Qüestió

Quaderns d'Estadística i Investigació Operativa

Any 1996, volum 20, núm. 2
Segona època

Entitats patrocinadores:

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

Entitat col·laboradora:

International Biometric Society

Generalitat de Catalunya
Institut d'Estadística de Catalunya

Any 1996, volum 20, núm. 2

Sumari

EDITORIAL

La consolidació de la nova estructura editorial i ampliació de les seccions de la revista iniciada en el primer número de «**Qüestiió**» del 1996 (Volum 20) té un reflex fidel en els continguts d'aquest segon número, en el qual s'hi apleguen onze articles originals repartits entre les seves quatre seccions temàtiques. En el número present, la secció dedicada a l'Estadística Oficial s'estructura monogràficament a l'entorn de la fiabilitat de la informació estadística, continuant d'aquesta manera l'experiència de números anteriors que apleguen originals centrats en una mateixa problemàtica.

La primera secció més genèrica sobre «Estadística General» conté, en aquesta ocasió, tres articles a l'entorn dels criteris de selecció en arbres de decisió, selecció de models i l'estimació de mitjanes de població seguint models lineals, respectivament. A continuació, la secció «Investigació Operativa» que, per primera vegada s'edita com una secció pròpia de la revista, incorpora dos articles, el primer dels quals tracta l'optimització en presència de solució degenerada, mentre que en el segon es presenta un resultat teòric sobre estabilitat no lineal amb dues variables i restriccions.

Sota el títol genèric d'«Estadística i Fiabilitat», la secció «Estadística Oficial» agrupa quatre treballs que tracten la fiabilitat de la informació estadística. Llevat del primer, que té un caràcter de síntesi, els altres tres originals ho fan en camps i amb mètodes molt diferents, de manera que al comparar-los sorgeix el tema metodològic fonamental següent: el concepte de fiabilitat, emprat en diversos àmbits estadístics i en particular en els treballs indicats, és sempre el mateix? per tal de respondre aquesta pregunta, «Estadística, Fiabilitat i Veritat» d'À. Costa, fonamenta el concepte de fiabilitat en el de veritat, on l'autor ens recorda que la discussió d'aquest darrer no és habitual en moltes comunitats de científics i d'estadístics, però que és necessària per aprofundir l'estudi de la fiabilitat, tal com recull en versió negativa l'acudit popular segons el qual la tercera classe de mentides són les estadístiques; el treball presenta diferents conceptes de veritat que ens ajuden a evitar confusions i formalitza el concepte representacional de Tarsky en un llenguatge lògic. La resta d'articles de la secció continua amb «Estimació de la variança mostral de l'enquesta de població activa» (M. Guillén i X. Martín) on s'estima aquesta variança calculant els errors mostrals d'unes semimostres reiterades, la dispersió de les quals avalua la fiabilitat de l'estimació. A continuació, l'«Estudi dels codis obtinguts per codificació automàtica enfront dels codis precodificats» (M. Delgado i J.A. Sánchez) té per objecte l'anàlisi de la fiabilitat de la precodificació a partir de la comparació de la distribució de freqüències de la variable activitats indicada del Cens de Població de Catalunya de 1991, que s'obté d'una precodificació amb la que es genera una codificació literal feta «a posteriori». En tercer lloc, a «Les revisions de les estimacions de la comptabilitat nacional» (J. Muñoz, E. Pons i J. Pons) s'estudia la fiabilitat de les magnituds macroeconòmiques que formen la Comptabilitat Nacional, obtingudes

com a síntesis comptables; la comparació entre les diverses estimacions provisionals i l'estimació final permet una anàlisi estadística per estudiar la seva fiabilitat.

Per la seva banda, la secció «Biometria» inclou un únic article centrat sobre processos evolutius i estudiant la variabilitat de la divergència mitjançant simulació. Aquest és el cinquè article publicat en aquesta secció i constitueix el primer original publicat sobre aspectes de recerca estadística aplicada a l'àmbit de la biologia. Referent a la Societat de Biometria, l'article «Sobre la Biometría en España» (publicat al Volum 20, núm. 1) contenia una petita errada sobre la composició nominal del Consell Directiu del Grup espanyol de la Societat, la versió correcta de la qual és: Carles M. Cuadras (president), Emilio Carbonell (vice-president), Fernando López Santoveña (secretari), M. Dolores Sánchez Muñoz (tresorer), Juan Luis Chorro (vocal), Rosa Estarelles (vocal i corresponsal del «Biometric Bulletin») i Martín Ríos (vocal). Cal afegir que la «Región Española de la International Biometric Society» disposa d'una pàgina web on s'hi pot recabar més informació mitjançant l'adreça següent: `http://www.iata.csic/IBSREsp/`.

En darrer lloc, la «Secció docent i problemes» inclou la publicació de dos nous enunciats de problemes i un article de F. Borrell a l'entorn de la quantificació d'informació electoral.

Estadística

QÜESTIIÓ, Vol 20, 2 pp. 147-166, 1996

ESTABILIDAD DE ALGUNOS CRITERIOS DE SELECCIÓN DE MODELOS

CARMEN GARCÍA OLAVERRI*

Universidad Pública de Navarra

En este artículo se comparan nueve criterios de selección de modelos. Se analiza si el número de modelos influye en la selección. Se estudia la estabilidad y la robustez de la selección. La comparación se lleva a cabo mediante simulación.

Stability of some model selection criteria

Keywords: Model selection criteria. Robustness.

1. INTRODUCCIÓN

La construcción de modelos estadísticos surge de la necesidad de explicar y predecir el comportamiento de fenómenos reales que dependen de distintas variables. Cuando para una misma evidencia muestral existen modelos alternativos surge el problema de la selección. ¿Cuál es el mejor modelo de todas las alternativas formuladas? ¿Tiene sentido seleccionar un modelo en función del uso posterior que se vaya a dar al mismo? ¿Es siempre necesaria una evaluación extramuestral? Para dar respuesta a las anteriores cuestiones se han definido en la literatura estadística distintos criterios de selección de modelos. Algunos de ellos son muy empleados y los paquetes estadísticos incluyen desde hace tiempo información de este tipo (es el caso de los criterios AIC (Akaike), Cp (Mallows), SBIC(Schwarz), cuyo calculo está incluido en el software de uso más común).

*Carmen García Olaverri. Departamento de Estadística e Investigación Operativa. Universidad Pública de Navarra. 31006. Pamplona.

La autora agradece las observaciones y sugerencias del Profesor C. Cuadras y de un evaluador anónimo.

-Article rebut l'octubre de 1994.

-Acceptat el setembre de 1995.

Los distintos métodos de selección de modelos han sido objeto de comparación en la literatura sin que exista una postura unánime sobre cual es la mejor forma de seleccionar el modelo óptimo. Gran parte de la controversia está basada en el hecho de que no todos los criterios han sido definidos con el mismo fin, es decir para todos los autores la idea de "mejor modelo"no es la misma. Sin embargo, parece claro que hay algunas propiedades deseables que todo criterio de selección debiera satisfacer; esto es, existen algunas formas objetivas de comparar los distintos criterios de selección y concluir cuál de ellos es el mejor , al menos en esa parcela. En este sentido, en la literatura estadística se ha comparado el comportamiento de los criterios de selección cuando cambia el tamaño muestral (Geweke y Meese, 1981), la estructura del proceso que generó los datos (Koehler y Murphree, 1988), el grado de colinealidad entre las variables o la distribución del término de error (Mills y Prasad, 1992; García Olaverri y Aznar, 1994).

Una cuestión que apenas se ha abordado ha sido la de analizar el comportamiento de los criterios según sea el número de modelos presentes en la comparación. Unicamente Geweke y Meese (1981) hacen referencia a esta cuestión, no tanto para estudiar como afecta a la selección el número de alternativas sino para poder comparar criterios aplicados a distintos tamaños muestrales.

Desde nuestro punto de vista, es relevante el saber si un determinado criterio proporcionará distintos resultados según sea el número de alternativas que se formulan en el proceso de selección. Es evidente que para ningún criterio es lo mismo seleccionar un modelo entre 2 alternativas que entre 14, pero parece natural que si se hacen conjeturas acerca de modelos muy alejados del verdadero Proceso Generador de Datos (PGD) su presencia o no en el proceso de selección no debiera provocar variaciones en la selección final.

En el presente trabajo presentamos un estudio comparado del comportamiento de distintos criterios de selección cuando se modifica el número de modelos presentes en la comparación siguiendo el siguiente esquema:

Imaginemos que se dispone de cierta evidencia muestral sobre una variable de nuestro interés (Y_t) y desconocemos cuál ha sido el modelo que la ha generado. Supongamos que se hacen distintas conjeturas sobre cuál es el modelo buscado:

$$\begin{aligned}
\mathbf{M}_1 &: \quad Y_t = \beta_1 X_{1t} + u_{1t} \\
\mathbf{M}_2 &: \quad Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + u_{2t} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\mathbf{M}_k &: \quad Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + u_{kt}
\end{aligned}$$

donde u_{it}, es la perturbación aleatoria de cada modelo ($i = 1, \ldots, k$), que supondremos satisface las condiciones de esfericidad. Las variables X_{it} se supondrán no estocásticas y con ausencia de multicolinealidad exacta.

Supongamos ahora que con esa misma evidencia muestral modificamos el número de modelos presentes en la comparación, ampliándolo e incluyendo en la selección además de $\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_K$, otros modelos $\mathbf{M}_{K+1}, \ldots, \mathbf{M}_P$ cada vez más alejados del verdadero PGD. La pregunta que nos hacemos es ¿Cambiarán en algo los criterios su forma de seleccionar por el hecho de haber incluido las nuevas alternativas ? ¿Hay algunos criterios que cambian más que otros?

Para dar respuesta a las anteriores cuestiones se ha realizado un exhaustivo ejercicio de simulación, del que mostramos algunos resultados, llegándose a la conclusión de que para algunos criterios el resultado final del proceso de selección es muy distinto según sea el número de modelos presentes en la comparación. Es decir, son criterios poco robustos pues si la selección se realiza en un ambiente de incertidumbre, los resultados de ésta serán bastante arbitrarios.

Los criterios de selección de modelos que vamos a comparar son los siguientes: $\overline{R}^2$ (Theil, 1961), Cp (Mallows, 1964), AIC (Akaike, 1973), CAT (Parzen, 1974), BIC (Sawa, 1978), SBIC (Schwarz, 1978), PC (Amemiya,1980), BEC (Geweke y Meese, 1981) y PEC (Aznar y García, 1993). Si la selección se realiza entre m^* modelos anidados: $\mathbf{M}_{m^*} \supset \cdots \supset \mathbf{M}_2 \supset \mathbf{M}_1$, para todos los criterios se trata de elegir aquel valor de $m(m = 1, \ldots, m^*)$, (es decir el modelo con m variables), que minimice las expresiones siguientes:

Tabla 1

Criterio	Elegir $m(m = 1, \ldots, m^*)$ que minimice:
$\overline{R}^2$:	$\hat{\sigma}_m^2 \cdot \dfrac{T}{T-m}$
Cp:	$\hat{\sigma}_m^2 + \dfrac{2m}{T-m^*} \cdot \hat{\sigma}_{m^*}^2$
AIC:	$\ln \hat{\sigma}_m^2 + \dfrac{2m}{T}$
BIC:	$\ln(\hat{\sigma}_m^2) + 2\left(\dfrac{m+2}{T-m^*}\right)\left(\dfrac{\hat{\sigma}_{m^*}^2}{\hat{\sigma}_m^2}\right) - \dfrac{2T}{(T-m^*)^2}\left(\dfrac{\hat{\sigma}_{m^*}^2}{\hat{\sigma}_m^2}\right)^2$
PC:	$\hat{\sigma}_m^2\left(\dfrac{T+m}{T-m}\right)$
SBIC:	$\ln(\hat{\sigma}_m^2) + m\dfrac{\ln T}{T}$
CAT:	$\sum_{j=1}^{m} \dfrac{(T-j)}{T^2} \cdot \hat{\sigma}_j^{-2} - \dfrac{(T-m)}{T} \cdot \hat{\sigma}_m^{-2}$
BEC:	$\hat{\sigma}_m^2 + m\hat{\sigma}_{m^*}^2 \dfrac{\ln T}{T-m^*}$
PEC:	$\left[\dfrac{1}{T_1}\sum_{p=1}^{T_1} \hat{\mathrm{Var}}(\hat{y}_{mp})\right] \cdot \left[\dfrac{1}{T_1}\sum_{p=1}^{T_1} e_{mp}^{*2}\right]$

Donde $\mathbf{M}_m$ y $\mathbf{M}_{m^*}$ indican modelos con m y m^* variables, siendo $\mathbf{M}_{m^*}$ el modelo más amplio de todos los presentes en la comparación; $\hat{\sigma}^2_m, \hat{\sigma}^2_{m^*}$ son los estimadores máximo verosímiles de la varianza del término de error en $\mathbf{M}_m$ y $\mathbf{M}_{m^*}$ respectivamente; T es el tamaño muestral.

En la expresión del criterio PEC, T_1 indica el número de observaciones que han quedado corroboradas (es decir, que pertenecen al intervalo de confianza de su correspondiente predicción); el segundo factor indica el error cuadrático medio de predicción que se comete al predecir, mediante el modelo de m variables, las T_1 últimas observaciones muestrales.

En las anteriores definiciones se observan grandes diferencias y, como consecuencia de ello en ocasiones se seleccionan distintos modelos según sea el criterio considerado. Como ya ha quedado indicado, estas diferencias son debidas a los distintos objetivos con que han sido definidos los criterios de selección. Por ejemplo los criterios SBIC y BEC han sido diseñados con el objetivo de ser consistentes (esto es, tender a seleccionar siempre el PGD a medida que el tamaño muestral aumenta), los criterios PC, Cp tienen como objetivo elegir aquel modelo que minimice el error cuadrático medio de predicción, el criterio AIC persigue seleccionar el modelo más próximo al PGD (en términos de la medida de Kullback - Leibler). Un estudio exhaustivo sobre los distintos objetivos en los procedimientos de selección puede verse en Aznar (1989).

En el problema que nos ocupa, entendemos que la influencia que pueda tener en un criterio el número de modelos en la comparación, va a depender en gran medida, de la tendencia que presente el criterio hacia la selección de modelos distintos al PGD. Es decir, si un criterio es consistente, en el sentido de que nunca tiende a elegir modelos distintos al PGD, el hecho de que en la comparación haya pocos o muchos modelos no afectará al resultado final. Por el contrario, si un criterio muestra tendencia, por ejemplo, hacia modelos más amplios que el PGD el hecho de que pueda elegir entre varias alternativas de este tipo modificará los resultados.

2. ANÁLISIS DE LA TENDENCIA HACIA MODELOS DISTINTOS AL PGD

En este apartado vamos a analizar cuál es la tendencia de los distintos criterios hacia modelos distintos al PGD y en base a los resultados formularemos hipótesis sobre el comportamiento de los criterios cuando cambian los términos de la comparación.

Por similitud en cuanto a la forma funcional estudiaremos separadamente los criterios «clásicos» y el criterio PEC. Consideremos aquellos en primer lugar. En

todos ellos se trata de elegir aquel modelo con m variables que minimice una expresión que siempre depende de $\hat{\sigma}_m^2$, estimador máximo verosímil de la varianza del término de error del modelo considerado.

Supongamos que los datos se han generado con el modelo de k variables $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$; con $\boldsymbol{u} \sim N(0, \sigma_k^2 I)$ donde $\boldsymbol{y}$, y $\boldsymbol{u}$ son vectores $T \cdot 1, \beta$ es un vector $k \cdot 1$ y $\boldsymbol{X}$ es una matriz $T \cdot k$.

Si consideramos otro modelo lineal de m variables: $\boldsymbol{y} = \boldsymbol{Z}\gamma + \boldsymbol{v}$; con $\boldsymbol{v} \sim N(0, \sigma_m^2 I)$ donde $\boldsymbol{y}$, y $\boldsymbol{v}$ son vectores $T \cdot 1$, γ es un vector $m \cdot 1$ y $\boldsymbol{X}$ es una matriz $T \cdot m$ y para este modelo calculamos $\hat{\sigma}_m^2 = \dfrac{\hat{\boldsymbol{v}}'\hat{\boldsymbol{v}}}{T}$ se tiene: $\hat{\boldsymbol{v}}'\hat{\boldsymbol{v}} = (\boldsymbol{y} - \boldsymbol{Z}\hat{\gamma}) = \boldsymbol{y}'\boldsymbol{M}_z\boldsymbol{y}$; con $\boldsymbol{M}_z = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ matriz idempotente; sustituyendo $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}$ se tiene:

$$E(\hat{\sigma}_m^2) = \sigma_k^2 \frac{(T-m)}{T} + \frac{\beta'\boldsymbol{X}'\boldsymbol{M}_z\boldsymbol{X}\beta}{T}$$

$$E(\hat{\sigma}_k^2) = \sigma_k^2 \frac{(T-K)}{T}$$

Si $m < k$, omisión de variables relevantes, siempre se cumple $E(\hat{\sigma}_m^2) > E(\hat{\sigma}_k^2)$ pues $\beta'\boldsymbol{X}'\boldsymbol{M}_z\boldsymbol{X}\beta$ es definida positiva, además $(T-m) > (T-k)$. Por lo tanto la presencia del estadístico $\hat{\sigma}_m^2$ penaliza fundamentalmente la selección de modelos anidados en el PGD; además, a medida que aumenta el tamaño muestral el resto de sumandos o factores que aparecen en la definición de los criterios son irrelevantes cuando se está comparando un modelo anidado en el PGD. Salvo en casos muy excepcionales como por ejemplo que el PGD tenga muchas variables o se disponga de muy reducido tamaño muestral, los denominados criterios clásicos apenas seleccionarán modelos anidados en el PGD. Por lo tanto podemos concluir que, para estos métodos, el número de modelos anidados en el PGD presentes en la comparación es, en general, irrelevante para la selección final.

Por el contrario si $m > k$, inclusión de variables irrelevantes, $E(\hat{\sigma}_m^2) = \sigma_k^2 \dfrac{(T-m)}{T}$ pues el sumando $\beta'\boldsymbol{X}'\boldsymbol{M}_z\boldsymbol{X}\beta$ se anula (si $m > k$ significa que la matriz de datos $\boldsymbol{Z}$ puede expresarse como $(\boldsymbol{X}, \boldsymbol{Z}^*)$ y reescribiendo $\boldsymbol{M}_z$ se obtiene fácilmente el resultado).

En este caso: $E(\hat{\sigma}_m^2) < E(\hat{\sigma}_k^2)$ lo que no significa necesariamente que se vaya a elegir el modelo amplio, pues en primer lugar $\hat{\sigma}_m^2$ aparece en muchos criterios corregido por el factor $(T-m)$ y en segundo lugar en la definición de los criterios aparecen otras expresiones que precisamente penalizan la selección de modelos muy amplios. No obstante, existe un «trade -off» entre esa penalización y el hecho de que el valor de $\hat{\sigma}_m^2$ pueda ser menor, ello implica que la probabilidad de seleccionar modelos más amplios que el PGD no se anula para los criterios clásicos. Unicamente para los criterios SBIC y BEC se ha comprobado en la literatura que asintóticamente esa probabilidad tiende a cero (Geweke y Meese, 1981).

Para muestras finitas, podemos concluir que al existir una tendencia no despreciable hacia modelos amplios, existe también la posibilidad de que la selección cambie en función del número de alternativas propuestas.

Veamos algunas comparaciones analíticas sobre la robustez que, en este aspecto, presentan algunos de los criterios. Si comparamos por ejemplo los criterios SBIC y AIC que presentan una forma funcional parecida se observa que el primero penaliza más que el segundo la inclusión de modelos amplios, con lo cual cabe esperar que SBIC presente un comportamiento más estable.

En efecto, supongamos que habiéndose generado los datos con el modelo de k variables se plantea la selección entre dos modelos con m_2 y m_1 variables con $m_2 > m_1 \geq k$.

$$\text{SBIC}(m_2) = \ln(\hat{\sigma}^2_{m_2}) + m_2 \frac{\ln T}{T}$$

$$\text{SBIC}(m_1) = \ln(\hat{\sigma}^2_{m_1}) + m_1 \frac{\ln T}{T}$$

Llamando $\nabla(\text{SBIC}) = \text{SBIC}(m_2) - \text{SBIC}(m_1)$, podemos expresar que el modelo con m_1 variables resultará elegido siempre que $\nabla(\text{SBIC}) \geq 0$. Es decir si

$$\nabla(\text{SBIC}) = \ln(\hat{\sigma}^2_{m_2}) - \ln(\hat{\sigma}^2_{m_1}) + (m_2 - m_1)\frac{\ln T}{T} \geq 0$$

Análogamente de la definición del criterio AIC, se deduce que según este criterio el modelo con m_1 variables resultará elegido siempre que

$$\nabla(\text{AIC}) = \ln(\hat{\sigma}^2_{m_2}) - \ln(\hat{\sigma}^2_{m_1}) + (m_2 - m_1)\frac{2}{T} \geq 0.$$

Es decir, para tamaños muestrales por encima de $T = 8$ el criterio SBIC penaliza más que el AIC la inclusión de variables irrelevantes.

Procediendo de modo análogo para los criterios BEC y C_p se tiene:

$$\nabla(\text{BEC}) = \hat{\sigma}^2_{m_2} - \hat{\sigma}^2_{m_1} + (m_2 - m_1)\hat{\sigma}^2_{m^*}\frac{\ln T}{T - m^*}$$

$$\nabla C_p = \hat{\sigma}^2_{m_2} - \hat{\sigma}^2_{m_1} + (m_2 - m_1)\hat{\sigma}^2_{m^*}\frac{2}{T - m^*}$$

Al igual que en la comparación anterior, se concluye que BEC penaliza más que C_p la inclusión de modelos amplios para $T > 8$.

Por último, si comparamos las condiciones correspondientes a PC y $\overline{R}^2$ se tiene:

$$\nabla(\text{PC}) = \hat{\sigma}^2_{m_2} - \hat{\sigma}^2_{m_1} + \left(\frac{T+m_2}{T-m_2}\right) - \left(\frac{T+m_1}{T-m_1}\right) \geq 0$$

$$\nabla(\overline{R}^2) = \hat{\sigma}^2_{m_2} - \hat{\sigma}^2_{m_1} + \left(\frac{T}{T-m_2}\right) - \left(\frac{T}{T-m_1}\right) \geq 0$$

siempre que se satisfaga la condición relativa a $\overline{R}^2$ se satisfará la de PC, por tanto éste seleccionará menos veces modelos muy amplios.

De las comparaciones establecidas anteriormente y de algunas equivalencias probadas en la literatura estadística (AIC $\approx$ PC $\approx$ CAT, SBIC $\approx$ BEC asintóticamente) podemos pensar en tres grandes categorías para los criterios clásicos. Por un lado estarían los criterios SBIC y BEC que apenas muestran tendencia a la sobreparametrización o subparametrización y por ello estimamos que no se verán apenas afectados por el número de modelos en la comparación. Por otro lado está el criterio $\overline{R}^2$ que penaliza muy poco la selección de modelos con muchas variables y por tanto, tendrá un comportamiento arbitrario dependiendo del número de modelos. En una situación intermedia estarían el resto de criterios.

En cuanto al criterio PEC su comportamiento es bien distinto del resto de modelos considerados. La definición del estadístico tiene dos partes una que mide la varianza del predictor y otra los errores de predicción cometidos.

Si la selección se plantea entre dos modelos anidados $\mathbf{M}_m \supset \mathbf{M}_k$ la varianza del predictor para el modelo más restringido $\mathbf{M}_k$ es siempre menor que la del modelo amplio $\mathbf{M}_m$, independientemente del proceso que haya generado los datos.

Si utilizamos el modelo amplio para predecir definimos el predictor como:

$$\hat{y}_{mp} = \boldsymbol{z}_p'\hat{\gamma} \quad \text{con} \quad \hat{\gamma} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

donde $\boldsymbol{z}_p'$ es un vector fila con m elementos cuyo correspondiente valor de y queremos predecir.

Si la predicción se hace con el modelo restringido el predictor será:

$$\hat{y}_{kp} = \boldsymbol{x}_p'\hat{\beta} \quad \text{con} \quad \hat{\beta} = (\boldsymbol{X}' - \boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

donde $\boldsymbol{x}_p'$ es un vector fila con k elementos Si el proceso generador de datos ha sido el modelo restringido se tiene que:

$$\text{Var}(\hat{y}_{kp}) = \sigma_k^2\boldsymbol{x}_p'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_p < \text{Var}(\hat{y}_{mp}) = \sigma_k^2\boldsymbol{z}_p'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_p$$

Si por el contrario los datos se hubieran generado con el modelo amplio:

$$\text{Var}(\hat{y}_{kp}) = \sigma_m^2\boldsymbol{x}_p'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_p < \text{Var}(\hat{y}_{mp}) = \sigma_m^2\boldsymbol{z}_p'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{z}_p$$

Obsérvese que al ser $m > k$, $Z = (X, Z^*)$ y $z_p' = (x_p', z_p^{*'})$ pudiendo escribirse:

$$z_p'(Z'Z)^{-1}z_p = x_p'(X'X)^{-1}x_p + C, \quad \text{con } C \text{ escalar positivo}$$

Por lo tanto, el primer factor en la definición del criterio PEC propicia el que se seleccionen modelos anidados en el PGD, penalizando únicamente la inclusión de variables irrelevantes.

En cuanto al segundo factor, denotando $e_{mp} = y_p - \hat{y}_{mp}$; $e_{kp} = y_p - \hat{y}_{kp}$ a los errores que se cometen al predecir con el modelo amplio y el restringido respectivamente se tiene que: si los datos fueron generados con el modelo restringido $\mathrm{E}(e_{mp}) = \mathrm{E}(e_{kp}) = 0$; mientras que si fueron generados con el amplio $\mathrm{E}(e_{kp}) \neq \mathrm{E}(e_{mp}) = 0$. Por esta razón este segundo factor penaliza los modelos restringidos en favor de los más amplios. No obstante el hecho de exigir previamente la fase de corroboración (acotación para los errores de predicción de datos muestrales) elimina cualquier modelo muy alejado del PGD.

En síntesis, cabe esperar que al realizar distintas comparaciones, incluyendo modelos cada vez más amplios que contienen al verdadero PGD, los criterios SBIC, BEC y PEC apenas modificarán la selección, los criterios AIC, PC, Cp, BIC y CAT lo harán de forma moderada y el criterio $\overline{R}^2$ mostrará gran variabilidad en la selección. Veamos a través de un ejercicio de simulación como se confirman algunos de estos resultados.

3. RESULTADOS DEL EJERCICIO DE SIMULACIÓN

El ejercicio de simulación se ha realizado mediante la elaboración de un programa en lenguaje FORTRAN utilizando la librería estadística IMSL, la descripción sobre la forma de generar las observaciones, el procedimiento de estimación y los criterios de selección empleados puede verse en (Garcia Olaverri, Aznar ,1994).

Para efectuar la selección se comparan m^* modelos anidados, entre los que siempre se encuentra uno con el mismo número de regresores que el PGD. Nuestro objetivo es analizar si los criterios seleccionan de distinta forma cuando se modifica m^*.

El ejercicio de simulación se realiza dentro del siguiente esquema:

En primer lugar se generan observaciones para las variables X_t, según el modelo AR(2):

$$X_t = 1.6X_{t-1} - 0.64X_{t-2} + \xi_t \quad \text{con} \quad \xi_t \sim \mathrm{N}(0,1).$$

En segundo lugar, se obtienen las observaciones de la variable Y_t a través del PGD.

Presentamos los resultados correspondientes a dos posibles PGD:

$$Y_t = X_t + u_t$$

$$Y_t = 1.25X_t + 1X_{t-1} + 0.8X_{t-2} + u_t \quad \text{con} \quad u_t \sim \text{N}(0;0.5)$$

en ambos modelos.[1]

Por último, se estiman distintos modelos para el conjunto de observaciones y se van aplicando los criterios de selección de modelos. Para los dos PGD anteriores realizamos distintos procesos de selección en los que se va modificando el número de modelos presentes en la comparación. Mostramos a continuación algunos resultados:

Por ejemplo, generando los datos con el modelo $Y_t = X_t + u_t$, utilizando una muestra de tamaño $T = 100$, var$(u_t) = 0.5$ y realizando 100 iteraciones, se obtuvieron los resultados que se muestran en la Tabla 2.

Como puede observarse, independientemente del objetivo con que fueran construídos los criterios, algunos de ellos muestran la propiedad de que apenas se ven afectados por el número de modelos presentes en la comparación (siempre que entre ellos haya un modelo con las mismas variables que el PGD), en esta situación están los criterios SBIC, BEC y PEC. En el extremo opuesto aparecen criterios como el $\overline{R}^2$ que varía de forma notable la selección cuando se incluyen más modelos en la comparación, aun cuando estén realmente alejados del PGD. En una situación intermedia se encuentran los criterios BIC, AIC, PC, C_p y CAT que modifican bastante la selección cuando se pasa de 2 a 5 modelos, pero apenas cambian cuando se formulan alternativas muy alejadas del PGD (comparación entre 10 o más modelos).

El ejercicio se ha realizado para distintos modelos, tamaños muestrales y número de alternativas en la comparación. Con el fin de mostrar lo más relevante de estos resultados presentamos en la Tabla 3 información sobre el número de veces que, en 100 iteraciones, se eligió el PGD según los distintos criterios, variando el número de modelos, m^*, presentes en la comparación, así como el tamaño muestral. La fila correspondiente a $m^* = 2$ indica el porcentaje de veces que se eligió el modelo con un

[1] Para generar las X_t se han probado otros modelos AR(2): $X_t = \rho_1 X_{t-1} + \rho_2 X_{t-2} + \xi_t$; modificando ρ_1, ρ_2 y Var (ξ_t). En general, cuanto mayor es la varianza muestral de las X_t aparece una tendencia más clara a seleccionar el PGD, sin embargo el grado de cumplimiento de la propiedad que nos ocupa (estabilidad de los criterios) se mantiene, es decir si un criterio es muy inestable cuando cambia el número de modelos en la comparación, seguirá siéndolo independientemente de los valores que tomen ρ_1, ρ_2 y Var(ξ_t); por tanto las tablas que mostramos si bien se refieren a una forma concreta de generar las variables X_t, son representativas de un conjunto más amplio de posibltes situaciones.

regresor (PGD) cuando se comparó únicamente con otro modelo (de dos regresores). La fila correspondiente a $m^* = 5$ indica el porcentaje de veces que se eligió el PGD cuando se comparó con otros cuatro modelos, cada uno de ellos incluyendo una variable más que el anterior. Omitimos por tanto, el detalle de cuantas veces se seleccionaron modelos distintos al PGD, pero la tendencia es, en todos los casos muy parecida a la mostrada en la Tabla 2.

Por otras simulaciones realizadas en trabajos anteriores es sabido que de todos los criterios considerados, sólo el PEC muestra alguna tendencia a la subparametrización, pues como ha quedado indicado, al depender la selección de la varianza del predictor, penaliza claramente la selección de modelos con muchas variables. En este primer ejercicio como no es posible comparar el PGD con modelos anidados en él, el criterio PEC muestra unos resultados óptimos pero su comportamiento general no es así. De hecho, como se verá en el siguiente ejemplo, cuando intervienen en la comparación modelos anidados en el PGD, el criterio PEC tiende a seleccionarlos de forma no despreciable.

Al final de cada columna se indica entre paréntesis la desviación típica de la variable «número de veces que, en 100 iteraciones, se eligió el PGD».

Como es natural una desviación típica pequeña nos informa de que la selección apenas se ve afectada por el hecho de modificar el número de modelos alternativos. Por el contrario, una desviación típica grande muestra la poca robustez del criterio ante cambios en el número de modelos a comparar.

Es interesante observar cómo los distintos criterios mejoran (en términos del número de veces que se elige el PGD) cuando aumenta el tamaño muestral, pero no debiera obviarse el hecho de que criterios tan empleados como AIC seleccionan un 24% de las veces modelos distintos al PGD cuando la selección se plantea entre 5 alternativas, incluso para un tamaño muestral de 400. Para este mismo tamaño muestral el número de veces que no se selecciona el PGD se reduce al 16% si sólo se comparan dos modelos.

Como es de esperar para tamaños muestrales pequeños las fluctuaciones en la selección son más evidentes, así para $T = 60$ se observa que la diferencia entre $\overline{R}^2$ y C_p o CAT es de un 19% cuando la selección es entre dos modelos pero alcanza más de un 40% cuando se comparan 16 especificaciones alternativas.

Los criterios SBIC, BEC y PEC se muestran claramente como los más robustos, la selección mejora a medida que aumenta el tamaño muestral, pero apenas se modifica si para un mismo tamaño muestral se formulan más alternativas.

Tabla 2

Porcentaje de veces que se elige un modelo con m variables (Habiéndose generado los datos con $m = 1$; PGD: $Y_t = X_t + u_t$)

$T = 100$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	66	81	78	78	78	78	96	96	100
$m = 2$	34	19	22	22	22	22	4	4	0

$T = 100$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	43	74	70	70	73	70	96	96	100
$m = 2$	17	13	15	15	13	15	4	4	0
$m = 3$	10	7	7	7	7	7	0	0	0
$m = 4$	10	2	3	3	2	3	0	0	0
$m = 5$	20	4	5	5	5	5	0	0	0

$T = 100$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	28	71	66	66	68	66	96	96	100
$m = 2$	10	14	14	14	13	14	4	4	0
$m = 3$	5	8	6	6	7	6	0	0	0
$m = 4$	5	1	3	3	2	3	0	0	0
$m = 5$	5	2	3	3	3	3	0	0	0
$m = 6$	2	1	1	1	0	1	0	0	0
$m = 7$	5	0	0	0	0	0	0	0	0
$m = 8$	14	1	3	3	3	3	0	0	0
$m = 9$	15	0	1	1	1	2	0	0	0
$m = 10$	11	2	3	3	3	2	0	0	0

$T = 100$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	20	72	65	65	68	65	96	96	100
$m = 2$	7	15	14	14	12	14	4	4	0
$m = 3$	4	8	6	6	6	6	0	0	0
$m = 4$	5	0	3	3	2	3	0	0	0
$m = 5$	3	0	3	3	3	3	0	0	0
$m = 6$	2	1	1	1	1	1	0	0	0
$m = 7$	3	0	0	0	0	0	0	0	0
$m = 8$	8	0	3	3	3	3	0	0	0
$m = 9$	9	0	1	1	1	1	0	0	0
$m = 10$	3	2	2	2	2	2	0	0	0
$m = 11$	6	1	1	1	1	1	0	0	0
$m = 12$	4	1	0	0	1	1	0	0	0
$m = 13$	5	0	0	0	0	0	0	0	0
$m = 14$	5	0	0	0	0	0	0	0	0
$m = 15$	8	0	1	1	0	0	0	0	0
$m = 16$	8	0	0	0	0	0	0	0	0

Tabla 3

Porcentaje de veces que se elige un modelo con las mismas variables que el PGD en funcion de m^* (numero de modelos anidados presentes en la comparacion)

PGD: $Y_t = X_t + u_t$

$T = 60$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$	58	78	77	77	77	77	93	93	98
$m^* = 5$	34	65	61	61	62	62	93	93	98
$m^* = 10$	24	63	57	58	60	60	93	92	96
$m^* = 16$	18	66	56	57	60	60	93	91	96
	(15.25)	(5.87)	(8.43)	(8.07)	(7.12)	(7.12)	(0.0)	(0.83)	(1)

$T = 200$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$	74	84	83	84	84	84	97	97	98
$m^* = 5$	38	74	73	73	73	73	97	97	98
$m^* = 10$	23	72	68	68	69	68	97	97	98
$m^* = 16$	18	71	66	66	68	67	97	97	97
	(21.91)	(5.16)	(6.57)	(6.98)	(6.34)	(6.75)	(0.0)	(0.0)	(0.43)

$T = 400$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$	72	85	84	84	85	85	98	98	100
$m^* = 5$	45	76	76	76	76	76	98	98	100
$m^* = 10$	33	73	73	73	73	73	98	98	98
$m^* = 16$	24	73	73	73	73	73	98	98	97
	(18.06)	(4.92)	(4.5)	(4.5)	(4.92)	(4.92)	(0.0)	(0.0)	(1.29)

Si observamos las fluctuaciones de la selección a través de las desviaciones típicas (escritas al final de cada tabla) se observa con nitidez la robustez de los distintos criterios. Insistimos en la cautela con que deben ser interpretados los resultados correspondientes al criterio PEC cuando, como en este caso no hay modelos anidados en el PGD.

Las Tablas 4 y 5 nos muestran los resultados de la simulación cuando la selección se plantea conjuntamente para modelos más amplios y más restringidos que el PGD. Consideremos el caso de que el PGD sea $Y_t = 1.25X_t + 1X_{t-1} + 0.8X_{t-2} + u_t$. El resto de variables (varianza, semilla, etc) son idénticas a las empleadas en el anterior ejercicio.

Tabla 4

Porcentaje de veces que se elige un modelo con m variables (Habiéndose generado los datos con m = 3;

$$Y_t = 1.25X_T + X_{t-1} + 0.8X_{t-2} + U_t)$$

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 2$	0	0	0	0	0	0	0	0	11
$m = 3$	100	100	100	100	100	100	100	100	87

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 3$	73	89	89	89	89	89	98	98	99
$m = 4$	27	11	11	11	11	11	2	2	1

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	0	0	0	0	0	0	0	0	0
$m = 2$	0	0	0	0	0	0	0	0	11
$m = 3$	57	78	78	78	78	78	97	97	88
$m = 4$	16	8	8	8	8	8	2	2	1
$m = 5$	27	14	14	14	14	14	1	1	0

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	0	0	0	0	0	0	0	0	0
$m = 2$	0	0	0	0	0	0	0	0	11
$m = 3$	31	73	68	68	69	68	97	97	89
$m = 4$	8	6	7	7	7	7	2	2	0
$m = 5$	6	10	9	9	9	9	1	1	0
$m = 6$	3	3	4	4	4	4	0	0	0
$m = 7$	7	1	3	3	3	3	0	0	0
$m = 8$	15	4	4	4	4	5	0	0	0
$m = 9$	18	1	2	2	2	2	0	0	0
$m = 10$	12	2	3	3	2	2	0	0	0

$T = 100$	$\bar{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m = 1$	0	0	0	0	0	0	0	0	0
$m = 2$	0	0	0	0	0	0	0	0	11
$m = 3$	23	74	68	68	69	68	97	97	89
$m = 4$	7	6	7	7	7	7	2	2	0
$m = 5$	3	8	8	8	9	8	1	1	0
$m = 6$	3	3	4	4	3	4	0	0	0
$m = 7$	4	1	3	3	2	3	0	0	0
$m = 8$	8	3	4	4	5	5	0	0	0
$m = 9$	11	1	2	2	1	1	0	0	0
$m = 10$	4	0	2	2	2	2	0	0	0
$m = 11$	6	1	1	1	1	1	0	0	0
$m = 12$	4	1	0	0	1	1	0	0	0
$m = 13$	5	0	0	0	0	0	0	0	0
$m = 14$	5	0	0	0	0	0	0	0	0
$m = 15$	8	0	1	1	0	0	0	0	0
$m = 16$	9	0	0	0	0	0	0	0	0

Tabla 5

Porcentaje de veces que se elige un modelo con las mismas variables que el PGD en funcion de m^*

$$PGD: \quad Y_t = 1.25X_t + X_{t-1} + 0.8X_{t-2} + u_t$$

$T = 60$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$ (2 vs.3)	100	100	100	100	100	100	100	100	82
$m^* = 2$ (3 vs.4)	72	82	79	79	80	80	92	94	96
$m^* = 5$	62	77	74	74	75	75	89	90	83
$m^* = 10$	31	73	61	61	64	63	89	88	83
$m^* = 16$	26	74	56	57	62	60	89	88	83
	(27.31)	(9.91)	(15.45)	(15.22)	(13.66)	(14.26)	(4.26)	(4.56)	(5.31)

$T = 200$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$ (2 vs.3)	100	100	100	100	100	100	100	100	93
$m^* = 2$ (3 vs.4)	72	82	82	82	82	82	96	96	98
$m^* = 5$	57	72	70	70	71	70	96	96	90
$m^* = 10$	29	66	64	64	65	64	96	96	91
$m^* = 16$	20	63	61	61	62	61	96	96	90
	(29.04)	(13.38)	(14.29)	(14.29)	(13.81)	(14.29)	(1.6)	(1.6)	(3)

$T = 400$	$\overline{R}^2$	BIC	AIC	PC	C_p	CAT	SBIC	BEC	PEC
$m^* = 2$ (2 vs.3)	100	100	100	100	100	100	100	100	95
$m^* = 2$ (3 vs.4)	68	85	84	84	84	84	100	100	98
$m^* = 5$	50	76	74	74	74	74	98	98	96
$m^* = 10$	35	70	69	69	70	69	98	98	96
$m^* = 16$	24	70	69	69	70	69	98	98	96
	(26.76)	(11.32)	(11.75)	(11.75)	(11.41)	(11.75)	(0.98)	(0.98)	(0.98)

Cuando indicamos $m^* = 2$ (2 vs. 3) nos referimos a la comparación entre dos modelos lineales anidados de 2 y 3 variables respectivamente, cuando indicamos $m^* = 2$ (3 vs. 4) nos referimos a la comparación entre dos modelos anidados de 3 y 4 variables. Como puede observarse, en la primera fila de cada tabla, excepto el criterio PEC que muestra cierta tendencia a la subparametrización, el resto de métodos seleccionan sin excepción el PGD. La situación cambia claramente cuando la selección se establece entre el PGD y un modelo más amplio que él (fila 2 de cada tabla).

Las filas restantes corresponden a la comparación de m^* modelos lineales anidados de $1,2,3,\ldots,m^*$ variables respectivamente. Al igual que en ejemplo anterior, mientras algunos criterios se ven muy poco afectados por el número de modelos en la comparación, otros cambian significativamente el resultado de la selección. Así, los criterios AIC, PC, C_p, CAT, son poco robustos en este sentido y el $\overline{R}^2$ nada robusto. Además para tamaños muestrales pequeños ($T = 60$), criterios como BIC, CAT o C_p pueden parecer semejantes si la selección se hace entre dos modelos, pero en un hipotético ambiente de más incertidumbre donde la selección se estableciera entre 10

ó más modelos, el criterio BIC es superior a los otros dos. Claramente por encima de todos ellos se sitúan los criterios, SBIC, BEC y PEC.

Hay que tener en cuenta que el hecho de realizar la selección entre modelos con muchas variables, puede plantear problemas de grados de libertad, que podrían eliminarse aumentando el tamaño muestral; queremos señalar, no obstante, que aun estando ambas cuestiones (tamaño muestral y tamaño del modelo) íntimamente relacionadas hay que separar los dos problemas.

Si observamos la selección desde un punto de vista verificacionista en el que exclusivamente estuviéramos interesados en conocer cuántas veces se elige cada modelo, es evidente que los criterios PEC, SBIC y BEC mantienen fija la selección, mientras que otros van disminuyendo el número de veces que se elige el PGD a medida que aumenta el número de modelos en la comparación.(obsérvense por ejemplo, las columnas correspondientes al criterio $\overline{R}^2$).

Entendemos que esta robustez relativa al número de modelos que entran en la comparación es una buena propiedad de los criterios PEC, SBIC y BEC frente a los otros métodos de selección.

BIBLIOGRAFÍA

[1] **Akaike, H.** (1969). «Fitting Autoregressive Models for Prediction». *Annals of Institute of Statistical Mathematics*, bf 21, 243–247.

[2] **Akaike, H.** (1974). «A New Look at the Statistical Model Identification». *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

[3] **Amemiya, T.** (1980). «Selection of Regressors». *International Economic Review*, **21**, 331–354.

[4] **Aznar, A.** (1989). *Econometric Model Selection: A New Approach*. Dordrecht: Kluwer Academic Publishers.

[5] **Chow, G.C.** (1981). «A Comparison of the Information and Posterior Probability Criteria for Model Selection». *Journal of Econometrics*, **16**, 21–33.

[6] **Geweke, J.** y **Meese, R.** (1981). «Estimating Regression Models of Finite but Unknown Order». *International Economic Review*, **22**, 55–70.

[7] **García Olaverri, C.** y **Aznar, A.** (1994). «Estudio comparado de la robustez de distintos criterios de selección de modelos econométricos ante cambios en la varianza». *Estadística Española*, **36, 136**, 287–318.

[8] **García Olaverri, C.** (1993). *El Criterio ACOR: nuevos desarrollos teóricos y resultados de un estudio de Monte-Carlo.* Tesis Doctoral. Universidad de Zaragoza.

[9] **Koehler, A. B.** y **Murphree, B. S.** (1988). «A Comparison of the Akaike and Schwarz Criteria for selecting Model Order». *Applied Statistics*, **37**, 187–195.

[10] **Mallows, C.L.** (1973). «Some Comments on C_p». *Technometrics*, **15**, 661–676.

[11] **Mills, J.A.** y **Prasad, K.** (1992). «A Comparison of Model Selection Criteria». *Econometric Reviews*, **11**, 201–223.

[12] **Parzen, E.** (1974). «Some Recent Advances in Time Series Analysis». *IEEE Transactions on Automatic Control*, **AC-19**, 723–730.

[13] **Sawa, T.** (1978). «Information Criteria for Discriminating among Alternative Regression Models». *Econometrica*, **46**, 1273–1282.

[14] **Schwarz, G.** (1978). «Estimating the dimension of a Model». *Annals of Statistics*, **6**, 461–464.

[15] **Theil, H.** (1961). *Economic Forecasts and Policy*. Amsterdam: North Holland.

ENGLISH SUMMARY:

STABILITY OF SOME MODEL SELECTION CRITERIA

Carmen García Olaverri

In this paper we compare the behavior of nine model selection criteria. The aim of the study is to analyze if the number of models in the comparison has any influence on the results of the selection procedure. Are the results of selection the same if we compare 2 or 10 models? Are all the selection criteria similar behavior in this aspect? To answer these questions we conduct a Monte - Carlo simulation study.

Let us suppose that we have sample information relative to the variables $Y_t, X_{1t}, X_{2t}, \ldots, X_{mt}$.

There are some relationships that can be established between the former set of variables, if we don't know what is the «best» specification we can formulate some alternative models:

$$\mathbf{M}_1: \quad Y_t = \beta_1 X_{1t} + u_{1t}$$
$$\mathbf{M}_2: \quad Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + u_{2t}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\mathbf{M}_{m^*}: \quad Y_t = \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_m^* X_{mt}^* + u_{mt}^*$$

and then, select the better model by means of a criterion.

The criteria to be compared are the following: (Theil, 1961), Cp (Mallows, 1964), AIC (Akaike, 1973), CAT (Parzen, 1974), BIC (Sawa, 1978), SBIC (Schwarz, 1978), PC (Amemiya, 1980), BEC (Geweke and Meese, 1981), PEC (Aznar y García 1994). To establish the comparison we have adopted the following notation:

The selection consists in choose one of the m^* linear and nested models $\mathbf{M}_{m^*} \supset \cdots \supset \mathbf{M}_2 \supset \mathbf{M}_1$, being:

- m: the number of variables contained in $\mathbf{M}_m$ model
- m^*: the number of variables contained in the biggest model considered $\mathbf{M}_{m^*}$.
- $\hat{\sigma}_m^2, \hat{\sigma}_{m^*}^2$: maximum likelihood estimates of the variance on the error term in $\mathbf{M}_m$ and $\mathbf{M}_{m^*}$ respectively.
- T: sample size.

Using this notation, the expressions of the criteria to be compared are the following:

Choose the m value ($m = 1, \ldots, m^*$) that minimizes:

- ***$\overline{R}^2$ Criterion:***

$$\hat{\sigma}_m^2 \cdot \frac{T}{T-m}$$

- ***C_p Criterion:***

$$\hat{\sigma}_m^2 + \frac{2m}{T-m^*} \cdot \hat{\sigma}_{m^*}^2$$

- ***AIC Criterion:***

$$\ln \hat{\sigma}_m^2 + \frac{2m}{T}$$

- ***BIC Criterion:***

$$\ln(\hat{\sigma}_m^2) + 2\left(\frac{m+2}{T-m^*}\right)\left(\frac{\hat{\sigma}_{m^*}^2}{\hat{\sigma}_m^2}\right) - \frac{2T}{(T-m^*)^2}\left(\frac{\hat{\sigma}_{m^*}^2}{\hat{\sigma}_m^2}\right)^2$$

- ***PC Criterion:***

$$\hat{\sigma}_m^2\left(\frac{T+m}{T-m}\right)$$

- ***SBIC Criterion:***

$$\ln(\hat{\sigma}_m^2) + m\frac{\ln T}{T}$$

- ***CAT Criterion:***

$$\sum_{j=1}^{m} \frac{(T-j)}{T^2} \cdot \hat{\sigma}_j^{-2} - \frac{(T-m)}{T} \cdot \hat{\sigma}_m^{-2}$$

- ***BEC Criterion:***

$$\hat{\sigma}_m^2 + m\,\hat{\sigma}_{m^*}^2 \frac{\ln T}{T-m^*}$$

- ***PEC Criterion:***

$$\left[\frac{1}{T_1}\sum_{p=1}^{T_1} \hat{\mathrm{Var}}(\hat{y}_{mp})\right] \cdot \left[\frac{1}{T_1}\sum_{p=1}^{T_1} e_{mp}^{*2}\right]$$

where the second factor indicates the mean squared prediction error corresponding to the last T_1 observations.

Following the definition of the criteria we can observe that there are some methods that assign a high penalty to the models with many variables. For these critera the number of models in the comparison will be irrelevant.

For example SBIC penalizes the big models more than AIC; BEC more than C_p and PC more than $\overline{R}^2$. So, we can expect a different behavior of the criteria.

SIMULATION RESULTS

A Monte-Carlo experiment is conducted in order to study the behavior of the nine compared criteria. The program was written in FORTRAN language using IMSL library. Robustness of the experiment was tested previously.

Here we have limited our attention to two models as Data Generating Process (DGP), one hundred replications were generated from each of the following models:

$$\begin{aligned} Y_t &= X_t + u_t \\ Y_t &= 1.25X + X_{t-1} + 0.8X_{t-1} + u_t \end{aligned}$$

The generating model for the regressors is $X_t = 1.6X_{t-1} - 0.64X_{t-2} + \xi_t$ where ξ_t is a standard normal variable with mean zero and a variance that can take different values. (Following a experiment from Geweke and Meese). We consider four possible sample sizes 60, 100, 200 and 400.

Once the data are generated we estimate m^* nested models (one of them containing the same variables as the true DGP) and we select the «best» model using the different selection criteria.

The results are summarized in TABLES 2-5. Tables 2 and 3 show the selection when the DGP is $Y_t = X_t + u_t$. Tables 4 and 5 show the results when DGP is the model $Y_t = 1.25X + X_{t-1} + 0.8X_{t-1} + u_t$.

For each criterion we evaluate how many times the DGP is selected and we repeat the selection changing the number of the models in the comparison. Tables 2 and 4 are refered to $T = 100$ and show the number of times that every model on the comparison has been selected. Tables 3 and 5 are relative to sample sizes $T = 60$, 200 and 400, and show the number of times that the DGP has been selected, m^* indicates the number of models being compared.

As can be seen there are some criteria as SBIC, BEC or PEC that rarely change the results of the comparison when the number of models is modified. So we can

conclude that they are robust in this aspect. On the other hand the $\overline{R}^2$ criterion shows a poor behavior. The BIC, AIC, PC, C_p and CAT criteria present an intermediate situation; when the comparison is established between two models they can seem similar, but when the number of models increases the BIC criterion is better than the other ones (see the correponding table to $T = 60$ in Table 5).

If we suppose that we have to select a model in a uncertain environment and we decide to establish the comparison between a few models, it is interesting to know what are the criteria that are «independent» on the number of models in the comparison and which are not. In this framework we conclude that the criteria having this good property are: SBIC, BEC and PEC; being the $\overline{R}^2$ criterion the worse.

Working with the simulation experiment we get another results well known in the literature of model selection criteria. The consistency of SBIC and BEC criteria; the tendency to overfit that does not vanish in AIC, PC, C_p, and CAT criteria being around 30% of selections, even when the sample size increases; and the fact that PEC criterion is more parsimonious than the other criteria.

QÜESTIIÓ, Vol 20, 2 pp. 167-186, 1996

CRITERIO DE SELECCIÓN DE UN ÁRBOL ÓPTIMO SEGÚN COEFICIENTES DE ASOCIACIÓN DERIVADOS DE χ^2

F.J. CANO SEVILLA*

A. MUNDUATE DEL RIO**

A. PÉREZ PRADOS***

Se analiza en primer lugar la variación que se produce en el valor del coeficiente de contingencia al realizarse un proceso de poda en un árbol de decisión T. Conocido este efecto, se define una cantidad criterio que combina linealmente el coeficiente de contingencia con el índice de simplicidad. A partir de esta cantidad criterio, se propone un método de obtención de un árbol óptimo para cada uno de los distintos valores del parámetro α de la combinación lineal. Para seleccionar el árbol óptimo, entre todos ellos, se utiliza el coeficiente de Tschuprow, dependiente de las dos medidas consideradas para la calidad del árbol.

Criterion for the selection of an optimum tree considering association coefficients obtained from the χ^2

Keywords: Árboles de decisión, proceso de poda, coeficiente de contingencia, coeficiente de Tschuprow, simplicidad, árbol óptimo.

Clasificación AMS: 62H30

* Dep. de Estadística e Investigación Operativa. Universidad Complutense de Madrid.

** Dep. de Física de Materiales. Universidad del País Vasco.

*** Dep. de Estadística e Investigación Operativa. Universidad Pública de Navarra.

– Article rebut el desembre de 1994.

– Acceptat el gener de 1996.

1. INTRODUCCIÓN

Es sabida la existencia de diversos métodos para la construcción de árboles de decisión, obtenidos a partir de un conjunto de datos para las variables cualitativas $\{V^j\}_{j=1,\ldots,J}$ así como para la variable Y, que se relaciona con las anteriores, definidas todas ellas sobre el conjunto de aprendizaje I.

Se conoce también la validez de los coeficientes de asociación obtenidos a partir del estadístico χ^2 (coeficiente de contingencia, coeficiente de Tschuprow y coeficiente de Cramer) para medir la utilidad de un árbol como predictor de la variable criterio Y. Considerando los nodos terminales de T como modalidades de una nueva variable, dichos coeficientes de asociación calculados sobre la matriz que cruza las modalidades de esta nueva variable con las de la variable criterio Y, son una medida de la asociación existente entre T e Y y en consecuencia de la utilidad de T como predictor de Y.

