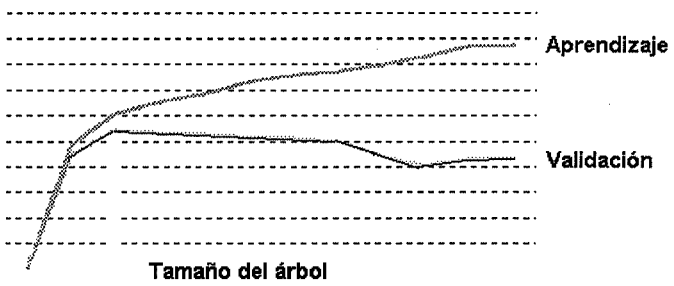


Figura 5. Calidad de un árbol en función de su tamaño.

validación, a partir de un cierto tamaño se estabiliza y puede llegar a decrecer, indicando que las particiones efectuadas más allá de este nivel son producto del azar.

6. CONCLUSIONES

La experiencia práctica muestra claramente la aptitud de las técnicas de minería de datos para resolver problemas empresariales. También es clara su aportación para resolver problemas científicos que impliquen el tratamiento de grandes cantidades de datos.

La minería de datos es, en realidad, una prolongación de una práctica estadística de larga tradición, la de Análisis de Datos. Existe, además, una aportación propia de técnicas específicas de Inteligencia Artificial, en particular sobre la integración de los algoritmos, la automatización del proceso y la optimización del coste.

A diferencia de la IA, que es una ciencia joven, en Estadística se viene aprendiendo de los datos desde hace más de un siglo, la diferencia consiste que ahora existe la potencia de cálculo suficiente para tratar ficheros de datos de forma masiva y automática. Esta es una realidad que cada vez será más habitual. Sin abandonar ninguno de los campos previamente abordados, la Estadística ha evolucionado de ocuparse de la contabilidad de los estados a ser la metodología científica de las ciencias experimentales, hasta ser un «problem solver» para las organizaciones modernas. Es por esta razón el énfasis dado a que los resultados sean accionables.

Por otro lado y en relación a la amplia panoplia de técnicas disponibles, conviene tener claro de que no existe la técnica más inteligente, sino formas inteligentes de utilizar una técnica y que cada uno utiliza de forma inteligente aquello que conoce. También que para la mayoría de problemas no existen diferencias significativas en los resultados obtenidos.

Por todo lo dicho, es nuestra opinión de que la minería de datos no es una moda pasajera, sino que se entronca en una vieja tradición estadística y que cada vez más debe servir para hacer más eficiente el funcionamiento de las organizaciones modernas, ayudar a resolver problemas científicos y ampliar los horizontes de la Estadística.

7. BIBLIOGRAFÍA

Adriaans, P. & Zantige, D. (1996). *Data mining*. Addison-Wesley.

Aluja, T. & Morineau, A. (1999). *Aprender de los datos: el análisis de componentes principales, una aproximación desde el data mining*. EUB. Barcelona.

Aluja, T. & Nafria, E. (1996). «Automatic segmentation by decision trees». *Proceedings on Computational Statistics COMPSTAT 1996,* ed. A. Prat. Physica Verlag.

Aluja, T. & Nafria, E. (1998*a*). «Robust impurity measures in Decision Trees». *Data Science, Classification and related methods*, ed. C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H-H. Bock and Y. Baba. Springer.

Aluja, T. & Nafria, E. (1998*b*). «Generalised impurity measures and data diagnostics in decision trees». *Visualising Categorical Data*, ed. Jörg Blasius and M. Greenacre. Academic Press.

Aluja, T. (2000). «Los nuevos retos de la estadística, el Data Mining». *Investigación y Marketing*, 68, 3, 34-38. AEDEMO.

Benzécri, J.-P. & *coll.* (1973). *La Taxinomie, Vol. I, L'Analyse des Correspondances, Vol. II*, Dunod, Paris.

Berry, M. J. A. & Linoff, G. (1997). *Data mining techniques for marketing, sales and customer support*. J. Wiley.

Beveridge, W. H. (1944). *Full employed in a free society*. George Allen and Unwin.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press.

Booker, L. B., Goldberg, D. E. & Holland, J. H. (1989). *Classifier systems and genetic algorithms*. Springer-Verlag.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Waldsworth International Group, Belmont, California.

Cameron, M. (1997). «Current influences of Computing on Statistics». *International Statistical Review*, 65, 3, 277-280.

Celeux, G. (Ed.) (1990). *Analyse discriminante sur variables continues*, coll. didactique, INRIA.

Celeux, G. & Lechevallier, Y. (1982). «Méthodes de Segmentation non Paramétriques». *Revue de Statistique Appliquée*, XXX(4), 39-53.

Celeux, G. & Nakache, J. P. (1994). *Analyse discriminante sur variables qualitatives.* Polytechnica.

Ciampi, A. (1991). «Generalized Regression Trees». *Computational Statistics and Data Analysis*, 12, 57-78. North Holland.

Cox, D. R. & Snell, E. J. (1982). *Applied Statistics. Principles and Examples.* Chapman and Hall.

Elder, J. F. & Pregibon, D. (1996). «A statistical perspective on Knowledge Discovery in Databases». *Advances in Knowledge Discovery and Data Mining*, 83-116. AAAI Press.

Fayad, U. M., Piatetsky-Shapiro, G. & Smuth, P. (1996). «From Data Mining to Knowledge Discovery: an overview». *Advances in Knowledge Discovery and Data Mining*, 1-36. AAAI Press.

Fisher, R. A. (1925). *Statistical Methods, Experimental Design and Scientific Inference.* Oxford Science Publications.

Friedman, J. H. (1991). «Multiple Adaptative Regression Splines». *Annals of Statistics* 19, 1-141.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison Wesley.

Greenacre, M. (1984). *Theory and Application of Correspondence Analysis.* Academic Press.

Gueguen, A. & Nakache, J. P. (1988). «Méthode de discrimination basée sur la construction d'un arbre de décision binaire». *Revue de Statistique Appliquée*, XXXVI (1), 19-38.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules.* J. Wiley.

Hand, D. J. (1998). «Data Mining: Statistics and more?». *The American Statistician*, 52, 2, 112-118.

Hand, D., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining.* The MIT Press.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer.

Kass, G. V. (1980). «An Exploratory Technique for Investigating Large Quantities of Categorical Data». *Applied Statistics*, 29, 2, 119-127.

Lebart, L., Morineau, A. & Piron, M. (1995). *Statistique exploratoire multidimensionnelle*, Dunod, Paris.

Lebart, L., Salem, A. & Berry, E. (1998). *Exploring Textual Data*, Kluwer, Boston.

Lebart, L. (1998). «Correspondence Analysis, Discrimination and Neural Networks». *Data Science, Classification and related methods*, ed. C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H-H. Bock and Y. Baba. Springer.

Lefébure, R. & Venturi, G. (1998). *Le data mining*. Eyrolles.

McCullagh, P. & Nelder, J. A. (1986). *Generalized Linear Models*. Chapman and Hall.

Mena, J. (1999). *Data Mining your website*. Digital Press.

Mola, F. & Siciliano, R. (1992). «A two-stage predictive splitting algorithm in binary segmentation». *Computational Statistics*, 1. Y. Dodge and J. Whittaker ed. Physica Verlag.

Murthy, S. K. (1998). «Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey». *Data Mining and Knowledge Discovery*, 2, 345-389.

Quinlan, J. (1988). *C4.5: Programs for machine learning*. Morgan Kaufman.

Rao, C. R. (1989). *Statistics and Truth*. CSIR, New Delhi.

Ripley, B. D. (1996). *Neural Networks and pattern recognition*. Wiley, New York.

Sarle, W. S. (1994). «Neural Networks and Statistical Models». *Proc. 9th. Annual SAS Users Group International Conference*. SAS Institute.

Sonquist, J. A. & Morgan, J. N. (1964). *The Detection of Interaction Effects*. Ann Arbor: Institute for Social Research. University of Michigan.

ENGLISH SUMMARY

DATA MINING, BETWEEN STATISTICS AND ARTIFICIAL INTELLIGENCE

TOMÀS ALUJA
Universitat Politècnica de Catalunya*

In the last decade a new concept had raised in the entrepreneurial side: data mining. Some companies have created data mining units directly linked to the CRM direction and in the professional forums data mining sessions have gained appeal. Data mining has appeared as a new discipline linked to Machine Learning, Artificial Intelligence and Data Bases, clearly differentiated from Statistics. On the other side, on the well-established statistics academia, data mining has been seen as the last fashion of a bad-known trend of data fishing and data dredging.

Is it really so? In this paper we will focus on the statistical roots of data mining, we will try to make and overview of the actual scope of data mining, we will present an application in the TV audience measure and we give some insights for the next future.

Keywords: Data mining, data analysis, modelling, artificial intelligence, KDD, neural networks, decision trees

AMS Classification (MSC 2000): 62-07, 68T10, 62P30

* Department of Statistics and Operations Research. Technological University of Catalonia (UPC). E-mail: tomas.aluja@upc.es

– Received April 2001.
– Accepted November 2001.

Statistics had risen in the beginning of XX century as a response to problems of society. Problems like defining a optimum fertiliser, the optimal conditions of production in an industry or assessing the effect of a drug, etc. Innovation always occurred due to stated problems. Anyway, we shall agree that statistics has been manipulating data for a most part of XX century without having a real computer device. Also a certain style of statistics installed in academia favoured a theoretical concept of the discipline. Nowadays, development in hardware have contributed to new and interesting sorts of data to analyse, which statistics should face. One of this problem is to come into knowledge the information hidden in the stored data by the information systems put in work for companies in the last two decades, coming up what is called the field of «data mining». It is no question to analyse small data files, but gigas or terabytes of data, with a precise goal, to take a managerial decision. This caused the appearance of data mining units in firms and its increasing interest in scientific meetings which devoted sessions to data mining. Anyway, data mining appeared linked to Artificial Intelligence, mainly machine learning, disciplines. Whereas it was considered by statisticians as a new version of the bad-known «data fishing» or «data dredging». It is really so?. I will establish that data mining stems from an old statistics tradition. In fact statistics lump together what can be called «data analysis» and «inferential statistics», the first being the first phase for the second (Cox, 1982, Rao, 1989). The difference between both approaches was wisely stated by Benzecry in his courses of 1964, «data is first, then follows the model», whereas for the «inferential» approach it is just the opposite. This is a sound difference, but it is clear that new problems arise, from the retail bar-code readers, transactions with a banking card, calls from a mobile telephone, or from the genome project, or satellite data, etc. This data very often is, as D. Hand (1998) pointed out, very large (huge), secondary, messy, with many missings, noisy and not representative. But it is absolute clear that statistics could play a central role for handling their associated uncertainty. These problems constitute a challenge for statisticians, pushing them to think again statistics, in particular the central issue of the tests of significance, and also to establish bridges of co-operation with our competitors of Artificial Intelligence, taking advantage of the strongness of both disciplines. Scientific disciplines, splitted in locked knowledge areas, have been developing isolated ones from others, leading to apparent different disciplines for the same problems. Here we show this applied to the case of Statistics and Artificial Intelligence, following the L. Lebart theory of two languages.
Finally we present a data mining application to the problem of finding good targets for TV program audience using decision trees. Decision trees, although they are not within the best classifier performers, like neural networks, generalised linear models, support vector machines, etc. have the appeal of being directly actionable, which in practice can overcome its shortcoming. Also for not very complex problems there is very little difference among different methods, making trees very useful for managerial applications. We follow the previously presented methodology (Aluja *et al*, 1998*b*) of building stable trees taking into account the individual contribution to impurity whithin the CART framework of tree building (Breiman *et al*, 1984).

Finally a point of humility from lord William Beveridge (1940) in this starving for knowledge:

> «Nobody believes a theory, except the one that has formulated it.
> Everybody believes a figure, except the one who has calculated it».

QÜESTIIÓ, vol. 25, 3, p. 499-508, 2001

APLICACIONES EMPRESARIALES DE DATA MINING

LLUÍS GARRIDO
JOSÉ IGNACIO LATORRE
Universitat de Barcelona*

El Data Mining, o ***extracción de información útil y no evidente de grandes bases de datos****, es una tecnología con un gran potencial para ayudar a las empresas a focalizar sus esfuerzos alrededor de la información importante contenida en sus «data warehouses».*

En este artículo analizaremos las ideas básicas que sustentan el Data Mining y, más concretamente, la utilización de redes neuronales como herramienta estadística avanzada. Presentaremos también dos ejemplos reales de la aplicación de estas técnicas: predicciones bursátiles y predicción de propagación de fuego en cables eléctricos.

Data Mining business applications

Palabras clave: Minería de datos, data mining, redes neuronales

Clasificación AMS (MSC 2000): 62-07, 62-09, 62H30, 68T05

* Departament d'Estructura i Constituents de la Matèria. Universitat de Barcelona. garrido@ecm.ub.es, latorre@ecm.ub.es.

– Recibido en abril de 2001.
– Aceptado en noviembre de 2001.

1. ¿QUÉ ES EL DATA MINING?

La mayoría de compañías tienen una gran cantidad de datos almacenados en sus ordenadores. Estos datos contienen una información que puede ser de gran utilidad para los resultados de la empresa. La gran abundancia de datos o su deficiente estructura puede hacer muy difícil extraer esta información útil. El objetivo del Data Mining [1] es la extracción de forma automática de información relevante, útil y no evidente contenida en dichos datos. Existen tres razones fundamentales por las cuales el Data Mining es una realidad en nuestros días:

- Avances tecnológicos en almacenamiento masivo de datos y CPU.
- Existencia de nuevos algoritmos para extraer información en forma eficiente.
- Existencia de herramientas automáticas que no hacen necesario el ser un experto en estadística, redes neuronales, o algoritmos matemáticos para convertirse en un «DataMiner».

2. ¿QUÉ PUEDE HACER EL DATA MINING?

Una empresa en posesión de unas bases de datos de calidad y tamaño suficiente puede emplear el Data Mining para generar nuevas oportunidades de negocio, dada su capacidad para proporcionar:

- **Predicción automática de comportamientos.**

  Generalmente se trata de problemas de clasificación. como ejemplo podemos citar el marketing dirigido. Data Mining usa los resultados de campañas de marketing realizadas anteriormente para identificar el perfil de los clientes que son más propensos a comprar el producto y de este modo permitirnos substituir el correo masivo por el correo dirigido.

- **Predicción automática de tendencias.**

  Basándonos en base de datos históricas, Data Mining creará un modelo para predecir las tendencias. Como ejemplos podemos citar la predicción de ventas en el futuro o la predicción en mercados de capitales.

- **Descubrimiento automático de comportamientos desconocidos anteriormente.**

  Las herramientas de Data Mining de visualización y clustering, permiten «ver» nuestros datos desde una perspectiva distinta y por ello descubrir nuevas relaciones entre ellos.

3. ¿CÓMO HACER DE DATA MINING? TÉCNICAS

Todo proyecto de Data Mining se desarrolla aplicando ciertas técnicas de especial interés en este campo. Las técnicas más utilizadas son:

- **Redes Neuronales.** Son modelos no-lineales inspirados en las redes de neuronas biológicas y se usan generalmente en problemas de clasificación y predicción. Discutiremos su estructura con un poco más de detalle en los ejemplos.
- **Árboles de decisión.** Son estructuras en forma de árbol que representan conjuntos de decisiones capaces de generar reglas para la clasificación de los datos.
- **Algoritmos genéticos.** Son modelos inspirados en la evolución de las especies y que se aplican generalmente en problemas de optimización. Permiten incluir fácilmente ligaduras complicadas que limitan la solución a un problema.
- **Clustering.** Métodos de agrupación de datos que nos permiten clasificar los datos por su similitud entre ellos. Son utilizadas con frecuencia para entender los grupos naturales de clientes en empresas o bancos.

4. METODOLOGÍA DE DATA MINING

Todo proyecto de Data Mining tiene unas fases bien definidas que van desde la definición del problema hasta la ejecución y evaluación del modelo, pasando por el estudio de los datos y la creación de dicho modelo. Dichas fases quedarán ilustradas en los dos ejemplos que se darán a continuación.

5. REDES NEURONALES

Una red neuronal artificial (ver por ejemplo [2, 3] para una introducción) es un modelo de computación inspirado en nuestros conocimientos sobre neurociencia, es decir, el estudio de las neuronas de nuestro sistema nervioso, aunque sin tratar de ser biológicamente realistas en detalle. En los últimos años estos modelos han experimentado un gran desarrollo gracias al descubrimiento de su excelente comportamiento en problemas de reconocimiento de patrones, predicción y clasificación, entre otros. Las redes neuronales artificiales son mecanismos matemáticos que aprenden a reconocer o clasificar patrones y, tal como lo hace nuestro propio cerebro, dicho aprendizaje no descansa sobre un modelo preconcebido sino que busca las correlaciones existentes entre las variables del problema que se está estudiando.

Para los ejemplos que seguirá hemos escogido redes neuronales multicapa entrenadas mediante el algoritmo de la back-propagation [4, 5]. Una arquitectura típica de di-

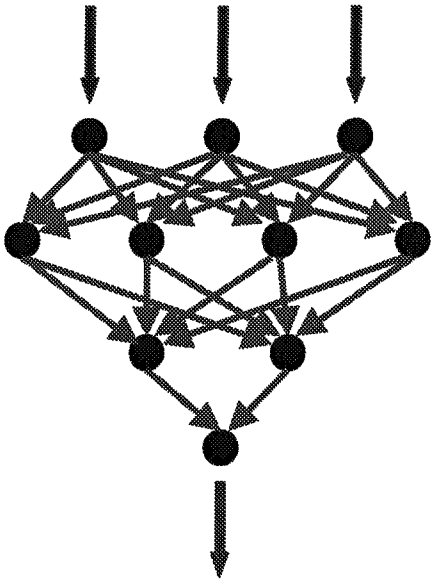


Figura 1. Esquema de una red neuronal multicapa sencilla.

chas redes es similar a la representada en la figura 1. Las neuronas (o unidades) están agrupadas en capas. Existe una primera capa por la que entra la información, una o varias capas ocultas que la procesan, y una capa de salida que proporciona los resultados de la red. Durante la fase de entrenamiento se le presenta a la red un conjunto de posibles valores de entrada juntamente con sus resultados de salida deseados, y el algoritmo de back-propagation busca automáticamente las correlaciones entre dichas entradas y sus salidas respectivas. Cuando el aprendizaje ha finalizado, la red puede ser aplicada sobre datos que no ha visto previamente, realizando sus propias predicciones.

En las redes neuronales multicapa se asocia un peso $W_{ij}^{(l)}$ a la conexión (sinapsis) existente entre la unidad j de la capa $l-1$ y la unidad i de la capa l. La salida de cada unidad $I_i^{(l)}$ se evalua utilizando la expresión

$$I_i^{(l)} = F\left(\sum_j W_{ij}^{(l)} I_j^{(l-1)} + B_i^{(l)}\right),$$

donde $B_i^{(l)}$ es el llamado umbral de la unidad i de la capa l. La función F se conoce como función de activación, y habitualmente se escoge como la función identidad ($F(x) = x$) o como una función sigmoidal ($F(x) = 1/(1+e^{-x})$). En este último caso es conveniente escalar las entradas y las salidas deseadas entre los valores 0 y 1.

Para entrenar una red neuronal con L capas se necesita un conjunto de datos de entrenamiento $\left\{(\mathbf{I}^{(L)\mu}, \mathbf{O}^{(L)\mu}), \mu = 1, \ldots, N\right\}$, donde $\mathbf{O}^{(L)\mu}$ representa la salida deseada correspondiente a la entrada $\mathbf{I}^{(0)\mu}$. El error de la red para estos datos es

$$E = \frac{1}{2}\sum_{\mu=1}^{N}\sum_{i}\left(I_i^{(L)\mu} - O_i^{(L)\mu}\right)^2,$$

donde $\mathbf{I}^{(L)\mu}$ representa la salida de la red después de presentarle una entrada $\mathbf{I}^{(0)\mu}$. La back-propagation intenta minimizar este error modificando la intensidad de los pesos y de los umbrales de forma iterativa, en un proceso conocido como de entrenamiento de la red neuronal. Existen otros métodos de entrenamiento, pero la back-propagation es, con diferencia, el más utilizado. De hecho, la back-propagation es un caso particular del bien conocido algoritmo de descenso por el gradiente, que está basado en una modificación de los pesos según la regla

$$W_{ij}^{(l)} = -\alpha\frac{\partial E}{\partial W_{ij}^{(l)}}$$

Es importante tener presente que la aplicación de las redes neuronales no se reduce únicamente a su entrenamiento con los datos disponibles: uno de los principales problemas es escoger previamente las variables más relevantes, ya que un exceso de variables puede introducir ruido que oculte las más significativas, mientras que su defecto puede provocar una falta de información.

6. PREDICCIONES BURSÁTILES

La aplicación de redes neuronales a la predicción de series temporales ha atraído la atención de mucha gente relacionada con los mercados financieros de todo el mundo. Sin embargo, el hecho de que la evolución de los mercados dependa de multitud de variables, muchas de las cuales son difíciles de cuantificar, hace que sea complicado encontrar situaciones en las que únicamente un análisis numérico de los datos históricos permita realizar buenas predicciones.

Como aplicación práctica de este tipo de redes neuronales, consideraremos el análisis de una acción en la bolsa noruega. Seguiremos paso a paso la metodología básica de la aplicación de esta técnica.

- *Definición del problema*

  Deseamos conocer la evolución de la acción PGS a dos días. El primer paso requiere decidir qué variables serán alimentadas a la red neuronal. Hemos optado por considerar como variables de entrada: la propia serie PGS, el índice Nasdaq, la cotización de

la misma empresa en el mercado americano (PGSUS), el índice de la bolsa noruega y los tipos de interés en Estados Unidos.

- *Preprocesamiento de las variables*

  La red neuronal podrá extraer información útil si las variables que la alimentan están correctamente preprocesadas. Hemos decidido considerar transformaciones de las variables del tipo considerado en el chartismo (medias móviles, ROC, ...) y en otras técnicas de análisis de datos (Gumbel,...)

- *Entrenamiento de las redes*

  Las redes neuronales son entrenadas evitando sobreentrenamiento y falta de generalización. Es conveniente considerar un entrenamiento que corresponda a mostrar cada patrón unos pocos miles de veces a la red neuronal. El entrenamiento breve no logra captar la ley que subyace al mercado. Un entrenamiento demasiado largo hace que la red aprenda los errores (no la ley general) de cada patrón. Una forma de evitar sobreentrenamiento es construir modelos alternativos con redes neuronales pequeñas. Una red pequeña no tiene grados de libertad suficientes para llegar a estar sobreentrenada.

- *Evaluación del modelo*

  Una vez que nuestra red neuronal es entrenada disponemos de un modelo de predicción. Podemos a continuación rastrear el éxito o fracaso de este modelo en los últimos meses. Es importante lograr que nuestro modelo además de ser rentable sea robusto frente pequeños cambios de variables o de estructura de la red neuronal. El modelo debe también describir con igual éxito tanto las subidas como las bajadas del mercado. Por último, el modelo debe proporcionar resultados similares a lo largo del tiempo. Si cualquiera de estos requisitos no se cumple podemos fácilmente haber construido un modelo que tendrá pobres resultados.

- *Definición de estrategia de actuación*

  El modelo neuronal debe considerarse como una herramienta de predicción que se complementa con otros métodos alternativos. Es muy posible que no todos los datos del mercado queden reflejados en las pocas series que alimentan a la red. Una correcta estrategia de actuación deberá contar con los condicionantes de la agresividad del inversor, de su liquidez, etc.

Los resultados obtenidos por las redes neuronales entrenadas siguiendo los pasos anteriores ofrecen resultados muy satisfactorios. Sobre un periodo de seis meses se logran beneficios de un 20%, ignorando comisiones de compra. Los beneficios se obtienen tanto en operaciones de compra como de venta (en el mercado de futuros).

Este ejemplo bursátil se ha implementado con la herramienta Stocknn (http://www.aernsoft.com) desarrollada conjuntamente por las empresas AERN y CAP GEMINI ERNST & YOUNG para la predicción de mercados de capitales con redes neuronales.

7. DATA MINING Y PROPAGACIÓN DE FUEGO

La aplicación de redes neuronales como herramienta de Data Mining tiene un vasto campo de aplicación. Presentamos un segundo ejemplo del empleo de redes neuronales para la predicción de la propagación de fuego en cables eléctricos. Como en el caso anterior, deseamos describir los pasos en que se organiza el análisis de este problema.

- *Definición del problema*

  Una severa normativa regula la construcción de cables eléctrico de forma que la propagación de un eventual fuego sea lo más reducida posible. El problema consiste en predecir la propagación de fuego en diferentes cables de nueva construcción a partir de los datos de sus componentes. Las variables utilizadas para alimentar la red neuronal serán, pues, las características geométricas del cable y las propiedades químicas de sus componentes.

- *Selección de variables*

  En este problema el número de variables es muy elevado. Cada material tiene muchas propiedades químicas. Se hace necesario realizar una selección de las variables más interesantes. Para ello hemos construido grandes redes neuronales que, una vez entrenadas, permiten establecer un criterio cuantitativo para evaluar la relevancia de cada variable. Un buen criterio es la suma de los valores absolutos de los pesos conectados a una cierta neurona i, p_i:

  $$p_i = \sum_j |W_{ij}^{(1)}|$$

  A mayor sea este peso, mayor será la relevancia de esta variable.

- *Entrenamiento de la red neuronal*

  El entrenamiento de una esta red neuronal sigue los pasos del ejemplo anterior. Debe evitarse el sobreentramiento y la falta de entrenamiento.

- *Construcción de un modelo alternativo*

  Para poder evaluar la bondad del modelo neuronal es preciso disponer de un modelo alternativo. En este caso, hemos optado por construir una regresión mulitlineal a las mismas variables que definen el modelo neuronal. La red debe captar las correlaciones no lineales y, por lo tanto, ser superior al modelo lineal.

- *Evaluación de los modelos*

  Para evaluar los modelos se procede a dividir la base de datos en tres partes: datos de entrenamiento, datos de validación y datos de predicción. La red neuronal y el modelo lineal son entrenados empleando los datos de entrenamiento. Luego se aplican ambos modelos en los datos de validación. Así se puede evaluar si ambos modelos funcionan con corrección y en qué medida la red neuronal es mejor que el modelo lineal. Una vez tenemos una correcta evaluación del modelo, podemos aplicarlo sobre los datos de predicción.

- *Clustering*

  Los cables forman grupos naturales con respecto a su comportamiento frente a la propagación del fuego. Hemos empleado la técnica de Análisis de Componentes Principales (PCA) que permite construir una proyección en dos dimensiones del espacio de variables que definen los cables. Cables próximos en cuanto a su construcción son próximos en su proyección.

Las redes neuronales entrenadas para la propagación de fuego logran correlaciones del orden de .8 entre predicción y realidad sobre el conjunto de validación. El porcentaje de acierto en el criterio de paso de normativa de fuego se halla por encima del 80% de los cables.

Este ejemplo de Data Mining en propagación de fuego se ha implementado con la herramienta QnnM (http://quantium.ecm.ub.es) el grupo Quantium de aplicaciones empresariales de redes neuronales.

8. REFERENCIAS

Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J. & Zanasi, A. (1997). *Discovering Data Mining from Concept to Implementation*, Book & Cd edition, (September 1997). [1]

Hertz, J. A.; Krogh, A. & Palmer, R. G. (1991). *Introduction to the theory of neural computation*, Addison-Wesley, Redwood City, California. [2]

Müller, B. & Reinhardt, J. (1991). *Neural networks: an introduction*, Springer-Verlag, Berlin. [3]

Rumelhart, D. E.; Hinton, G. E. & Williams, R. J. (1986). «Learning representations by back-propagating errors», *Nature*, 323, 533. [4]

Werbos, P. (1974). *Beyond regression: new tools for prediction and analysis in the behavioral sciences*, Ph. D. thesis, Harvard University. [5]

ENGLISH SUMMARY

DATA MINING BUSINESS APPLICATIONS

LLUÍS GARRIDO
JOSÉ IGNACIO LATORRE
Universitat de Barcelona*

Data Mining, can be defined as the ***art of extracting non-obvious, useful information from large databases.*** *This emerging field brings a set of powerful techniques which are of relevance for companies to focus their efforts in taking advantage of their data warehouses.*

In this paper we analyse the basic ideas behind Data Mining and we pay special attention to the use of neural networks as an advanced statistical tool. We also present two real world applications of these techniques to stock market predictions and to the study of fire propagation in cables.

Keywords: Data Mining, Neural Networks

AMS Classification (MSC 2000): 62-07, 62-09, 62H30, 68T05

* Departament d'Estructura i Constituents de la Matèria. Universitat de Barcelona. garrido@ecm.ub.es, latorre@ecm.ub.es.

- Received April 2001.
- Accepted November 2001.

A company may keep large amounts of high quality information which can be analysed using Data Mining techniques to uncover new business opportunities, as for instance

- **Targeted Marketing**

  This is a typical classification problem. Data Mining analyses results from previous marketing efforts to identify customer profiles and their predisposition to buy new products. These techniques are a profitable substitute for massive spamming.

- **Characterization and clustering**

  Data can be analysed to visualize and cluster them. Market segmentation and isolation of relevant categories of customers are better characterize using non-linear data mining techniques.

Data Mining projects are developed using a number of advanced statistical techniques. We here mention a few of them:

- **Neural Networks**

  They provide non-linear models inspired in biological neural networks and they are used in problems of classification and prediction. In the paper we discussed practical examples of how to use them.

- **Decision trees**

  These are tree structures representing sets of decisions that can generate rules for data classification.

- **Genetic algorithms**

  These models are inspired in the evolution of species and are applied to optimisation problems. They allow for easily introducing constraints which limit the solution to a problem.

- **Clustering**

  These methods allow for a classification of data given similarities in their patterns. They are used to understand natural groups of customers in large companies and banks.

From the practitioner point of view Data Mining follows well defined phases. The project must be defined, data must be cleaned, several alternative models must be built and a strict validation must check the goodness of the final result. These examples and techniques are discussed in more detail in the paper.

QÜESTIIÓ, vol. 25, 3, p. 509-520, 2001

PRACTICAL DATA MINING IN A LARGE UTILITY COMPANY*

GEORGES HÉBRAIL*

We present in this paper the main applications of data mining techniques at Electricité de France, the French national electric power company. This includes electric load curve analysis and prediction of customer characteristics. Closely related with data mining techniques are data warehouse management problems: we show that statistical methods can be used to help to manage data consistency and to provide accurate reports even when missing data are present.

Keywords: Data mining, statistical methods, data warehouses, applications, load curve analysis, utility company

AMS Classification (MSC 2000): 62-07, 62P30, 62M45, 62M10, 62H30

* This is a reprinted version of the article published in Compstat 2000 Proceedings (ISBN 3-7908-1326-5).

* Electricite de France, R&D Division. 1, Av. du Général de Gaulle. 92141 Clamart, France.

- Received May 2001.
- Accepted November 2001.

1. INTRODUCTION AND SHORT PRESENTATION OF EDF

Electricité de France (EDF) is the French national electric power company which is in charge of the generation, transmission and distribution of electric power in France. In 1999, the electric power market was deregulated in France so that now EDF does not have anymore the monopoly on the market of large customers. In parallel with the deregulation, EDF develops its activity in other countries in Europe and in the world. EDF is a large company, gathering nearly 115 000 employees with around 30 million customers in France.

Acting in a deregulated environment is quite different from the previous situation where EDF had a monopolistic position on the French electric power market. The company currently develop new tools based on data mining techniques to be able to face this competition. Actually, EDF developed in the past many tools in order to optimize its activity and reduce the cost of generation, transmission and distribution of electric power: many of these tools designed to solve technical problems are adapted to solve commercial problems.

In this paper, we present the main developments of data mining techniques at EDF in the area of customer relationship management. Data mining techniques are also used in other domains, like power plant maintenance, electrical network operation, human resource management, but are not described here. In the paper, we will not make a special distinction between the terms «*data mining*» and «*statistical*» methods since in our opinion most of efficient data mining techniques are based on well-known statistical methods, the main difference being related to their scalability to large volumes of data.

In Section 2, we present the development of data mining techniques for dealing with electric power load curves of individual customers, i.e. curves representing their electric power consumption. In particular, we present a software developed in our Division which performs clustering of such curves interactively: this software is based on the Self Organizing Map (SOM) approach (see Kohonen (1990)). In this section, we also present a new approach we are studying to analyze long time series associated with the consumption of one customer: pattern extraction techniques are applied to build a symbolic description of the time series.

Section 3 is devoted to the presentation of data mining techniques applied to predict missing data in customer databases. Some fields in customer databases are partially filled, like for instance the possible use of electric power for water heating. Logistic regression methods are applied to predict missing information, from customers with non missing data.

In Section 4, applications related to interactions between statistical methods and data warehouses are described. We show that standard statistical methods can be used for two very useful goals: detection and characterization of records which fail consistency

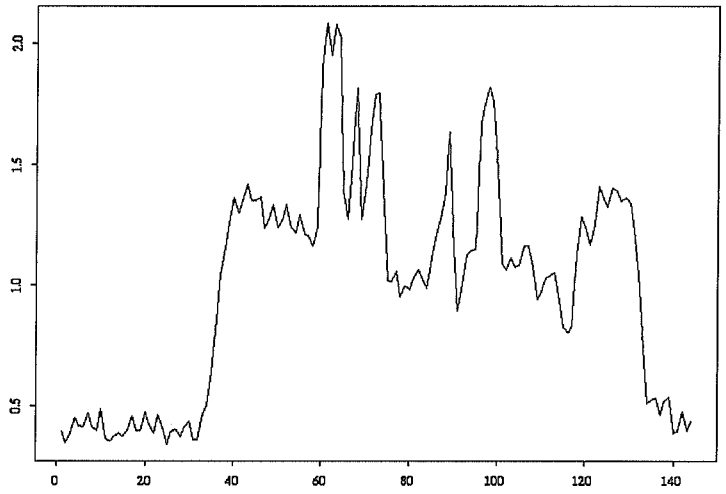


Figure 1. Example of a daily load curve.

checking of the data warehouse, and production of accurate reports even when data is missing, by statistical adjustment.

2. ELECTRIC LOAD CURVE ANALYSIS

2.1. Typical available data

Electric power sales for a customer are represented by a time series (also called a *load curve*) figuring the electric power consumption for every period of time. Availability of such data deeply depends on the type of customer. General public small customers (like residential ones) are poorly described since a communicating meter is too expensive regarding to their consumption: for these customers there are only a few points of the curve every year.

For larger customers, a communicating meter is often available for many reasons: the billing is done every month, the consumption is high and justifies the communicating meter investment, a detailed record of consumption is necessary because prices depend on the period. For these customers, a load curve is available figuring points every 10 minutes. Figure 1 shows an example of such a curve for a period of one day. The curve is generally available all over the year. In the rest of this section, we describe applications when such data is available.

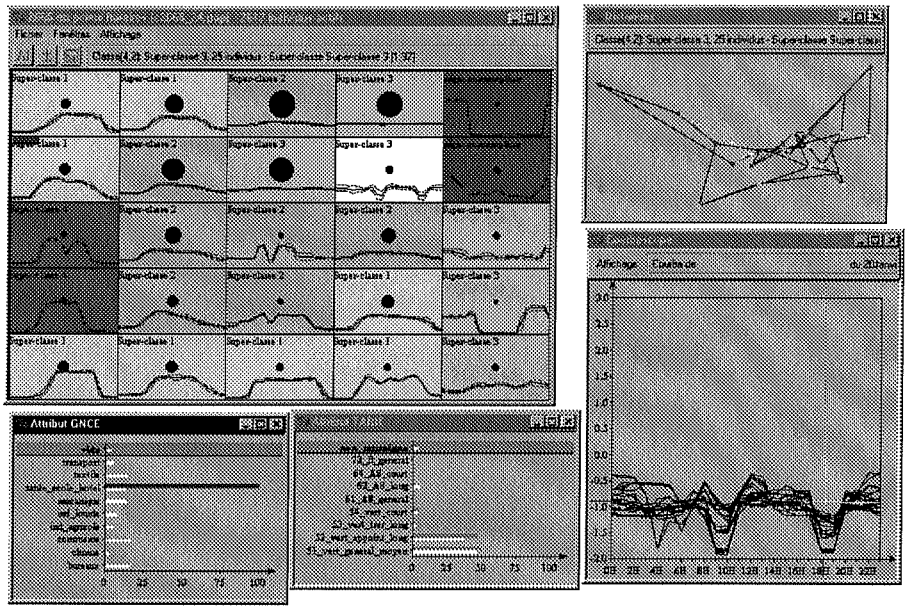


Figure 2. Main screen of the Courboscope software.

2.2. Goals of load curve analysis

Load curves have been analyzed at EDF for more than 20 years, with different goals of study. Analyzed curves may correspond to individual customer curves or to aggregates over a geographical sector. Applications may be either understanding the behavior of customers in relation to some external characteristics, or predicting consumption in a short or long term perspective.

In this paper, we focus on the exploratory data analysis of a set of load curves, each curve being associated with a single customer. The goal is to understand customer consumption behavior all along the year, in relation to his usage of electric power, to the pricing policy, and to the characteristics of the customer (mainly its activity).

2.3. The «Courboscope» software

The «Courboscope» (*courbe* is the French word for curve) is a software which has been developed a few years ago in our R&D Division. It is described in detail in Debrégeas and Hébrail (1998). This software is a tool for clustering sets of curves.

Curves are here assumed to be described by the same number of points (for instance 24 points for a daily hourly measured curve). Each curve is thus considered as a point in a p-dimensional euclidian space (3^{24} for daily curves). The euclidian distance is used as a measure of dissimilarity between curves. If this distance is not accurate, a preliminary process normalizes the curves so that the euclidian distance becomes appropriate.

Curves are also described by additional attributes, such the name of the customer, his activity, the day in the week, the month, the season, the pricing option, ...

The user's task associated with this software is to build clusters of curves, where curves belonging to the same clusters have similar shapes. The clustering method is the SOM (Self-Organized Map) method proposed by Kohonen, see Kohonen (1990). This method, based on a neural network approach, produces clusters organized in a spatial way so that clusters which are close in the map have similar global shapes (see the upper left window in Figure 2). This special arrangement of clusters is quite interesting because it allows to perform clustering with a large number of clusters (for instance $10 \times 10 = 100$ clusters): the result is still readable.

Beyond construction of clusters, the software helps the user to give an interpretation to each cluster. A sophisticated and easy to use interface enables the user both to examine details of each cluster (lower right window in Figure 2), and to characterize each cluster with external attributes (lower left windows).

This software is used by load curve analysts who bring out typical shapes of curves (usually on a basis of daily or weekly curves), associated with some external information, typically a tariff, an activity, or an electric power usage. The result of this process is what we call an *interpreted map*, where a label is associated with each cluster cell.

Interpreted maps can be used by other (operational) people through another module of the software which is designed to classify new curves into the interpreted map. Main applications here are default detection and pricing advice. Two kinds of expertise are thus mixed together: the analyst expertise by the means of the interpreted map, and the expertise of people on the field through this classification module, which enables them to check if the use of the euclidian distance does not get completely crazy.

The Courboscope software is not dedicated to electric power load curves but is generic. It can be used with any type of curves described by any additional attributes.

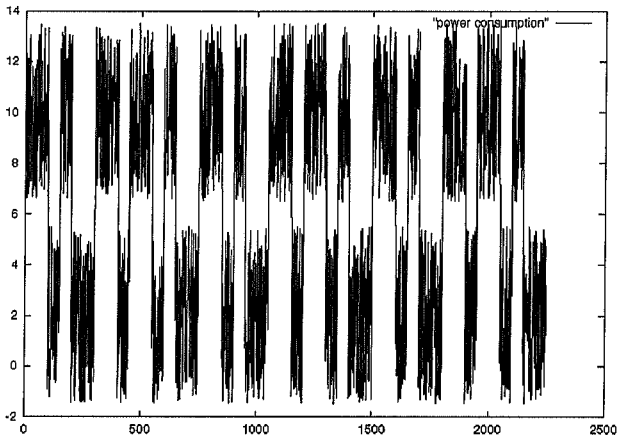
2.4. Towards a symbolic analysis of curves

The approach described in the previous section shows several restrictions:

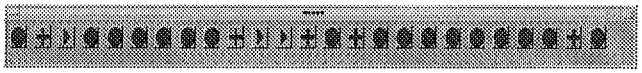
- all curves have to be of the same length and described by points at the same time positions,

Initial one year long curve

(only 2300 of the 8760 are represented below)



Symbolic representation of the one year long curve



Typical curves associated with symbols

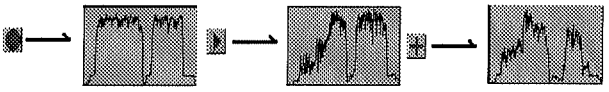


Figure 3. Example of a symbolic representation of curves.

- long curves (for instance time series describing one month or one year) cannot be treated because the euclidian distance in a 8000 dimension space is not meaningful,
- multi-dimensional curves (i.e. several distinct values available for every time tick) cannot be processed easily.

We are currently working on another approach where long curves (i.e. time series) are transformed into a symbolic form (see Hugueney and Hébrail (1999)). The basic process is the following:

- selection of a window width (manually or guided by a Fourier transform),
- generation of a set of fixed length curves obtained by sliding a fixed length window all along the long curve,
- clustering of the fixed length curves with the Courboscope, in order to obtain clusters of fixed length curves,

– transformation of the long curve into a succession of symbols, each symbol being associated with a cluster of fixed length curves.

Figure 3 shows an example where a curve of one year long (8760 hourly points) is replaced by a succession of 26 from a set of 3 different symbols. Here, each symbol represents a two week typical curve.

Once a long curve has been transformed into such a symbolic form, several methods can be used, in an exploratory perspective:

– visualization of the symbolic representation: this provides a very synthetic view of the long curve,
– application of data mining methods such as sequence analysis, in order to discover frequent subsequences (see Agrawal and Srikant (1995)),
– application of multivariate data analysis methods in the case of multi-dimensional curves.

Manual editing of the symbolic representation is also very interesting. Editing can be done at two different levels:

– edition of the sequence of symbols, the application here is to simulate a change in the customer's activity schedule,
– edition of the curves associated with symbols, the application here is to simulate a change in the customer's equipment.

Once curves have been edited at the symbolic level, a simple process can reconstitute a predicted curve which reflects changes done at the symbolic level. The real-world application here is to simulate changes in customer behavior in relation to pricing options.

3. PREDICTION OF CUSTOMER CHARACTERISTICS

In this section, we describe data mining applications which are not specific to electric power distribution. These applications are related to the marketing activity of electric power sales within general public customers.

There are two main ideas underlying these applications:

– information which is necessary for a good marketing activity cannot be obtained for all general public customers: the cost is prohibitive,
– customer databases contain enough data to infer missing information using statistical methods.

We only do here a short presentation of these applications since they are quite standard in the data mining field.

3.1. Brief overview of customer databases

EDF sells electric power to nearly 30 million general public customers in France. There are around 100 local customer databases, each database describing an average of 300 000 customers of the same geographical area. These databases have been designed primarily for billing purposes: they contain information about the contract, the consumption and the premises where power is delivered.

Some years ago, these databases have been enriched by additional fields to improve customer relationship and to allow marketing actions to be more efficient. This fields are filled out with the stream of contacts between the company and its customers. Consequently, they are currently partially filled.

In parallel, national surveys are carried out in order to acquire more information on a small number of customers. These surveys also contain a copy of information available in our databases for the corresponding customers (identity of each customer is lost for ethical reasons).

3.2. Prediction of customer characteristics

Data mining techniques are used to predict some customer characteristics when they are missing. The basic method we use is logistic regression.

Two different problems are addressed:

- application of logistic regression within a local database: partially filled fields define the prediction variables and the learning sample gathers customers with non missing values.
- application of logistic regression to predict values of fields available in surveys but not in customer databases. Values are predicted for all customers in each local database.

3.2.1. Prediction within local databases

As written before, local databases describe an average of 300 000 general public customers. Some fields describing them are partially filled because these fields have been introduced in the database only some years ago. Depending on the database, these fields are filled out for 10% to 50% of the customers. This is usually enough to build a prediction model from a sample gathering the customers with non missing values.

Several fields —like for instance the presence/absence of electric heating of water— have been predicted successfully with error rates from 10 to 30%, depending on the field and on the local database. This precision is enough to feed marketing campaigns.

It is important to note that one model is built for each local database, in order to take into account specific local characteristics of the populations. A special attention has to be given to the fact that in this case the sample is biased. A reweighting is necessary and is based on a stratification of the local database.

3.2.2. Projection of national survey data to local customer databases

The other application is prediction (within each local customer database) of fields which are available only in national surveys. Survey data contain both answers to questionnaires and customer data extracted from our databases: this enables to build prediction models applicable to local customer databases.

In this case, the sample is not made of customers from the local database on which the prediction is performed. Actually, customers of national surveys are picked up randomly from all local databases. So, there is only a very small number of customers from each local database in the survey. This number is not sufficient to build a model for each database. A model is built from the whole survey and applied to the different local databases.

We have done preliminary experiments which show that prediction is possible but with an error rate which is still too high for our needs. We are currently running more experiments on this problem to improve the method.

4. STATISTICS AND DATA WAREHOUSES

In this section, we examine the joint use of statistical techniques and data warehouse facilities. We are currently working in two directions:

- quality of data in *relational* data warehouses,
- estimation of missing data in *multi-dimensional* data warehouses.

All applications described in this section correspond to very basic statistical methods but are really useful. The only difficulty is that the volume of data may become very large: algorithms must be scalable.

4.1. Quality of data warehouses

In this section, we focus on data warehouses where the database model is the *relational* model, i.e. the database is a collection of standard data tables.

First, simple statistics are systematically computed on every table of the data warehouse in order to give a diagnosis of the quality of data. The following basic statistics appear to be very useful to detect errors: number of values, number of different values, and number of missing values for every column. Barcharts, boxplots and histograms on some specified variables (like the contract type or the power consumption) are also of good help, especially to detect outliers which often correspond to errors in the database.

Secondly, we carry out a project addressing the problem of detecting errors or inconsistencies in relational data warehouses. This project is based on a database point of view: a set of constraints which should be satisfied by the database is defined. A program generates for each table of the database the set of records which violate the constraints. Here, statistical methods (see Morineau (1984) for instance) are used to characterize error records against others, using all columns of the table. This helps the user to understand the reason of errors: for instance, it may reveal that records in error correspond to bills of October or to customers of some geographical area.

4.2. Estimation of missing data in data warehouses

In Section 3.2.1, we have seen that logistic regression is used to predict values for customers showing missing values on some partially filled fields. This is prediction of missing data at the detailed level. In this section, we address the problem of estimation of aggregated data from a database containing missing values at the detailed level.

Data warehouses now offer a *multi-dimensional* model of data in addition to the relational model. This new model allows to define data structures which are matrices with several dimensions. In real-world applications, usual dimensions are *Product*, *Store*, and *Time*. For each product, store, and date, data usually consist in sales defined by several numerical values such as the number of sold items, the amount of the transaction, ... For each dimension, hierarchies are defined in order to build syntheses of data, for instance: products are grouped into types of products, brands, or store departments; stores are grouped into geographical areas and regions; time is aggregated at the day, month, quarter, and year levels.

Once sales data have been collected from stores, the user can use the dimension hierarchies to build interactively any aggregated array of data: for instance the total sales of food by region and month for year 1999. The database system ensures that the response time is good even if the volume of data is huge (millions of transactions for a large corporate department store). This facility assumes that all data are available, i.e. there are no missing values in the detailed data.

Within our local databases, we are currently building such a multi-dimensional database centered on electric power sales. This database is intended to be used by marketing people in order to have an idea of the sales by geographical area, time period, and

type of customer. As mentioned in Section 3, there are many missing values in the customer characteristics which are not necessary for billing activity. This is the case for information about usages of electric power (heating, water-heating, type of appliances, ...). At the detailed level of data, prediction can be done with a logistic regression. At aggregated levels, prediction of aggregate values can also be done with statistical methods but the approach is different: it is related to sample adjustment to take into account the fact that non missing values constitute a biased sample.

We are currently working on the problem of adding statistical inference capabilities to multi-dimensional databases, in order to be able to query the database as if there were no missing values. An error has to be defined and evaluated for each query, depending on the level of aggregation of the query.

5. CONCLUSION

In this paper, we have presented several applications of data mining and statistical methods in the context of a large utility company.

Some of them are rather specific to electric power distribution, or by extension to utilities. The methods in this case are related to the analysis of load curves of large customers. The Courboscope software allows to analyze short period curves (day or week) to understand customer behavior. Symbolic curve analysis is intended to analyze longer curves (one year for instance): it is an active research area for us.

There are also other applications of statistical methods related to load curves, in order to predict the consumption demand in the future. These applications were not described in this paper and are based on ARIMA and neural network models (see for instance Mangeas and Muller (1997)).

Other applications presented in the paper are related to general public customers. Better knowledge of these customers is achieved by analysis of customer databases and national surveys. Here statistical methods allow prediction of customer characteristics either at a detailed or aggregated level. Work on projecting survey data on customer databases is an active research area for us.

6. REFERENCES

Agrawal, R. & Srikant, R. (1995). «Mining Sequential Patterns, *Proceedings of 11th Int'l Conf. on Data Engineering (DE'95)*, Taipei, Taiwan.

Debrégeas, A. & Hébrail, G. (1998). «Interactive Interpretation of Kohonen Maps Applied To Curves», in *KDD'98, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, New-York, 179-183, AAAI Press.

Hugueney, B. & Hébrail, G. (1999). «Construction de descriptions symboliques de courbes numériques». *EDF R&D Division internal report*, n° HI-23/99-025.

Kohonen, T. (1995). *Self-organizing maps*, Berlin: Springer.

Mangeas, M. & Muller, C. (1997). «An automatic search of feed forward neural network architecture based on genetic algorithms: application to the short term load forecasting». In *ISAP'97, Proceedings of the International Conference on Intelligent System Applications for Power Systems*, Corea.

Morineau, A. (1984). «Note sur la Caractérisation Statistique d'une Classe et les Valeurs-test». *Bulletin du CESIA*, vol. 2, n° 1-2, Paris.

Investigació Operativa

QÜESTIIÓ, vol. 25, 3, p. 523-549, 2001

SECUENCIACIÓN DINÁMICA DE SISTEMAS DE FABRICACIÓN FLEXIBLE MEDIANTE APRENDIZAJE AUTOMÁTICO: ANÁLISIS DE LOS PRINCIPALES SISTEMAS DE SECUENCIACIÓN EXISTENTES

PAOLO PRIORE*
DAVID DE LA FUENTE*
JAVIER PUENTE*
ALBERTO GÓMEZ*

Una forma habitual de secuenciar de modo dinámico los trabajos en los sistemas de fabricación es mediante el empleo de reglas de secuenciación. Sin embargo, el problema que presenta este método es que el comportamiento del sistema de fabricación dependerá de su estado, y no existe una regla que supere a las demás en todos los posibles estados que puede presentar el sistema de fabricación. Por lo tanto, sería interesante usar en cada momento, la regla más adecuada. Para lograr este objetivo, se pueden utilizar sistemas de secuenciación que emplean aprendizaje automático que permiten, analizando el comportamiento previo del sistema de fabricación (ejemplos de entrenamiento), obtener el conocimiento necesario para determinar la regla de secuenciación más apropiada en cada instante. En el presente trabajo se realiza una revisión de los principales sistemas de secuenciación, existentes en la literatura, que utilizan aprendizaje automático para variar de forma dinámica la regla de secuenciación empleada en cada momento.

Dynamic scheduling of flexible manufacturing systems through machine learning: an analysis of the main scheduling systems

Palabras clave: Secuenciación dinámica, aprendizaje automático, sistemas de fabricación flexible, reglas de secuenciación, simulación

Clasificación AMS (MSC 2000): 68T20, 68M20

* Dpto. de Administración de Empresas y Contabilidad. Escuela Técnica Superior de Ingenieros Industriales e Informáticos de Gijón. Universidad de Oviedo.

– Recibido en diciembre de 1999.
– Aceptado en julio de 2001.

1. INTRODUCCIÓN

La secuenciación de trabajos, que forma parte del proceso de control en un sistema de fabricación, es necesaria cuando un conjunto común de recursos debe ser compartido, para fabricar una serie de productos durante el mismo período de tiempo. El objetivo de la secuenciación es la asignación eficiente de máquinas y otros recursos a los trabajos, o a las operaciones contenidas en éstos, y la determinación del momento en el que cada uno de estos trabajos debe procesarse (Shaw *et al.*, 1992).

Los tiempos de procesamiento en los sistemas de fabricación flexible (*FMSs*[1]) son casi determinísticos puesto que las operaciones son controladas por ordenador y procesadas, en su mayor parte, por máquinas de control numérico; asimismo, las preparaciones entre operaciones consecutivas están automatizadas. Por lo tanto, si no hay perturbaciones en el *FMS*, se puede predecir el resultado y es suficiente un sistema de secuenciación fijo y «off-line».

Sin embargo, debido a la llegada de nuevos trabajos (en ocasiones urgentes), averías en las máquinas y otras perturbaciones, el estado de un *FMS* puede no ser predecible. Esta naturaleza incierta y dinámica sugiere que un sistema de secuenciación de tipo «off-line» no es el más adecuado. Asimismo, los *FMSs* son más sensibles a las perturbaciones que los sistemas de fabricación convencionales puesto que sus componentes están más sincronizados e integrados y son más interdependientes. Por lo tanto, se requiere una respuesta inmediata a los cambios en los estados del *FMS*, mediante la utilización de un sistema de secuenciación en tiempo real. Si los estados del *FMS* cambian de modo dinámico, la secuenciación de trabajos debe hacerse en función del estado actual de éste (Jeong y Kim, 1998).

El resto del artículo está organizado de la siguiente forma. En primer lugar, se define el aprendizaje automático y se describen los tipos principales de sistemas de secuenciación de *FMSs* existentes en la literatura. Posteriormente, para superar el problema que presenta las reglas de secuenciación cuando se emplean de forma estática, se definen dos métodos para modificar estas reglas de modo dinámico. Uno de ellos está basado en la utilización de un modelo de simulación, mientras que el otro emplea «conocimiento de secuenciación» del sistema de fabricación. A continuación, se realiza una revisión de los principales trabajos que utilizan este último método. Finalmente, se especifican una serie de carencias comunes de los sistemas de secuenciación basados en el conocimiento cuya resolución constituye una fuente de futuras líneas de investigación.

[1]Del inglés, Flexible Manufacturing System.

2. APRENDIZAJE AUTOMÁTICO

El aprendizaje automático, que pertenece al campo de la inteligencia artificial, permite resolver problemas mediante el empleo del conocimiento obtenido de problemas resueltos en el pasado similares al actual (Michalski *et al.*, 1983). Una representación esquemática de lo expuesto anteriormente se muestra en la figura 1. Los tipos principales de algoritmos dentro del aprendizaje automático son las redes neuronales (Bishop, 1995), el aprendizaje inductivo (Quinlan, 1993) y el razonamiento basado en casos (CBR[2]; Watson, 1997). La diferencia fundamental entre estos tipos de algoritmos radica en la forma en que se almacena el conocimiento. Así, en las redes neuronales, el conocimiento se traduce en una serie de pesos y umbrales que poseen las neuronas. En cambio, en el aprendizaje inductivo, el conocimiento se transforma en un árbol de decisión o un conjunto de reglas. Por último, en el CBR, el conocimiento está formado por una base de casos compuesta por los problemas resueltos en el pasado.

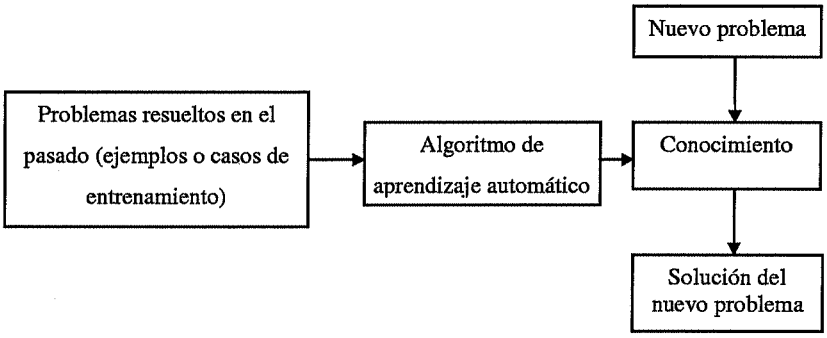


Figura 1. Esquema general del problema de aprendizaje.

Los casos o problemas resueltos en el pasado (también denominados ejemplos o casos de entrenamiento) se pueden representar por tablas atributo-valor, como la que se muestra en la tabla 1, donde hay un atributo especial que se denomina la clase (la solución del problema resuelto). Los atributos representan las características del problema. El objetivo que tiene un algoritmo de aprendizaje automático es tratar de aprender a clasificar nuevos casos, similares a los de entrenamiento, de los que se conocen los valores de todos los atributos excepto la clase. En esta situación, el término clasificar se utiliza en el sentido más literal; es decir, determinar cuál es la clase de un nuevo caso o ejemplo.

[2]Del inglés, Case-Based Reasoning.

Tabla 1. Tabla atributo-valor de los ejemplos de entrenamiento.

Ejemplos	*Atributo 1*	*Atributo 2*	.....	*Atributo m*	*Clase*
Ejemplo 1	A_{11}	A_{12}	.....	A_{1m}	C_1
Ejemplo 2	A_{21}	A_{22}	.....	A_{2m}	C_2
.....	.....	.....	.....	.....	.....
Ejemplo n	A_{n1}	A_{n2}	.....	A_{nm}	C_n

3. SISTEMAS DE SECUENCIACIÓN EN LOS FMSs

Los sistemas de secuenciación en los *FMSs* se pueden dividir en las siguientes categorías:

1. Sistemas basados en métodos analíticos.
2. Sistemas basados en métodos heurísticos.
3. Sistemas basados en simulación.
4. Sistemas basados en inteligencia artificial.

Los sistemas analíticos formulan el problema de secuenciación de un *FMS* como un modelo de optimización con restricciones, en términos de una función objetivo y restricciones explícitas. Posteriormente, se resuelve el modelo utilizando alguno de los algoritmos de resolución existentes (ver por ejemplo, Han *et al.*, 1989; Hutchison *et al.*, 1989; Kimemia y Gershwin, 1985; Lashkari *et al.*, 1987; Shanker y Rajamarthandan, 1989; Shanker y Tzen, 1985; Stecke, 1983; Wilson, 1989).

En general estos problemas son de tipo NP-completo (Garey y Johnson, 1979), por lo que normalmente se proponen algoritmos de tipo heurístico y «off-line» para su resolución (Chen y Yih, 1996; Cho y Wysk, 1993). Asimismo, estos modelos analíticos poseen simplificaciones que no siempre son válidas en la práctica (de hecho, Basnet y Mize (1994) afirman que algunos modelos son tan singulares que parece que los problemas son ideados para encajar en el modelo y no viceversa) y no son eficientes para problemas de tamaño razonable.