Por otra parte, diferentes criterios para determinar el árbol óptimo entre los construídos a partir de una colección de datos pueden encontrarse en la literatura. Así, por ejemplo, el criterio de complejidad (Breiman, 1984) utiliza una cantidad criterio que combina el error de resustitución y la simplicidad del árbol medida a través del número de sus nodos terminales; el criterio del error esperado (Niblett, 1987) se basa en la probabilidad de error al asignar un nuevo ejemplo a una modalidad de Y en un nodo x de T. El criterio de contribución (Cuesta, 1989) es una generalización del criterio de complejidad considerando la contribución de cada nodo interior a la calidad global del árbol.

Teniendo en cuenta todo ello, y conocida la variación que un proceso de poda produce en el valor del estadístico χ^2 (Pérez Prados y otros, 1994) se estudia la modificación que ésta induce en el coeficiente de contingencia. A la vista de este resultado se plantea la utilización de este coeficiente para la definición de una nueva cantidad criterio para la selección del árbol óptimo.

Esta cantidad criterio combina linealmente dos medidas de la calidad del árbol: el coeficiente de contingencia y la simplicidad y permite construir un proceso de selección del árbol óptimo, de forma que para cada valor del parámetro α de la combinación lineal se obtiene un árbol óptimo. Además el coeficiente de Tschuprow que depende tanto del estadístico χ^2 como del número de nodos terminales de T permite seleccionar entre los anteriores. Con lo cual mediante este planteamiento se consideran dos de los coeficientes de asociación obtenidos a partir de χ^2 como medidas de la calidad, junto con la simplicidad medida a través del número de nodos terminales.

2. CONCEPTOS FUNDAMENTALES

2.1. Estructura en árbol

Sea un conjunto de variables $\{V^j\}_{j=1,\ldots,J}$, llamadas *variables explicativas*, cuyos valores son conocidos para los n elementos de un conjunto I, llamado *conjunto de aprendizaje*, extraído de una población total $\mathfrak{I}$. Relacionada con ellas se considera una variable Y, llamada *variable criterio* también conocida para los elementos de I; se supone que esta variable es cualitativa y el conjunto de sus modalidades se representa por $Y = \{y_k\}_{k=1,\ldots,c}$.

Partiendo de los datos conocidos para las variables anteriores puede construirse lo que se denomina *estructura en árbol*, que es aquella que presenta distintos niveles de asociación de los elementos del conjunto de aprendizaje, correspondientes a diferentes grados de homogeneidad, de acuerdo con la información dada por el conjunto de variables explicativas.

En un árbol T los *nodos o vértices* se corresponden con subconjuntos de I. Se representa por x un nodo cualquiera de T, siendo n_x el número de ejemplos de I situados en x; cada una de las ramas que partiendo de un vértice llega directamente a otro es un *arco* y una sucesión de arcos consecutivos se llama *camino.*

Dados dos nodos x_1 y x_2 de un árbol T, si existe un arco que partiendo de x_1 llega a x_2, se dice que x_1 es *nodo generador* de x_2 y x_2 es *nodo sucesor* de x_1.

Son *nodos terminales* de un árbol T aquellos que no tienen nodos sucesores. Todos los demás nodos de T se llaman *nodos interiores*; en particular, el nodo interior que no tiene nodo generador se llama *nodo inicial o nodo raíz* y se representa x_0. Para el árbol T, se representa $\tilde{T}$ el conjunto de sus nodos terminales y $\tilde{T}^0$ el de sus nodos interiores.

2.2. Proceso de poda

Dado un nodo x de un árbol T, T_x representa *la rama de T generada por x* o subárbol engendrado por el nodo x en el árbol T, es decir el árbol formado por la parte de T que contiene a x y a todos sus nodos sucesores hasta llegar a los correspondientes nodos terminales; $(T_x)^*$ representa dicha rama eliminado el nodo x.

Se llama poda del subárbol T_x de T al hecho de considerar en T el nodo x como terminal eliminando toda su rama engendrada. El árbol así obtenido se representa por $T' = T - (T_x)^*$ y se llama *subárbol podado de T.*

2.3. Coeficientes de contingencia y de Tschuprow

Partiendo de la matriz $(\tau, Y)_{\text{card}\,\tau \cdot c}$ que cruza las modalidades de la variable criterio con los nodos terminales, los coeficientes de contingencia y de Tschuprow obtenidos a partir del estadístico χ^2 calculado sobre dicha matriz, supuesto que $\text{card}\,\tau > 1$ y $c > 1$, ya que en otras condiciones existiría un único nodo terminal y una sola modalidad para la variable criterio, son válidos para medir la utilidad del árbol T como predictor de la variable criterio. Estos coeficientes se definen por:

$$\textit{Coeficiente de contingencia:} \quad \text{CP} = \sqrt{\frac{\frac{\chi^2}{n}}{1+\frac{\chi^2}{n}}}$$

$$\textit{Coeficiente de Tschuprow:} \quad T = \frac{\frac{\chi^2}{n}}{\sqrt{(m-1)(c-1)}}$$

siendo

χ^2: el valor del estadístico calculado sobre $(\tau, Y)_{\text{card}\,\tau \cdot c}$

m: el número de nodos terminales de T.

c: el número de modalidades de la variable criterio Y.

2.4. Simplicidad

Si en cada nodo terminal se asignan los elementos de I a la modalidad de la variable criterio Y que en dicho nodo presenta mayor proporción, cada camino que une el nodo raíz con un nodo terminal es una caracterización para la modalidad asignada a dicho nodo terminal. En consecuencia, la determinación de la modalidad de Y que le corresponde a un elemento cualquiera será más simple cuanto menor sea el número de caracterizaciones. Por este motivo se considera como una medida de la calidad del árbol, la simplicidad del mismo calculada a través del número de sus nodos terminales, definiéndose esta simplicidad $\mathcal{M}(T)$ en la siguiente forma: $\mathcal{M}(T) = \text{card}\,\tau$.

3. EFECTO DE UN PROCESO DE PODA EN EL COEFICIENTE DE CONTINGENCIA

Como se ha indicado, este coeficiente de asociación viene dado por:

$$CP = \sqrt{\frac{\frac{\chi^2}{n}}{1+\frac{\chi^2}{n}}}$$

Es inmediado que, si en el árbol T se realiza la poda de la rama engendrada por el nodo interior x, la variación producida en esta medida viene dada por:

$$\Delta CP = (CP)' - CP$$

donde $(CP)'$ es el valor del coeficiente de contingencia para el árbol podado $T' = T - (T_x)^*$ y CP es el correspondiente a T.

Analizando la variación del cuadrado de este coeficiente se tiene:

$$\Delta(CP)^2 = (CP')^2 - (CP)^2 = \frac{\frac{(\chi^2)'}{n}}{1+\frac{(\chi^2)'}{n}} - \frac{\frac{\chi^2}{n}}{1+\frac{\chi^2}{n}} = \frac{\frac{\chi^2+\Delta\chi^2}{n}}{1+\frac{\chi^2+\Delta\chi^2}{n}} - \frac{\frac{\chi^2}{n}}{1+\frac{\chi^2}{n}} =$$

$$= \frac{\left(1+\frac{\chi^2}{n}\right)\frac{\Delta\chi^2}{n} - \frac{\chi^2}{n}\frac{\Delta\chi^2}{n}}{\left(1+\frac{\chi^2}{n}\right)^2 + \frac{\Delta\chi^2}{n}\left(1+\frac{\chi^2}{n}\right)} = \frac{\frac{\Delta\chi^2}{n}}{\left(1+\frac{\chi^2}{n}\right)^2 + \frac{\Delta\chi^2}{n}\left(1+\frac{\chi^2}{n}\right)} =$$

$$= \frac{1}{\frac{\left(1+\frac{\chi^2}{n}\right)^2}{\frac{\Delta\chi^2}{n}} + \left(1+\frac{\chi^2}{n}\right)}$$

(1)
$$\Delta(CP)^2 = \frac{1}{\frac{\left(1+\frac{\chi^2}{n}\right)^2}{\frac{\Delta\chi^2}{n}} + \left(1+\frac{\chi^2}{n}\right)}$$

donde se ha supuesto que $\Delta\chi^2 \neq 0$ ya que en caso contrario, resulta que $\Delta(CP) = 0$.

Proposición 1

El coeficiente de contingencia no aumenta al realizarse la poda de un subárbol T_x cualquiera de T

Demostración

De acuerdo con (1) se tiene:

$$\Delta(\mathrm{CP})^2 = \frac{1}{\dfrac{\left(1+\dfrac{\chi^2}{n}\right)^2}{\dfrac{\Delta\chi^2}{n}} + \left(1+\dfrac{\chi^2}{n}\right)}$$

Pero: $$\Delta\chi^2 = \frac{n}{\sum\limits_{s\in\tilde{T}_x} n_{s\cdot}} \sum_{k=1}^{c} \left(\frac{1}{n_{\cdot k}} \sum_{\substack{s,s'\in\tilde{T}_x \\ s<s'}} \frac{-(n_{s'k}n_{s\cdot} - n_{s'\cdot}n_{sk})^2}{n_{s'\cdot}n_{s\cdot}} \right)$$

al realizarse la poda de un subárbol cualquiera T_x de T (Pérez Prados y otros, 1994) y en consecuencia $\Delta\chi^2 \leq 0$, y en nuestro caso $\Delta\chi^2 < 0$, ya que es por hipótesis $\Delta\chi^2 \neq 0$.

Pero esta reducción que se produce en χ^2 deberá ser en todo caso inferior al valor inicial del estadístico, en consecuencia, $\Delta\chi^2 \geq -\chi^2$, luego:

$$\frac{1+\dfrac{\chi^2}{n}}{\dfrac{\Delta\chi^2}{n}} + 1 \leq \frac{1+\dfrac{\chi^2}{n}}{-\dfrac{\chi^2}{n}} + 1 \leq 0$$

Por lo tanto: $\Delta(\mathrm{CP})^2 \leq 0$, y puesto que:

$$\Delta(\mathrm{CP}) = \frac{\Delta(\mathrm{CP})^2}{(\mathrm{CP}) + (\mathrm{CP})'}$$

y el coeficiente de contingencia es siempre positivo se tiene que $\Delta(\mathrm{CP}) \leq 0$.

■

Proposición 2

El valor del coeficiente de contingencia para el árbol $T_{máx}$ viene dado por:

(2) $$(\mathrm{CP})_{T_{\mathrm{máx}}} = \sqrt{1 - \frac{1}{c}}$$

Demostración

Este resultado se obtiene directamente de la definición del coeficiente de contingencia, teniendo en cuenta que $(\chi^2)_{T_{máx}} = (c-1)n$.

$$(\mathrm{CP})_{T_{máx}} = \sqrt{\frac{\frac{(c-1)n}{n}}{1+\frac{(c-1)n}{n}}} = \sqrt{\frac{c-1}{1+(c-1)}} = \sqrt{1-\frac{1}{c}}$$

■

4. CRITERIO DE UTILIDAD RELATIVA SEGÚN COEFICIENTES DE ASOCIACIÓN

4.1. Definiciones y propiedades fundamentales

Dado un árbol cualquiera T se considera la cantidad $S_\alpha(T)$ definida por la siguiente expresión:

$$S_\alpha(T) = \mathrm{CP}(T) - \alpha \mathcal{M}(T) \tag{3}$$

donde $\mathrm{CP}(T)$ es el coeficiente de contingencia; α es un número real positivo o nulo y $\mathcal{M}(T)$ la simplicidad.

Obsérvese que $S_\alpha(T)$ combina una medida de la utilidad de T con su simplicidad.

Por otra parte, se conoce también que el coeficiente de Tschuprow depende tanto de χ^2 como de la simplicidad de T. Por lo tanto podrán combinarse ambas medidas para obtener un árbol óptimo.

Definición 1

Se dice que T' es un subárbol óptimamente podado de T si el valor de la cantidad criterio a él asociado es el mayor entre los correspondientes a todos los subárboles de T.

Definición 2

Dados T_1 y T_2 dos árboles cualesquiera obtenidos a partir de I se dice que T_1 es mejor que T_2 según el coeficiente de Tschuprow, si se verifica que: $T(T_1) > T(T_2)$

Definición 3

Dos árboles cualesquiera T_1 y T_2 obtenidos a partir de I se dice que son equivalentes según el coeficiente de Tschuprow si se verifica que: $T(T_1) = T(T_2)$

Definición 4

Dado un conjunto A de subárboles podados de un árbol T, se dice que $\{T_e\}$ es el conjunto de mejores árboles de A según el coeficiente de Tschuprow, si para todos su elementos se verifica que:

$$T(T_e) \geq T(T_i) \qquad \forall\, T_i/T_i \in A$$

Propiedades

❶ Para el caso particular del árbol trivial T_1 que contiene únicamente el nodo raíz x_0, se tiene:

$$S_\alpha(T_1) = \mathrm{CP}(T_1) - \alpha \mathcal{M}(T_1) = -\alpha$$

❷ Si el árbol T corresponde al árbol $T_{\text{máx}}$ en el sentido de que todos los elementos de cada nodo terminal pertenecen a una misma modalidad de Y, entonces:

$$S_\alpha(T_{\text{máx}}) = \mathrm{CP}(T_{\text{máx}}) - \alpha \mathcal{M}(T_{\text{máx}}) = \sqrt{1 - \frac{1}{c}} - \alpha \,\mathrm{card}\, \mathfrak{T}_{\text{máx}}$$

Según el resultado obtenido en la proposición 1, el máximo valor del coeficiente de contingencia se alcanza en el árbol $T_{\text{máx}}$, por ser máximo χ^2 en dicho árbol; en consecuencia, cualquier árbol obtenido a partir de $T_{\text{máx}}$ mediante un proceso de división de uno o varios de sus nodos terminales reducirá el valor de la utilidad $S_\alpha(T)$, salvo en el caso $\alpha = 0$ en el cual no se producirá ninguna modificación. En este último caso:

$$S_0(T) = \mathrm{CP}(T) \leq \sqrt{1 - \frac{1}{c}}$$

Para cada α será un árbol óptimamente podado de $T_{\text{máx}}$ según este criterio, cualquiera que maximice el valor de $S_\alpha(T)$.

Dado un árbol cualquiera T, si mediante un proceso de poda del nodo x, se obtiene el árbol T', la variación producida en la utilidad $S_\alpha(T)$ será:

$$\begin{aligned} \Delta_x S_\alpha(T) &= S_\alpha(T') - S_\alpha(T) = \mathrm{CP}(T') - \mathrm{CP}(T) + \alpha(\mathrm{card}\, \mathfrak{T} - \mathrm{card}\, \mathfrak{T}') = \\ &= \Delta_x \,\mathrm{CP}(T) + \alpha(\mathrm{card}\, \mathfrak{T}_x - 1) \end{aligned} \tag{4}$$

donde $\Delta_x \text{CP}(T) \leq 0$ cualquiera que sea el nodo x, de acuerdo con los resultados de la proposición 1.

En consecuencia, el incremento producido en la utilidad $S_\alpha(T)$ al introducirse la poda de un nodo x, consta de dos términos, el primero de ellos es negativo o nulo y el segundo positivo, salvo en el caso $\alpha = 0$ en el cual se anula.

4.2. Sucesión de árboles según los valores de α

Se trata de obtener ahora una sucesión de árboles que optimicen, para los distintos valores de α, la utilidad $S_\alpha(T)$. Las demostraciones de los lemas se hallan en el apéndice.

Lema 1

«El único árbol óptimamente podado de $T_{\text{máx}}$ según $S_0(T)$ es el propio $T_{\text{máx}}$».

Por lo tanto, de acuerdo con este lema, el primer elemento de la secuencia que queremos obtener es: $T_0 = T_{\text{máx}}$.

Proceso de obtención de la secuencia de subárboles óptimamente podados del árbol máximo

- ***Paso inicial*** $\alpha = 0$

  En este caso según (4)

$$\Delta_x S_0(T) = \Delta_x \text{CP}(T) \leq 0$$

cualquiera que sea el nodo x elegido. Por tanto serán árboles óptimamente podados de $T_{\text{máx}}$ para $\alpha = 0$ todos aquellos que correspondan al mismo valor del coeficiente de contingencia que dicho árbol $T_{\text{máx}}$.

- ***Pasos sucesivos*** $\alpha > 0$ y creciente

  A partir de T_0 cualquier poda que se introduzca dará lugar a una reducción del valor del estadístico χ^2 y, como consecuencia, a una reducción del valor del coeficiente de contingencia; pero como α es creciente, el término positivo de $\Delta_x S_\alpha(T)$ crece, hasta que al llegar a un determinado valor de α que se indica α_1, el valor de $\Delta_x S_{\alpha_1}(T)$ es nulo, lo que indica que para ese valor α_1 existe al menos una poda posible tal que el nuevo árbol posee una utilidad S_{α_1} igual que la que corresponde a T_0; por lo cual, para α_1, cualquiera de los subárboles obtenidos mediante estas podas serán subárboles óptimamente podados. Entre ellos se elige T_1 como el mejor de acuerdo con el criterio de Tschuprow.

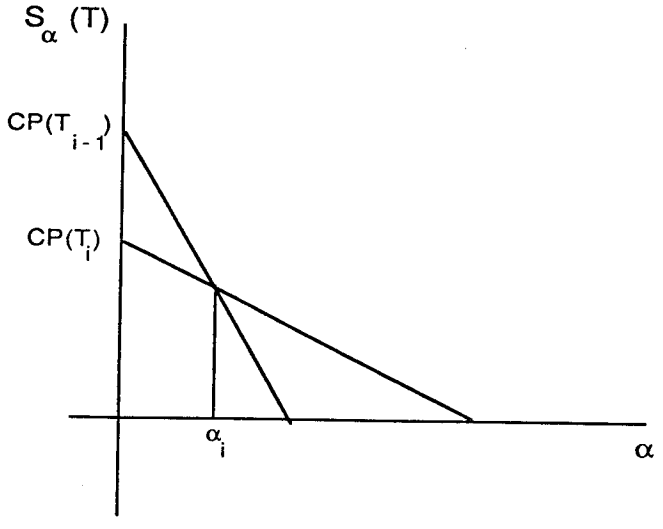
Se continúa con el crecimiento de α hasta llegar a un nuevo valor α_2 para el cual el valor de la utilidad $S_{\alpha_2}(T_1)$ coincide con el valor de la misma correspondiente a otro u otros árboles podados de $T_{\text{máx}}$. Entre éstos se elige el mejor según el criterio de Tschuprow y éste será el nuevo árbol de la sucesión.

Repitiendo el proceso se obtendría la sucesión completa de árboles $T_0 = T_{\text{máx}}$, $T_1, T_2, \ldots, T_r = \{x_0\}$ todos ellos podados de $T_{\text{máx}}$, junto con los valores de α asociados.

Lema 2

«Los árboles de la sucesión $T_0 = T_{\text{máx}}$, $T_1, T_2, \ldots, T_r$ anterior verifican: $\text{card}\boldsymbol{\tau}_i < \text{card}\boldsymbol{\tau}_{i-1} \quad \forall\, i = 1, \ldots, r$».

Puede verse la interpretación gráfica de este lema en la figura siguiente, teniendo en cuenta que la pendiente de la recta $S_\alpha(T) = \text{CP}(T) - \alpha\,\text{card}\boldsymbol{\tau}$ es $m = -\text{card}\boldsymbol{\tau}$.



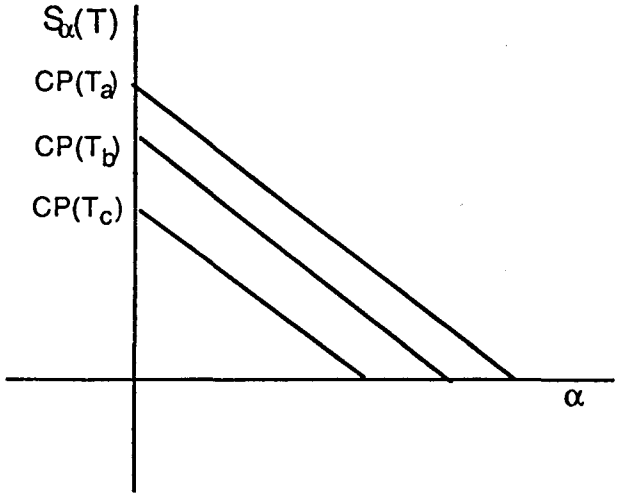
Por lo tanto, de acuerdo con este criterio de utilidad relativa según coeficientes de asociación, en el que se trata de obtener la sucesión de subárboles óptimamente podados de $T_{\text{máx}}$ para valores crecientes de α, será preciso en primer lugar, calcular el conjunto de todos los subconjuntos podados de $T_{\text{máx}}$ y sus correspondientes valores del coeficiente de contingencia CP(T) así como el número de sus nodos terminales $\text{card}\boldsymbol{\tau}$, valor este último que verificará $1 \leq \text{card}\boldsymbol{\tau} \leq \text{card}\boldsymbol{\tau}_{\text{máx}}$; a partir de estos resultados se realiza una partición del conjunto de subárboles en función del valor de

card$\mathfrak{T}$. Esta partición constará por lo tanto de card$\mathfrak{T}_{máx}$ elementos de la forma:

$$C^i = \{T \leq \mathfrak{p}T_{máx} / \text{card}\mathfrak{T} = c^i\}$$

donde $T \leq \mathfrak{p}T_{máx}$ indica que T se obtiene mediante un proceso de podas realizadas a partir de $T_{máx}$.

Las representaciones $(\alpha, S_\alpha(T))$ para los árboles de C^i pueden observarse en la figura siguiente:



Lema 3

«En cada uno de los subconjuntos C^i de subárboles podados de $T_{máx}$ anteriores, el conjunto de árboles de dicho conjunto que pueden ser subárboles óptimamente podados, o bien está formado por un único elemento o bien los elementos que lo forman son de igual utilidad para cualquier valor de α y equivalentes según Tschuprow».

Se representa T^i el elemento de C^i que verifica las condiciones para poder ser subárbol óptimamente podado de $T_{máx}$, que evidentemente es el subárbol óptimo entre los del conjunto C^i.

De acuerdo con el lema anterior la selección de subárboles óptimamente podados de $T_{máx}$ deberá realizarse en el conjunto de subárboles T^i anteriores, conjunto que contará con un número de elementos igual a card$\mathfrak{T}_{máx}$ y que se representa $\mathcal{S}$. En consecuencia, se tienen card$\mathfrak{T}_{máx}$ ecuaciones de la forma:

$$S_\alpha(T^i) = \text{CP}(T^i) - \alpha\,\text{card}\mathfrak{T}^i \quad T^i \in \mathcal{S} \quad i = 1,2,\ldots,\text{card}\mathfrak{T}_{máx}$$

Para $\alpha = 0$ se tiene que:

$$S_0(T_{\text{máx}}) = \text{CP}(T_{\text{máx}}) = \sqrt{1 - \frac{1}{c}}$$

$$T_0 = T_{\text{máx}}$$

a partir de este árbol se calculan los valores de α para los cuales se verifica:

$$S_\alpha(T^i) = S_\alpha(T_{\text{máx}})$$

para cada uno de los árboles $T^i \in \mathcal{S}$. Esto significa:

$$\text{CP}(T^i) - \alpha^{i0}\,\text{card}\,\tau^i = \sqrt{1 - \frac{1}{c}} - \alpha^{i0}\,\text{card}\,\tau_{\text{máx}} \rightarrow$$

$$\rightarrow \quad \alpha^{i0} = \frac{\sqrt{1 - \frac{1}{c}} - \text{CP}(T^i)}{\text{card}\,\tau_{\text{máx}} - \text{card}\,\tau^i}$$

El mínimo de ellos será α_1 y T_1 el árbol correspondiente del conjunto $\mathcal{S}$, que será el segundo elemento de la sucesión de árboles óptimos.

$$\alpha_1 = \text{mínimo} \left\{ \alpha^{i0}/\alpha^{i0} = \frac{\sqrt{1 - \frac{1}{c}} - \text{CP}(T^i)}{\text{card}\,\tau_{\text{máx}} - \text{card}\,\tau^i} \right\}$$

$$T_1 \in \mathcal{S}/S_{\alpha_1}(T_1) = S_{\alpha_1}(T_0)$$

Se opera análogamente con T_1, para la obtención de α_2 y T_2, teniendo en cuenta ahora que según el lema 2 se verifica que $\text{card}\,\tau_2 < \text{card}\,\tau_1$.

$$S_\alpha(T^i) = S_\alpha(T_1)$$

para cada uno de los árboles $T^i \in \mathcal{S}/\text{card}\,\tau^i < \text{card}\,\tau_1$ y por lo tanto:

$$\text{CP}(T^i) - \alpha^{i1}\,\text{card}\,\tau^i = \text{CP}(T_1) - \alpha^{i1}\,\text{card}\,\tau_1 \rightarrow$$

$$\rightarrow \quad \alpha^{i1} = \frac{\text{CP}(T_1) - \text{CP}(T^i)}{\text{card}\,\tau_1 - \text{card}\,\tau^i}$$

El mínimo de ellos será α_2, siendo T_2 su árbol correspondiente.

$$\alpha_2 = \text{mín} \left\{ \alpha^{i1}/\alpha^{i1} = \frac{\text{CP}(T_1) - \text{CP}(T^i)}{\text{card}\,\tau_1 - \text{card}\,\tau^i} \right\}$$

$$T_2 \in S/S_{\alpha_2}(T_2) = S_{\alpha_2}(T_1)$$

Continuando con el proceso se llega a la obtención del árbol T_r formado únicamente por el nodo raíz. La forma general del proceso en su fase $j-$ésima será:

$$S_\alpha(T^i) = S_\alpha(T_{j-1}) \quad T^i \in S/\text{card}\,\tau^i < \text{card}\,\tau_{j-1} \rightarrow$$

$$\rightarrow \quad \text{CP}(T^i) - \alpha^{i(j-1)}\,\text{card}\,\tau^i = \text{CP}(T_{j-1}) - \alpha^{i(j-1)}\,\text{card}\,\tau_{j-1} \rightarrow$$

$$\rightarrow \quad \alpha^{i(j-1)} = \frac{\text{CP}(T_{j-1}) - \text{CP}(T^i)}{\text{card}\,\tau_{j-1} - \text{card}\,\tau^i}$$

$$\alpha_j = \text{mín}\left\{\alpha^{i(j-1)}/\alpha^{i(j-1)} = \frac{\text{CP}(T_{j-1}) - \text{CP}(T^i)}{\text{card}\,\tau_{j-1} - \text{card}\,\tau^i}\right\}$$

$$T_j \in S/S_{\alpha_j}(T_j) = S_{\alpha_j}(T_{j-1})$$

Es decir, que para ese valor de α ambos subárboles corresponden a la misma utilidad.

Si en alguno de los casos el árbol T_j que verifica la condición anterior para α_j no es único, se elige entre ellos el de menor número de nodos terminales, ya que para $\alpha_j < \alpha < \alpha_{j+1}$ este árbol será el óptimo.

Como consecuencia de todo ello, si se consideran todos los valores de $\alpha \geq 0$ se obtiene una secuencia de árboles óptimos, de forma que para cada intervalo (α_i, α_{i+1}) el subárbol podado de $T_{\text{máx}}$ que es óptimo de acuerdo con este criterio de utilidad es T_i. Según se ha realizado la construcción de la sucesión de árboles, para los valores de α_i se verifica:

$$S_{\alpha_i}(T_{i-1}) = S_{\alpha_i}(T_i)$$

con lo cual para los valores de α correspondientes a los extremos de los intervalos (α_i, α_{i+1}), se elige entre los dos subárboles correspondientes al mismo valor de la utilidad S_α, el que sea mejor según el criterio de Tschuprow. En estas condiciones se puede considerar una asociación entre los valores de α y los árboles óptimos T en la siguiente forma:

$\alpha = 0 \quad \rightarrow \quad T = T_0 = T_{\text{máx}}$

$\alpha \in (\alpha_i, \alpha_{i+1}) \quad \rightarrow \quad T = T_i$ término $i-$ésimo de la sucesión $T_0, T_1, \ldots, T_r$

$\alpha = \alpha_i \quad \rightarrow \quad T = T_{i-1}$ si $T(T_{i-1}) > T(T_i)$

$T = T_i$ si $T(T_{i-1}) < T(T_i)$

$\alpha > \alpha_r \quad \rightarrow \quad T = T_r = \{x_0\}$

4.3. Selección de un árbol entre los de la sucesión

A partir de los resultados anteriores, si el valor del parámetro α es conocido, el árbol podado de $T_{\text{máx}}$ óptimo de acuerdo con los criterios S_α y T queda perfectamente determinado. Esto, sin embargo, no ocurre si el valor de α no es conocido. En este caso, es necesario seleccionar uno entre los árboles de la sucesión $T_0, T_1, T_2, \ldots, T_r$. Para ello es preciso, en primer lugar, conocer la existencia o no de posibles limitaciones bien en cuanto a la importancia relativa de las medidas de calidad utilizadas, o bien en cuanto a la forma del árbol; la presencia o no de condiciones puede dar lugar a los siguientes casos:

❶ Se presentan limitaciones en cuanto a los valores de α, en cuyo caso la selección se realizará únicamente entre los árboles de la sucesión $T_0, T_1, T_2, \ldots, T_r$ que correspondan a esos valores de α, pudiendo incluso darse el caso particular en que los valores posibles de α lleven asociado un único árbol, con lo cual el árbol óptimo queda directamente determinado. En caso de existir varios, se selecciona uno, como en el caso general, pero entre los de la subsecuencia obtenida.

❷ Se presentan limitaciones en cuanto al número de nodos terminales del árbol, que no deben exceder un valor determinado. En este caso, se produce una reducción en el número de árboles posibles de la sucesión de subárboles óptimos de $T_{\text{máx}}$. La situación es por lo tanto, desde el punto de vista de la selección, análoga a la del apartado anterior.

❸ No existen limitaciones y, en consecuencia, se presenta el caso general, en el que hay que seleccionar un árbol entre los de la sucesión $T_0, T_1, T_2, \ldots, T_r$. Para ello, puede considerarse el valor del coeficiente de Tschuprow. Se elige entre los árboles óptimos según el criterio de utilidad que combina el coeficiente de contingencia con la simplicidad, el mejor según el coeficiente de Tschuprow.

5. CONCLUSIONES

Teniendo en cuenta los resultados anteriores puede concluirse la importancia de los coeficientes de asociación en la selección del árbol óptimo.

Partiendo del criterio que combina linealmente el coeficiente de contingencia con la simplicidad, según la cantidad criterio $S_\alpha(T) = \text{CP}(T) - \alpha\mathcal{M}(T)$, la sucesión de árboles óptimos para valores crecientes de $\alpha \geq 0$ se inicia con el árbol T_0, máximo en el sentido de que en cada nodo terminal todos los elementos pertenecen a la misma

modalidad de la variable criterio Y y entre ellos el que tenga un menor número de nodos terminales. A partir de él, mediante sucesivas podas pueden obtenerse subárboles podados de T_0 con número de nodos terminales decrecientes; agrupando éstos según el número de nodos terminales, en cada uno de los conjuntos formados, únicamente un árbol, el que corresponde al mayor valor del coeficiente de contingencia, puede ser subárbol óptimamente podado. Considerando estos árboles, a medida que los valores de α crecen, de acuerdo con la cantidad criterio, se determina para cada valor de α el que corresponde al mayor valor de la utilidad, teniendo en cuenta que el número de nodos terminales de cada árbol de la sucesión $T_0, T_1, T_2, \ldots, T_r$ es necesariamente menor que el que le corresponde al árbol que le precede en la sucesión.

Obtenida la sucesión de árboles, teniendo en cuenta las posibles restricciones existentes, bien en cuanto al valor del parámetro o del número de nodos terminales del árbol, se selecciona el que proporciona un mayor valor para el coeficiente de Tschuprow.

Entre las numerosas aplicaciones para las que los resultados anteriores son válidos está en estudio una referida a índices de ocupación, en particular en el campo de la enseñanza. Considerando como variables tanto las calificaciones en sucesivos cursos escolares, como las que hacen referencia a otras actividades extraescolares, familiares, etc. medidas sobre alumnos de una etapa escolar en sucesivos años, y conocidos los resultados que esos alumnos obtienen y la «plaza» que en cada momento ocupan, pueden determinarse árboles de clasificación entre los que se seleccionará el óptimo según el método planteado, que permitirá preveer las necesidades para cursos posteriores y planificar sobre ellas. Podrán obtenerse también otras consecuencias que posibiliten una mejora de la calidad de enseñanza y orientación del alumnado.

APÉNDICE

Demostración lema 1

Sea T un árbol cualquiera podado de $T_{\text{máx}}$, será T un árbol óptimamente podado si verifica:

$$S_0(T) = S_0(T_{\text{máx}}) = \text{CP}(T_{\text{máx}}) = \sqrt{1 - \frac{1}{c}}$$

es decir, si $\Delta_x S_0(T_{\text{máx}}) = 0$.

Pero:

$$\Delta_x S_0(T_{\text{máx}}) = \Delta_x \text{CP}(T_{\text{máx}}) = \frac{1}{\text{CP}(T_{\text{máx}}) + \text{CP}(T)} \frac{\frac{\Delta\chi^2}{n}}{\left(1+\frac{\chi^2}{n}\right)^2 + \frac{\Delta\chi^2}{n}\left(1+\frac{\chi^2}{n}\right)}$$

Luego $\Delta_x S_0(T_{\text{máx}}) = 0$ exige $\Delta\chi^2 = 0$, pero por las propiedades de χ^2 se conoce que no existe ninguna poda posible en $T_{\text{máx}}$ que verifique esta condición, y en consecuencia no existe ningún árbol podado de $T_{\text{máx}}$ que mantenga el valor del coeficiente de contingencia, luego el único árbol óptimamente podado de $T_{\text{máx}}$ para el criterio $S_0(T)$ es el mismo $T_{\text{máx}}$.

■

Demostración lema 2

Si T_i y T_{i-1} son dos árboles cualesquiera de la sucesión de árboles óptimos anterior, se verifica que:

$$S_\alpha(T_i) < S_\alpha(T_{i-1}) \quad \text{si } 0 \leq \alpha < \alpha_i$$

$$S_\alpha(T_i) = S_\alpha(T_{i-1}) \quad \text{si } \alpha = \alpha_i$$

$$S_\alpha(T_i) > S_\alpha(T_{i-1}) \quad \text{si } \alpha > \alpha_i$$

Considerando el caso particular $\alpha = 0$ se obtiene:

$$\text{CP}(T_i) < \text{CP}(T_{i-1})$$

Pero si $\alpha = \alpha_i > 0$ ⟶

$$\longrightarrow \quad \text{CP}(T_i) - \alpha_i \,\text{card}\,\mathbf{T}_i = \text{CP}(T_{i-1}) - \alpha_i \,\text{card}\,\mathbf{T}_{i-1} \quad \longrightarrow$$

$$\longrightarrow \quad \alpha_i = \frac{\text{CP}(T_i) - \text{CP}(T_{i-1})}{\text{card}\,\mathbf{T}_i - \text{card}\,\mathbf{T}_{i-1}} > 0;$$

en consecuencia:

$$\text{card}\,\mathbf{T}_i - \text{card}\,\mathbf{T}_{i-1} < 0 \quad \longrightarrow \quad \text{card}\,\mathbf{T}_i < \text{card}\,\mathbf{T}_{i-1}$$

■

Demostración lema 3

Sean T_1 y T_2 dos subárboles de C^i, por tanto:

$$S_\alpha(T_1) = \text{CP}(T_1) - \alpha \,\text{card}\,\mathbf{T}_1 = \text{CP}(T_1) - \alpha c^i$$

$$S_\alpha(T_2) = \text{CP}(T_2) - \alpha \,\text{card}\,\mathbf{T}_2 = \text{CP}(T_2) - \alpha c^i$$

Luego si para algún α_1 es T_1 un subárbol óptimamente podado de $T_{máx}$:

$$S_{\alpha_1}(T_1) \geq S_{\alpha_1}(T_2) \rightarrow \text{CP}(T_1) \geq \text{CP}(T_2) \rightarrow$$

$$\rightarrow S_\alpha(T_1) \geq S_\alpha(T_2) \quad \text{cualquiera que sea } \alpha$$

es decir, pueden presentarse dos posibilidades:

❶ $S_\alpha(T_1) > S_\alpha(T_2)$ cualquiera que sea α, con lo cual T_2 no puede ser un subárbol óptimamente podado de $T_{máx}$, es decir, T_1 es el único subárbol óptimamente podado de $T_{máx}$ entre los de C^i.

❷ $S_\alpha(T_1) = S_\alpha(T_2)$ con lo cual para cualquier valor de α la utilidad de T_1 y T_2 es la misma. Además:

$$T(T_1) = \frac{\frac{\chi^2(T_1)}{n}}{\sqrt{(c^i-1)(c-1)}} = \frac{\frac{\chi^2(T_2)}{n}}{\sqrt{(c^i-1)(c-1)}} = T(T_2)$$

Luego T_1 y T_2 son equivalentes según el criterio de Tschuprow.

■

BIBLIOGRAFÍA

[1] **Breiman, L., Friedman, J.H., Olshen, R.A.** y **Stone, Ch.J.** (1984). *Classification and Regression Trees.* Wadsworth & Brooks. Monterey, California.

[2] **Ciampi, A., Chang, C.H., Hogg, S.** y **McKineey, S.** (1987). *Recursive partition: A versatile method for exploratory data analysis in biostatistics.* In Proceedings from Joshi Festschrift, G. Umphrey (ed), 23–50. Amsterdam: Nort-Holland.

[3] **Ciampi, A.** (1989). «Generalized Regression Trees». *Computational Statistics and Data Analysis*, **12, 1**, 732–764.

[4] **Cuesta, P.** (1989). *Inducción en bancos de datos cualitativos.* Tesis Doctoral. Facultad de Matemáticas. Universidad Complutense de Madrid.

[5] **Goodman, L.** y **Kruskal, W.** (1954). «Measures of Association for Cross Classifications». *JASA*, **49**, 732–764.

[6] **Hartigan, J.A.** (1975). *Clustering Algorithms.* Wiley Publication.

[7] **Matusita, K.** (1956). «Decision rule, based on the distance for the classification problem». *Annals Inst. Statist. Math.*, **8**, 67–77.

[8] **Munduate, A.** (1993). *Cuestiones notables en la construcción y comparación de árboles de decisión.* Tesis Doctoral. Departamento de Métodos Estadísticos. Universidad Pública de Navarra.

[9] **Pérez Prados, A., Munduate del Río, A.** y **Cano Sevilla, F.J.** (1994). «El estadístico χ^2 en la selección del árbol óptimo (I): Proceso de poda». *Cuadernos de Bioestadística y sus Aplicaciones Informáticas,* **12, 1**, 5–17.

[10] **Pérez Prados, A., Munduate del Rio, A.** y **Cano Sevilla, F.J.** (1994). «El estadístico χ^2 en la selección del árbol óptimo (II): Criterio de Selección». *Cuadernos de Bioestadística y sus Aplicaciones Informáticas*, **2, 1**, 18–38.

[11] **Quinlan, J.R.** (1986). «Induction of Decision Trees». *Machine Learning*, **1**, 81–106.

[12] **Quinlan, J.R.** (1988). «Decision trees and multi-valued attributes». *Machine Intelligence*, **11**, 305–319.

ENGLISH SUMMARY:

CRITERION FOR THE SELECTION OF AN OPTIMUM TREE CONSIDERING ASSOCIATION COEFFICIENTS OBTAINED FROM THE χ^2

F.J. Cano Sevilla, A. Munduate del Río and A, Pérez Prados

There are various recognised methods for the construction of decision trees, starting from a set of data for $\{V^j\}$ variables and the Y variable related to the previous series, and defined on the learning group Y as a whole. Furthermore, the validity of association coefficients obtained from the χ^2 statistic (contingency coefficient, Tschuprow's coefficient and Cramer's coefficient) is also recognised for measuring the usefulness of a tree as a predictor of the Y variable criterion.

Taking this into account, we propose a criterion for selecting an optimum tree, considering a complexity criterion which combines the contingency coefficient whit the simplicity of the tree measured by the number of its terminal nodes.

Therefore, starting from the variation that a pruning process produces in the value of the statistic χ^2 (Pérez Prados *et al.*, 1994), the modification that this induces in the contingency coefficient is studied. The result obtained indicates that under no circumstances does this coefficient increase as a result a pruning process, the maximum value corresponding to the $T_{\text{máx}}$ tree, where in each of the terminal nodes all the elements belong to the same Y modality.

This behaviour of the contingency coefficient being known, the following complexity criterion is considered:

$$S_\alpha(T) = \text{CP}(T) - \alpha\mathcal{M}(T)$$

$\text{CP}(T)$ being the contingency coefficient, α a real positive or null number and $\mathcal{M}(T)$ the simplicity obtained based on the number of terminal nodes. A method of obtaining an optimum tree for each of the different values of parameter α is developed in accordance with this complexity criterion $S_\alpha(T)$ together with Tschuprow's coefficient. By establishing an α value any subtree which maximises the $S_\alpha(T)$ value would be an optimally pruned one.

The variation produced in the complexity criterion through the process of pruning the x node of the T tree is:

$$\Delta_x S_\alpha(T) = \Delta_x \text{CP}(T) + \alpha(\text{card}\,\mathbf{T}_x - 1)$$

where $\Delta_x CP(T)$ is the variation produced in the contingency coefficient. It therefore consists of two terms, the first is negative or null and the second positive, except in the case of $\alpha 00$, in which it is cancelled.

In accordance with this result a series of optimally pruned subtrees $T_{máx}$ is obtained for increasing values of α. For $\alpha = 0$ the only optimally pruned subtree $T_{máx}$ is itself, so T_0 is the first element of the succession. On the basis of this element, any pruning done will lead to a reduction in the contingency coefficient value. Furthermore, by considering growing α values the positive term of $\Delta_x S_\alpha(T)$ grows. Therefore, on reaching a certain α level (shown as α_1), the value of $\Delta_x S_{\alpha_1}(T)$ is null. This indicates that there is at least one possible pruning for α_1, so the maximum subtree. Among the subtrees which verify this condition, which will all be optimally pruned subtrees $T_{máx}$ for α_1, the one which corresponds to the highest value of Tschuprow's coefficient is chosen.

The growth of α is continued until a new value (α_2) is reached, for which the value of the $S_\alpha(T)$ utility coincides with the value of the one corresponding to another or other $T_{máx}$ pruned trees. The best is chosen among theses according to Tschuprow's criterion and this will be the new tree of the succession.

Repeating the process, the following succession of trees is obtained: $T_0, T_1, T_2, \ldots, T_r$ (all pruned from $T_{máx}$), together with the associated α value. These trees have a decreasing number of terminal nodes, T_r being the one consisting exclusively of the root node of $T_{máx}$.

If, by taking $T_{máx}$, successive prunings determine all the possible pruned subtrees, and grouping them according to the number of terminal nodes they have, in each of the sets created only one tree (corresponding to the higher contingency coefficient) can be the optimally pruned subtree.

Considering these trees, and bearing in mind that in the $T_0, T_1, T_2, \ldots, T_r$ succession, $\text{card}\,\mathbf{T}_i < \text{card}\,\mathbf{T}_{i-1}\ \forall i = 1, \ldots, r$ is determined for growing α values the tree corresponding to a higher utility value for each α value is necessarily lower than the one corresponding to the preceding tree.

If the value of the α parameter is known, the optimum tree is the one from the succession to which this value corresponds, otherwise one is selected from among all the trees, taking the value of Tschuprow's coefficient into account for each one.

QÜESTIIÓ, Vol 20, 2 pp. 187-212, 1996

SMALL-AREA ESTIMATION USING ADJUSTMENT BY COVARIATES

N.T. LONGFORD*

Linear regression models with random effects are applied to estimating the population means of indirectly measured variables in small areas. The proposed method, a hybrid with design- and model-based elements, takes account of the area-level variation and of the uncertainty about the fitted regression model and the area-level population means of the covariates. The method is illustrated on data from the U. S. Department of Labor Literacy Surveys and is informally validated on two states, Mississippi and Oregon, for which statewide surveys have been conducted.

Keywords: Effective sample size, linear regression, random effect, sampling variation.

1. INTRODUCTION

Large-scale educational surveys, such as the U. S. Department of Labor and National Adult Literacy Survey, the National Education Longitudinal Survey, and the National Assessment of Educational Progress in the United States, often provide abundant information about their target populations and certain subpopulations, but cannot be used directly for inference about small areas, such as states, counties, or census tracts. States or smaller administrative units often contract out smaller-scale surveys for their jurisdictions. Such surveys often do not utilize any information from the national surveys. As a result, some duplication in collection of information takes

*Nicholas T. Longford. Department of Medical Statistics, De Montfort University, The Gateway, Leicester LE1 9RH, England. Email: ntl@dmu.ac.uk

–Article rebut el maig de 1995.

–Acceptat el juny de 1996.

place. Considerable savings could be achieved and more information extracted if information about a small area contained in the national and small-area surveys could be combined, or indeed, if inference about a small area, based on the national sample, would use information ('borrow strength') from the other small areas.

This paper explores the extent to which information from national surveys can be used for inference about smaller units. We discuss in detail inference for states but the approach is equally applicable to other jurisdictions.

In a typical setup, a regression equation is fitted to the outcome variable y in the survey data, using an appropriately selected vector of covariates $\mathbf{x}$, yielding an estimate $\hat{\boldsymbol{\beta}}$ of the vector of regression parameters $\boldsymbol{\beta}$. Let the mean vector of the covariates $\mathbf{x}$ for small area j be $\mathbf{x}^{(j)}$, and let $\hat{\mathbf{x}}^{(j)}$ be its estimate, obtained not necessarily from the same survey. When no values of the outcome variable in area j are available,

$$\hat{y}^{(j)} = \hat{\mathbf{x}}^{(j)}\hat{\boldsymbol{\beta}} \tag{1}$$

is the obvious estimator (predictor) of the mean of the outcome variable y in area j. When area j is represented in the survey, an estimator of the area mean, $\bar{y}^{(j)}$, based solely on the data from the area, can be combined with the synthetic estimator $\hat{y}^{(j)}$,

$$\hat{y}_c^{(j)} = a_j\bar{y}^{(j)} + (1 - a_j)\hat{y}^{(j)},$$

with the area-specific coefficient a_j chosen so as to minimize the mean squared error of the combined estimator $\hat{y}_c^{(j)}$.

This paper gives details of the prediction procedure outlined above and describes an application to the Job Training Partnership Act (JTPA) and the U. S. Employment Service and Unemployment Insurance (ES/UI) surveys administered by the U. S. Department of Labor in 1989–90. For details of these surveys, see Kirsch and Jungeblut (1992). The prediction procedure is an application of the approach of Battese, Harter, and Fuller (1988), and is here extended to account for sampling weights and uncertainty about the population mean of the covariates.

Section 2 summarizes the two-level (random-effects) regression model on which the predictions are based. Section 3 describes the adaptation of this model and its model fitting algorithms for sampling weights. Sections 4 and 5 give the minimal details of the datasets and the variables used. The methods are illustrated in Section 6 on examples that compare the prediction for Mississippi and Oregon based on the national surveys, with the estimates of the population means from the statewide sample surveys. Section 7 summarizes the paper, outlines a way of assessing the information about a small area in the national sample, and discusses how estimators for a small area can be combined.

2. LINEAR REGRESSION AND ITS USE IN PREDICTION

For an outcome variable y consider the random-effects regression model

$$y_{ij} = \mathbf{x}_{ij}\beta + \delta_j + \varepsilon_{ij}, \tag{2}$$

where the subscripts i and j denote the elementary unit (subject) $i = 1, \ldots, n_j$ within area (e.g., state) $j = 1, \ldots N_2$; the random terms δ_j and ε_{ij} are mutually independent random variables with centered normal distributions and respective variances σ_2^2 and σ_1^2. Zero expectation of δ_j is not a restrictive assumption because a non-zero mean would be confounded in the regression $\mathbf{x}\beta$. Let p be the number of regression parameters (i.e., the length of the vector of covariates $\mathbf{x}_{ij}$). It is assumed throughout that the first component of $\mathbf{x}$ is equal to unity for each subject. The choice of the variables in $\mathbf{x}$ is a well-appreciated problem involving balancing the requirements of model parsimony and adequacy.

The random term δ_j can be interpreted as the deviation of area j from the national mean, after an adjustment for the covariates. The area-level variance σ_2^2 is a summary measure of the (adjusted) differences among the areas and it plays an important role in prediction for a state. To illustrate this, suppose the regression parameters β are known exactly. The synthetic predictor for area j with the known population mean vector of $\mathbf{x}$ equal to $\mathbf{x}^{(j)}$, is $\mathbf{x}^{(j)}\beta$. This predictor is not exact, though, because the 'true' value of the mean $\bar{y}^{(j)}$ is

$$\mathbf{x}^{(j)}\beta + \delta_j + \frac{1}{m_j}\sum_{i}^{m_j}\varepsilon_{ij}, \tag{3}$$

where m_j is the population size of area j ($m_j > n_j$, unless a full census is taken in area j). When the area is not represented in the data ($n_j = 0$), no information about δ_j is available. Then the mean squared error of the predictor in (3) is $\delta_j^2 + \sigma_1^2/m_j$, and so its lower bound (approximate value for large m_j) is δ_j^2. The expectation of this lower bound over the small areas, σ_2^2, represents a component of uncertainty about the prediction for each area. If prediction based on a random-effects model is to be used for areas not represented in the data, the variables $\mathbf{x}$ should be selected so that the adjusted area-level variance σ_2^2 is as small as possible. When an area is represented in the survey, the data for area j, but also the data for the other areas, contain information about δ_j.

2.1. Sampling variation of the prediction

Suppose the sampling variance matrix of $\hat{\beta}$ is Σ_b, and it is estimated by $\hat{\Sigma}_b$. Details of estimating β and Σ_b are given in Section 3.

Let $\hat{\mathbf{x}}^{(j)}$ be an estimator of the mean vector of the covariates $\mathbf{x}$ for area j, with its sampling variance matrix Σ_S estimated by $\hat{\Sigma}_S$. In the illustration in Section , $\hat{\mathbf{x}}^{(j)}$ is the ratio estimator and its sampling variance matrix is derived assuming a weighted random sampling design. The predictor $\hat{y}^{(j)}$ given in (1) involves products of random variables (parameter estimates and sample means). If $\hat{\mathbf{x}}^{(j)}$ were subject to no error, that is, $\Sigma_S = \mathbf{0}$, the sampling variance of the predictor would be

$$\text{var}\left(\hat{y}^{(j)}\right) = \mathbf{x}^{(j)}\Sigma_b\left(\mathbf{x}^{(j)}\right)^\top. \tag{4}$$

When the mean $\hat{\mathbf{x}}^{(j)}$ is based on data not used in estimating β, $\hat{\mathbf{x}}^{(j)}$ and $\hat{\beta}$ are independent. Then

$$\text{var}\left(\hat{y}^{(j)}\right) = \text{tr}(\Sigma_S\Sigma_b) + \mathbf{x}^{(j)}\Sigma_b\left(\mathbf{x}^{(j)}\right)^\top + \beta^\top\Sigma_S\beta, \tag{5}$$

which may be much greater than the variance in (4); poorer information about the covariates causes poorer prediction.

Next, suppose $\hat{\beta}$ and $\hat{\mathbf{x}}^{(j)}$ are both normally distributed, and let Σ_{bS} be their covariance matrix. Then

$$\begin{aligned}\text{var}\left(\hat{y}^{(j)}\right) &= \text{tr}(\Sigma_S\Sigma_b) + \mathbf{x}^{(j)}\Sigma_b\left(\mathbf{x}^{(j)}\right)^\top + \beta^\top\Sigma_S\beta \\ &\quad + \text{tr}\left(\Sigma_{bS}^2\right) + 2\mathbf{x}^{(j)}\Sigma_{bS}\beta.\end{aligned} \tag{6}$$

See Appendix for proof. Equation (5) is obtained from (6) by setting $\Sigma_{bS} = \mathbf{0}$. Estimating the covariance matrix Σ_{bS} is a problem, especially when the data for the small area is used for estimating β and the estimator $\hat{\beta}$ has a complex form. Typically, no area constitutes a large proportion of the data, and so the correlations of $\hat{\beta}$ and $\hat{\mathbf{x}}^{(j)}$ are small. Then the terms involving Σ_{bS} in (6) can be ignored and (5) applies.

When $\hat{\mathbf{x}}^{(j)}$ and $\hat{\beta}$ are unbiased the expectation of the estimator $\hat{y}^{(j)}$ is $\mathbf{x}^{(j)}\beta + \text{tr}(\Sigma_{bS})$, and so the conditional mean squared error of $\hat{y}^{(j)}$, given δ_j, is

$$\mathbf{E}\left\{\left(\hat{y}^{(j)} - y^{(j)}\right)^2 \mid \delta_j\right\} = \text{var}\left(\hat{y}^{(j)}\right) + \left\{\delta_j - \text{tr}(\Sigma_{bS})\right\}^2. \tag{7}$$

When Σ_{bS} is not known the size of the bias and mean squared error can be inferred by substituting a range of plausible values for the matrix Σ_{bS} and for the deviation δ_j. When the small area is not represented in the survey sample $\Sigma_{bS} = \mathbf{0}$, and there is no information about δ_j. Then the typical mean squared error is obtained by averaging over the marginal distribution of δ_j:

$$\mathbf{E}\left[\mathbf{E}\left\{\left(\hat{y}^{(j)} - y^{(j)}\right)^2 \mid \delta_j\right\}\right] = \text{var}\left(\hat{y}^{(j)}\right) + \sigma_2^2. \tag{8}$$

For a cluster represented in the survey, the deviation δ_j can be estimated as its estimated conditional expectation given the data, see Section 3.1.

3. SAMPLING WEIGHTS

The sampling weights are an important feature of the sampling design of a survey. Inferences based on data from such a survey have to take account of the weights. For instance, the commonly used estimator of the mean of a simple random sample $y_i, i = 1, \ldots N$, is $\bar{y} = N^{-1} \sum_i y_i$. Its counterpart for independent data with sampling weights w_i is the 'weighted' mean, or the ratio estimator,

$$\bar{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i}.$$

Similarly, the regression parameter vector β, estimated for independent observations with equal weights as

$$\hat{\beta} = \left(\sum_i \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_i y_i \mathbf{x}_i,$$

has the 'weighted' version

$$\hat{\beta}_w = \left(\sum_i w_i \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_i w_i y_i \mathbf{x}_i.$$

For estimation of the variances σ_2^2 and σ_1^2, as well as of the sampling variance matrix for $\hat{\beta}_w$, it is essential to use an appropriate normalization of the weights. Potthoff, Woodbury, and Manton (1992) show that the normalization in which the total of weights and the total of their squares are equal is appropriate. That is, the weights w_i are replaced by

$$w_i^* = w_i \frac{\sum_{i'} w_{i'}}{\sum_{i'} w_{i'}^2}.$$

It is assumed throughout that the weights have been normalized in this fashion, and the asterisk * on w is dropped. The total of the normalized weights, $N_w = \sum_i w_i$, can be interpreted as the *effective sample size*, that is, the size of a simple random sample that would contain the same amount of information about the population mean as the sample at hand. It can be shown that $N_w \leq N$, and equality holds only when the weights are constant.