Las dificultades de la aplicación de los sistemas analíticos en los problemas de secuenciación, han propiciado la investigación de múltiples heurísticas. Éstas, suelen tomar la forma de reglas de secuenciación y habitualmente se emplean para secuenciar los trabajos en los sistemas de fabricación de modo dinámico. Estas heurísticas ordenan los diversos trabajos que compiten por el uso de una máquina dada, mediante diferentes esquemas de prioridad; así, a cada trabajo se le asigna un índice de prioridad y aquel que posea el menor índice se selecciona en primer lugar.

Hasta la fecha, muchos investigadores (ver por ejemplo, Baker, 1984; Blackstone *et al.*, 1982; Kim, 1990; Panwalkar y Iskander, 1977; Ramasesh, 1990; Russel *et al.*, 1987; Vepsalainen y Morton, 1987) han evaluado, mediante simulación, el comportamiento de los sistemas de fabricación con diferentes reglas de secuenciación, concluyendo que dicho comportamiento depende de múltiples factores como el criterio de eficiencia, la configuración del sistema de fabricación, la carga de trabajo, etc. (Cho y Wysk, 1993). Con la llegada de los *FMSs*, surgen numerosos estudios que analizan el comportamiento de estos sistemas con las reglas de secuenciación (ver por ejemplo, Choi y Malstrom, 1988; Denzler y Boe, 1987; Egbelu y Tanchoco, 1984; Henneke y Choi, 1990; Montazeri y Van Wassenhove, 1990; Stecke y Solberg, 1981; Tang *et al.*, 1993).

Debido al comportamiento variable de los sistemas de fabricación, sería interesante modificar las reglas de secuenciación dinámicamente, en el momento apropiado, dependiendo de las condiciones del sistema de fabricación. A priori, se espera que este método sea superior a utilizar una regla de secuenciación de forma constante, por dos razones. En primer lugar, porque es capaz de identificar la mejor regla para un escenario de fabricación determinado. Así, debido a esta capacidad de selección, el sistema de fabricación debe comportarse al menos tan bien como con la mejor de las reglas de secuenciación candidatas considerada. En segundo lugar, este método puede adaptar su selección de forma dinámica a los escenarios cambiantes. Esta adaptabilidad permite secuenciar los trabajos con una eficiencia incluso superior a la de la mejor regla de secuenciación (Shaw *et al.*, 1992).

Para modificar de forma dinámica las reglas existen, básicamente, dos tipos de sistemas de secuenciación en la literatura. En el primero, la regla se determina, en el momento apropiado, simulando un conjunto de reglas de secuenciación seleccionadas de antemano y eligiendo la mejor (ver por ejemplo, Ishii y Talavage, 1991; Jeong y Kim, 1998; Kim y Kim, 1994; Wu y Wysk, 1989). Los principales inconvenientes que presenta este sistema basado en simulación son los siguientes:

1. El tiempo necesario para realizar las simulaciones con el conjunto de reglas candidatas que puede dificultar la secuenciación en tiempo real.
2. Cambios en el sistema de fabricación muy frecuentes. Dado que la simulación con cada una de las reglas se hace hasta el final del período de simulación considerado, puede que no exista coincidencia entre la regla propuesta y la realmente necesaria ya que la regla elegida se utiliza durante un período de tiempo inferior al que se empleó durante la simulación.
3. No se dispone de mecanismos que eviten modificaciones innecesarias de las reglas de secuenciación ante cambios de tipo transitorio.
4. No se obtiene ningún tipo de conocimiento acerca del sistema de fabricación.

En el segundo tipo de sistema de secuenciación, perteneciente al campo de la inteligencia artificial, se emplea un conjunto de simulaciones previas del sistema de fabricación

(ejemplos de entrenamiento) para determinar cuál es la mejor de las reglas de secuenciación en cada posible estado del sistema de fabricación. Estos casos de entrenamiento se utilizan para entrenar un algoritmo de aprendizaje automático, con el objeto de obtener conocimiento acerca del sistema de fabricación. Finalmente, este conocimiento se utiliza para tomar decisiones inteligentes en tiempo real.

4. SISTEMAS DE SECUENCIACIÓN BASADOS EN EL CONOCIMIENTO

Para que un sistema de secuenciación en tiempo real, que modifique de forma dinámica las reglas de secuenciación, funcione adecuadamente debe cumplir dos características conflictivas:

1. La selección de reglas debe tener en cuenta una variedad de información, en tiempo real, acerca del sistema de fabricación.
2. La elección debe realizarse en un corto período de tiempo de modo que las operaciones no se retrasen.

Una forma de conseguir estas características es utilizar alguna clase de conocimiento acerca de las relaciones entre el estado del sistema de fabricación y la regla a emplear en ese momento. Por lo tanto, es útil emplear «conocimiento de secuenciación» del sistema de fabricación para ahorrar tiempo y alcanzar una respuesta rápida en un entorno que cambia dinámicamente (como es el caso de un *FMS*). Sin embargo, uno de los problemas más difíciles de resolver en un sistema de secuenciación basado en el conocimiento es la adquisición de éste. Para ello, se utilizan algoritmos de aprendizaje automático que reducen el esfuerzo en la determinación del conocimiento necesario para realizar las decisiones de secuenciación.

De todas formas, para que este conocimiento sea útil, es necesario que los ejemplos de entrenamiento y el propio algoritmo de aprendizaje sean los adecuados. Asimismo, para obtener los ejemplos de entrenamiento, son cruciales los atributos seleccionados (Chen y Yih, 1996). Sin embargo, existen, al menos, cuatro razones por las cuales un sistema de secuenciación basado en el conocimiento puede tener un comportamiento inferior a la mejor de las reglas utilizada de forma individual:

1. El conjunto de entrenamiento es un subconjunto del universo de todos los casos posibles. De todos modos, siempre se pueden observar los escenarios en los que el sistema de secuenciación no funciona de forma adecuada y añadirlos como ejemplos de entrenamiento.
2. El comportamiento del sistema de fabricación depende del número y rango de los atributos de control considerados para diseñar los ejemplos de entrenamiento.

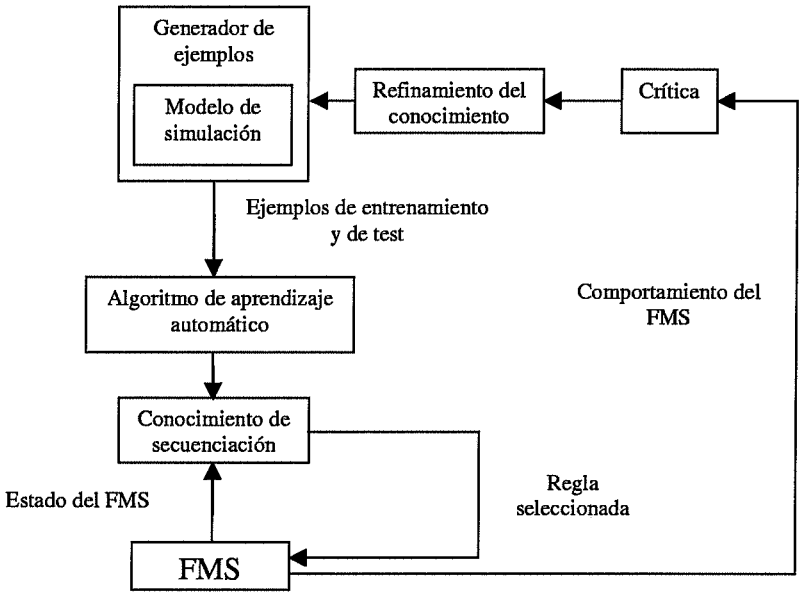


Figura 2. Esquema general de un sistema de secuenciación basado en conocimiento.

3. Una regla puede ser adecuada en una simulación durante un período de tiempo largo, para un conjunto dado de atributos, y no ser apropiada cuando se aplica de modo dinámico.
4. Existen escenarios, o estados del *FMS*, en los que el sistema de secuenciación no determina la regla adecuada que se debe utilizar.

El esquema general de un sistema de secuenciación basado en el conocimiento se muestra en la figura 2. Las etapas fundamentales de este sistema de secuenciación son las siguientes:

1. Creación de un conjunto de ejemplos de entrenamiento y de test mediante el generador de ejemplos. Para ello, es necesario definir los atributos adecuados que identifiquen el estado del sistema de fabricación. Obviamente, como no es posible tener en cuenta a todos ellos, se deben de elegir los más significativos. Los atributos seleccionados se denominan atributos de control siendo los valores utilizados de éstos, aquellos que se presentan con más frecuencia en el sistema de fabricación que se estudia. La clase o solución de cada ejemplo de entrenamiento o de test se obtiene a partir de la regla de secuenciación (o combinación de ellas, si existe más de un tipo de decisión que se debe tomar) que genere el mejor comportamiento en el sistema de fabricación. Para poder realizar lo anterior, se debe construir un modelo

de simulación del sistema de fabricación, y probar para cada conjunto de valores de los atributos de control (ejemplo de entrenamiento o de test) el comportamiento del sistema de fabricación con las diversas reglas de secuenciación que se pretenden utilizar.

2. Determinación del «conocimiento de secuenciación» mediante un algoritmo de aprendizaje automático.
3. Determinación de la regla de secuenciación más adecuada (o combinación de ellas, si existe más de un tipo de decisión), mediante el «conocimiento de secuenciación», dependiendo de los valores que presenten los atributos (estado del sistema de fabricación) en cada momento.
4. Comparación del comportamiento del sistema de fabricación utilizando el «conocimiento de secuenciación» y la mejor regla de secuenciación, o combinación de ellas. Si este segundo método produce un comportamiento del sistema de fabricación superior, se debe regresar al primer paso para refinar el «conocimiento de secuenciación».

A continuación, se analizan diversos sistemas de secuenciación basados en el conocimiento, que modifican de forma dinámica la regla de secuenciación empleada en cada momento. Estos sistemas, según el tipo de algoritmo de aprendizaje automático utilizado, se pueden dividir en las siguientes categorías:

1. Sistemas que no utilizan algoritmos de adquisición de conocimiento.
2. Sistemas basados en aprendizaje inductivo.
3. Sistemas basados en redes neuronales.
4. Sistemas mixtos. En este caso se utiliza una combinación de distintos tipos de algoritmos de aprendizaje.

4.1. Sistemas de secuenciación que no utilizan algoritmos de adquisición de conocimiento

Thesen y Lei (1986) sugieren un sistema experto para secuenciar robots en un sistema de galvanizado flexible. Los autores realizan un conjunto de simulaciones previas utilizando diversas reglas de secuenciación para estudiar el comportamiento del sistema de fabricación en diversos escenarios, obteniendo, de este modo, 38 ejemplos de entrenamiento. Sin embargo, el conocimiento acerca del sistema de fabricación no se alcanza mediante ningún procedimiento de adquisición automática, sino que se efectúa de modo manual, inspeccionando directamente los resultados de las simulaciones. Los autores comprueban que el sistema de fabricación aumenta el número de piezas producidas en unos porcentajes que varían entre un 7% y un 30%.

Sarin y Salgame (1990) definen un sistema experto para realizar secuenciación de tipo dinámico. El sistema de secuenciación tiene, al comienzo de un período de tiempo determinado, una secuencia ordenada de los trabajos que se realizan durante ese período y actúa cuando se produce un cambio. Los cambios se clasifican en diferentes grupos: avería de una máquina, llegada de un trabajo urgente, llegada de un nuevo lote de trabajos, falta de material, absentismo laboral, finalización de un trabajo en una máquina y cambio de turno. El sistema de secuenciación propuesto consta de los siguientes elementos: un «conocimiento de secuenciación», una base de datos global, una interface con el usuario y un bloque de control. El conocimiento está dividido en varios grupos, cada uno de los cuales contiene reglas que se encargan de resolver distintos tipos de problemas en función del cambio que ocurra en el sistema de fabricación. Las reglas representan las heurísticas de un secuenciador de trabajos humano.

La base de datos global contiene información de los diferentes trabajos y turnos existentes en el momento actual. El grupo de reglas que necesite información, acude a la base de datos global. Por último, el bloque de control, en forma de árbol («meta-reglas» o «conocimiento acerca del conocimiento»), elige el grupo de reglas adecuado al nuevo problema que origina el cambio en el sistema de fabricación. Del mismo modo, los autores presentan un sistema de secuenciación integrado que consta de dos módulos. El primero, que se apoya en la programación matemática, determina una secuencia de trabajos de tipo predictivo como punto inicial de partida. El segundo, el sistema experto, ante cualquier cambio que se produzca, retoma el control para ejecutar una secuenciación dinámica o reactiva en función de la nueva situación. Finalmente, los autores señalan que este sistema de secuenciación integrado aún no ha sido implementado en un caso real.

Chandra y Talavage (1991) presentan un sistema de secuenciación denominado *EXPERT*, formado por un conjunto de reglas de decisión. La información utilizada en el proceso de decisión es el nivel de congestión del sistema de fabricación, la preferencia de una pieza por una máquina, la criticidad de la pieza (indica la capacidad de la pieza para cumplir sus fechas de entrega) y el objetivo actual del sistema de fabricación. Los autores afirman que, en principio, es interesante el objetivo de maximizar el ritmo de «progreso de trabajo», aunque exista el peligro de que algunos trabajos se retrasen (sobre todo si el sistema es muy utilizado y son muchos los trabajos críticos). La excesiva preocupación por los trabajos críticos, puede empeorar todo el sistema; por lo tanto, se elige el objetivo de maximizar el ritmo de «progreso de trabajo» como criterio primario.

Por otra parte, los trabajos se dividen en grupos (preferencia alta, mediana y baja) en lugar de clasificarlos de forma individual. El sistema de secuenciación propuesto elige el trabajo que se asigna a la máquina empezando por los de preferencia alta, persiguiendo el objetivo primario y buscando oportunidades para mejorar el secundario (minimizar el número de trabajos retrasados) al mismo tiempo. En determinados casos, se inspeccionan los trabajos disponibles en un futuro cercano. En caso de empate o si no se logra

una decisión clara se aplica la regla *SPT*[3]. En la etapa experimental se comprueba que el sistema *EXPERT* es superior a las reglas de secuenciación convencionales.

Sabuncuoglu y Hommertzheim (1992) sugieren un algoritmo dinámico para secuenciar trabajos en máquinas y *AGVs*[4]. El algoritmo propuesto se basa en la idea de que un trabajo no debe ser asignado a una máquina si tiene que esperar por un *AGV* en la siguiente operación y viceversa. Para ello, se utilizan diversos esquemas de prioridad (o reglas) e información acerca del sistema de fabricación (niveles en las colas, número de piezas en el sistema, estado de las máquinas, etc.) y de los trabajos (tiempos de operación, número de operaciones, etc.). El algoritmo consta de dos partes fundamentales: un conjunto de procedimientos para secuenciar trabajos en las máquinas y otro para secuenciar los trabajos en los *AGVs*. Éstos últimos comprueban la existencia de estaciones bloqueadas o vacías de piezas en el «buffer» central.

Los autores comparan el algoritmo propuesto con las dos mejores combinaciones de reglas de secuenciación en máquinas y *AGVs*, utilizando como criterios de comportamiento el tiempo medio de una pieza en el sistema y el retraso medio. Para ello, proponen diversos escenarios variando los niveles de carga, la capacidad de las colas, la variable *F*[5] (Baker, 1984), el tipo de distribución de los tiempos de procesamiento y el criterio de comportamiento. Se comprueba que el algoritmo funciona mejor que las mejores combinaciones de reglas cuando la carga de las máquinas es elevada (si es baja, en las colas apenas hay piezas, con lo cual no influye la regla seleccionada) y el tamaño de las colas pequeño. En estas condiciones, la mejora obtenida es superior al 12%.

Pierreval y Mebarki (1997) presentan una metodología heurística, denominada *SFSR* (Shift From Standard Rules), para modificar de forma dinámica las reglas de secuenciación, teniendo en cuenta dos criterios de comportamiento (uno primario y otro secundario). La heurística *SFSR* chequea el estado del sistema de fabricación cuando se vuelve disponible un recurso o entra un nuevo trabajo. Mediante el empleo de unas reglas definidas de antemano que son función de unos parámetros que se deben optimizar, se puede detectar la presencia de determinadas anomalías en el sistema de fabricación (por ejemplo, la saturación de una máquina, un trabajo que permanece demasiado tiempo en el sistema, etc.). Los valores óptimos de los parámetros se calculan mediante el método de Hooke y Jeeves (1961). Si no existen anomalías en el sistema de fabricación se utilizan reglas de tipo estándar, tomadas de la literatura en función del criterio que se vaya a optimizar, que permiten determinar la regla de secuenciación que se empleará.

---

[3]Del inglés, Shortest Processing Time. Esta regla elige el trabajo que tiene el menor tiempo de procesamiento.

[4]Del inglés, Automatic Guided Vehicles.

[5]La variable F (flow allowance factor) determina la holgura de las fechas de entrega.

En caso contrario, se utilizan reglas definidas por los autores, que dependen de los criterios que se optimizan, de la anomalía detectada y del estado del sistema. En general, la metodología propuesta supera, en el criterio primario, a la alternativa de usar una regla de forma constante; y si no, lo compensa con el criterio secundario. Los porcentajes de mejora varían entre el 12.3% y el 33.8%. El principal inconveniente de la metodología propuesta es que las reglas de tipo estándar se definen de acuerdo con los resultados de investigaciones previas presentadas en la literatura. Un enfoque alternativo sería generarlas mediante aprendizaje inductivo, para tener en cuenta las particularidades del sistema de fabricación bajo estudio. También se podría utilizar aprendizaje inductivo para generar los otros tipos de reglas empleadas en *SFSR*.

4.2. Sistemas de secuenciación basados en aprendizaje inductivo

Pierreval y Ralambondrainy (1990) sugieren una técnica de aprendizaje inductivo, denominada *GENREG*, para obtener reglas heurísticas que permitan conocer el comportamiento del sistema de fabricación ante diferentes reglas de secuenciación y estados del sistema de fabricación. En la metodología propuesta se obtiene una regla con cada ejemplo de entrenamiento. Debido a que el número de reglas es muy grande, se utiliza posteriormente *GENREG* para generalizarlas y, de esta forma, reducir su número. La metodología se aplica en una configuración sencilla, de tipo «flow shop» con dos máquinas, utilizando 198 ejemplos de entrenamiento. Sin embargo, las reglas obtenidas mediante *GENREG* no se emplean de forma dinámica.

Shaw *et al.* (1992) presentan un sistema de secuenciación denominado *PDS* (Pattern-Directed Scheduling) para secuenciar trabajos en un *FMS* que utiliza aprendizaje inductivo. En este caso, el algoritmo de aprendizaje utilizado es *ID3* (Quinlan, 1986). En la etapa de adquisición de conocimiento se utilizan 130 ejemplos de entrenamiento, empleando como criterio de comportamiento el retraso medio de los trabajos. El sistema de secuenciación propuesto genera una reducción media en los retrasos de los trabajos del 11.5%.

Los autores observan que la máxima efectividad del sistema de secuenciación se obtiene cuando el número de cambios en los estados del sistema de fabricación (patterns) es de medio a razonablemente alto, y no existe un estado claramente dominante. Este rango también corresponde a valores de retraso que varían entre bajos y moderadamente altos. Asimismo, el número de máquinas alternativas para procesar una operación dada no debe de ser muy elevado. Los autores afirman que todas estas características se presentan en la mayor parte de los *FMSs* reales.

Nakasuka y Yoshida (1992) proponen un sistema de secuenciación denominado *LADS* (Learning-Aided Dynamic Scheduler) que incorpora un algoritmo de aprendizaje inductivo con dos características que lo diferencian de los algoritmos convencionales. En

primer lugar, un nuevo criterio para decidir como separar los grupos de datos, ya que debido al ruido que presentan éstos en los problemas de secuenciación, no interesa dividirlos de forma que pertenezcan a una sola «clase» pues el número de datos en cada grupo sería muy pequeño. La segunda característica del algoritmo propuesto, es la generación de combinaciones lineales de los atributos facilitados al sistema. El sistema de secuenciación propuesto se utiliza en un sistema de fabricación sencillo «flow shop» con 3 máquinas, con el objetivo de minimizar el «makespan» (tiempo total invertido para realizar todos los trabajos) y mantener el retraso por debajo de un valor determinado. Los autores observan que el sistema de secuenciación propuesto es superior (en los dos criterios) a utilizar una regla de forma constante.

Piramuthu *et al.* (1993) definen un sistema de secuenciación para secuenciar tareas de modo dinámico. El sistema propuesto, así como los ejemplos en los que lo aplica, es similar al presentado por los autores en otros trabajos (ver por ejemplo, Shaw *et al.*, 1992; Piramuthu *et al.*, 1994). Lo que aporta este artículo, con respecto a otros publicados por los mismos autores, es un marco teórico más elaborado de cada una de las partes que forman el sistema de secuenciación de trabajos.

Piramuthu *et al.* (1994) definen un sistema para secuenciar trabajos, mediante aprendizaje inductivo, en un sistema de fabricación flexible de tipo «flow-shop», utilizando como criterio de comportamiento el tiempo medio de una pieza en el sistema. Empleando *C*4.5 (Quinlan, 1993) como algoritmo de aprendizaje, se generan dos árboles de decisión, mediante 66 ejemplos de entrenamiento. El primer árbol se utiliza para secuenciar los trabajos en las propias máquinas, y el segundo, para decidir el orden de entrada de los trabajos en el sistema de fabricación. Del mismo modo, se presenta un procedimiento de refinamiento de los árboles de decisión, que consiste en incluir en el conjunto de entrenamiento, aquellos casos que el sistema clasifica mal.

Los autores observan que la incorporación de un árbol de decisión, para seleccionar la regla de secuenciación, no mejora los resultados de forma significativa con respecto a la alternativa de usar sólo el árbol de decisión para fijar el orden de entrada de las piezas, utilizando una regla de secuenciación de forma constante. Además, esta metodología es particularmente útil cuando el tamaño del «buffer» de entrada es limitado y pequeño, y existe una gran variación en los tiempos de procesamiento de las piezas en las máquinas cuello de botella.

4.3. Sistemas de secuenciación basados en redes neuronales

Chen y Yih (1996) definen una metodología para determinar los atributos más importantes como paso previo a la construcción de sistemas de secuenciación basados en el conocimiento. El mecanismo de identificación de atributos consta de tres pasos:

1. Recopilación de datos a través de la simulación del sistema de fabricación.

2. Construcción, mediante redes neuronales de tipo «backpropagation», de funciones de «mapeo atributo-comportamiento» del sistema de fabricación.
3. Selección de los atributos esenciales.

Para ello, se omite un atributo y se mide la diferencia entre la salida original y la obtenida con el atributo omitido. Los atributos que son importantes, tienen una diferencia mayor que aquellos con un nivel de significación más bajo. En el estudio experimental se usan 20 atributos candidatos (utilizando la variabilidad del atributo que se define como el cociente entre la varianza y la media de éste) tomados de la literatura (Nakasuka y Yoshida, 1992; Cho y Wysk, 1993; Chiu, 1994), 6 medidas de comportamiento y 10 reglas de secuenciación. Se generan 1300 ejemplos de entrenamiento para cada tipo de regla y la correspondiente red neuronal, seleccionando los 10 atributos más importantes.

Finalmente los autores comparan 3 redes neuronales con nodos de entrada formados por grupos alternativos de atributos (con los 10 atributos seleccionados, con los 20 atributos iniciales y con los 10 restantes) y nodos de salida correspondientes a las reglas que se eligen. Se verifica que la capacidad de generalización de la red obtenida con los atributos significativos es un 9% superior a la de la red formada por los 20 atributos iniciales. Los autores afirman que usar tantos atributos como sea posible, para construir una base de conocimiento, no mejora su capacidad de generalización o predicción. Cuantos más atributos se incluyan, el esfuerzo necesario para el desarrollo de la base de conocimiento aumenta, y su estructura se vuelve más compleja. Los mayores defectos de la metodología propuesta son que no identifica atributos importantes si éstos no están inicialmente considerados y que el proceso se debe de repetir si cambian las medidas de comportamiento.

Sun y Yih (1996) emplean un sistema de secuenciación que utiliza una red neuronal de tipo «backpropagation» en cada una de las máquinas, para la elección de la regla de secuenciación más adecuada en un entorno multicriterio. En cada punto de decisión, cuando una máquina necesita seleccionar un nuevo trabajo, un módulo de ajuste determina, en función de los valores deseados y actuales, la importancia relativa de cada criterio de comportamiento. La red neuronal, tomando como entrada el valor proporcionado por el módulo de ajuste y el estado actual del sistema de fabricación, proporciona la regla de secuenciación más adecuada. Los autores utilizan alrededor de 1000 ejemplos de entrenamiento para cada red neuronal y comprueban que el sistema de secuenciación propuesto genera en el sistema de fabricación, como término medio, una mejora superior al 4.2% con respecto a la alternativa de utilizar la mejor de las reglas. Asimismo, el sistema de secuenciación presenta una gran adaptabilidad ante cambios en la elección de los criterios de comportamiento considerados como prioritarios. El mayor defecto de este estudio es que el sistema de fabricación considerado no tiene rutas flexibles y el número de piezas es limitado.

Min *et al.* (1998) proponen un sistema de secuenciación que utiliza redes neuronales «competitivas». En este caso, se emplean como atributos las diferencias entre los valores, en los diferentes intervalos de tiempo, de los criterios de comportamiento y de las variables de estado del sistema de fabricación. Simulando el sistema de fabricación durante un largo período de tiempo, los ejemplos de entrenamiento se obtienen con estas diferencias, modificando las reglas de forma aleatoria. En la etapa de entrenamiento se utilizan 3500 ejemplos y se definen 40 «clases» o nodos en la red. Posteriormente, se emplean las redes neuronales para obtener las «clases» a partir de los ejemplos de entrenamiento.

El sistema de secuenciación funciona en tiempo real de la siguiente forma. En primer lugar, el usuario fija unas diferencias como objetivo y se identifica la «clase» mediante la red neuronal. A continuación, dentro de todos los ejemplos de entrenamiento, se buscan aquellos que tengan la misma «clase» y las mismas variables de decisión (reglas de secuenciación) del intervalo anterior. Si no se encuentra este ejemplo, lo cual es probable, se elige para cada variable de decisión, la regla más utilizada dentro de la correspondiente «clase». Una característica interesante de este sistema de secuenciación es que se emplean las reglas de secuenciación anteriores para hallar las actuales. El sistema de secuenciación propuesto se compara con otro que elige las reglas de forma aleatoria (en realidad, se hacen 10 réplicas del sistema de secuenciación aleatorio y se elige la mejor de ellas), siendo superior el comportamiento del primero. Los principales defectos del sistema de secuenciación son la falta de un método que sistemáticamente busque un número óptimo de nodos de salida de la red neuronal y que se compara con un sistema de secuenciación aleatorio y no con la mejor combinación posible de las reglas propuestas.

4.4. Sistemas de secuenciación mixtos

Wu y Wysk (1988) proponen un sistema de control y secuenciación denominado *MPECS* (Multipass Expert Control System), que combina sistemas expertos, simulación y aprendizaje inductivo. El sistema de secuenciación propuesto consta de tres módulos: un módulo de secuenciación inteligente, un simulador del sistema de fabricación y un módulo de control de célula. El primer elemento, a su vez, está compuesto de una base conocimiento, un motor de inferencia y un módulo de aprendizaje. La base tiene conocimiento de tipo «declarativo» (información acerca del estado del sistema de fabricación, reglas y heurísticas de secuenciación) y de «procedimiento» (criterios de tipo general, en forma de reglas, para seleccionar las reglas de secuenciación). El motor de inferencia es un mecanismo de búsqueda para elegir las reglas adecuadas del conocimiento de «procedimiento». Por último, el módulo de aprendizaje genera un conjunto de reglas a través de ejemplos de entrenamiento que asocian reglas de secuenciación, medidas de comportamiento y características del sistema de fabricación. Las reglas creadas por el módulo de aprendizaje se envían al conocimiento de «procedimiento».

El módulo de secuenciación inteligente se activa cuando llega un nuevo trabajo, o si se produce una anomalía en el sistema de fabricación. El simulador tiene como misión examinar las reglas de secuenciación sugeridas por el módulo de secuenciación inteligente y elegir la mejor. Por último, el módulo de control posibilita la secuenciación en la célula física; además, la mayor parte de la información de ésta, que es esencial para el control de la célula, se obtiene y manipula a través de este módulo. Los autores señalan que el sistema de secuenciación propuesto produce una mejora en el comportamiento del sistema de fabricación que varía entre el 2.3% y el 29.3% cuando se compara con la alternativa de utilizar una regla de secuenciación de forma constante.

Rabelo y Alptekin (1989) definen un sistema de secuenciación, denominado *ISS/FMS* (Intelligent Scheduling System for *FMS*), para secuenciar trabajos en un *FMS* que consta de tres módulos básicos. El primero de ellos, es un sistema experto que basándose en determinada información (datos de los trabajos a realizar, restricciones impuestas por el taller, estado de la célula, etc.) decide la regla heurística que se debe utilizar. Para ello, tiene en cuenta los datos facilitados por una red neuronal y un módulo de análisis estadístico que estudian casos pasados obtenidos mediante un estudio de simulación. El segundo módulo ejecuta un procedimiento heurístico (Kiran y Alptekin, 1989) que depende de dos coeficientes determinados por una red neuronal, en función de las características del problema de secuenciación que se resuelva. El tercer componente elige la mejor de las soluciones determinadas por los dos anteriores módulos.

Cho y Wysk (1993) presentan un sistema de secuenciación denominado *IWC* (Intelligent Workstation Controller) que utiliza redes neuronales de tipo «backpropagation» y un simulador basado en el trabajo de Wu y Wysk (1989). La red neuronal tiene 7 nodos de entrada, correspondientes al estado del sistema de fabricación, y 9 nodos de salida, uno para cada regla de secuenciación considerada. La red se entrena, con 90 ejemplos tomados de la literatura, teniendo en cuenta diversas configuraciones de las capas ocultas y distintos ritmos de aprendizaje. El simulador, con las dos mejores reglas que proporciona la red, en función del estado del sistema de fabricación, selecciona la mejor de ellas. Asimismo, los autores calculan de forma experimental, mediante un conjunto de simulaciones, la ventana de simulación más adecuada para el criterio de comportamiento elegido. Se observa que *IWC* es superior al empleo de una regla de forma constante, aunque el porcentaje de mejora no supera nunca el 3%.

Li y She (1994) emplean un sistema de secuenciación que utiliza «análisis cluster» (Evert, 1980) y aprendizaje inductivo. Mediante 600 ejemplos de entrenamiento, que generan de forma uniforme del espectro de decisión, y «análisis cluster», se establecen 7 «clases» con valores de comportamiento similares. Posteriormente, con un algoritmo similar a *C*4.5, establecen dos conjuntos de reglas que determinan la «clase», en función de los atributos de decisión o de comportamiento. Una forma que proponen los autores de utilizar este «conocimiento de secuenciación» es fijar unas condiciones de comportamiento y determinar la «clase» a la cual corresponden. Con el otro conjunto

de reglas se determinan, una vez conocida la «clase», las variables de decisión que se van a tomar. Sin embargo, no se compara esta metodología con ninguna alternativa para comprobar su funcionamiento.

Chiu y Yih (1995) sugieren un sistema de secuenciación que utiliza aprendizaje inductivo y algoritmos genéticos. Éstos últimos se emplean para buscar un conjunto de ejemplos de entrenamiento que posea buena calidad. Para ello, en cada punto de decisión, se elige la mejor regla de secuenciación, y ésta, junto con el estado del sistema de fabricación, forma un caso de entrenamiento. Por otra parte, el algoritmo de aprendizaje puede modificar el árbol de decisión cuando se presentan nuevos ejemplos, sólo si el cambio es significativo. Los autores comprueban que el sistema de secuenciación propuesto es superior a utilizar una regla de secuenciación de forma constante. El mayor defecto del sistema de secuenciación presentado es la necesidad de cambiar el «conocimiento de secuenciación» inducido con pequeñas modificaciones en el sistema de fabricación.

Quiroga y Rabelo (1995) resuelven el problema de secuenciación de trabajos en una máquina, mediante aprendizaje inductivo (*ID3*), redes neuronales «backpropagation» y lógica «borrosa». Para ello, se utilizan 358 casos de entrenamiento y 198 de test, siendo el nivel de aciertos superior al 90% en las tres metodologías. El aprendizaje inductivo y la lógica «borrosa» presentan la ventaja de que generan reglas que son inteligibles para el ser humano, cosa que no ocurre con las redes neuronales. Sin embargo, éstas son las menos sensibles a ruidos o datos incompletos y presentan el mayor porcentaje de aciertos (98.8%).

Bowden y Bullington (1996) sugieren un sistema de secuenciación denominado *GARDS* (Genetic Algorithm Rule Discovery System), para determinar estrategias de control utilizando algoritmos genéticos. *GARDS* contiene tres bloques fundamentales:

1. Un modelo de simulación para analizar el comportamiento del sistema de fabricación con las distintas estrategias generadas.
2. Un algoritmo que determina la regla más adecuada, dentro de una estrategia o plan, para el estado actual del sistema de fabricación.
3. Un algoritmo genético que intenta, mediante los operadores tradicionales de cruce y mutación, mejorar los planes iniciales, eligiendo el mejor para el control del sistema de fabricación.

El sistema de secuenciación propuesto se prueba en dos configuraciones de distinta complejidad con el objetivo de minimizar el número de trabajos retrasados. Se observa que *GARDS* mejora el comportamiento del sistema de fabricación con respecto a diversos métodos clásicos heurísticos (por ejemplo, enviar un trabajo a la cola de la máquina con menor número de trabajos).

Lee *et al.* (1997) proponen un sistema de secuenciación que también emplea aprendizaje inductivo y algoritmos genéticos. La primera técnica se utiliza, generando un árbol de

Tabla 2. Clasificación de las referencias según la metodología empleada.

Metodología	*Referencias*
Sistemas basados en métodos analíticos	Han *et al.* (1989); Hutchison *et al.* (1989); Kimemia y Gershwin (1985); Lashkari *et al.* (1987); Shanker y Rajamarthandan (1989); Shanker y Tzen (1985); Stecke (1983); Wilson (1989).
Sistemas basados en métodos heurísticos	Choi y Malstrom (1988); Denzler y Boe (1987); Egbelu y Tanchoco (1984); Henneke y Choi (1990); Montazeri y Van Wassenhove (1990); Stecke y Solberg (1981); Tang *et al.* (1993)
Sistemas basados en simulación	Ishii y Talavage (1991); Jeong y Kim (1998); Kim y Kim (1994); Wu yWysk (1989);
Sistemas basados en inteligencia artificial	Bowden y Bullington (1996); Chandra y Talavage (1991); Chen y Yih (1996); Chiu y Yih (1995); Cho y Wysk (1993); Kim *et al.* (1998); Lee *et al.* (1997); Li y She (1994); Min *et al.* (1998); Nakasuka y Yoshida (1992); Pierreval y Mebarki (1997); Pierreval y Ralambondrainy (1990); Piramuthu *et al.* (1993); Piramuthu *et al.* (1994); Quiroga y Rabelo (1995); Rabelo y Alptekin (1989); Sabuncuoglu y Hommertzheim (1992); Sarin y Salgame (1990); Shaw *et al.* (1992); Sun y Yih (1996); Thesen y Lei (1986); Wu y Wysk (1988)

decisión mediante *C*4.5, para seleccionar la regla más adecuada que controle el flujo de entrada de trabajos en el sistema de fabricación. Por otra parte, los algoritmos genéticos se emplean para seleccionar las reglas de secuenciación más apropiadas en cada una de las máquinas del sistema de fabricación. Los autores verifican el sistema de secuenciación propuesto en dos sistemas de tipo «job shop» (uno de ellos con una máquina cuello de botella) utilizando el retraso medio como criterio de comportamiento, y comprueban que supera a la mejor combinación de reglas utilizadas de forma constante, en un porcentaje que varía entre el 20.34% y el 25.28%. Sin embargo, los tiempos requeridos (26 y 168 minutos para el primer y segundo caso, respectivamente) son bastante elevados para que el sistema de secuenciación funcione en tiempo real.

Kim *et al.* (1998) sugieren un sistema de secuenciación, ampliando un trabajo anterior (Min *et al.*, 1998), que emplea redes neuronales «competitivas» y aprendizaje inductivo. Se aplica esta última técnica, una vez obtenidas las «clases» mediante las redes neuronales, para expresar el conocimiento en forma de árbol y reglas de producción. Los autores utilizan 99.999 casos de entrenamiento y fijan una red con 100 grupos o «clases». El sistema de secuenciación funciona en tiempo real del mismo modo que el sistema presentado previamente por los mismos autores (Min *et al.*, 1998). La única diferencia es que la «clase» se identifica mediante las reglas de producción obtenidas del programa de aprendizaje inductivo *C*4.5. Los autores comparan este sistema de secuenciación con otro que solamente utiliza la red neuronal «competitiva», comprobando la

Tabla 3. Clasificación de las referencias según el algoritmo de aprendizaje automático empleado.

Algoritmo de aprendizaje automático	*Referencias*
No utiliza	Chandra y Talavage (1991); Pierreval y Mebarki (1997); Sabuncuoglu y Hommertzheim (1992); Sarin y Salgame (1990); Thesen y Lei (1986)
Aprendizaje inductivo	Nakasuka y Yoshida (1992); Pierreval y Ralambondrainy (1990); Piramuthu *et al.* (1993); Piramuthu *et al.* (1994); Shaw *et al.* (1992)
Redes neuronales	Chen y Yih (1996); Min *et al.* (1998); Sun y Yih (1996)
Mixto	Bowden y Bullington (1996); Chiu y Yih (1995); Cho y Wysk (1993); Kim *et al.* (1998); Lee *et al.* (1997); Li y She (1994); Quiroga y Rabelo (1995); Rabelo y Alptekin (1989); Wu y Wysk (1988)

superioridad del primero debido al algoritmo de «podado» de los árboles de *C*4.5 que maneja de forma eficiente el ruido en los datos. Este sistema de secuenciación posee los mismos defectos que el sugerido en Min *et al.* (1998).

A modo de recapitulación, en la tabla 2 se presenta un resumen de las distintas aproximaciones existentes en la literatura, clasificadas según la metodología empleada. Por otro lado, en la tabla 3 se muestra una recopilación de los diferentes sistemas de secuenciación existentes que modifican de forma dinámica las reglas de secuenciación, clasificados según el tipo de algoritmo de aprendizaje automático utilizado.

5. LIMITACIONES DE LA LITERATURA EXISTENTE Y LÍNEAS DE INVESTIGACIÓN FUTURAS

A partir de los sistemas de secuenciación basados en el conocimiento presentados anteriormente, que utilizan algoritmos de aprendizaje automático, se detectan, en general, una serie de carencias o características deseables comunes. Éstas, pueden originar líneas de investigación en el campo de la secuenciación dinámica de sistemas de fabricación, mediante la modificación de la regla de secuenciación empleada. Estas carencias son las siguientes:

1. Comparación de diversos tipos de algoritmos de aprendizaje automático. En los sistemas de secuenciación presentados en la literatura se utiliza un algoritmo o, en determinados casos, una combinación de ellos. Sin embargo, no existe un estudio comparativo que determine cuál es el mejor de ellos. Por otra parte, debido a la gran disparidad de los *FMSs* utilizados en la literatura revisada, no es posible intuir cuál

de los algoritmos presentados es el más adecuado para la resolución de este tipo de problemas de secuenciación.

2. Utilización del *CBR* en los sistemas de secuenciación. Estos algoritmos poseen una gran eficacia de clasificación, a pesar de su sencillez. Sin embargo, ninguno de los sistemas de secuenciación presentados utiliza CBR; por lo tanto, sería interesante probar su idoneidad en los problemas de secuenciación.
3. Determinación del número de ejemplos de entrenamiento óptimo. En ninguno de los sistemas de secuenciación considerados se calcula el número de ejemplos necesario para entrenar el algoritmo de aprendizaje automático de forma óptima. Por otra parte, tampoco se especifica si los ejemplos de test son iguales, parecidos, o muy distintos de los de entrenamiento. Sin embargo, el error de clasificación del «conocimiento de secuenciación» y, por lo tanto, el comportamiento del sistema de fabricación, depende en gran medida del número de ejemplos de entrenamiento considerado. Por lo tanto, es necesario estudiar el error de clasificación en función del número de ejemplos considerado y elegir un tamaño adecuado del conjunto de entrenamiento.
4. Selección de un período de supervisión adecuado. En general, en la literatura existente, no se realiza estudio alguno para determinar el período de supervisión apropiado para cada criterio de comportamiento. Sin embargo, la frecuencia utilizada para chequear los atributos de control y decidir si se cambian, o no, las reglas de secuenciación, es un tema de vital importancia que determina el comportamiento del sistema de fabricación.
5. Determinación de un mecanismo o filtro que amortigüe los estados transitorios. En determinadas ocasiones, el sistema de fabricación alimentado con el «conocimiento de secuenciación» no se comporta tal como se esperaba, y es superado por la alternativa de utilizar la mejor combinación de reglas de secuenciación de forma constante. Este fenómeno se puede explicar por el hecho de que el sistema de secuenciación reacciona de forma precipitada ante cambios en los atributos de control que sólo son transitorios en el tiempo. Por ello, se propone utilizar filtros de tipo digital que permitan amortiguar estos escenarios transitorios en los atributos de control. En la mayor parte de los trabajos estudiados no se tiene en cuenta este mecanismo y, cuando se considera, no se analizan los diversos tipos de filtros digitales existentes y su relación con el período de supervisión.
6. Generación de nuevos atributos de control mediante un algoritmo que permita crear atributos que sean combinación de los iniciales. En algunos casos, para seleccionar las reglas de secuenciación más adecuadas, es preciso chequear relaciones del tipo: la utilización de la máquina 1 es menor que la de la máquina 2. Para poder lograr estas relaciones, sería necesario definir combinaciones aritméticas de los atributos básicos iniciales. Sin embargo, a menudo, tales combinaciones no son conocidas de antemano y sólo se encuentran, en sistemas de fabricación muy sencillos, después de examinar, en detalle, los resultados de simulación.

7. Incorporación de un simulador. El comportamiento del sistema de fabricación, podría mejorar si se incorporase un simulador que determine la mejor regla de entre las que el «conocimiento de secuenciación» considere más importantes. En ocasiones, el «conocimiento de secuenciación» determina que ante unos valores dados de los atributos de control existen dos, o más, reglas de secuenciación que podrían ser, en principio, adecuadas. En estos casos, en los cuales la decisión por parte del «conocimiento de secuenciación» no es clara, la incorporación del simulador sería bastante útil.
8. Refinamiento de la base de conocimiento. La base de conocimiento, una vez desarrollada, no es estática. Por lo tanto, sería interesante establecer un procedimiento que modifique el conocimiento automáticamente si se producen cambios importantes en el sistema de fabricación. La misión principal del módulo de refinamiento es descubrir deficiencias en la base de conocimiento y añadir casos de entrenamiento que cubran dichas deficiencias. Éstas, se pueden presentar en determinados rangos de valores de los atributos de control. Para solucionar este problema, se requiere «cubrir» estos rangos con nuevos casos de entrenamiento, de forma que el nuevo «conocimiento de secuenciación» obtenido, sea capaz de tratar estas situaciones.

6. CONCLUSIONES

En este artículo, se realiza una revisión de la literatura existente sobre secuenciación dinámica de *FMSs*, mediante aprendizaje automático, en la cual se modifica la regla de secuenciación empleada. Al efectuar esta revisión, se ha detectado una serie de carencias en los sistemas de secuenciación estudiados. En primer lugar, no se ha encontrado una comparación de los distintos algoritmos de aprendizaje automático existentes para determinar cuál es el mejor tipo de algoritmo para resolver esta clase de problemas de secuenciación. Asimismo, tampoco se ha utilizado el razonamiento basado en casos como algoritmo de aprendizaje automático, a pesar de su sencillez y gran eficacia de clasificación. Por otro lado, no se ha determinado el número óptimo de ejemplos de entrenamiento necesario para obtener un «conocimiento de secuenciación» con un error de clasificación pequeño.

Del mismo modo, en los sistemas de secuenciación estudiados, no se ha considerado, en general, la selección de un período de supervisión ni la utilización de filtros que amortigüen los estados transitorios del sistema de fabricación. De igual forma, tampoco se ha tenido en cuenta, en la mayor parte de los sistemas de secuenciación estudiados, la incorporación de un generador de nuevos atributos de control y de un simulador que apoye al «conocimiento de secuenciación» cuando éste no pueda determinar la regla de secuenciación que se debe utilizar. Asimismo, no se han encontrado, en general, módulos de refinamiento de la base de conocimiento que permitan modificar el

«conocimiento de secuenciación» si se producen cambios importantes en el sistema de fabricación. Por último, reseñar que sería interesante, como futuro trabajo, diseñar un sistema de secuenciación que incorpore las carencias anteriormente señaladas y medir el efecto de cada una de ellas en el comportamiento del sistema de fabricación.

7. REFERENCIAS

Baker, K. R. (1984). «Sequencing rules and due-date assignments in a job shop». *Management Science*, 30, 9, 1093-1103.

Basnet, C. Mize, J. H. (1994). «Scheduling and control of flexible manufacturing systems: a critical review». *International Journal Computer Integrated Manufacturing*, 7, 6, 340-355.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Blackstone, J. H.; Phillips, D. T. & Hogg, G. L. (1982). «A state-of-the-art survey of dispatching rules for manufacturing job shop operations». *International Journal of Production Research*, 20, 1, 27-45.

Bowden, R. & Bullington, S. F. (1996). «Development of manufacturing control strategies using unsupervised machine learning ». *IIE Transactions*, 28, 319-331.

Chandra, J. & Talavage, J. (1991). «Intelligent dispatching for flexible manufacturing». *International Journal of Production Research*, 29, 11, 2259-2278.

Chen, C. C. & Yih, Y. (1996). «Identifying attributes for knowledge-based development in dynamic scheduling environments». *International Journal of Production Research*, 34, 6, 1739-1755.

Chiu, C. (1994). *A learning-based methodology for dynamic scheduling in distributed manufacturing systems*, PhD thesis. School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

Chiu, C. & Yih, Y. (1995). «A learning-based methodology for dynamic scheduling in distributed manufacturing systems». *International Journal of Production Research*, 33, 11, 3217-3232.

Cho, H. & Wysk, R. A. (1993). «A robust adaptive scheduler for an intelligent workstation controller». *International Journal of Production Research*, 31, 4, 771-789.

Choi, R. H. &Malstrom, E. M. (1988). «Evaluation of traditional work scheduling rules in a flexible manufacturing system with a physical simulator». *Journal of Manufacturing Systems*, 7, 1, 33-45.

Denzler, D. R. &Boe, W. J. (1987). «Experimental investigation of flexible manufacturing system scheduling rules». *International Journal of Production Research*, 25, 7, 979-994.

Egbelu, P. J. & Tanchoco, J. M. A. (1984). «Characterization of automated guided vehicle dispatching rules». *International Journal of Production Research*, 22, 3, 359-374.

Evert, B. (1980). *Cluster Analysis*. New York: Heinemann.

Garey, M. & Johnson, D. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York Freeman.

Han, M.; Na, Y. K. & Hogg, G. L. (1989). «Real-time tool control and job dispathing in flexible manufacturing systems». *International Journal of Production Research*, 27, 1257-1267.

Henneke, M. J. & Choi, R. H. (1990). «Evaluation of FMS parameters on overall system performance». *Computer Industrial Engineering*, 18, 1, 105-110.

Hooke, R. & Jeeves, T. A. (1961). «Direct search solution of numerical and statistical problems». *Journal of the Association of Computer Machines*, 8, 212-229.

Hutchison, J.; Leong, K.; Snyder, D. & Ward, F. (1989). «Scheduling for random job shop flexible manufacturing systems». *Proceedings of the Third ORSA/TIMS Conference on Flexible Manufacturing Systems*, 161-166.

Ishii, N. & Talavage, J. (1991). «A transient-based real-time scheduling algorithm in FMS». *International Journal of Production Research*, 29, 12, 2501-2520.

Jeong, K.-C. & Kim, Y.-D. (1998). «A real-time scheduling mechanism for a flexible manufacturing system: using simulation and dispatching rules». *International Journal of Production Research*, 36, 9, 2609-2626.

Kim, Y.-D. (1990). «A comparison of dispatching rules for job shops with multiple identical jobs and alternative routings». *International Journal of Production Research*, 28, 5, 953-962.

Kim, M. H. & Kim, Y.-D. (1994). «Simulation-based real-time scheduling in a flexible manufacturing system». *Journal of Manufacturing Systems*, 13, 2, 85-93.

Kim, C.-O.; Min, H.-S. & Yih, Y. (1998). «Integration of inductive learning and neural networks for multi-objective FMS scheduling». *International Journal of Production Research*, 36, 9, 2497-2509.

Kimemia, J. & Gershwin, S. B. (1985). «Flow optimization in flexible manufacturing systems». *International Journal of Production Research*, 23, 81-96.

Kiran, A. & Alptekin, S. (1989). «A tardiness heuristic for scheduling flexible manufacturing systems». *15th Conference on Production Research and Technology: Advances in Manufacturing Systems Integration and Processes* (pp. 559-564), University of California at Berkeley, Berkeley, CA.

Lashkari, R. S.; Dutta, S. P. & Padhye, A. M. (1987). «A new formulation of operation allocation problem in flexible manufacturing systems: mathematical modelling and computational experience». *International Journal of Production Research*, 25, 1267-1283.

Lee, C.-Y.; Piramuthu, S. & Tsai, Y.-K. (1997). «Job shop scheduling with a genetic algorithm and machine learning». *International Journal of Production Research*, 35, 4, 1171-1191.

Li, D.-C. & She, I.-S. (1994). «Using unsupervised learning technologies to induce scheduling knowledge for FMSs». *International Journal of Production Research*, 32, 9, 2187-2199.

Michalski, R. S.; Carbonell, J. G. & Mitchell, T. M. (1983). *Machine Learning. An Artificial Intelligence Approach.* Palo Alto, CA: Tioga Press.

Min, H.-S.; Yih, Y. & Kim, C.-O. (1998). «A competitive neural network approach to multi-objective FMS scheduling». *International Journal of Production Research*, 36, 7, 1749-1765.

Montazeri, M. & Wassenhove, L. N. V. (1990). «Analysis of scheduling rules for an FMS». *International Journal of Production Research*, 28, 4, 785-802.

Nakasuka, S. & Yoshida, T. (1992). «Dynamic scheduling system utilizing machine learning as a knowledge acquisition tool». *International Journal of Production Research*, 30, 2, 411-431.

Panwalkar, S. S. & Iskander, W. (1977). «A survey of scheduling rules». *Operations Research*, 23, 5, 961-973.

Pierreval, H. & Mebarki, N. (1997). «Dynamic selection of dispatching rules for manufacturing system scheduling». *International Journal of Production Research*, 35, 6, 1575-1591.

Pierreval, H. & Ralambondrainy, H. (1990). «A simulation and learning technique for generating knowledge about manufacturing systems behaviour». *Journal of the Operational Research Society*, 41, 6, 461-474.

Piramuthu, S.; Raman, N. & Shaw, M. J. (1994). «Learning-based scheduling in a flexible manufacturing flow line». *IEEE Transactions on Engineering Management*, 41, 2, 172-182.

Piramuthu, S.; Raman, N.; Shaw, M. J. & Park, S. (1993). «Integration of simulation modeling and inductive learning in an adaptive decision support system». *Decision Support Systems*, 9, 127-142.

Quinlan, J. R. (1986). «Induction of decision trees». *Machine Learning*, 1, 1, 81-106. Reprinted in J. W. Shavlik and T. G. Dietterich (eds.), *Readings in Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1991. Reprinted in B. G. Buchanan and D. Wilkins (eds.), *Readings in Knowledge Acquisition and Learning*. San Mateo, CA: Morgan Kaufmann, 1992.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Quiroga, L. A. & Rabelo, L. C. (1995). «Learning from examples: a review of machine learning, neural networks and fuzzy logic paradigms». *Computers & Industrial Engineering*, 29, 561-565.

Rabelo, L. C. & Alptekin, S. (1989). «Integrating scheduling and control functions in computer integrated manufacturing using artificial intelligence». *Computers & Industrial Engineering*, 17, 101-106.

Ramasesh, R. (1990). «Dynamic job shop scheduling: a survey of simulation studies». *OMEGA: The International Journal of Management Science*, 18, 1, 43-57.

Russel, R. S.; Dar-El, E. M. & Taylor, B. W. (1987). «A comparative analysis of the COVERT job sequencing rule using various shop performance measures». *International Journal of Production Research*, 25, 10, 1523-1540.

Sabuncuoglu, I. & Hommertzheim, D. L. (1992). «Dynamic dispatching algorithm for scheduling machines and automated guided vehicles in a flexible manufacturing system». *International Journal of Production Research*, 30, 5, 1059-1079.

Sarin, S. C. & Salgame, R. R. (1990). «Development of a knowledge-based system for dynamic scheduling». *International Journal of Production Research*, 28, 8, 1499-1512.

Shanker, K. & Rajamarthandan, S. (1989). «Loading problem in FMS: part movement minimization». *Proceedings of the Third ORSA/TIMS Conference on Flexible Manufacturing Systems*, 99-104.

Shanker, K. & Tzen, Y. J. (1985). «A loading and dispatching problem in a random flexible manufacturing system». *International Journal of Production Research*, 23, 579-595.

Shaw, M. J.; Park, S. & Raman, N. (1992). «Intelligent scheduling with machine learning capabilities: the induction of scheduling knowledge». *IIE Transactions*, 24, 2, 156-168.

Stecke, K. E. (1983). «Formulation and solution of nonlinear integer production planning problems for flexible manufacturing systems». *Management Science*, 29, 3, March, 273-288.

Stecke, K. E. & Solberg, J. (1981). «Loading and control policies for a flexible manufacturing system». *International Journal of Production Research*, 19, 5, 481-490.

Sun, Y.-L. & Yih, Y. (1996). «An intelligent controller for manufacturing cells». *International Journal of Production Research*, 34, 8, 2353-2373.

Tang, L.-L.; Yih, Y. & Liu, C.-Y. (1993). «A study on decision rules of a scheduling model in an FMS». *Computer in Industry*, 22, 1-13.

Thesen, A. & Lei, L. (1986). «An expert system for scheduling robots in a flexible electroplating system with dynamically changing workloads». *Proceedings of the Second ORSA/TIMS Conference on Flexible Manufacturing Systems*, (pp. 555-566). Amsterdam: Elsevier Science Publishers.

Vepsalainen, A. P. J. & Morton, T. E. (1987). «Priority rules for job shops with weighted tardiness costs». *Management Science*, 33, 8, 1035-1047.

Watson, I. (1997). *Applying Case-Based Reasoning: Techniques for Enterprise Systems*. San Francisco, CA: Morgan Kaufmann Publishers.

Wilson, J. M. (1989). «An alternative formulation of the operation-allocation problem in flexible manufacturing systems». *International Journal of Production Research*, 27, 1405-1412.

Wu, S.-Y. D. & Wysk, R. A. (1988). «Multi-pass expert control system-a control/scheduling structure for flexible manufacturing cells». *Journal of Manufacturing Systems*, 7, 2, 107-120.

Wu, S.-Y. D. & Wysk, R. A. (1989). «An application of discrete-event simulation to online control and scheduling in flexible manufacturing». *International Journal of Production Research*, 27, 9, 1603-1623.

ENGLISH SUMMARY

DYNAMIC SCHEDULING OF FLEXIBLE MANUFACTURING SYSTEMS THROUGH MACHINE LEARNING: AN ANALYSIS OF THE MAIN SCHEDULING SYSTEMS

PAOLO PRIORE*
DAVID DE LA FUENTE*
JAVIER PUENTE*
ALBERTO GÓMEZ*

A common way of dynamically scheduling jobs in a flexible manufacturing system (FMS) is by means of dispatching rules. The drawback of this method is that the performance of the manufacturing system depends on the state the manufacturing system is in at each moment, and no one rule exists that overrules the rest in all the possible states that the manufacturing system may be in. It would therefore be interesting to use the most appropriate dispatching rule at each moment. To achieve this goal, a scheduling system which uses machine learning can be used. By means of this technique, and by analysing the previous performance of the manufacturing system (training examples), knowledge is generated that can be used to decide which is the most appropriate dispatching rule at each moment in time. This paper provides a review of the main scheduling systems that use machine learning to vary the dispatching rule dynamically that have been described in the literature.

Keywords: Dynamic Scheduling, Machine Learning, Flexible Manufacturing Systems, Dispatching Rules, Simulation

AMS Classification (MSC 2000):

* Dpto. de Administración de Empresas y Contabilidad. Escuela Técnica Superior de Ingenieros Industriales e Informáticos de Gijón. Universidad de Oviedo.

- Received December 1999.
- Accepted July 2001.

A common way of dynamically scheduling jobs in a flexible manufacturing system (*FMS*) is by means of dispatching rules. The problem of this method is that the performance of these rules depends on the state the system is in at each moment, and no one rule exists that overrules the rest in all the possible states that the system may be in. It would therefore be interesting to use the most appropriate dispatching rule at each moment. To achieve this goal, a scheduling approach which uses machine learning can be used. By means of this technique, analysing the previous performance of the system (training examples), a knowledge is generated that can be used to decide which is the most appropriate dispatching rule at each moment in time. In this paper, a review of the main machine learning-based scheduling approaches described in the literature is presented.

Basically, two approaches can be found in the literature to modify dispatching rules dynamically. Firstly, the rule is selected at the appropriate moment by simulating a set of pre-established dispatching rules and selecting the one that provides the best performance. In the second approach, belonging to the field of artificial intelligence, a set of earlier system simulations (training examples) is used to determine which is the best rule for each possible system state. These training cases are used to train a machine learning module to acquire knowledge about the manufacturing system. Such knowledge is then used to make intelligent decisions in real time. These systems are normally said to be knowledge-based.

Several knowledge-based approaches that dynamically modify the dispatching rule being used at a specific instance are reviewed next. Depending on the type of machine learning algorithm used, these approaches can be divided into the following categories:

1. Approaches that do not use knowledge acquisition algorithms.
2. Inductive learning-based approaches.
3. Neural network-based approaches.
4. Mixed approaches. Here a combination of different types of learning algorithm is applied.

Inductive learning algorithms or neural networks are used in many of the above approaches to acquire knowledge. The advantage of neural networks over inductive learning is that the class need not be predefined (when the number of decision variables is high, the number of simulations to be done to determine the class is great). However, the knowledge of the network is not intelligible for the user and it is difficult for him or her to modify the knowledge. It is similarly easy to identify which attributes really affect performance of the dispatching rules considered by using an inductive learning algorithm (Shaw *et al.*, 1992).