When $\sigma_2^2 = 0$, the linear regression model in (2) simplifies to the ordinary regression. Then the estimators of the residual variance and the sampling variance of $\hat{\beta}$,

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{N_w - p} \sum_i (y_i - \mathbf{x}_i \hat{\beta})^2 \\ \text{var}(\hat{\beta}) &= \hat{\sigma}^2 \left(\sum_i w_i \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1}, \end{aligned}$$

are *approximately* unbiased.

Strictly speaking, the assumptions about the sampling weights listed in Potthoff *et al.* (1992) are not satisfied because the weights are random variables (owing to poststratification). However, we concur with Potthoff *et al.* that the consequences of randomness are not severe. Although the poststratified weights are derived by a complex process of adjustment for a number of background variables aggregated at various levels, the absolute changes of the weights are insubstantial relative to variation of the design weights.

3.1. Fitting random-effects models

The random-effects model in (2) can be fitted by the Fisher scoring algorithm. We list the relevant equations for the case of equal weights, and then describe the adaptation for unequal sampling weights.

Let $\mathbf{y}$ be the $N \times 1$ vector of outcomes, $\mathbf{X}$ the $N \times p$ matrix of regressors, $\mathbf{e} = \mathbf{y} - \mathbf{X}\beta$ the vector of residuals, and $\mathbf{V} = \text{var}(\mathbf{y})$ the variance matrix of the outcomes.

The first and second-order partial derivatives of the log-likelihood

$$l = -\frac{1}{2} \left\{ N \log(2\pi) + \log(\det \mathbf{V}) + \mathbf{e}^\top \mathbf{V}^{-1} \mathbf{e} \right\} \tag{9}$$

with respect to the regression parameters are

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{e} \\ \frac{\partial^2 l}{\partial \beta \partial \beta^\top} &= -\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}. \end{aligned}$$

An iteration of the Fisher scoring algorithm updates a current estimate $\hat{\beta}_{old}$ to obtain the 'new' estimate

$$\hat{\beta}_{new} = \hat{\beta}_{old} + \left(\mathbf{X}^\top \mathbf{V}_{old}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{V}_{old}^{-1} \mathbf{e}_{old},$$

where $\mathbf{V}_{old}$ and $\mathbf{e}_{old}$ are equal to $\mathbf{V}$ and $\mathbf{e}$ evaluated for the current values of the parameter estimates. Substitution for $\mathbf{e}_{old}$ and elementary algebra yield

$$\hat{\beta}_{new} = \left(\mathbf{X}^\top \mathbf{V}_{old}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{V}_{old}^{-1} \mathbf{y}; \tag{10}$$

the Fisher scoring and Newton-Raphson algorithms coincide with the generalized least squares.

In evaluating (10), inversion of large matrices is avoided by exploiting the pattern of the variance matrix $\mathbf{V}$. First, $\mathbf{V}$ is block-diagonal, with blocks

$$\mathbf{V}_j = \sigma_1^2 \mathbf{I}_{n_j} + \sigma_2^2 \mathbf{J}_{n_j}$$

correspoding to areas ($\mathbf{I}_m$ and is the $m \times m$ unit matrix and $\mathbf{J}_m = \mathbf{1}_m \mathbf{1}_m^\top$ the $m \times m$ matrix of ones). Next,

$$\begin{aligned} \det(\mathbf{V}_j) &= \sigma_1^{2n_j}(1 + n_j\tau) \\ \mathbf{V}_j^{-1} &= \sigma_1^{-2}\left(\mathbf{I}_{n_j} - \frac{\tau}{1 + n_j\tau}\mathbf{J}_{n_j}\right), \end{aligned}$$

where $\tau = \sigma_2^2/\sigma_1^2$.

Some advantage is gained by using the variance ratio τ instead of σ_2^2. Letting $\mathbf{W} = \sigma_1^{-2}\mathbf{V}$, the variance σ_1^2 can be separated out in the log-likelihood l,

$$l = -\frac{1}{2}\left\{N\log(2\pi\sigma_1^2) + \log(\det \mathbf{W}) + \sigma_1^{-2}\mathbf{e}^\top \mathbf{W}^{-1}\mathbf{e}\right\}. \tag{11}$$

The first-order partial derivative with respect to σ_1^2 has the root

$$\hat{\sigma}_1^2 = \frac{\mathbf{e}_{old}^\top \mathbf{W}_{old}^{-1}\mathbf{e}_{old}}{N}. \tag{12}$$

Finally, noting that $\partial \mathbf{W}/\partial\tau = \mathrm{diag}_j\{\mathbf{J}_{n_j}\}$,

$$\begin{aligned} \frac{\partial l}{\partial \tau} &= -\frac{1}{2}\sum_j \mathbf{1}_{n_j}^\top \mathbf{W}_j^{-1}\mathbf{1}_{n_j} + \frac{1}{2\sigma_1^2}\sum_j \left(\mathbf{e}_j^\top \mathbf{W}_j^{-1}\mathbf{1}_{n_j}\right)^2 \\ -\mathbf{E}\left(\frac{\partial^2 l}{\partial \tau^2}\right) &= \frac{1}{2}\sum_j \left(\mathbf{1}_{n_j}^\top \mathbf{W}_j^{-1}\mathbf{1}_{n_j}\right)^2, \end{aligned}$$

where $\mathbf{e}_j$ is the subvector of $\mathbf{e}$ corresponding to area j and $\mathbf{W}_j = \sigma_1^{-2}\mathbf{V}_j$. At each iteration the current estimate of τ is updated as

$$\hat{\tau}_{new} = \hat{\tau}_{old} - \left\{ \mathbf{E}\left(\frac{\partial^2 l}{\partial \tau^2} \right) \right\}^{-1} \frac{\partial l}{\partial \tau}, \tag{13}$$

with the right-hand side evaluated at the current solution $(\hat{\beta}_{old}, \hat{\sigma}^2_{1,old}, \hat{\tau}_{old})$. The iterations, consisting of (10), (12), and (13), are terminated when a convergence criterion is satisfied. Such a criterion can be based on the change of the log-likelihood, $l_{new} - l_{old}$, on the size of the corrections for the estimates, the norm of the score vector, or a combination of these criteria.

The algorithm requires the following statistics: the totals of squares and cross-products, $(\mathbf{y},\mathbf{X})^\top(\mathbf{y},\mathbf{X})$, and within-area totals $(\mathbf{y}_j, \mathbf{X}_j)^\top \mathbf{1}_{n_j}$. The algorithm is adapted for sampling weights by replacing these summaries by their weighted versions, $(\mathbf{y}_j, \mathbf{X}_j)^\top \mathbf{w}_j$ and $(\mathbf{y},\mathbf{X})^\top \mathrm{diag}(\mathbf{w})(\mathbf{y},\mathbf{X})$, where $\mathbf{w}_j$ is the $n_j \times 1$ vector of sampling weights for area j, and $\mathbf{w}$ is the $N \times 1$ vector of weights for the entire sample. Note that the area-level sample size n_j is replaced by the total weight $\sum_i w_{ij}$.

An integral part of the algorithm is estimation of the realized values of δ_j as the conditional expectations and of their precision as the conditional variances of δ_j given the data and the parameter estimates:

$$\begin{aligned} \mathbf{E}(\delta_j \mid \mathbf{y}_j) &= \frac{\tau \mathbf{e}_j^\top \mathbf{1}_{n_j}}{1 + n_j \tau} \\ \mathrm{var}(\delta_j \mid \mathbf{y}_j) &= \frac{\sigma_2^2}{1 + n_j \tau}. \end{aligned}$$

The weighted versions of these equations are obtained by replacing $\mathbf{e}_j^\top \mathbf{1}_{n_j}$ with $\sum_i e_{ij} w_{ij}$ and n_j with $\sum_i w_{ij}$.

The Fisher scoring algorithm can be straightforwardly adjusted for restricted maximum likelihood estimation (REML), see Harville (1974). The log-likelihood in (11) is adjusted by the term

$$\Delta l_R = \frac{1}{2} \log \left\{ \det \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X} \right) \right\},$$

and the scoring vector and information matrix by the corresponding partial derivatives. For instance, in the equation for $\hat{\sigma}_1^2$, (12), the denominator is reduced from N to $N - p$, taking account of the uncertainty about the p regression parameter estimates. The approach based on the best linear unbiased prediction (BLUP, Kackar and Harville, 1984) can also be adapted for sampling weights. These approaches rely on normality of the random terms and on linearity. For alternative approaches, see Beran and Hall (1992) and references therein.

4. DATA

Data from two national surveys of adult literacy, ES/UI and JTPA, administered by the U. S. Department of Labor, and data from surveys from the states of Mississippi and Oregon are available. The surveys are designed to assess the nature and extent of the literacy skills of the U. S. adult population aged 16 and over.

The survey instrument, common to all four surveys, consists of a 15-minute background questionnaire and a 45-minute set of exercises. The design of each survey is that of a stratified multi-stage clustered sample. The target population in the national surveys was stratified to seven geographical regions and U. S. states were drawn from each region with replacement as the primary sampling units. There were some minor differences in the definitions of the clusters between the two national surveys. The sampling weights were derived by adjustment of the design weights due to poststratification. For the purposes of illustration we treat the poststratified weights as if they were the design weights, and we ignore all levels of clustering except state level.

The sample sizes and the effective sample sizes (N_w) are given in Table 1. The coefficient of variation of the weights, given in the third row of the table, is defined as the ratio of the sample variance and the square of the sample mean of the weights,

$$\rho = \frac{\text{var}(w_i)}{\bar{w}^2}.$$

Table 1

Raw and effective sample sizes in the adult literacy surveys

	Survey			
	ES/UI	JTPA	Mississippi	Oregon
Sample size	3277	2501	1804	1993
Effective sample size	1403	1046	1629	1854
Coefficient of variation	1.34	1.39	0.11	0.08

It is a useful indicator of how much smaller the effective sample size is compared to the raw sample size. Constant weights correspond to $\rho = 0$. In the national surveys (ES/UI and JTPA) the weights vary considerably; for instance, the ratio of the largest and smallest weights is 86.5 and 53.5 in ES/UI and JTPA, respectively. In the statewide surveys, these ratios are only 10.6 (Mississippi) and 3.9 (Oregon). The

histograms of the normalized sampling weights are drawn in Figure 1. Each national survey dataset contains less information, in the sense of the effective sample size, than either dataset from the state surveys, even though the former have greater (raw) sample sizes.

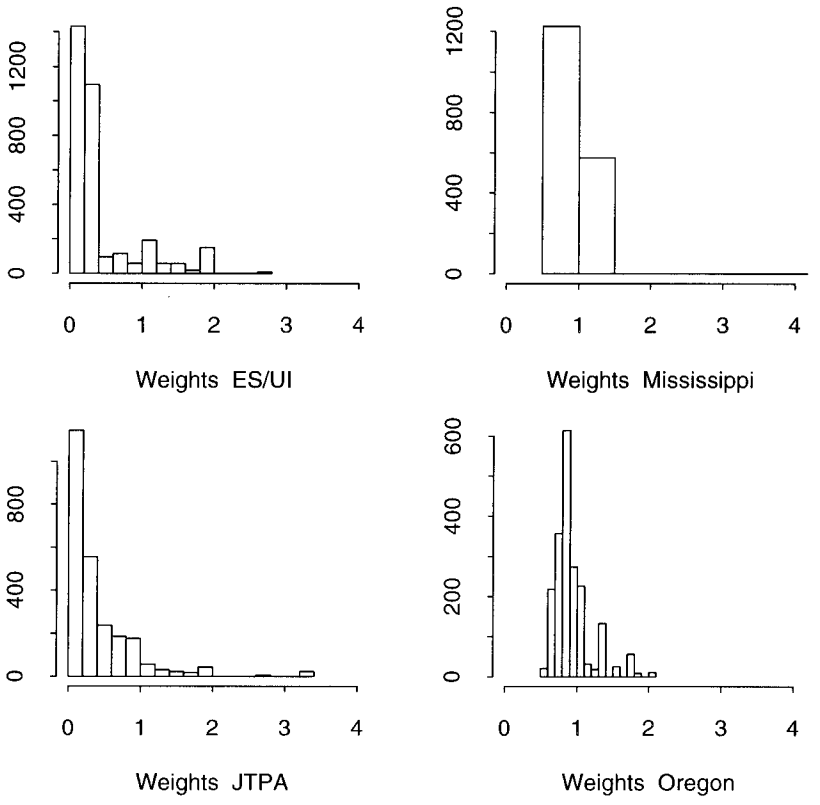


Figure 1. ***Histograms of the normalized sampling weights for the four surveys. The same scale for the horizontal axis is used in all four panels.***

The principal outcome variable, called the (literacy) *proficiency score*, is defined on a scale common to all four surveys. The probability of correct response of subject i with proficiency score θ_i to question k with item parameters (a_k, b_k, c_k) is modelled as

$$P(Z_{ik} = 1 \mid \theta_i; a_k, b_k, c_k) = c_k + (1 - c_k)\frac{\exp(a_k + b_k\theta_i)}{1 + \exp(a_k + b_k\theta_i)};$$

the parameters a_k, b_k, and c_k can be interpreted as the difficulty, the discrimination, and the probability of guessing, respectively. In the item response model applied, the

responses Z_{ik} are assumed to be conditionally independent given θ_i, and a normal prior distribution is imposed for $\{\theta_i\}$. Note that the proficiency scores $\{\theta_i\}$ are confounded with the item parameters $\{a_k\}$ and $\{b_k\}$, and so the mean and variance of the prior distribution for $\{\theta_i\}$ are set merely to ensure identifiability. The proficiencies as well as their estimates are in the range $(-\infty, +\infty)$; the value of zero is of no special significance.

The survey subjects' scores on this scale are estimated using a marginal maximum likelihood approach (Bock and Aitkin, 1981, and Mislevy and Bock, 1983). Inferences about the proficiency scores derived by regarding the estimates $\hat{\theta}_i$ as the true values θ_i are likely to underestimate the precision because they ignore the substantial sampling variance of each estimate $\hat{\theta}_i$. To take account of the uncertainty associated with the *estimated* proficiency scores, a set of five *imputed values* are randomly drawn from the approximation to the posterior distribution of the proficiency scores. The decision to use five imputed values was based on extensive simulations. Any analysis of the proficiency scores (e.g., regression) involves identical analyses using each set of the five imputed values. Let $\hat{\beta}_h$, $h = 1, \ldots, 5$, be a quintet of such estimates. Then the estimate that refers to the proficiency scores is

$$\hat{\beta} = \frac{1}{5}\sum_h \hat{\beta}_h .$$

The standard errors for the parameters that refer to the proficiency scores are obtained similarly, but they have to be inflated by the variance of the estimates across the five analyses. Suppose s_h^2 is the estimated sampling variance of $\hat{\beta}_h$. Then the sampling variance of $\hat{\beta}$ is estimated as

$$s^2 = \frac{1}{5}\sum_h s_h^2 + \frac{1}{4}\sum_h \left(\hat{\beta}_h - \hat{\beta}\right)^2 .$$

The main purpose of the study is to validate the outlined method; improvement in the estimation of population means for the particular two states is of lesser importance. For this purpose, we estimate the population means for the two states based on the national survey data and the covariates for the state-wide surveys, treating the within-state data on y as a 'hold-out' dataset. These estimates and their standard errors are then compared with their counterparts based solely on the imputed values for the within-state samples. In the concluding section we discuss how such pairs of estimators can be combined.

The histograms of the first sets of imputed values are displayed in Figure 2. The mean of the sample for Oregon is somewhat higher than that for Mississippi, and the sample variances in the state surveys are smaller than their national counterparts. There is a perceptible difference in the distributions for the national surveys (JTPA

has smaller variance than ES/UI), although it may be accounted for by the varying weights. These comparisons carry over from the imputed values to the proficiency scores.

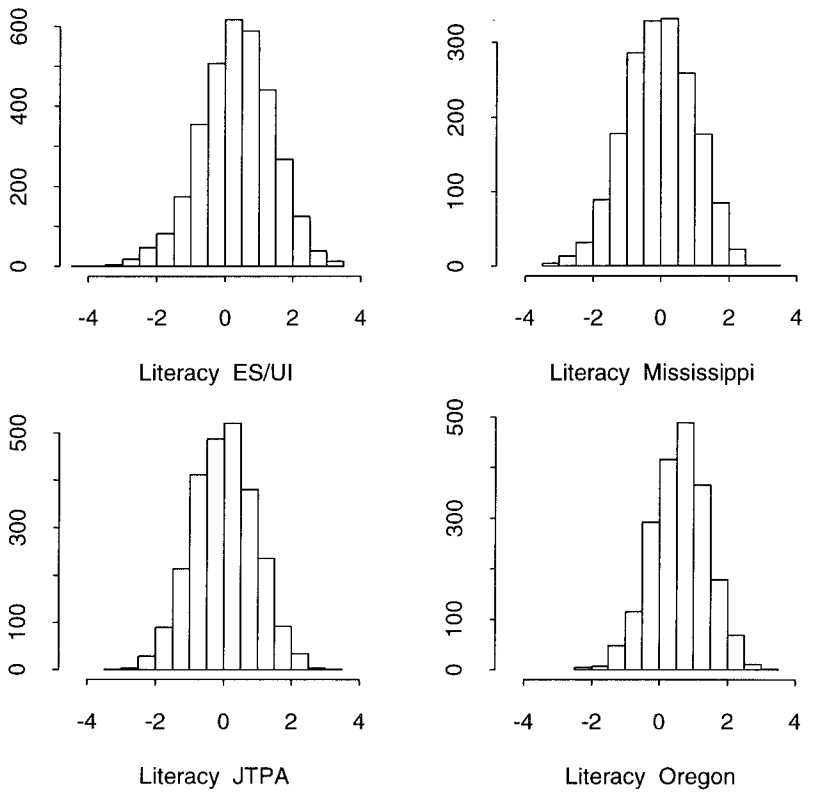


Figure 2. ***Histograms of the first sets of imputed values for the four surveys. The same scale for the horizontal axis is used in all four panels.***

5. ANALYSIS

For illustration, we describe the prediction for Mississippi and Oregon using the following set of predictor variables:

Sex (dichotomous);
Ethnicity (six nominal categories);

Educational level (five ordinal categories);
Age category (five ordinal categories);
More English (dichotomous);
More Mathematics (dichotomous);
Mother's education (quantitative);
Father's education (quantitative);
Personal income (quantitative);
Household income (quantitative);
C.Age (quantitative, in years);
Q.Age (quantitative, in years^2/1000).

The variable *Sex* is coded 0 for men and 1 for women. The other dichotomous variables have values 2 ('thinks more English/Mathematics would be useful') and 1 ('does not think so'). The variables related to income and parents' education are defined on the integer scale 1–11. See Kirsch and Jungeblut (1992) for details. The variable *C.Age* is the subject's age in years, and *Q.Age* is defined as $C.Age^2/1000$. For parsimony, they are used as quantitative variables, so that they are represented only by one parameter each.

The means and proportions (as applicable) for the covariates and for the four datasets are given in Table 2. The regression model fit for the ES/UI data using the first set of imputed values and the set of covariates listed above, is summarized in Table 3. For comparison, the ordinary least squares fit (OLS) and the maximum likelihood (MLE) fit are given. The likelihood ratio test statistic, equal to the difference of the OLS and MLE deviances (the values of
-2 log-likelihood), can be used to assess the significance of the state-level variance σ_2^2. The value of this statistic is $3380.50 - 3354.89 = 25.61$; its approximate (asymptotic) null-distribution is χ_1^2. Thus, the state-level variation is highly significant. Note however, that the t-statistic for σ_2^2 is nominally not significant at the 5 per cent level.

To avoid problems with missing data, listwise deletion was applied; all records which contain a missing observation for at least one of the covariates were deleted. This reduced the sample size for the ES/UI data from 3277 to 3219, and the effective sample size from 1402.5 to 1378.4. Listwise deletion for the same covariates in the JTPA dataset resulted in reduction of the effective sample size from 1047.6 to 1023.9.

A model-based method for full use of incomplete data in random effects models is described in Longford (1995). Its implementation requires specification of a model for the process that gives rise to missing data. It may reduce the bias due to the informative nature of this process, but improvement in efficiency of the prediction is likely to be insubstantial because of the low proportion of missing data. The means and proportions in Table 2 refer to the entire sample; the corresponding figures for the listwise deleted samples differ insubstantially.

Table 2

Weighted sample means and proportions of the covariates

	Survey							
	ES/UI		JTPA		Mississippi		Oregon	
Variable	Proprt'n or mean	Count	Proprt'n or mean	Count	Proprt'n or mean	Count	Proprt'n or mean	Count
Sex 1	0.56	1756	0.41	1008	0.47	761	0.50	1064
2	0.44	1515	0.58	1484	0.53	1043	0.50	929
Ethnicity 1	0.63	2394	0.69	1556	0.66	1300	0.92	1845
2	0.12	375	0.21	663	0.32	473	0.01	16
3	0.20	384	0.06	159	0.01	20	0.03	57
4	0.02	40	0.00	17	0.00	2	0.02	28
5	0.01	48	0.03	76	0.00	5	0.02	28
6	0.02	36	0.01	30	0.00	4	0.01	19
Education 1	0.03	135	0.07	202	0.13	217	0.02	34
2	0.18	619	0.33	871	0.19	349	0.15	282
3	0.59	2006	0.55	1295	0.48	872	0.56	1104
4	0.19	513	0.06	130	0.19	362	0.27	570
5	0.00	4	0.00	3	0.00	1	0.00	1
Age categ. 1	0.10	314	0.17	489	0.09	140	0.09	138
2	0.18	616	0.19	485	0.11	173	0.12	183
3	0.22	727	0.21	505	0.13	238	0.17	301
4	0.32	1059	0.31	733	0.27	477	0.35	778
5	0.17	546	0.10	259	0.40	776	0.28	593
More Engl. 1	0.57	1792	0.66	1717	0.52	889	0.39	745
2	0.42	1461	0.33	762	0.47	901	0.61	1246
More Maths 1	0.69	2231	0.79	2002	0.61	1054	0.52	1003
2	0.30	1020	0.20	473	0.38	732	0.48	988
Mother's Ed.	3.77		3.67		4.38		4.31	
Father's Ed.	4.37		4.32		5.07		4.86	
Pers. Income	3.29		2.29		3.99		3.86	
Hous. Income	5.04		3.41		4.95		5.55	
Age (years)	33.84		30.67		42.64		37.91	
Quadr. age	1287.91		1061.96		2113.72		1612.45	

Note: For categorical variables the proportions are accompanied by the counts of subjects in the category. The counts of all the categories of a variable do not add up to the sample size because of missing data.

Table 3

Regression model fits for ES/UI using OLS and MLE

	OLS		MLE	
Parameter	Estimate	St. error	Estimate	St. error
Intercept	–1.148	(0.470)	–1.171	(0.465)
Sex 2–1	–0.199	(0.050)	–0.204	(0.052)
Ethnicity 2–1	–0.648	(0.080)	–0.641	(0.064)
3–1	–0.522	(0.067)	–0.449	(0.072)
4–1	–0.702	(0.174)	–0.679	(0.173)
5–1	–0.575	(0.212)	–0.592	(0.210)
6–1	–0.124	(0.200)	–0.102	(0.199)
Education 2–1	0.806	(0.155)	0.813	(0.153)
3–1	1.184	(0.148)	1.187	(0.147)
4–1	1.685	(0.158)	1.686	(0.157)
5–1	–0.111	(0.689)	–0.130	(0.681)
Age category 2–1	–0.287	(0.117)	–0.287	(0.116)
3–1	–0.257	(0.172)	–0.271	(0.170)
4–1	–0.373	(0.250)	–0.386	(0.247)
5–1	–0.198	(0.322)	–0.021	(0.318)
More English	0.547	(0.063)	0.539	(0.063)
More Maths	–0.055	(0.067)	–0.053	(0.067)
Mother's educ.	0.0190	(0.0093)	0.0187	(0.0092)
Father's educ.	0.0037	(0.0078)	0.0043	(0.0077)
Personal income	0.031	(0.014)	0.030	(0.014)
Household income	0.033	(0.011)	0.034	(0.011)
Age (years)	0.02236	(0.02810)	0.02288	(0.02777)
Quadratic age	–0.00038	(0.00029)	–0.00038	(0.00029)
σ_1^2	0.7414		0.7404	
τ			0.0098	(0.0083)
Deviance	3380.50		3354.89	

Table 4

Regression model fits for JTPA using OLS and MLE

	OLS		MLE	
Parameter	Estimate	St. error	Estimate	St. error
Intercept	–1.232	(0.470)	–1.244	(0.465)
Sex 2–1	–0.149	(0.052)	–0.148	(0.052)
Ethnicity 2–1	–0.620	(0.064)	–0.621	(0.064)
3–1	–0.699	(0.108)	–0.703	(0.107)
4–1	0.338	(0.380)	0.332	(0.375)
5–1	–0.0036	(0.160)	–0.0016	(0.159)
6–1	–0.548	(0.313)	–0.550	(0.310)
Education 2–1	0.374	(0.105)	0.374	(0.104)
3–1	0.839	(0.103)	0.832	(0.102)
4–1	1.257	(0.149)	1.254	(0.148)
5–1	0.545	(1.989)	0.547	(1.962)
Age category 2–1	0.047	(0.110)	0.049	(0.109)
3–1	–0.035	(0.172)	–0.034	(0.170)
4–1	0.032	(0.254)	0.029	(0.251)
5–1	0.077	(0.334)	0.073	(0.330)
More English	0.404	(0.062)	0.406	(0.062)
More Maths	–0.157	(0.072)	–0.154	(0.071)
Mother's educ.	0.0176	(0.0089)	0.0178	(0.0088)
Father's educ.	0.0018	(0.0074)	0.0018	(0.0073)
Personal income	–0.022	(0.011)	–0.021	(0.011)
Household income	0.003	(0.010)	0.0033	(0.0096)
Age (years)	0.03458	(0.03044)	0.03545	(0.03009)
Quadratic age	–0.00046	(0.00034)	–0.00047	(0.00033)
σ_I^2	0.5841		0.5838	
τ			0.0043	(0.0084)
Deviance	2297.88		2274.13	

Some of the estimated regression parameters in Table 3 are difficult to interpret. For instance, the parameters associated with the categories of *Education* are not in monotone order. However, this is of little importance since improved prediction is the sole purpose of the fitted regression model. Note that some of the variables (e.g., *Age category*) can be deleted from the model without substantial deterioration of the fit. Also, category 5 of *Education* is associated with very large standard error because it is represented by very few subjects. The category could be collapsed with category 4. For assessing the importance of a set of variables the likelihood ratio test is preferable to separate t-tests for each variable. Refinement of the model is dealt with in Section 6.1.

Instead of the state-level variance σ_2^2, the variance ratio $\tau = \sigma_2^2/\sigma_1^2$ and its standard error are estimated. Thus, the estimated state-level variance is $\hat{\sigma}_2^2 = 0.7404 \times 0.0098 = 0.00726$, and the corresponding standard deviation is $\hat{\sigma}_2 = \sqrt{0.00726} = 0.085$. This can be interpreted as the expected difference between the regression for a randomly drawn state and the 'average' regression given by the regression parameter vector β. The state-level variance is a source of uncertainty of the prediction for each state not represented in the survey.

The regression model fits for the JTPA data using the first set of imputed values are given in Table 4, in the same format as in Table 3. To conserve space, the regression model fits for the other sets of imputed values are not given. The estimated regression parameters for the proficiency scores are obtained by averaging over the five analyses. They are of little interest in the present context, because the predictions based on each set of imputed values will be averaged to obtain the prediction based on the proficiency scores. This way, uncertainty in estimation of the proficiency scores is allowed to permeate through all the stages of prediction.

In general, there is a lot of variation in the estimated parameters across the imputed values, as well as between the datasets. This may be of little consequence, though, because the substantially different regression parameter estimates may yield very similar predictions.

From each MLE model fit the inverse of the information matrix is stored, because it is used in estimation of the sampling variance of the prediction.

6. PREDICTION

In this section we describe prediction of the means of the proficiency scores for Mississippi and Oregon, using the national surveys and the covariate information from the surveys for these states. The equations for the predicted means and their standard errors, assuming accurately observed proficiency scores, are given by (1) and

(5). The five predictions (one for each set of imputed values) are then combined to obtain estimates which take into account the uncertainty in estimating the proficiency scores of the sampled individuals. The within-state means $\mathbf{x}^{(j)}$ were estimated by the ratio estimator, and their standard errors were obtained under the assumption of weighted random sampling design (we failed to obtain information that would identify the clusters in the state-wide surveys).

The two panels in Table 5 summarize this prediction for Mississippi and Oregon. Of principal interest is the right-most column, generated by the results based on the sets of imputed values. For comparison, the weighted sample mean ('observed' mean) and its sampling standard error from the respective statewide surveys are given in the last two lines of each panel.

Table 5

Prediction of the means of the proficiency scores for Mississippi and Oregon

		Mississippi					
		Imputed value					Prof-cy
Survey		1	2	3	4	5	score
ES/UI	Mean	0.135	0.112	0.169	0.154	0.161	0.146
	St. error	0.065	0.068	0.062	0.060	0.058	0.067
JTPA	Mean	0.067	0.084	0.099	0.109	–0.006	0.071
	St. error	0.084	0.081	0.082	0.083	0.085	0.095
Miss.	Obs. mean	–0.117	–0.122	–0.108	–0.127	–0.124	–0.120
	St. error	0.026	0.026	0.026	0.027	0.026	0.027
		Oregon					
ES/UI	Mean	0.624	0.603	0.625	0.629	0.641	0.624
	St. error	0.043	0.047	0.039	0.035	0.032	0.042
JTPA	Mean	0.486	0.516	0.539	0.495	0.467	0.501
	St. error	0.062	0.059	0.060	0.061	0.063	0.067
Oregon	Obs. mean	0.575	0.561	0.567	0.579	0.572	0.571
	St. error	0.019	0.019	0.019	0.019	0.019	0.020

The weighted sample means for both Mississippi and Oregon are at the extremes of the distribution for the within-state means of proficiency scores. The weighted within-state sample means in the ES/UI dataset are greater than that of Mississippi for all states that are represented by 25 or more subjects. In the JTPA dataset only Missouri (–0.29) has a lower weighted sample mean than Mississippi. All the within-state means in the JTPA dataset are lower than the mean for Oregon; in the ES/UI dataset Maryland (0.62) and Utah (0.63) have a slightly higher mean than Oregon, and the mean for Massachussets is 0.89. The weighted (national) sample means in ES/UI and JTPA are 0.30 and 0.04, respectively.

The predictions for Mississippi (0.146 and 0.071) are much higher than the observed weighted mean (–0.120), and the estimated standard errors for the prediction and the sample mean are too small to account for the discrepancy. For Oregon, the prediction appears to be much more successful; the discrepancy of the prediction from the observed mean is well within the estimated sampling error. Note that the standard errors for the prediction for Oregon are smaller that those for Mississippi. More detailed analysis of the sources of uncertainty in prediction can be carried out by comparing the three components of the sampling variance in (5).

The standard errors for prediction take no account of the (estimated) state-level variance. The estimates of the state-level variances for the respective surveys ES/UI and JTPA, averaged over the five analyses, are 0.00322 (standard error 0.00480) and 0.00095 (0.00414). These variances, if taken at face value, are by no means ignorable. Combined with the standard errors quoted in Table 5, on average, in the sense of (8), they inflate the standard errors for Mississippi from 0.067 to 0.087 (prediction based on ES/UI data), and from 0.095 to 0.099 (JTPA). The corresponding increases for Oregon are from 0.042 to 0.070 (ES/UI), and from 0.067 to 0.074 (JTPA). Note that the variance σ_2^2 is estimated with relatively little precision in both surveys.

6.1. Refinement of the model

The regression model given by the covariates $\mathbf{x}$ is a key element of the prediction procedure. Adequate model fit and small state-level variation are likely to be achieved by supplementing the covariates listed in Table 3 with further variables. Substantive information about the descriptive power and small reduction of the data by listwise deletion are two important criteria for selecting such variables.

First, we consider supplementing the regression model with the following covariates (abbreviated names, and for categorical variables the number of categories, are given in parentheses):

Enrolled in school? (*Sch?*, 2);
High school diploma? (*H.S.*, 2);

Military service? (*Mil.S.*, 2);
Registered to vote? (*RgVot*, 2);
How often use math on the job? (*MatU*, 5);
Reading skills good enough for your job? (*ReadJ*, 3);
Writing skills good enough for your job? (*WritJ*, 3);
Math skills good enough for your job? (*MathJ*, 3);
Better job if more English training? (*MorEn*, 2);
Better job if more math training? (*MorEn*, 2);
How often read English newspaper? (*ENws*, 5);
How many people in household? (*#Hhld*);
How often read/use reports on the job? (*Reps*);
How often write memos on the job? (*Mems*).

A number of other variables, related to the use of English language, employment, income, education, and housing, were not considered because they were either defined only for a small fraction of the population (e.g., not applicable for many subjects), or they were not collected in the statewide surveys.

Listwise deletion using these variables led to a reduction of the dataset by about 900 subjects in both datasets; from 3277 to 2367 for ES/UI, and from 2501 to 1604 for JTPA. Inclusion in the regression model of the variables listed above resulted in small changes of the state-level variance for both datasets; the within-state variance was reduced by about 13 per cent for the ES/UI dataset, and by about 8 per cent for the JTPA dataset. Most of the regression parameters are nominally statistically significant (at the 5 percent level). For brevity, details of the regression fits are omitted.

The standard errors of the prediction based on these variables are slightly higher than for the reduced model. Improvement in the model fit is negated by the reduced effective sample size. The estimates for Mississippi are 0.278 (standard error 0.067) based on the ES/UI survey, and 0.135 (0.105) based on JTPA. They are even more distant from the weighted means based on the respective statewide surveys than the estimates given in Table 5. The predictions for Oregon are changed only slightly (0.665 for ES/UI and 0.513 for JTPA).

The 900 or so subjects discarded by listwise deletion for each dataset are informative subsamples of the respective datasets. For instance, the weighted mean for the 1604 included subjects for Mississippi is –0.021, while the weighted mean for the entire sample of 2501 subjects is –0.124. The corresponding means for Oregon are 0.587 and 0.520.

Clearly, too high a price is paid for inclusion of many variables in the prediction model. In the next round of refinement we drop the quadratic age term (it has a low t-ratio for all model fits), collapse the categories 4 – 6 of *Ethnicity* (Native Americans, Asian Americans, and 'Other', each with small counts), collapse the categories 4 and

5 of *Education* (very few subjects in category 5), discard variables *H.S.*, *Mil.S.*, *MatU*, *ReadJ*, *WritJ*, *MathJ*, *ENws*, *#Hhld*, *Reps*, and *Mems* because their values are missing for large numbers of subjects, and/or they are not important in the regression model.

Table 6

Regression model fits for ES/UI and JTPA; maximum likelihood estimates

	ES/UI		JTPA	
Parameter	Estimate	St. error	Estimate	St. error
Intercept	–0.236	(0.245)	–0.345	(0.264)
Sex 2–1	–0.197	(0.050)	–0.125	(0.069)
Ethnicity 2–1	–0.703	(0.079)	–0.689	(0.078)
3–1	–0.510	(0.069)	–0.555	(0.080)
(>3)–1	–0.443	(0.116)	0.367	(0.145)
Education 2–1	0.751	(0.159)	0.690	(0.171)
3–1	1.091	(0.151)	1.051	(0.151)
(>3)–1	1.528	(0.160)	1.489	(0.168)
Age category 2–1	–0.028	(0.106)	–0.032	(0.104)
3–1	0.060	(0.120)	0.089	(0.120)
4–1	0.058	(0.163)	0.126	(0.171)
5–1	0.198	(0.262)	0.244	(0.269)
In School?	–0.259	(0.072)	–0.212	(0.077)
Reg. Voter?	–0.234	(0.055)	–0.181	(0.068)
More English	0.461	(0.066)	0.437	(0.065)
More Maths	–0.018	(0.067)	–0.041	(0.071)
Engl. pps 2–1	0.062	(0.057)	0.067	(0.057)
3–1	–0.140	(0.083)	–0.127	(0.082)
4–1	–0.231	(0.110)	–0.165	(0.122)
5–1	–0.329	(0.175)	–0.293	(0.178)
Mother's educ.	0.0163	(0.0093)	0.0176	(0.0086)
Father's educ.	–0.0026	(0.0077)	–0.0030	(0.0075)
Personal income	0.026	(0.014)	0.015	(0.013)
Household income	0.027	(0.011)	0.021	(0.011)
Age (years)	–0.0128	(0.0067)	–0.0109	(0.0069)
σ_I^2	0.7332		0.5825	
τ	0.0068	(0.0078)	0.0039	(0.0078)

Table 7

Predictions for Mississippi and Oregon based on the original model (see Table 5) and the refined model

State	ES/UI		JTPA		State	
Model	Mean	St. error	Mean	St. error	Mean	St. error
Mississippi						
Original	0.146	(0.067)	0.071	(0.095)	–0.120	(0.027)
Refined	0.140	(0.060)	0.091	(0.084)	–0.107	(0.027)
Oregon						
Original	0.624	(0.042)	0.501	(0.067)	0.571	(0.020)
Refined	0.648	(0.039)	0.499	(0.067)	0.570	(0.021)

The new regression model contains 25 regression parameters (degrees of freedom). The respective raw and effective sample sizes after listwise deletion are 3089 and 1323 for ES/UI, and 2364 and 991 for JTPA (compare with Table 1). Thus, the selected model is a compromise of adequacy (more covariates), and small loss of observations due to listwise deletion.

Table 6 contains the maximum likelihood fits to the two national survey datasets (averaged over the five sets of imputed values). Although the covariates included at the last (refinement) stage are nominally significant, their impact on the model fit is only marginal, especially for the JTPA dataset. The subject-level variance estimates are reduced insubstantially, and the variance ratio is reduced only for ES/UI (by about 30 per cent).

The predictions for Mississippi and Oregon using the original and the refined models are summarized in Table 7. The predictions of the state means, based on the refined model differ from their counterparts based on the original model (Table 5) only slightly, and the differences in predictions are trivial compared to the standard errors. The standard errors are reduced somewhat, but the conclusions drawn using the original model are not affected.

We see that even though the model fit can be improved without sacrificing a large number of records the improvement in prediction is insubstantial and a large apparent bias (for Mississippi) remains unexplained. This 'robustness' feature of the prediction model is very desirable because model selection has to rely to a large extent on expediency (availability of data) rather than familiarity with the subject matter, underlying theory, or formal statistical procedures.

7. CONCLUSION

A prediction method for estimating within-state means of literacy proficiency scores from national surveys, based on a random-effects regression model, is presented and illustrated on two states, Mississippi and Oregon. The method combines the advantages of the design-based and model-based approaches by incorporating sampling weights, by imposing a model which captures the clustered structure of the data, and by using linear regression to reduce the residual variation at both cluster and elementary levels. For the two states statewide survey data are available, and so the predictions from national surveys can be compared with more reliable estimates from the state-wide surveys. The estimated standard errors of the prediction can be compared with the standard errors of the (weighted) sample means to assess the usefulness of the prediction.

For instance, the predictions for Mississippi, based on the ES/UI and JTPA data, have about 2.5 and 3.5 times greater standard errors than the sample mean. This can be interpreted that information about Mississippi in the national surveys is equivalent to that in a sample about ($2.5^2 =$) six times (ES/UI) and twelve times (JTPA) smaller than the sample collected in Mississippi. The corresponding factors for Oregon are smaller, about four for ES/UI and eleven for JTPA. These factors can be adjusted for unequal (effective) sample sizes in the obvious manner. Although these conclusions are contingent on the selected regression model, the prediction appears to be fairly robust with respect to model specification.

The only iterative component of the prediction procedure is the fitting of the random coefficient model. Using the 'weighted' version of the Fisher scoring algorithm (Jennrich and Schluchter, 1986; Longford, 1987) no problems with convergence or multiple local maxima arise; usually less than twelve iterations are required to achieve convergence using any reasonable criterion for convergence.

The method can be extended to multiple layers of clustering (see Longford, 1987) and to non-normally distributed data by application of generalized linear models with random coefficients (Longford 1994). In the latter case it is assumed that the regression estimator $\hat{\beta}$ and the estimator $\bar{x}$ of the state's mean are (approximately) normally distributed. Owing to the asymptotic theory, this is a realistic assumption.

The predictions based on the two national surveys can be combined. The coefficients of the convex combination of the predictions can be determined so as to minimize the standard error of the combined estimator. The coefficients of this convex combination are inversely proportional to the variances of the components. Thus, the combined estimator for Mississippi has coefficients 0.67 (ES/UI) and 0.33 (JTPA), and the resulting prediction for the state is 0.121 (standard error 0.054). Of course, prediction from one or several national surveys can also be combined with an esti-

mate based on the data from the state. In the case of Mississippi this would lead to only a marginal improvement because the estimate based on the Mississippi survey is far superior to the prediction from either national survey, or their combination. The combined prediction for Oregon is equal to 0.590 (standard error 0.035), very close to the weighted sample mean of 0.571.

ACKNOWLEDGEMENTS

Norma Norris and Kate Pashley assisted me with orientation in the databases. Discussions with Irwin Kirsch, Peter Pashley, and Kentaro Yamamoto are acknowledged. The last revision of the manuscript benefited from the referees' comments and suggestions.

APPENDIX: SAMPLING VARIANCE OF THE PREDICTOR $\hat{y}^{(j)}$

Suppose $\hat{\mathbf{x}}^{(j)} \sim \mathcal{N}(\mathbf{x}^{(j)}, \Sigma_S)$, $\hat{\beta} \sim \mathcal{N}(\beta, \Sigma_b)$ and $\text{cov}(\hat{\mathbf{x}}^{(j)}, \hat{\beta}) = \Sigma_{bS}$. By conditioning on $\hat{\beta}$ we obtain

$$\begin{aligned}
\text{var}\left(\hat{y}^{(j)}\right) &= \mathbf{E}_\beta\left\{\text{var}\left(\hat{\mathbf{x}}^{(j)}\hat{\beta} \mid \hat{\beta}\right)\right\} + \text{var}_\beta\left\{\mathbf{E}\left(\hat{\mathbf{x}}^{(j)}\hat{\beta} \mid \hat{\beta}\right)\right\} \\
&= \mathbf{E}_\beta\left\{\hat{\beta}^\top\left(\Sigma_S - \Sigma_{bS}^\top\Sigma_b^{-1}\Sigma_{bS}\right)\hat{\beta}\right\} \\
&\quad + \text{var}_\beta\left[\left\{\mathbf{x}^{(j)} + \left(\hat{\beta} - \beta\right)^\top\Sigma_b^{-1}\Sigma_{bS}\right\}\hat{\beta}\right].
\end{aligned}$$

For an arbitrary $p \times p$ matrix of constants $\mathbf{A}$ we have the following identities:

$$\begin{aligned}
\mathbf{E}\left(\hat{\beta}^\top\mathbf{A}\hat{\beta}\right) &= \beta^\top\mathbf{A}\beta + \text{tr}(\mathbf{A}\Sigma_b) \\
\text{cov}\left(\hat{\beta}, \hat{\beta}^\top\mathbf{A}\hat{\beta}\right) &= \Sigma_b(\mathbf{A} + \mathbf{A}^\top)\beta \\
\text{var}\left(\hat{\beta}^\top\mathbf{A}\hat{\beta}\right) &= \text{tr}(\mathbf{A}\Sigma_b\mathbf{A}\Sigma_b) + \text{tr}\left(\mathbf{A}\Sigma_b\mathbf{A}^\top\Sigma_b\right) \\
&\quad + 2\beta^\top\mathbf{A}\Sigma_b\mathbf{A}\beta + 2\beta^\top\mathbf{A}\Sigma_b\mathbf{A}^\top\beta.
\end{aligned}$$

The versions of these identities for a symmetric matrix $\mathbf{A}$ are well-known, see, e.g., Seber (1977, Ch. 2). Their general versions are derived by application of the 'symmetric' versions to $\mathbf{A} + \mathbf{A}^\top$, noting that $\hat{\beta}^\top\mathbf{A}\hat{\beta} = \hat{\beta}^\top\mathbf{A}^\top\hat{\beta}$. Substitution of these identities

for appropriate matrices **A** yields

$$\begin{aligned}\operatorname{var}\left(\hat{y}^{(j)}\right) &= \beta^{\top}\Sigma_S\beta + \operatorname{tr}\left(\Sigma_b\Sigma_S\right) - \beta^{\top}\Sigma_{bS}^{\top}\Sigma_b^{-1}\Sigma_{bS}\beta - \operatorname{tr}\left(\Sigma_{bS}^{\top}\Sigma_b^{-1}\Sigma_{bS}\Sigma_b\right) \\ &\quad + \left(x^{(j)} - \beta^{\top}\Sigma_b^{-1}\Sigma_{bS}\right)\Sigma_b\left(x^{(j)} - \beta^{\top}\Sigma_b^{-1}\Sigma_{bS}\right)^{\top} \\ &\quad + 2\left(x^{(j)} - \beta^{\top}\Sigma_b^{-1}\Sigma_{bS}\right)\left(\Sigma_{bS} + \Sigma_b\Sigma_{bS}^{\top}\Sigma_b^{-1}\right)\beta \\ &\quad + \operatorname{tr}\left(\Sigma_{bS}^2\right) + \operatorname{tr}\left(\Sigma_b^{-1}\Sigma_{bS}\Sigma_b\Sigma_{bS}^{\top}\right) + 2\beta^{\top}\Sigma_b^{-1}\Sigma_{bS}^2\beta \\ &\quad + 2\beta^{\top}\Sigma_b^{-1}\Sigma_{bS}\Sigma_b\Sigma_{bS}^{\top}\Sigma_b^{-1}\beta,\end{aligned}$$

from which equation (6) follows directly.

REFERENCES

[1] **Battese, G. E., Harter, R. M.,** and **Fuller, W. A.** (1988). «An error-components model for prediction of county crop areas using survey and satellite data». *Journal of American Statistical Association* **83**, 28–36.

[2] **Beran, R.,** and **Hall, P.** (1992). «Estimating coefficient distributions in random coefficient regressions». *Annals of Statistics* **20**, 1979–1984.

[3] **Bock, R. D.** and **Aitkin, M.** (1981). «Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm». *Psychometrika* **46**, 443–459.

[4] **Harville, D. A.** (1974). «Bayesian inference for variance components using only error contrasts». *Biometrika* **61**, 383–385.

[5] **Jennrich, R. I.** and **Schluchter, M. D.** (1986). «Unbalanced repeated-measures models with structured covariance matrices». *Biometrics* **42**, 805–820.

[6] **Kackar, R. N.** and **Harville, D. A.** (1984). «Approximations for standard errors of estimators of fixed and random effects in mixed linear models». *Journal of the American Statistical Association* **79**, 853–862.

[7] **Kirsch, I. S.** and **Jungeblut, A.** (1992). *Profiling the literacy proficiencies of JTPA and ES/UI populations*. Final Report to the Department of Labor. Educational Testing Service, Princeton, NJ.

[8] **Longford, N. T.** (1987). «A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects». *Biometrika* **74**, 817–827.

[9] **Longford, N. T.** (1994). «Logistic regression with random coefficients». *Computational Statistics and Data Analysis* **17**, 1–15.

[10] **Longford, N. T.** (1995). «Random coefficient models and missing data». Submitted.

[11] **Mislevy, R. J.** and **Bock, R. D.** (1983). *BILOG: Item analysis and test scoring with binary logistic models*. Scientific Software, Inc., Mooresville, IN.

[12] **Potthoff, R. F., Woodbury, M. A.** and **Manton, K. G.** (1992). «Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models». *Journal of the American Statistical Association* **87**, 383–396.

[13] **Seber, G. A. F.** (1977). *Linear Regression Analysis*. Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, New York.

Investigació Operativa

QÜESTIIÓ, Vol 20, 2 pp. 215-222, 1996

A STABILITY THEOREM IN NONLINEAR BILEVEL PROGRAMMING**

SHOU-YANG WANG*, QIAN WANG† and LUIS COLADAS URIA‡

In this short paper, we are concerned with the stability of nonlinear bilevel programs. A stability theorem is proven and an example is given to illustrate this theorem.

Keywords: Bilevel programming, stability.

Bilevel programming, a nested optimization problem, emerged as the appropriate model to formulate a hierarchical decision making situation where the higher level in the hierarchy can only influence rather than dictate the choices of the lower level (Bard, 1984; Bialas and Karwan, 1984; Wang and Lootsma, 1994). Most of the investigations in this field are focused on optimality conditions and algorithms (see comments made in Chen and Florian, 1995; Wang, Wang and Romano-Rodríguez, 1994). Since a parametric solution or error bounds on a solution with perturbed data are typically of great interest both in practical applications and in theoretical characterizations (Fiacco, 1983), to study stability of an optimal solution to a bilevel programming problem is certainly a very important topic in bilevel programming.

** Supported in part by NSFC and MADIS. This paper was completed during the first author's stay at Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, Spain. He is grateful to the financial support for his visit by Consejo Superior de Investigaciones Científicas (CSIC) and Universidad de Santiago de Compostela. The authors are very grateful to the referees' valuable comments and suggestions.

* Institute of Systems Science, Chinese Academy of Sciences, Beijing 100080, China.

† Department of Mathematics, University of Nebraska at Omaha, Omaha, Nebraska 68182-0243, USA.

‡ Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, 15706 Santiago de Compostela, Spain.

- Article rebut el juliol de 1995.

- Acceptat el maig de 1996.

The bilevel programming problem with parameter considered in this paper is stated as $(BLP(\varepsilon))$:

$$\begin{aligned} &\underset{x\in X}{\text{minimize}}\, F(x,y,\varepsilon) \text{ where } y \text{ solves} \\ &\quad \underset{y\in Y}{\text{minimize}}\, f(x,y,\varepsilon) \\ &\quad\quad \text{subject to } g_i(x,y,\varepsilon)\le 0, \quad i=1,\cdots,m \end{aligned}$$

where $x\in R^{n_1}$ and $y\in R^{n_2}$ are controlled by the leader and the follower respectively, X and Y are closed convex sets in R^{n_1} and R^{n_2} respectively, ε is a parameter vector in R^k, $F(x,y,\varepsilon)$ and $f(x,y,\varepsilon)$ are the objective functions of the leader and the follower respectively, $g_i(x,y,\varepsilon), i=1,\cdots,m$, are the constraint functions.

When the parameter vector ε is identical with the zero vector (*i.e.*, no data is perturbed in the model), the problem $(BLP(\varepsilon))$ can be written as (BLP):

$$\begin{aligned} &\underset{x\in X}{\text{minimize}}\, F(x,y) \text{ where } y \text{ solves} \\ &\quad \underset{y\in Y}{\text{minimize}}\, f(x,y) \\ &\quad\quad \text{subject to } g_i(x,y)\le 0, \quad i=1,\cdots,m \end{aligned}$$

where $F(x,y)=F(x,y,0)$, $f(x,y)=f(x,y,0)$ and $g_i(x,y)=g_i(x,y,0), i=1,\cdots,m$. This type of bilevel programming problems have been extensively studied by many authors. Refer to Ben-Ayed (1993) and Wang and Wang (1994) for a survey.

For a given $x\in X$, we denote the inner program as $P(x,\varepsilon)$ and define

$$Y(x,\varepsilon)=\{y \mid y \text{ is a minimum of } P(x,\varepsilon)\},$$

$$\bar{S}(\varepsilon)=\{(x,y)\in X\times Y \mid g_i(x,y,\varepsilon)\le 0, i=1,\cdots,m, \text{ and } y\in Y(x,\varepsilon)\}$$

and

$$L(x,y,\varepsilon,u)=f(x,y,\varepsilon)+\sum_{i=1}^{m} u_i g_i(x,y,\varepsilon)$$

We make the following assumptions:

(i) the bilevel problem is well–posed, *i.e.*, $Y(x,\varepsilon)$ is a singleton and the unique element is denoted as $y(x,\varepsilon)$;

(ii) F, f and $g_i(i=1,\cdots,m)$ are twice continuously differentiable in y, their gradients with respect to y and $g_i(i=1,\cdots,m)$ are continuously differentiable in both x and ε, f is convex in y;

(iii) for any x, the second–order sufficient conditions for a minimum of $P(x,\varepsilon)$ holds at $y(x,\varepsilon)$, with associated Lagrange multipliers $u(x,\varepsilon)$ *i.e.*, for any $s\neq 0$ that satisfies

$$s^T \nabla_y g_i(x, y(x,\varepsilon),\varepsilon) = 0, i \in I_1(x,\varepsilon)$$
$$s^T \nabla_y g_i(x, y(x,\varepsilon),\varepsilon) \leq 0, i \in I_2(x,\varepsilon)$$

$s^T \nabla^2_{yy} L(x, y(x,\varepsilon,\varepsilon, u(x,\varepsilon)))s > 0$ holds where $I_1(x,\varepsilon) \triangleq \{j \mid g_j(x, y(x,\varepsilon),\varepsilon) = 0, u_j(x,\varepsilon) > 0\}$ and $I_2(x,\varepsilon) \triangleq \{j \mid g_j(x, y(x,\varepsilon),\varepsilon) = 0, u_j(x,\varepsilon) = 0\}$;

(iv) the gradients $\nabla_y g_i(x, y(x,\varepsilon),\varepsilon), i \in I_0(x,\varepsilon) \triangleq \{j \mid g_j(x, y(x,\varepsilon),\varepsilon) = 0\}$ are linearly independent;

(v) strict complementary slackness holds, *i.e.*, $u_i(x,\varepsilon) > 0$ when $i \in I_0(x,\varepsilon)$;

(vi) $F(x,y,\varepsilon)$ is continuous on $X \times Y \times R^k$ and X and Y are compact.

A pair $(x^*(\varepsilon), y^*(\varepsilon))$ is said to be *an optimal solution to* $(BLP(\varepsilon))$ if it satisfies (i) $y^*(\varepsilon) \in Y(x^*(\varepsilon),\varepsilon)$ and (ii) $F(x^*(\varepsilon), y^*(\varepsilon),\varepsilon) \leq F(x,y,\varepsilon)$ for any $(x,y) \in \tilde{S}(\varepsilon)$.

Define

$$M(x) = \begin{bmatrix} \nabla^2_{yy} L(x, y(x,0), 0, u(x,0)) & \nabla_y g_1(x, y(x,0), 0), & \cdots & \nabla_y g_m(x, y(x,0), 0) \\ u_1 \nabla_y g_1^T(x, y(x,0), 0) & g_1(x, y(x,0), 0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_m \nabla_y g_m^T(x, y(x,0), 0) & 0 & \cdots & g_m(x, y(x,0), 0) \end{bmatrix}$$

and

$$N(x) = [-\nabla^2_{\varepsilon x} L(x, y(x,0), 0, u(x,0)), -u_1 \nabla_\varepsilon g_1(x, y(x,0), 0), \cdots, -u_m \nabla_\varepsilon g_m(x, y(x,0), 0)]^T.$$

Lemma 1

For any given $x \in X$, $M(x)$ is nonsingular.