Moreover, a series of common defects are generally noticeable in the above-mentioned knowledge-based approaches that use machine learning algorithms. These defects might

suggest lines of research in the field of dynamic scheduling of manufacturing systems, based on modifying the dispatching rule used. These research lines are the following:

1. A comparison of the different machine learning methodologies. One methodology, or in some cases a combination of them, is used in the approaches presented in the literature. However, there is no comparative study to determine which is the best of them. Furthermore, it would be interesting to use case-based reasoning (CBR) as a machine learning methodology in scheduling systems.
2. Selection of an adequate period of monitoring. In the literature available no study is generally made to calculate the monitoring period for each performance criterion.
3. The determination of a mechanism that mitigates transitory states. In most works that were studied this mechanism is not taken account of, and when it is, variable time is not used.
4. Determination of the optimum number of training examples. In most approaches that were considered the number of examples required to train the machine learning algorithm optimally is not calculated. On the other hand, it's not specified if test examples are similar or very different to the training examples.
5. Generation of new attributes. Development of an algorithm that serves to create new attributes that are linear combinations of the initial ones.
6. Incorporation of a simulator. The scheduling system's performance could be improved if a simulator were incorporated to determine the best rule from amongst the machine learning system considers most important.
7. Refinement of the knowledge base. It would be interesting to establish a procedure to automatically modify knowledge when important changes in the manufacturing system occur.

Estadística Oficial

QÜESTIIÓ, vol. 25, 3, p. 553-580, 2001

EL CENS DE POBLACIÓ: UN ASSAIG D'INTERPRETACIÓ MITJANÇANT DATA MINING

CRISTINA GUISANDE ALLENDE*
FRANCESC SUBIRADA CURCO**

En aquest article es presenten els resultats d'un treball fet conjuntament entre l'Institut d'Estadística de Catalunya (Idescat), el Centre de Supercomputació de Catalunya (Cesca) i IBM. Les tres institucions van realitzar una anàlisi de les dades del Cens de població i habitatge de 1991 segons la metodologia Data Mining. Del conjunt de tècniques que es poden aplicar amb l'Intelligent Miner, es van fer servir les tècniques de segmentació i d'associació que s'han aplicat als individus segons el tipus de llar al que pertanyen.

Population Census: an interpretation using Data Mining

Paraules clau: Data Mining, estadística oficial, associacions, segmentació, anàlisis de dades, aplicació, mètodes estadístics

Classificació AMS (MSC 2000): 62-07, 62P25, 62H30, 62H20

* Institut d'Estadística de Catalunya. Via Laietana, 58. 08003 Barcelona. E-mail: cguisande@idescat.es
** CEPBA-IBM Research Institute. Jordi Girona, 1-3. 08034 Barcelona. E-mail: frsubirada@es.ibm.com

- Rebut l'abril de 2001.
- Acceptat el novembre de 2001.

1. INTRODUCCIÓ

L'Institut d'Estadística de Catalunya (Idescat), el Centre de Supercomputació de Catalunya (Cesca) i IBM van signar un conveni de col·laboració per tal de realitzar una anàlisi de les dades del Cens de població i habitatge de 1991 segons la metodologia Data Mining.

L'Idescat va aportar l'arxiu estadístic del Cens de població i habitatge, IBM va contribuir amb la tecnologia de tractament de les dades i per la seva banda, Cesca va aportar l'equip informàtic adient per procedir a l'execució del projecte.

L'objectiu d'aquest article és presentar els principals resultats d'aquesta experiència on es va aplicar Data Mining a informació censal. L'article s'estructura de la següent manera: en primer lloc es realitza una breu descripció dels recursos informàtics utilitzats, seguidament es presenten les característiques de les dades censals i el seu tractament i, finalment, s'examinen els principals resultats; en un annex final es recullen els trets bàsics de les operacions de Mineria de Dades.

2. LA MINERIA DE DADES

Durant la col·laboració entre l'Institut d'Estadística de Catalunya, el Cesca i IBM es va utilitzar la següent infrastructura:

1. A nivell de Programari:

El producte IBM anomenat «DB2 Intelligent Miner for Data». Aquest és un producte que permet aplicar operacions de Mineria de Dades als registres per poder trobar en ells informació que prèviament es desconeix.

2. A nivell de Maquinari:

Un ordinador IBM de procés paral·lel anomenat «System Paralel 2» (SP2). Aquest ordinador permet la utilització conjunta de diferents processadors per resoldre un sol problema. En el cas concret de Mineria de Dades dóna la possibilitat de treballar amb grans bases de dades (per exemple el cens de població) en un temps raonable.

Respecte a les operacions que es poden realitzar aquestes són: segmentació, classificació, predicció i anàlisi de relacions. Una breu descripció de les mateixes es presenta en un annex al final d'aquest article. Per les característiques de les dades censals descrites més endavant solament es van poder aplicar dues de les operacions esmentades: la segmentació i dins de l'anàlisi de relacions, l'associació.

3. LES DADES CENSALS

Un cens de població es pot definir com el conjunt d'operacions de recollida, elaboració, valoració i anàlisi de les dades de caràcter demogràfic, cultural, econòmic i socials de tots els habitants d'un país amb referència a un moment o període determinat. Per aquest motiu, acostuma a ser una de les tasques fonamentals dels instituts d'estadística de tots els països. Aquestes dades es recullen cada deu anys aproximadament d'acord amb la normativa internacional de l'ONU i també de la Comunitat Europea.

A Catalunya l'últim cens es va realitzar prenent per referència les zero hores de l'1 de març de 1991. Comprèn les persones que tenien fixada la seva residència a Catalunya en el moment censal i aquelles que tot i no residir-hi, s'hi trobaven en la data de referència.

El cens de població és una operació quasi insubstituïble per al recompte de la població i disposa de la màxima protecció legal per tal de salvaguardar la confidencialitat de les respostes dels ciutadans al qüestionari censal. La legislació vigent protegeix i garanteix la confidencialitat de les dades censals individuals sotmeses a secret estadístic que no poden ser divulgades, fins i tot a altres administracions o organismes públics no sotmesos al compliment del secret estadístic. La protecció legal assegura el tractament operatiu de la informació mitjançant el qual aquesta no podrà ser tractada de forma personalitzada.

Cal recordar que el cens de població és la principal font de dades demogràfiques i és una eina molt valuosa per a la planificació social i econòmica. Recull informació de la totalitat dels habitants en una data concreta, es a dir, —les dades són de tipus transversal—, dóna com a resultat el que s'acostuma a descriure com «la fotografia» de la població en un moment determinat.

El qüestionari del Cens de 1991 va incloure 31 preguntes relatives a les característiques de les persones i 6 temes en relació a l'habitatge.

Els principals aspectes que investiga es refereixen a:

a) característiques geogràfiques: lloc de residència
b) característiques demogràfiques: sexe, edat, estat civil, lloc de naixement, nacionalitat, relació de parentiu, fecunditat, data del casament, migració, mobilitat.
c) característiques socials o culturals: nivell d'instrucció, estudis en curs, coneixement del català.
d) característiques econòmiques: relació amb l'activitat, sector d'activitat, situació professional, professió o ocupació.
e) característiques de l'habitatge: data de construcció, règim de tinença de l'habitatge, superfície, nombre d'habitacions, instal·lacions de l'habitatge.

4. TRACTAMENT DE LES DADES

La informació recollida en el Cens de població no és uniforme per a la totalitat de la població. Hi ha alguns temes que solament poden tenir resposta per part d'un sector de la població, així per exemple els relatius a l'activitat econòmica es refereixen a la població de 16 anys i més, el nivell d'instrucció a les persones de 10 anys i més, el coneixement del català a les de 2 anys i més, la fecunditat a les dones de 12 anys i més, etc. També hi ha altres que es refereixen al conjunt de la llar com ara les referides a les característiques de l'habitatge.

Per a l'aplicació de les tècniques de la mineria de dades, ha tingut lloc un procés d'anàlisi, de prova i selecció de les variables.

S'han descartat algunes variables perquè presenten un comportament molt semblant en el conjunt de la població, no discriminen, com són algunes de les referides a les instal·lacions de l'habitatge. En altres casos, a pesar del baix pes relatiu de determinades variables, aquestes constitueïen un element important a l'hora de determinar grups específics (nínxols), per exemple les referides a la població estrangera que en aquest any representa l'1%.

Per a un millor tractament i posterior anàlisi algunes variables continues es van agrupar en intervals (edat, any d'arribada a Catalunya, data de construcció de l'habitatge) és a dir, es van transformar en discretes i altres en variables categòriques com és el canvi de municipi de residència en els últims deu anys, com indicador de migració en el decenni anterior al cens.

També es van combinar variables com per exemple, relació amb l'activitat amb situació professional i es van crear noves variables com el nombre de persones per llar. Finalment, les variables utilitzades, presentades segons els temes investigats pel cens, són les següents:

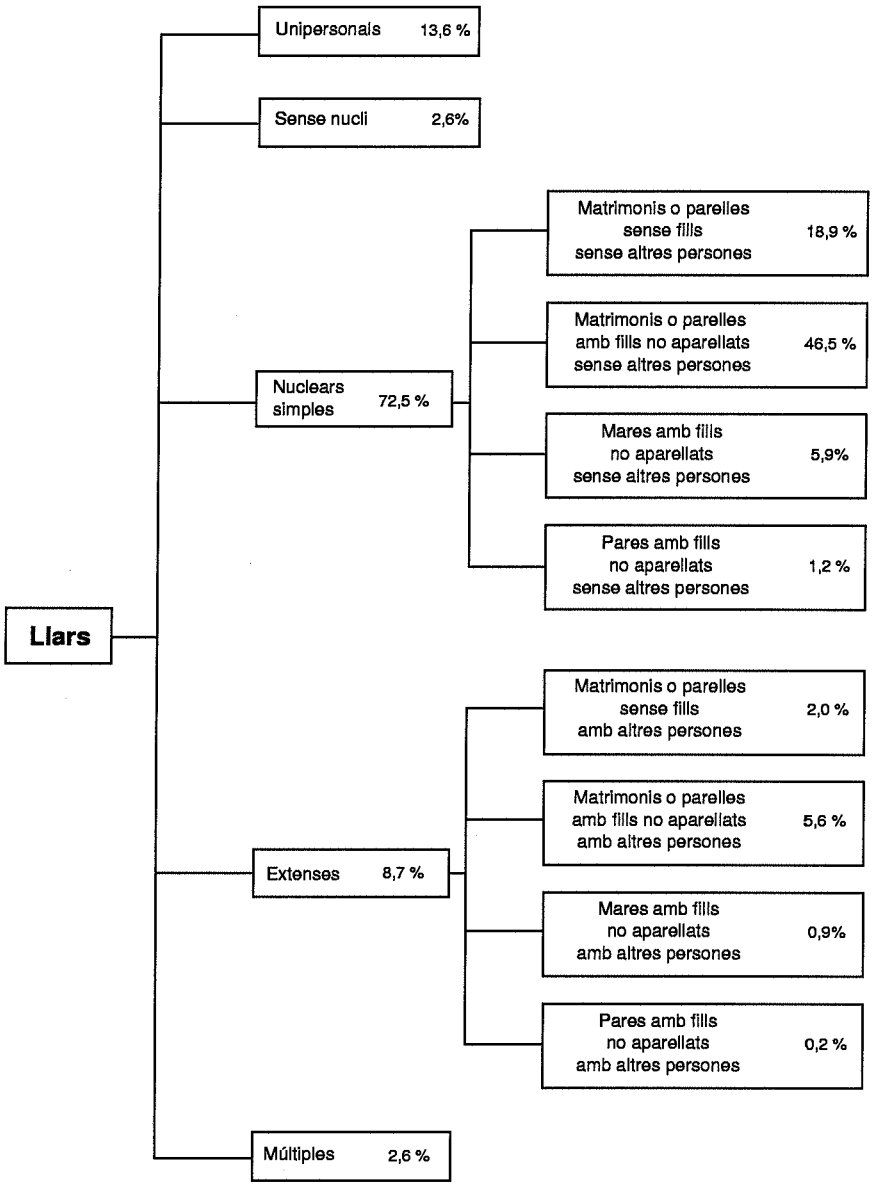
a) característiques geogràfiques i demogràfiques: lloc de naixement, edat, sexe, estat civil, nacionalitat, província de residència, canvi de residència als últims deu anys, any d'arribada a Catalunya, formes de convivència, nombre de persones per llar.

b) característiques socials o culturals: nivell d'instrucció, estudis en curs, coneixement del català.

c) característiques econòmiques: relació amb l'activitat, sector d'activitat, situació professional, professió o ocupació, mitjà de transport utilitzat per anar a treballar o estudiar.

d) característiques de l'habitatge: superfície de l'habitatge, nombre d'habitacions per llar, any de construcció de l'habitatge, règim de tinença de l'habitatge.

Quadre 1. Tipologia de llars.



Les variables referides a llars i famílies no es poden aplicar de manera directa com es recull al cens, per tant requereixen abans de ser utilitzades una ordenació, agrupació i classificació d'acord a una tipologia determinada. En aquest cas s'aplica una tipologia estàndard[1].

Les tècniques de segmentació i d'associació s'han aplicat als individus segons el tipus de llar al que pertanyen. L'interès per aquesta aproximació té la seva justificació en el fet de que molts dels comportaments demogràfics i una part important dels socials i econòmics estan sota la influència d'aquestes formes d'agrupació. D'altra banda, l'Institut d'Estadística de Catalunya, per primera vegada, ha obtingut informació detallada de les estructures familiars catalanes per a qualsevol àmbit territorial a partir de les dades del Cens de població i habitatge de 1991.

Per tal de situar en la seva perspectiva els resultats aquí presentats, convindrà fer un breu recordatori de l'estructura de les llars catalanes l'any 1991. El Cens de població va registrar aquest any 1.933.144 llars familiars on vivien 5.982.674[2] persones. Les característiques més notables d'aquestes llars (quadre 1) són, per una banda, l'elevada nuclearitat (73%) i, d'altra, un significat grau de complexitat familiar, reflectit en la presència de llars extenses i múltiples (11%). Tot i així, s'ha d'assenyalar que el pes de les llars unipersonals (14%), cada vegada més important, és un indici dels canvis que tenen lloc dins la composició familiar. El model dominant de família nuclear és el resultat d'un elevat percentatge de parelles amb fills (47%) i sense fills (19%).

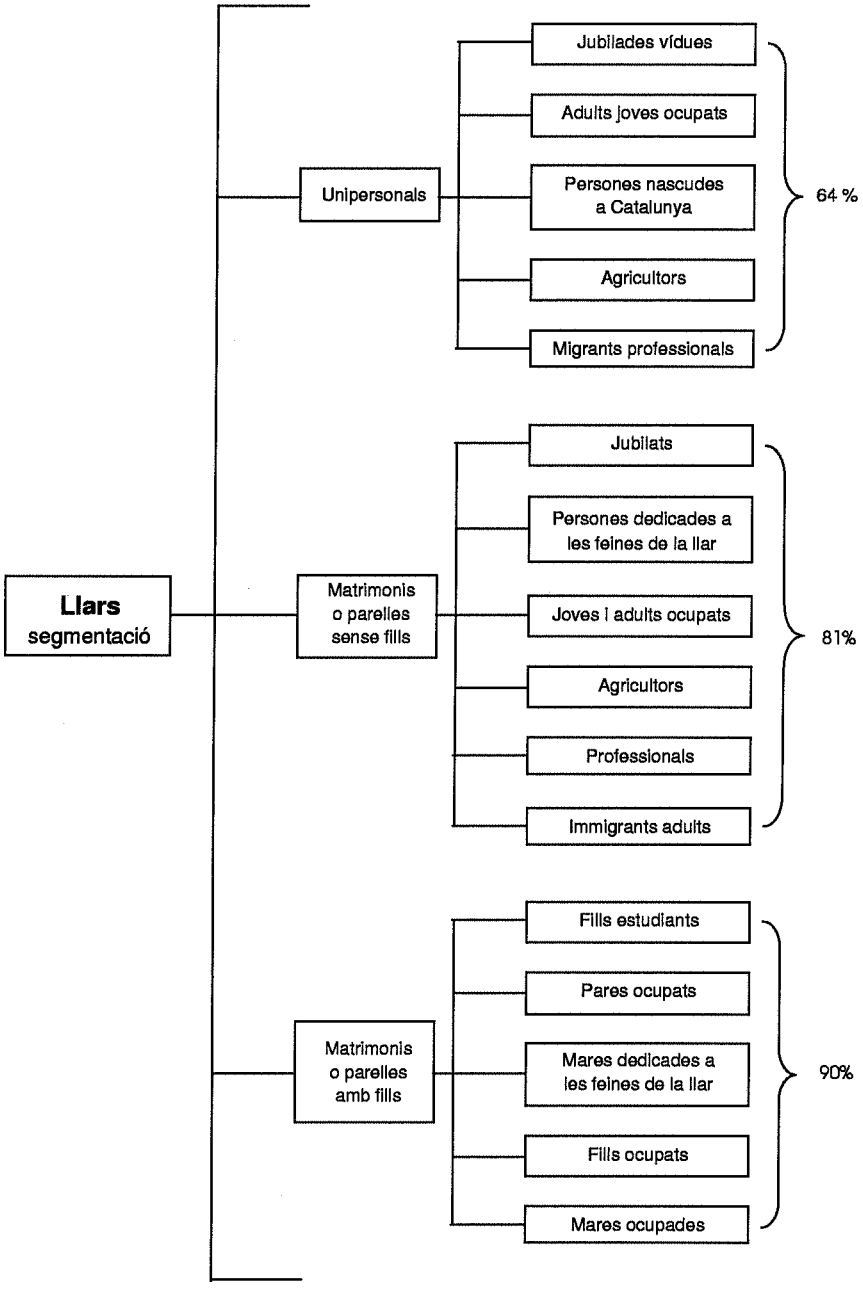
5. SEGMENTACIÓ: PRINCIPALS RESULTATS

Del conjunt d'estructures familiars es presentaran els resultats més rellevants obtinguts de l'aplicació de la tècnica de segmentació a les llars formades per una sola persona –les unipersonals–, i de les nuclears simples formades per matrimonis o parelles sense fills i matrimonis o parelles amb fills. Els tres tipus representen el 79% de les llars i concentren el 75% de la població de Catalunya a l'any 1991.

[1]La tipologia de llars que s'ha utilitzat té com a base, fonamentalment, la identificació i quantificació dels nuclis familiars i la presència o absència d'altres persones. Els nuclis poden estar composats per un matrimoni o parella sense fills, un matrimoni o parella amb fills, i nuclis formats per un sol progenitor, denominats també monoparentals. Així es distingeixen cinc tipus de llars: les unipersonals, formades per persones que viuen soles, les sense nucli, constituïdes per dues o més persones emparentades o no però que no formen un nucli familiar; les nuclears simples formades per nuclis en absència d'altres persones; les nuclears extenses compostes per nuclis amb la presència d'altres persones i per últim, les múltiples constituïdes per dos o més nuclis familiars.

[2]Aquesta dada no inclou les persones residents en llars de col·lectius.

Quadre 2. Resultats de la segmentació.



En una primera visió de conjunt (quadre 2) pot apreciar-se, pel que fa a les llars unipersonals, com els cinc segments reunits al quadre agrupen el 64% d'aquest tipus de llar; tot i que, com s'examinarà més endavant, el pes de cada segment és desigual amb una major freqüència de llars formades per dones jubilades vídues.

En el cas dels matrimonis sense fills, un nombre de segments molt proper a l'anterior —sis— agrupen un percentatge de llars més elevat, del 81%. També es constatarà la desigual distribució interna on els tres primers d'aquests segments comprenen la major part d'aquell valor, el 74%.

Finalment, respecte els matrimonis o parelles amb fills, un total de cinc segments tornen a reunir un percentatge molt elevat de llars d'aquest tipus, el 90%. S'aprecia, a l'igual que en el cas dels matrimonis sense fills, la presència d'un agrupament de població més homogeni, format ara per fills d'estudiants, pares ocupats, mares dedicades a les feines de la llar, fills ocupats i mares ocupades.

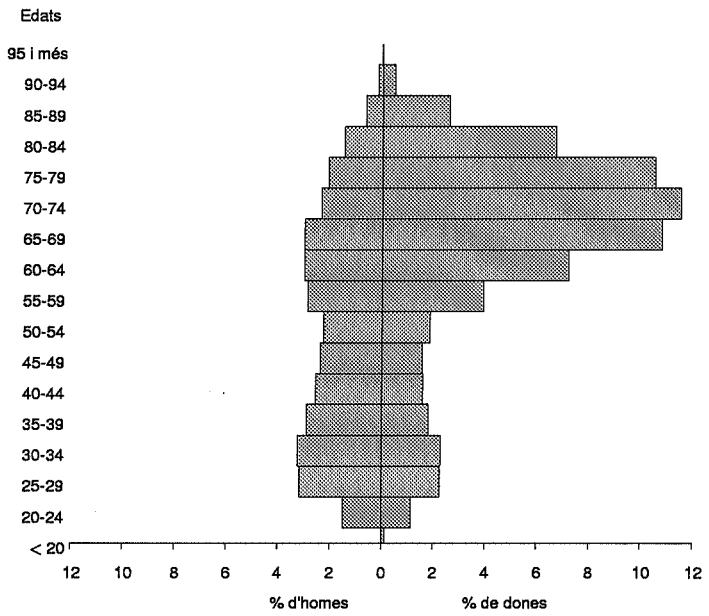
5.1. Llars unipersonals

L'estructura per edat i sexe de la població que viu a les llars unipersonals presenta una distribució dels seus efectius molt irregular (gràfic 1). Les diferències per edat i sexe són el resultat de diferents factors, entre els quals es destaquen, les distintes edats en què homes i dones comencen una relació de parella, l'efecte de les separacions i divorcis i les diferències de mortalitat entre homes i dones.

En aquesta estructura queden representats, com a mínim, tres moments del cicle de vida de les famílies. La solitud a edats joves, quan generalment aquesta és una elecció i en molts casos no resultarà definitiva, marca el moment de l'emancipació familiar. A les edats adultes, especialment pels homes, és el resultat d'una ruptura familiar i a les edats més elevades quan és fruit de la imposició vital, que es manifesta en la contracció del nucli familiar per mort d'algun del dos cònjuges, de manera més freqüents, dels marits.

Aquesta composició per edats condicionarà de manera significativa la determinació dels segments. A partir de l'aplicació d'aquest tipus de tècnica es constata (quadre 3) que la piràmide d'edats ha quedat «fragmentada» en diferents grups, el primer dels quals representa el 43% de la població, mentre que el següent més allunyat del primer agrupa a un 12% i els restants són petits segments amb reduït pes dins del conjunt de llars.

El primer segment representa fonamentalment la fase de contracció del nucli familiar. Es composa de dones vídues pensionistes que majoritàriament no tenen estudis o el nivell d'instrucció correspon a estudis primaris. El coneixement del català guarda estreta



Font: Institut d'Estadística de Catalunya (1994). *Cens de població 1991. Vol. 17.*

Gràfic 1. Estructura de la població de les llars unipersonals. Catalunya 1991.

Quadre 3. Llars unipersonals.

43%	Jubilades vídues
12%	Adults joves ocupats
5%	Persones nascudes a Catalunya
3%	Agricultors
1%	Migrants professionals

relació amb la naturalesa de la població. Aproximadament la meitat van néixer a Catalunya i les immigrants van arribar a aquesta Comunitat Autònoma abans del 1960 o entre 1960 i 1970. En conjunt la població d'aquest segment, no ha canviat de municipi de residència als últims deu anys. Aquestes dones viuen en habitatges de propietat, de construcció antiga, i resideixen principalment a la província de Barcelona.

El segon grup representa el 12% de les llars unipersonals. Està format de manera equilibrada per homes i dones, amb edats compreses entre els 30 i 50 anys, predominen les persones solteres, i en segon lloc els separats o divorciats. Presenten un nivell d'instrucció alt (secundaris i titulació superior), treballen al sector serveis, com professionals o tècnics i són assalariats amb caràcter fix. Van néixer a Catalunya, resideixen principalment a la província de Barcelona i viuen en habitatges de lloguer. El 10% ha canviat de municipi de residència els últims deu anys.

El tercer segment, amb un 5% del total de llars unipersonals es composa per homes i dones que van néixer a Catalunya. L'estat civil predominant és el de solter o vidu, fonamentalment són pensionistes, amb estudis primaris o secundaris, amb alt coneixement de la llengua catalana, i resideixen a la província de Barcelona en habitatges de lloguer.

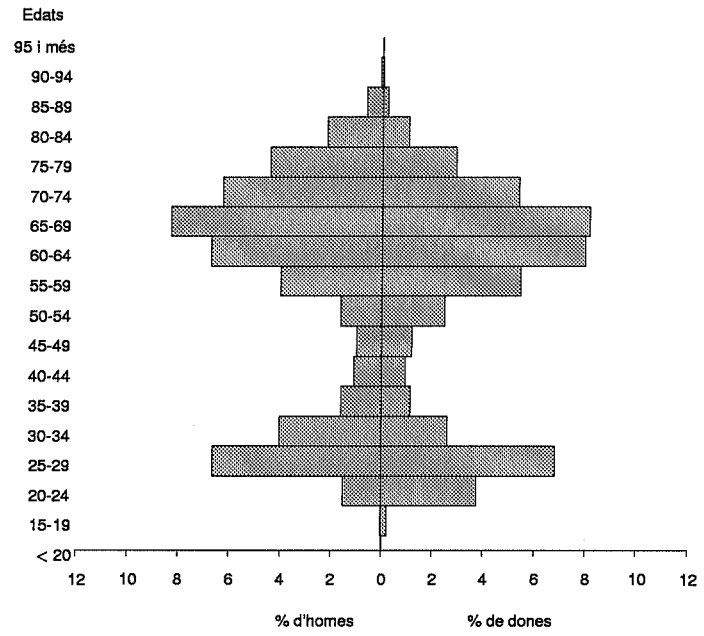
Dels restants segments, s'han de destacar al menys dos: el primer representa als agricultors, que és un grup que tornarà a aparèixer en els següents tipus de llars i el segon, els migrants professionals.

El segment dels agricultors es composa per un 3% de les persones de les llars unipersonals. Són homes i dones de 65 anys i més, vidus o solters que van néixer a Tarragona, Lleida i Girona, tenen estudis primaris, declaren un coneixement del català alt, resideixen en habitatges de propietat i no han canviat de municipi als últims deu anys.

El segment que representa als migrants professionals, constitueixen l'1% del total de llars unipersonals. Són homes solters, amb titulació superior i alta mobilitat als últims deu anys. Tenen edats entre els 30 i 50 anys, estan ocupats, són professionals i tècnics, el 71% van néixer a la resta d'Espanya i el 3% són de nacionalitat estrangera. Resideixen principalment a la província de Barcelona, viuen en habitatges de lloguer, han arribat a Catalunya en l'última dècada i el seu coneixement de català és alt.

5.2. Matrimonis o parelles sense fills

Mostren una estructura per edats molt envellida com a conseqüència de la falta d'infants. La piràmide (gràfic 2) presenta dos sortints molt marcats: un pels voltants dels 25-29 anys i un altre entre els 65-69 anys. Aquests valors modals representen dos moments del cicle de vida familiar. El primer, correspon a les edats joves i assenyalaria la fase de



Font: Institut d'Estadística de Catalunya (1994). *Cens de població 1991.*

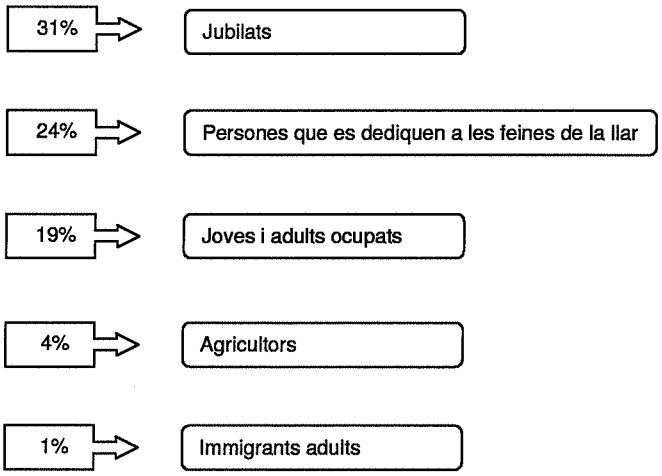
Gràfic 2. Estructura de la població de les llars de parelles sense fills. Catalunya 1991.

formació de la família i el segon, a les edats grans, la fase de contracció, és a dir, quan els fills han marxat de la casa dels pares. La variació d'aquests grups estarà en funció de diferents factors entre els quals s'han de destacar el calendari de la nupcialitat (o de la formació de la parella), de la fecunditat, de l'edat d'emancipació dels fills i del descens de la mortalitat.

Dels resultats obtinguts s'ha de destacar que els tres primers segments representen el 74% del total de la població dels matrimonis o parelles sense fills. El primer el composen els homes jubilats, el segon les dones dedicades a les feines de la llar i el tercer, les persones joves i adultes ocupades. Entre els restants segments, de menor importància relativa destaquen els agricultors, els professionals i els immigrants adults (quadre 4).

El primer dels segments esmentats inclou el 31% de la població d'aquest tipus de llar. Conté una alta representació d'homes casats, pensionistes, per tant persones de 65 anys i més, el 54% dels quals van néixer a Catalunya. Els que treballen o han treballat anteriorment, més de la meitat ho fan com treballadors industrials. El nivell d'estudis és baix, no han migrat als últims deu anys, viuen en habitatges de propietat, i resideixen majoritàriament a la província de Barcelona.

Quadre 4. Matrimonis o parelles sense fills.



El segon segment mencionat representa el 24% de la població i està compost fonamentalment per dones casades que treballen en les feines de la llar. Tenen edats de 65 anys i més o entre 51 i 64 anys, predominen les persones sense estudis o estudis primaris amb baix coneixement de la llengua catalana. No han canviat de municipi de residència els últims deu anys i majoritàriament viuen a la província de Barcelona en habitatges de propietat.

El tercer dels principals segments, amb un 19% de la població, es composa bàsicament per joves de menys de 30 anys, i persones adultes-joves d'entre 31 i 50 anys, dels quals més de la meitat són homes. Van néixer a la província de Barcelona, declaren un nivell molt alt de català i presenten una titulació més elevada que els casos anteriors: secundaris i superiors. Majoritàriament estan ocupats al sector serveis, es desplacen a treballar o a estudiar en cotxe o taxi i la seva situació professional correspon a una persona que treballa amb un contracte fix. No han deixat mai de residir a Catalunya i viuen a la província de Barcelona en habitatges de compra, que han estat construïts entre 1960-1986. Un elevat percentatge han canviat de municipi als últims deu anys i hi ha un 8% de persones solteres que formen parelles de fet.

En aquest tipus de llar, també destaquen com segments diferenciats els corresponents als agricultors, als professionals i als immigrants adults.

Els agricultors representen el 4% de la població. La variable que els determina en primer lloc és el sector d'activitat i la professió o ocupació. Treballen en l'agricultura i la seva situació professional és d'empresaris o treballadors per comte propi que no contracten personal. Són majoritàriament homes, casats, el 80% són persones majors de 65 anys, que van néixer a Catalunya, actualment viuen a Tarragona, Lleida i Girona,

en habitatges de propietat. Predominen les persones amb estudis primaris encara que és significatiu el percentatge de les sense estudis, el coneixement de català és molt alt i no han canviat de municipi als últims deu anys.

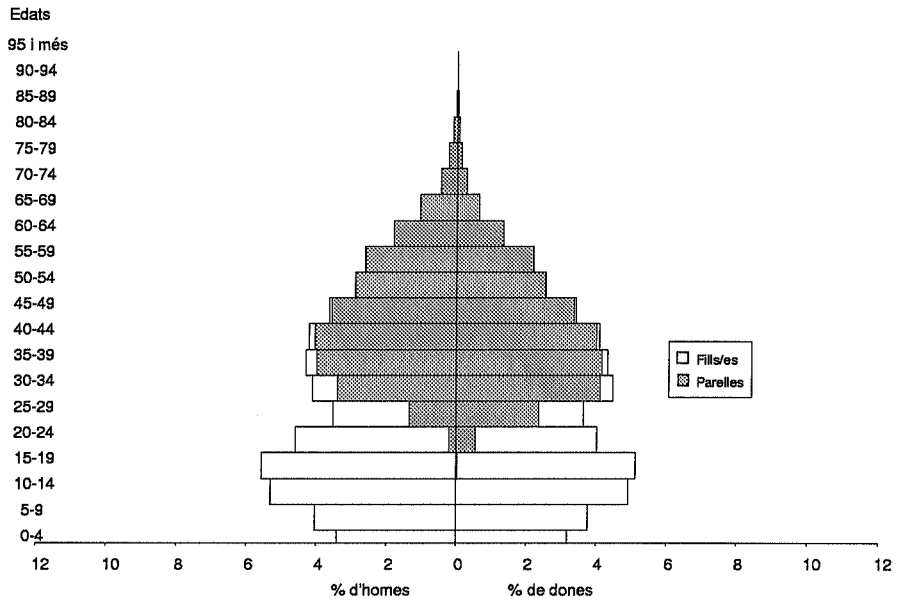
El segment que representa als professionals, inclou al 2% de la població. Ho formen adults joves (entre 30 i 50), dels quals més de la meitat són dones. El 9% de les persones són estrangeres i una alta proporció són parelles de fet (27%). El 92% tenen estudis superiors, són professionals o tècnics que treballen al sector serveis amb caràcter fix i utilitzen el cotxe o taxi per anar a treballar. La meitat han canviat de municipi de residència els últims deu anys, han viscut fora de Catalunya, tenen un coneixement molt alt de català i aproximadament la meitat van néixer a aquesta Comunitat Autònoma. Viuen en la província de Barcelona, en habitatges de recent construcció, de lloguer o en propietat amb pagament pendent amb una superfície d'entre 90 i 110 metres quadrats.

Finalment els immigrants adults representen l'1% del conjunt de matrimonis o parelles sense fills. Presenten una alta mobilitat, ja que el 72% han canviat de municipi als últims deu anys. Predominen els homes, pensionistes amb un nivell d'instrucció correspon a titulació secundària i superior. Aproximadament, la meitat del segment el composen persones de nacionalitat estrangera, dels quals més de la quinta part són de la Unió Europea. Han arribat a Catalunya entre 1981 i 1991, resideixen fonamentalment a la província de Girona i viuen en habitatge de propietat. Les persones que treballen ho fan al sector serveis, amb caràcter fix.

5.3. Matrimonis o parelles amb fills

Representen el grup més nombrós de tota la població amb un 59% dels seus habitants. La forma de la piràmide (gràfic 3) guarda relació amb el tipus d'estructura de la llar, és a dir, el fet que siguin parelles que conviuen amb els seus fills fa que el pes de les persones grans resulti molt baix. S'aprecia, també, una reducció de la piràmide al voltant dels 30 anys, edat que correspon a l'emancipació dels fills, o a la formació de noves parelles. De fet la piràmide pot dividir-se en dues parts: una primera comprèn des de la base fins als 25 anys i són majoria els fills i la segona, a partir d'aquesta edat, ocupada fonamentalment pels pares.

Aquest tipus de llars és el més homogeni dels investigats, ja que cinc segments reuneixen el 90% de la població, ells són: els fills estudiants, els pares ocupats, les mares dedicades a les feines de la llar, els fills ocupats i les mares ocupades (quadre 5).



Font: Institut d'Estadística de Catalunya (1994). *Cens de població 1991.*

Gràfic 3. Estructura de la població de les llars de parelles amb fills. Catalunya 1991.

Quadre 5. Matrimonis o parelles amb fills.

35%	Fills estudiants
21%	Pares ocupats
15%	Mares dedicades a les feines de la llar
11%	Fills ocupats
8%	Mares ocupades

El primer segment, representa el 35% de la població que viu a les llars de matrimonis o parelles amb fills i està compost pels fills estudiants. La variable més significativa en la determinació del segment és l'edat: el 70% són menors de 16 anys i el 30% restant són joves d'edats compreses entre els 16 i 30 anys. Són fills, homes i dones, estudiants, solters, que es caracteritzen en què majoritàriament van néixer a Catalunya (76% a la província de Barcelona), no han migrat d'aquesta Comunitat Autònoma, viuen en llars de 4 o més persones, presenten un coneixement de català molt alt, viuen en habitatges en propietat, són de nacionalitat espanyola, el 75% resideixen a la província de Barcelona.

S'ha de destacar que si bé és un segment representatiu dels infants i adolescents, hi ha un 30% de joves entre 16 i 30 anys que tenen les mateixes característiques que els infants. També s'ha d'indicar que és l'únic segment on el sexe és la variable menys important en la determinació del mateix.

El segon segment correspon als pares ocupats, amb un 21% de la població d'aquest tipus de llar. Són homes, pares, majoritàriament d'edats compreses entre els 31 i 50 anys, casats, ocupats, que treballen amb caràcter fix, al sector d'indústria-energia, com treballadors industrials. Aproximadament la meitat van néixer i no han deixat de residir a Catalunya, sols el 8% van canviar de municipi als últims deu anys. El nivell d'instrucció predominant és el d'estudis primaris, seguit pel secundari, el coneixement de català és alt, viuen principalment en llars de 4 persones, en habitatges de propietat, construïdes entre 1960-1986 i resideixen a la província de Barcelona.

El tercer segment, el formen les mares dedicades a les feines de la llar amb un 15% de la població de les llars de matrimonis o parelles amb fills. Són dones casades, principalment d'edats compreses entre els 31 i 50 anys i en menor proporció d'entre 51 i 64 anys. Treballen a la llar en feines no remunerades, presenten un nivell d'estudis primaris o sense estudis i un alt coneixement de la llengua catalana. Més de la meitat d'aquestes persones no van néixer a Catalunya i les que són immigrants van arribar a aquesta Comunitat Autònoma entre 1961-1970. Viuen en llars de 4 o més persones, en habitatges de propietat, construïdes entre 1960-1986, a la província de Barcelona. Sols un 10% han canviat de municipi de residència els últims deu anys.

El quart segment, amb un 11% de la població de les llars de matrimonis o parelles amb fills representa als fills ocupats. Aquests fills, tenen edats compreses entre els 16 i 30 anys, l'estat civil és el de solters i el 60% són homes. Respecte a l'activitat econòmica, són assalariats que treballen amb caràcter eventual (52%) o fix (39%), principalment al sector serveis. El nivell d'estudi predominant és el de secundari, declaren un coneixement molt alt de català i un 16% estan cursant estudis. Fonamentalment són catalans que no han canviat de municipi de residència en els últims deu anys, viuen en llars de 4 o més persones, en habitatges de propietat i el 82% resideixen a la província de Barcelona.

El cinquè segment, està integrat per mares ocupades i representen el 8% de la població de les llars de matrimonis o parelles amb fills. Són dones, mares, casades, de 31 a 50 anys, que estan ocupades, treballen al sector serveis, com assalariades amb caràcter fix, la distribució de les professions és molt homogènia encara que predominen les administratives. El nivell d'estudis amb major pes és el secundari, el coneixement de català és alt, viuen en llars de 3-4 persones. El 63% van néixer a Catalunya, la mobilitat dels últims deu anys és baixa. Resideixen a la província de Barcelona en habitatges de propietat.

6. ASSOCIACIONS: PRINCIPALS RESULTATS

Aquesta tècnica permet conèixer l'associació entre diferents elements. D'una banda, s'obté la intensitat de la relació i d'altra, la freqüència amb que es detecta aquesta regla d'associació en el conjunt que s'analitza. En el cas de les dades censals, la recerca d'associacions s'ha de fer en cada llar per cada variable per separat. Això permet obtenir les associacions entre les diferents categories de la variable escollida.

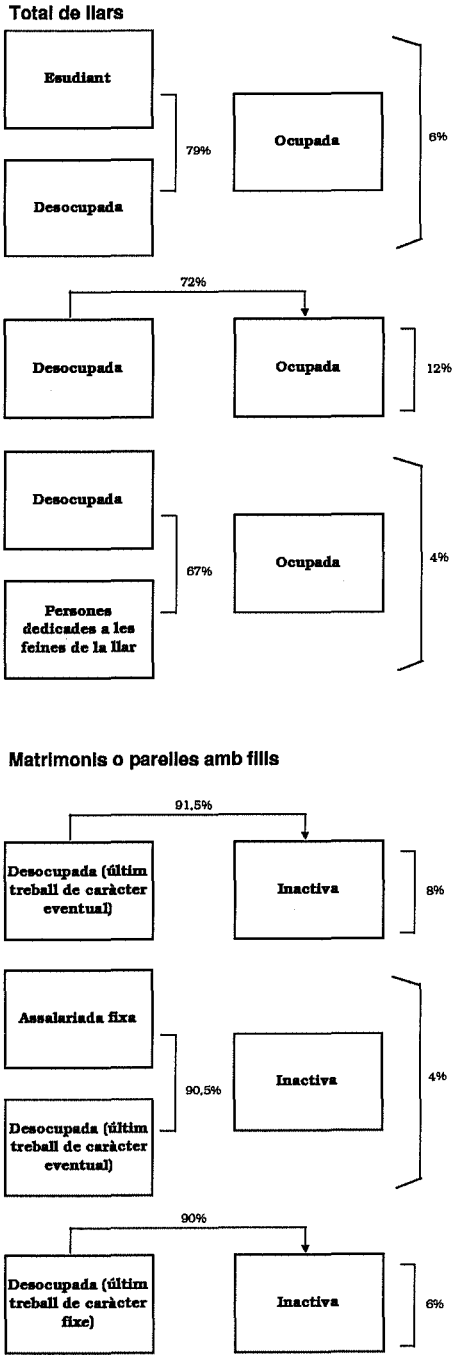
Dins les limitacions que en l'aplicació d'aquesta tècnica convé assenyalar és que l'associació només pot fer-se per cada variable per separat i no detecta els casos en què en una mateixa llar es repeteixen algunes categories de la variable. Aquestes restriccions es podrien solucionar creant unes noves que resultaran de la combinació de diferents variables, o que permetin registrar el nombre de vegades que alguna categoria de la variable en qüestió es repeteix.

Les dades censals presenten la particularitat que moltes de les associacions possibles o són conegudes o tenen un escàs valor analític. Per exemple si s'analitza la variable estat civil, per una banda es podria esperar en llars on hi ha fills, associacions del tipus casats (els pares)-solters (els fills), i d'altra banda, no enregistraria els casos on a les llars hi ha dues o més persones del mateix estat. Tampoc resultarien de major interès associacions del tipus casat/casada-vidu/vídua, atès que ja acostumen a estar prou tractades en els estudis sobre pautes familiars de convivència i/o corresidència.

En aquesta ocasió, per a l'aplicació d'aquesta tècnica es va seleccionar la variable «Relació amb l'activitat econòmica». L'estudi de l'activitat econòmica de les persones a nivell de les llars, ha estat sovint assenyalada com una de les millors vies per conèixer la incidència de l'atur a les famílies, avaluar el grau de dependència familiar i poder analitzar com el grup familiar col·labora en la reducció dels efectes de l'atur. També s'ha creat una variable que combina la relació amb l'activitat econòmica amb la situació professional.

Al moment d'aplicar aquesta segona tècnica no es va poder seguir el mateix enfocament utilitzat per la segmentació. Així, resulta del tot impossible el seu ús per a les llars

Quadre 6. Resultats de les associacions.



unipersonals i té molt poc sentit aplicar-la a les formades per matrimonis o parelles sense fills i sense altres persones, pel fet d'estar constituïdes per dues persones. Per aquestes raons s'han buscat associacions per al conjunt de llars de Catalunya que eren 1.933.044 i per al tipus de llars de major importància relativa que són els matrimonis o parelles amb fills que agrupen el 46,5% de les llars.

Al quadre 6 són reunits els resultats principals d'aquesta aplicació per als dos grups de llars esmentats. Així, pel que fa al total de llars catalanes es presenten les associacions amb una major incidència, en ordre decreixent. En aquestes tres s'observa l'existència d'una persona desocupada relacionada amb una altra en diverses situacions respecte l'activitat, tot indicant també el percentatge de llars on aquesta associació té lloc.

En un 6% de les llars catalanes, la presència d'un desocupat s'acompanya el 79% dels casos d'un estudiant i ambdues situacions són relacionades amb una persona ocupada. Aquest darrer status associat ara a l'individu desocupat assoleix un valor del 72% i afecta a un 12% de les llars. Convé assenyalar com aquest resultat, tot i que acotat al moment censal de 1991, il·lustra l'abast d'una tipologia social relativa als efectes de l'atur en les famílies. En aquest mateix sentit pot apreciar-se com en un 4% de les llars, un 67% dels casos la persona desocupada s'associa a una dedicada a les feines de la llar, sempre amb la presència d'un membre empleat. Aquestes dues situacions descrites, poden considerar-se en concordança amb allò observat, a partir d'altres fonts estadístiques, pels estudis sobre la composició de les famílies, segons la relació dels seus membres amb l'activitat. En particular, el fet que a moltes de les llars amb alguns membres desocupats com a mínim un s'ha mantingut empleat en les etapes de major desocupació de finals dels anys vuitanta i primers de la dècada dels noranta.

Pel que fa als matrimonis o parelles amb fills, el quadre 6 ofereix unes modalitats d'associació molt intensa, al voltant del 90% dels casos i totes relacionades amb persones declarades inactives, en aquest cas, mestresses de casa o fills estudiants. En un 8% d'aquest grup de llars, l'individu desocupat, però amb un darrer treball eventual, és present en el 91,5% de casos amb un d'inactiu. Un percentatge molt proper presenta en el 6% de les llars, l'associació relativa ara a un últim treball de caràcter fix. Totes dues situacions semblarien detectar grups amb certes dificultats en relació als efectes de l'atur i es diferencien de la tercera obtinguda, on si bé en un nombre menor de llars, el 4%, ara la persona desocupada s'associa a una d'assalariada, amb un treball fix.

7. CONSIDERACIONS FINALS

Fins ara, les aplicacions de Data Mining s'han efectuat principalment amb informació del sector financer, hospitalari, farmacèutic, serveis, entre altres, com també s'ha utilitzat per extreure informació de texts. En canvi, menys freqüent havia estat la seva aplicació al tractament de grans bases de dades sociodemogràfiques com pot ser un cens

de població. Gràcies a la col·laboració de l'Idescat, el Cesca i IBM, s'ha pogut realitzar l'aplicació de Data Mining a l'àmbit de l'estadística oficial. Per les característiques de les dades solament es van aplicar les tècniques de segmentació i d'associació.

Com a primer aspecte d'interès fruit d'aquests procediments pot afirmar-se que, amb la mineria de dades, s'ha tingut l'oportunitat de treballar amb la totalitat de la població (6.000.000 de registres), s'ha fet una lectura integral de la informació i s'han determinat grups de dades homogènies. Així, s'ha pogut resumir i sintetitzar un gran volum d'informació i també quantificar la intensitat de relacions entre les variables.

Convé assenyalar que moltes de les variables que recull el cens de població presenten un comportament condicionat al cicle vital de les persones i/o de les famílies. L'edat i el sexe són atributs fonamentals, i per això, alguns dels resultats obtinguts eren previsibles. Ara bé, l'aplicació de la segmentació ha constatat que de vegades l'edat i/o el sexe són les variables que defineixen el segment, però en altres casos la naturalesa de la població, l'ocupació, el sector d'activitat, etc., són aquelles que han tingut el major pes principal en la determinació del mateix.

L'aproximació clàssica en l'anàlisi de la informació censal acostuma a partir de l'estructura per sexe i edat de la població i, sobre aquesta base, continua amb l'estudi de les característiques demogràfiques, socials i econòmiques associades. En aquesta ocasió, mitjançant la utilització d'aquestes tècniques ha estat possible operar simultàniament amb un major nombre de variables i poder identificar modalitats d'agrupaments més ajustats a la natura heterogènia que caracteritza tota població. Aquesta nova perspectiva cal considerar-la com una anàlisi no convencional però complementària de l'estadística oficial.

L'Institut d'Estadística de Catalunya té previst continuar en aquesta línia de treball amb l'aplicació de la mineria de dades a la informació que proporciona l'Estadística de població de 1996 i les dades del proper cens del 2001, que permetrà aprofundir en l'anàlisi de la dimensió temporal i territorial de les estadístiques censals.

Annex: breu descripció de les operacions de Mineria de Dades

Segmentació

Es tracta d'agrupar automàticament registres que comparteixen característiques similars. Per exemple ajuda a saber quines agrupacions lògiques existeixen en el cens de Catalunya segons les característiques o el comportament de les persones estudiades.

La segmentació es basa en dues tècniques de les que es detallen les seves característiques fonamentals:

- Segmentació basada en Anàlisi Relacional:

– El número de segments es determina durant l'execució de l'algoritme.
– Processa tant variables quantitatives com qualitatives.
– Maximitza la similaritat entre els membres d'un mateix segment i les diferències entre els membres de segments diferents.
– Segmenta en base a mètriques de similaritat, no de distància.
– És eficient per a la detecció de conjunts molt petits de registres.

- Segmentació Neuronal:

– És necessari predefinir el número de segments que es vol obtenir, així com la seva distribució bidimensional.
– Processa tan variables qualitatives com quantitatives, però funciona millor amb les darreres.
– És eficient quan es vol particionar una població imposant certa relació entre els segments obtinguts.

Classificació

Es tracta de predir un resultat entre diferents possibilitats i d'explicar els factors de classificació. Per exemple saber quines característiques tindrà una llar catalana en la que els seus membres tinguin coneixement del català parlat i escrit.

La classificació es basa en dues tècniques que són els arbres de decisió i les xarxes neuronals i les seves principals característiques són:

- Classificació basada en Arbres de Decisió:

– Model de classificació en forma d'arbre de decisió binària.

- Treballa amb la tecnologia SLIQ (Supervised learning In Quest), processant tant variables quantitatives com qualitatives.
- Incorpora una tècnica de podat basada en el principi de la mínima longitud de descripció (MLD), que proporciona arbres més senzills.
- Es escalable i pot processar conjunts amb independència del número de classes, atributs i registres.

• Classificació Neuronal:

- Basada en xarxes neuronals de propagació cap endarrere.
- Detecta de forma automàtica la topologia més adequada per a cada problema. També permet especificar una concreta.
- Realitza una anàlisi de sensibilitat per a detectar les variables més significatives per a cada tipologia.

Predicció

Es tracta de predir un valor per un registre entre un conjunt continu de possibilitats. Per exemple saber quina probabilitat hi ha que una mesura sanitària concreta sigui necessària per les llars catalanes d'unes característiques determinades.

Les principals característiques de les dues tècniques en què es basa la predicció són les següents:

• Funcions de Base Radial:

- Poden processar variables quantitatives i qualitatives a la vegada.
- Detecta el número de centroides òptim, predefinint el número màxim d'aquests i el número mínim de registres assignats a cada centre.
- Funciona especialment bé quan l'estructura de les dades tendeix a agrupar-se en conjunts, ja que implementa cert tipus de segmentació.

• Predicció Neuronal:

- Basada en xarxes neuronals de propagació cap endarrere.
- Detecta de forma automàtica la topologia més adequada para cada problema, encara que permet especificar una concreta.
- Permet predir dades en forma de sèries temporals.
- Permet implementar regressió logística.

Anàlisi de Relacions

Es tracta relacionar característiques o comportaments que estan relacionats entre si. Per exemple saber quines variables soci-cultural-econòmiques van normalment juntes.

L'anàlisi de relacions es basa en tres tècniques de les que es detallen les seves característiques fonamentals:

- Anàlisi d'Associacions:

– Detecta elements en una transacció que impliquen la presència d'altres elements en la mateixa.
– Determina les afinitats entre elements en forma de regles d'associació XY, facilitant una sèrie de mètriques com son el suport, la confiança, el tipus de la regla, etc.
– Permet incorporar taxonomies de productes, permetent la detecció d'associacions a diferents nivells.

- Patrons seqüencials:

– S'introdueix la variable temps. Es detecten patrons entre transaccions, el que permet determinar elements en una transacció que impliquen la presencia d'altres elements en una altra transacció.

- Anàlisi de similaritat en series temporals:

– Detecta totes les ocurrències de seqüències similars en una col·lecció de sèries temporals.

Com a complement a les operacions/tècniques de mineria de dades el producte «DB2 Intelligent Miner for Data» incorpora les següents funcions de preprocés i funcions estadístiques que van ser de gran utilitat en el desenvolupament del projecte.

- Funcions de preprocés de dades:

– Recollida de dades: extraccions de bases de dades i d'altres fitxers.
– Selecció de dades: filtrat, preparació de mostres, agregació, projecció, agrupació i, resum de dades.
– Transformació: conversió, codificació, discretització de dades.
– Fusió i enllaç de taules.
– Consistència: verificació, filtrat i descart de dades.
– Derivació: càlcul de nous descriptors a partir de les dades disponibles.
– Processat de sentències SQL contra bases de dades.

- Funcions estadístiques:

- Elaboració de descriptives univariants i bivariants; càlcul de percentils, ANOVA, i proves F.
- Anàlisi de components principals i anàlisi de factors.
- Ajust de corbes per a sèries temporals.
- Regressió lineal múltiple i regressió polinòmica.
- Anàlisis d'ocurrències.

BIBLIOGRAFIA

Adriaans, P. & Zantinge, D. (1996). *Data Mining*. Addison-Wesley: Harlow England.

Bigus, J. (1996). *Data Mining with Neural Networks*. McGraw-Hill: New York.

Cabena, Peter, *et al.* (1997). *Discovering Data Mining: from concept to implementation*. Prentice-Hall: New Jersey.

Hinde, A. (1998). *Demographic Methods*. Arnold: London.

Institut d'Estadística de Catalunya (1994). *Cens de població 1991*. Generalitat de Catalunya: Barcelona.

Institut d'Estadística de Catalunya (1997). *Llars i famílies a Catalunya 1991*. Generalitat de Catalunya: Barcelona.

Leridon, H. & Toulemon, L. (1997). *Démographie. Approche statistique et dynamique des populations*. Economica: Paris.

ENGLISH SUMMARY

POPULATION CENSUS: AN INTERPRETATION USING DATA MINING

CRISTINA GUISANDE ALLENDE*
FRANCESC SUBIRADA CURCO**

This article presents the results of a research job carried out jointly by the Institut d'Estadística de Catalunya (Idescat), the Centre de Supercomputació de Catalunya and IBM. They analysed data from the 1991 population and households Census using Data Mining methodology. Segmentation and association techniques were used on individuals according to the type of household they belonged to.

Keywords: Data Mining, official statistics, association, segmentation, data analysis, applied statistics, statistical methods

AMS Classification (MSC 2000): 62-07, 62P25, 62H30, 62H20

* Institut d'Estadística de Catalunya. Via Laietana, 58. 08003 Barcelona. E-mail: cguisande@idescat.es
** CEPBA-IBM Research Institute. Jordi Girona, 1-3. 08034 Barcelona. E-mail: frsubirada@es.ibm.com

- Received April 2001.
- Accepted November 2001.

1. INTRODUCTION

The Institut d'Estadística de Catalunya (Idescat), the Centre de Supercomputació de Catalunya (Cesca) and IBM signed a cooperation agreement to carry out a research on data obtained from the 1991 population and households Census using Data Mining. This article aims to present the main results of this experience, where Data Mining was applied on census information.

2. DATA MINING

The following items were used while conducting the research:

1. Softwarewise:

IBM's DB2 Intelligent Miner for Data. With this program, Data Mining operations can be performed on records in order to find out previously unknown information.

2. Hardwarewise:

IBM's System Parallel 2 (SP2). This parallel process computer allows the simultaneous use of several microprocessors to address a single problem. When using Data Mining it can deal with big databases in a reasonable amount of time.

The following operations can be performed: segmentation, classification, forecasting and analysis of relationships. Because of the nature of census data, only two of the aforementioned operations were carried out: segmentation and, as regards analysis of relationships, association.

3. CENSUS DATA

The main source of demographic data is the population census, and that makes it a very valuable tool for social and economic planning. It collects information about all the inhabitants on a given date –cross-section data– and it gives out the so called «picture» of the population at a given moment.

The survey for the 1991 Census included 31 questions relating to individuals and 6 others relating to housing.

The main objects of investigation were:

a) geographical aspects: place of residence.

b) demographic aspects: sex, age, marital status, place of birth, nationality, relationship of the dwellers, fertility, date of marriage, migration, mobility.

c) social or cultural aspects: education, knowledge of Catalan.

d) economic aspects: work, sector, professional status, occupation.

e) housing aspects: date of construction, property, area, number of rooms, fittings and appliances.

4. DATA PROCESSING

A process of selection, testing and analysis of the variables has to take place before using data mining techniques.

Some variables, like household equipment, have been discarded for they don't really discriminate, having always very similar values. Other variables, despite their relative low weight against the whole of the population, turned out to be important for determining specific groups (population niches), for example foreign population, which represented 1% of the total population.

In order to optimise the data processing, some continuous variables were grouped in intervals (age, date of arrival in Catalonia, date of construction of the house), thus turning discrete, while others were turned into categorical variables, for example the change of municipality of residence during the previous ten years as a migration indicator. Other variables were combined, like work and professional status, and yet other new ones were created, like number of individuals in the household.

Segmentation and association techniques have been applied on individuals according to the type of household to which they belong. What makes this approach interesting is that many demographic and some social and economic aspects are quite influenced by some types of household composition. Also, the Institut d'Estadística de Catalunya has, for the first time, obtained detailed information from the 1991 population Census about family structures for any region in the Catalonian territory.

5. SEGMENTATION: MAIN RESULTS

This article presents the most outstanding results obtained from applying segmentation to households formed by a single person and to simple family units formed by couples

with or without children. As it turns out, these three types represented 79% of all the households and 75% of the population of Catalonia back in 1991.

A quick overview –figure 2– reveals, as regards single-person households, that the five segments grouped in the figure represent as much as 64% of these households, although each segment has a different relative weight, the most frequent being households formed by retired widows.

Regarding couples without children, a similar number of of segments –six– represent an even higher percentage of households (81%). They also show an uneven distribution, with the first three segments answering for 74% of the total.

Finally, for couples with children a total of five segments once again stand for 90% of the total. As with couples without children, a more homogeneous population grouping –formed by students' children, working fathers, housewives, working sons or daughters and working mothers– is revealed.

6. ASSOCIATIONS: MAIN RESULTS

The use of this technique reveals associations between different elements. On one hand, we obtain a measure of the intensity of the relation and, on the other hand, the frequency with which this rule of association is detected. When it comes to census data, search for associations must be done separately for each variable in each household. By doing that, we obtain the associations between different categories of the chosen value.

It should be noted that, among other limitations of this technique, association can only be done on each saparate variable and it won't detect those cases where some categories of the variable are repeated in a household. These restrictions could be avoided by creating new ones, either by combining different variables, or by registering the number of times a given category of that variable is repeated.

The variable selected for the application of this technique was «Relationship to economic activity». The study of the individuals' economic activities householdwise has often been pointed out as one of the best markers for revealing the impact of redundancy on families, evaluate the degree of family dependence and analyse the way the family unit contributes to reducing the effects of redundancy. Another variable, which combines relationship to economic activity with professional status, has been created.

When it came to using this second technique, we couldn't use the same approach as we did with segmentation. Moreover, it is absolutely impossible to use it with households formed by a single individual, and makes little sense with couples without children or other dependants, since they're made of only two persons. Therefore, we have looked for associations for the most important type of households, which are those formed by

couples with children, amounting up to 46.5% of the 1,933,044 total households censed in Catalonia.

7. FINAL CONSIDERATIONS

So far, Data Mining had been applied mainly on information from the financial, hospital, pharmaceutical, and service sectors, as well as to extract information from texts. It had seldom been applied to large sociodemographic databases like a population census. Thanks to the cooperation between Idescat, Cesca and IBM, Data Mining has been used on official statistical information. Because of the nature of the data only segmentation and association techniques were used.