Proof

Without loss of generality, let $I_0(x,0) = \{1, \cdots, p\}, I \setminus I_0(x,0) = \{p+1, \cdots, m\}$. Denote

$$G = (\nabla_y g_1(x, y(x,0), 0), \cdots, \nabla_y g_p(x, y(x,0), 0)),$$
$$\bar{G} = (\nabla_y g_{p+1}(x, y(x,0), 0), \cdots, \nabla_y g_m(x, y(x,0), 0)),$$
$$U = diag(u_1(x,0), \cdots, u_p(x,0))$$

and

$$D = diag(g_{p+1}(x, y(x,0), 0), \cdots, g_m(x, y(x,0), 0)).$$

Then

$$M(x) = \begin{pmatrix} \nabla^2_{yy} L(x, y(x,0), 0, u(x,0)) & G & \bar{G} \\ UG^T & 0 & 0 \\ 0 & 0 & D \end{pmatrix}.$$

From Assumption (v) and the assumption of $I_0(x,0)$, $u_i(x,0) > 0, i = 1, \cdots, p$ and $g_j(x, y(x,0), 0) < 0, j = p+1, \cdots, m$. So the matrices U and D are nonsingular. Hence, it is only required to show that matrix

$$\begin{pmatrix} \nabla^2_{yy} L(x, y(x,0), 0, u(x,0)) & G \\ UG^T & 0 \end{pmatrix}$$

is nonsingular. This is equivalent to prove that the following system

$$\nabla^2_{yy} L(x, y(x,0), 0, u(x,0))s - Gz = 0 \tag{1a}$$

$$UG^T s = 0 \tag{1b}$$

has the unique solution $s = 0, z = 0$.

From (1b), we get $G^T s = 0$. Hence, s satisfies Assumption (iii). Multiplying (1a) by s, we have

$$s^T \nabla^2_{yy} L(x, y(x,0), 0, u(x,0))s - s^T Gz = 0,$$

$$s^T \nabla^2_{yy} L(x, y(x,0), 0, u(x,0))s = 0.$$

By Assumption (iii), we get $s = 0$. Thus, $Gz = 0$. Owing to Assumption (iv), the column rank of G is full. Hence, $z = 0$. Therefore, $M(x)$ is nonsingular.

■

Lemma 2

For any given $\bar{x} \in X$, the following first-order approximation

$$\begin{bmatrix} y(\bar{x}, \varepsilon) \\ u(\bar{x}, \varepsilon) \end{bmatrix} = \begin{bmatrix} y(\bar{x}, 0) \\ u(\bar{x}, 0) \end{bmatrix} + M(\bar{x})^{-1} N(\bar{x})\varepsilon + o(\|\varepsilon\|) \tag{2}$$

holds in a neighborhood of $\varepsilon = 0$.

Proof

From Assumption (iii), we know that

$$\nabla_y L(x, y, \varepsilon, u) = 0 \tag{3a}$$

$$u_i g_i(x, y, \varepsilon) = 0, i = 1, \cdots, m \tag{3b}$$

hold at $(\bar{x}, y(\bar{x},0), 0, u(\bar{x},0))$. By Lemma 1, the inverse of the Jacobian of the vector-valued function $(\nabla_y L(x,y,\varepsilon,u), u_1 g_1(x,y,\varepsilon), \cdots, u_m g_m(x,y,\varepsilon))$ with respect to (y,u) exists. Hence, the assumptions of the implicit function theorem with respect to (3) are satisfied and we can conclude that in a neighborhood of $\varepsilon = 0$, there exists a unique continuously differentiable function $(y(\bar{x},\varepsilon), u(\bar{x},\varepsilon))$ satisfying (3). This implies that for any ε near 0, $y(\bar{x},\varepsilon)$ is a Kuhn–Tucker point of $P(x,\varepsilon)$ with associated Lagrange multipliers $u(\bar{x},\varepsilon)$.

The gradient of $(y(\bar{x},\varepsilon), u(\bar{x},\varepsilon))$ with respect to ε at $\varepsilon = 0$ is $M(\bar{x})^{-1}N(\bar{x})$. So the conclusion of this lemma holds.

■

Lemma 3

$F(x,y,\varepsilon)$ is uniformly continuous on $X \times Y \times N_0(\varepsilon)$ and $M(x)^{-1}N(x)$ is uniformly bounded on X, where $N_0(\varepsilon)$ is a neighborhood of $\varepsilon = 0$.

Proof

It is not difficult to show $M(x)$ and $N(x)$ are continuous on X. Since $M(x)$ is nonsingular for all $x \in X$, $M(x)^{-1}N(x)$ is continuous on X. Hence, we can get this result from the properties that continuous functions are uniformly continuous and uniformly bounded on compact sets.

■

Let (x^*, y^*) be the unique optimal solution of problem $(BLP(0))$. Then, we can prove the following main result.

Theorem 1

Suppose Assumption (i)–(vi) are satisfied. Then for any given positive number ν, there exists a δ such that when $\|\varepsilon\| < \delta$,

$$|F(x^*(\varepsilon), y^*(\varepsilon), \varepsilon) - F(x^*, y^*, 0)| < \nu.$$

Proof

Let $(x^*(\varepsilon), y^*(\varepsilon))$ be the optimal solution of $(BLP(\varepsilon))$, then $y^*(\varepsilon) = y(x^*(\varepsilon), \varepsilon)$. Denote $\sigma_1 = |F(x^*(\varepsilon), y^*(\varepsilon), \varepsilon) - F(x^*(\varepsilon), y(x^*(\varepsilon), 0), 0)|$ and $\sigma_2 = |F(x^*, y(x^*, \varepsilon), \varepsilon) - F(x^*, y^*, 0)|$. Since

$$F(x^*(\varepsilon), y(x^*(\varepsilon), 0), 0) \geq F(x^*, y^*, 0)$$

and

$$F(x^*,y(x^*,\varepsilon),\varepsilon) \geq F(x^*(\varepsilon),y^*(\varepsilon),\varepsilon),$$

$$|F(x^*(\varepsilon),y^*(\varepsilon),\varepsilon) - F(x^*,y^*,0)| \leq \max\{\sigma_1,\sigma_2\}.$$

From Lemma 2, for $x^*(\varepsilon)$ and x^*, we have the following first-order approximations in a neighborhood of $\varepsilon = 0$,

$$\begin{bmatrix} y(x^*(\varepsilon),\varepsilon) \\ u(x^*(\varepsilon),\varepsilon) \end{bmatrix} = \begin{bmatrix} y(x^*(\varepsilon),0) \\ u(x^*(\varepsilon),0) \end{bmatrix} + M(x^*(\varepsilon))^{-1}N(x^*(\varepsilon))\varepsilon + o(\|\varepsilon\|) \tag{4}$$

$$\begin{bmatrix} y(x^*,\varepsilon) \\ u(x^*,\varepsilon) \end{bmatrix} = \begin{bmatrix} y(x^*,0) \\ u(x^*,0) \end{bmatrix} + M(x^*)^{-1}N(x^*)\varepsilon + o(\|\varepsilon\|) \tag{5}$$

Because of the uniformly continuity of $F(x,y,\varepsilon)$, for any given positive number v, there exists a δ_1 such that when $|(y^*(\varepsilon),\varepsilon) - (y(x^*(\varepsilon),0),0)| < \delta_1$, we have $\sigma_1 < v$. By the uniformly boundedness of $M(x)^{-1}N(x)$ and (4), there exists a δ_2 such that when $\|\varepsilon\| < \delta_2$, $|(y^*(\varepsilon),\varepsilon) - (y(x^*(\varepsilon),0),0)| < \delta_1$ holds. With an almost same analysis, we can find a δ_3 such that when $\|\varepsilon\| < \delta_3, \sigma_2 < v$. Let $\delta = \min\{\delta_2,\delta_3\}$. We can conclude that for any given v, there exist a δ such that when $\|\varepsilon\| < \delta$, $|F(x^*(\varepsilon),y^*(\varepsilon),\varepsilon) - F(x^*,y^*,0)| < v$.

■

Now we give an example to illustrate the above theorem.

Example 1

Consider the following bilevel programming problem $(P(\varepsilon))$:

$$\underset{x\in X}{\text{minimize}}\, x^2 + 2\varepsilon_1 y \text{ where } y \text{ solves}$$

$$\underset{y\in Y}{\text{minimize}}\, y$$

$$\text{subject to } y - x \geq \varepsilon_2$$

where $x \in R^1$ and $y \in R^1$ are controlled by the leader and the follower respectively, $X = \{x \in R^1 \,|\, |x| \leq M\}$ and $Y = \{y \in R^1 \,|\, |y| \leq 2M\}$, M is a given positive number and ε is a parameter vector in R^2 satisfying $|\varepsilon_2| < M$.

For any given $x \in X$, the unique optimal solution of the inner problem

$$\underset{y\in Y}{\text{minimize}}\, y$$

$$\text{subject to } y - x \geq \varepsilon_2$$

is $y^*(x,\varepsilon) = x + \varepsilon_2$. So the problem $(P(\varepsilon))$ can be reformulated as

$$\underset{x \in X}{\text{minimize}}\, x^2 + 2\varepsilon_1(x + \varepsilon_2).$$

It can be easily shown that the optimal solution of this minimization problem is $x^*(\varepsilon) = -\varepsilon_1$ and the optimal objective value $F(x^*(\varepsilon), y^*(\varepsilon)) = 2\varepsilon_1\varepsilon_2 - \varepsilon_1^2$. When $\varepsilon_1 = \varepsilon_2 = 0$, $(P(\varepsilon))$ is reduced to the following bilevel programming problem $(P(0))$:

$$\underset{x \in X}{\text{minimize}}\, x^2$$

$$\underset{y \in Y}{\text{minimize}}\, y$$

$$\text{subject to } y - x \geq 0.$$

It is obvious that the optimal objective value of this problem is $F(x^*, y^*, 0) = 0$.

It is not hard to verify that Assumptions (i) - (vi) are satisfied. By Theorem 1, for any given positive number ν, there exists a δ such that when $\|\varepsilon\| < \delta$,

$$|F(x^*(\varepsilon), y^*(\varepsilon), \varepsilon) - F(x^*, y^*, 0)| < \nu.$$

In fact, if we choose $\delta = \sqrt{\frac{\nu}{3}}$, then

$$|F(x^*(\varepsilon), y^*(\varepsilon), \varepsilon) - F(x^*, y^*, 0)| = |2\varepsilon_1\varepsilon_2 - \varepsilon_1^2|$$

$$\leq |\varepsilon_1|^2 + 2|\varepsilon_1\varepsilon_2| \leq \|\varepsilon\|^2 + 2\|\varepsilon\|^2 = 3\|\varepsilon\|^2 < \nu.$$

REFERENCES

[1] **Bard, J.F.** (1984). «Optimality conditions for the bilevel programming problem», *Naval Research Logistics Quarterly*, **31**, 13–26.

[2] **Ben–Ayed, O.** (1993). «Bilevel linear programming», *Computers and Operations Research*, **20**, 485–501.

[3] **Bialas, W.F. and Karwan, M.H.** (1984). «Two–level linear programming», *Management Science*, **30**, 1004–1020.

[4] **Chen, Y. and Florian, M.** (1995). «The nonlinear bilevel programming problem: formulations, regularity and optimality conditions», *Optimization*, **32**, 193–209.

[5] **Fiacco, A.V.** (1983). *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, New York.

[6] **Wang, Q. and Wang, S.Y.** (1994). «Bilevel programs with multiple potential reactions», *Journal of Systems Science and Systems Engineering*, **3**, 269–278.

[7] **Wang, S.Y. and Lootsma, F.A.** (1994). «A hierarchical optimization model of resource allocation», *Optimization*, **28**, 351–361.

[8] **Wang, S.Y., Wang, Q. and Romano–Rodríguez, S.** (1994). «Optimality conditions and an algorithm for linear–quadratic bilevel programs», *Optimization*, **31**, 127–139.

QÜESTIIÓ, Vol 20, 2 pp. 223-238, 1996

INTERPRETACIÓN DE LOS PRECIOS SOMBRA EN PRESENCIA DE DEGENERACIÓN

TERESA LEÓN* y VICENTE LIERN†

Universitat de València

El propósito de nuestro trabajo es analizar la interpretación económica de las variables duales como «precios sombra» cuando la solución posible básica óptima del problema de programación lineal es degenerada. Resolvemos algunos ejemplos que ilustran esta interpretación usando el programa LINDO. Finalmente planteamos una modificación a la propuesta de Gal (1986) para llevar a cabo el análisis de sensibilidad en presencia de degeneración.

An interpretation of the shadow prices under degeneracy

Keywords: Precios Sombra, Programación Lineal, Simplex Paramétrico.

1. INTRODUCCIÓN

El objetivo de la Investigación Operativa en las Facultades o Escuelas de Ciencias Económicas y Empresariales es capacitar al futuro titulado para resolver problemas reales en el ámbito de la empresa. Sin embargo, los ejemplos que manejamos en las aulas, suelen adolecer de un excesivo academicismo, que si bien resulta útil para la comprensión, puede ser engañoso por ignorar situaciones que en la práctica son habituales. En este sentido, nos parece interesante partir de una situación que en las clases «nos conviene» evitar cuando hacemos análisis de la sensibilidad: **las soluciones óptimas degeneradas**.

*Departamento de Estadística e Investigación Operativa, Universitat de València. Doctor Moliner, 50. 46100-Burjassot (València). Tel: (6) 386 43 54. Fax: (6) 386 47 35.

†Dep. Economía Financiera y Matemática. Artes Gráficas 29. Universitat de València. 46010 València. Tel: (6) 386 40 50, Fax: (6) 386 43 64, **E-mail:** Vicente.Liern **at** uv.es

–Article rebut l'abril de 1995.

–Acceptat el juny de 1996.

Las técnicas del análisis de sensibilidad convierten el modelo de Programación Lineal en una herramienta útil para la planificación de problemas reales. En numerosas ocasiones, cuando se formula un problema real como uno de programación lineal, tanto los coeficientes de la matriz de restricciones como los de la función y los términos independientes son estimaciones basadas en la experiencia actual y en la predicción de condiciones futuras. Sin embargo, es frecuente que, para llevar a cabo estas estimaciones, se manejen datos recogidos de forma poco rigurosa. Por tanto, es razonable mantener ciertas dudas respecto a la interpretación de los resultados obtenidos al optimizar, y antes de convertir una solución óptima en una política de actuación, deberá comprobarse que su comportamiento sigue siendo bueno para otras representaciones plausibles del problema real.

El análisis de sensibilidad permite, entre otras cosas, evaluar el impacto de la inclusión de nuevos productos en un plan de producción o determinar el precio, a partir del cual, un producto cuya fabricación no resultaba interesante pasa a ser competitivo.

En particular, las soluciones óptimas del problema dual del modelo lineal que se ha construido, suelen utilizarse para predecir la variación que se produciría en el valor de la función objetivo si se aumenta o disminuye el término independiente de una restricción. Por ejemplo, si nos centramos en un problema de maximización de beneficios sujeto a restricciones de capacidad o limitación de recursos, se utilizaría para determinar el aumento en los beneficios que se produciría al conseguir recursos adicionales que se han agotado al diseñar la política de producción óptima.

La idea fundamental de este trabajo es remarcar la diferencia que existe en la interpretación del análisis de sensibilidad de los términos independientes y de las variables duales cuando la solución posible básica (SPB) óptima del modelo lineal es degenerada y cuando no lo es. Una cuestión a tener en cuenta es que de la lectura de los listados de las soluciones que proporcionan la mayoría de paquetes comerciales, no se desprende, a simple vista, si estamos ante una SPB óptima degenerada o no. Además, debido a que el análisis más conocido y directo es el que puede hacerse en el caso no degenerado, es fácil comprender que se han producido errores de interpretación en algunas situaciones reales (véase Rubin y Wagner (1990)).

2. PLANTEAMIENTO DEL PROBLEMA

Sea (P) el problema lineal

$$(P)\quad \begin{array}{ll} Max & CX \\ \text{s.a} & A_o X \leq b \\ & X \geq 0_n \end{array}$$

donde C es un vector fila, A_o es una matriz $m \times n$ y b un vector columna fijos, y $X \in \mathbb{R}^n$ es el vector de variables del problema.

Introduciendo variables de holgura $h_j \geq 0$, $1 \leq j \leq m$, (P) puede escribirse en forma stándard (P') como sigue:

$$\begin{array}{lll} \text{(P')} & Max & cx \\ & \text{s. a} & Ax = b \\ & & x \geq 0_{n+m} \end{array}$$

donde $c = (C, 0_m)$, $A = (A_o, I_m)$, $x = \begin{pmatrix} X \\ h \end{pmatrix}$.

Sea $\bar{x} = \begin{pmatrix} x_B \\ x_N \end{pmatrix} = \begin{pmatrix} B^{-1}b \\ 0 \end{pmatrix}$ una solución posible básica de (P'). Se dice que $\bar{x}$ es ***degenerada*** si existe alguna variable básica x_{B_j} de modo que $\bar{x}_{B_j} = 0$. En este caso existen $\{B_1, ...B_k\}$ submatrices invertibles $m \times m$ de la matriz de restricciones A tales que $B_j \bar{x} = b$ para $1 \leq j \leq k$.

El problema dual de (P) que representaremos por (D) es

$$\begin{array}{lll} \text{(D)} & Min & wb \\ & \text{s. a} & wA_o \geq C \\ & & w \geq 0_m^T \end{array}$$

que por supuesto es equivalente a (D'), el dual de (P')

$$\begin{array}{lll} \text{(D')} & Min & wb \\ & \text{s. a} & wA \geq c. \end{array}$$

Una base B de (P') es ***dual posible*** si $w = c_B B^{-1} A$ es una solución posible de (D').

- Si la SPB óptima de (P'), x^* es no degenerada, tiene asociada una única base dual posible, y por tanto, una única solución óptima dual $w^* = (w_1^*, .., w_m^*)$. Cada w_i^* se interpreta como el índice de aumento (res. de disminución) del valor de la función objetivo en el óptimo si el término independiente de la restricción i–ésima, b_i, se incrementa (resp. se disminuye). De aquí que, en el contexto del análisis de sensibilidad a las variables duales se les llame «precios sombra», «precios duales» o «valores marginales».

- Por el contrario, si la SPB óptima de (P'), x^* es degenerada puede tener más de una solución óptima dual asociada. Entonces, ¿qué interpretación económica debemos dar a las variables duales?

3. INTERPRETACIÓN DE LAS VARIABLES DUALES EN PRESENCIA DE DEGENERACIÓN

Supongamos que x^* es una SPB óptima degenerada del problema primal y $w^* = (w_1^*, ..., w_m^*)$ es solución óptima del dual (D), vamos a ver que w_i^* es una cota superior del aumento de la función objetivo si aumenta en una unidad el recurso i–ésimo, y una cota inferior a la disminución del valor de la función objetivo si disminuye en una unidad el recurso i–ésimo. Para ello vamos a considerar un resultado más general cuya demostración puede hacerse, por ejemplo, siguiendo a Shapiro (1984).

Recordemos que si $v(b) = \text{Max } \{CX \mid \quad A_o X \leq b, \quad X \geq 0\}$ denota el coste máximo del problema primal como función de b, entonces $v(.)$ es cóncava en $\mathbb{R}^m$.

Dado $b_o \in \mathbb{R}^n$ de modo que $v(b_o)$ es finito, una dirección $d \neq 0_m$ diremos que es una ***dirección posible*** si existe $\theta^* \geq 0$ de manera que $v(b_o + \theta d)$ es finito para cualquier $\theta \in [0, \theta^*]$.

Entonces, se tiene el siguiente resultado:

Teorema 1

Supongamos que $v(b_o)$ es finito y que $d \neq 0$ es una dirección posible de $v(.)$ en b_o. Entonces, la derivada direccional de $v(.)$ en b_o en la dirección de d existe y viene dada por

$$\nabla v(b_o; d) = Min\{w^* d \mid w^* \text{ es solución óptima de } (D_o)\},$$

donde D_o es el dual del problema primal (P) cuando $b = b_o$.

Este resultado tiene unas consecuencias inmediatas muy apropiadas para nuestra intención:

a) Si el conjunto de soluciones óptimas de D_o se reduce a un único vector w^* entonces $\nabla v(b_o; d) = w^* d$ para cualquier dirección posible d.

b) Si además el conjunto de direcciones posibles es $\mathbb{R}^m - \{0_m\}$, entonces $v(.)$ es diferenciable y $\dfrac{\partial v}{\partial b_i}(b_o) = w_i^*$ y obtenemos ***los precios sombra***.

c) Si (D_o) tiene más de una solución óptima,

$$\nabla v(b_o; e_i) = Min\{w_i^* \mid w^* = (w_1^*, ..., w_m^*) \text{ es solución óptima de } (D_o)\}.$$

Entonces, dado un problema de la forma $(P)\,\mathbf{Max}\{CX \mid \quad A_o X \leq b, \quad X \geq 0\}$, para hallar el aumento en el valor de la función objetivo que se produciría si b_i aumenta

una unidad tendríamos que hallar el

$$Min\{w_i^* \mid w^* \text{ es solución básica óptima de } (D)\}$$

y por tanto cada w_i^* sería una cota superior para ese aumento. Siempre y cuando e_i (resp. $-e_i$) sea una dirección posible, es decir que el problema resultante de aumentar una cantidad positiva (resp. negativa) el término independiente b_i no sea imposible.

Por consiguiente, si hay más de una solución dual óptima, será posible definir dos precios sombra $p_i^+ = \nabla v(b_o; e_i)$ y $p_i^- = \nabla v(b_o; -e_i)$, y se tendrá que

$$\begin{aligned} (1) \quad p_i^+ &= \text{Min}\{w_i^* \mid w^* = (w_1^*, ..., w_m^*) \text{ es solución básica óptima de (D)}\}. \\ (2) \quad p_i^- &= \text{Max}\{w_i^* \mid w^* = (w_1^*, ..., w_m^*) \text{ es solución básica óptima de (D)}\}. \end{aligned}$$

Veámoslo con un ejemplo.

(P_1) **Max** $3x_1 + 4x_2$

Sujeto a

$$x_1 + 3x_2 \leq 15$$

$$2x_1 + x_2 \leq 10$$

$$2x_1 + 3x_2 \leq 18$$

$$x_1 + x_2 \leq 7$$

$$4x_1 + 5x_2 \leq 40$$

$$x_1, x_2 \geq 0$$

La solución óptima de (P_1) es

$$(3) \qquad x^* = (x_1^*, x_2^*, \ldots, x_7^*)^T = (3,4,0,0,0,0,8)^T, \qquad z^* = 25,$$

siendo x_3, x_4, x_5, x_6, x_7 las variables de holgura.

Y las soluciones básicas óptimas del dual son

$$(4) \qquad \begin{aligned} w^1 &= (0,0,1,1,0) \\ w^2 &= (\tfrac{1}{2},0,0,\tfrac{5}{2},0) \\ w^3 &= (1,1,0,0,0) \\ w^4 &= (0,\tfrac{1}{4},\tfrac{5}{4},0,0). \end{aligned}$$

Los precios sombra pueden ser calculados según (1), (2) obteniéndose:

$$p_1^+ = 0, p_2^+ = 0, p_3^+ = 0, p_4^+ = 0, p_5^+ = 0, \tag{5}$$

$$p_1^- = 1, p_2^- = 1, p_3^- = \frac{5}{4}, p_4^- = \frac{5}{2}, p_5^- = 0. \tag{6}$$

Los paquetes de Programación Lineal hallan ***una*** de estas soluciones y en general no suelen avisar de que ***existen otras***. En particular, si para resolver (P_1) utilizamos el paquete LINDO encontramos w^3 (véase el Apéndice 1).

Pese a que el listado no diferencia las variables básicas de las no básicas, se puede reconocer que la solución óptima de (P_1) es degenerada porque al leer el listado de los rangos en los que la base no cambia, al variar un término independiente, se observa, por ejemplo, que no es posible disminuir $b_3 = 18$.

En general, al resolver el problema (P') se encontrará una de las dos situaciones siguientes:

- Si la SPB óptima de (P') es no degenerada, entonces para cada $b_r \quad 1 \leq r \leq m$ existe un intervalo $\Lambda_r = [\underline{\lambda}_r, \overline{\lambda}_r]$ con $\underline{\lambda}_r < 0, \overline{\lambda}_r > 0$ de modo que si el r–ésimo término independiente es $b_r + \lambda_r$ con $\lambda_r \in \Lambda_r$, la base óptima continúa siéndolo. Además, la función valor óptimo es de la forma

$$z^*(\lambda_r) = a + p\lambda_r, \qquad \lambda_r \in \Lambda_r,$$

  por lo que $p_r^+ = p_r^- = p$.

- Si la SPB óptima de (P') es degenerada, alguno de los extremos del rango en que es posible variar b_r, sin que cambie la base, es 0. Esto es debido a que la función valor óptimo es de la forma

$$z^*(\lambda_r) = \begin{cases} a + p_r^-\lambda_r & \lambda_r \in [\lambda_{1r}, 0] \\ \\ a + p_r^+\lambda_r & \lambda_r \in [0, \lambda_{2r}] \end{cases}$$

donde $\lambda_{1r} \leq 0, \quad \lambda_{2r} > 0$.

Como se ha partido de un problema (P) de maximización con restricciones de «menor o igual», cuando b_r sobrepase un cierto valor, $p_r^+ = 0$ y $\lambda_{2r} = +\infty$. Por el contrario, si hacemos disminuir demasiado b_r, el problema pasa a ser imposible.

La discusión para un PL de minimización con restricciones de «mayor o igual» sería simétrica.

4. ANÁLISIS DE SENSIBILIDAD

El ejemplo anterior, por ser sencillo, permitía generar todas las soluciones duales óptimas. Sin embargo, no será posible utilizar este procedimiento para calcular los precios sombra de problemas más complejos.

En el caso general, para calcular p_i^+ y p_i^- proponemos la utilización del Simplex Paramétrico (véase Murty (1983)) que resuelve problemas en los que los términos independientes están en función de un parámetro λ, es decir

$$\text{Max}\{cx \mid Ax = b + \lambda b^*, \quad x \geq 0\}, \tag{7}$$

donde $b, b^* \in \mathbb{R}^m$ y $\lambda \in \mathbb{R}$.

Haciendo uso de este algoritmo se calculan p_i^+ y p_i^- considerando $b^* = e_i$ y $b^* = -e_i$ respectivamente. En el Apéndice 2 mostramos su utilidad para el ejemplo (P_1).

Visto que la información que proporcionan las variables duales es diferente, cabe preguntarse por la interpretación económica del análisis de sensibilidad en presencia de degeneración. Si la solución posible básica óptima x^* tiene más de una base óptima, ¿qué sentido económico tiene hallar los rangos para los términos independientes en los que la base actual sigue siendo óptima?

Supongamos que estamos interesados concretamente en la restricción r–ésima, es decir b_r, y sea $\tilde{B}$ el conjunto

$$\tilde{B} = \{B_j : B_j \text{es una base óptima asociada a} x^*, \quad 1 \leq j \leq k\}. \tag{8}$$

Para cada B_j podemos determinar

$$\Lambda_r^{(j)} = \{\lambda_r \in \mathbb{R} : x_i^{(j)}(\lambda_r) \geq 0, \quad 1 \leq i \leq n\} = [\underline{\lambda}_r^{(j)}, \overline{\lambda}_r^{(j)}], \tag{9}$$

donde $x_i^{(j)}(\lambda_r)$ representa la componente i–ésima de $x^{(j)}(\lambda_r)$ que es la solución óptima parametrizada considerando como base óptima B_j.

Para hacer el análisis de sensibilidad con respecto a b_r, Gal (1986) propone determinar la ***región crítica global*** $\Lambda_r = \bigcup_{j=1}^{k} \Lambda_r^{(j)}$ de manera que para todo $\lambda_r \in \Lambda_r$, al menos una base B_j permanezca óptima.

Esta propuesta es adecuada pero con ella se perderá información si no se especifica para cada base dual posible la solución óptima parametrizada. Nosotros añadimos

que debería determinarse el comportamiento de la solución en cada intervalo $\Lambda_r^{(j)}$, $1 \leq j \leq k$. Veámoslo en el siguiente ejemplo:

(P_2) **Max** $x_1 + x_2$

Sujeto a

$$x_1 + 3x_2 \leq 15$$

$$x_1 \leq 3$$

$$2x_1 + x_2 \leq 10$$

$$x_1, x_2 \geq 0$$

La solución óptima de (P_2) es $x^{*T} = (x_1^*, ..., x_5^*)^T = (3,4,0,0,0)^T$, donde x_3, x_4, x_5 son variables de holgura. Tiene tres bases asociadas, una que no es dual posible $B_1 = (a_2, a_1, a_3)$ y dos que son ***duales posibles***

$$B_2 = (a_4, a_1, a_2), \quad B_3 = (a_5, a_1, a_2). \tag{10}$$

- Si estamos interesados en la variación de b_1, es decir $b_1 + \lambda_1$, $\lambda_1 \in \mathbb{R}$, se obtienen los siguientes resultados:

Para la base B_2:

$$\begin{pmatrix} x_1^*(\lambda_1) \\ x_2^*(\lambda_1) \\ x_3^*(\lambda_1) \\ x_4^*(\lambda_1) \\ x_5^*(\lambda_1) \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \lambda_1 \begin{pmatrix} -1/5 \\ 2/5 \\ 0 \\ 1/5 \\ 0 \end{pmatrix}$$

Entonces $z^*(\lambda_1) = 7 + 1/5\lambda_1$, $p_1^+ = 1/5$ y el parámetro λ_1 debe estar en $\lambda_1 \in [0, 15]$.

Para la base B_3:

$$\begin{pmatrix} x_1^*(\lambda_1) \\ x_2^*(\lambda_1) \\ x_3^*(\lambda_1) \\ x_4^*(\lambda_1) \\ x_5^*(\lambda_1) \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \lambda_1 \begin{pmatrix} 0 \\ 1/3 \\ 0 \\ 0 \\ -1/3 \end{pmatrix}$$

Entonces $z^*(\lambda_1) = 7 + 1/3\lambda_1$, $p_1^- = 1/3$ y el parámetro λ_1 debe estar en $\lambda_1 \in [-12, 0]$.

La región crítica global propuesta por Gal sería $\lambda_1 \in [-12, 15]$. Sin embargo, es evidente que la solución óptima depende de λ_1 de diferente forma si $\lambda_1 \in [0, 15]$ o si $\lambda_1 \in [-12, 0]$.

- El análisis de la variación de los otros dos términos independientes sería análogo.

En la práctica, si utilizamos el paquete LINDO, podemos obtener los precios sombra y sus rangos de validez haciendo uso del mandato «PARA».

Por último, creemos necesario hacer una breve mención a la ***redundancia***, como causa de degeneración. La redundancia puede aparecer en cualquiera de las fases de modelización.

Por ejemplo, en el análisis de postoptimalidad, cambiar los términos independientes de un problema lineal puede convertir en redundante a una restricción que no lo era y viceversa. Sin embargo, estas restricciones redundantes, que pueden influir en los precios sombra, no deben omitirse puesto que podría dar lugar a interpretaciones erróneas.

5. CONCLUSIONES

Al llevar a cabo el análisis de sensibilidad de los términos independientes de un problema de programación lineal a partir del listado obtenido mediante un paquete informático, aparecen bajo la denominación de «dual prices» o «shadow prices» unas variables cuya interpretación varía según el carácter degenerado o no degenerado de la solución posible básica óptima obtenida. Por tanto, es imprescindible distinguir ante qué tipo de solución óptima nos hallamos.

En el caso no degenerado, los «dual prices» indican la conveniencia o no de adquirir una cantidad adicional de un recurso que se agota y a qué precio. Sin embargo, si la solución es degenerada, esta interpretación no es válida, y lo que proporcionan es una estimación arriesgada de los beneficios esperados al adquirir unidades extra del recurso. En este segundo caso aconsejamos que el análisis de sensibilidad se lleve a cabo según se expone en el epígrafe 4.

APÉNDICE 1: SOLUCIÓN DE P_1

Si resolvemos el problema P_1 con el paquete de programación lineal LINDO se obtiene lo siguiente:

```
MAX      3 X1 + 4 X2
SUBJECT TO
       2)    X1 + 3 X2 <=    15
       3)    2 X1 + X2 <=    10
       4)    2 X1 + 3 X2 <=    18
       5)    X1 + X2 <=    7
       6)    4 X1 + 5 X2 <=    40
END

LP OPTIMUM FOUND AT STEP       2

       OBJECTIVE FUNCTION VALUE

       1)     25.0000000

 VARIABLE        VALUE          REDUCED COST
       X1         3.000000          .000000
       X2         4.000000          .000000

      ROW    SLACK OR SURPLUS     DUAL PRICES
       2)          .000000         1.000000
       3)          .000000         1.000000
       4)          .000000          .000000
       5)          .000000          .000000
       6)         8.000000          .000000

NO. ITERATIONS=       2

RANGES IN WHICH THE BASIS IS UNCHANGED:

                        OBJ COEFFICIENT RANGES
VARIABLE        CURRENT        ALLOWABLE         ALLOWABLE
                  COEF         INCREASE          DECREASE
      X1       3.000000         5.000000          1.666667
      X2       4.000000         5.000000          2.500000

                        RIGHTHAND SIDE RANGES

     ROW         CURRENT        ALLOWABLE         ALLOWABLE
                   RHS          INCREASE          DECREASE
       2       15.000000          .000000         10.000000
       3       10.000000          .000000          5.000000
       4       18.000000         INFINITY           .000000
       5        7.000000         INFINITY           .000000
       6       40.000000         INFINITY          8.000000
```

APÉNDICE 2: USO DEL SIMPLEX PARAMÉTRICO

Partimos de una tabla óptima del problema original (P) (nótese que podemos haber obtenido cualquiera de las tablas óptimas asociadas a la solución óptima primal degenerada), por tanto los costes reducidos $z_j - c_j$ son no negativos:

Base	$\overline{b}$	$x_1 x_2 \ldots x_{n-1} x_n$
x_B	$B^{-1}b$	$Y = B^{-1}A$
	$c_B\overline{b}$	$c_B B^{-1}A - c$

(T_1)

donde hemos escrito los vectores columna

$$Y_j = \begin{pmatrix} y_{1j} \\ \ldots \\ y_{mj} \end{pmatrix} = B^{-1}a_j, \quad 1 \leq j \leq n$$

siendo a_j la columna j–ésima de A.

Si consideramos el vector de términos independientes parametrizado por $\lambda \in \mathbb{R}^+$, se tiene $b' = b + \lambda b^*$. Multiplicando por B^{-1} se obtendrá:

$$B^{-1}b' = B^{-1}b + \lambda B^{-1}b^*,$$

que denotaremos por

$$\overline{b}' = \overline{b} + \lambda \overline{b}^*.$$

La nueva tabla será la siguiente:

Base	$\overline{b}$	$\overline{b}^*$	$x_1 x_2 \ldots x_{n-1} x_n$
x_B	$B^{-1}b$	$B^{-1}b^*$	$Y = B^{-1}A$
	$c_B\overline{b}$	$c_B\overline{b}^*$	$c_B B^{-1}A - c$

(T_2)

T_2 será óptima mientras continúe siendo primal posible o equivalentemente mientras que $\overline{b}_i + \lambda \overline{b}_i^*$ sea no negativo para $1 \leq i \leq n$, y esto ocurre si y sólo si $\lambda \in [\underline{\lambda}, \overline{\lambda}]$, donde

$$\underline{\lambda} = \begin{cases} -\infty & \text{si} \quad \overline{b}_i^* \leq 0, \qquad 1 \leq i \leq m \\ \text{Max}\{-\overline{b}_i/\overline{b}_i^* : \quad \overline{b}_i^* > 0\} & \text{en otro caso} \end{cases}$$

$$\overline{\lambda} = \begin{cases} +\infty & \text{si} \quad \overline{b}_i^* \geq 0, \qquad 1 \leq i \leq m \\ \text{Min}\{-\overline{b}_i/\overline{b}_i^* : \quad \overline{b}_i^* < 0\} & \text{si} \exists i \mid \quad \overline{b}_i^* < 0 \end{cases}$$

Cuando $\lambda \notin [\underline{\lambda}, \overline{\lambda}]$ se actualizará la tabla mediante el dual del simplex.

En el ejemplo (P_1) una de las tablas óptimas del problema es

Base	$\overline{b}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_2	4	0	1	0	0	1	−2	0
x_4	0	0	0	0	1	1	−4	0
x_1	3	1	0	0	0	−1	3	0
x_3	0	0	0	1	0	−2	3	0
x_7	8	0	0	0	0	−1	−2	1
	25	0	0	0	0	1	1	0

(T_3)

Vamos a calcular por ejemplo p_3^-. Para ello debemos hacer $b^* = -e_3$, entonces se tiene

$$b'_3 = 18 - \lambda_3, \quad \lambda_3 > 0$$

La tabla sería la siguiente:

Base	$\overline{b}$	$\overline{b}^*$	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_2	4	−1	0	1	0	0	1	−2	0
x_4	0	−1	0	0	0	1	1	−4	0
x_1	3	1	1	0	0	0	−1	3	0
x_3	0	2	0	0	1	0	−2	3	0
x_7	8	1	0	0	0	0	−1	−2	1
	25	−1	0	0	0	0	1	1	0
x_2	4	−1/2	0	1	0	−1/2	1/2	0	0
x_6	0	1/4	0	0	0	−1/4	−1/4	1	0
x_1	3	1/4	1	0	0	3/4	−1/4	0	0
x_3	0	5/4	0	0	1	3/4	−5/4	0	0
x_7	8	3/2	0	0	0	−1/2	−3/2	0	1
	25	−5/4	0	0	0	1/4	5/4	0	0

(T_4)

De donde directamente se obtiene que

$$\begin{pmatrix} x_1^*(\lambda_3) \\ x_2^*(\lambda_3) \\ x_3^*(\lambda_3) \\ x_6^*(\lambda_3) \\ x_7^*(\lambda_3) \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 0 \\ 0 \\ 8 \end{pmatrix} + \lambda_3 \begin{pmatrix} 1/4 \\ -1/2 \\ 5/4 \\ 1/4 \\ 3/2 \end{pmatrix} \text{ es solución para } \lambda_3 \in [0,8].$$

La función objetivo en función de λ_3 será

$$z(\lambda_3) = 25 - 5/4\lambda_3,$$

y por tanto

$$p_3^- = 5/4,$$

que, desde luego, coincide con el dado en (6).

Operando de igual forma con los restantes e_i se obtendrían todos los precios sombra.

REFERENCIAS BIBLIOGRÁFICAS

[1] **Gal, T.** (1986). «Shadow Prices and Sensivity Analysis in Linear Programming under Degeneracy». *OR Spectrum,* **8**, 59–71.

[2] **Karwan, M.H., Lotfi, V., Telgen, J.** y **Zionts, S.** (1983). *Redundancy in Mathematical Programming.* Lecture Notes in Economics and Mathematical Systems. Springer–Verlag. Berlin.

[3] **Murty, K.G.** (1983). *Linear Programming.* John Wiley and sons. New York.

[4] **Rubin, D.S.** y **Wagner, M.H.** (1990). «Shadow Prices: Tips and Traps for Managers and Instructors». *Interfaces,* **20**, 150–157.

[5] **Shapiro, J.F.** (1979). *Mathematical Programming: Structures and Algorithms.* John Wiley and sons. New York.

[6] **Shapiro, R.D.** (1984). *Optimization Models for Planning and Allocation: Text and Cases in Mathematical Programming.* John Wiley and sons.

[7] **Schrage, L.** (1989). *Linear, Integer and Quadratic Programming with LINDO. User's Manual.* (3ª edición). The Scientific Press. San Francisco.

ENGLISH SUMMARY:

AN INTERPRETATION OF THE SHADOW PRICES UNDER DEGENERACY

Teresa León y Vicente Liern

1. INTRODUCTION

For practical purposes, knowing the behaviour of the solution of the linear programming program due to changes in the data is as important as obtaining the optimal solution itself.

The main goal of this paper is to remark the difference between the sensivity analysis of the right hand side coefficients and the interpretation of the dual variables when the optimal basic feasible solutions are degenerate and when they are not. Moreover, as the best known and direct analysis is for the nondegenerate case, one could be tempted to use it also for the degenerate case.

2. STATEMENT OF THE PROBLEM

Let (P) be the linear program **max** $\{CX : A_o X \leq b, X \geq 0_n\}$ where $C \in \mathbb{R}^n$, $A_o \in M_{m\times n}$, $b \in \mathbb{R}^m$ are parameters and $X \in \mathbb{R}^n$ is the decisions vector. Its dual problem can be stated as **min** $\{wb : wA_o \geq b, w \geq 0_m^T\}$, and we denote by (P') the standard form of (P).

Let $\overline{x} = (x_B, x_N)^T = (B^{-1}b, 0_N)^T$ be a basic feasible solution (BFS) of (P'), $\overline{x}$ is degenerate if less than m of its components are greater than zero. A basis, B, is dual feasible if $w = c_B B^{-1}A$ is a feasible solution of the dual problem.

Let x^* be the optimal BFS of (P'), then we have that

- If x^* is non degenerate, then it has only one dual feasible basis associated and therefore only one optimal dual solution w^*. Every w_i^* can be seen as the rate of increasing or decreasing of the objective function by unit of b_i. And from the point of view of the economic interpretation of the dual variables they are called shadow prices, dual prices or marginal values.

- If x^* is degenerate it could have more than one dual optimal solution associated. Then, how should we interpret the dual variables?

3. INTERPRETATION OF THE DUAL VARIABLES UNDER DEGENERACY

Let x^* be a degenerate optimal BFS of (P') and $w^* = (w_1^*, w_2^*, \ldots, w_m^*)$ an optimal solution of its dual, then w_i^* is an upper bound on the objective function value increase when b_i increases by one unit and a lower bound on the objective function decrease when b_i decreases by one unit.

Then, for a problem of maximization of benefits subject to capacity constraints if we want to compute the benefits increase when an extra unit of the i–th resource is obtained then we should find **min** $\{ w_i^* : w^*$ optimal solution of (D)$\}$.

Analogously, for the case that we have $b_i - 1$ units of the i–th resource (and the problem still remains feasible) we find **max** $\{ w_i^* : w^*$ optimal solution of (D)$\}$. Thus we must consider two «shadow prices» p_i^+ and p_i^-.

The sensivity analysis performed by most commercial packages returns for each of the constraints a dual solution and for each of the coefficients in the objective function and for each of the right hand side coefficients of the constraints a certain validity interval.

The point is that in general these packages do not include a message saying that more than one optimal dual solution exists when it is the case and we could make fallatious interpretations of their outputs.

However, we can recongnize this situation by a simple examination. If the optimal BFS of (P') is degenerate at least one of the interval limits is zero. This is so because the objective function value is of the following form:

$$z^*(\lambda_r) = \begin{cases} a + p_r^- \lambda_r & \lambda_r \in [\lambda_{1r}, 0] \\ \\ a + p_r^+ \lambda_r & \lambda_r \in [0, \lambda_{2r}] \end{cases}$$

4. SENSIVITY ANALYSIS

In order to compute p_i^+, p_i^- and also the validity ranges we can use the parametric simplex algorithm for programs in which the right hand side coefficients depend on a parameter: **max** $\{CX : A_o X = b + \lambda b^*, X \geq 0_n\}$ where $b, b' \in \mathbb{R}^m$, $\lambda \in \mathbb{R}$.

We obtain the required information by taking $b^* = e_i$ where $e_i = (0,..,1,..0)^T$ and $b^* = -e_i$ for every $i = 1,2,\ldots,m$.

In practice, if we use for instance the LINDO package to solve our linear programming problem we do it by the command «PARA».

5. CONCLUSIONS

If the optimal BFS of the problem **max** $\{CX : A_o X \leq b, X \geq 0_n\}$ is nondegenerate, the dual prices indicate the usefulness of purchasing an extra amount of a binding resource. However, under degeneracy this interpretation could be non valid and what «dual prices» would provide is a risky estimation of the expected benefits when extra resources are acquired.

Estadística Oficial:

Fiabilitat a l'Estadística Oficial (I)

QÜESTIIÓ; Vol 20, 2 pp. 241-257, 1996

ESTADÍSTICA, FIABILITAT I VERITAT

ÀLEX COSTA*

Institut d'Estadística de Catalunya

A la secció d'Estadística Oficial d'aquest número de ***Questiió*** *es presenten tres treballs agrupats sota un títol genèric: «Estadística i fiabilitat». Resulta destacable que si bé és cert que en els tres casos és natural parlar de fiabilitat, la diversitat de procediments implicats indica que aquest terme té un sentit diferent en cada cas. Enfront d'aquesta diversitat de procediments cal plantejar-se si realment aquests articles tenen alguna relació entre ells o no. Dit d'altra forma: si de debò el concepte de fiabilitat emprat en cada un d'ells té un significat equivalent.*

En aquest treball s'explica aquesta diversitat sobre la base de relacionar els conceptes de fiabilitat i de veritat. A partir d'aquí s'analitza el diferent sentit que, en el pensament filosòfic, ha tingut el concepte de veritat. La tesi del treball és que l'estadística d'inferència i descriptiva s'adapta al concepte de veritat com a correspondència, mentre que el treball de comptabilitat nacional comporta uns enunciats d'una veritat molt més propera al concepte de coherència. Aquesta és la base explicativa de la diversitat en els procediments dels tres treballs.

Statistics, reliability and truth

Keywords: Fiabilitat, veritat, correspondència, coherència, teoria semàntica, teoricitat, estadística, comptabilitat nacional.

*Agraeixo al professor Eduard Bonet la lectura i els comentaris a aquest treball. Els seus suggeriments han millorat la part introductòria d'aquestes línies. Qualsevol error o imprecisió és imputable únicament a l'autor.

– Article rebut el setembre de 1996.

– Acceptat l'octubre de 1996.

1. INTRODUCCIÓ

A la secció d'Estadística Oficial d'aquest número de Qüestiió es presenten tres treballs agrupats sota un títol genèric: «Estadística i fiabilitat». Malgrat que efectivament en els tres existeix una avaluació de la fiabilitat de la informació estadística, es fan servir enfocaments molt diferenciats.

El treball de M. Guillén i X. Martín té una òptica clàssica: es tracta de conèixer la fiabilitat d'una enquesta de disseny complex, l'enquesta de població activa (EPA). La fiabilitat de les estimacions queda avaluada en termes de la seva dispersió. Un segon article, el de M. Delgado i J.A. Sánchez, compara la distribució de freqüències d'una variable del cens de població de Catalunya de 1991 segons una precodificació i segons una codificació de literals feta a posteriori. En aquest cas, es vol determinar la fiabilitat de la precodificació a partir de la hipòtesi que és millor la informació literal codificada posteriorment. Finalment, J. Muñoz, E. Pons i J. Pons aborden el tema de les revisions en les estimacions de la comptabilitat nacional (CN), estudiant fins a quin punt aquestes revisions són importants i són aleatòries. Segons un article de «The Economist» [1], la fiabilitat del treball estadístic en l'àmbit de la comptabilitat nacional queda reflectida per la importància de les revisions dels resultats.

Davant d'aquesta diversitat de procediments, cal plantejar-se si realment aquests treballs tenen alguna relació entre ells o no. Dit d'altra forma, si de debò el concepte de fiabilitat emprat en cada un d'ells té un significat equivalent.

En primer lloc cal observar que aquests treballs responen a tres camps del treball estadístic clarament diferenciats: La inferència estadística més típica en el treball sobre l'EPA, la pura descripció segons dos procediments alternatius en l'estudi censal i, finalment, les revisions de les estimacions en el treball d'ajust comptable. Si bé és cert que en els tres casos resulta natural parlar de fiabilitat, resulta també evident que la diversitat de procediments implicats indica que aquest terme té un sentit diferent en cada cas.

Al meu entendre estem enfrontats a un tema d'un considerable interès metodològic. En aquest punt pot ser recomanable recordar què vol dir, en el llenguatge ordinari, la paraula «fiabilitat». Segons el diccionari[2], la fiabilitat és la mesura de la confiança que hom pot tenir en el funcionament correcte d'un sistema. Si en comptes d'un sistema es fa referència a unes dades estadístiques, el «funcionament correcte» s'hauria de substituir per la «correcció de la informació». Resulta aleshores natural identificar aquesta correcció de la informació amb la seva «veritat». És a dir, la fiabilitat és la

[1] «The Economist», 7 de setembre de 1991.

[2] Institut d'Estudis Catalans (1995) «Diccionari de la llengua catalana».

mesura de la confiança que hom pot tenir en la veracitat d'unes dades estadístiques. En aquest sentit fiabilitat pot considerar-se sinònim de versemblança.

El pas donat des de la fiabilitat a la veritat no sembla especialment discutible i, tanmateix, pot generar importants reticències. Si es fa una anàlisi sense prejudicis, resulta obvi que darrera el concepte de fiabilitat es troba el concepte de veritat. És immediat reconèixer un estret vincle entre l'estadística i la veritat (i la mentida, com recorda el conegut acudit sobre les mentides i l'estadística).

Ara bé, el concepte de veritat està amagat en bona part del pensament contemporani, per considerar-se la veritat relativa a les persones (relativisme subjectiu) o relativa a les cultures (relativisme històric o antropològic). Aquest relativisme genera un important malestar enfront d'un concepte de veritat més absolut, que en canvi no aflora en considerar termes més tècnics o professionals, com el de la fiabilitat. Malgrat tot, s'ha considerat ineludible rescatar el concepte de veritat, ja que ens obre les portes al pensament filosòfic, que és on es buscarà l'explicació a la diversitat d'enfocaments en els treballs estadístics referenciats.

Segons l'anterior transcripció, els tres treballs que es presenten tenen una preocupació comú: determinar fins a quin punt les informacions de l'EPA, del cens de població (a partir de la precodificació) i de la CN són veritat. Aquests treballs pretenen facilitar respostes a preguntes aparentment semblants, com per exemple les següents:

És veritat que la taxa d'atur de l'economia catalana l'any 1995 va ser del 19,8%? És veritat que a Catalunya l'any 1991 hi havien 264,5 mil persones treballant a la indústria tèxtil? És veritat que el 1994 l'increment real del PIB va ser del 3%?

Les preguntes són molt semblants, i resulta indiferent si es formulen en termes de veritat o en termes de grau de veritat o versemblança. Tanmateix, els procediments de determinació de les respostes són tan diferents que hom pot sospitar que el concepte de veritat emprat en cada cas és diferent. Aquesta diferent significació del concepte de veritat pot resultar, a primera vista, un fet difícil d'assumir.

Per poder afrontar aquesta qüestió, el tema de l'estadística i la veritat[3], ens podem abocar breument al problema de la veritat en el pensament filosòfic, per tal de decidir amb coneixement de causa si darrera la inferència estadística, la pura descripció o els ajustos de la CN s'amaguen veritats conceptualment diferents.

[3]Aquest és un tema clàssic, encara que tractat amb una òptica normalment diferent de la d'aquestes línies. Es pot citar el llibre de R.v. Mises (1948): «Probabilidad, Estadística y Verdad», centrat en el sentit de la probabilitat. Recentment s'ha publicat un recull de treballs breus de Rao, C.R. (1994) «Estadística y verdad», a l'entorn del paper de l'estadística en la investigació científica.

2. APUNT SOBRE LES CONCEPCIONS CLÀSSIQUES DE LA VERITAT

El concepte de veritat és un concepte molt atractiu i també de moltes connotacions filosòfiques. Què vol dir, des d'un punt de vista filosòfic, «veritat»? Segurament es pot omplir una biblioteca amb les respostes a aquest interrogant. De totes formes, per començar, es pot contestar sintèticament a aquesta pregunta seguint J. Ferrater Mora[4], diferenciant entre la veritat entesa com una entitat o veritat entesa com a predicat d'un enunciat.

La veritat entesa com a entitat té històricament dues grans significacions, en hebreu i en grec. En hebreu veritat és allò en què es pot confiar, la seguretat, la confiança, la fidelitat. Aquestes virtuts, per a l'antic hebreu, només les té Déu. És a dir: Déu és la veritat. En canvi, en l'idioma grec, en l'idioma humanista dels relats homèrics (on els homes són els factors del seu propi destí), la veritat és la realitat, la realitat immòbil, permanent. La veritat és oculta, no se'ns mostra directament, l'hem de desvelar: és la realitat enfront de l'aparença, del miratge dels sentits. En l'idioma del primer monoteïsme, la veritat és Déu. En l'idioma bressol del pensament filosòfic i científic, la veritat és la realitat, a la qual cal treure-li el vel[5].

A partir d'Aristòtil la veritat és un atribut d'un pensament o d'un enunciat. «Dir d'allò que és que no és, o d'allò que no és que és, és fals; mentre que dir d'allò que és que és, i d'allò que no és que no és, és veritat». Aquesta forma d'entendre la veritat, més propera a la veritat de la lògica, identifica la veritat no amb una entitat, sinó amb un predicat.

El plantejament aristotèlic dóna cobertura a diverses concepcions de la veritat, fins a cert punt oposades. En primer lloc, la teoria de la correspondència, la veritat és la correspondència entre pensament i realitat, idea que queda recollida en l'aforisme escolàstic «adaequatio rei et intellectus».

La teoria clàssica de la veritat com a correspondència, malgrat la seva càrrega de sentit comú, presenta alguna debilitat. Existeixen enunciats que són veritat amb independència de la realitat i, per tant, sense correspondència aparent amb els fets. Per exemple: «tres és igual a tres», o «tres més dos són cinc». Aquestes veritats es deriven purament d'un conveni entre signes. Aquest és un problema limitat, atès que es refereix a una tipologia força concreta d'enunciats. En canvi, els enunciats perceptuals («ara plou», «Joan és més alt que en Pere») semblen adaptar-se de forma molt natural al concepte clàssic de veritat com a correspondència. Entre els enunciats purament formals i els perceptuals, es troben enunciats derivats d'altres, compostos, que deriven la seva veritat dels enunciats més bàsics, perceptuals. La veritat d'aquests

[4]Josep Ferrater Mora (1980) «Diccionario de Filosofía», IV.

[5]Aquest plantejament sobre el sentit de la veritat en l'antic hebreu i grec és de H.F. Soden (1927).

enunciats compostos sembla no atentar contra el concepte de correspondència, si bé el seu contacte amb la realitat no és tan directe com en els enunciats perceptuals.