It should be noted that using these procedures (data mining) has given us the chance to work with the total population (6,000,000 records), achieve an integral reading of the information, and determine groups of homogenous data. Moreover, it has allowed us to summarize and synthetize a great volume of information, and also to quantify the intensity of the relatioships between variables.

The Institut d'Estadística de Catalunya plans to continue using data mining on information provided by 1996 population Statistics and on data from the 2001 Census, which will help us to further deepen the analisys of census statistics in both dimensions, time and space.

Secció Docent i Problemes

SECCIÓ DOCENT I PROBLEMES

La «Secció docent i problemes» té l'objectiu de publicar articles de caire docent, difícilment publicables en revistes de recerca. A cada número de *Qüestiió* s'inclouen d'un a tres problemes i les solucions es donen en el número següent.

Els lectors poden proposar problemes amb les solucions pertinents i enviar-los a *Qüestiió,* que farà una selecció i en publicarà els més adequats, fent la corresponent referència a l'autor.

També seran ben rebudes solucions alternatives a les propostes fetes per l'autor dels problemes. L'editorial es reservarà, però, el dret a publicar-les.

SOLUCIÓ AL PROBLEMA PROPOSAT AL VOLUM 25 N. 2

PROBLEMA N. 89

It is well known that $(n-1)S \sim W_p(V, n-1)$. See, e.g., Anderson (1958, sections 3.3 and 7.2). This means that this seemingly non-central Wishart variate is, in fact, a central Wishart variate. A quick way to see this is the following. Write $(n-1)S = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' = X'MX$, with $X := (x_1, \ldots, x_n)$, $M := I_n - n^{-1} 1_n 1_n'$, 1_n being an $(n \times 1)$ vector with n unit elements.

As M is symmetric idempotent, its Schur decomposition is $M = TT'$, with $T'T = I_{n-1}$ and $T' 1_n = 0$. This yields then $(n-1)S = Y'Y$, with $Y' := X'T$. Write $Y' = (y_1, \ldots, y_{n-1})$. Clearly $\mathcal{D}(\mathrm{vec}Y') = \mathcal{D}(\mathrm{vec}X'T) = \mathcal{D}[(T' \otimes I_p(\mathrm{vec}X')] = (T' \otimes I_p)$ $\mathcal{D}(\mathrm{vec}X')(T \otimes I_p) = (T' \otimes I_p)(I_n \otimes V)(T \otimes I_p) = T'T \otimes V = I_{n-1} \otimes V$. The $n-1$ vectors $y_1, \ldots, y_{n-1}$ are seen to be uncorrelated. Because of normality they are independent. Further $EY' = (EX')T = \mu 1_n' T = 0$. Given the definition of the central Wishart we conclude that $(n-1)S \sim W_p(V, n-1)$.

It is also well known that $E[(n-1)S]^{-1} = (n-p-2)^{-1}V^{-1}$ so that $ES^{-1} = (n-1)(n-p-2)^{-1}V^{-1}$.

See, e.g. Legault-Giguère (1974, Lemma B6) or Neudecker (2001).

In a recent article Fang, Kollo & Parring (2000) give an approximation

$$E\,\mathrm{vec}S^{-1} = \mathrm{vec}V^{-1} + (2n)^{-1} \left(\mathrm{vec}\Pi \otimes I_{p^2}\right)' \mathrm{vec}B' + 0(n^{-1}),$$

with

$$\Pi := \left(I_{p^2} + K_{pp}\right)(V \otimes V)$$

and

$$B := (I_p \otimes K_{pp} \otimes I_p)\left[I_{p^2} \otimes \mathrm{vec}V^{-1} + \left(\mathrm{vec}V^{-1}\right) \otimes I_{p^2}\right]\left(V^{-1} \otimes V^{-1}\right)$$

(We added a tranposition sign to B in the result. It was apparently lost in the process.)

A little bit of straightforward algebra shows that

$$\left(\mathrm{vec}\Pi \otimes I_{p^2}\right)' \mathrm{vec}B' = 2\,(p\Pi)\,\mathrm{vec}V^{-1}.$$

Hence the approximation boils down to

$$ES^{-1} = n^{-1}(n+p+1)V^{-1} + 0(n-1).$$

References

Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

Fang, K.-T., Kollo, T. & Parring, A.-M. (2000). «Approximation of the non-null distribution of generalized T^2-statistics», *Linear Algebra Appl.*, 321, 27-46.

Legault-Giguère, M.A. (1974). *Multivariate normal estimation with missing data*, M. Sc. Thesis, Mc Gill University, Montréal, Québec, Canada.

Neudecker, H. (2001). «Some applications of the matrix Haffian in connection with differentiable matrix functions of a central Wishart variate», *Qüestiió*, 25.2, 187-210.

Heinz Neudecker

PROBLEMES PROPOSATS

PROBLEMA N. 90

Let X be a continuous random variable such that $P(X > 0) = 1$ and X follows the same distribution as $1/X$, that is, the probability distribution function F satisfies

$$F(x) = 1 - F(1/x) \qquad x > 0.$$

Prove:

1) The median is $M = 1$.
2) The mean μ, if exists, satisfies $\mu > 1$.
3) The variance σ^2, if exists, satisfies $\sigma^2 \geqslant \mu^2 - 1$.
4) If μ and σ^2 exist then

$$\int_0^1 \frac{F^{-1}(t)}{F^{-1}(1-t)}\, dt = \sigma^2 + \mu^2.$$

C. M. Cuadras
Universitat de Barcelona

PROBLEMA N. 91

Let X be a random variable with standard logistic distribution function

$$F(x) = \left(1 + e^{-x}\right)^{-1}, \qquad -\infty < x < +\infty.$$

Let $G(x)$ be an absolutely continuous function. Suppose that $E\,(G'(X))$ and $\operatorname{var}(G(X))$ exist. Prove the following inequality

$$3\left(E\,G'(X)\right)^2 \leqslant \operatorname{var}(G(X))$$

with equality if G is F.

C. M. Cuadras
Universitat de Barcelona

Comentaris de llibres

PROBABILITATS

Marta Sanz i Solé

Edicions de la Universitat de Barcelona, 1999
220 pàgines

El llibre «Probabilitats» de Marta Sanz, constitueix una referència important, tant en el món científic com en la cultura catalana. El seu valor intel·lectual i el seu interès formatiu provenen de la competència científica i de l'experiència didàctica de la seva autora. El text es construeix a partir d'una sèrie de decisions clares i encertades sobre el públic al qual s'adreça, els objectius que persegueix, els temes que inclou o deixa fora, el rigor amb el qual els tracta i la forma que té d'esglaonar les dificultats, des de temes senzills a propietats sofisticades.

Marta Sanz, en la introducció del llibre, destaca que la teoria de la probabilitat pot resultar atractiva, i jo hi afegiria fascinant, per molta gent que tenen interessos diversos. A tall d'exemple, es refereix als científics de diverses especialitats que treballen amb mesures i mètodes matemàtics, als tecnòlegs que construeixen models probabilístics, als ciutadans que viuen en un món ple d'estadístiques i a les persones, molt nombroses, que tenen curiositat intel·lectual per a intentar resoldre problemes de jocs.

Immediatament després, l'autora ens indica que el text s'adreça de manera prioritària als estudiants de primer cicle de la llicenciatura de matemàtiques. Amb aquestes idees pretén que el llibre interessi alhora als lectors que es mouen per una curiositat general i, prioritàriament, als que es preparen per a exercir l'ofici de matemàtic. Tots sabem que un projecte d'aquesta mena té moltes dificultats i limitacions. Tanmateix, el text combina els dos objectius de manera diferent en cadascuna de les seves parts, tot seguint una estratègia molt meticulosa.

Analitzem primer l'estratègia i els mèrits del primer capítol. El llibre no requereix coneixements previs de probabilitat i, en aquest sentit, és autosuficient. Naturalment, dóna per sabudes les propietats elementals de les operacions amb conjunts, de les aplicacions entre conjunts i de l'anàlisi matemàtica. Però aquests coneixements no suposen cap dificultat ja que s'adquireixen al final del batxillerat o en un primer curs d'universitat.

El llibre s'obre amb una introducció als espais de probabilitat, dels quals en presenta prudentment les propietats bàsiques i no cau en la temptació d'estendre's innecessàriament en la teoria de la mesura i de la integració. Després de les lleis generals passa a estudiar els espais de probabilitat finits i les propietats elementals de la combinatòria

que necessita. Els exemples que ens ofereix són molt ben triats: per una banda ens mostren els models probabilístics fonamentals i per altra banda ens estimulen la curiositat, especialment en el cas que fa referència a les eleccions del Barça.

En conclusió, el primer capítol del llibre ens ofereix les propietats i els models bàsics de la probabilitat que són necessaris, tant per a un grup molt ampli de lectors com per als estudiants de matemàtiques. La diferència d'aquest capítol amb d'altres textos que tenen un contingut semblant rau en un enfocament rigorós i un llenguatge potent. El marc conceptual que introdueix està pensat per a fonamentar estudis més avançats. De fet el llibre és una preparació, sense elements superflus ni marrades innecessàries, per a encarar, en cursos posteriors, els temes de la teoria de la probabilitat més recents. En aquest sentit aconsegueix una economia important de recursos.

Per a ressenyar les parts posteriors del llibre i continuar la comparació amb d'altres tipus de textos que són molt freqüents, convé destacar dues classes d'aplicacions de la teoria de la probabilitat. Per una banda, les ciències naturals i socials i l'enginyeria empren la teoria de la probabilitat a través de l'estadística inferencial clàssica. Aquesta disciplina s'ocupa dels problemes de les mostres, de l'estimació i del test. Els llibres que corresponen a aquesta orientació destaquen les lleis febles dels grans nombres, que estableixen l'aproximació de les freqüències cap a les probabilitats, i empren les distribucions de probabilitat, especialment la de la llei normal i les que se'n deriven, per analitzar les dades empíriques observades.

Per altra banda, la física, l'economia i la teoria de les telecomunicacions s'interessen per a fenòmens sotmesos a la probabilitat tot al llarg del temps, és a dir, pels processos estocàstics, i per a l'anàlisi de sèries temporals. Aquests temes han conduït al desenvolupament de les teories de la integració i de les equacions diferencials estocàstiques, que constitueixen una branca avançada de la recerca matemàtica actual. Es tracta d'un camp en el qual hi treballen investigadors de ciències molt diverses. El podem veure com una especialitat de la física o de l'enginyeria. Però, bàsicament, l'hem de considerar com un desenvolupament actual avançat de l'anàlisi matemàtica.

Al costat del seu contingut i objectius propis, el llibre de Marta Sanz dóna als lectors instruments conceptuals que els permetran, si ho desitgen, entrar més endavant en temes de recerca. Amb aquesta finalitat, presenta acuradament els temes de la convergència quasi segura i de les lleis fortes dels grans nombres, que molts llibres d'estadística no necessiten. Personalment, crec que és molt important entendre amb profunditat, més enllà de les simples fórmules, la diferència entre la convergència en probabilitat i la quasi segura. El llibre també inclou una demostració del teorema del límit central, que molts llibres més senzills ometen, encara que sigui una propietat bàsica per a la pràctica de l'estadística.

Les diferències entre els textos orientats cap a l'estadística inferencial i cap als processos probabilístics, no ens han de fer oblidar que científicament estem en el mateix camp i que encara que donem més rellevància a uns temes, els altres també hi són pre-

sents. Les comunitats científiques de probabilitat i d'estadística han d'estar fortament integrades. Els temes de la convergència quasi segura i dels processos estocàstics també s'inclouen en els fonaments teòrics de moltes parts de l'estadística inferencial. Molts físics, economistes i enginyers no en poden prescindir.

Per totes les raons exposades en aquesta ressenya, el llibre «Probabilitats» és altament recomanable, a més a més dels esforçats estudiants de matemàtiques, pels professors i estudiants de moltes matèries relacionades amb l'estadística. Marta Sanz i el grup de recerca de la Universitat de Barcelona són reconeguts internacionalment en la investigació de les equacions diferencials estocàstiques. L'autoritat científica de l'autora i el seu esforç d'ordenació i de clarificació són una garantia del valor d'estudi o de consulta del llibre.

Eduard Bonet

STATISTICIANS OF THE CENTURIES

C. C. Heyde i E. Seneta

Springer-Verlag, New York, 2001
500 pàgines

El llibre *Statisticians of the centuries* és un recull de biografies curtes de 103 estadístics nascuts abans del segle XX.

Segons podem llegir en el Prefaci, el llibre neix a partir d'una iniciativa de l'International Statistics Intitute (ISI), entitat fundada l'any 1885 que ha esdevingut una de les més representatives de l'activitat estadística internacional. Un comitè de l'ISI va consultar extensivament la comunitat estadística i, particularment, destacats membres de l'Institut per arribar a la llista dels 103 estadístics inclosos en el llibre, amb la restricció prèvia que fossin nascuts abans del segle XX. De la joventut de la disciplina ens dóna la idea el fet que la història comprengui només quatre segles, i que el 85% dels estadístics inclosos siguin nascuts en els segles XVIII o XIX.

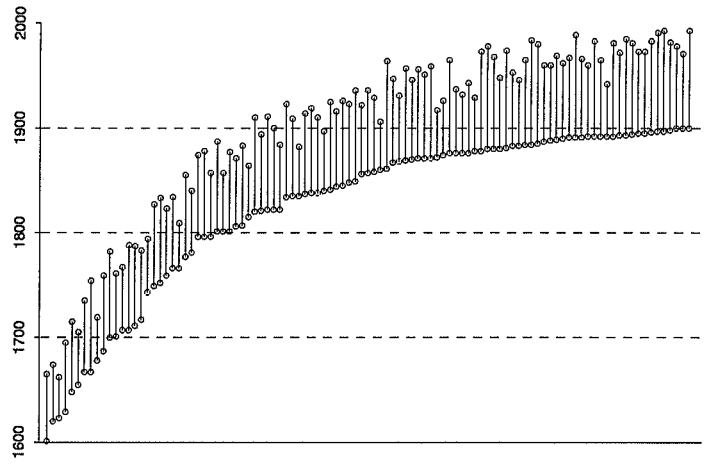


Figura 1. Cada línia vertical s'estén al llarg dels anys que va viure cadascun dels estadístics biografiats en el llibre, ordenats, com en el llibre, cronològicament. Per exemple, el tercer, de vida tan curta, és Blaise Pascal.

Adreçat a un ampli públic, el llibre se centra en les persones i en el seu context històric i sociològic. Al seu través, presenta els conceptes, mètodes i idees que conformen el cos de la disciplina. Setanta-cinc autors d'arreu del món aporten les biografies incloses en el llibre. Cadascuna d'elles està encapçalada amb una fotografia i un breu sumari i va acompanyada de les corresponents referències bibliogràfiques. El llibre comença amb un capítol sobre la Probabilitat abans de Pascal, que cobreix des de De Vetula, un poema en tres llibres atribuït a Ovidi (43 ac-17 dc) que conté la primera descripció escrita de càlculs probabilístics, fins a l'obra dels italians del segle XVI.

La distribució dels autors segueix l'ordre cronològic del seu naixement i els agrupa segons el segle en què van desenvolupar les seves principals aportacions. Així, trobem 11 autors del segle XVII, 16 en el segle XVIII, 21 en el segle XIX i 54 en el XX. El gràfic que acompanya aquestes línies visualitza clarament el creixement del nombre de científics involucrats en la disciplina en els seus pocs segles d'existència. De la quantitat de disciplines que incideixen en l'estadística en parlarem tot seguit.

Els textos biogràfics són molt interessants i ben documentats. Estan escrits, la majoria, amb la intenció d'arribar a qualsevol persona que conegui les beceroles de la disciplina. Més enllà dels detalls purament biogràfics, que en molts casos són ben interessants, els textos situen els personatges en el seu context històric i fan esment de les seves relacions amb altres científics i amb el desenvolupament d'altres disciplines científiques. És aquest aspecte el que apareix transversalment en tota l'obra: sense ser en cap moment argument central, destaca la quantitat d'àmbits d'aplicació que han proporcionat a l'estadística, no només problemes a resoldre, sinó científics que per resoldre'ls han aportat a la disciplina principis metodològics que conformen el seu cos actual.

Està clar que una ressenya sobre un llibre com aquest no pot acabar sense extreure'n algunes dades. Destaca el fet que només dues dones (citem-les: Florence Nightingale, la fundadora de la Creu Roja, i Gertrude Cox) apareixen en el llistat de biografiats. Els anys de vida dels estadístics d'aquesta insigne mostra, que potser no és representativa però sí, molt significativa, van des dels 39 als 103, sí 103, amb una mitjana de 76 anys. Pel que fa als llocs d'origen, en un sentit ampli del terme, hi trobem 26 estadístics de les Illes Britàniques, 19 francesos, 13 centreeuropeus (la variabilitat de certes fronteres no permet precisar més en aquest grup), 3 italians i 3 més dels Països Baixos. Els 10 estadístics de Rússia o Ucraïna contrapesen els altres tants dels Estats Units d'Amèrica (també n'hi ha un del Canadà). Representants d'altres continents en trobem 2 de l'Índia i 3 d'Austràlia, lloc d'origen dels editors del llibre.

En un intent, bastant més agosarat i molt menys científic, d'identificar les àrees de treball dels biografiats, i en una lectura somera, hem comprovat que només 53 dels 103 poden ser qualificats com a matemàtics, quedant, doncs, un bon nombre d'estadístics que hem de qualificar com a aplicats, si no fos que en la majoria dels casos és al revés: són científics d'altres àrees que, necessitats de mètodes estadístics, han fet importants aportacions metodològiques a l'àrea. En trobem 9 de les Ciències Naturals en sentit am-

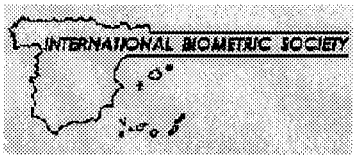
pli (incloent Física, Geofísica, Química, Meteorologia), 6 en diferents branques de Matemàtica Aplicada (Actuarials, Assegurances, Astronomia, Finances), 38 en Ciències Socials (la meitat economistes, però també demògrafs, sociòlegs o de Ciència Política), 18 de Ciències de la vida (Biòlegs, epidemiòlegs, metges, psicòlegs), 4 que han treballat en enginyeria o en la indústria i, finalment, 7 dedicats a l'estadística oficial.

Naturalment, les xifres sumen més que 103, per tant, molts dels estadístics relacionats centraren el seu treball en més d'una àrea. Entre tots, n'hem trobat només 14 que no els hem sabut qualificar de res més que d'estadístics, en el sentit que la metodologia estadística va ser el seu principal camp d'actuació aplicada a tants camps diferents que no podem identificar-los amb cap d'ells, i tampoc els veiem pròpiament com a matemàtics. Cal no oblidar que tots ells són nascuts abans de començar el segle XX. Entre els nascuts en aquest segle potser trobaríem una altra distribució d'interessos.

Frederic Udina

Ressenyes d'activitats institucionals

Sociedad Española de Biometría



Región Española
de la Sociedad Internacional de Biometría

Spanish Region
of the International Biometric Society

http://www.iata.csic.es/ibsresp

La **Sociedad Española de Biometría/Región Española de la Sociedad Internacional de Biometría** (abreviadamente **SEB** o ***REsp***) tiene como objetivos promover, impulsar y difundir el desarrollo y la aplicación de los métodos matemáticos y estadísticos a la biología, medicina, psicología, farmacología, agricultura y otras ciencias afines (ciencias relacionadas con los seres vivos). Cualquier profesional o alumno de estas disciplinas puede ser miembro de la SEB.

Consejo Directivo

Presidenta:	María Jesús Bayarri García (Medicina)
Vicepresidenta:	Guadalupe Gómez Melis (Biología)
Secretario y Tesorero:	Fernando López Santoveña (Agronomía)
Vocal en calidad de Miembro del Consejo de la IBS:	Emilio A. Carbonell Guevara (Agronomía)
Vocales:	Juan Luis Chorro Gascó (Psicología)
	Juan Ferrándiz Ferragud (Biología)
	Purificación Galindo Villardón (Medicina)
	Eduardo García Cueto (Psicología)
	José Luis González Andújar (Agronomía)
	Alex Sánchez Plá (Biología)
Corresponsal de la REsp en el «Biometric Bulletin» de la IBS:	María Luz Calle Rosingana

FIRST BARCELONA WORKSHOP ON SURVIVAL ANALYSIS

Barcelona, Spain: June 12th - 14th, 2002

The TES Institute
Training of European Statisticians

GENERAL INTRODUCTION

The TES Institute is a non-profit making association of 17 European National Statistical Institutes and the University Centre of Luxembourg. It offers a post-graduate vocational training programme for statisticians and other groups working in the statistical environment. The TES Institute today boasts a solid foundation of statistical training experience, built up since the beginning of the nineties.

Our mission is to provide world-wide cutting edge lifelong international training for professional statisticians. The training emphasises the inter country know-how transfer and focuses on the dissemination of best practices in terms of providing statistics compliant with both European and International Statistical System standards.

The training programmes of the TES Institute are designed for both public and private sector statisticians in the broadest sense of the word i.e. university graduates of any discipline involved in statistical work within Statistical Institutes, European Institutions, Government Agencies, National or private Banks, Enterprises, ...

Training methods are problem-oriented and based on a twofold track approach combining both teaching of theoretical concepts and practical sessions using real life cases. Moreover, the «learning-by-doing» process provides an opportunity to respond in an effective way to recent scientific and technological developments as well as new requirements from the labour market.

The training programmes offered by the TES Institute provide both theoretical and practical background but the courses have all a very strong applied character.

These programmes also offer participants the opportunity to meet colleagues form all over Europe and other countries since the TES Institute has extended its activities to the Central European, Mediterranean Basin and TACIS countries.

The above characteristics represent the basic conditions to acquire sharper competence in their work environment and highlight the European dimension of their activity.

After ten years of existence, the programme became entire part of the statistical world. For the time being, around 500 participants coming from more than 30 countries are trained every academic year. Such an interest is mainly due to the large number of courses on offer. Indeed, the TES portfolio comprises more than 80 courses of short duration all at post-graduate level.

The TES Institute offers a coherent set of interrelated courses in the following domains:

- Official Statistics: General Issues (OSG)
- Official Economic Statistics (ECO)
- Official Social Statistics (SOC)
- Data Collection and Survey Methodology (DAT)
- Applied Statistical Analysis (ASA)
- Publication and Dissemination of Statistics (PDS)
- Management in a Statistical Institute (MSI)
- Statistical Information Systems (SIS)

To reach the widest audience possible, the TES Institute currently offers courses in English, French, German, Arabic and Russian.

After a few years of co-operation with the Central European countries, the TES Institute has recently extended the co-operation to the MEDSTAT and TACIS region. Such an internationalisation is the direct result of the growing importance of training as a part of the current technological and intellectual development. Therefore, as far as statistics and economics are concerned, it is of the utmost importance to extend the best national practices to an international level.

It is obvious that the TES programmes should be considered as a complement and not a substitute to the training provided at national level.

In brief, one may say that by offering training opportunities which are complementary to the ones provided at national level, the TES Institute is offering a new approach of the subsidiarity concept.

TRAINING

The Core Programme 2002 started in January will end in November 2002. The overview of the remaining courses can be found at the end of the present paper.

The TES Institute is currently preparing the Core Programme 2003 that will start in January 2003 and offer 28 courses. The final programme will be available by September 2002.

Beside the execution of the subsidised annual vocational training programme –which is only one of our many activities– the TES Institute also offers Special Courses that may be repetition of courses in the Core Programme or tailor-made and organised at the request of a country or a group of countries. In this context the TES Institute is active through the PHARE programme for Central European Countries, through the MEDSTAT programme in the countries of the Mediterranean Basin and through the TACIS programme in a number of CIS countries.

CONSULTING

In addition to the vocational training activities, we will continue to develop and extend our consulting activities in the above mentioned regions and elsewhere in order to maintain and solidify our position in the market of professional training and staff development for statisticians. These consulting activities consist in either *competence building* by the training of «in-country» trainers with multiplicative effects of the training or *institutional building* by the development of training centres, improvement of vocational training system and curricula. For the time being, the TES Institute has been involved in consulting activities with the Russian Federation, Ukraine and Kazakhstan.

RESEARCH

In various Research Institutes, Universities and National Statistical Institutes of the Member States of the European Union and other countries, research is taking place covering statistical methodology and statistical techniques such as data collection methods, data transmission, storage and warehousing, statistical data analysis, etc. TES is confident that, being closely involved in dissemination of the results of activities in the field of research and technological development, will not only help the Commission in its future policy on R&D but will also contribute to a better communication between the various players in statistical research. As a result, the TES Institute has decided to play an active role in European research projects in various fields of statistics.

Within this framework, the TES Institute is participating in a consortium, in the 5th Framework Programme of the Commission, focusing on the creation and the development of on-line training courses (VL-CATS Project). On the basis of possibilities offered by existing tools, the VL-CATS project will provide access over the Internet, to teaching and educational material with a special focus on Official Statistics. Through this project, experts will be given the opportunity to share and disseminate their statistical knowledge on a wider scale. Along with EU and EFTA countries, statisticians from accessing countries and other Central European countries will also be given the opportunity to produce statistics compliant with the European Statistical System standards.

The main objectives of VL-CATS Project are:

- To create a virtual library of all reference material available in Official Statistics;
- To define and implement standards for the development of distance courses in the various fields of Official Statistics;
- To develop a controlled environment that will support the virtual library and give selective access to training courses;
- To develop a service for update and quality assurance of VL-CATS.

The TES Institute has been also involved in a second research project called ASSO. The ASSO project represents a unique environment for the development of tools helping people to better interpret results yielded by data analysis. The aim of the project is to use Symbolic Data and Symbolic Data Analysis to solve problems of Statistical Offices (NSIs), in particular: Large database treatment, Confidentiality, Missing data, Metadata modelling, Quality control and accurate data interpretation.

The goals of the ASSO project are:

- Offer user-friendly dissemination of statistical data through Symbolic Data
- Use, create, propagate and design statistical concepts by Symbolic Objects
- Improve the building of Symbolic Data from a relational database or directly through edition
- Introduce structured statistical Metadata ensuring better quality of statistical results
- Allow to retrieve new Symbolic Data and to propagate results on the database
- Add new measures of dissimilarity between Symbolic Objects
- Add consensus between classifications of Symbolic Objects
- Add new Symbolic Analysis methods of unsupervised and supervised classification of Symbolic Data and Symbolic Objects
- Give new tools for the study of quality, stability and robustness of the Symbolic Analysis
- Give new tools for visualising Symbolic Objects and/or the results of Symbolic Analysis

PUBLICATIONS

The TES Institute is regularly publishing articles on its current activities in periodic Newsletter of several Statistical Institutes and some statistical journals as Qüestiió (Quaderns d'Estadística i Investigació Operativa), edited by the Institut d'Estadística de Catalunya.

The TES Institute has started the production of TES Manuals on subjects covered by the vocational training programme:

- The first manual available at the TES institute presents *The role of statistics in a democracy.*
- The second manual to be published soon will cover the *Index numbers for spatial and temporary comparisons.*
- Further manuals will cover topics of *sampling techniques, seasonal adjustment methods.*

TES NEWSLETTER

The TES Institute has disseminated its Newsletter named «Facts & Visions». This newsletter is considered as an important information tool between the TES Institute and its partners from all over Europe focusing on matters related to training and staff development in the statistical world.

We hope that our partners will use «Facts & Visions» to report on their own activities. So, may we encourage you to use your keyboards for European information purposes.

Copies of «Facts & Visions» are available at the TES Institute as well as guidelines for the submission of articles.