L'época moderna, tanmateix, ha trencat amb aquesta concepció clàssica de la veritat, a favor d'una idea de veritat com a sistema coherent de coneixements. L'expressió més forta d'aquest trencament es troba a l'idealisme de Hegel. La veritat no es troba mai en l'expressió d'un fet aïllat, la veritat no és parcial, és una totalitat indivisible que només es pot predicar del conjunt del coneixement científic o del sistema filosòfic. Aquesta concepció pot semblar forçada i metafísica, però és el resultat de la reflexió que genera la maduració de la física en mans de Newton. Enfront de l'empirisme radical de Hume, segons el qual totes les nostres idees, base del coneixement científic, són còpia de les impressions sensibles, Kant afirma que les sensacions no signifiquen res sense l'estructuració en l'espai i el temps i les categories de l'enteniment. L'èmfasi fet per Kant en l'organització sistemàtica de l'experiència fa que quedi limitada la importància de les lleis individuals empíriques i l'inductivisme, i que resulti fonamental el caràcter hipotètic-deductiu del coneixement[6].

El posicionament de Kant respecte de la base empírica de la ciència té un efecte directe sobre la concepció de la veritat. En el paradigma de la veritat com «adaequatio rei et intellectus» sembla que la «cosa» perd autonomia enfront de la ment, atès que no té significació si no és acompanyada per les categories de l'enteniment. Aleshores la veritat és conformitat entre el coneixement i les categories de l'enteniment. La relació directa i senzilla entre un enunciat i la realitat queda, a partir de Kant, qüestionada. Sobre aquesta base Hegel es refereix a la veritat com quelcom absolut, no particionable en petits fets atòmics, i que només es pot atribuir al conjunt del coneixement científic o filosòfic.

Aquests dos enfocaments bàsics sobre la veritat, la teoria clàssica de la correspondència i la teoria de la coherència idealista, troben en la filosofia contemporània versions més elaborades, i també teories que poden situar-se a mig camí entre ambdues.

La versió moderna de la concepció de la correspondència va ser formulada per Russell, en el marc de l'atomisme lògic. Les teories idealistes sobre la veritat en termes de coherència tenen a Bradley com a continuador més destacat. Amb un posicionament intermedi es pot identificar el corrent de pensament nord-americà pragmatista, amb Peirce, James i Dewey. Segons aquesta darrera òptica, la veritat és el final de la investigació científica, que és la que genera el coneixement útil, que produeix tecnologia, que funciona (it works)[7]. L'aportació més innovadora, que ha estat

[6]John Losee (1979) «Introducción histórica a la filosofía de la ciencia».

[7]Les obres de referència són Russell (1918) «The philosophy of logical atomism», Bradley (1914) «Essays on Truth and Reality», Peirce (1877) «The fixation of belief», James (1907) «Pragmatism» i Dewey (1929) «Experience and Nature».

considerada com la «teoria científica de la veritat», és la teoria semàntica d'Alfred Tarski.

3. LA TEORIA SEMÀNTICA DE LA VERITAT

La teoria de Tarski[8] es considera la més evolucionada tècnicament de les teories de la veritat contemporànies, i aplica les eines de la moderna lògica formal per a la seva formulació. Una bona part de les anàlisis de la segona meitat del segle sobre el concepte de veritat, ha tingut com a punt de referència la teoria semàntica de Tarski.

La teoria semàntica originalment té una presentació molt formalitzada, i la seva significació substantiva ha estat objecte de controvèrsia. Encara que alguns autors han considerat que aquesta teoria és poc rellevant (Black), la majoria dels filòsofs, com Popper, la valoren com una aportació determinant sobre el tema de la veritat. Tampoc la seva presentació no formal és trivial. Es poden trobar exposicions a Quine (1977) i Haack (1982).

Un primer aspecte que s'ha d'aclarir és que en cap moment ens importa el criteri per determinar la veritat o falsedat d'un enunciat, sinó que del que es tracta és de conèixer el sentit que té dir que l'enunciat és veritat. Es busca una definició del sentit de veritat, no els criteris de determinació de la veritat o verificació de l'enunciat.

Tarski planteja, en primer lloc, la possibilitat de definir la veritat en el llenguatge ordinari i arriba a la conclusió que és impossible fer-ho sense caure en paradoxes (per exemple: «aquest enunciat és fals»). La raó és la universalitat d'aquest llenguatge, que inclou tant els enunciats com les formes de referir-se a aquests, i també predicats semàntics com «veritat» o «falsedat». Per evitar aquest problema cal definir la veritat a partir de la definició de dos llenguatges: el llenguatge objecte O i el meta-llenguatge M. D'aquesta forma es podrà definir en el llenguatge M la veritat dels enunciats del llenguatge O. Un esquema de Tarski que exemplifica aquest doble nivell de llenguatge i que ha estat molt referenciat és:

«la neu és blanca» és veritat si i només si la neu és blanca.

A més del doble nivell de llenguatge, existeix un segon element bàsic en la definició semàntica de la veritat: el concepte de satisfacció, vinculat als enunciats oberts del llenguatge objecte O. Per exemple, la seqüència <Cèsar, Gàl·lia> satisfà l'enunciat

[8]Les presentacions originals són Tarski (1931) «The concept of truth in formalised languages» i Tarski (1944) «The semantic conception of truth». Les presentacions més intuitives es troben a Haack, S. (1982) «Filosofía de las lógicas», i també a Quine, W.V. (1977) «Filosofía de la lógica».

obert d'O «x va conquerir y». Els membres d'aquesta seqüència són objectes segons el llenguatge O, corresponent a hipòtesis ontològiques d'aquest llenguatge. Aquests objectes no necessàriament formen part del llenguatge: són valors en el domini de les seves variables (x,y). La relació entre un enunciat d'O i la seqüència d'objectes no es pot expressar en O, cal fer servir el meta-llenguatge M, mitjançant l'esquema (a):

<Cèsar, Gàl·lia> satisfà «x va conquerir y» si i només si Cèsar va conquerir Gàl·lia.

La veritat semàntica queda definida com el predicat d'M en relació a un enunciat d'O, que estableix la satisfacció de l'enunciat enfront de qualsevol seqüència d'objectes. Per exemple, l'enunciat «$x = x$» satisfà qualsevol seqüència.

A més d'aquestes veritats formals, es pot arribar a veritats empíriques per la via de tancar enunciats oberts amb quantificadors del tipus «Existeix un x tal que...», (s'escriu «(Ex)»). D'aquesta forma, l'enunciat obert «x va conquerir y» es pot tancar i s'arriba a l'enunciat «(Ex)(Ey) (x va conquerir y)». Aquest enunciat és tancat perquè la quantificació fa que no tingui lloc per situar el valor de les variables. Això fa que o sigui satisfet per totes les seqüències o per cap. Si existeix almenys un parell d'objectes $< X, Y >$ que satisfà l'enunciat obert, aleshores el tancat és veritat. En cas contrari és fals.

És important observar que l'enunciat «(Ex)(Ey) (x va conquerir y)» no recull el sentit de l'esquema (a), atès que només requereix l'existència d'un parell qualsevol que pugui satisfer l'enunciat obert (per exemple, es fa veritat amb <Jaume I, València>). Per expressar la veritat que correspon a l'esquema (a), s'ha d'identificar el parell <Cèsar, Gàl·lia>. Això ens porta a l'esquema (b) següent:

«(Ex)(Ey) (x va conquerir y) & (Ex) (x= Cèsar) & (Ey) (y= Gàl·lia)»
és veritat si i només si Cèsar va conquerir Gàl·lia.

Ha estat imprescindible ampliar l'enunciat, per tal d'identificar el valor concret de les variables i reproduir el sentit de l'enunciat original. Aquesta identificació comporta passar d'un a tres enunciats. Tanmateix s'ha de dir que aquest és el cas més senzill.

A l'esquema (b) apareixen enunciats del tipus «(Ex)($x = X$)», on X és un valor de la variable x. Aquesta mena d'enunciats són inacceptables en certs llenguatges objecte, per exemple, en el càlcul de predicats de primer ordre, que és un dels casos més significats de la definició de Tarski. A més, fins i tot si s'admet el predicat «=X», aquest pot ser un predicat compost, especialment en llenguatges teòrics. Per exemple, en física el predicat «x és un electró» és clarament derivat, de forma que ens referim a una partícula amb una determinada massa, càrrega elèctrica, spin, etc. De la mateixa forma, tornant al llenguatge ordinari, si no acceptem les fórmules «x=Cèsar»

i «y=Gàl·lia» o els predicats implicats es consideren compostos, aleshores hem de definir uns descriptors de Cèsar: DC1, DC2, ... DCm i uns descriptors de Gàl·lia: DG1, ...DGn[9]. D'aquesta forma s'arriba a l'esquema (c):

«(Ex)(Ey) (x va conquerir y) & (Ex)(DC1x) & ... & (Ex)(DCmx) & (Ey)(DG1y) & ... & (Ey)(DGny)» és veritat si i només si Cèsar va conquerir Gàl·lia.

L'expressió de la veritat de l'enunciat original «Cèsar va conquerir Gàl·lia» ens porta a considerar $1+m+n$ enunciats atòmics. El desplegament de l'enunciat en $1+m+n$ sembla forçat en enunciats molt perceptuals del llenguatge natural. Per exemple, en «la neu és blanca» o «Joan és més alt que en Pere». No és tant artificiós en enunciats sobre fets històrics. Finalment, la traducció d'enunciats del tipus «(Ex)($x=X$)» a enunciats més bàsics s'adapta de forma natural a hipòtesis existencials teòriques, com ara les referides a l'electró o qualsevol altra entitat d'una teoria científica.

En resum, el mètode de Tarski de definició de la veritat porta a uns esquemes de dos llenguatges, el llenguatge objecte O i el meta-llenguatge M, en els que la veritat queda expressada en termes de satisfacció de totes les seqüències. En l'operació de tancar els enunciats oberts, cal fer servir uns quantificadors existencials de tal forma que l'enunciat en estudi, malgrat ser un enunciat aparentment senzill, pot desdoblar-se en una llarga cadena d'enunciats, aflorant una també llarga cadena de predicats.

La teoria semàntica ha estat molt estudiada. El propi Tarski la considera un desenvolupament de la concepció d'Aristòtil, però creu que no es pot alinear amb una de les concepcions filosòfiques tradicionals de la veritat, la correspondència o la coherència. La neutralitat de la teoria semàntica respecte de les posicions tradicionals en el problema de la veritat s'ha d'entendre en termes que no atorga tota la raó a només una de les posicions clàssiques. En realitat, la teoria semàntica dóna llum al concepte de veritat, mostrant el sentit i la compatibilitat de les dues concepcions clàssiques sobre la veritat.

En primer lloc i de forma principal, la concepció semàntica recull la idea de correspondència, en el sentit de satisfacció. Per aquesta raó Popper va acollir l'aportació de Tarski com l'aclariment definitiu de la teoria de la correspondència[10]. Aquesta lectura de Popper està fonamentada, però és parcial. Certament, la satisfacció pot ser interpretada com una versió moderna de l'adeqüació dels escolàstics, però l'adeqüació semàntica no és ja un vincle entre una realitat transparent i externa enfront d'un llenguatge universal. La teoria semàntica és molt més prudent. La correspondència es

[9]S'aplica la teoria de les descripcions de Russell (1905) «On denoting».
[10]Popper (1960) «On the sources of knowledge and ignorance».

dóna entre l'enunciat d'un llenguatge i seqüències que són domini de valors de variables del mateix. Aquestes seqüències en principi no formen part del llenguatge objecte en sentit estricte. Són allò que diu el llenguatge que existeix, les seves hipòtesis ontològiques. En aquest sentit, si bé és cert que la teoria semàntica reforça la concepció de la correspondència per l'ús que fa del concepte de satisfacció, queda en canvi alineada amb la crítica idealista (kantiana) respecte del paper de les percepcions en la construcció del coneixement. La realitat no és un conjunt de percepcions netes, deslligades del llenguatge. La realitat es presenta com un conjunt de valors de les variables que defineix el llenguatge. En un conegut aforisme de Quine «ésser és ésser el valor d'una variable».

La teoria semàntica no només il·lumina els conceptes de realitat i adeqüació. També resulta rellevant respecte de la relació entre veritat i llenguatge. La veritat és relativa a un llenguatge en el seu conjunt i la seva definició depèn de la sintaxi d'aquest llenguatge (es defineix veritat en O). L'esquema (c) permet entendre posicionaments coherentistes no només en l'aspecte de pèrdua de contacte amb una realitat transparent i independent del llenguatge. També mostra que la veritat d'un enunciat, per senzill que pugui ser en apariència, pot dependre de molts enunciats concatenats. La veritat de l'enunciat original depèn de la veritat d'altres enunciats. Aquesta situació porta al concepte de veritat com a coherència. En primera instància, perquè la simple descomposició de l'enunciat en molts altres confereix a la seva veritat un sentit de coherència, de no contradicció, entre els enunciats i entre els predicats que afloren. En aquest nivell, el sentit de la coherència no és molt fort, i es pot argumentar que és reduïble a la veritat com a correspondència d'un nombre més o menys gran d'enunciats. En un segon nivell, però, la coherència cobra una força singular.

En efecte, en determinats casos els predicats compostos no porten a uns predicats bàsics, sinó que presenten definicions circulars. Es a dir, els predicats compostos porten a d'altres predicats que no es poden considerar bàsics. Un cas típic és el dels conceptes teòrics en el llenguatge científic. Per exemple, els conceptes de massa i força en la mecànica clàssica. La seva definició és simultània, i encara que el concepte de força sembla més complex o abstracte que el de massa, aquest darrer no és definible sense comptar amb el concepte de força. El problema detectat és, de fet, molt conegut: el problema de la circularitat en la definició dels conceptes teòrics, la dificultat de definir-los sobre una base empírica clara i indiscutible (per tant, la dificultat en la fixació de la seva correspondència amb la realitat). Resulta destacable, de totes formes, que el plantejament de Tarski no només no resol aquest problema, sinó que l'amplifica.

L'esquema (c) amplia l'àmbit d'aquest problema de circularitat en la definició dels conceptes, de forma que reforça la concepció de la veritat com a coherència. El desdoblament recollit en l'esquema indica que qualsevol enunciat, si implica variables referides a entitats caracteritzades per predicats no bàsics, condueix a uns conceptes

compostos que poden tenir definicions creuades. Per tant, apareix amb intensitat el sentit de la veritat com a coherència, una coherència no reduïble, però tampoc incompatible, amb la correspondència.

En definitiva, el desenvolupament de Tarski justifica, explica, tant la concepció de la correspondència amb la realitat (hipotètica) com la coherència, en un sentit dèbil o més fort.

Per acabar, la teoria semàntica també permet determinar en quin tipus d'enunciats i llenguatges objecte serà preponderant cadascuna d'aquestes òptiques. Si el llenguatge objecte O admet com a enunciats bàsics expressions del tipus «$(Ex)(x = X)$» és a dir, termes singulars, aleshores es produeix el petit desdoblament que s'ha vist en l'esquema (b), i la correspondència resultarà determinant. L'acceptació de termes singulars és típica del llenguatge ordinari. En canvi, si el llenguatge O és més restrictiu, resultarà inevitable el desdoblament mostrat a l'esquema (c) i, en conseqüència, s'ampliarà tant l'àmbit de la coherència en sentit dèbil com la possible aparició de definicions circulars de predicats, és a dir, la coherència en sentit fort. Aquesta darrera és la situació característica dels llenguatges teòrics. Per aquesta raó, històricament, en estudiar enunciats del llenguatge ordinari, ha estat molt convincent la teoria de la correspondència i, en canvi, en estudiar els enunciats de la ciència ha cobrat força el sentit de la veritat com a coherència.

4. VERITAT I ESTADÍSTICA: ENTRE LA CORRESPONDÈNCIA I LA COHERÈNCIA

Una vegada feta aquesta excursió pel concepte filosòfic de veritat, podem tornar al punt de partida. Els treballs que es publiquen en aquesta secció de la revista pretenen avaluar fins a quin punt les informacions de l'EPA, del cens de població (a partir de la precodificació) i de la CN són fiables, en darrer terme, fins a quin punt són veritat. Malgrat que les preguntes són molt semblants, els procediments de determinació de les respostes són tant diferents que vàrem sospitar que el concepte de veritat emprat en cada cas podia ser diferent.

El treball més clàssic, el de Guillén i Martín, fa referència a enunciats del tipus: «la taxa d'atur de l'economia catalana l'any 1995 va ser del 19,8%», formulat a partir d'una investigació per mostreig. Aquest enunciat és compost, i es pot traduïr a d'altres relatius al nombre de persones desocupades i actives residents al territori català, en mitjana al llarg de l'any 1995. Des d'un punt de vista lògic, aquests enunciats només impliquen tres predicats: resident, desocupat i actiu, i quantificacions existències referides a una variable que té com a valors les persones. Els predicats

són compostos però poden ser traslladats a respostes en un qüestionari. D'aquesta forma, «*x* és un aturat» quedarà traslladat a respostes en el qüestionari, per exemple: «*x* declara cercar feina activament». Aquests enunciats són bàsics des del punt de vista del llenguatge ordinari, en el mateix sentit que ho és «la neu és blanca». D'altra banda, el quantificador existencial es refereix a persones i, per tant, cau plenament en l'àmbit del llenguatge ordinari i requereix un desplegament de l'enunciat semblant al de l'esquema (b) del punt anterior.

En definitiva, des de la perspectiva del llenguatge ordinari, que és el meta-llenguatge habitual, la veritat de l'enunciat es correspon amb el sentit de la veritat com a correspondència: es tracta de comptabilitzar quantes persones satisfan els enunciats oberts que se'n deriven de l'enunciat original. La presència de la coherència és molt mínima, ja que ni la cadena d'enunciats bàsics generada per l'enunciat original és complexa ni apareixen conceptes teòrics amb definicions circulars.

En la mesura en que el valor de la taxa d'atur (19,8%) ha estat inferit a partir d'una mostra i que l'enunciat es refereix al conjunt de la població, el sentit de la correspondència resulta aquí d'una gran claretat: l'enunciat construït sobre la base de la inferència és veritat si i només si diu allò que s'obtindria d'una investigació censal, d'una exploració del conjunt de la població en estudi[11].

El treball de Delgado i Sánchez també respon de forma molt clara a la veritat com a correspondència. En aquest cas no es tracta d'inferència estadística, sinó de la comparació de dues estadístiques descriptives. El diagnòstic, pel que fa al concepte de veritat, és exactament el mateix que en el treball anterior. L'única diferència és que aquí sí que serà possible conèixer el valor poblacional per recompte de les dades censals literals i posterior codificació. Això no afecta el sentit de la veritat dels enunciats, sinó que afecta la verificació o contrastació de la veritat, qüestió diferent a la que aquí ens ha ocupat.

L'article de Muñoz, Pons i Pons ens mostra una tercera cara de l'estadística. No és ni inferència ni estadística descriptiva. Es tracta del camp que, en l'àmbit de la producció d'informació estadística, és més modern i molt puixant: la comptabilitat econòmica. En aquest camp de treball estadístic no es parla d'inferència, sinó d'ajust o síntesi d'informacions. Un enunciat típic d'aquest àmbit del treball estadístic és el següent: «el 1994 l'increment real del PIB de l'economia catalana va ser del 3%».

[11] L'anàlisi del sentit de la veritat en un enunciat derivat de la inferència estadística hauria de tractar específicament el concepte de probabilitat. Per aquesta raó, és evident que la presentació feta és parcial. Ara bé, cal dir que malgrat la rellevància de la probabilitat en les relacions entre estadística i veritat, aquesta no és la situació en abordar el tema de la veritat com a correspondència o coherència, que és el centre d'interès d'aquest treball. De totes formes val la pena apuntar que el concepte semàntic de veritat, com a límit de la satisfacció dels enunciats, obre una porta de forma molt natural a la definició semàntica de la probabilitat com a mesura de veritat, és a dir, casos no límits de veritat o falsedat.

La referència a l'economia catalana en el primer treball va ser transformada en una referència a persones residents en el territori de Catalunya i a uns enunciats bàsics sencills. Aquestes persones i els predicats que corresponen als enunciats bàsics formen part, de forma natural, del llenguatge ordinari que es fa servir com a meta-llenguatge.

Quina és la situació en el cas de la comptabilitat nacional? Els aspectes a considerar són tres. En primer lloc, si l'enunciat és compost, la seva descomposició en enunciats bàsics des del punt de vista de la pròpia comptabilitat. En un segon estadi, cal considerar les variables que apareixen en l'enunciat obert. En cas que aquest tipus d'enunciat sigui inacceptable o el predicat implicat sigui compost, aleshores s'haurà de desdoblar l'enunciat. Finalment, en desdoblar l'enunciat es poden presentar o no definicions circulars.

Pel que fa al primer punt, resulta evident que l'enunciat considerat sobre el PIB no és bàsic, sinó derivat d'altres més senzills. L'aritmètica de l'obtenció del PIB està definida en el sistema de comptabilitat nacional vigent en el conjunt dels països d'Europa, el SEC. Aquesta no és una representació científica en sentit habitual: és una representació convencional. Els predicats implicats en la determinació del PIB són totes aquelles operacions que formen part del seu càlcul. En el sistema de comptes actualment vigent aquestes operacions són més d'un centenar, integrades en una sèrie de comptes referits a diferents sectors institucionals: empreses, administració pública, famílies, etc. La complexitat d'aquesta derivació és força elevada, de manera que en aquesta primera fase ja es pot parlar com a mínim d'una important presència de coherència en sentit dèbil.

En segon lloc, les variables dels enunciats són les anomenades unitats institucionals: tota entitat econòmica que pugui tenir propietats, actius i passius i participar en operacions econòmiques. Són unitats, des d'una família a una gran societat anònima, passant per una administració pública o una fundació. Aquestes unitats queden caracteritzades per unes operacions que reflecteixen la seva activitat econòmica. En el llenguatge de la comptabilitat nacional, el predicat «*x* és una unitat institucional» és compost, i s'ha de transformar en una cadena d'enunciats amb uns predicats derivats de la comptabilitat de les unitats institucionals. Aquesta comptabilitat s'ajusta a la comptabilitat empresarial o pública establerta legalment. Aquesta situació ens porta a un desdoblament molt important dels enunciats, equivalent al mostrat a l'esquema (c) del punt anterior.

En molts casos, l'atribució de comptabilitat a les unitats institucionals és purament teòrica, especialment en el cas de les famílies. Això fa que els valors de la seva activitat no es puguin determinar més que agregadament i gràcies als equilibris comptables de l'economia total, que equiparen les macromagnituds d'oferta i de demanda, la producció i el consum. Aquesta situació comporta una circularitat de fet, de forma que la coherència entre els agregats substitueix la satisfacció dels enunciats indivi-

duals referits a moltes unitats institucionals. Es detecta aquí el sentit de coherència més fort, no reduïble a la satisfacció d'enunciats particulars.

L'anterior situació comporta que, des de la perspectiva del llenguatge ordinari, la veritat dels enunciats de comptabilitat nacional tingui una càrrega bàsicament de coherència respecte de les regles comptables i de la informació comptable i administrativa que es deriva de l'activitat de les unitats institucionals.

En síntesi, els enunciats de l'àmbit de la descripció o la inferència en l'estadística econòmica tenen un sentit de veritat equivalent als enunciats del llenguatge ordinari i s'adapten de forma natural al concepte de correspondència. En canvi, els enunciats de la comptabilitat nacional s'assemblen més als de les teories substantives: tenen unes variables amb uns predicats allunyats del llenguatge ordinari i presenten, en la seva determinació, la circularitat pròpia dels conceptes teòrics. En conseqüència, la veritat d'aquests enunciats resulta molt més propera a la veritat com a coherència.

L'estadística és un conjunt de mètodes i tècniques que tenen una aplicació quasi universal. Els mètodes estadístics, a més, són molt diversos i poden situar-se justament tant en l'àmbit de la matemàtica aplicada com en el de les ciències socials. Una mostra paradigmàtica d'aquesta diversitat és la comparació dels mètodes de la descripció i de la inferència estadística i els mètodes de la comptabilitat nacional.

La descripció i la inferència estadística s'adapten a qualsevol llenguatge enriquint-lo sense introduïr en el mateix nous conceptes. Per aquesta raó, en els enunciats sobre la realitat econòmica derivats de la descripció o de la inferència estadística, el sentit de la veritat és la correspondència o adeqüació amb la realitat més familiar, la del llenguatge ordinari. En canvi, la comptabilitat nacional és una representació convencional, amb una forta càrrega de conceptes sustentats per una realitat que no és la del llenguatge ordinari, sinó la d'altres representacions també convencionals, la comptabilitat de les unitats institucionals. Per aquesta raó, els enunciats de la comptabilitat poden ser veritat o falsos, però aquesta atribució no té una correspondència directa amb la realitat familiar del llenguatge ordinari, sinó amb la coherència entre uns enunciats que provenen de les regles de la pròpia comptabilitat nacional i de normatives de comptabilitat, empresarial o pública.

En aquestes línies s'ha intentat mostrar que la diversitat del sentit de veritat que contenen els tres articles que es presenten en aquest número de Qüestiió, no és una situació anormal o atípica de l'estadística. Al contrari, aquesta diversitat, que porta de la veritat com a correspondència a la veritat com a coherència, queda plenament recollida en les anàlisis que sobre la veritat es troben tant en el pensament filosòfic clàssic com en el concepte modern més evolucionat de veritat, el que es deriva de l'anàlisi semàntica dels enunciats.

ANNEX: PRESENTACIÓ DE LA TEORIA SEMÀNTICA DE LA VERITAT

Aquesta presentació és una mica més formal que la del text i en aquest sentit s'assembla més a l'original de Tarski. Se segueixen les exposicions de Haack i Quine.

La definició de Tarski de veritat té la següent seqüència:

(a) especificació del llenguatge objecte O, per al que es definirà la veritat,

(b) especificació del meta-llenguatge M,

(c) definició del predicat «satisfà en O»,

(d) definició del predicat «veritat en O» a partir del predicat «satisfà en O».

(a) Sintaxis d'O:

variables: $x1, x2, x3, \ldots$

predicats: F, G, ... (per a un nombre donat d'arguments)

connectives oracionals: $-$, &

quantificador (E...)

parèntesis (,)

Oració atòmica d'O: F$x1$, G$x1x2$, etc...

Definició de fórmules ben formades (fbf)

i) Tota oració atòmica és una fórmula ben formada (fbf)

ii) Si A és una fbf, $-$A és una fbf

iii) Si A i B són fbfs, (A&B) és una fbf

iv) Si A és una fbf, (Ex)A és una fbf

v) Res més és una fbf

(b) Sintaxi d'M: és el català més el llenguatge O

(c) Definició de satisfà:

X, Y són seqüències d'objectes

$X(i)$ és l'objecte i–èsim de la seqüència X

A, B són oracions d'O

❶ Per a oracions atòmiques:

per a predicats 1-posició

per a tot i,X: X satisfà «Fxi» si i només si $X(i)$ és F

per a predicats de 2-posicions

per a tot i,j,X: X satisfà «Gxi,xj» si i només si $X(i)$ i $X(j)$ es troben en la relació G

i així per a tot predicat,

❷ Per a altres fbfs:

per a tot X, A: X satisfà «$-$A» si i només si X no satisfà «A»

per a tot X,A,B: X satisfà «A&B» si i només si X satisfà A i X satisfà B

per a tot X,A,i: X satisfà «(Ex)A» si i només si existeix una seqüència Y que satisfà «A» i aquesta Y és tal que $X(j) = Y(j)$ per a tot j, sent j diferent d'i.

*(d) **Definició de veritat:***

Una oració tancada d'O és veritat si és satisfeta per a totes les seqüències.

De la mateixa forma, és falsa en el cas de no ser satisfeta per cap seqüència.

ENGLISH SUMMARY:

STATISTICS, RELIABILITY AND TRUTH

Àlex Costa

The Official Statistics section of this number of Qüestiió contains three studies under the generic title «Statistics and reliability». The first article, by Guillén and Martín, discusses the reliability of a survey with a complex design about the proportion of the population in work. A second article, by Delgado and Sánchez, compares the frequency distributions of one variable of Catalonia's 1991 population census, between a pre-codification and a codification of literal data made *a posteriori*. The aim was to determine the reliability of the pre-codification, starting from the hypothesis that literal information codified *a posteriori* is better. Finally, Muñoz, E. Pons and J. Pons tackle the question of reviewing the calculations involved in national accounting. They study the extent to which these reviews are useful or random.

If it is true that it is natural to talk of reliability in these three cases, it is also clear that the diversity of procedures involved suggests that this term has a different meaning in each case. Although there is, in all three papers, an evaluation of the reliability of statistical information, the concept is focused in very different ways. This diversity of procedures should pose the question whether these studies really have anything in common or not. To put it another way, whether the concepts of reliability used in each one of the studies have equivalent meaning.

In this study, diversity is explained by associating the concepts of reliability and truth. The different meanings of the concept of truth in philosophical thought are then analysed. Thanks to a brief historical reference, two distinct concepts are identified: truth as correspondence, which is summed up in the scholastic aphorism «adaequatio rei et intellectus», and truth as coherence, with Hegel's idealism as its strongest expression, by which truth can only be predicated in function of an entire system of philosophical or scientific knowledge.

These two concepts of truth are explained and made compatible by Tarski's semantic analysis of the terms, which has paved the way for what is known as the semantic theory of truth. This theory begins from the concept of satisfaction in order to support the meaning of truth as correspondence. However, it also shows the meaning of truth as coherence, by breaking down the terms and theorising concepts with crossed definitions.

The terms closest to ordinary language display the correspondence meaning of truth; whereas more theoretical terms are true in a sense closer to the concept of coherence.

The statistical studies in this issue fit this diversity in the meaning of truth. In the study on statistical inference and in the descriptive study, the type of term, with substantive concepts characteristic of ordinary language, suggests truth as correspondence. However, national accounting contains concepts which are much more theoretical, in that terms used for discussing G.D.P., for example, are compound, divide into many other terms and, moreover, cannot be calculated without the prior assumption that the national accounting itself is correct. It is for this reason that truth in national accounting terms depends strongly on coherence.

QÜESTIIÓ, Vol 20, 2 pp. 259-272, 1996

ESTIMACIÓ DE LA VARIÀNCIA MOSTRAL A L'ENQUESTA DE POBLACIÓ ACTIVA

MONTSERRAT GUILLÉN

XAVIER MARTÍN

Universitat de Barcelona i
Institut d'Estadística de Catalunya

L'Enquesta de Població Activa constitueix una de les principals fonts d'informació sobre el mercat de treball. Aquest treball tracta de com incorporar el disseny mostral complex de l'Enquesta en l'estimació dels errors mostrals de resultats referits a Catalunya. En primer lloc, es descriu el mètode d'estimació emprat per l'Instituto Nacional de Estadística. Seguidament, s'introdueixen els dos enfocaments fonamentals per a l'estimació dels errors en la submostra de Catalunya. Per acabar, es comenten alguns resultats sobre el nombre total d'ocupats.

Estimation of sampling variance of the Spanish Labour Force Survey

Keywords: Mostreig de disseny complex, Mercat de treball, Estimació de la variància.

* Aquest article ha estat elaborat dins el projecte **Implementació de programes en l'entorn SAS de fiabilitat i millores en l'estudi metodològic.** Hem d'agrair la col·laboració de Luis Antonio del Barrio Martín, cap de l'Área de Estadística de la Actividad, el Empleo y el Paro, i de Florentina Alvarez Alvarez, Subdirectora de Estadísticas Demográficas, de l'INE.

– Article rebut el desembre de 1995.

– Acceptat l'abril de 1996.

1. INTRODUCCIÓ

La principal font d'informació sobre el mercat de treball d'un país són les seves estadístiques de la població activa. Així ho reconeix, per exemple, l'Organització Internacional del Treball (OIT) quan recomana incloure les estadístiques de la població actualment activa i dels seus components en els programes d'estadístiques contínues [1]. En el cas de la Unió Europea, a més a més, aquesta recomanació de l'OIT es converteix en una obligació: els països membres han de dur a terme un cop l'any una enquesta sobre les forces de treball seguint estrictament les directrius metodològiques establertes per l'Eurostat. Cal tenir present que una de les missions que compleix l'enquesta comunitària sobre les forces de treball és la de proporcionar indicadors comparables que permetin repartir entre els països membres els recursos dels fons estructurals (Fons Social Europeu, Fons Europeu de Desenvolupament Regional) [2,3].

A Espanya, l'Enquesta de Població Activa (EPA) és una investigació per mostreig adreçada a les famílies que realitza l'Instituto Nacional de Estadística (INE) al conjunt de l'Estat des de l'any 1964. En l'actualitat, l'Enquesta de Població Activa proporciona resultats trimestrals i és l'EPA corresponent al segon trimestre la que constitueix l'enquesta comunitària sobre les forces de treball. Entre aquests resultats s'inclouen estimacions de les principals magnituds per comunitats autònomes però amb un nivell de detall molt inferior al de les estimacions del conjunt d'Espanya.

L'Institut d'Estadística de Catalunya és l'organisme estadístic oficial de Catalunya. L'interès de disposar d'informació per a l'anàlisi del mercat de treball a Catalunya va obligar l'Institut d'Estadística de Catalunya a escollir entre dues possibilitats: la primera, portar a terme la seva pròpia enquesta sobre les forces de treball; la segona, aprofitar els resultats de l'Enquesta de Població Activa de l'INE. Es va optar per la segona possibilitat atès que, entre d'altres, són criteris de decisió de l'Institut, per un costat, que la metodologia de les operacions estadístiques permeti la comparació dels seus resultats amb altres estadístiques similars, i, per un altre, que no resulti una duplicitat amb altres estadístiques ja existents [4]. A més, el pes específic de Catalunya a la mostra de l'INE abonava la idea que era possible una tabulació més rica de l'Enquesta de Població Activa que la disponible fins a aquell moment [5]. Per emprendre la tasca d'explotació de les cintes de l'EPA, però, calia primer dotar-se d'un instrument de càlcul dels errors de mostreig de les estimacions. Aquest instrument era necessari per tal que el disseny de la tabulació estigués definit, a més de per criteris d'interès temàtic, també per consideracions sobre la fiabilitat de la informació. Fruit d'aquesta necessitat es va desenvolupar la investigació que aquí es presenta.

2. DISSENY MOSTRAL DE L'ENQUESTA

L'Instituto Nacional de Estadística utilitza en l'Enquesta de Població Activa un mostreig bietàpic amb estratificació de les unitats de primera etapa.

Les **seccions censals** (que constitueixen les unitats de primera etapa) s'estratifiquen primer segons un criteri geogràfic que té en compte la província i el tipus de municipi al què pertanyen. Després, les seccions censals de cada estrat s'agrupen per formar subestrats segons la seva categoria sòcio-econòmica.

Els **habitatges familiars** principals i els allotjaments fixos constitueixen les unitats de segona etapa. La selecció d'habitatges dins de cada secció es porta a terme per un procediment sistemàtic amb arrencada aleatòria. A cada habitatge es recull informació de totes les persones que hi tinguin la seva residència habitual.

La distribució de les seccions entre les províncies, els estrats i els subestrats es realitza de la forma següent:

❶ per a les províncies s'utilitza una afixació de compromís entre la proporcional i la uniforme (per a garantir una mínima representació de totes les províncies i per tal que el nombre de seccions sigui un múltiple de 12 —nombre de seccions que pot visitar un enquestador en un trimestre—);

❷ per als estrats d'una província s'utilitza una afixació proporcional a la població lleugerament corregida (s'incrementa el nombre de seccions de determinats estrats en els què s'espera que hi haurà major dispersió en les característiques estudiades);

❸ entre els subestrats d'un estrat s'utilitza una afixació proporcional a la població.

3. CÀLCUL DELS ERRORS DE MOSTREIG DE L'INE

L'Instituto Nacional de Estadística utilitza el mètode de les semimostres reiterades per a calcular els errors de mostreig de l'Enquesta de Població Activa [6]. Aquest mètode estima la variància d'un determinat resultat (indicador de l'error de mostreig) de la forma següent: calcula de forma repetida l'estimació que interessa utilitzant diferents parts (meitats) de la mostra per mesurar, després, la variabilitat d'aquestes estimacions.

L'INE fa 40 reiteracions utilitzant el següent procediment:

a) Es divideix la mostra original en dues semimostres complementàries fent parelles de seccions a dins de cada estrat.

b) S'assigna després la primera secció de cada parella a 20 reiteracions aleatòriament i la segona a les 20 restants. D'aquesta forma es garanteix que en totes les reiteracions hi participa la meitat de la mostra total i que cada secció apareix en la meitat de les reiteracions.

4. INFORMACIÓ DISPONIBLE A L'ENQUESTA

L'Institut d'Estadística de Catalunya rep regularment els fitxers trimestrals de l'Enquesta de Població Activa de Catalunya. Aquests fitxers, però, no contenen tota la informació necessària per a poder aplicar el procediment utilitzat per l'INE en l'estimació dels errors de mostreig[1].

En efecte, com s'ha vist en l'apartat anterior, el procediment de l'INE exigeix conèixer la secció i estrat a què pertany cada individu. L'estrat es pot deduir a partir de la província i el factor d'elevació associat a cada registre. La secció, malauradament, no és identificable de la informació continguda en el fitxer, per la qual cosa és impossible definir semimostres basades en parelles de seccions.

Per tal de poder construir un sistema similar a l'utilitzat per l'INE, es va establir provisionalment un criteri alternatiu que consisteix en realitzar parelles de famílies o habitatges.

5. ESTIMACIÓ DE TOTALS I DE LA SEVA VARIÀNCIA

Com a primera aproximació a l'estimació dels errors mostrals (variància i error estàndard) suposarem que l'EPA té un disseny de mostra aleatòria simple. Al següent apartat es presenta la manera de dur a terme aquesta aproximació i, després, la forma d'implementar un mètode de remostreig.

[1]En el moment de publicació d'aquest article, aquesta mancança s'haurà probablement solventat i, per tant, es podrà millorar el procediment que es descriu en els propers apartats.

5.1. Sota el supòsit de mostra aleatòria simple

En primer lloc, cal esmentar que la forma de mesurar la fiabilitat dels resultats obtinguts es farà mitjançant l'estimació de la variància del resultat desitjat.

El principal interès de l'explotació de l'Enquesta de Població Activa resideix en proporcionar l'estimació de xifres totals de persones en una determinada situació (el nombre d'ocupats, per exemple). També és habitual voler aquest total desagregat segons les diferents categories d'una variable classificatòria (per exemple, el nombre d'ocupats al sector agrari). Haurem de ser capaços, per tant, de calcular un interval de confiança o un marge que informi sobre la fiabilitat d'un estimador d'aquestes característiques.

En general, interessa acompanyar la taula d'estimacions de totals de persones en una determinada situació a Catalunya (per exemple, ocupades) classificades segons les categories de dues variables (sexe i sector, per exemple) d'una altra taula de les mateixes dimensions amb les estimacions de l'error associat a cada una de les cel·les de la taula original.

Cada individu de l'enquesta té associat un factor d'elevació que denotarem per w, i la variable estudiada que anomenarem X prendrà el valor 1 si l'individu té les condicions que es desitgen totalitzar o bé 0 si no és així.

En aquesta situació, l'estimació del total és:

$$\hat{X} = \sum_{i=1}^{n} x_i w_i.$$

El pes w_i és inversament proporcional a la probabilitat d'ésser inclòs a la mostra. L'elevació de cada individu indica el nombre d'individus de la població que representa.

Quan suposem que el mostreig realitzat és aleatori simple, l'expressió utilitzada per calcular la variància del total és:

$$\mathrm{Var}(\hat{X}) = \sum_{i=1}^{n} \frac{(x_i w_i n - \hat{X})^2}{n(n-1)}.$$

Substituint l'expressió de $\hat{X}$, la variància es pot escriure com:

$$\mathrm{Var}(\hat{X}) = \sum_{i=1}^{n} \frac{(x_i w_i n - \sum_{j=1}^{n} x_j w_j)^2}{n(n-1)},$$

que es pot corregir, finalment, per a tenir en compte la fracció mostrejada. En aquest cas, cal recordar que una bona aproximació per a conèixer el volum total de població

estudiada que intervé en el càlcul és:

$$\hat{N} = \sum_{i=1}^{n} w_1.$$

Per tant, la fracció de la mostra es pot aproximar per $\frac{n}{\hat{N}}$, i, finalment, l'expressió de la variància esdevé:

$$\widehat{\text{Var}(\hat{X})} = \left(1 - \frac{n}{\hat{N}}\right) \sum_{i=1}^{n} \frac{(x_i w_i n - \sum_{j=1}^{n} x_j w_j)^2}{n(n-1)}.$$

Com que el que es desitja estimar és l'error de mostreig d'un total d'individus que compleixen una condició determinada (per exemple, ser ocupats) la variable d'interès és de tipus dicotòmic (1 - ser ocupat, 0 - no ser-ho) i, llavors, l'expressió es pot reduir a:

$$(1) \quad \widehat{\text{Var}(\hat{X})} = \left(1 - \frac{n}{\hat{N}}\right) \frac{\sum_{i=1}^{n_1} (x_i w_i n - \sum_{j=1}^{n_1} x_j w_j)^2 + (n - n_1)\left(\sum_{i=1}^{n_1} x_i w_i\right)^2}{n(n-1)},$$

on n_1 és el nombre d'individus pels quals X pren el valor 1, i no s'utilitzen els individus pels quals X pren el valor 0.

Per obtenir una referència completa, es pot consultar [7].

5.2. Remostreig per a l'estimació de la variància

L'aproximació al càlcul de les variàncies basada en el supòsit de mostra aleatòria simple és incorrecte, donat que el disseny mostral de l'Enquesta de Població Activa s'allunya d'aquesta situació: es tracta d'un disseny complex. Existeixen dues components bàsiques que així ho determinen; en primer lloc, l'estratificació i, en segon lloc, la selecció de les diferents unitats mostrals en les diverses etapes del mostreig (seccions i famílies). L'entrevista a tots els membres d'una mateixa família és especialment rellevant, ja que cal esperar que les característiques sòcio-econòmiques dels membres seran iguals i que el comportament individual estarà relacionat, trencant així la independència entre les observacions. És conegut que els efectes d'aquests trets de la mostra són de caire oposat; si bé l'existència d'estrats adequats redueix la variància, l'existència de conglomerats l'augmenta, i és difícil *a priori* determinar la magnitud d'ambdues distorsions.

Wolter [8] presenta diferents mètodes que es poden emprar per a estimar la variància en situacions com la descrita. Els mètodes de remostreig són adequats quan el

nivell de complexitat del disseny (o dels estadístics d'interès) dificulta una aproximació analítica. Entre les diferents alternatives hi ha el mètode jackknife, el bootstrap i les semimostres equilibrades, a part de les expansions en sèrie de Taylor. Els tres primers mètodes es basen en obtenir estimacions a partir de repetir un mostreig de la mostra completa i en analitzar la variabilitat dels seus resultats. El procediment d'expansió en sèrie s'utilitza habitualment quan la principal dificultat es troba en l'estadístic que es vol calcular, el qual s'aproxima linealment, per a calcular la variància de la seva linealització.

En aquest treball ens proposem utilitzar un mètode de remostreig, que reprodueixi el disseny i permeti una aproximació de la variància. A tal efecte, i a la vista de la informació continguda en les dades, podem establir un mètode d'estimació de la variància per semimostres, semblant a l'aplicat per l'INE, on els parells no es formin a nivell de seccions sinó de famílies.

En general, per a poder aplicar el mètode d'estimació de la variància per semimostres, cal dividir la mostra original en dues parts complementàries. Per realitzar l'estimació de la variabilitat de l'estadístic es pren la meitat dels individus de la primera part i l'altra meitat de la segona (l'INE pren la meitat de les seccions de la primera part i l'altra meitat de la segona) i, multiplicant per dos el resultat obtingut, s'aconsegueix una estimació del total que es desitja. Si es reitera aquest procés un cert nombre de vegades en les quals els individus (seccions) se seleccionen aleatòriament, es poden aconseguir diverses estimacions del total i calcular la seva variabilitat.

L'estructura general del disseny mostral de la mostra completa ha de ser incorporada en la formació de les semimostres utilitzades en cada reiteració. Així, es garanteix que la semimostra tingui unes característiques similars a les de la mostra principal.

Per a poder formar semimostres en el cas del segment de dades per Catalunya i tenint en compte la impossibilitat de poder identificar les seccions, s'ha optat per construir les semimostres de la manera més fidel al disseny de la mostra original. Seguint les recomanacions de Wolter [8], i després d'analitzar alguns exemples reals de formació de les semimostres, hem resolt dividir les famílies en dos grups diferents dins de cada estrat (és a dir, fer parells de famílies). Per tant, actuem com si es tractés d'un mostreig estratificat de famílies i no de seccions. Com que no disposem de cap criteri de proximitat entre les famílies d'un mateix estrat, establim que els parells es formen de forma seqüencial en el fitxer d'observacions. Els dos grups de la mostra queden determinats quan diem que el primer grup el formen les famílies parells i que el segon grup les senars[2].

[2]S'ha previst provar un procediment aleatori de formació de parells de famílies, per tal d'evitar que l'ordre de l'arxiu provoqués biaix en els resultats.

Per a poder construir les reiteracions, en lloc d'utilitzar una assignació aleatòria dels grups dels estrats a cada reiteració, utilitzem les matrius Hadamard que asseguren que finalment a cada semimostra es té en compte un dels grups de l'estrat i que cada grup no apareix en més semimostres que el seu complementari (veure [8]). Les matrius Hadamard determinen quin dels dos grups de famílies de l'estrat intervé en la reiteració concreta. La seva utilització recomana prendre tantes reiteracions com el menor múltiple de quatre estrictament superior al nombre d'estrats. Com que el nombre d'estrats en les dades és 26, realitzarem 28 reiteracions.

A fi d'aplicar la metodologia d'estimació de la variància en la situació anterior s'ha elaborat una macro-funció en llenguatge SAS [9,10] suficientment general per ser aprofitada en explotacions similars.

6. RESULTATS

A continuació presentem els resultats obtinguts en l'explotació de les dades de l'Enquesta de Població Activa del primer trimestre de 1990. Es desitja obtenir una taula on aparegui el total estimat d'ocupats a Catalunya per sector d'activitat i sexe. Seguidament, cal donar una estimació del seu error.

Taula 1

Distribució de freqüències mostrals de la població ocupada per sectors. 1r trimestre de 1990.

Sector d'activitat	Freqüència	Percentatge
Agricultura	562	7.2
Energia i Aigua	99	1.3
Indústria bàsica	327	4.2
Trans. Metalls	703	9.0
Prod. Aliment.	341	4.4
Tèxtil, Cuiro	500	6.4
Fusta, Paper	476	6.1
Construcció	783	10.0
Comerç, Hosteleria	1617	20.7
Transports, Comunic.	459	5.9
Fin., Asseg. Llog.	462	5.9
A.P. Enseny. i San.	1478	18.9
Total	7807	100.0

Les Taules 1 i 2 mostren les distribucions de freqüències mostrals de la població ocupada per sectors i per sexe.

Taula 2

Distribució de freqüències mostrals de la població ocupada per sexes. Ir trimestre de 1990.

Sexe	Freqüència	Percentatge
Homes	5170	66.2
Dones	2637	33.8
Total	7807	100.0

6.1. Estimació del total i de la seva variància segons l'expressió de mostra aleatòria simple

A la Taula 3 es presenten els resultats de l'estimació de totals d'ocupats a Catalunya per sectors i sexe.

Taula 3

Estimació de total d'ocupats per sectors i sexe. Ir trimestre de 1990.

Sector d'activitat	Homes	Dones	Total
Agricultura	70832	19221	90053
Energia i Aigua	22394	2957	25351
Indústria bàsica	84430	23383	107813
Trans. Metalls	217623	32837	250460
Prod. Aliment.	67334	20501	87836
Tèxtil, Cuiro	79125	85959	165085
Fusta, Paper	116826	26163	142988
Construcció	180011	6791	186802
Comerç, Hosteleria	241506	175781	417288
Transports, Comunic.	122240	19337	141578
Fin., Asseg. Llog.	85958	435509	129468
A.P. Enseny. i San.	161420	243548	404969
Total	1449701	699989	2149690

Taula 4

Error estàndard relatiu del total d'ocupats. 1r trimestre de 1990.

Sector d'activitat	Homes	Dones	Total
Agricultura	5.52	10.09	4.83
Energia i Aigua	12.55	35.76	11.84
Indústria bàsica	7.02	13.78	6.24
Trans. Metalls	4.43	11.15	4.10
Prod. Aliment.	7.48	13.24	6.50
Tèxtil, Cuiro	7.47	6.81	5.01
Fusta, Paper	5.87	12.23	5.28
Construcció	4.42	21.55	4.33
Comerç, Hosteleria	3.90	4.45	2.88
Transports, Comunic.	5.75	14.18	5.32
Fin., Asseg. Llog.	6.64	9.48	5.42
A.P. Enseny. i San.	4.81	3.89	2.98
Total	1.45	2.19	1.09

L'error estàndard relatiu es calcula com $E.S.R. = 100\frac{S}{\hat{X}}$, on S és l'error estàndard estimat i $\hat{X}$ el total estimat.

Taula 5

Semiamplitud de l'intèrval de confiança (95%) del nombre d'ocupats. 1r trimestre de 1990.

Sector d'activitat	Homes	Dones	Total
Agricultura	7665	3801	8519
Energia i Aigua	5508	2073	5883
Indústria bàsica	11620	6316	13191
Trans. Metalls	18913	7177	20147
Prod. Aliment.	9872	5321	11187
Tèxtil, Cuiro	11587	11479	16213
Fusta, Paper	13450	6273	14793
Construcció	15604	2869	15848
Comerç, Hosteleria	18449	15339	23580
Transports, Comunic.	13783	5374	14756
Fin., Asseg. Llog.	11193	8082	13744
A.P. Enseny. i San.	15217	18590	23643
Total	41125	30058	46108

Taula 6

Coeficients de variació del total d'ocupats. 1r trimestre de 1990. (Mètode de semimostres.)

Sector d'activitat	Homes	Dones	Total
Agricultura	6.74	9.16	5.52
Energia i Aigua	12.34	18.48	9.91
Indústria bàsica	5.52	12.25	5.85
Trans. Metalls	6.21	17.47	4.70
Prod. Aliment.	8.82	12.11	8.66
Tèxtil, Cuiro	4.41	5.64	3.69
Fusta, Paper	4.12	7.13	3.27
Construcció	4.38	13.90	3.97
Comerç, Hosteleria	5.89	2.98	3.73
Transports, Comunic.	4.73	11.98	4.86
Fin., Asseg. Llog.	3.87	98.47	3.63
A.P. Enseny. i San.	5.39	3.50	3.96
Total	1.12	1.50	0.91

L'error estàndard relatiu es calcula com $E.S.R. = 100\frac{S}{\hat{X}}$, on S és l'error estàndard estimat i $\hat{X}$ el total estimat.

Taula 7

Semiamplitud l'intèrval de confiança (95%) del nombre d'ocupats. 1r trimestre de 1990. (Mètode de semimostres.)

Sector d'activitat	Homes	Dones	Total
Agricultura	9360	3451	9748
Energia i Aigua	5416	1071	4926
Indústria bàsica	9130	5614	12369
Trans. Metalls	26493	11245	23091
Prod. Aliment.	11640	4867	14905
Tèxtil, Cuiro	6834	9499	11935
Fusta, Paper	9431	3654	9164
Construcció	15462	1850	14526
Comerç, Hosteleria	27869	10255	30499
Transports, Comunic.	11322	4542	13474
Fin., Asseg. Llog.	6526	7220	9218
A.P. Enseny. i San.	17062	16709	31446
Total	31801	20631	38181

Quan es realitza el tractament com si la mostra fos aleatòria simple, es consideren tots els individus en el mateix estrat.

La Taula 4 mostra els errors estàndard relatius (segons la fórmula clàssica (1)) corresponents als totals estimats de la Taula 3.

Mitjançant la Taula 4 és senzill identificar aquelles caselles on l'error mostral és més gran en relació al nombre d'ocupats estimat. Alternativament es presenta la Taula 5, on es mostra la meitat de l'amplada de l'intèrval de confiança calculat al 95%.

Sumant i restant la semiamplitud de la Taula 5 al total d'ocupats, obtindríem els límits superior i inferior de l'intèrval de confiança.

6.2. Estimació de la variància segons el mètode de semimostres reiterades

Per brevetat, reproduirem aquí els resultats obtinguts corregint la fracció mostrejada[3], els corresponents errors estàndard relatius en percentatge i la semiamplitud de l'interval de confiança al 95%.

En general, els errors de mostreig són inferiors quan s'utilitza el mètode de les semimostres, però aquest efecte es deu essencialment a la inclusió del factor de correcció per fracció mostrejada (per més detalls veure [9]).

Amb l'aproximació de les variàncies per remostreig, que té en compte l'estratificació i també l'agrupació dels individus en famílies (encara que no en seccions), el corresponent efecte de correlació (donada la manca d'independència entre els individus d'una mateixa família) provoca estimacions dels errors superiors als derivats de l'expressió de mostreig estratificat. Malgrat que l'expressió de mostra aleatòria simple, que ignora la totalitat del disseny d'Enquesta de Població Activa, proporciona una estimació dels errors conservadora, en alguns casos no és prou acurada. Per exemple, en la casella d'agricultura-homes, el mètode de les semimostres, que incorpora bona part del disseny mostral, proporciona una estimació de l'error mostral superior.

Convé basar les conclusions de fiabilitat sobre els resultats del total d'ocupats en aquesta metodologia d'estimació d'errors per remostreig, donat que incorpora més informació sobre el procés d'obtenció de la mostra. Cal esperar que el procediment que s'ha presentat es realitzi més acuradament quan es pugui tenir en compte la informació sobre seccions censals i estratificació.

[3]Aquesta correcció per fracció mostrejada redueix lleugerament la variància estimada, i en conseqüència els errors estàndard i els errors estàndar relatius

REFERÈNCIES

[1] **Mehran, F. i Hussmanns, R.** (1990). *Surveys of economically active population, employment, unemployment and underemployment.* An ILO manual on concepts and methods, Oficina Internacional del Treball, Ginebra.

[2] **Eurostat** (1992). *Enquête sur les forces de travail. Méthodes et définitions*, Oficina Estadística de les Comunitats Europees, Luxemburg.

[3] **Comissió de les Comunitats Europees** (1988). *L'enquête sur les forces de travail comme instrument de la politique de l'emploi*, CECA-CEE-CEEA, Luxemburg.

[4] **Llei 30/1991, de 13 de desembre,** *Pla estadístic de Catalunya 1992-1995*, Diari Oficial de la Generalitat de Catalunya núm. 1539 de 10 de gener de 1992.

[5] **Institut d'Estadística de Catalunya** (diversos anys). *Mercat de treball. Ampliació de resultats anuals de l'Enquesta de Població Activa*, Barcelona.

[6] **INE** (1996). *Encuesta de Población Activa. Informe Técnico.* Area de Diseño de Muestras y Evaluación de Encuestas de Población y Hogares. Madrid.

[7] **Cochran, W.** (1977). *Sampling Techniques.* John Wiley.

[8] **Wolter, K.M.** (1986). *Introduction to variance estimation.* Springer-Verlag.

[9] **Martín, X. i Guillén, M.** (1992). *Estimación de totales y su error de muestreo en la Encuesta de Población Activa. Document de Treball.* Institut d'Estadística de Catalunya.

[10] **SAS Institute** (1990). *SAS/IML Software. Usage and Refererence Version 6. First edition.* SAS Institute. Cary, NC.

ENGLISH SUMMARY:

ESTIMATION OF SAMPLING VARIANCE OF THE SPANISH LABOUR FORCE SURVEY

Montserrat Guillén and Xavier Martín

The Enquesta de Població Activa is one of the main sources f or information about the labour market. This article discusses how to incorporate the complex sample design of this survey in the estimation of sampling errors for results refered to Catalonia.

The first section introduces the reader to the context of active population surveys. The role of the *Institut d'Estadística de Catalunya* in producing reliable statistical results within this framework is emphasized.

In the second section, a description of the sample design used to obtain the *Enquesta de Població Activa* is detailed. It is a two-step design with stratification. Furthermore, once a household is selected in the sample, all the individuals in the household are interviewed.

Section 3 explains how the *Instituto Nacional de Estadística* estimates the variance for the results obtained from the latter survey. A half-sample method is used.

Section 4 summarizes the type of information that is received in the files corresponding to the *Enquesta de Població Activa de Catalunya*, which do not allow to use the variance estimation method that is reproduced in section 3.

In section 5, the general theoretical framework for the estimation of totals and their variance is introduced. A resampling approach is also described, and the article explains how to apply the half-sampling method to the subsample which is restricted to Catalonia.

Finally, section 6 shows some of the results obtained using the methodology introduced in the previous section and some comments about the estimation of the total working population are presented. The discussion includes a comparison of the variance estimates under two different hypothesis; namely, the simple random sample approach and the half-samples method that takes into account the sample design.

A list of related references is also included at the end.

QÜESTIIÓ, Vol 20, 2 pp. 273-291, 1996

ESTUDI DELS CODIS OBTINGUTS PER CODIFICACIÓ AUTOMÀTICA ENFRONT DELS CODIS PRECODIFICATS. APLICACIÓ PER AL CPH/91 DE LA VARIABLE ACTIVITAT

MIQUEL DELGADO ALZAMORA*
JOSEP ANTON SÁNCHEZ CEPEDA*

Institut d'Estadística de Catalunya

Al cens de població i habitatge de 1991 es va donar la situació de que les preguntes d'ocupació i activitat es responien mitjançant una descripció —literal— i un codi precodificat. A la fase de tractament de la informació censal es realitza el procés de codificació automàtica d'aquests literals amb la qual cosa s'aconsegueix tenir per cadascuna d'aquestes preguntes dues respostes codificades.

Aquesta situació ens porta a l'estudi de la doble resposta amb la finalitat de constatar si els enquestats codifiquen bé les seves preguntes o pel contrari tenen confussions.

L'objectiu de l'estudi és avaluar la coincidència i estabilitat de les respostes obtingudes amb cada sistema i establir quins problemes es produeixen amb cada tipus de classificació, problemes derivats alguns de l'excessiva generalització a les categories del precodificat, de l'existència d'ambigüitat o d'absència del literal on el precodificat ens pot servir d'ajut per classificar-lo. Per fer-ho definirem uns paràmetres d'estudi, els valors dels quals ens determinaran l'agrupació i/o separació de codis de les variables.

Analysis of automatic coding values versus respondent's values. Application for Activity and Occupation.

Keywords: Cens de població i habitatge, pregunta codificada, codificació automàtica.

*Miquel Delgado Alzamora, Josep Anton Sánchez Cepeda. Institut d'Estadística de Catalunya. Departament d'Economia i Finances. Generalitat de Catalunya. Via Laietana, 58. 08003 Barcelona.

– Article rebut el novembre de 1995.

– Acceptat el juny de 1996.

1. INTRODUCCIÓ

En el Cens de Població i Habitatge de 1991 es va donar la situació de que les preguntes d'Ocupació i Activitat es podien respondre de dues formes, per una banda es posava una de les categories d'una llista de codis (resposta precodificada) i l'altra alternativa era que la persona responia mitjançant una descripció (literal obert).

L'Institut d'Estadística de Catalunya pren la decisió de gravar les dues respostes a cada variable i realitzar la codificació dels literals oberts, mitjançant el corresponent procés de codificació.

A la variable Activitat i a l'Ocupació es procedeix a la codificació dels literals oberts en un dels codis possibles de les classificacions oficials vigents en el moment de la codificació: La Classificació Nacional d'Activitats Econòmiques de 1974 (CNAE-74) i la Classificació Nacional d'Ocupacions (CNO-79), ambdues a un nivell de desagregació de tres dígits.

Aquesta situació ens porta a l'estudi de la doble resposta per a cada pregunta amb la finalitat de constatar si els enquestats codifiquen bé les seves pròpies respostes o pel contrari tenen confusions en alguns casos. Es parteix de la base de que la descripció donada per l'enquestat com a resposta és el valor més fiable; davant la necessitat de codificar-lo es pren com a vàlid el valor resultant de la codificació automàtica enfront del codificat per l'enquestat, per haver estat sotmesa la codificació automàtica a un procés exhaustiu de depuració de les variants.

Tots dos sistemes de classificació tenen avantatges i desavantatges. El precodificat aporta facilitat en el moment de respondre per part de la persona, rapidesa i un menor cost a l'hora de la gravació, però té l'inconvenient de la seva excessiva generalització al categoritzar les variables, amb la conseqüent pèrdua de detall. A més, la taxonomia escollida per classificar cada variable va ser poc afortunada, pel fet d'haver utilitzat unes classificacions a mig camí entre les velles i les noves (aquestes últimes no estaven encara vigents en el moment de l'operació). D'altra banda, les classificacions basades en el codificat aporten un major nivell de detall en la classificació al basar-se en el literal que la pròpia persona descriu, però hi ha problemes si un literal és molt ambigu no podent-se ubicar en una de les categories.

L'objectiu de l'estudi és avaluar la coincidència i l'estabilitat de les respostes obtingudes amb cada sistema i establir quins problemes es produeixen amb cada tipus de classificació, problemes derivats alguns de l'excessiva generalització a les categories del precodificat, de l'existència d'ambigüitat o d'absència del literal on el precodificat ens pot servir d'ajut per classificar-lo. Per fer-ho definirem uns paràmetres d'estudi, els valors dels quals ens determinaran l'agrupació i/o separació de codis de les variables.

2. PARÀMETRES D'ESTUDI

Considerem una variable que té n codis possibles com a resposta. Disposar de dues respostes alternatives per a cada pregunta genera una matriu de dimensions $n \times n$ on a les columnes hi posem els n codis possibles provinents de la precodificació i a les files els n codis provinents de la codificació automàtica. D'aquesta matriu calculem el tant per cent de coincidència entre la resposta segons el precodificat i el total de respostes segons la codificació automàtica. Considerem el següent exemple amb $n = 4$.

A. Matriu de codi absoluts

	Precodificats				
Codificació	$P1$	$P2$	$P3$	$P4$	total
$C1$	5000	1000	50	50	6100
$C2$	900	1000	50	50	2000
$C3$	250	200	1000	50	1500
$C4$	50	50	50	10000	10150

B. Matriu de percentatges

	Precodificats			
Codificació	$P1$	$P2$	$P3$	$P4$
$C1$	81.9	16.4	0.8	0.8
$C2$	45.0	50.0	2.5	2.5
$C3$	16.6	13.3	66.6	3.3
$C4$	0.5	0.5	0.5	98.5

La diagonal correspondrà, per a cada codi, amb el tant per cent de casos ben precodificats.

1. Paràmetre «diag»: Per a cada codi el tant per cent de casos que estan ben codificats. En el nostre exemple quedaria, per a cada codi:

codi i	*diag (i)*
$C1$	81.9
$C2$	50.0
$C3$	66.6
$C4$	98.5

És interessant, tanmateix, saber per a cadascun dels codis (els anomenarem receptors) com es distribueixen la resta de codis que havien estat prèviament precodificats (els anomenarem donants). Es pot donar la situació en què els casos erròniament precodificats, als que anomenarem residus, es distribueixin entre diferents codis o es concentrin en uns pocs. També mesurarem si els residus es distribueixen de forma homogènia o si alguns destaquen per sobre d'altres.

2. Paràmetre «acum»: Per a cada codi, el nombre de codis (inclòs aquest mateix) necessaris per explicar el 95% dels casos. Mesura si els casos precodificats erròniament per a cada codi es concentren en pocs o molts codis i, per tant, si el codi en qüestió es confon amb pocs o molts.

codi i	*acum (i)*
$C1$	2
$C2$	2
$C3$	3
$C4$	1

3. Distribució «MIJO»: Per a cada codi es calcula el pes del residu sobre la diagonal sempre que el pes sobrepassi un llindar de .05. Mesura els codis que recullen un pes important sobre el codi de la diagonal i la seva distribució.

codi i	*mijo1(i)*	*mijo2(i)*	*mijo3(i)*
$C1$	2	01	01
$C2$	9	05	05
$C3$	25	02	05
$C4$	005	005	005

4. Distribució «Donant»: Es composa dels codis donants de la distribució anterior.

codi i	*donant1(i)*	*donant2(i)*	*donant3(i)*
$C1$	2		
$C2$	1		
$C3$	1	2	
$C4$			

5. Paràmetre «nmijo»: De la distribució Donant, per a cada codi quants residus superen el llindar.

codi i	*acum (i)*
$C1$	1
$C2$	1
$C3$	2
$C4$	0

Categoritzarem el paràmetre «diag» perquè sigui més orientatiu. Les categories que hem realitzat són:

1. Molt atractora o receptora: > 95
2. Atractora: de .85 a .95
3. Bastant atractora: de .7 a .85
4. Poc atractora: de .5 a .7
5. Molt poc atractora: $< .5$

El mètode per a l'obtenció dels paràmetres està programat en SAS i produeix com a output:

- La matriu de precodificació-codificació automàtica en codis absoluts
- La matriu de precodificació-codificació automàtica en percentatges
- Per a cada codi, els percentatges i donants ordenats en ordre decreixent i el paràmetre «diag».
- Les distribucions Mijo i donant amb el paràmetre Nmijo.
- Els paràmetres Diag, pcdonant i acum sense parametritzar i parametritzats.

Amb la distribució donant es genera un graf per visualitzar millor els resultats.

3. ESTUDI DE LA VARIABLE ACTIVITAT

Apliquem el mètode a la codificació de l'«Activitat de l'Establiment» del CPH/91. El nombre de persones que responen que són ocupats, desocupats amb treball anterior o jubilats —que són els únics que han de respondre a la pregunta en qüestió— puja a 3.150.056. D'aquests no hi consta el literal per 165.762, no hi consta el precodificat per 241.378, —en 152.798 no hi consta cap dels dos—, no s'ha pogut codificar el literal per a 399.802 i s'han codificat 2.495.912. Aquests últims són els que prenem per al present estudi. (Els resultats es troben a l'annex 2.)

Els valors de la diagonal principal de la matriu de percentatges oscil·len entre el 95,87% del codi 01 al 42,35 del codi 07, essent la mitjana general del 74,40% amb desviació del 17,35.

Comencem observant l'output dels paràmetres acum, receptor i donant. Podem observar el següent:

- Hi ha 3 codis amb receptor = 5 i acum = 1, és a dir, codis que tenien més d'un 95% dels casos ben precodificats i a més no actuen com a donants d'altres codis. Això vol dir que no recullen casos precodificats erròniament que corresponen a altres codis. Aquests codis —01,06 i 23— no presentaran problemes de confusió.

- En el cas extrem es troben codis amb receptor = 1, és a dir, codis que tenien ben precodificats menys del 50% dels casos. Podem distingir, altrament, dos grups:

    1. Els codis 03 i 07 tenen acum 3 i 4 respectivament, significant que els codis erròniament precodificats provenen de pocs codis.
    2. El codi 15 té acum igual a 11 i, per tant, els seus casos erronis es troben dispersos entre molts codis.

- El paràmetre receptor presenta la següent distribució:

  Val 5 en 3 casos, 4 en 7, 3 en 9, 2 en 7 i 1 en 3; per tant els codis de la variable Activitat que estaven ben precodificats s'igualen amb els que no ho estaven.

- Destaquen com a codis molt dispersos el 19 i el 24 que tenen acum 16 i 14 respectivament, els quals resulten poc atractors.

Passem ara a les distribucions Mijo i donants en les quals poden observar:

- Destaquen els codis 03 i 07, que tenen un mijol superior a 1, el que significa que hi ha un altre codi que recull més casos precodificats que ell mateix. Això

és indicatiu de gran confusió entre parelles de codis: el 03 amb el 04 i el 07 amb el 17. El següent mijol notable —.6— corresponamb el codi 09 que s'emparella amb el 03.

- Hi ha 12 codis en què cap residu sobrepassa l'.1, però si es baixa el límit al .05 es mantenen 9 d'ells. És a dir, per aquests codis no hi ha cap codi donant suficientment significatiu.

- Per alguns codis hi ha un residu notablement més fort que els altres —cas del codi 03, 07, 09 i 13— però per a altres la importància dels residus està més repartida entre ells sense que en destaqui cap.

  El següent pas és estudiar el graf resultant d'unir les variables que apareixen en la distribució Mijo com a receptors amb les que apareixen en la donant com a tals, el que ens donarà una idea de les confusions que es produeixen entre els diferents codis. Hi ha 9 codis —1, 2, 6, 8, 18, 21, 22, 23 i 25— que queden aïllats: tots tenen un codi alt del paràmetre receptor i baix del paràmetre donant com era d'esperar. La resta es reparteix en 3 grups inconnexes, 2 d'ells petits i un altre més difícil d'interpretar.

  Es pot donar una explicació dels grups resultants:

  1. El grup format pels codis 05-20-19 on s'ha confós el que és, primerament el comerç al detall amb el comerç a l'engròs i seguidament tot el que és comerç amb la indústria alimentària i del tabac.
  2. El format pels codis 24-26-27-28-29, les relacions que s'estableixen entre una sèrie de serveis que s'ha prestat a confusió, essent els «Altres serveis» un sac per algun d'ells. Apareix entre aquests codis l'Administració Pública i la confusió entre el que és i no és la seva pròpia activitat.
  3. El format pels altres codis que no queden aïllats i dels quals es poden veure 3 subgrups més diferenciats:
     3.1. El format pels codis 03-04-09-10 on es confon que són tot tipus d'indústries extractives amb el refinament i tractament dels productes que s'extreuen d'elles.
     3.2. El format pels codis 07-15-17-16 on es confon el que és la indústria de la fusta i el suro amb la fabricació de mobles, producte de matèries plàstiques i també amb el que és la fusteria, fontaneria, etc.
     3.3. El format pels codis 11-12-13-14. Les relacions que es formen aquí giren al voltant dels codis de l'apartat d'indústries metàl·liques i les seves transformacions, amb contínues confusions entre uns i altres, ja que en la majoria de casos per a una bona codificació es necessitarien sistemes de codificació amb un nivell alt de detall.

L'explicació que s'obté per a cada un dels codis a partir de tota la informació disponible és la següent:

Codi 01. *Agricultura, ramaderia, caça i silvicultura (producció i serveis annexes)*

Resulta un codi molt atractor —diag(1) = 95.87 acum(1) = 1—. No apareix en la distribució Mijo ni en la distribució donant, això vol dir que no rep significativament de cap altre codi —el més important és el 13 del que rep un .9%— i no dóna a altres codis de forma significativa —el codi 02 rep d'ell un 4.51%—. Com a conclusió podem afirmar que aquesta variable ha estat ben precodificada.

Codi 02. *Pesca i piscicultura*

Resulta un codi atractor —diag(2) = 89.24 i acum(2) = 3—. No apareix a la distribució Mijo ni a la distribució donant, és a dir, no rep significativament de cap codi —el més important és el 01 d'on rep un 4.51%— i no dóna significativament a altres codis —el codi 20 rep d'ell un .39%—. Com a conclusió és pot afirmar, com abans, que aquesta variable també ha estat ben precodificada per la gent.

Codi 03. *Extracció de combustibles sòlids, petroli, gas natural i minerals radioactius*

Resulta un codi poc atractor —diag(3) = 44.89 i acum(3) = 3—. Aquest codi d'acum ens informa que el que resta per arribar al 95% es distribueix entre només 2 codis més. A la seva distribució Mijo destaca, sobretot, el codi 04 de qual rep un 45.52% i el 10 del que rep un 4.9%; a la vegada actua com a donant del codi 9 d'una forma important.

Es pot concloure que, en gran mesura, aquest codi ha estat confós principalment amb el 04 i en un grau menor amb el 09.

Codi 04. *Resta d'indústries extractives: ferro i minerals metàl·lics no energètics*

Aquest és un codi força atractor —diag(4) = 78.74 i acum(4) = 6—. Amb això sabem que amb 5 codis més arribem al 95%, és a dir, estan bastant repartits els codis del quals rep, sobressortint el 10 de qui rep un 9.3% perquè els altres ja deixen de ser significatius. Important és el fet de ser donant del codi 03 de forma que sobrepassa, inclús, el codi de diag(3). Aquest codi, a més de ser confós amb el 10 —fabricació de productes químics— provoca confusió amb el 03 per ser donant molt important d'ell.

Codi 05. *Indústries de productes alimentaris, begudes i tabac*

Codi força atractor —diag(5) = 78.2 i acum(5) = 5—. No dóna de forma significativa a cap codi, però rep un 11% del 20 —Comerç al detall—, per tant, s'ha confós amb el 20.

Codi 06. *Indústries tèxtil, cuir, sabateria i confeccions tèxtils*

Resulta molt atractor —diag(6) = 95.3 i acum(6) = 1—. No apareix a la distribució Mijo ni a la distribució donant, això vol dir que no rep significativament de cap altre codi ni dóna a altres codis.

Podem concloure que aquest codi ha estat ben precodificat per la gent.

Codi 07. *Indústria de la fusta i el suro*

Resulta un codi molt poc atractor —diag(6) = 42.35 i acum(6) = 4—. Aquest codi d'acum ens informa que el que resta per arribar al 95% es distribueix entre només 3 codis més. A la seva distribució Mijo destaca sobretot el codi 17 —Construcció— del que rep un 7.12%, a la vegada actua com a donant del codi 15 d'una forma significativa. Entre els codis 15 i 07 hi ha una confusió recíproca, actuant mútuament com a donants i com a receptors.

Codi 08. *Indústries de paper, arts gràfiques, edició i reproducció de suports i gravats*

És un codi atractor —diag(8) = 90.1 i acum(8) = 5—. No apareix a la distribució Mijo ni a la distribució donant, no rep significativament de cap altre codi i no dóna a altres codis de forma significativa. Els casos precodificats erròniament d'aquest codi són el 10% però es distribueixen entre els altres de forma homogènia.

Codi 09. *Coqueries*

Codi poc atractor —diag(9) = 50.89 i acum(9) = 6—. No és un codi significatiu per a cap altra codi però rep del 3 -30.97% —i del 10 -5.3%—. Poc més de la meitat dels codis han estat ben precodificats i ha estat confós amb els codis 03 i 10.

Codi 10. *Fabricació de productes químics, fibres artificials i sintètiques, productes minerals no metàl·lics*

Codi força atractor —diag(10) = 77.05 i acum(10) = 8—. Aquest codi és molt donant ja que apareix com a tal per a 4 codis —03, 04, 09 i 15—. Possiblement aquest codi té un camp de significat ampli que porta a la confusió amb diferents codis. D'altra banda, rep del 17 un 9.25% i els altres 7 codis que ens indica acum estan molt repartits. Podem considerar aquest codi com molt poc definit i molt ampli.

Codi 11. *Producció de metalls*

Codi força atractor —diag(11) = 72.28 i acum(11) = 6—. Rep del codi 12 un 17.18% però dóna un 14.19% dels casos a aquest mateix codi, indicant-nos que hi ha una mútua confusió entre ambdós codis.

Codi 12. *Fabricació de productes metàl·lics, construcció de màquines, equip i material mecànic*

Codi poc atractor —diag(12) = 62.2 i acum(12) = 11—, trobant-se els casos erròniament precodificats molt repartits, a excepció dels provinents de l'11. Apareix com a donant dels codis 11, 13, i 14, sent tots ells de fabricació d'algun material o equip, fet que pot portar a confusió.

Codi 13. *Fabricació d'equip i material elèctric, electrònic i òptic*

Còdi poc atractor —diag(13) = 50.64 i acum(13) = 12—, necessita de molts donants per recollir el 95% dels casos. No actua com a donant però sí com a receptor dels codis 17 —19.09%— i 12 —8.9%—. Codi poc definit que ha estat precodificat amb molts altres codis. La confusió amb el 17 —Construcció— és de difícil explicació.

Codi 14. *Fabricació de material de transport*

Codi bastant atractor —diag(14) = 74.05 i acum = 7—. No actua com a donant però sí rep del 12 un 9.4%, presentant una lleugera confusió amb ell.

Codi 15. *Fabricació de productes de suro i matèries plàstiques. Altres indústries manufactureres*

Codi molt poc atractor —diag(15) = 49.46 i acum(15) = 11—. Necessita, també, de molts donants per recollir el 95% dels casos. Actua com a donant del 07 —ja s'explica anteriorment— i és receptor dels codis 10, amb un 10.93%, del 07 amb un 9.19% i del 17 amb un 5.62%. Codi molt repartit entre altres codis. Existeix una confusió mútua entre el 15 i el 17. —els dos treballen amb la fusta— i confusió amb el 10 i el 17 significativa.

Codi 16. *Producció, transport i distribució d'energia elèctrica, gas i aigua*

Codi bastant atractor —diag(16) = 70.97% i acum(16) = 8—. No actua com a donant de cap altre codi però rep del 17 un 11.9%. La confusió pot originar-se per la definició dels codis 16 i 17.

Codi 17. *Construcció*

Codi atractor —diag(17) = 92.28 i acum(17) = 5—. No és receptor però sí, en canvi, és sumament donant, significant que a ell han anat a parar força casos d'altres codis; per a 4 codis dels 5 que dóna —07, 10, 13 i 16— és el principal codi donant i, a més, dóna de forma significativa al 15. Notem que la definició del codi pot confondre a la gent d'altres activitats en ser molt àmplia.

Codi 18. *Venda, manteniment i reparació de vehicles a motor. Gasolineres*

Codi bastant atractor —diag((18) = 77.29 i acum(18) = 7)—. No és receptor ni donant de cap codi. Es considera un codi ben precodificat.

Codi 19. *Comerç a l'engròs i intermediaris*

Codi poc atractor —diag(19) = 59.19 i acum(19) = 16—. Destaca el seu codi d'acum que ens indica que s'havia precodificat amb molts altres codis, del qual sobressurt el 20 amb un 8.6%, a més, actua a la vegada com a donant del mateix 20. La confusió mútua del 19 i el 20 és clara en ser tots dos diferents tipus de comerç.

Codi 20. *Comerç al detall i reparacions d'efectes personals i béns domèstics*

Codi poc atractor —diag(20) = 68.82 i acum(20) = 12—, trobant-se els casos erròniament precodificats molt repartits, amb l'excepció dels que provenen del 19. És donant del 05 i del 19.

Codi 21. *Hotels, restaurants i bars*

Codi atractor —diag(21) = 93.54 i acum(21) = 2—. No apareix a la distribució Mijo ni a la distribució donant, és a dir, no rep significativament de cap altre codi i no dóna a altres codis de forma significativa. Es pot concloure que aquesta variable ha estat ben precodificada per la gent.

Codi 22. *Transport i activitats annexes. Comunicacions*

Codi atractor —diag(22) = 82.45 i acum(22) = 7—, els casos erròniament precodificats estan molt repartits però cap d'ells és significatiu. No apareix a la distribució Mijo ni a la distribució donant, és a dir, no dóna a un altre codi de forma significativa ni rep de cap codi significativament. Com a conclusió es pot afirmar que aquest codi ha estat ben precodificat per la gent.

Codi 23. *Institucions financeres i assegurances*

Codi molt atractor —diag(23) = 95.47 i acum = 1—. No apareix a la distribució Mijo ni en la donant, per tant, ni rep ni dóna a cap altra codi de forma significativa. Es pot concloure que ha estat ben precodificat per la gent.

Codi 24. *Activitats immobiliàries i de lloguer de béns. Serveis prestats a les empreses*

Codi poc atractor —diag(24) = 52.59 i acum(24) = 14—. Destaca el codi d'acum que ens indica que s'havia precodificat com molts altres codis, sobresortint significativament el 27 amb un 11.55% i el 29 amb un 8.65%. No actua com a donant de cap codi.

Codi 25. *Educació*

Codi atractor amb diag(25) = 93.19 i acum(25) = 2. No apareix a la distribució Mijo ni a la donant, és a dir, no dóna a un altre codi de forma significativa ni rep de cap codi significativament. Com a conclusió es pot afirmar que aquest codi ha estat ben precodificat per la gent.

Codi 26. *Sanitat, serveis veterinaris i assistència social*

Codi atractor amb diag(26) = 87.64 i acum(26) = 3. No actua com a receptor però sí és donant del codi 27, ja que inclou a la seguretat social.

Codi 27. *Administracions públiques, Defensa i Seguretat Social*

Codi atractor amb diag(27) = 83.22 i acum(27) = 4. És receptor del codi 26 i donant del 24.

Codi 28. *Servei domèstic*

Codi atractor amb diag(28) = 93.26 i acum(28) = 2. No és receptor de cap codi però sí donant del 29 amb un 20.27% de forma significativa.

Codi 29. *Altres serveis recreatius, culturals i esportius. Representacions diplomàtiques*

Codi poc atractor amb diag(29) = 56.65 i acum(29) = 12 i els seus codis erronis es reparteixen entre molts codis, fet esperat ja que es tracta d'una activitat definida com «altres». Actua com a donant del codi 24 amb un 8.65% i del 28 amb un 20.27%.

Annex 1

Correspondència entre la classificació censal[1] i la utilitzada a la codificació automàtica per la variable Activitat[2]

CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91
011	1	341	13	474	8	812	23
012	1	342	13	475	8	813	23
013	1	343	13	481	15	814	23
014	1	344	13	482	15	819	23
015	1	345	12	491	15	821	23
016	1	346	13	492	15	822	23
019	1	347	13	493	15	823	23
021	1	351	13	494	15	831	23
022	1	352	13	495	15	832	23
023	1	353	13	501	17	833	24
024	1	354	13	502	17	834	24
029	1	355	13	503	17	841	24
030	1	361	14	504	17	842	24
040	1	362	14	611	19	843	24
051	1	363	14	612	19	844	24
052	1	371	14	613	19	845	24

[1]La classificació censal de la variable activitat de l'establiment és la que es troba en el mateix qüestionari del Cens de Població i Habitatge la qual conté els 29 codis possibles amb els quals els enquestats precodificaven l'activitat en la que treballaven.

[2]La classificació utilitzada per codificar automàticament els literals d'activitat és la CNAE a 3 dígits.

CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91
061	2	372	14	614	19	846	24
062	2	381	14	615	19	849	24
111	3	382	14	616	19	851	24
112	3	383	14	619	19	852	24
113	3	389	14	621	19	853	24
114	9	391	13	629	19	854	24
121	3	392	13	631	19	855	24
122	3	393	13	632	19	856	24
123	3	399	13	633	19	859	24
124	3	411	5	634	19	861	24
130	9	412	5	635	19	869	24
140	3	413	5	636	19	911	27
151	16	414	5	637	19	912	27
152	16	415	5	638	19	913	27
153	16	416	5	639	19	914	27
160	16	417	5	641	20	915	27
211	4	418	5	642	20	916	27
212	4	419	5	643	20	917	27
221	11	420	5	644	20	921	29
222	11	421	5	645	18	922	29
223	11	422	5	646	18	931	25
224	11	423	5	647	20	932	25
231	4	424	5	648	20	933	25
232	4	425	5	651	21	934	25
233	4	426	5	652	21	935	25
234	4	427	5	653	21	936	25
239	4	428	5	654	21	937	25
241	10	429	5	661	21	941	26
242	10	431	6	662	21	942	26
243	10	432	6	663	21	943	26
244	10	433	6	669	21	944	26

CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91	CNAE	CPH/91
245	10	434	6	671	20	945	26
246	10	435	6	672	18	946	26
247	10	436	6	679	20	951	26
249	10	437	6	711	22	952	29
251	10	439	6	712	22	953	29
252	10	441	6	721	22	954	29
253	10	442	6	722	22	955	29
254	10	451	6	723	22	959	29
255	10	452	6	724	16	961	29
311	11	453	6	729	22	962	29
312	12	454	6	731	22	963	29
313	12	455	6	732	22	964	29
314	12	456	6	733	22	965	29
315	12	461	7	741	22	966	29
316	12	462	7	742	22	967	29
319	12	463	7	751	22	968	29
321	12	464	7	752	22	969	29
322	12	465	7	753	22	971	29
323	12	466	7	754	22	972	29
324	12	467	7	755	22	973	29
325	12	468	15	756	22	979	29
326	12	471	8	761	22	980	28
329	12	472	8	762	22	990	29
330	13	473	8	811	23		

Annex 2

Resultats de la variable activitat

1. Matriu de correspondències entre els codis precodificats i els codificats. % horitzontals.

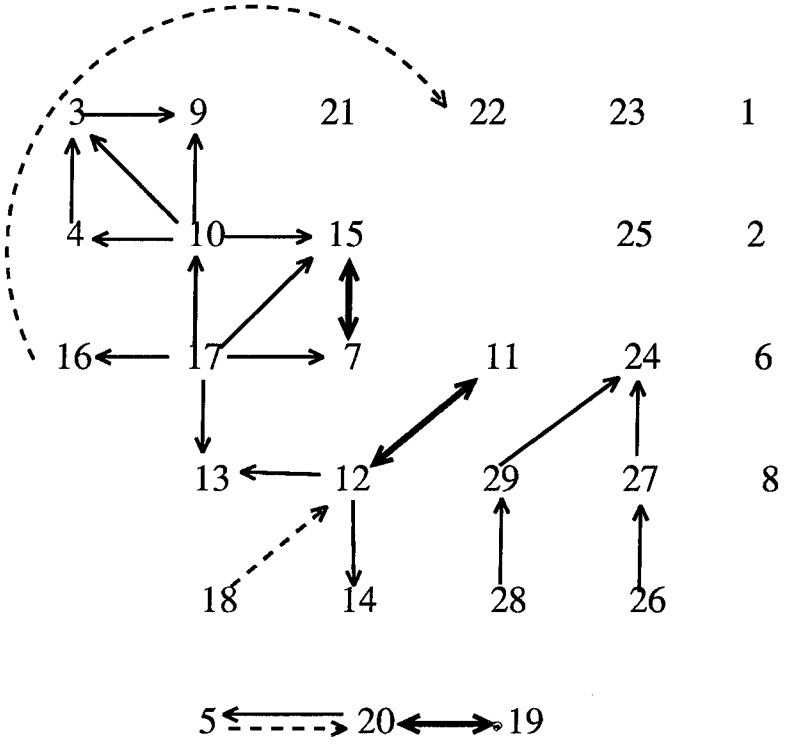
	Precodificat																												
Codificat	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	95,87	0,16	0,04	0,03	0,47	0,04	0,21	0,01	0,03	0,08	0,07	0,12	0,91	0,01	0,03	0,04	0,15	0,02	0,21	0,22	0,17	0,04	0,01	0,04	0,01	0,03	0,32	0,04	0,62
2	4,52	89,24	0,13	0,01	1,43	0	0,13	0,01	0,07	0,01	0	0,12	0,78	0	0,08	0,01	0,03	0,04	0,29	2,58	0,03	0,12	0,03	0,00	0,00	0,01	0,22	0,00	0,13
3	0,12	0,03	44,89	45,52	0,17	0,21	0,14	0,02	0,75	4,9	1,17	0,24	0,14	0,09	0,09	0,49	0,33	0,07	0,19	0,16	0,00	0,07	0,02	0,07	0,02	0,03	0,03	0,00	0,03
4	0,26	0,12	1,27	78,74	1,01	0,4	1,3	0	0,26	9,37	0,89	0,35	0,03	0,17	0,23	0,14	4,18	0,03	0,14	0,00	0,00	0,87	0,00	0,00	0,00	0,00	0,09	0,00	0,14
5	2,31	0,14	0,11	0,05	78,2	0,51	0,1	0,15	0,29	1,2	0,02	0,15	0,07	0,04	0,26	0,09	0,38	0,16	2,68	11,04	0,94	0,48	0,01	0,02	0,07	0,08	0,06	0,03	0,36
6	0,03	0,03	0,03	0,04	0,18	95,3	0,16	0,15	0,04	0,44	0,03	0,12	0,01	0,03	0,28	0,14	0,79	0,06	0,31	1,31	0,01	0,02	0,00	0,01	0,01	0,04	0,01	0,03	0,38
7	0,24	0,07	0,03	0,05	0,11	0,27	42,35	0,17	0,04	0,47	0,12	3,39	0,11	0,1	7,12	0,20	43,51	0,13	0,30	0,78	0,01	0,13	0,01	0,07	0,03	0,00	0,05	0,02	0,10
8	0,06	0,08	0,11	0,03	0,17	0,56	0,69	90,1	0,11	0,43	0,1	0,39	0,15	0,04	0,58	0,03	0,73	0,20	0,45	0,42	0,01	0,54	0,02	0,50	0,07	0,01	0,16	0,07	3,17
9	0,22	0,09	30,97	0,58	2,23	0,22	0,18	0,13	50,89	5,3	0,22	0,22	0,04	0,09	0,22	2,32	0,40	3,70	0,85	0,09	0,00	0,85	0,00	0,00	0,00	0,00	0,13	0,00	0,04
10	0,26	0,08	0,15	0,58	0,31	1,71	0,28	0,45	0,28	77,05	0,66	0,91	0,14	0,13	2,18	0,23	9,25	0,26	1,10	1,78	0,02	0,14	0,01	0,04	0,03	1,03	0,12	0,06	0,72
11	0,11	0,02	0,03	2,44	0,02	0,22	0,18	0,15	0,06	1,65	72,38	17,18	0,86	0,98	0,95	0,97	0,94	0,25	0,32	0,04	0,02	0,05	0,03	0,01	0,01	0,02	0,06	0,02	0,02
12	0,14	0,05	0,03	1,01	0,16	0,97	0,66	0,57	0,03	0,92	14,19	62,2	3,61	2,82	0,99	1,26	3,46	4,06	0,59	1,23	0,08	0,19	0,02	0,10	0,01	0,02	0,04	0,03	0,55
13	0,09	0,12	0,08	0,07	0,04	0,29	0,11	1,45	0,05	2,24	1,14	8,89	50,64	1,12	1,79	4,01	19,09	0,88	1,31	2,41	0,02	0,39	0,06	0,75	0,05	1,58	0,15	0,02	1,14
14	0,06	0,19	0,02	0,11	0,07	0,47	0,27	0,06	0,03	0,49	1,05	9,42	0,66	74,05	1,44	0,69	0,79	7,29	0,87	0,60	0,02	0,91	0,03	0,07	0,03	0,00	0,05	0,02	0,23
15	0,1	0,08	0,05	0,06	1,01	4,27	9,19	3,17	0,14	10,92	0,68	3,78	1,68	0,73	49,46	0,21	5,62	0,78	1,30	4,08	0,04	0,31	0,10	0,35	0,08	0,03	0,05	0,03	1,70
16	0,14	0,03	2,14	0,21	0,84	0,3	0,1	0,04	1,09	2,37	0,12	0,93	4,78	0,1	0,23	70,97	11,94	0,33	0,40	0,63	0,01	0,67	0,02	0,11	0,00	0,31	0,84	0,04	0,29
17	0,09	0,03	0,04	0,16	0,03	0,09	0,19	0,06	0,02	0,49	0,09	0,75	0,34	0,27	0,93	1,03	92,28	0,16	0,47	0,39	0,03	0,49	0,01	0,33	0,01	0,01	0,47	0,24	0,50
18	0,06	0,02	0,21	0,1	0,07	0,2	0,05	0,17	0,4	0,26	0,52	6,66	0,22	5,49	0,23	1,15	1,20	77,30	1,44	2,72	0,13	0,40	0,04	0,04	0,01	0,01	0,25	0,09	0,57
19	0,91	0,3	0,13	0,4	4,61	2,29	1,32	2,85	0,39	2,11	1,04	2,54	0,95	0,23	0,79	0,72	1,47	1,88	59,19	8,61	0,21	3,96	0,45	0,93	0,02	0,29	0,29	0,04	1,10
20	0,49	0,4	0,03	0,03	4,71	2,59	0,31	0,93	0,92	0,73	0,1	0,93	0,99	0,13	0,81	0,13	0,71	2,17	9,93	68,62	0,72	0,23	0,05	0,22	0,05	1,85	0,09	0,08	1,05
21	0,11	0,07	0,01	0,01	1	0,02	0,02	0,01	0,05	1,39	0,07	0,08	0	0,01	0,02	0,01	0,08	0,03	0,14	0,62	93,54	0,13	0,02	0,16	0,26	0,14	0,11	0,26	1,61
22	0,1	0,31	0,04	0,04	0,5	0,21	0,19	0,12	0,03	0,2	0,09	0,42	0,38	0,45	0,07	5,23	0,62	1,67	1,00	0,43	0,34	82,45	0,15	0,51	0,03	0,08	2,69	0,13	1,53
23	0,03	0,05	0,11	0,01	0,05	0,12	0,29	0,17	0,01	0,03	0,01	0,05	0,07	0,01	0,01	0,01	0,02	0,02	0,04	0,06	0,05	0,14	95,47	1,01	0,12	0,88	0,94	0,15	0,10
24	0,57	0,22	0,06	0,29	0,18	0,25	0,41	1,86	0,07	0,3	0,3	0,64	2,24	0,14	0,1	0,41	3,54	0,80	1,52	2,24	0,98	4,21	3,21	52,59	0,56	0,60	11,55	1,54	8,65
25	0,19	0,06	0,03	0,64	0,09	0,07	0,03	0,04	0	0,17	0,01	0,04	0,02	0,03	0,08	0,02	0,11	0,02	0,03	0,03	0,10	0,08	0,11	0,38	93,19	0,77	2,31	0,29	1,07
26	0,46	0,33	0,02	0,05	0,03	0,11	0,06	0,01	0	0,37	0,03	0,08	0,03	0,01	0,04	0,13	0,06	0,02	0,05	0,16	0,42	0,18	0,14	0,20	3,70	87,64	4,18	0,47	1,02
27	0,16	0,08	0,01	0,04	0,15	0,27	0,47	0,05	0,01	0,05	0,2	0,06	0,03	0,01	0,04	0,10	0,41	0,03	0,12	0,24	0,14	0,89	0,44	0,94	0,61	10,30	83,22	0,18	0,76
28	0,14	0,06	0	0	0	0,09	0	0,11	0,01	0,1	0	0,76	0,01	0,01	0	0,01	0,12	0,33	0,08	0,37	0,37	0,23	0,17	0,21	0,69	0,57	0,16	93,26	2,11
29	0,32	0,15	0,16	0,04	0,08	1,19	0,14	1,58	0,15	0,67	0,08	0,87	0,44	0,05	0,16	0,23	0,82	0,45	0,61	1,92	0,97	1,21	0,30	3,76	1,16	1,82	3,77	20,27	56,65

2. Quadre resum

Codi	*Diag*	*Nmijo*	$Mijo_1$	$Mijo_2$	$Mijo_3$	$Donant_1$	$Donant_2$	$Donant_3$	*Acum*	*Receptor*
01	95,87	0	,	,	,	.	.	.	1	5
02	89,24	0	,	,	,	.	.	.	3	4
03	44,89	2	1,01	0,11	,	4	10	.	3	1
04	78,74	1	0,12	,	,	10	.	.	6	3
05	78,20	1	0,14	,	,	20	.	.	5	3
06	95,30	0	,	,	,	.	.	.	1	5
07	42,35	2	1,03	0,17	,	17	15	.	4	1
08	90,10	0	,	,	,	.	.	.	5	4
09	50,89	2	0,61	0,10	,	3	10	.	6	2
10	77,05	1	0,12	,	,	17	.	.	8	3
11	72,38	1	0,24	,	,	12	.	.	6	3
12	62,20	1	0,23	,	,	11	.	.	11	2
13	50,64	2	0,38	0,18	,	17	12	.	12	2
14	74,05	1	0,13	,	,	12	.	.	7	3
15	49,46	3	0,22	0,19	0,11	10	7	17	11	1
16	70,97	1	0,17	,	,	17	.	.	8	3
17	92,28	0	,	,	,	.	.	.	5	4
18	77,30	0	,	,	,	.	.	.	7	3
19	59,19	1	0,15	,	,	20	.	.	16	2
20	68,62	1	0,14	,	,	19	.	.	12	2
21	93,54	0	,	,	,	.	.	.	2	4
22	82,45	0	,	,	,	.	.	.	7	3
23	95,47	0	,	,	,	.	.	.	1	5
24	52,59	2	0,22	0,16	,	27	29	.	14	2
25	93,19	0	,	,	,	.	.	.	2	4
26	87,64	0	,	,	,	.	.	.	3	4
27	83,22	1	0,12	,	,	26	.	.	4	3
28	93,26	0	,	,	,	.	.	.	2	4
29	56,65	1	0,36	,	,	28	.	.	12	2

3. Graf de relacions entre codis[3]

Graf de valors donants i receptors per a la variable activitat



[3]La direcció de la fletxa indica que el codi de sortida dóna al codi de destí de forma significativa.

ENGLISH SUMMARY:

ANALYSIS OF AUTOMATIC CODING VALUES VERSUS RESPONDENT'S VALUES. APPLICATION FOR ACTIVITY AND OCCUPATION

Miquel Delgado Alzamora and Josep Anton Sánchez Cepeda

In the Census of Population 1991, the questions related to «Occupation or profession» and to «Activity of the stablishment or place of work» had to be answered twice: the first question had to be answered with a description and the second one by coding this description according to a list supplied with the questionnaire (this will be called the respondent's code). In the case of «Occupation» the list had 20 items and the list of «Activity» had 29 items. In the Institut d'Estadística de Catalunya both the description and the respondent's code were recorded in order to be compared.

At the stage of information processing a process of automatic coding of the descriptions made by respondents is carried out; for «Occupation» answer, a 3 digit number according to the National Classification of Occupations (CNO) is got, and for «Activity» answer a 3 digit number according to the National Classification of Economic Activities (CNAE) is got. These three digit numbers are translated into their corresponding codes of the 20 and 29 item lists. Those resulting codes will be called automatic codes. So, two coded answers are available for each question: the first one, precoded by the respondant, and the second one, the result of the automatic coding. This situation leads us to the study of the double response in order to check if the respondents make mistakes in coding their own responses or they provide good codings and to detect the confusing items. It is supposed that the description supplied by the respondent is the most reliable; so, the automatic code is taken as the valid value versus the respondent's code. This assumption can be made because in the automatic coding process, the dictionary with the variants has been studied deeply.

The aim of this study is to check the values for each variable in order to discriminate answers leading to no confussion and to group the more mixed up answers. In order to do this, a square matrix A is defined; the element a_{ij} indicates the number of answers with automatic code value «i» and respondent's code value «j». The diagonal gives the number of matches and the rest, the mismatches or residues. Some parameters are defined for each value of each variable; their values will determine which codes are to be grouped or separated. These parameters are:

1. Diag: rate of well coded values; calculated as the weight of the diagonal over the row.

2. Acum: number of necessary values of the row to achieve 95% of success.

3. MIJO: for each value a_{ij}, the weight of the residue over the diagonal a_{ii}.

4. Donant distribution: for each row, values of MIJO distribution bigger than 0.05; they will be the donant values. A graph is made to a better understanding of the values.

5. Nmijo: for each row, the number of residues exceeding the theshold.

QÜESTIIÓ, Vol 20, 2 pp. 293-324, 1996

LES REVISIONS DE LES ESTIMACIONS DE LA COMPTABILITAT NACIONAL

JESÚS MUÑOZ MALO*,
ERNEST PONS FANALS**
JORDI PONS NOVELL**

Aquest treball té com a finalitat introduir el tema de la fiabilitat aplicada a la comptabilitat nacional. Les estimacions macroeconòmiques que s'elaboren a través d'aquesta metodologia de síntesi comptable són difícils d'avaluar des d'una perspectiva de fiabilitat perquè són el resultat d'integrar un conjunt de estadístiques de base on s'ha de garantir la congruència interna dels resultats. La literatura econòmica recull poques alternatives per aproximar la fiabilitat dels comptes nacionals, però una de les més difoses és el mètode de les revisions. Atès que existeixen diferents estimacions comptables segons el seu grau de provisionalitat, és possible considerar la magnitud de les revisions, fins arribar a les dades definitives, com un indicador de la qualitat de les estimacions. Es presenta una anàlisi de les revisions de la comptabilitat espanyola (1981-92) a través d'un conjunt de tècniques estadístiques: anàlisi descriptiva (desviació típica, coeficient de correlació, etc), anàlisi d'aleatorietat temporal (test de ràfegues, test Von Neuman, etc) i descomposició de la variació de les revisions (anàlisi de la variança).

Revisions of national accounting estimates

Keywords: Comptabilitat nacional, fiabilitat, revisions.

Els autors volen expressar els seu agraïment a Àlex Costa per l'atenta lectura d'aquest article i els comentaris aportats.

* Jesús Muñoz Malo. Institut d'Estadística de Catalunya.

** Ernest Pons Fanals i Jordi Pons Novell. Universitat de Barcelona.

- Article rebut el març de 1996.
- Acceptat el juliol de 1996.

1. INTRODUCCIÓ

L'objectiu d'aquest article és analitzar les revisions de les principals magnituds macroeconòmiques elaborades en el marc de la comptabilitat nacional, per fer possible alguna mesura de la fiabilitat d'aquestes estimacions. Com es ben conegut, aquests agregats són fonamentals, entre altres objectius, per analitzar la realitat econòmica, prendre decisions vinculades a la formulació de polítiques econòmiques i facilitar les comparacions internacionals.

Aquest treball presenta tres parts diferenciades. La primera és una breu explicació sobre la comptabilitat nacional. Malgrat la seva importància, la difusió de les seves característiques és limitada i per aquest motiu és interessant aprofundir en la seva naturalesa, explicar com s'elabora i així disposar d'elements de judici que permetin avaluar globalment aquestes estimacions.

La segona part presenta sintèticament el problema de la fiabilitat a la comptabilitat nacional. La literatura econòmica sobre aquest tema reconeix la dimensió del problema, però es suggereixen poques vies per abordar-lo des d'una perspectiva quantitativa. Una aproximació provisional és el mètode de les revisions.

La tercera part és l'anàlisi dels resultats de la comptabilitat nacional espanyola publicats per l INE des de 1980-1992. S'han estudiat un conjunt de sèries (PIB i els seus components des del punt de vista de l'oferta i la demanda) amb l'objectiu d'aprofundir en la naturalesa d'aquestes revisions a través d'una bateria de test estadístics.

2. LES ESTIMACIONS DE LA COMPTABILITAT NACIONAL

La comptabilitat nacional és una tècnica de síntesi estadística que té com a finalitat principal la descripció de les característiques d'una economia durant un període temporal de referència mitjançant un conjunt congruent d'operacions comptables.

La metodologia que s'utilitza per a l'elaboració d'aquestes macromagnituds es basa en els criteris definits al *Sistema europeu de comptes econòmics integrats* (SEC). Aquest marc metodològic és una adaptació a la realitat europea del Sistema de Comptes de les Nacions Unides i en ell es defineixen el conjunt d'operacions i comptes que permeten un coneixement articulat de l'activitat econòmica d'un territori i les seves relacions amb l'exterior.

Malgrat els esforços internacionals per homogeneïtzar la metodologia de la comptabilitat nacional, la seva aplicació als comptes nacionals de cada país ha estat fins el present molt flexible. El motiu és l'existència de grans desigualtats en l'estat de situació de l'estadística de base entre països, la qual cosa dificulta una comparació exhaustiva dels resultats[1].

El procediment que s'utilitza per construir la comptabilitat nacional és complex i molt estructurat. En síntesi, es tracta d'integrar un volum important d'informació estadística de base, no sempre congruent entre si, de manera que els resultats finals siguin coherents. Així, per exemple, per mesurar la producció d'una branca industrial determinada no és suficient amb disposar dels resultats d'una enquesta estadística específica. Aquesta ha d'ésser la font estadística principal, però també s'ha d'analitzar si aquests resultats són compatibles amb altres variables econòmiques connexes com, per exemple, el consum i el sector exterior, atès que a la comptabilitat nacional sempre s'ha de garantir el cumpliment d'identitats comptables del tipus: *producció = consum/inversió + exportacions − importacions*. En conseqüència, cal que les fonts estadístiques implicades (consum / inversió, producció i sector exterior, en aquest cas) siguin prou satisfactòries per a garantir unes estimacions comptables precises.

A mesura que s'ha anat avançant en l'estadística de base i s'ha aprofundit en aquelles àrees on la realitat estadística era més deficient, s'ha possibilitat la maduració i el perfeccionamient del propi sistema de comptes. Així, en última instància, la comptabilitat nacional de cada país reflecteix de manera sintètica la qualitat de les estadístiques necessàries per a la seva elaboració.

Un problema important associat a les estimacions comptables prové de la dificultat de satisfer unes demandes socials que exigeixen que aquesta informació sigui àmplia, exhaustiva, rigurosa i amb un retard de disponibilitat mínim. En última instància és un plantejament irresoluble perquè una comptabilitat nacional amb aquests atributs requereix un període llarg d'elaboració, difícil d'escorçar. L'única via de resoldre parcialment aquesta situació és assumir un cert sacrifici d'amplitud i precisió en els resultats a canvi de reduir significativament el termini de disponibilitat i, per tant, d'ampliar la seva utilitat i impacte social.

Els instituts oficials d'estadística seguint les directrius derivades del Sistema Europeu de Comptes han recullit aquesta idea i tendeixen a presentar resultats amb retards limitats. Això els obliga a *qualificar* les seves estimacions en funció del moment en què es publiquen, assumint implícitament una provisionalitat de les dades durant

[1] Aquesta situació canviarà de manera radical en el marc de la Unió Europea en un futur proper. La importància d'algunes dades macroeconòmiques claus per definir el grau d'integració europea (dèficit i deute de les administracions públiques), així com la necessitat de determinar les contribucions financeres al pressupost comunitari en funció de macromagnituds com el producte interior brut ha obligat a dotar de rang normatiu la nova versió del SEC i assegurar la màxima exhaustivitat i homogeneïtat en el seu càlcul.

un termini ampli de temps. Concretament, l'Instituto Nacional de Estadística (INE) presenta quatre tipus d'estimacions a la seva comptabilitat anual: primera estimació, avanç, provisional i definitiva. El seu calendari és el següent[2]:

Primera estimació[3]	T + 80 dies
Estimació Avanç	T + 8 mesos
1ª Estimació Provisional	T+ 20 mesos
2ª Estimació Provisional	T+ 32 mesos
Estimació definitiva	T + 44 mesos

El grau d'informació disponible vuitanta dies després del període de referència és necessàriament provisional. Per aquest motiu les primeres estimacions es basen en indicadors conjunturals no sempre disponibles en la totalitat de la sèrie necessària i per tant obliguen a elaborar prediccions d'algunes variables.

Les estimacions avanç també s'utilitzen com a informació de base estadística conjuntural, però amb referència anual completa, i s'incorporen altres estadístiques complementàries que permeten millorar la qualitat de les estimacions i ampliar la desagregació d'algunes operacions.

En el moment en què s'elaboren les estimacions provisionals la informació estadística permet quantificar pràcticament totes les operacions comptables. Algunes estimacions són pràcticament tancades, però només quan es disposa de tot el conjunt d'estadístiques estructurals disponibles i es possibilita una àmplia desagregació per branques d'activitat l'estimació passarà a qualificar-se com a definitiva.

3. LA MESURA DE LA FIABILITAT A LA COMPTABILITAT NACIONAL

La pròpia naturalesa de les estimacions de comptatibilitat nacional dificulta extraordinàriament com determinar la magnitud de l'error implícit a aquestes estimacions. Una enquesta permet mesurar *a priori* l'error que assumeix a partir de la grandària de la mostra, sota la hipòtesi de veracitat de les dades declarades i qualitat del directori sobre el qual es determina la mostra. La comptabilitat nacional, en canvi, al fonamentar les seves dades en tot tipus d'informació estadística (enquestes, registres administratius, etc) i requerir ajustos per garantir la coherència de les dades a nivell agregat, fa molt difícil avaluar quantitativament la precisió de les seves estimacions.

[2]Vegeu Quevedo, J. (1995).

[3]Simultània a la publicació del quart trimestre de la comptabilitat trimestral.

La literatura econòmica ha plantejat la problemàtica de la qualitat de les estimacions de comptabilitat nacional a través de dues línies d'argumentació[4]: la primera és el mètode de l'error residual i la segona el mètode de les revisions.

El mètode de l'error residual (*statistical discrepancy*) va ser proposar pel propi R. Stone, un dels impulsors més significatius dels comptes nacionals. El fonament d'aquesta idea es basa en les tres vies de construcció comptable del producte interior brut. Aquesta macromagnitud bàsica es pot calcular com a agregació dels valors afegits de les diferents branques productives (mètode de la producció), com a suma de les diferents possibles destinacions de la renda generada, consum i/o inversió (mètode de la despesa) i com a agregació de les rendes percebudes pels diferents agents vinculats al procés productiu: remuneració d'assalariats i excedent brut d'explotació (mètode de la renda).

Totes tres estimacions proporcionen dades del producte interior brut diferenciades, atès que provenen de fonts estadístiques ben heterogènies. Des d'un punt de vista teòric, però, haurien d'ésser idèntiques. Tenint present aquest fet, R. Stone proposava que la magnitud d'aquestes diferències (els «errors residuals») es podia considerar com un índex de la fiabilitat dels comptes nacionals. Així, una elevada convergència entre estimacions independents a través del mètode de la producció i la despesa (en definitiva, els components del PIB per a l'oferta i la demanda), reflectiria una satisfactòria qualitat global del sistema de comptabilitat nacional.

Hi ha diversos problemes que limiten l'operativitat d'aquest mètode. El primer és que aquesta informació no és disponible, atès que els instituts d'estadística realitzen els ajustos pertinents per garantir la congruència de totes les estimacions i, per tant, presenten el PIB final i no els que inicialment han obtingut a través de les dues vies de càlcul. Però, encara que es disposés d'aquesta informació, només tindria una validesa global i no permetria analitzar les estimacions dels diferents components del PIB.