GENERAL INFORMATION

For further details on any of the above mentioned topics, please contact directly Ms. Valérie Vandewalle (*Head of Curriculum Development*):

Phone: (352) 29.85.85.34
Fax: (352) 29.85.29
E-mail: vvandewalle@tes-institute.lu

Code	Course Title	Days	Course Leader	Location	From	To
DAT-102/2002	Survey non response: reduction, weighting and imputation	5	Peter Lynn	London	21-jan-02	25-jan-02
SOC-109/2002	Systems of education statistics	3	Spyridon Pilos	Luxembourg	4-feb-02	6-feb-02
MSI-105/2002	Planning and management of statistical programmes and structures	4	John Scrivener	London	18-feb-02	21-feb-02
PDS-001/2002	Basic principles of publication and dissemination of statistical products	5	Ed Swires-Hennessey	London	4-mar-02	8-mar-02
ECO-202/2002	The european system of account (ESA95) – Sector accounts	3	Brian Newson	Luxembourg	11-mar-02	13-mar-02
DAT-203/2002	Estimation for small areas	3	T.M. Fred Smith & Danny Pfeffermann	Barcelona	18-mar-02	20-mar-02
ECO-101/2002	Nomenclatures, classifications and their harmonisation	4	Niels Langkjaer	Luxembourg	18-mar-02	21-mar-02
SIS-203/2002	Geographical information systems	4	Marco Painho	Lisbon	8-apr-02	11-apr-02
ECO-205/2002	The European system of accounts (ESA95) – Quarterly accounts	3	Roberto Barcellan	Luxembourg	15-apr-02	17-apr-02
PDS-102/2002	Effective presentation of official statistics to news media	5	John Wright	London	15-apr-02	19-apr-02
OSG-001/2002	The European statistical system	3	Eurostats Experts	Luxembourg	22-apr-02	24-apr-02
ASA-102/2002	Price index theory and price statistics	5	Peter von der Lippe	Düsseldorf	27-may-02	31-may-02
SOC-106/2002	Introduction to demographic data and their analysis	5	Otto Andersen	Copenhagen	27-may-02	31-may-02
ECO-102/2002	Business registers: set-up, use and maintenance	5	Yngve Bergström	Oslo	3-jun-02	7-jun-02
ECO-001/2002	Pratique de la statistique des comptes nationaux	10	Maryvonne Lemaire	Paris	3-jun-02	14-jun-02
DAT-002/2002	Sampling techniques and practice	10	Prof. T.M. Fred Smith	Southampton	10-jun-02	21-jun-02
SOC-001/2002	Systems of social statistics	5	Pieter Everaers	Voorburg	17-jun-02	21-jun-02
SIS-202/2002	Measurement of the quality of statistics	3	Raoul Depoutot	Paris	24-jun-02	26-jun-02
SOC-108/2002	Systems of health statistics	4	Jacques Bonte	Luxembourg	16-sep-02	19-sep-02
PDS-103/2002	The European statistical system	4	Eurostat Experts	Luxembourg	22-apr-02	24-apr-02
OSG-003/2002	Confidentiality and protection of privacy	3	Photis Nanopoulos	Luxembourg	30-sep-02	2-oct-02
ECO-001/2002	National accounts statistics in practice	10	Wim van Nunspeet	Voorburg	7-oct-02	18-oct-02
SIS-150/2002	Statistics on the information society	3	Douglas Koszerek & Richard Deiss	Luxembourg	21-oct-02	23-oct-02
DAT-103/2002	New advanced technologies for data collection	3	Uwe Kunzler	Luxembourg	28-oct-02	30-oct-02
ECO-152/153	Economic accounts for agriculture	3	Ulrich Eidmann	Luxembourg	4-nov-02	6-nov-02
ECO-153/2002	Structural business statistics	4	Sven Egmose	Copenhagen	18-nov-02	21-nov-02
SIS-151/2002	New tools and methods of data transmissions for IT specialists	4	Didier Hardy	Luxembourg	19-nov-02	22-nov-02
ECO-125/2002	Economic analysis and flash estimates: application for SNA statistics in practice	4	Gian Luigi Mazzi	Luxembourg	25-nov-02	28-nov-02

OSG Official Statistics: General Issues
ECO Official Statistics: Economic Statistics
SOC Official Statistics: Social Statistics
DAT Data Collection and Survey Methodology
ASA Applied Statistical Analysis
SIS Statistical Information Systems
PDS Publication and Dissemination of Statistics
MSI Management in a Statistical Institute

tes

Institute

«ESTIMATION FOR SMALL AREAS»
(DAT-203/2002)

COURSE LEADER

T.M.Fred SMITH

Danny PFEFFERMANN

OBJECTIVE

The course supplements the foundational course DAT-003 on sampling techniques in the direction of estimation and inference for small areas.

TRAINING METHODS

Lectures, exercises, presentation of case studies.

TARGET GROUP

Statisticians responsible for survey design and analysis.

ENTRY QUALIFICATIONS

University degree with a solid knowledge of mathematical statistics. Knowledge of sampling theory equivalent to the course DAT-003 and sound knowledge of the teaching language (see below).

EXPECTED OUTPUT

Participants should have an understanding of the conceptual issues involved in small area estimation and should be able to apply the main techniques presented.

CONTENTS

THE PROBLEM EXAMINED

- Examples
- Estimation for domains
- Estimation for small domains or small areas
- Inter-censal estimates for small areas

BACKGROUND TECHNIQUES

- Randomisation inference, the Horvitz-Thompson estimator and its properties
- Borrowing strength (information) from other areas
- ANOVA methods
- Using auxiliary information
- Regression methods
- Synthetic estimators
- Model-based methods: components of variance, regression, structure preserving estimators (SPREE).

DEMOGRAPHIC METHODS

- Inter-censal estimator
- Types of auxiliary information
- Symptomatic accounting techniques.

CASE STUDIES

- To illustrate the methods throughout the course, participants should bring examples from their agencies
- Problems of evaluation.

REQUIRED READING

- Revision of basic sampling theory from a text such as Cochran: *Sampling techniques*, Wiley, 1977.
- Särndal, C-E, Swensson, B & Wretman, J.: *Model assisted survey sampling*, Springer Verlag, 1992.

SUGGESTED READING

Platek R. *et al.* (eds): *Small area statistics: an international symposium*, Wiley, 1987.

REQUIRED PREPARATION

The participants are requested to write a short summary both of their own activity at their organisation and of the organisation's practices, problems and experiences in the field of the course, in addition to any course specific requirements.

PRACTICAL INFORMATION

Date: 18-20 March 2002

Duration: 3 days

Training site: Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58. 08003 Barcelona. ESPAÑA

Language: English

Training Fee: 650 €

Registration: until 15 December 2001

First Barcelona Workshop on Survival Analysis

Barcelona, Spain: June 12th-14th, 2002

Research and Training in Failure Time Methods in the New Millennium

The Statistics and Operations Research Department of the Universitat Politècnica de Catalunya, together with the Biostatistics Department of Harvard University and the Spanish Region of the International Biometric Society, are very pleased to announce a three-days workshop on Survival Analysis in Barcelona from June 12th to 14th, 2002. The workshop will consist on invited methodological sessions, a round table on how to train future generations of biostatisticians, a short course on optimization methods for statistics and a session of contributed papers in poster form.

The topics to be covered in the invited sessions include:

- Sampling and Observational Structures. What happens under various conditioning schemes?
- Multivariate and Covariate Methods and Models. Application of the indirect method developed in econometrics to analyze noisy recurrent event data.
- Alternatives to the Cox model. Using additive hazards regression models in multistate modeling.
- Issues on Interval-censored data. Current status data with competing risks.
- Event History Analysis. Cost or Quality of Life processes related to event histories. Structural models, PH version.
- New challenges in survival studies with missing data

Confirmed invited speakers are (in alphabetical order):

P. K. Andersen, University of Copenhagen, Denmark
D. Commenges, Université de Bordeaux II, France
E. Goetghebeur, Ghent University, Belgium
N. Jewell, University of California, Berkeley, USA
J. Kalbfleisch, University of Michigan, USA
N. Keiding, University of Copenhagen, Denmark
J. Klein, University of Wisconsin, USA
S. Lagakos, Harvard University, USA
J. Lawless, University de Waterloo, Canada
S. F. Nielsen, University of Copenhagen, Denmark
R. Prentice, University of Washington, USA
B. Turnbull, Cornell University, USA
L. J. Wei, Harvard University, USA

Local organizer commitee:

Guadalupe Gómez, Universitat Politècnica de Catalunya, Barcelona
M. Luz Calle, Universitat de Vic, Vic
Olga Julià, Universitat de Barcelona, Barcelona
Carles Serrat, Universitat Politècnica de Catalunya, Barcelona

Sponsors:

Institut Estadística de Catalunya (Idescat)
Servei d'Estadística de la Universitat Autònoma de Barcelona

The total number of participants will be limited to approximately 60 people.

If you are interested in the event and want to be included in our mailing list for future information, please send an e-mail to carles.serrat@upc.es. The web site of the meeting is still under construction but, meanwhile you can start visiting, and enjoying,

Barcelona from the following web pages:

http://www.bcn.es/english/ihome.htm
http://www.bcn.es/english/turisme/saber/frames/ilsaber.htm
http://www.bcn.es/english/barcelon/cultura/frames/ilcultu.htm
http://www.bcn.es/english/barcelon/transport/itransp.htm

Informació per als autors i lectors

NORMES PER A LA PRESENTACIÓ D'ARTICLES A *QÜESTIIÓ*

La revista accepta, per a la seva publicació, articles originals no sotmesos en cap altra revista dins els àmbits de l'Estadística, la Investigació Operativa, l'Estadística Oficial i la Biometria. Els articles poden ser teòrics o aplicats, incloent aspectes computacionals i/o de caire docent, i es publiquen exclussivament en anglès, acompanyats d'un breu resum en català.

Tots els originals destinats a les esmentades seccions temàtiques de *Qüestiió* seran sotmesos sistemàticament a una avaluació prèvia a càrrec d'especialistes independents i/o membres del Consell Editor, llevat d'aquells que siguin convidats per la revista i de les reimpressions d'articles. El resultat de l'avaluació serà comunicat a l'autor principal als efectes d'eventuals correccions formals o dels seus continguts.

Per a totes les trameses d'originals, la revista emetrà un acusament de recepció la data del qual figurarà com a «data de rebuda» en la publicació de l'article. Per la seva banda, la «data d'acceptació» de l'article serà la data de recepció de la versió definitiva.

Per a la presentació de articles, l'autor trametrà a l'adreça postal de la Secretaria de *Qüestiió* (Institut d'Estadística de Catalunya) dues còpies del treball complet en DIN A4, a una sola cara i a doble espai i, paral·lelament, per correu electrònic (questiio@idescat.es) la primera pàgina de l'article en format PDF o PS que contindrà el títol, el nom de l'autor o autors, l'afiliació i l'adreça completa, així com un resum de 75-100 paraules seguit de les principals paraules clau i la seva adscripció a la classificació MSC2000 de la American Mathematical Society. Abans de sotmetre els articles a la revista, s'aconsella els autors que revisin la correcció lingüística de l'anglès dels textos.

Les referències bibliogràfiques es faran indicant el cognom de l'autor seguit de l'any de la publicació entre parèntesi [i.e.: Mahalanobis (1936), Rao (1982b)] i seran llistades alfabèticament al final de l'article; les referències múltiples d'un mateix autor s'ordenaran cronològicament. Les notes explicatives es numeraran correlativament i han d'aparèixer al peu de la pàgina corresponent. Les taules i figures també es numeraran correlativament en el text i seran reproduïdes directament dels originals tramesos en cas que no sigui possible la seva autoedició.

Una vegada avaluat satisfactòriament l'article caldrà que l'autor el trameti en versió impresa i electrònica segons les instruccions que li seran indicades pel responsable de l'avaluació del seu treball. Es recomana que aquesta versió final es trameti preferiblement en el processador de textos LaTeX 2ε [subsidiàriament, es poden trametre els textos i les taules en Word Perfect —versió 6.0A o anterior— o ASCII]; en el cas de figures, diagrames o gràfics es recomanen els formats adients per als programes-editors PS, EPS o PCX. Els autors han de garantir la correspondència exacta entre la versió impresa i la còpia electrònica. La Secretaria de Qüestiió posa a disposició dels autors que ho sol·licitin plantilles en aquest format per a la seva edició. Per a les referències adequades de la classificació de l'AMS: enllaç amb la *2000 Mathematics Subject Classification (MSC2000)*. Per a més informació sobre l'ús i criteris per a l'adopció d'aquesta classificació podeu consultar prèviament *Classificació de l'AMS per als articles publicats a Qüestiió* (www.idescat.es/publicacions/questiio/questiio.stm).

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE SUBMISSION OF ARTICLES FOR *QÜESTIIÓ*

The journal welcomes submission of articles and contributions that are not being considered for publication in any other journal in the fields of Statistics, Operational Research, Official Statistics or Biometrics. Articles may be theoretical or applied, including teaching aspects and applications, and will be published exclusively in English, with a brief abstract in Catalan enclosed.

All originals assigned to the thematic sections of *Qüestiió* mentioned will be systematically reviewed by independent referees and/or members of the Editorial Board, who will send a report to the main author of the article in order to correct, if necessary, any formal or content aspects. The articles invited by the journal and articles reprinted will be excluded from this evaluation process.

For all submissions, the journal will issue a receipt corresponding to the submission date, which will appear as «date received» in the final publication of the article. The «acceptance date» of the article, which will appear in its final publication, will be the date of the journal receiving the final version.

For the presentation of original articles, the author should send, to the postal address of the Secretary of *Qüestiió* (Institut d'Estadística de Catalunya), two copies of the complete article, typed on A4 sheets, one side of the paper only, double spaced and with wide margins. At the same time, the author should send by electronic mail (questiio@idescat.es) the first page of the article on PDF or PS format including the title, the name of the author or authors, their affiliation, full address as well as an abstract of the paper (75-100 words) followed by the keywords (in the original language) and its assignation to the AMS classification. Before submitting the papers to the magazine, authors are advised to revise their English linguistic correctness.

Bibliographical references should state the author's name followed by the year of publication in brackets [e.g.: Mahalanobis (1936), Rao (1982b)] and they should be listed at the end of the article in alphabetical order; multiple references to the same author should be given in chronological order. Footnotes should be numbered in the article and appear at the foot of the corresponding page. Figures and tables are to be numbered in consecutive order in the text using Arabic numerals and will be directly reproduced from the originals submitted if it is not impossible to print them electronically.

Once the evaluation of the article had been successful, the author is required to send it with both a paper and an electronic copy, according to the instructions given by the person responsible for the evaluation of his work. This final version should be processed by LaTeX 2_ε, preferably, or, failing that, by Word Perfect (6.0A or earlier) or ASCII for text and tables; for figures, diagrams or graphs, the appropriate formats of PS, EPS or PCX software tools are strongly recommended. Authors must ensure that the electronic copy version and the one on paper are exactly the same. The Secretary of Qüestiió can send, by request of the authors, the LATEX for manuscript preparation. For the appropriate AMS classification references link to *2000 Mathematics Subject Classification (MSC2000)* and for more information about the criteria to use the *AMS Clasification for the articles published by Qüestiió* (www.idescat.es/publicacions/questiio/questiio.stm).

QÜESTIIÓ
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS INSTITUCIONALS A *QÜESTIIÓ*

Qüestiió convida les entitats patrocinadores, les institucions col·laboradores, els organismes públics i privats, i tota la comunitat científica vinculada a l'estadística o la investigació operativa, a la publicació d'anuncis institucionals sobre cursos, seminaris, congressos i activitats similars que, preferentment, tinguin lloc en el nostre país. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les entitats interessades, de manera que Qüestiió no fa una cerca sistemàtica d'esdeveniments d'aquesta naturalesa, ni té cap ànim d'exhaustivitat en les ressenyes d'activitats finalment publicades.

Una vegada aprovada la inclusió dels anuncis sol·licitats es procedirà a la seva publicació, i es reproduirà directament dels originals tramesos amb les mides adequades i la màxima qualitat tipogràfica possible; en aquest cas, Qüestiió no procedeix a cap mena de procés d'autoedició de la versió impresa que l'anunciant hagi tramès. Si els originals es trameten en els mateixos termes electrònics exigits per als articles (vegeu «Normes per a la presentació d'articles a Qüestiió»), la revista procedirà a la seva autoedició. Si es desitja una qualitat superior a la reproducció simple o l'autoedició, o bé la seva publicació en color, els sol·licitants hauran de posar-se en contacte amb la Secretaria de Qüestiió per tal de trametre els fotolits dels textos originals corresponents.

La disposició dels textos i les figures adjuntes dels anuncis han de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de Qüestiió es reserva la decisió final pel que fa a la seva publicació.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la inserció en els números/volums de Qüestiió que prèviament s'hagi establert. La revista no es fa responsable dels retards, per part de l'anunciant, que impedeixin la publicació de l'anunci en els termes previstos.

Mònica M. Jaime
Secretaria de Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR INSTITUTIONAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió invites all sponsor entities, collaborating institutions, other public and private bodies and the entire scientific community related to Statistics or Operations Research to submit institutional advertisements on courses, seminars, congress and similar activities that will be held, preferably in our country. These will be accepted in English, French, Catalan or any of the other official languages in Spain. The initiative should always come from the entities interested in advertising them so that Qüestiió's aim is not to do a systematic search of these events and therefore does not publish a comprehensive list of such activities.

Once their insertion is approved the advertisements will be reproduced from the most accurate photocopy of the originals sent by the advertiser to Qüestiió in paper copy, with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editing process to the printed version that the advertiser has sent. If the original advertisements are sent in the same electronic format requested by the articles (please see «Guidelines for the submission of articles for Qüestiió») the journal will print it directly from the file. If a better quality than the simple reproduction or automatic printing or a colour version of the adverts is desired, the authors should contact the Secretary of Qüestiió in order to negotiate this.

The typesetting of texts and figures in the advertisement should have maximum intelligibility and clearness, neither compressing the information too much nor using formats or letter fonts that are too small. Furthermore, the information has to be reliable, without errors and respectful of the people and institutions. The management of Qüestiió has the right to a final decision concerning the insertion of the advertisement.

Advertisers commit themselves to give the text/materials on request in order to insert them in the issues of Qüestiió that have been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published on the agreed terms.

Mònica M. Jaime
Secretaria de Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona
Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

NORMES PER A LA PUBLICACIÓ D'ANUNCIS PRIVATS O AMB FINALITAT COMERCIAL A *QÜESTIIÓ*

Qüestiió accepta la publicació d'anuncis privats o amb finalitat comercial sobre productes, serveis o altres eines promocionals a l'entorn de l'estadística o la investigació operativa. Els textos poden presentar-se en anglès, francès, català o en qualsevol altra llengua oficial a l'Estat espanyol. Les iniciatives per a una possible publicació sempre són a instància de les organitzacions que hi estiguin interessades, de manera que Qüestiió no fa una cerca sistemàtica de novetats o productes d'aquesta naturalesa ni té cap ànim d'exhaustivitat en els anuncis finalment publicats.

Els anuncis en **blanc i negre** s'elaboren a partir de la fotocòpia més acurada possible dels originals que trameti l'anunciant en versió impresa, amb les mides adequades i la màxima qualitat tipogràfica. Per tant, en aquest cas la revista no efectua cap procés d'edició ulterior respecte de la versió impresa que l'anunciant hagi tramès. Alternativament, si els anuncis originals es trameten en els mateixos termes formals exigits per als articles (vegeu «Normes per a la presentació d'articles a Qüestiió»), la revista procedirà a la seva autoedició. Igualment, si es desitja una qualitat superior a la reproducció simple, els sol·licitants hauran de trametre els fotolits dels originals corresponents o encarregar-los a Qüestiió, que els facturarà separadament.

Els anuncis **en color** requereixen els fotolits dels textos originals, que poden ser subministrats directament per l'anunciant o bé encarregats per la revista a compte de l'anunciant; en el segon cas, l'anunciant ha de trametre a la revista els originals impresos en color amb la màxima qualitat, per tal de filmar-los amb les millors garanties i condicions. El cost dels fotolits realitzats per Qüestiió serà sempre a càrrec de l'anunciant, a qui se li repercutirà l'import i l'IVA d'aquests, juntament amb les tarifes que corresponen a la modalitat d'anunci per la qual hagi optat.

La disposició dels textos i figures adjuntes dels anuncis ha de procurar la màxima intel·ligibilitat i claredat expositiva, sense atapeir la informació ni utilitzar formats o fonts de lletres excessivament petites. D'altra banda, la publicitat ha de ser fidedigna, exempta d'enganys i respectuosa amb les persones i institucions. En qualsevol cas, la direcció de Qüestiió es reserva la decisió final de la seva inclusió.

L'anunciant es compromet a lliurar els textos/materials amb l'antelació que se li indiqui per a la seva inserció en el(s) número(s)/volum(s) de Qüestiió que prèviament s'hagi establert. La revista no es fa responsable dels retards per part de l'anunciant que impedeixin la publicació de l'anunci en els termes previstos.

Imports:

1 pàgina en color (un número aïllat):	125.000 PTA + IVA
1 pàgina en color (tres números consecutius):	200.000 PTA + IVA
1 pàgina en blanc i negre (un número aïllat):	30.000 PTA + IVA
1 pàgina en blanc i negre (tres números consecutius):	50.000 PTA + IVA
1/2 pàgina en blanc i negre (un número aïllat):	20.000 PTA + IVA
1/2 pàgina en blanc i negre (tres números consecutius):	35.000 PTA + IVA

Mides opcionals dels anuncis:

1 pàgina sencera (espai intern):	19.0 cm. × 12.3 cm.
1 pàgina sencera (espai extern):	23.8 cm. × 17.0 cm.
1/2 pàgina (espai intern):	9.5 cm. × 12.3 cm.
1/2 pàgina (espai extern):	11.9 cm. × 17.0 cm.

Forma de Pagament:

- Transferència bancària al compte: 2013-0100-53-0200698577
- Xec bancari nominatiu a l'Institut d'Estadística de Catalunya
- Pagament amb targeta de crèdit

El pagament serà per l'import total de la factura corresponent, on hi figurarà el cost dels fotolits en el cas que l'edició de l'anunci hagi estat a càrrec de l'Institut. En el cas que l'anunciant necessiti una factura proforma, només cal que ho faci saber amb l'antelació suficient.

Correspondència:

Mònica M. Jaime
Secretaria de Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

GUIDELINES FOR THE PRIVATE OR COMMERCIAL ADVERTISEMENTS IN *QÜESTIIÓ*

Qüestiió accepts for their publication both private and commercial advertisements on products, services or other promotional tools related to statistics or operational research and will be accepted in English, French, Catalan or any of the official languages in Spain. The initiatives should always come from entities interested in advertising them so that Qüestiió's aim is not to do a systematic search of news and therefore does not publish a comprehensive list of such private or profit activities.

The **black and white** advertisements are made out from the most accurate photocopy of the originals sent by the advertiser to Qüestiió in paper copy with the appropriate size and at the best possible typographic quality. Therefore, in this case the journal does not elaborate any further editorial process to the printed version that the advertiser has sent. Alternatively, if the original advertisements are sent in the same formal terms required by the articles (please see «Guidelines for the submission of articles for Qüestiió»), the journal will proceed to its autoedition. In the same way, if a better quality than the simple reproduction is wanted, the authors should send the photolits of the corresponding original texts or, on the other hand, order to Qüestiió their fulfilment, which will be invoiced separately from the rates charged as advertisements.

The advertisements **In colour** need the photolits of the original texts, which can be provided directly by the advertiser or requested by Qüestiió to the advertiser charge; in the second case, the advertiser must sent to the journal the originals printed in colour with the best possible quality, so that they can be filmed at the best conditions and guarantees. The cost of the photolits made by Qüestiió will always be charged to the advertiser together with the VAT derived from it, plus the prices corresponding to the type of the advertisement that has been chosen.

The set up of texts and figures of the advertisement should provide the maximum intelligibility and clearness, neither squeezing together the information nor using set ups or letter types that are too small. On the other hand the publicity has to be reliable, without fraud and respectful to the persons and institutions. The direction of Qüestiió has the right of the last decision concerning the insertion of the advertisement.

The advertiser committs himself to give the texts/materials on request, in order to insert them in the issue(s) of Qüestiió that had been previously agreed. The journal is not responsible for any delay from the announcer that could prevent the advertisement from been published in the agreed terms.

Rates:

1 colour page (only one issue):	125.000 PTA + VAT
1 colour page (three consecutive issues):	200.000 PTA + VAT
1 black and white page (only one issue):	30.000 PTA + VAT
1 black and white page (three consecutive issues):	50.000 PTA + VAT
1/2 black and white page (only one issue):	20.000 PTA + VAT
1/2 black and white page (three consecutive issues):	35.000 PTA + VAT

Advertisement sizes (optional):

1 full page (internal space):	19.0 cm. × 12.3 cm.
1 full page (external space):	23.8 cm. × 17.0 cm.
1/2 page (internal space):	9.5 cm. × 12.3 cm.
1/2 page (external space):	11.9 cm. × 17.0 cm.

Payment:

- A bank transfer to account number: 2013-0100-53-0200698577
- A bank cheque to Institut d'Estadística de Catalunya
- Charge on a credit card

The payment should be for the amount shown at the invoice, where it will be shouwn the total cost of the photolits, in case that Qüestiió would be in charge of the filmation of the advertisement. If advertiser need a pro-forma invoice, he should let us know some time in advance so that Qüestiió could send it to the proper address.

Mail address:

Mònica M. Jaime
Secretaria de Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Tel: +34-93 412 15 36
Fax: +34-93 412 31 45
E-mail: questiio@idescat.es

Butlleta de subscripció a la revista *Qüestiió*

Nom i cognoms _____

_____

Empresa/Institució _____

_____

Adreça _____

_____

Codi postal _____ Ciutat _____

Tel. _____ Fax _____ NIF _____

Data _____

Signatura

Desitjo subscriure'm a ***Qüestiió*** per a l'any 2002

Preu de subscripció vigent:

– Estat espanyol: 21,64 €/any (3.600 Pta) (IVA inclòs)
– Estranger: 24,04 €/any (4.000 Pta) (IVA inclòs)

Forma de pagament

☐ Transferència al compte 2013-0100-53-0200698577

☐ Domiciliació bancària al compte número

☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

☐ Xec nominatiu a l'Institut d'Estadística de Catalunya

☐ Gir postal

☐ En efectiu

Retorneu aquesta butlleta (o una fotocòpia) a:

Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Preu de números solts (actuals i endarrerits):

– Estat espanyol: 9,02 €/exemplar (1.500 Pta) (IVA inclòs)
– Estranger: 10,22 €/exemplar (1.700 Pta) (IVA inclòs)

Exemplar per a l'entitat bancària

Autorització de domiciliació bancària per al pagament de les subscripcions anuals de la revista ***Qüestiió***

El sotasignat _____

autoritza el Banc/Caixa _____

Adreça _____

Codi postal _____ Ciutat _____

a abonar les subscripcions a la revista ***Qüestiió*** amb càrrec al seu compte

número ☐☐☐☐ ☐☐☐☐ ☐☐ ☐☐☐☐☐☐☐☐☐☐

Data _____

Signatura

Qüestiió
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Novetats editorials en matèria estadística de la Generalitat de Catalunya
gener-desembre 2001

- **Anuari Estadístic de Catalunya, 1992-2000 CD-ROM**
  gener 2001, 18,03 €
  ISBN 84-393-5332-4

- **Xifres de Catalunya 2001**
  (quadríptic divulgatiu en català, castellà, francés, anglès i alemany)

- **Estadística de població 1996 Vol. 6 Lloc de naixement i nacionalitat de la població. Dades comarcals i municipals**
  octubre 2000, 9,02 €, 274 pp.,
  ISBN 84-393-5264-6

- **Estadística de població 1996. Vol. 11 Professions de la població. Dades comarcals i municipals**
  febrer 2001, 7,51 €, 173 pp.,
  ISBN 84-393-5360-X

- **Estadística de població 1996. Vol. 14 Estructures familiars de la població. Dades comarcals i municipals**
  febrer 2001, 9,02 €, 247 pp.,
  ISBN 84-393-5361-8

- **Moviments migratoris 1998 Dades comarcals i municipals**
  gener 2001, 9,62 €, 184 pp.,
  ISBN 84-393-5341-3

- **Planificació i coordinació de l'estadística catalana**
  desembre 2000, 11,42 €, 350 pp.,
  ISBN 84-393-5201-8

- **Estadística de biblioteques 1998 Característiques bàsiques**
  febrer 2001, 7,51 €, 131 pp.,
  ISBN 84-393-5359-6

- **Localització de l'activitat econòmica 1998. Empreses, professionals, establiments i superfícies**
  febrer 2001, 10,22 €, 365 pp.,
  ISBN 84-393-5364-2

- **Anuari Estadístic del Departament de Política Territorial i Obres Públiques 1998 (CD-ROM)**
  Departament de Política Territorial i Obres Públiques. Secretaria General
  2000, 9,02 €, 246 pp.,
  ISBN 84-393-5332-4

- **Mercat de treball 2000 Ampliació de resultats anuals de l'enquesta de població activa**
  juny 2001, 10,22 €
  ISBN 84-393-5448-7

- **Moviments migratoris 1999 Dades comarcals i municipals**
  juny 2001, 9,62 €
  ISBN 84-393-5447-9

- **Estadística de població 1996 Vol. 12 Fluxos de la mobilitat obligada per treball i estudi. Dades comarcals i municipals**
  maig 2001, 13,52 €
  ISBN 84-393-5423-1

- **Estadística de població 1996 Vol. 13 Localització de l'ocupació laboral. Dades comarcals i municipals**
  maig 2001, 7,51 €
  ISBN 84-393-5424-X

- **Comptes de les administracions públiques de Catalunya 1997**
  maig 2001, 7,81 €
  ISBN 84-393-5414-2

- **Anuari estadístic de Catalunya 2001**
  juliol 2001, 19,23 €, 842 pp.,
  ISBN 84-393-5503-3

- **Anuari Estadístic de Catalunya, 1992-2001 CD-ROM**
  desembre 2001, 19,23 €
  ISBN 84-393-5536-X

- **Estadística, producció i comptes de la indústria 1999**
  octubre 2001, 9 €, 416 pp.
  ISBN: 84-393-5550-5

- **6. Material i equipament elèctric, electrònic i òptic**
  octubre 2001, 6,60 €, 100 pp.
  ISBN: 84-393-5549-1

- **Moviment natural de la població 1999. Dades comarcals i municipals**
  octubre 2001, 8,70 €, 106 pp.
  ISBN: 84-393-5551-3

LLIBRERIES DE LA GENERALITAT

Barcelona
Rambla dels Estudis, 118 (tel. 93 302 64 62)
llibrbcn@correu.cattel.com

Girona
Gran Via de Jaume I, 38 (tel. 972 22 72 67)
llibrgi@ibernet.com

Lleida
Rambla d'Aragó, 43 (tel. 973 28 19 30)
llibrlle@ibernet.com

Madrid
Blanquerna. Llibreria catalana.
Serrano, 1 (tel. 91 431 00 22)
blanquerna@nauta.es

PUNT DE VENDA
Puigcerdà
Plaça del Rec, 5 (tel. 972 88 05 14)

VENDA PER CORREU
Apartat 2800, 08080 Barcelona
eadop@correu.gencat.es
Plaça del Rec, 5 (tel. 972 88 05 14)

Publicacions de la Generalitat. Apartat de correus 2800, 08080 Barcelona

Nom i cognoms _____

Empresa / Institució _____

Professió _____ E-mail _____

Adreça _____

Població _____ CP _____

NIF / DNI _____ Telèfon _____

Desitjo rebre els volums

_____

_____

○ Carregueu l'import a la meva targeta de crèdit Signatura

☐ American Express ☐ 6000

☐ Master Charge ☐ Visa

○ Contra reemborsament

Núm. de targeta |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|

Data de caducitat |_|_|_|_|

DOGC