La segona línia argumental per plantejar la mesura dels errors en el marc de la comptabilitat nacional és el mètode de les revisions. El plantejament de la questió prové també d'un economista reconegut com Franco Modigliani qui va suggerir que les diferències entre estimacions comptables definitives i provisionals es podien considerar un primer indicador de la qualitat de les estimacions dels agregats comptables. Així, en la mesura que les primeres estimacions s'apropin a les definitives, la valoració del mètode d'estimació ha d'ésser favorable. Encara que aquest mètode no permet resoldre el problema conceptual de la mesura dels errors a les estimacions comptables, sí possibilita la validació d'una sèrie d'hipòtesis vinculades amb aquesta problemàtica. Les primeres estimacions de la comptabilitat nacional són bones aproximacions del que seran els resultats definitius? Les revisions tenen algun biaix o són

[4]Una síntesi es pot trobar a Arkhipoff (1991).

estrictament aleatòries? Les variacions de les revisions són homogènies per a totes les variables considerades? El següent epígraf pretén respondre aquestes preguntes i analitzar la situació en el cas de la comptabilitat nacional d'Espanya.

4. LES REVISIONS A LA COMPTABILITAT NACIONAL ESPANYOLA

4.1. Introducció

L'objectiu d'aquest apartat és analitzar les revisions de les estimacions comptables que l'INE elabora de l'economia espanyola, a través d'un conjunt de contrastos estadístics. En primer lloc es mostra l'aplicació d'un conjunt de tècniques estadístiques descriptives per mesurar la dispersió i proximitat de les diferents estimacions al llarg del temps (desviació típica, coeficient de correlació simple i creuat). A continuació s'analitza l'aleatorietat temporal de les revisions a través de diversos tests i, finalment, es presenta la descomposició de la variació de la revisió a través de l'anàlisi de la variança.

La informació de base utilitzada correspon a les primeres estimacions, avanços i estimacions definitives del PIB i els seus principals components d'oferta i demanda. Les sèries comprenen des del 1980 fins al 1992[5] i es presenten en l'annex les seves taxes de variació a preus corrents (C), preus constants (K) i els corresponents deflactors (D). S'ha de tenir present que les estimacions en nivell corresponen a diferents bases (70, 80 i 86), però s'ha obviat la problemàtica del canvi de base en calcular els creixements amb base homogènia.

També cal esmentar que les primeres estimacions de les taxes de creixement de l'any t s'obtenen a partir de l'avanç de l'any $t-1$, i que les taxes de creixement de l'avanç de l'any t s'obtenen a partir de la dada provisional de l'any $t-1$.

Les variables considerades són les següents:

PIB: producte interior brut

Components d'oferta:

- VABA: valor afegit brut agrari
- VABI: valor afegit brut industrial
- VABC: valor afegit brut de la construcció
- VABS: valor afegit brut dels serveis

[5]Per a garantir l'existència d'una sèrie prou àmplia que permeti millorar la significació estadística dels resultats que es presenten, s'ha considerat com a estimació definitiva la corresponent a 1992 però *strictu sensu* és una segona estimació provisional.

Components de demanda:

CPR: consum privat
CPU: consum públic
FBCF: formació bruta de capital fix
EXP: exportacions de béns i serveis
IMP: importacions de béns i serveis.

4.2. Anàlisi descriptiva de les revisions

Aquest apartat pretén obtenir una primera valoració del comportament de les revisions comptables espanyoles. S'utilitzaran, en primer lloc, un conjunt d'instruments estadístics per avaluar la significació de les diferències entre les primeres estimacions i les definitives. Finalment, s'inclou una valoració sobre el guany de precisió dels avanços respecte de les primeres estimacions.

4.2.1. Desviació típica de les revisions

Taula 1

Desviació típica de les revisions

	Desviació típica relativa		**Desviació típica**	
	C	K	C	K
PIB	0,035	0,108	0,430	0,308
IMP	0,048	0,198	0,658	1,705
CPR	0,049	0,188	0,560	0,476
EXP	0,081	0,220	1,172	1,485
CPU	0,101	0,191	1,389	0,934
FBCF	0,111	0,165	1,470	0,797
VABS	0,089	0,125	1,180	0,411
VABI	0,233	0,503	2,315	1,110
VABC	0,288	0,465	3,875	1,765
VABA	0,327	1,942	2,072	1,800

Aquesta primera taula mostra dues mesures de dispersió que permeten valorar la bondat de les primeres estimacions comptables respecte els resultats definitius. La desviació típica mesura el grau de variabilitat de les revisions en termes absoluts i la desviació típica relativa indica el mateix ponderant la desviació típica per la mitjana de l'estimació corresponent. D'aquesta manera es faciliten les comparacions de variables que mantenen diferències significatives a les seves taxes de creixement.

Moltes de les consideracions d'interès que es desprenen de la taula 1 es confirmen per tots dos indicadors. En primer lloc s'ha de destacar el diferent comportament de la desviació del PIB i la dels seus components. Amb independència de la unitat de mesura o la valoració a preus corrents o constants, la dispersió del PIB respecte els valors definitius és força limitada i inferior a qualsevol dels seus components, fins i tot els millor estimats. Aquest fet mostra una evidència interessant: les bones estimacions del PIB no s'expliquen perquè els seus components tinguin una revisió limitada, sinó pel fet que les revisions dels seus components es compensen entre sí.

També s'ha de destacar la magnitud de les revisions de les variables de l'oferta a preus constants, així com l'asimetria de les revisions del valor afegit brut dels serveis (VABS) i les de la resta de branques. Mentre que les primeres presenten desviacions molt limitades (12,5% en termes relatius), les revisions del valor afegit brut de la indústria i construcció són molt importants (50,3% i 46,5%, respectivament). Finalment el VAB agrari presenta els pitjors resultats (194%), malgrat s'ha de considerar que en aquest cas la dada és distorsionada pel fet que la mitjana de creixement del període és petit (0,9%) i per tant dispara el resultat quan es mesura en termes relatius. Si s'utilitza la desviació típica absoluta el resultat és lleugerament inferior al vab de la construcció.

El comportament diferenciat entre variables d'oferta i demanda sembla tenir connexió amb les diferents fonts estadístiques utilitzades. Per a l'estimació directa d'algunes variables de la demanda (CPU, EXP, IMP) la utilització de fonts administratives (duanes, balança de pagaments, comptabilitat pública) és determinant. També influeix en menor grau i indirectament a través del concepte de disponibilitats per a l'estimació de les altres variables (CPR, FBCF). Aquest fet sembla contribuir a l'estabilitat d'aquestes estimacions en la mesura que evita la discontinuïtat de fonts i accelera la conversió de dades provisionals en definitives. En canvi, la importància de les revisions de les branques industrials, agràries i de la construcció es pot relacionar amb la discontinuïtat de fonts estadístiques entre les dues fases de l'estimació. Els indicadors conjunturals que s'utilitzen per la primera estimació (índex de producció industrial, habitatges iniciats, licitacions, etc) són substituïts per estadístiques estructurals ajustades i es generen diferències significatives d'estimació.

Hi ha un conjunt de valoracions que són força diferents segons s'utilitzi la desviació típica absoluta o relativa. En concret, els components del PIB des del punt de vista de la demanda, a preus constants, presenten unes dispersions relatives molt

estables, a l'entorn del 20% i són força inferiors a les dels valors afegits bruts sectorials, a excepció del VAB dels serveis. En termes de desviació típica absoluta, però, hi ha unes revisions significatives entre les exportacions i importacions (1,5 i 1,7 respectivament) que contrasten amb la resta de components, amb valors inferiors a la unitat. Les diferències en els creixements mitjans de les exportacions i importacions (6,8% i 8,6%, respectivament) són importants amb els de la resta de variables de la demanda (2,5% per al consum, 4,9% per al consum públic i 4,8% per a la formació de capital) i això explica aquest comportament asimètric. En les variables de l'oferta no es produeix aquest efecte perquè la dispersió de creixements entre variables és molt inferior.

La diferència entre la dispersió de les revisions a preus corrents i constants també s'ha de ressenyar. En termes absoluts la dispersió és significativament superior a preus corrents que a preus constants. Només quan les fonts estadístiques de les variables provenen de registres administratius a preus corrents, com passa amb les estimacions del consum públic (pressupostos liquidats de les administracions públiques) i comerç exterior (duanes i balança de pagaments), s'ha d'esperar que les estimacions a preus corrents presentin revisions menys significatives que les corresponents a preus constants. Aquest efecte es confirma per a les exportacions i importacions, però no per al consum públic.

En canvi, si es consideren les dades de dispersió en termes relatius els resultats són radicalment diferents: per a totes les variables la dispersió és superior a preus constants. La limitada variabilitat dels diferents deflactors sembla facilitar les estimacions corrents.

4.2.2. Correlacions de les revisions

La taula 2 mostra els coeficients de correlació entre les diferents estimacions a preus corrents, constants i deflactors per a una mateixa variable. Els resultats són interessants i ajuden a explicar la seqüència lògica d'algunes estimacions comptables. *A priori* s'ha d'esperar que tota modificació dels preus i de les estimacions a preus constants impliqui una revisió significativa i del mateix signe de l'estimació a preus corrents.

La gran majoria de variables mostren una correlació molt accentuada entre aquestes revisions, significatives al 5% a partir d'un valor superior al 0,49. S'ha d'assenyalar que les correlacions entre preus corrents i constants són superiors per als components de la demanda que per als de l'oferta. I, en canvi, passa el contrari amb la relació de preus i el deflactor.

També cal destacar els casos en què alguna d'aquestes correlacions presenta un valor no significatiu: vab agrari, consum privat, exportacions i importacions. En el cas

de les exportacions i importacions és previsible aquest resultat. Atès que les dades de duanes i balança de pagaments es valoren a preus corrents i són la font directa de les estimacions de la demanda exterior comptable, la problemàtica d'aquestes es concentra en l'estimació dels corresponents deflactors per assolir els resultats a preus constants. Aquest fet provoca que les revisions de preus no hagin d'estar tan correlacionades amb les estimacions a preus corrents com la resta de variables i, en canvi, s'hagi d'esperar una correlació positiva amb les estimacions a preus constants. Els coeficients de correlació confirmen aquest efecte per tots els casos esmentats excepte per la relació preus i variació a preus constants de les exportacions. Pel sector agrari, l'absència de correlació significativa entre les revisions de les estimacions a preus constants i corrents pot provenir de l'efecte creuat amb els preus, atès que la correlació negativa entre els creixements reals agraris i els preus neutralitza l'efecte total. Finalment, la baixa correlació entre el deflactor del consum privat i el creixement a preus corrents d'aquesta variable pot explicar-se per l'estabilitat de les estimacions del deflactor del consum privat.

Taula 2

Matriu de correlacions de les revisions entre primera estimació i definitiva

		C	J	D			C	K	D
VABA	C	1,00	0,20	0,67	CPR	C	1,00	0,93	0,24
	K		1,00	-0,59		K		1,00	-0,14
	D			1,00		D			1,00
VABI	C	1,00	0,61	0,86	CPU	C	1,00	0,72	0,70
	K		1,00	0,13		K		1,00	0,00
	D			1,00		D			1,00
VABC	C	1,00	0,60	0,88	FBCF	C	1,00	0,69	0,59
	K		1,00	0,14		K		1,00	−0,02
	D			1,00		D			1,00
VABS	C	1,00	0,78	0,84	EXP	C	1,00	0,90	−0,11
	K		1,00	0,45		K		1,00	−0,25
	D			1,00		D			1,00
PIB	C	1,00	0,60	0,38	IMP	C	1,00	0,70	−0,37
	K		1,00	−0,46		K		1,00	−0,91
	D			1,00		D			1,00

4.2.3. Correlacions creuades de les revisions

Taula 3.1

Matriu de correlacions creuada de les revisions. Variables d'oferta

		VABI		VABC		VABS	
		C	K	C	K	C	K
VABA	C	0,40	0,43	−0,03	−0,16	−0,22	−0,51
	K	−0,22	0,27	0,22	0,41	−0,21	−0,12
VABI	C			0,27	−0,11	−0,50	−0,61
	K			0,28	−0,20	−0,82	−0,82
VABC	C					−0,63	−0,20
	K					−0,26	−0,02

Taula 3.2

Matriu de correlacions creuada de les revisions. Variables de demanda

		CPU		FBCF		EXP		IMP	
		C	K	C	K	C	K	C	K
CPR	C	0,16	0,15	−0,14	−0,08	0,02	−0,08	0,42	0,28
	K	0,14	0,29	−0,17	−0,10	−0,20	−0,25	0,43	0,35
CPU	C			0,24	0,30	0,38	0,36	0,50	0,56
	K			0,08	0,36	−0,13	−0,20	0,23	0,42
FBCF	C					0,52	0,21	−0,41	0,04
	K					0,43	0,16	−0,32	0,09
EXP	C							0,19	0,37
	K							0,46	0,41

Aquest deflactor té com a font estadística directa l'índex de preus al consum, el qual es calcula amb periodicitat mensual i obté el caràcter de definitiu quatre mesos més tard de la data de referència. Per tant, les seves revisions són força limitades.

La taules 3.1 i 3.2 mostren les correlacions creuades entre les diferències de la primera estimació i la definitiva per les variables d'oferta i demanda. *A priori* s'ha d'esperar que no hi hagi cap relació significativa entre variables creuades perquè les

estimacions són independents. Només en alguns sectors seria lògic obtenir una associació lineal positiva en les revisions, en la mesura que hi hagin relacions estructurals prou evidents, com per exemple és el cas del consum públic amb el valor afegit brut dels serveis no destinats a la venda i la formació bruta de capital en construcció amb el valor afegit brut de la branca construcció. A la matriu de correlació creuada entre variables d'oferta i demanda no s'evidencia cap vincle d'aquesta naturalesa. Els resultats de la matriu de correlació per a la demanda són congruents amb els esperats ja que molt poques relacions entre variables superen el límit de significació al 95% (les importacions amb el consum públic i les exportacions amb la formació bruta de capital). Per a les variables de l'oferta els resultats són singulars. D'una banda, n'hi ha una correlació en el límit de la significació entre el valor afegit brut agrari i dels serveis. D'altra, es manifesta una correlació negativa molt intensa (−0,82, quan el límit per al 99% de significació és 0,65) entre el valor afegit brut industrial i dels serveis, tant a preus corrents com constants. També la correlació és negativa i intensa entre el vab dels serveis i la construcció (−0,63). La magnitud i el signe d'aquestes correlacions semblen qüestionar la independència d'aquestes revisions.

4.2.4. Les estimacions avanç versus *les primeres estimacions*

Un últim apartat de l'estadística descriptiva de les revisions introdueix una comparativa entre les estimacions avanç i primeres estimacions, per mesurar en quin grau les estimacions avanç milloren les primeres estimacions. Si es tenen present les diferències de calendari entre ambdues estimacions s'hauria de suposar que la simple substitució d'indicadors provisionals per altres més definitius hauria de garantir certa convergència entre estimacions avanç i definitives. El simple resultat de mesurar aquest efecte, a través del nombre d'anys en què les estimacions avanç milloren les estimacions, proporciona resultats sorprenents: només per set anys dels dotze considerats les estimacions avanç són millors que les primeres estimacions[6]. Si s'analitza per tipus de variables s'observa una situació comparativament pitjor per les variables

[6]El nombre d'anys on l'estimació avanç és millor que la primera estimació es presenta a continuació:

	C	K
VABA	8	7
VABI	8	8
VABC	8	9
VABS	8	9
CPR	8	6
CPU	5	5
FBK	6	9
EXP	7	8
IMP	6	9
PIB	6	9

de demanda que per les d'oferta. Hi ha una variable, el consum públic, que presenta unes estimacions avanç que només milloren en cinc anys dels dotze de la sèrie respecte de les primeres estimacions. Aquest fet és paradoxal perquè *a priori* és una de les variables que hauria de millorar nítidament amb les estimacions avanç. En el termini de les primeres estimacions (t+80 dies) la Intervención General de la Administración del Estado, l'organisme competent en matèria de comptabilitat pública, pot disposar d'unes primeres liquidacions de les administracions públiques vinculades a l'administració central (incloent organismes autònoms administratius i seguretat social), però és més difícil disposar d'informació del conjunt de l'administració territorial (comunitats autònomes i corporacions locals). S'hauria d'esperar que vuit mesos després de l'any de referència, la informació pressupostària de l'administració central, i sobretot de l'administració territorial, fos més àmplia i acurada i, per tant, s'apropés més als resultats definitius. Es podria entendre que els resultats no impliquin cap avenç, però és més difícil d'explicar que les primeres estimacions elaborades amb informació més provisional siguin més properes a les definitives que les estimacions avanç.

4.3. Anàlisi de l'aleatorietat temporal

L'objectiu d'aquest apartat consisteix en estudiar si existeix algun tipus de sistematicitat temporal en les revisions experimentades per les taxes de creixement de les diferents magnituds de la comptabilitat nacional espanyola, o si pel contrari es pot acceptar que el comportament d'aquestes revisions al llarg del temps és totalment aleatori. Per realitzar aquesta anàlisi s'ha utilitzat un instrument descriptiu com és el primer coeficient d'autocorrelació i tres estadístics que permeten contrastar la hipòtesi d'aleatorietat en les revisions: l'estadístic Q de Ljung-Box, el número de ràfegues i el test de Von Neuman.

En la taula 4 es mostren els coeficients de la funció d'autocorrelació simple corresponents al primer retard[7] de cada una de les sèries de revisions. Tot i que aquest valor només té un significat purament descriptiu és destacable el fet de que la majoria de coeficients mostren una baixa autocorrelació en les sèries de revisions, fet que pot indicar una absència de sistematicitat en el procés de revisió de les magnituds de la

[7] Aquests coeficients s'obtenen de la següent manera:

$$r_x(1) = \frac{\sum_{t=2}^{n} (x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum_{t=1}^{n} (x_t - \bar{x})^2}$$

on x_t és el valor de cada una de les revisions efectuades per a cada una de les magnituds analitzades en el treball.

comptabilitat nacional.

Taula 4

Funció d'autocorrelació simple (FAS)

	Definitiu/Primera estimació	
	C	K
VABA	−0,31	−0,17
VABI	0,21	0,41
VABC	−0,08	−0,40
VABS	0,02	−0,55
PIB	−0,10	−0,14
CPR	−0,52	−0,32
CPU	−0,09	0,20
FBCF	0,14	0,08
EXP	−0,08	−0,20
IMP	−0,18	−0,04

Taula 5

Estadístic de Ljung-Box (10 retards)

	Definitiu/Primera estimació	
	C	K
VABA	4,46	14,84
VABI	13,22	6,09
VABC	11,08	11,09
VABS	2,71	13,30
PIB	11,65	11,75
CPR	6,96	4,71
CPU	8,13	5,76
FBCF	3,31	5,91
EXP	8,35	6,08
IMP	7,93	9,62

Així mateix, es pot utilitzar el concepte de l'autocorrelació per realitzar un contrast d'aleatorietat a partir de l'estadístic Q de Ljung-Box[8]. En la taula 5 es mostra el valor d'aquest estadístic corresponent al retard número 10 per les diferents sèries de revisions analitzades. Atès que el valor crític de l'estadístic al 95% de confiança és de 19,675, es confirma l'aleatorietat temporal per totes les variables considerades.

A la literatura estadística s'han desenvolupat dos tests estadístics que permeten contrastar la hipòtesi d'aleatorietat en un cas com el que s'està analitzant en aquest estudi. El primer és l'anomentat test de ràfegues. A diferència de l'anterior aquest és un test qualitatiu, basat no en la magnitud de les revisions sinó en el signe d'aquestes.

Una ràfega[9] és una successió de símbols de la mateixa classe limitada per símbols de diferent classe. En el nostre cas es disposa de dos tipus de signes «+» i «-». Si les dues classes d'observacions procedeixen aleatòriament d'una mateixa població, els signes «+» i «-» apareixeran barrejats i, per tant, el número de ràfegues serà gran. En canvi, si els símbols es repeteixen sovint, és a dir, hi ha poques ràfegues, això és contrari a la noció d'aleatorietat.

Si es representa per $\gamma = n_1/n$, la proporció d'elements del tipus «+» en la mostra (i, per tant, $1-\gamma$ és la proporció d'elements del tipus «-»), Wald i Wolfowitz van demostrar que la variable aleatòria número de ràfegues, R, segueix una distribució asimptòtica normal donada per:

$$R \sim N\left[2\gamma(1-\gamma)n;\ 2\gamma(1-\gamma)\sqrt{n}\right]$$

Aquesta aproximació és acceptable si $n_1 > 10$ i $n_2 > 10$, però per al cas de grandàries mostrals més reduïdes la distribució de R està tabulada i es poden consultar les taules adients. En la taula 6 es mostra el número de ràfegues de cada una de les sèries de revisions. No s'han reproduït els valors crítics[10] ja que aquests són diferents per a cada una de les sèries ja que depenen de n_1 i n_2 que varien en cada cas, però en cap de les sèries de signes analitzades es pot rebutjar amb una confiança del 95% que aquestes s'hagin generat de manera aleatòria.

Tot i que s'han utilitzat els valors crítics tabulats específicament per grandàries mostrals reduïdes, els resultats d'aquest test s'han de prendre amb precaució ja que grandàries mostrals tan petites provoquen que la potència del test no sigui molt elevada.

[8] Vegeu Ljung i Box (1978).

[9] La denominació anglesa d'aquest concepte és *run*. En castellà s'ha traduït per *racha* o per *tendencia*. En aquest treball s'ha optat per traduir-lo per ràfega.

[10] Vegeu, per exemple, Ruiz-Maya i Martín Pliego (1995).

Taula 6

Test de ràfegues

	Definitiu/Primera estimació	
	C	K
VABA	4	9
VABI	8	6
VABC	8	7
VABS	5	6
PIB	7	
CPR	9	10
CPU	6	6
FBCF	7	5
EXP	4	6
IMP	4	8

L'última tècnica aplicada per valorar el grau d'aleatorietat temporal de les revisions és el test de Von Neuman. Aquest parteix del fet que amb observacions de caràcter temporal es disposa de dos estimadors no esbiaixats de la variància. El primer és la quasivariància mostral s^2, que és un estimador no esbiaixat:

$$s^2 = \frac{\sum_{t=1}^{n} (x_t - \bar{x})^2}{n-1}$$

A més, en el cas de independència serial, el quadrat mig de les diferències successives d^2:

$$d^2 = \frac{\sum_{t=2}^{n} (x_t - x_{t-1})^2}{2(n-1)}$$

és també un estimador no esbiaixat de la variància.

L'estadístic de contrast és:

$$r = \frac{d^2}{s^2} = \frac{\sum_{t=2}^{n} (x_t - x_{t-1})^2}{\sum_{t=1}^{n} (x_t - \bar{x})^2}$$

Si és certa la hipòtesi nul·la d'aleatorietat i independència en les observacions, r tendirà a prendre valors pròxims a la unitat ja que s^2 i d^2 seran similars. En cas contrari, si existeix alguna associació entre els valors consecutius de x_i (si cada x_i tendeix a estar pròxim al seu inmediat anterior) el numerador de l'estadístic d^2 serà més petit que s^2 i l'estadístic r també es farà petit. En el cas d'aleatorietat es pot demostrar que:

$$E(r) = 1 \qquad V(r) = \frac{n-2}{n^2-1}$$

En el cas $n > 20$, una bona aproximació de la distribució asimptòtica de r s'obté de la següent manera:

$$r \sim N\left(1; \sqrt{\frac{n-2}{n^2-1}}\right)$$

Taula 7

Test de Von Neuman

	Definitiu/Primera estimació	
	C	K
VABA	1,13	0,99
VABI	1,16	0,69
VABC	1,03	1,28
VABS	0,73	1,07
PIB	0,75	1,34
CPR	1,38	1,35
CPU	0,57	0,80
FBCF	0,89	0,94
EXP	1,10	1,15
IMP	0,80	1,03

Tot i disposar d'una grandària mostral inferior a 20, s'ha calculat l'estadístic r per cada una de les sèries de revisions. Els resultats estan recollits en la taula 7, i mostren

com no hi ha cap rebuig de la hipòtesi d'independència serial[11], és a dir, s'accepta la hipòtesi d'aleatorietat temporal en les revisions de les magnituds considerades en aquest estudi.

4.4. Descomposició de la variació de la revisió

Com a complement de l'anàlisi descriptiva i de l'estudi de l'aleatorietat de les revisions s'ha considerat interessant descomposar la variabilitat en la grandària de les revisions mitjançant una anàlisi de la variància, ja que de l'evolució de les revisions es deriven un conjunt d'interrogants que un estudi d'aquest tipus pot ajudar a respondre.

És interessant utilitzar alguna eina estadística que permeti contrastar si hi ha diferències significatives en les revisions de les diferents estimacions. Per fer-ho es considera com a base de dades el conjunt de revisions que s'han produït per cada una de les diferents magnituds i per cada un dels anys, tant a preus constants com a corrents, i tant respecte de la primera estimació com respecte de l'avanç. En aquest sentit, la variació total de la mostra Q_T és:

$$Q_T = \sum_{t=1}^{n} (x_t - \bar{x})^2$$

on x_t és el valor de cada una de les revisions efectuades per a cada una de les magnituds.

La base d'una anàlisi de la variància consisteix en descomposar aquesta variació total en la variació que es pot atribuir a cada un dels factors que *a priori* es considera que poden ser explicatius d'aquesta variabilitat. Una vegada definits aquests factors es contrasta la seva significació. En el cas de les revisions s'han considerat com a factors importants els quatre següents:

A. El factor macromagnitud.
B. El factor preus corrents/preus constants.
C. El factor any.
D. El factor avanç/primera estimació.

Es designa per n_A, n_B, n_C i n_D el número de possibilitats de cada un dels factors; és a dir, n_A és el número de macromagnituds analitzades (en el nostre cas 10), n_B és igual a 2 ja que es disposa de dades valorades a preus corrents i a preus constants, n_C recull el número d'anys (12) i finalment, n_D és 2 ja que es disposa de les revisions de les dades de la primera estimació i de l'avanç. Habitualment, en una anàlisi de la

[11] L'interval d'acceptació de la hipótesi d'independència amb un nivell de confiança del 95% és [0,48, 1,52].

variància les diferents possibilitats de cada factor s'anomenen nivells del factor. Per tant n_A, n_B, n_C i n_D són el número de nivells o possibilitats de cada factor.

Es designa per X_{ijkl} a la dada que correspon al $i-$èssim nivell del primer factor (macromagnitud), $j-$èssim nivell del segon factor (corrents o constants), $k-$èssim nivell del tercer factor (any) i $l-$èssim nivell del quart factor (avanç o primera estimació). Per exemple X_{3241} és la revisió de la primera estimació ($l = 1$) de la macromagnitud número 3 ($i = 3$ correspon al Valor Afegit Brut a la Construcció) pel quart any ($k = 4$ correspon a l'any 1984) i valorada a preus constants ($j = 2$).

Aleshores, la variabilitat deguda al factor macromagnitud és:

$$Q_A = n_B \cdot n_C \cdot n_D \sum_i (x_{i...} - \bar{x})^2$$

on $x_{i...}$ són els valors mitjans de totes les revisions que fan referència a la magnitud $i-$èssima (nivell i del primer factor). De la mateixa manera, la variabilitat deguda a cada un dels altres factors és:

$$Q_B = n_A \cdot n_C \cdot n_D \sum_j (x_{.j..} - \bar{x})^2$$

$$Q_C = n_A \cdot n_B \cdot n_D \sum_k (x_{..k.} - \bar{x})^2$$

$$Q_D = n_A \cdot n_B \cdot n_C \sum_l (x_{...l} - \bar{x})^2$$

Es demostra que la variació total es pot descomposar com:

$$Q_T = Q_A + Q_B + Q_C + Q_D + Q_R$$

on Q_R és una variació residual[12].

Sota la hipòtesi de normalitat cada una d'aquestes sumes de quadrats corregida pels graus de llibertat adequats segueix una distribució χ^2. Atès que aquestes distribucions són estocàsticament independents es pot contrastar, per exemple, la significació del primer factor utilitzant l'estadístic:

$$F_A = \frac{Q_A/(n_A - 1)}{Q_R/(n - n_A - n_B - n_C - n_D + 3)}$$

que sota la hipòtesi nul·la de no significativitat del primer factor segueix una distribució F de Fisher-Snedecor. Per contrastar la significació dels altres factors s'utilitzen estadístics amb expressions equivalents.

[12]Aquesta no és estrictament la variació residual ja que conté la variació deguda a les possibles interaccions entre diversos factors. Prèviament s'ha obtingut que cap d'aquestes interaccions és significativa, i, per tant, s'han agrupat tota la resta de termes en un terme residual total. Per a una anàlisi més detallada d'aquest procediment estadístic es pot consultar Scheffe (1959).

Taula 8

Anàlisi de la variança

Factor	Suma quadrats (1)	G. de llibertat (2)	(1)/(2)	Estadístic	Valor crític
Variables	41,579	10	4,158	2,59	1,83
Preus constants/ corrents	0,036	1	0,036	0,02	3,84
Any	159,906	11	14,537	9,05	1,75
Avanç/primera estimació	5,535	1	5,5535	3,45	3,84
Residus	809,528	504	1,606		
Total	1016,584	527	1,929		

La taula 8 recull el resultat d'aquesta anàlisi de la variància[13]. Els factors macromagnitud i any són clarament significatius tal com era previsible *a priori*. Es pot destacar el resultat que no hi ha diferències significatives en la importància de les revisions de les magnituds a preus corrents i a preus constants. Així mateix el factor avanç/primera estimació no és significatiu[14].

La no significació d'aquest últim factor porta a la conclusió que les revisions més importants es produeixen en la dada definitiva respecte a les dades anteriors i no entre l'avanç i la primera estimació, és a dir, que entre la primera estimació i l'avanç d'una magnitud es produeixen diferències poc significatives.

5. CONSIDERACIONS ENTORN DE LES ESTIMACIONS CATALANES

Les estimacions comptables de l'economia catalana tenen una limitada sèrie històrica. Hi ha un conjunt de factors que expliquen aquest fet. D'una banda ha estat necessari un canvi en les prioritats de política estadística a nivell estatal, derivada

[13]No és sorprenent que la variació residual suposi un gran percentatge de la variació total, ja que en la grandària de les revisions hi influeixen molts factors dels considerats. En tot cas, i encara que no es tenen en compte altres variables rellevants, això no és incompatible amb contrastar la significació dels factors considerats

[14]L'estadístic es troba molt proper al valor crític, fet que provoca certa incertesa.

de la descentralització política i les directrius de la política regional comunitària, per possibilitar la implantació d'una Comptabilitat regional d'Espanya on es presentin les principals dades macroeconòmiques per comunitats autònomes.

Aquesta informació, malgrat ésser de gran utilitat, presenta dos grans inconvenients: l'important retard en la seva disponibilitat limita molt la seva projecció pública i l'absència d'estimacions a preus constants impedeix una anàlisi acurada de l'evolució econòmica regional.

L'Institut d'Estadística de Catalunya va iniciar la seva activitat de producció estadística en l'àmbit econòmic en un període relativament recent[15]. Dins les seves prioritats es va definir la necessitat de disposar d'estimacions comptables amb un retard assumible i també la presentació de resultats a preus corrents i constants. Aquestes estimacions són disponibles en la publicació anual *Evolució de les principals macromagnituds de l'economia catalana.*

La necessitat d'explicitar les diferències entre les diverses estimacions disponibles per un mateix any va motivar que l'Institut publiqués, a partir de l'edició corresponent a l'any 1993 i amb caràcter anual, un annex dins la seva publicació dedicat a la revisió de les pròpies estimacions, on es presenten les diferències anuals de les principals macromagnituds i els factors principals que les expliquen.

L'absència d'unes sèries de macromagnituds catalanes prou àmplies a preus corrents i constants i que incorporin juntament amb les estimacions definitives les estimacions prèvies, amb el seu corresponent grau de provisionalitat, impedeix avaluar la magnitud i dispersió de les revisions. En conseqüència, també impossibilita l'elaboració d'una anàlisi estadística equivalent a la duta a terme amb les estimacions espanyoles.

6. CONCLUSIONS

Les principals conclusions que es poden extreure de l'anàlisi estadística elaborada per a les estimacions de la comptabilitat nacional d'Espanya es poden sintetitzar en els següents punts:

1. Les estimacions del PIB global a preus constants presenten unes revisions de poca magnitud (10,8% en termes de desviació típica relativa). Aquest fet implica una valoració molt positiva, però s'ha de matisar pel fet que aquestes

[15]L'Institut d'Estadística de Catalunya va ésser creat pel decret 341/89, però la llei que regula el Pla estadístic de Catalunya no es va aprovar fins l'any 1991.

petites revisions del PIB no es basen en unes revisions equivalents del seus components, sinó en la seva compensació. És a dir, els components del PIB des de l'oferta i la demanda tenen desviacions significatives però de diferent signe i, per tant, l'efecte net implica unes revisions mínimes. Concretament, la mitjana de revisió dels components del PIB, excloent el sector agrari, supera el 25%.

Aquesta idea es confirma amb els resultats de les correlacions creuades de les revisions. En aquest cas es detecten alguns valors estadísticament significatius com, per exemple, la intensa correlació negativa entre les revisions de les estimacions del valor afegit industrial i els serveis (0,82).

2. Els components de la demanda presenten unes revisions baixes, a l'entorn d'un 20% de desviació típica relativa, i molt homogènies per les diferents variables. En canvi, les revisions dels components de l'oferta són clarament superiors i més disperses (indústria i construcció per sobre del 45%), a excepció del sector serveis (12,5%).

3. Les estimacions a preus corrents són més estables que a preus constants. Per a totes les variables la desviació típica relativa és inferior en les revisions a preus corrents que a preus constants. En canvi si es mesura a través de la desviació típica absoluta, només les exportacions i importacions presenten revisions més petites a preus corrents que a preus constants.

4. Les estimacions avanç no representen una millora generalitzada respecte de les primeres estimacions. I en algun cas singular, com el consum públic, arriben a representar una estimació divergent respecte de la definitiva.

5. No s'aprecia cap evidència de dependència temporal a les revisions. Els test de Ljung- Box, de ràfegues i Von Neuman confirmen l'aleatorietat temporal per totes les variables.

Annex

Taxes de variació del PIB i dels components d'oferta i demanda (1981-92)

Comptabilitat Nacional d'Espanya

Taxes de creixement interanual

Dades definitives

	VABA			VAB			VABC			VABS			PIB		
	C	K	D	C	K	D	C	K	D	C	K	D	C	K	D
81	−2,56	−10,40	8,76	14,03	0,50	13,46	9,90	−2,52	12,75	15,33	1,50	13,63	14,11	0,44	13,60
82	17,74	2,20	15,21	11,28	−0,54	11,88	16,50	2,54	13,62	16,56	2,10	14,17	14,68	0,88	13,68
83	10,35	6,24	3,86	14,20	2,67	11,24	6,88	0,10	6,78	15,24	1,86	13,13	14,24	2,22	11,76
84	19,89	8,61	10,39	13,77	1,08	12,55	0,24	−6,16	6,82	15,95	2,65	12,95	12,94	1,47	11,62
85	6,19	3,11	2,99	10,74	2,09	8,47	10,42	2,24	8,00	11,09	2,31	8,57	11,06	2,31	8,55
86	4,02	−9,10	14,43	21,53	5,61	15,07	23,28	5,93	16,38	11,69	3,54	9,01	14,55	3,28	10,92
87	8,52	11,59	−2,75	8,82	5,85	2,81	14,02	8,33	5,25	11,25	4,61	6,35	11,82	5,64	5,85
88	8,03	2,52	5,37	8,05	4,85	3,05	21,63	10,16	10,42	11,51	4,77	6,43	11,17	5,23	5,65
89	1,47	−6,72	8,78	9,38	3,40	5,78	24,90	13,81	9,74	12,41	5,30	6,75	12,07	4,75	6,99
90	7,87	3,36	4,37	6,21	1,85	4,28	21,03	9,86	10,17	13,61	3,91	9,34	11,36	3,61	7,48
91	−3,97	−1,83	−2,17	7,02	1,40	5,54	11,95	3,26	8,42	11,61	2,94	8,42	9,54	2,36	7,01
92	−9,65	−2,11	−7,70	0,62	−1,76	2,42	−0,44	−4,59	4,35	12,47	3,07	9,13	7,47	0,69	6,73

Comptabilitat Nacional d'Espanya

Taxes de creixement interanual

Dades definitives

	CPR			CPU			FBCF			EXP			IMP		
	C	K	D	C	K	D	C	K	D	C	K	D	C	K	D
81	14,03	−0,88	15,05	15,05	1,46	13,39	19,25	1,17	17,87	27,99	6,94	19,69	22,63	−3,53	27,11
82	14,99	0,66	14,23	18,71	6,52	11,44	11,28	−2,46	14,09	20,41	6,69	12,85	18,34	4,84	12,87
83	12,87	0,29	12,55	17,84	3,88	13,44	9,90	−2,37	12,57	28,54	9,96	16,90	21,20	−0,30	21,56
84	10,55	−0,38	10,97	11,56	2,88	8,44	2,93	−5,79	9,26	25,77	11,73	12,57	10,35	−1,03	11,50
85	10,80	2,38	8,22	13,29	4,64	8,27	12,04	4,14	7,58	9,68	2,75	6,75	10,28	6,22	3,83
86	12,67	3,61	8,74	14,42	5,73	8,23	16,08	9,99	5,54	−0,66	1,29	−1,92	−2,51	16,48	−16,30
87	11,83	5,79	5,71	15,01	8,85	5,66	19,40	14,04	4,69	8,97	6,33	2,49	21,04	20,11	0,77
88	9,96	4,76	4,98	8,67	4,03	4,46	21,11	14,08	6,16	8,50	5,33	3,01	15,57	14,34	1,07
89	12,74	5,76	6,61	15,31	8,29	6,48	19,40	13,60	5,11	7,60	2,99	4,47	19,92	17,21	2,31
90	10,35	3,61	6,51	13,54	5,64	7,47	13,40	6,94	6,03	4,97	3,22	1,69	6,55	7,86	−1,21
91	9,48	2,92	6,38	14,52	6,56	7,47	5,89	0,88	4,97	9,99	7,87	1,96	8,65	9,01	−0,33
92	8,76	2,18	6,44	13,64	4,04	9,22	−0,96	−3,89	3,04	10,63	7,28	3,13	8,22	6,86	1,27

Comptabilitat Nacional d'Espanya

Taxes de creixement interanual

Avanç

	VABA			VABI			VABC			VABS			PIB		
	C	K	D	C	K	D	C	K	D	C	K	D	C	K	D
81	−5,38	−10,89	6,18	12,95	0,00	12,95	7,45	−1,95	9,58	14,56	1,34	13,05	13,60	0,39	13,16
82	16,59	1,29	15,10	11,13	−0,61	11,81	12,01	1,10	10,80	16,33	2,10	13,94	14,76	1,25	13,35
83	8,96	3,06	5,72	14,17	2,93	10,92	6,17	−2,86	9,30	15,80	1,94	13,59	14,63	2,45	11,89
84	20,50	9,99	9,56	14,10	1,53	12,38	6,56	−4,00	11,00	13,92	2,30	11,36	14,05	2,05	11,77
85	6,24	1,34	4,83	11,76	2,10	9,46	10,30	1,80	8,35	11,30	2,34	8,76	11,18	2,24	8,74
86	1,72	−11,03	14,33	21,18	5,03	15,38	23,10	5,93	16,21	13,88	3,67	9,85	14,57	3,28	10,94
87	7,00	9,61	−2,38	6,90	5,07	1,75	28,45	10,42	16,33	13,06	4,76	7,92	11,79	5,54	5,92
88	8,47	3,70	4,59	9,51	4,20	5,10	25,66	12,50	11,70	11,33	4,55	6,48	11,76	5,16	6,28
89	2,60	−6,90	10,20	7,85	3,44	4,26	24,28	13,70	9,30	13,37	5,42	7,54	12,10	4,82	6,95
90	4,31	2,46	1,81	5,38	1,39	3,94	21,22	10,30	9,90	14,14	4,23	9,51	11,21	3,61	7,34
91	−3,25	−2,29	−0,99	5,29	0,56	4,70	11,80	3,51	8,01	11,98	3,39	8,32	9,31	2,29	6,86
92	−7,14	−2,14	−5,11	1,69	−1,07	2,79	−0,98	−4,38	3,55	11,60	2,82	8,54	7,35	0,80	6,50

Comptabilitat Nacional d'Espanya

Taxes de creixement interanual

Avanç

	CPR			CPU			FBCF			EXP			IMP		
	C	K	D	C	K	D	C	K	D	C	K	D	C	K	D
81	13,25	−1,30	14,75	16,30	2,00	14,03	17,60	1,00	16,44	27,99	7,90	18,62	22,63	−4,00	27,74
82	14,94	0,60	14,25	19,51	6,20	12,53	9,69	−1,81	11,72	21,03	7,10	13,01	18,50	4,54	13,35
83	13,04	0,64	12,33	17,51	3,77	13,23	10,35	−1,03	11,50	28,66	8,00	19,13	20,46	−0,17	20,67
84	9,98	−1,14	11,25	13,64	1,96	11,45	7,51	−3,60	11,52	29,95	14,95	13,05	10,04	0,04	9,99
85	10,25	1,81	8,29	11,41	3,10	8,06	12,04	3,86	7,87	9,89	2,87	6,82	9,54	5,35	3,97
86	12,74	3,66	8,76	13,74	5,10	8,22	15,92	9,56	5,81	−0,99	0,99	−1,96	−3,37	15,44	−16,30
87	11,15	5,45	5,40	15,03	8,69	5,83	20,61	14,62	5,23	8,46	5,87	2,45	20,65	20,41	0,20
88	11,37	5,95	5,12	9,47	4,14	5,12	20,39	14,20	5,42	10,22	7,46	2,56	17,62	18,91	−1,08
89	12,53	5,53	6,63	14,56	7,60	6,47	19,20	13,71	4,83	7,42	2,91	4,39	19,69	17,00	2,30
90	10,37	3,74	6,38	11,67	4,22	7,14	13,44	6,89	6,13	4,97	3,22	1,69	6,55	7,80	−1,16
91	9,52	3,10	6,24	10,72	4,21	6,25	6,91	1,60	5,23	8,71	6,56	2,02	8,56	8,91	−0,32
92	8,66	2,09	6,44	12,40	3,76	8,33	−1,57	−3,92	2,44	10,08	6,72	3,15	8,05	6,58	1,38

Comptabilitat Nacional d'Espanya

Taxes de creixement interanual

Primera estimació

	VABA			VABI			VABC			VABS			PIB		
	C	K	D	C	K	D	C	K	D	C	K	D	C	K	D
81	−3,88	−6,53	2,83	10,54	0,03	10,52	12,32	0,00	12,32	15,12	1,85	13,03	13,41	0,30	13,07
82	17,32	1,30	15,81	10,97	−0,50	11,52	10,85	0,50	10,30	17,31	2,10	14,90	14,64	1,31	13,16
83	10,55	4,01	6,29	14,33	3,00	11,00	10,42	−1,48	12,08	14,55	2,00	12,30	14,58	2,28	12,03
84	21,00	9,99	10,00	13,22	2,30	10,68	7,89	−3,50	11,80	13,68	2,08	11,36	13,86	2,31	11,28
85	7,06	2,12	5,15	11,71	2,55	9,43	7,89	−0,42	8,13	11,40	2,43	9,35	11,02	2,09	9,22
86	4,03	−9,00	14,32	15,40	3,48	11,52	16,59	5,97	10,02	13,90	4,12	9,40	14,88	3,46	11,04
87	10,60	9,50	1,00	10,30	4,50	5,55	14,28	10,00	3,89	11,96	4,52	7,12	11,57	5,23	6,03
88	8,47	3,70	4,59	9,51	4,20	5,10	25,66	12,50	11,70	12,04	5,20	6,50	11,76	5,30	6,13
89	5,60	−4,00	9,99	8,80	4,00	4,61	24,44	12,40	10,71	13,32	5,31	7,61	12,42	4,95	7,12
90	4,14	2,80	1,30	5,78	1,63	4,08	21,26	10,40	9,84	13,92	4,04	9,50	11,25	3,66	7,32
91	−3,25	−2,29	−0,98	5,25	0,75	4,47	11,97	3,73	7,94	12,11	3,46	8,36	9,39	2,37	6,86
92	−5,50	−0,48	−5,05	3,54	0,51	3,01	−2,35	−4,56	2,31	10,56	2,19	8,19	7,09	0,98	6,05

Comptabilitat Nacional d'Espanya

Taxes de creixement interanual

Primera estimació

	CPR			CPU			FBCF			EXP			IMP		
	C	K	D	C	K	D	C	K	D	C	K	D	C	K	D
81	13,28	−1,10	14,54	15,50	2,00	13,24	18,93	1,50	17,17	28,13	8,06	18,57	23,32	−3,20	27,40
82	14,86	0,40	14,40	18,20	6,09	11,42	10,40	−1,29	11,85	21,00	7,00	13,08	18,29	3,40	14,40
83	13,09	0,70	12,30	16,74	4,09	12,15	11,88	−1,48	13,57	28,00	7,59	18,97	20,30	−0,61	21,03
84	10,11	−0,80	11,00	12,60	3,01	9,31	7,50	−3,00	10,83	29,84	15,40	12,51	10,29	0,99	9,21
85	9,94	1,52	8,51	12,96	4,54	8,34	12,31	4,12	10,61	10,67	3,60	13,70	10,34	6,13	4,32
86	13,10	4,00	8,75	15,08	6,00	8,57	18,16	10,80	6,64	−0,15	1,12	−1,25	−2,21	14,97	−14,94
87	11,10	5,50	5,31	15,50	9,00	5,96	20,38	14,51	5,13	9,18	6,55	2,47	21,62	22,55	−0,76
88	11,20	5,82	5,08	10,25	4,65	5,35	20,39	14,20	5,42	10,22	7,46	2,56	17,62	18,91	−1,08
89	12,51	5,44	6,71	12,93	5,60	6,94	20,04	13,24	6,00	9,32	5,20	3,92	20,00	15,98	3,47
90	10,40	3,72	6,44	11,68	4,41	6,96	13,38	6,74	6,22	5,93	4,19	1,67	6,85	8,10	−1,16
91	9,19	2,87	6,15	12,44	5,14	6,94	6,93	1,63	5,22	10,51	8,43	1,91	8,99	9,36	−0,33
92	8,71	2,40	6,16	10,90	4,00	6,63	−1,28	−2,98	1,75	10,28	6,41	3,63	8,20	6,75	1,36

REFERÈNCIES

[1] **Arkhipoff, O.** (1991). «Esbozo de una metrología estadística económica y social». *Información Comercial Española*, **698**.

[2] **Cristóbal, A.** i **Quilis, E.** (1990). «Un análisis de las revisiones de los agregados de la contabilidad nacional (óptica del gasto)». *Boletín Trimestral de Coyuntura*, **35**, 37–51.

[3] **Frank, H.** i **Althoen, S.C.** (1994). *Statistics. Concepts and Applications*. Cambridge University Press. New York.

[4] **Ljung, G.** i **Box, G.E.P.** (1978). «On a measure of lack of fit in time series models». *Biometrika*, **65**, 133–137.

[5] **Quevedo, J.** (1995). «Rasgos básicos de la economía española a través de la Contabilidad Nacional». *Papeles de economía española*, **62**.

[6] **Ruiz-Maya, L.** i **Martín Pliego, F.J.** (1995). *Estadística II. Inferencia*. Editorial AC, Madrid.

[7] **Scheffe, H.** (1959). *The analysis of variance*. John Wiley and Sons, New York.

ENGLISH SUMMARY:

REVISIONS OF NATIONAL ACCOUNTING ESTIMATES

Jesús Muñoz Malo, Ernest Pons Fanals and Jordi Pons Novell

The aim of this study is to analyze the revisions of the main macroeconomic variables prepared within the framework of national accounting, in order to make it possible to measure the reliability of these estimates.

It is divided into three different parts. The first one is a brief explanation of national accounting. This is a technique of statistical synthesis, the main aim of which is the description of the characteristics of an economy by means of a congruent set of accounting operations. The methodology is homogeneous for European countries and is based on the definitions, operations and accounts of the SEC (European integrated economic accounts system). The strict comparability of results between different countries is, however, influenced by the differences between the statistical system of each country.

By aiming to reconcile speed and rigour, the statistical institutes have to «qualify» their estimates in accordance with their provisional nature (Spanish accounting prolongs this period for 44 months). The different estimates for the same economic variable over this period demonstrate the scope of the revisions, the core around which the application of statistical comparisons is carried out in order to analyze these estimates.

The second chapter briefly introduces the problem of the reliability of national accounting. The scarce economic literature dealing with this subject mentions two methods, residual error and revisions. The former is based on the possibility of estimating gross domestic product using three independent methods (production, expenditure and income) and considers that the closeness of the different estimates made from independent statistical sources is a good indicator of the quality of the national accounts. This system is not operative because the estimates which are published are not the original ones but rather the reconciled ones, and, moreover, this method would only allow the evaluation of the aggregate result (GDP), but not of its components. The second method, with revisions, comes from the idea that the closeness of the first estimates to the definitive ones indicates a superior quality of estimate. Although it does not solve the essence of the problem of reliability, this method is an initial approach to validate whether the first accounting estimates are good approximations of the definitive results.

The third part includes the analysis of the revisions of Spanish national accounting for the period 1981-1992, starting from the GDP series and from their supply and demand components, at constant, current prices. The statistical techniques used are divided into three groups: dispersion measurements, analysis of temporary randomness and breakdown of the revisions.

The first group measures the dispersion between the first estimates and the definitive results by means of the typical deviation and the simple, crossed correlation coefficient. The different behaviour of the GDP and of its components stands out among the main results, given that the revisions of the GDP are very small while, on the contrary, all of its components display greater, in some cases very significant, revisions. The reason for this paradoxical situation is the existence of revisions with the opposite sign in the different components, which lead to the compensation of the revisions of the aggregate variable (GDP). This fact is confirmed by the crossed correlations between variables. By branches of activity, the stability of services contrasts with the considerable revisions of industry, construction and, above all, the agricultural sector. On the other hand, by components of demand, the variables have a more homogeneous behaviour.

A second group of techniques allows the analysis of the temporary randomness by means of three instruments: Ljung-Box's Q statistic, the run's test and Von Neuman's test. All three contrasts allow the hypothesis of temporary randomness of the revisions considered to be accepted.

Finally, through the analysis of the variance, the variation of the revisions is broken down into the following four factors: macromagnitude, year, current/constant prices and advance/first estimate. The result indicates the significance of the first two factors and the lack of significance of the others.

Biometria

QÜESTIIÓ, Vol 20, 2 pp. 327-346, 1996

SOBRE LA SIMULACIÓN DE PROCESOS DE EVOLUCIÓN MOLECULAR: CONSIDERACIONES SOBRE LA DERIVACIÓN Y CONTRASTACIÓN DE UN ESTIMADOR DE LA VARIANZA DE LA DIVERGENCIA NUCLEOTÍDICA A PARTIR DE FRAGMENTOS DE RESTRICCIÓN

SANTIAGO F. ELENA, ANDRÉS MOYA y
FERNANDO GONZÁLEZ CANDELAS

Una de las áreas de la Biología que ha experimentado mayor expansión en los últimos años es la investigación de procesos evolutivos mediante la aplicación de técnicas de la Biología Molecular. La puesta en marcha de programas de obtención de información cartográfica de genes y de secuencias de los mismos no ha hecho más que aumentar esta tendencia. La investigación de procesos evolutivos lleva emparejada el contraste de hipótesis alternativas, tales como el orden de división de los linajes en una reconstrucción filogenética o las estimas de distancias entre los nodos de un árbol filogenético. Existen modelos evolutivos que, bajo supuestos más o menos restrictivos, han permitido la construcción de tests paramétricos de contraste de hipótesis. Una de las dificultades que se encuentra en el desarrollo de estos tests es la derivación de las propiedades de los estadísticos correspondientes. En tales situaciones es muy frecuente recurrir a la simulación de procesos de evolución para, así, contrastar la bondad de los estadísticos. En este trabajo exponemos la derivación de un estimador de la varianza de la divergencia nucleotídica a partir de la comparación de los fragmentos producidos por la digestión con endonucleasas de restricción del DNA de especies descendientes de un ancestro común y los resultados de las simulaciones realizadas para comprobar su bondad, prestando especial atención al efecto que tienen sobre los resultados de la simulación distintos parámetros inicialmente considerados no relevantes y la importancia de establecer controles rigurosos e independientes sobre las simulaciones.

Santiago F. Elena, Andrés Moya y Fernando González Candelas. Departament de Genètica i Servei de Bioinformàtica. Facultat de Biología. Universitat de València «Estudi General». Dr. Moliner 50, 46100 Burjassot, València.

–Article rebut el setembre de 1995.

–Acceptat el setembre de 1996.

On the simulation of molecular evolutionary processes: considerations on the derivation and confirmation of an estimator for the variance of nucleotide divergence estimated from restriction fragments

Keywords: Simulation, nucleotide divergence, restriction fragment data, parameter estimation, delta method.

1. INTRODUCCIÓN

A lo largo del proceso evolutivo todo par de especies descendientes de un ancestro común acumulan cambios en sus moléculas de DNA que son, aproximadamente, proporcionales al tiempo transcurrido desde su divergencia. El estudio de estos cambios entre pares de secuencias contemporáneas es la base de una nueva metodología para la reconstrucción de procesos filogenéticos, disciplina conocida como Sistemática Molecular. Además, en cualquier población de individuos contemporáneos de cualquier especie se producen de forma constante nuevas mutaciones, material básico para la evolución, que quedan igualmente plasmadas como cambios en las moléculas de DNA de los distintos individuos. El estudio de las relaciones entre las distintas poblaciones de una especie, los patrones de intercambio genético entre ellas, la difusión de aquellas variantes especialmente favorables bajo circunstancias ambientales particulares, etc., estudios que habitualmente corresponden al ámbito de la Genética de Poblaciones, también pueden basarse en la comparación de los patrones de variación entre moléculas de DNA al nivel o niveles (individual, poblacional, regional, etc.) adecuados. De esta forma, la relación entre la Genética de Poblaciones y la Teoría de la Evolución se amplía desde la fundamentación teórica y formal que representa la primera para la segunda, hasta compartir una misma metodología experimental sobre la que contrastar y basar sus avances. Una excelente introducción a las distintas ramas de la Biología, tanto académicas como aplicadas, que utilizan estas técnicas de estudio de la variación genética a nivel molecular se halla en Avise (1994).

Para analizar estos cambios en las secuencias de DNA, el investigador dispone de diversas técnicas experimentales, de complejidad y coste crecientes aproximadamente a medida que aumenta la calidad de la información que proporcionan. Así, la secuenciación, es decir, la determinación de la secuencia de nucleótidos de parte de la molécula de DNA, proporciona el mayor grado de información posible, pero a costa de un mayor esfuerzo metodológico y económico y de una reducción muy drástica en la proporción de genoma estudiado. Con el fin de reducir los costes, así como para aumentar la fracción de genoma analizado aún a costa de perder parte de la información, se disponen de técnicas alternativas basadas, muchas de ellas, en la digestión con enzimas de restricción.

Una enzima de restricción es una proteína que reconoce una secuencia específica de DNA, normalmente de pequeña longitud (4 ó 6 nucleótidos) provocando un corte en la molécula de DNA nativa. Los fragmentos generados por este corte pueden ser separados mediante electroforesis en función de su tamaño y ser comparados con los producidos mediante el mismo procedimiento en otros individuos de la misma o distintas especies. Con la combinación adecuada de enzimas de restricción en una misma reacción de digestión es posible, además, establecer los puntos específicos de corte de cada uno de ellos a lo largo del genoma analizado. La primera de las técnicas se conoce como análisis del polimorfismo en la longitud de los fragmentos de restricción (RFLPs) mientras que la segunda se conoce como análisis de los sitios de restricción.

Otras técnicas de análisis de la variación en las secuencias de DNA que son en algunos aspectos asimilables a las anteriores utilizan secuencias más o menos largas de cebadores para amplificar fragmentos de la molécula original mediante la reacción en cadena de la polimerasa (PCR). La separación de estos fragmentos permite, de nuevo, la comparación entre distintas moléculas de DNA. Una de estas técnicas, que emplea cebadores relativamente cortos (10 pares de bases) que sirven para amplificar fragmentos aleatorios del DNA, es conocida como RAPDs-PCR (Hadrys, Balick y Schierwater, 1992). Esta técnica es una de las de utilización más frecuente en los últimos años.

Las diferencias genéticas obtenidas mediante RAPDs-PCR no son susceptibles del análisis cuantitativo necesario para obtener medidas de distancias filogenéticas entre taxones con divergencias medias (al menos intergenéricas) debido a diversos problemas ampliamente discutidos por Clark y Lanigan (1993) y por Lynch y Milligan (1994). No obstante, es posible realizar comparaciones filogenéticas entre poblaciones de una misma especie o entre especies muy próximas entre sí, en el ámbito de lo que Avise (1994) define como Filogeografía. En lo fundamental, la estimación de la divergencia nucleotídica a partir del análisis de bandas de RAPDs compartidas entre dos genomas es formalmente similar a la propuesta por Nei y Li (1979) para estimar ese mismo valor mediante el análisis de fragmentos de restricción. Esta similitud fue puesta de manifiesto por Clark y Lanigan (1993).

Dada la incertidumbre asociada al proceso de estimación de la divergencia nucleotídica, se hace necesario disponer de una buena estimación de la varianza del estimador de la divergencia nucleotídica empleado. De los diversos estimadores propuestos para datos obtenidos mediante análisis de los fragmentos de restricción, por comparación sólo del tamaño de los mismos y no de los sitios donde se producen los cortes, el más frecuentemente empleado es el desarrollado por Nei y Li (1979). Nei y Miller (1990) propusieron un método de remuestreo mediante «bootstrap» para obtener una estimación de la varianza de ese estimador, mientras que González-Candelas, Elena y Moya (1995) han propuesto el uso de un estimador analítico.

2. EL MODELO DE CAMBIO EVOLUTIVO EN LOS SITIOS DE RESTRICCIÓN

Nei y Li (1979) desarrollaron un método para estimar el número, d, de sustituciones nucleotídicas por sitio entre dos secuencias de DNA cuando los datos de que se dispone son los fragmentos de restricción. Para ello se basaron en el siguiente modelo de cambio evolutivo de los sitios de restricción.

Sea $n(t)$ el número de sitios de restricción en la molécula (o porción de ella) de DNA considerada en el instante t y sea $n(0) = n_0$. Se supone

(1) que el contenido esperado G+C permanece constante y

(2) que la sustitución de nucleótidos se produce de forma aleatoria según un proceso Poisson con tasa de sustitución λ por unidad de tiempo (año o generación).

A medida que transcurre la evolución, algunos de los sitios de restricción originales desaparecerán mientras que aparecerán otros nuevos. Denotemos por $n_1 t$ y $n_2 t$ respectivamente esos valores. En ese caso podemos escribir $n(t) = n_1(t) + n_2(t)$. La probabilidad de que un sitio original permanezca inalterado al cabo de un tiempo t viene dada por $P = e^{-r\lambda t}$, por lo que la esperanza de $n_1(t)$ es $n_0 \cdot e^{-r\lambda t}$.

Para obtener la esperanza de $n_2(t)$ pensemos que el enzima de restricción considerado reconoce una secuencia de r nucleótidos (habitualmente $r = 4$ ó 6). La probabilidad de que esta secuencia haya sufrido uno o más cambios en el tiempo t es $1 - P$ y sea a la probabilidad de que la nueva secuencia generada sea un sitio de restricción. El valor de a está relacionado con la distribución de frecuencias en el equilibrio de los cuatro nucleótidos en la molécula de DNA correspondiente y en la proporción con que aparecen los mismos en la secuencia diana de la enzima de restricción. Si en la molécula de DNA considerada existen m_T secuencias posibles de longitud r, entonces la esperanza de $n_2(t)$ es $m_T a(1 - P)$. En ese caso, el valor esperado, $E[n]$, de $n(t)$ es

$$E[n] = n_0 P + m_T a(1 - P). \tag{1}$$

Su varianza puede obtenerse teniendo en cuenta que n_1 se distribuye binomialmente y que n_2 sigue una distribución Poisson, siendo ambos valores independientes:

$$\mathrm{Var}[n] = n_0 P(1 - P) + m_T a(1 - P). \tag{2}$$

Consideremos ahora la divergencia entre dos linajes evolutivos o poblaciones X e Y. Asumimos que sus DNAs se derivan de una secuencia ancestral común a partir del instante 0. Sean n_{X1} y n_{X2} el número de sitios de restricción ancestrales y el de

nuevos sitios, respectivamente, en el linaje X, con $n_X = n_{X1} + n_{X2}$, y sean n_{Y1}, n_{Y2} y n_Y los valores correspondientes en el linaje Y. Sea n_{XY} el número de sitios idénticos compartidos por los dos linajes. Dada la baja probabilidad de que se produzcan *de novo* y de forma independiente dos nuevos sitios de restricción idénticos en los linajes X e Y, consideraremos que todos los sitios compartidos se derivan de sitios presentes ya en la secuencia ancestral común. Bajo este supuesto n_{XY} sigue una distribución binomial cuya media y varianza vienen dadas por n_0P^2 y $n_0P^2(1-P^2)$, respectivamente.

Consideremos ahora la relación entre los fragmentos de restricción compartidos entre las dos especies. Para que un fragmento de DNA se conserve a lo largo de t generaciones se precisan dos condiciones:

(1) los dos sitios flanqueantes al fragmento deben permanecer inalterados y

(2) no debe aparecer ningún nuevo sitio de restricción en su interior.

La probabilidad del primer suceso es obviamente P^2, y la del segundo es $(1-b)^{(m-r+1)}$, donde $b = a(1-P)$ y m es el número de nucleótidos en ese fragmento. Teniendo en cuenta que para que el fragmento siga siendo compartido las anteriores condiciones deben cumplirse en los dos linajes considerados, la proposición de fragmentos compartidos por ambos será

$$F = \left(\frac{1}{n_0}\right) \sum_{i=1}^{n_0} p^4 (1-b)^{2(m_i - r+1)} \tag{3}$$

Ahora bien, en la práctica no puede aplicarse esta fórmula porque se desconocen tanto n_0 como m_i. No obstante, bajo ciertos supuestos simplificadores adicionales (Nei y Li, 1979), se puede derivar la siguiente aproximación

$$F \approx \frac{P^4}{3-2P} \tag{4}$$

Usando $P = e^{-r\lambda t}$ y $d = 2\lambda t$, se establece una relación entre F y d. El número d representa el número esperado de sustituciones nucleotídicas entre los dos linajes al cabo de un tiempo t.

El estimador máximo verosímil de F es

$$\hat{F} = \frac{2m_{XY}}{m_X + m_Y} \tag{5}$$

donde m_X y m_Y representan el total de fragmentos de restricción observados en las secuencias X e Y, respectivamente, y m_{XY} representa el número de tales fragmentos que son compartidos por ambas secuencias.

Para poder realizar inferencias sobre la igualdad o no de dos estimaciones de d, es necesario disponer de una estimación de sus errores respectivos. En el presente trabajo, en primer lugar, describimos el desarrollo de una expresión analítica aproximada del estimador de $\mathrm{Var}(\hat{d})$, a partir de un estimador ya conocido de la varianza de $\hat{F}$ (Nei y Tajima, 1981) y, en segundo lugar, comprobamos la validez de esta expresión mediante una simulación numérica, comparando los resultados con los obtenidos para otros estimadores.

Nei y Tajima (1981) derivaron la siguiente expresión para la varianza muestral de $\hat{F}$:

$$\widehat{\mathrm{Var}}(\hat{F}) = \frac{1}{\overline{m}}\left\{\hat{F}\left(1-\hat{F}\right) - \hat{F}^2\left(1-\sqrt{\hat{F}}\right)\left[1+\frac{1}{2}\left(1-\sqrt{\hat{F}}\right)\right]\right\} \tag{6}$$

donde $\overline{m} = \dfrac{m_X + m_Y}{2}$ es el número promedio de fragmentos de restricción observados en las dos secuencias analizadas. Este valor es estimador del número de fragmentos de restricción presentes originalmente en la secuencia ancestral.

3. DESARROLLO DEL ESTIMADOR DE VARIANZA DE $\hat{d}$

Asumimos que se ha obtenido una estimación empírica de F a partir de la digestión con enzimas de restricción de dos secuencias X e Y y que los fragmentos obtenidos son separados y comparados por tamaños (lo que nos permite obtener los valores m_X, m_Y y m_{XY} antes indicados). A partir de esa estimación, utilizando la ecuación (4) se obtiene por iteración una estimación de P, con lo que finalmente se puede estimar la divergencia nucleotídica, $\hat{d}$, entre ambas secuencias.

Los detalles de la derivación del estimador de la varianza del estimador de d pueden encontrarse en González-Candelas, Elena y Moya (1995). No obstante, delineamos aquí las ideas generales empleadas. Aplicamos en varias ocasiones la fórmula aproximada de Fisher para la obtención de la varianza de un parámetro o método delta (Fisher, 1925), reteniendo hasta el momento de tercer orden en la expansión de Taylor de (1) para aumentar la precisión de la estimación. Utilizamos la expresión (4) antes mencionada para obtener una estimación de la varianza del estimador de F a partir de los datos empíricos antes mencionados. Además, tenemos en cuenta que $\mathrm{E}(m_X) = \mathrm{E}(m_Y) = \overline{m}P$, que $\mathrm{E}(m_{XY}) = \mathrm{E}(m_{XY}\, m_X) = \mathrm{E}(m_{XY}\, m_Y) = \overline{M}P^2$, que $\mathrm{Cov}(m_{XY}\, m_X) = \mathrm{Cov}(m_{XY}\, m_Y) = \overline{m}P^2(1-P)$, y que $\mathrm{Cov}(m_X, m_Y) = 0$, siendo P la probabilidad de que un fragmento dado aparezca compartido por un par de secuencias y $\overline{m}$ el número de fragmentos de restricción en la secuencia ancestral (Nei y Li, 1979).

Con todo ello se obtiene la siguiente expresión para el estimador de la varianza del estimador de d:

$$\widehat{\mathrm{Var}}(\hat{d}) = \frac{(3-2\hat{P})^4}{9r^2\hat{P}^8(2-\hat{P})^2}\left[\mathrm{Var}(\hat{F}) - \frac{72-117\hat{P}+64\hat{P}^2-12\hat{P}^3}{6\hat{P}^4(2-\hat{P}^2)}\mu_3(\hat{F})\right] \tag{7}$$

donde

$$\mu_3(\hat{F}) = \frac{\hat{F}\left(1-\hat{F}\right)\left(1-2\hat{F}\right)}{\overline{m}^2} - \frac{\hat{F}^3\left\{\overline{m}\sqrt{\hat{F}}\left(1-\sqrt{\hat{F}}\right)\left(1-2\sqrt{\hat{F}}\right)+\sqrt[4]{\overline{m}\left(1-\sqrt{\hat{F}}\right)}\right\}}{4\overline{m}^3} -$$

$$-\frac{3\hat{F}^2\left(1-\overline{m}\sqrt{\hat{F}}-2\overline{m}\hat{F}+2\overline{m}^2\sqrt{\hat{F}^3}\right)}{4\overline{m}^4} \tag{8}$$

Para ciertas aplicaciones nos bastará con tener una estimación de la varianza del estimador de d, pero en otras ocasiones estamos interesados en realizar un constraste de hipótesis directamente sobre el valor estimado. Para la construcción del test es imprescindible conocer la distribución del estimador, y no sólo su varianza, lo que es imposible. No obstante, podemos suponer que las estimaciones de d se comportan de forma asintóticamente normal, por ser transformaciones suaves de estimadores de máxima verosimilitud. Supongamos que se han calculado las estimaciones $\hat{d}_a$ y $\hat{d}_b$ y sus varianzas respectivas a partir de m_a($m_a = m_X + m_Y - m_{XY}$, si consideramos las secuencias X e Y según lo indicado anteriormente) y m_b fragmentos, respectivamente, entre pares de secuencias independientes. En ese caso, el estadístico

$$t'_s = \frac{\hat{d}_a - \hat{d}_b}{\sqrt{\dfrac{\widehat{\mathrm{Var}}(\hat{d}_a)}{m_b} + \dfrac{\widehat{\mathrm{Var}}(\hat{d}_b)}{m_a}}} \tag{9}$$

se distribuye aproximadamente como una t de Student con $m_a + m_b - 2$ grados de libertad.

4. COMPROBACIÓN POR SIMULACIÓN DE LA BONDAD DEL ESTIMADOR DE $\mathrm{Var}(\hat{d})$

Para comprobar la bondad del procedimiento de derivación y para comprobar las estimaciones de la varianza del estimador de divergencia nucleotídica obtenidas

por los análisis de fragmentos de restricción, se hicieron simulaciones siguiendo el procedimiento descrito por Li (1981). Junto a la evaluación del estimador desarrollado, hemos procedido a comprobar la bondad tanto del estimador de divergencia nucleotídica basado en sitios de restricción (Nei y Li, 1979) como del calculado directamente sobre la secuencia nucleotídica (Jukes y Cantor, 1969). Al realizar estas comparaciones pretendíamos establecer controles adicionales sobre el proceso que se estaba simulando, dada la conocida validez de estos estimadores.

Se simuló la evolución de una secuencia de DNA que da lugar a dos secuencias derivadas para distintos valores de la tasa de divergencia. En este modelo solamente se ha considerado la aparición de mutaciones por sustitución nucleotídica y no por deleciones o inserciones. En la simulación se consideró una única tasa de sustitución nucleotídica, la misma para cambios transicionales que para transvesiones, por lo que se sigue el modelo evolutivo propuesto por Jukes y Cantor (1969) para considerar la superposición de sustituciones en una misma posición. Este modelo, sólo es aceptable para valores bajos de divergencia ($d < 0.1$), lo que corresponde con las condiciones para las que es adecuado estimar la divergencia a partir de fragmentos de restricción (Nei y Li, 1979; Li, 1981). Tras la simulación del proceso evolutivo, se procedió a simular la digestión con enzimas de restricción y a comparar los fragmentos resultantes de las mismas.

Los análisis de simulación se basaron en tres secuencias nucleotídicas aleatorias de longitudes 1000, 10000 y 100000 pares de bases respectivamente, de composición nucleotídica equiprobable. Las tasas de evolución, $2\lambda t$, variaron en el intervalo 0.002 y 0.10 sustituciones por nucleótido y unidad de tiempo. Para cada valor de la tasa de evolución y longitud del genoma, se realizaron 2000 réplicas. Todo el proceso de simulación fue implementado en un programa escrito en Pascal estándar que se ejecutó en una estación de trabajo DEC-AXP 3000-400 en entorno Open VMS.

Se simularon dos procesos de digestión diferentes, uno empleando un enzima que reconoce dianas de 4 nucleótidos ($r = 4$) y otro con diez enzimas que reconocen dianas de 6 nucleótidos ($r = 6$). Las secuencias diana de las enzimas de restricción empleadas se generaron también aleatoriamente al principio de cada simulación. El mismo procedimiento de simulación se realizó también manteniendo constantes las enzimas de restricción empleadas en todos los casos, obteniéndose resultados semejantes a los anteriores y que no son detallados aquí. La divergencia entre cada par de secuencias se estimó por tres vías:

(i) a partir de la longitud de los fragmentos obtenidos en ambas restricciones empleando las ecuaciones descritas en el apartado 2,

(ii) a partir de datos de sitios de restricción usando la ecuación 5.42 de Nei (1987, página 101) y

(iii) a partir de las secuencias nucleotídicas empleando el estimador de Jukes y Cantor (1969). Las varianzas para cada estimación de divergencia entre pares de secuencias se obtuvieron de acuerdo con la ecuación (7) para datos de fragmentos de restricción, según la ecuación 5.45 de Nei (1987, página 101) para datos de sitios de restricción y según la ecuación 5.4 de Nei (1987, página 66) para los datos directos de secuencia nucleotídica.

5. RESULTADOS DE LAS SIMULACIONES

Los valores de las divergencias promedio estimados directamente a partir de los datos de secuencia, con la corrección de Jukes-Cantor, y las estimaciones de sus varianzas se encuentran en la figura 1. Los datos estimados mediante el uso de sitios de restricción se muestran en las figuras 2 (para un enzima con $r = 4$) y 3 (para 10 enzimas con $r = 6$), mientras que los valores estimados a partir de los fragmentos de restricción se muestran en las figuras 4 y 5, respectivamente. En los dos últimos casos se muestra en cada figura tanto el valor promedio de las 2000 varianzas proporcionadas por las estimaciones en cada réplica, como la varianza de las 2000 estimaciones de divergencia nucleotídica calculadas usando sitios y fragmentos de restricción para cada valor de la tasa de evolución.

La fiabilidad del proceso evolutivo simulado para todas las tasas y longitudes de secuencia puede ser apreciada a partir de las estimaciones de divergencia nucleotídica obtenidas mediante el estimador de Jukes y Cantor (figura 1). En la figura se pone de manifiesto la utilidad del método de Monte-Carlo empleado para comprobar la validez de la expresión 5.4 dada por Nei (1987, página 66) para estimar la varianza del estimador de divergencia nucleotídica a partir de las secuencias. Teniendo en cuenta la validez del proceso evolutivo simulado y del método de contraste empleado, las estimaciones de divergencia evolutiva a partir de sitios y fragmentos de restricción muestran las siguientes características.

Para datos de sitios de restricción, se obtuvieron mejores estimaciones usando un sólo enzima de $r = 4$ que usando 10 enzimas de $r = 6$. Esta observación es válida para todas las longitudes de secuencia empleadas en la simulación, aunque generalmente se obtienen mejores estimaciones empleando secuencias más largas. Empleando un enzima de $r = 4$, y para $L = 1000$ y $L = 10000$ con $2\lambda t \leq 0.07$, generalmnte se obtiene una ligera sobreestimación de $\mathrm{Var}(\hat{d})$ empleando la ecuación 5.45 de Nei (1987). Para $L = 100000$ y para cualquier tasa de evolución, así como para $L = 10000$ y tasas $2\lambda t \geq 0.07$, es habitual encontrar subestimaciones. Cuando se usan 10 enzimas de $r = 6$, las estimaciones de $\hat{d}$ son aceptables únicamente cuando

$L = 100000$, obteniéndose una clara subestimación para los demás valores de L. Estas subestimas no pueden ser atribuidas a un error en la simulación, como se demostró en el párrafo anterior al coincidir los valores estimados mediante secuencia con la corrección de Jukes y Cantor. No existe un patrón claro para las correspondientes varianzas, donde se pueden observar pequeñas desviaciones entre las estimaciones y los valores obtenidos en las simulaciones para todas las longitudes y tasas de evolución empleadas.

El uso de datos de longitud de fragmentos de restricción para estimar divergencias nucleotídicas presenta algunos problemas. Primero, las divergencias nucleotídicas pueden ser estimadas correctamente a partir de la combinación adecuada de longitudes de secuencia y número de fragmentos. En nuestro caso, esto se consigue usando o bien $L = 10000$ y un solo enzima de $r = 4$ o bien $L = 100000$ y diez enzimas de $r = 6$. Todas las demás combinaciones de longitud de secuencia y enzimas originan series subestimaciones de la divergencia simulada. Para secuencias de longitud pequeña o para un número de fragmentos generados también pequeño, esto puede ser debido al pequeño número de fragmentos compartidos que aparecen. Para el caso de secuencias largas, la explicación reside en la redundancia de fragmentos no homólogos con igual longitud generados cuando $2\lambda t \geq 0,02$. Para los dos casos en los que las estimas de divergencia fueron aceptables ($L = 1000, r = 4$ y $L = 100000, r = 6$), las estimas de varianza obtenidas usando la ecuación (7) siempre subestiman las varianzas obtenidas mediante las réplicas de Monte-Carlo (figuras 4 y 5). Esto produce un aumento en la probabilidad de error de tipo I si se acepta que las varianzas obtenidas a partir de las 2000 estimas de d son una buena aproximación al valor real de la varianza.

En resumen, para las tasas de divergencia analizadas y el modelo evolutivo empleado, la estimación de la divergencia mediante el estimador de Jukes-Cantor para datos de secuencia es prácticamente exacta. Para datos de sitios de restricción, las estimaciones son adecuadas cuando se emplea un enzima de $r = 4$ y no lo son cuando se usan 10 enzimas con $r = 6$ excepto para $L = 100000$. El mismo patrón se observa para las estimaciones de divergencia cuando se emplean las longitudes de los fragmentos originados en las digestiones, si bien con una ligera pérdida de precisión respecto a la estimación con sitios. Para las estimaciones de las varianzas de los correspondientes estimadores se observa un buen ajuste en el caso de sitios de restricción pero una ligera sobrestima cuando se emplea el estimador desarrollado por González-Candelas, Elena y Moya (1995) para el caso de fragmentos de restricción.

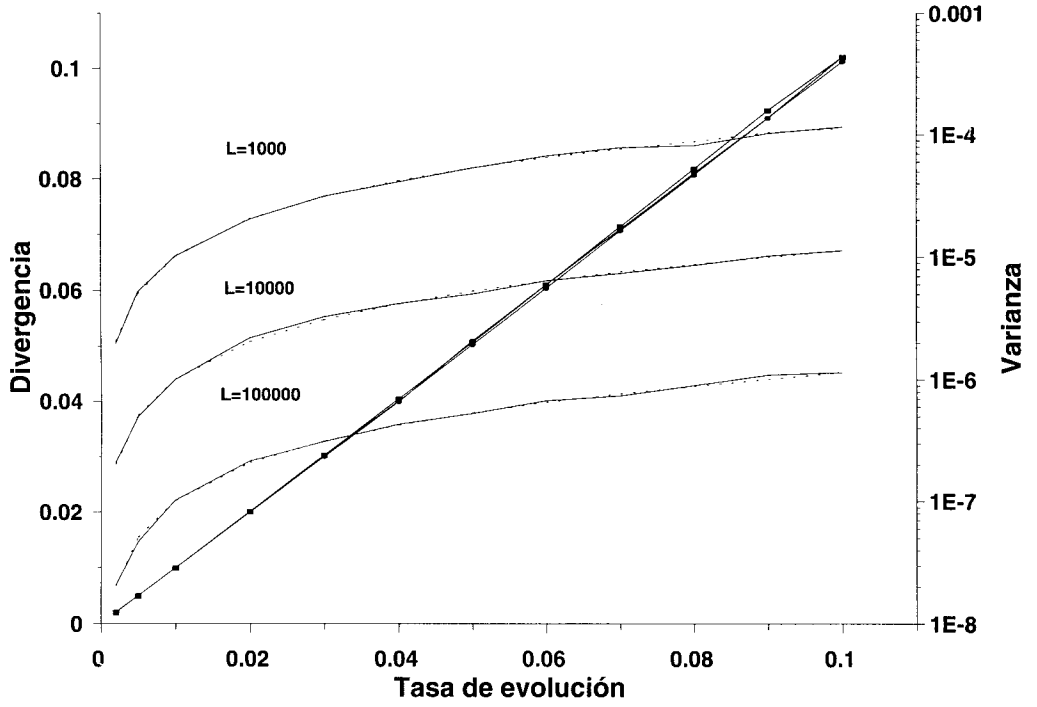


Figura 1

Resultados de la simulación para el análisis directo de la secuencia nucleotídica utilizando el método de Jukes-Cantor.

(a) En el eje de ordenadas principal se muestran las divergencias estimadas para los tres valores de longitud de secuencia considerados (L = 1000, L = 10000 y L = 100000). En este caso no se muestra el valor esperado de la divergencia por ser prácticamente coincidente con el obtenido en las simulaciones.

(b) En el eje de ordenadas secundario (escala logarítmica) se muestran las estimaciones de las varianzas del estimador de divergencia nucleotídica según la ecuación 5.4 dada por Nei (1987). En todos los casos se dan los resultados de la simulación para las tres longitudes de secuencia estudiadas (líneas continuas) y junto a ellas los valores de varianza muestral obtenidos en las correspondientes simulaciones (líneas punteadas).

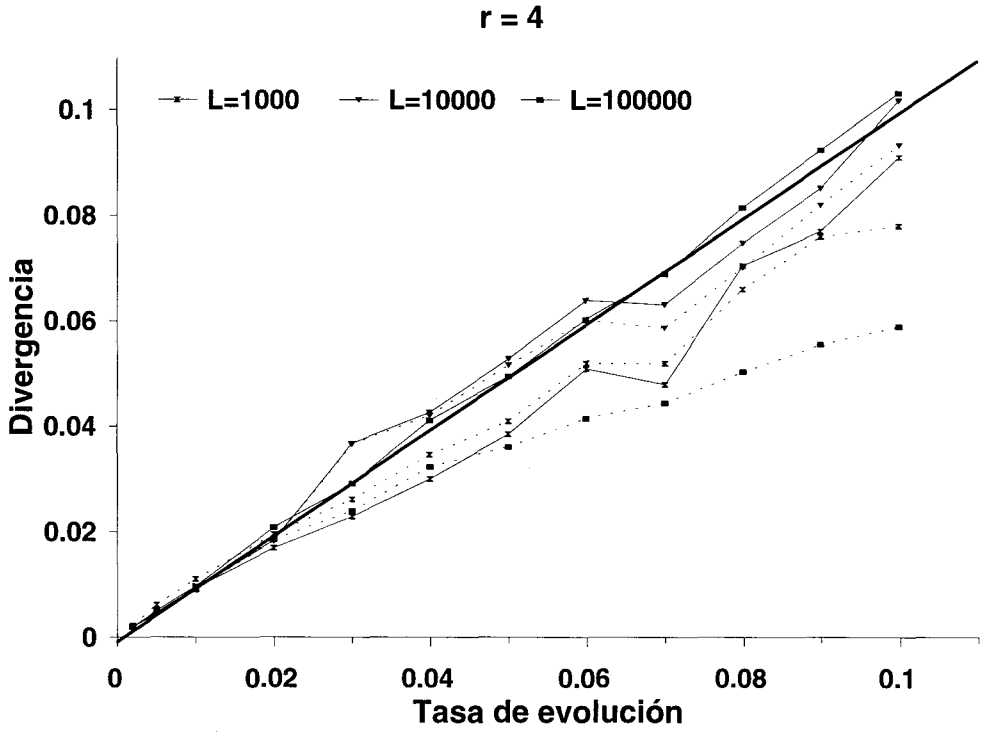


Figura 2

Resultados de la simulación para el análisis de sitios de restricción empleando 1 solo enzima que reconoce dianas de 4 nucleótidos.

(a) Líneas continuas: divergencias estimadas mediante el análisis de sitios de restricción.

(b) Líneas punteadas: divergencias estimadas mediante el análisis de fragmentos de restricción. En todos los casos se muestran los resultados para las tres longituds genómicas estudiadas. La línea gruesa bisectriz indica el valor esperado de la divergencia nucleotídica.

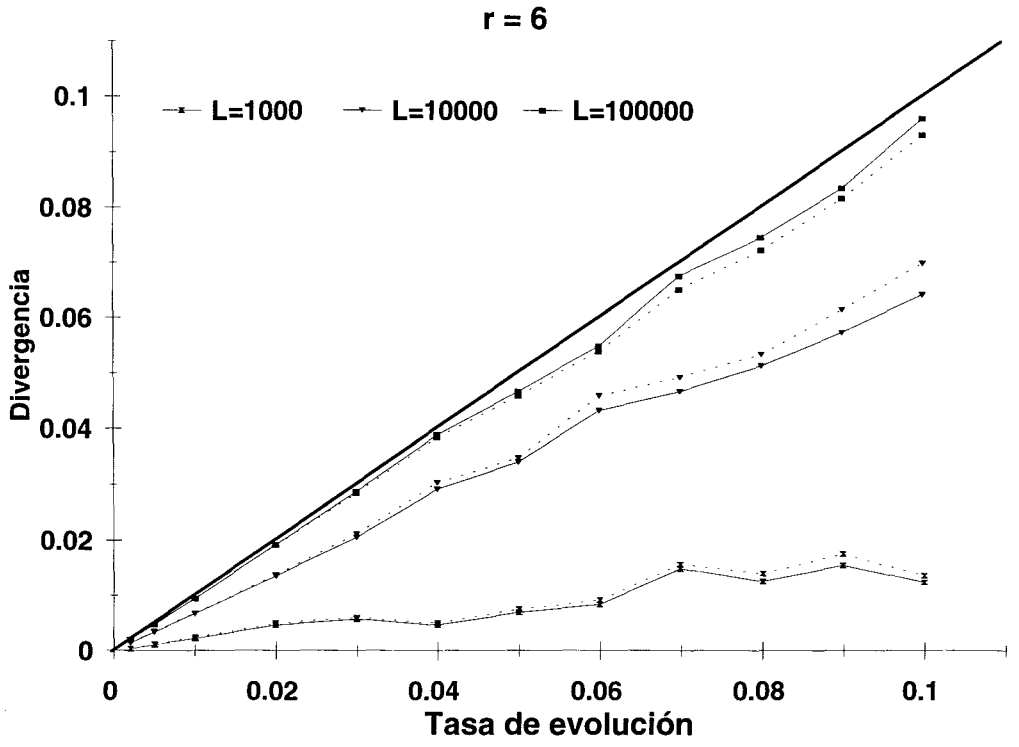


Figura 3

Igual que la figura 2, pero empleando 10 enzimas que reconocen una diana de 6 nucleótidos de longitud.

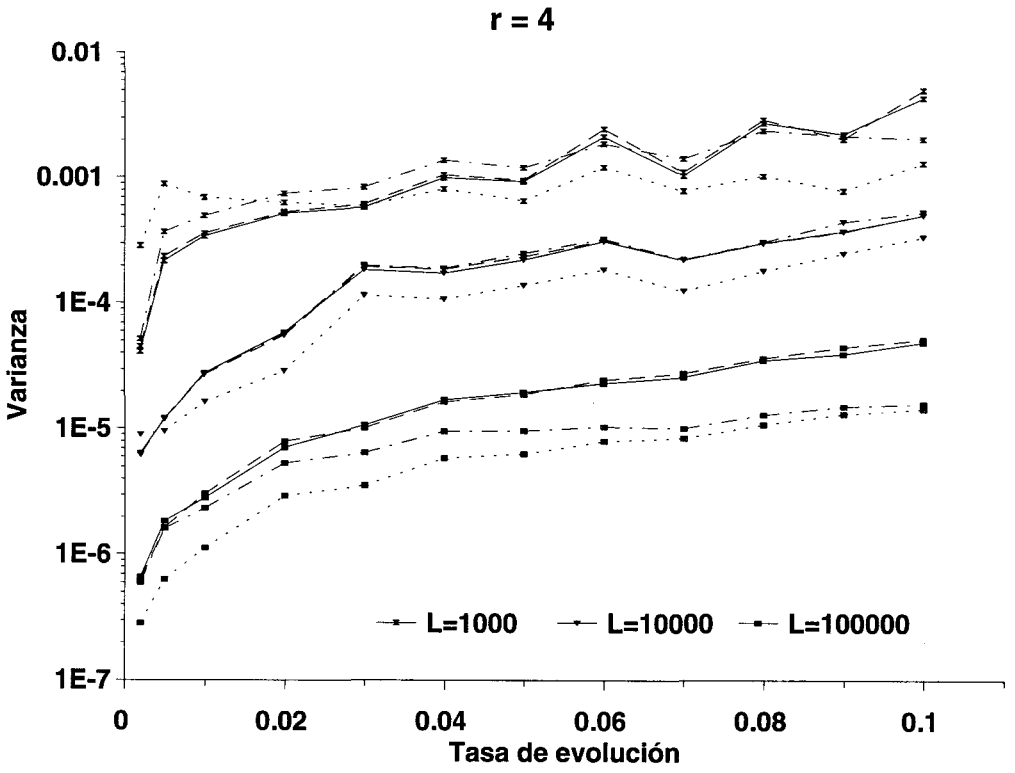


Figura 4

Estimaciones de la varianza de los estimadores de divergencia para sitios y fragmentos de restricción según la ecuación 5.45 de Nei (1987) y la ecuación (7), respectivamente tras la simulación por el método de Monte-Carlo. Para cada longitud de genoma considerado, los puntos unidos por una línea continua representan las estimaciones de la varianza para el estimador de divergencia a partir de sitios de restricción obtenidos a partir de las simulaciones y la línea discontinua el promedio de las 2000 estimaciones correspondientes según la ecuación 5.45 de Nei (1987). De igual forma, las líneas que alternan puntos y rayas corresponden a la varianza entre las 2000 réplicas del estimador de divergencia nucleotídica a partir de fragmentos de restricción y las líneas punteadas corresponden al promedio de las 2000 estimaciones de la varianza según la ecuación (7).

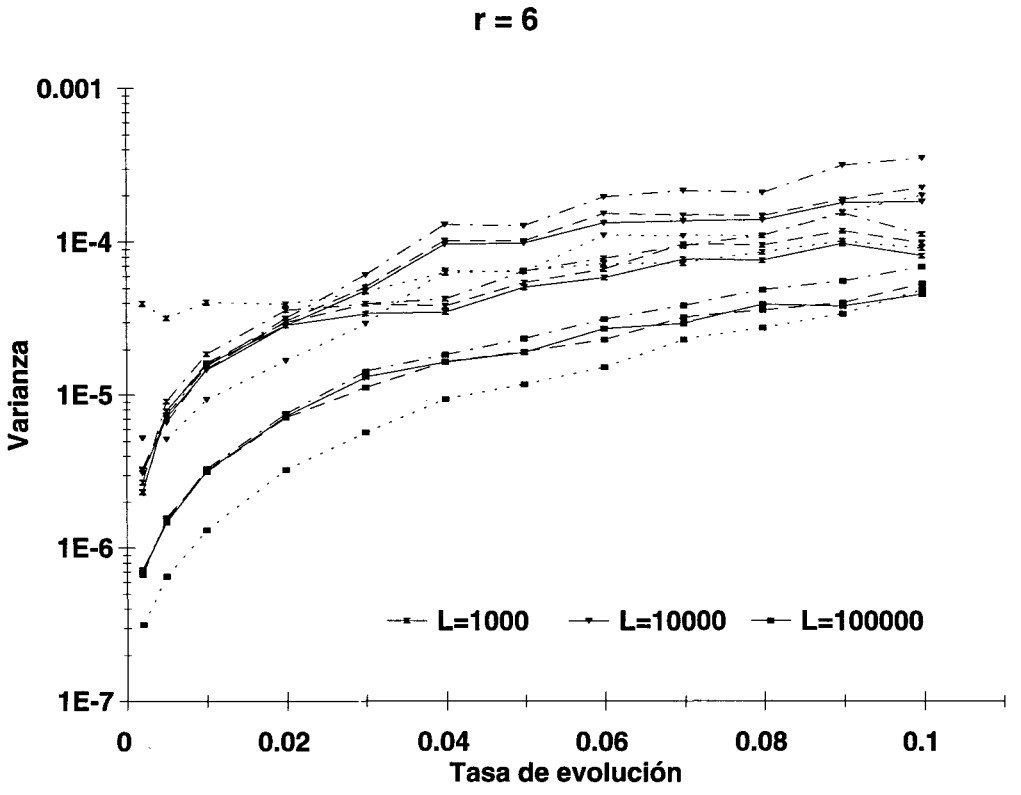


Figura 5

Igual que la figura 4, pero empleando 10 enzimas que reconocen una diana de 6 nucleótidos de longitud.

7. AGRADECIMIENTOS

Estamos en deuda con los Drs. J. Ferrándiz y M. Sendra por sus valiosos comentarios y sugerencias, con el Dr. W.-H. Li por habernos incitado con sus sugerencias a profundizar en nuestro análisis y con dos revisores de este artículo por sus indicaciones que, en todo caso, han contribuido a mejorarlo. S.F.E. ha sido becario predoctoral de la Consellería d'Educació i Ciència de la Generalitat Valenciana. Este trabajo ha sido subvencionado por los proyectos PB93-0690 y PB93-0350 de la DGICYT.

8. BIBLIOGRAFÍA

[1] **Avise, J.C.** (1994). *Molecular Markers, Natural History and Evolution.* Chapman & Hall, New York.

[2] **Clark, A.** y **Lanigan, C.M.S.** (1993). «Prospects for estimating nucleotide divergence with RAPDs». *Mol. Biol. Evol.*, **10**, 1096–1111.

[3] **Fisher, R.A.** (1925). *Statistical methods for research workers.* Edición 13ª. Hafner, New York.

[4] **González Candelas, F., Elena, S.F.** y **Moya, A.** (1995). «Approximate variance of nucleotide divergence between two sequences estimated from restriction fragment data». *Genetics*, **140**, 1443–1446.

[5] **Hadrys, H., Balick, M.** y **Schierwater, B.** (1992). «Applications of random amplified polymorphic DNA (RAPD) in molecular ecology». *Mol. Ecol.*, **1**, 55–63.

[6] **Jukes, T.H.** y **Cantor, C.R.** (1969). «Evolution of protein molecules». En *Evolution of genes and proteins*, 191–207. Sunderland MA, Sinauer Associates.

[7] **Li, W.-H.** (1981). «A simulation study on Nei and Li's model for estimating DNA divergencie from restriction enzyme maps». *J. Mol. Evol.*, **17**, 251–255.

[8] **Lynch, M.** y **Milligan, B.G.** (1994). «Analyzing population genetic structure with RAPD markers». *Mol. Ecol.*, **3**, 91–100.

[9] **Nei, M.** (1987). *Molecular Evolutionary Genetics.* Columbia University Press, New York.

[10] **Nei, M.** y **Li, W.-H.** (1979). «Mathematical model for studying genetic variation in terms of restriction endonucleases». *Proc. Natl. Acad. Sci. USA*, **76**, 5269–5273.

[11] **Nei, M.** y **Miller, J.C.** (1990). «A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data». *Genetics,* **125**, 873–879.

[12] **Nei, M.** y **Tajima, F.** (1981). «DNA polymorphism detectable by restriction endonucleases». *Genetics,* **97**, 145–163.

ENGLISH SUMMARY:

ON THE SIMULATION OF MOLECULAR EVOLUTIONARY PROCESSES: CONSIDERATIONS ON THE DERIVATION AND CONFIRMATION OF AN ESTIMATOR FOR THE VARIANCE OF NUCLEOTIDE DIVERGENCE ESTIMATED FROM RESTRICTION FRAGMENTS

Santiago F. Elena, Andrés Moya and Fernando González-Candelas

1. INTRODUCTION

In the last few years an increasing number of studies that make use of DNA fragments to characterize genetic divergences between individuals, populations, and species are being employed. This is the case, for instance, of RFLPs or RAPDs-PCR and related methods. In order to make statistical inferences about the equality or not of two divergence estimates, it is necessary to know its variance. Nei and Li (1979) developed an statistical method to obtain the divergence between two DNA sequences using this kind of information. However, they did not develop an estimate for the variance of their divergence estimate. In the present work, we have developed such an estimator.

We have started from the expression for the divergence derived by Nei and Li (1979). Applying Fisher's delta method (Fisher, 1925) over their expressions and retaining up to third order moments in the Taylor's expansion we have derived an approximate estimator for the variance of the nucleotide divergence (equation 7).

In order to test the accuracy of our expression, a computer simulation following the procedure described by Li (1981) has been carried out. The simulation analyses were based on three random DNA sequences of different lengths. These sequences were made to evolve, allowing only for nucleotide substitutions, giving rise 2000 pairs of derived sequences for each evolutionary rate. The resulting pairs were compared by three different methods:

(i) Using the Jukes-Cantor correction for estimating nucleotide divergence,

(ii) using Nei's model for divergence estimation with restriction sites data (Nei, 1987) and

(iii) using only restriction fragment length data. The first method gave us an idea of the reliability of the simulated process: we found a good agreement between expected and predicted divergence values and also for their variances.

For restriction site data, better estimates are obtained using one single four cutter than ten six-cutter enzymes. This is independent of the length of the sequence used in the simulation process. The use of fragment lengths of the sequence used in the simulation process. The use of fragments lengths presents several problems. For instance, nucleotide divergences can only be estimated reliably from a combination of the right sequence length and number and type of enzymes, although underestimates have usually been found.

The need for exhaustive controls and exploration of all parameters, independently of their incorporation into the final expressions, during simulations aimed at checking analytical derivations is discussed.

SECCIÓ DOCENT I PROBLEMES

La "SECCIÓ DOCENT I PROBLEMES" a la revista QÜESTIIÓ té l'objectiu d'incloure una secció on es publiquen articles de caire docent, difícilment publicables en revistes de recerca. A cada número de QÜESTIIÓ s'inclourà d'un a tres problemes i les solucions es donaran en el número següent.

Els lectors poden, si ho volen, proposar problemes amb les solucions pertinents i enviar-los a QÜESTIIÓ, que farà una selecció i en publicarà els més adequats, fent la corresponent referència a l'autor.

També seran ben rebudes solucions alternatives a les propostes fetes per l'autor dels problemes; l'editorial es reservarà, però, el dret a publicar-les.

PROBLEMES PROPOSATS

PROBLEMA N° 60

Dado el estimador r_u de la razón de totales $R = X/Y$, definido por:

$$r_u = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{y_i} = \frac{1}{n}\sum_{i=1}^{n} r_i \qquad (r_i = x_i/y_i)$$

siendo n el tamaño muestral, x_i e y_i el valor de la variable de interés y auxiliar en la i–ésima unidad seleccionada con probabilidades proporcionales a y_i con reemplazamiento de una población finita de tamaño N. Demostrar que:

a) r_u es insesgado para R.

b) $V(r_u) = \frac{1}{n}\sum_{i=1}^{N}(r_i - R)^2\frac{y_i}{Y}$.

c) Un estimador insesgado de $V(r_u)$, es

$$s^2(r_u) = \hat{V}(r_u) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(r_i - r_u)^2.$$

M. Ruiz Espejo

UNED

PROBLEMA N° 61

a) Probar que todo esquema muestral de «probabilidades de inclusión proporcionales al tamaño» (IPPS) normalizado p_i (> 0, para todo $i = 1,2,\ldots,N$), siendo

$$\sum_{i=1}^{N} p_i = 1,$$

de tamaño muestral $n = 2$ y cuya probabilidad de primera selección es p_i para la unidad $i(= 1,2,\ldots,N)$, es sin reemplazamiento.

b) Dar las condiciones necesarias y suficientes para que exista un único esquema muestral IPPS de tales características con $n = 2$, para poblaciones finitas de tamaño $N = 3$.

M. Ruiz Espejo

UNED

QÜESTIIÓ, Vol 20, 2 pp. 351-369, 1996

ELECCIONS A GIRONA: UN EXEMPLE D'ESTUDI D'UNA TAULA TERNÀRIA

F. BORRELL THIÓ

I.B. Salvador Espriu - Salt

L'article mostra una part dels resultats obtinguts en una aplicació de l'Anàlisi Factorial Múltiple en l'estudi d'una taula ternària de contingència. No es preten donar una interpretació exhaustiva dels resultats electorals, sinó resaltar que, en aquest estudi, els dos tipus de inèrcia en què es descomposa la inèrcia total de la taula (Inèrcia INTER *i Inèrcia* INTRA*) són quasibé ortogonals, la qual cosa facilita molt la seva interpretació. Per aquesta raó aquest exemple es pot utilitzar dins del context docent de l'Anàlisi de dades.*

Elections in Girona: an example of a study of a ternary table

1. INTRODUCCIÓ

El present article va sorgir d'una aplicació de l'Anàlisi Factorial Multiple en l'estudi d'una taula ternària de contingència (Ecofier, B. - Pagès, J. 1992), en què es creuen 3 variables qualitatives que anomenarem I, J i T. Una manera d'analitzar aquests tipus de taules és considerar les dades com una successió de taules de freqüència binàries, que resulten del creuament de dues de les variables, I i J, indexades per l'altre variable T que juga un paper diferent.

L'inconvenient més gran que presenta aquest enfoc és el fet que no s'analitza directament la taula ternària, sinó que s'estudien algunes taules binàries construides a partir d'ella, la qual cosa en dóna una visió parcial. No obstant això, la característica més novedosa que s'aporta amb aquesta metodologia és la descomposició de la inèrcia

- Article rebut el juny de 1995.
- Acceptat el juny de 1996.

total de la taula en inèrcia INTER i inèrcia INTRA, com s'explicarà posteriorment, la qual cosa ajuda a entendre les relacions entre les tres variables.

La tècnica bàsica d'anàlisi és l'AFC (Anàlisi factorial de correspondències) ja que es treballa amb diferents tipus de taules binàries de contingència. Cadascun dels distints AFC es correspon a una manera diferent d'estudiar la successió indexada de taules binàries, i la interpretació conjunta d'aquestes anàlisis permeten veure les característiques essencials de la taula ternària.

Alguna d'aquestes AFC, sobretot la que es coneix com a ANÀLISI GLOBAL o de LES INÈRCIES INTER + INTRA, és de difícil interpretació, ja que els dos tipus d'inèrcia (INTER i INTRA) estan barrejades, i s'ha d'anar molt en compte a l'hora d'interpretar els resultats. El que fa interessant aquest estudi és, com es veurà més endavant, que la inèrcia INTER i la inèrcia INTRA són ortogonals, la qual cosa facilita molt la interpretació, i és per això que aquest estudi em sembla força vàlid com a aplicació didàctica per ajudar a entendre millor aquest tipus de metodologia.

2. PRESENTACIÓ DE LES DADES

La taula que es va analitzar és la que porta per títol: *Girona barri per barri* que va publicar el diari EL PUNT, el dimecres 9 de juny de 1993. En ella es fa una relació del nombre de vots obtinguts pels diferents partits en els barris de Girona, a les eleccions generals dels anys 1989 i 1993.

Si s'analitza la taula es veu que hi intervenen tres variables: I = «Barris», J = «Partits» i T = «Anys». Aquesta última variable agafa dues modalitats 1989 i 1993 i juga un paper secundari respecte a les altres dues, és la variable que indexa la successió formada per dues taules que creuen les variables, I i J. Per tant es pot separar la taula en dues, una amb els resultats de l'any 89 i l'altra amb els resultats del 93. Sobre aquestes dues taules es farà l'anàlisi comparativa.

A continuació es relacionen les modalitats de cadascuna de les variables juntament amb les abreviatures utilitzades en els gràfics:

Taula 1

Girona barri per barri. (Extracte de la taula publicada)

Barri	El.	Abs.	CiU	Psc	Erc	Pp	Ic
Casernes	93	299	361	316	91	193	47
	89	419	339	297	27	112	55
Palau-sacosta	93	924	1105	639	283	451	137
	89	492	440	186	42	138	65
Pedret	93	108	145	141	46	55	22
	89	212	161	158	5	49	25
Sant Ponç	93	490	259	415	55	158	57
	89	596	206	254	19	50	32
Sant Narcís	93	1195	1210	1239	306	568	188
	89	1403	845	804	74	300	154
Eixample	93	1856	3600	1068	776	1899	262
	89	2739	3630	839	239	1193	304
Devesa-Mercadal	93	708	1391	395	310	609	89
	89	1074	1410	329	101	420	124
Carme-Vista Alegre	93	594	813	447	220	239	71
	89	650	519	299	48	108	53
La Rodona	93	780	1031	710	294	460	135
	89	1124	848	537	93	232	144
Barri Vell	93	830	593	631	176	269	118
	89	1522	930	683	99	240	152
Can Gibert del Pla	93	328	184	678	53	82	70
	89	442	141	547	24	47	82
Santa Eugènia	93	1595	1526	1765	422	547	295
	89	2070	1264	1395	137	238	294
Montilivi	93	615	553	623	139	321	99
	89	732	361	460	22	165	66
Vila-roja-	93	783	131	1351	13	129	113
Font de la Pólvora	89	1046	80	1229	9	34	46
Montjuïc-Sant Daniel	93	347	600	396	202	168	85
	89	512	466	314	48	88	102
Pont Major	93	532	497	717	108	129	73
	89	811	411	565	35	67	105
Germans Sàbat-Taialà	93	459	254	1146	29	99	139
	89	650	168	992	8	45	116

Variable I = «Barris» (17 modalitats):

(Cas) Casernes	(Dev) Devesa-Mercadal	(Mvi) Montilivi
(Pal) Palau-sacosta	(Car) Carme-Vista Alegre	(Vro) Vila-roja-F. Pólvora
(Ped) Pedret	(Rod) La Rodona	(Mon) Montjuïc-S. Daniel
(Spo) St. Ponç	(Bve) Barri Vell	(Pon) Pont Major
(Sna) St. Narcís	(Cgp) Can Gibert del Pla	(Gsa) Germans Sabat-Tai
(Eix) Eixample	(Seu) Sta. Eugènia	

Variable J = «Partits» (6 modalitats):

(Abs) Abstenció	(Psc) P.S.C.	(Pp) P.P.
(Ciu) C. i U.	(Erc) E.R.C.	(Ic) I.C.

3. ANÀLISI INTER + INTRA O GLOBAL

Aquesta és una de les formes habituals d'analitzar aquestes successions de taules. Consisteix en fer una Anàlisi Factorial de Correspondències (AFC) de la taula formada per la juxtaposició per files o per columnes de tota la successió de taules. Així, per exemple, en juxtaposar per files ens queda una taula amb el mateix nombre de files que cadascuna de les taules de la successió, i amb un nombre de columnes igual al nombre de columnes de les taules pel nombre de taules.

Concretament, en la successió que treballem nosaltres es poden juxtaposar les dues taules (tal com es veu a la Figura 1): per files, és a dir, es treballa amb una taula de 17 files, que corresponen a les modalitats dels barris, i 12 columnes, 6 corresponents a les modalitats dels partits de la taula de 1989 i les altres 6 corresponents al 93; per columnes es traballa amb una taula de 6 columnes, corresponents a les modalitats dels partits, i 34 files, 17 corresponents als resultats dels barris l'any 1989 i les altres 17 corresponents als resultats del 1993.

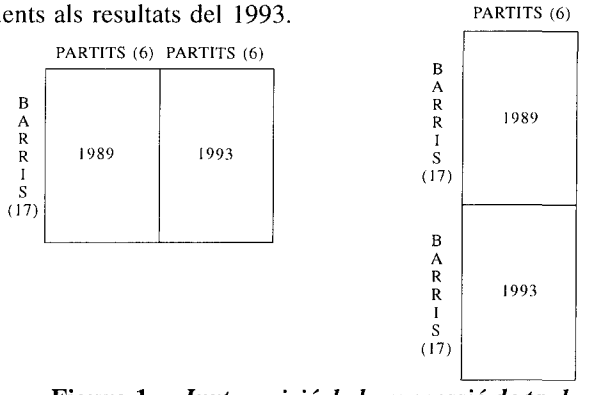


Figura 1. ***Juxtaposició de la successió de taules***

Com ja es veu per la configuració de les matrius, en aquesta anàlisi el paper de les dues variables I i J no és simètric.

Ens centrarem en la taula juxtaposada per files, que és la més interessant de presentar amb l'objectiu d'explicar la metodologia, ja que és la que té una descomposició quasi ortogonal de les seves inèrcies INTER i INTRA, tal com hem comentat abans. Ja s'enten que en una anàlisi completa s'ha de fer un estudi semblant per columnes, però no el comentem ja que la separació entre els dos tipus d'inèrcia no és tan clara.

AFC de les taules juxtaposades per files, on es projecta la taula suma com a suplementària.

Com ja s'ha indicat abans es fa un AFC d'una taula de 17 files i 12 columnes, i es completa amb la projecció, com a columnes suplementàries, de les 6 columnes de la taula suma de totes les taules, que corresponen a les posicions baricèntriques dels diferents partits. És a dir, si ens mirem el núvol de punts de les 18 columnes, tenim 12 columnes actives que corresponen a les modalitats dels partits a les dues eleccions i 6 columnes suplementàries que representen les posicions baricèntriques de cada partit, tenint en compte les dues eleccions. Per tant, es pot considerar el núvol complet dividit en 6 subnúvols, cadascun d'ells compost pels dos perfils actius dels barris corresponents al mateix partit (ex: IC-89, IC-93) i un perfil suplementari que correspon al baricentre d'aquestes dues modalitats.

Així, la inèrcia total de la taula juxtaposada per files es pot descompondre segons el principi de Huygens, és a dir:

INÈRCIA TOTAL = INÈRCIA INTER + Suma de les INÈRCIES INTRA

La inèrcia inter és la del núvol dels baricentres dels 6 subnúvols, i la inèrcia intra és la inèrcia de les modalitats de cada subnúvol respecte al seu baricentre, per tant en aquesta AFC es farà palesa, a més de la inèrcia inter partits, la inèrcia intra partits, és a dir, la diferència entre els perfils dels barris de les modalitats dels partits al 89 i al 93. D'aquí ve el nom d'*Anàlisi Inter + Intra.*

Anem a veure, doncs, la inèrcia total d'aquesta taula i com es descompon en factors:

TAULA JUXTAPOSADA PER FILES
INÈRCIA TOTAL: 0.14495

Factor	1	2	3	4
Inèrcia Total	0.118	0.0134	0.0064	0.0047
Inèrcia Percentual	81.6	9.3	4.4	3.2

La resta de factors es poden obviar.

Normalment, els dos tipus d'inèrcies queden més o menys barrejades en tots els factors, la qual cosa dificulta molt l'anàlisi dels resultats, a menys que, com es veurà en aquest exemple, la major part de la inèrcia intra sigui ***ortogonal*** a la inter i es concentri a un eix. Aquest és el fet que fa més interessant l'exemple que es presenta.

Per tal de fer evident el percentatge d'inèrcia de cada tipus a cadascun dels factors es calculen dos índexs d'ajuda:

❶ ***Contribució a la inèrcia dels 6 subnúvols dels partits***

A la taula 2 es mostra per tot l'espai i per cada factor, el percentatge d'inèrcia inter i d'inèrcia intra, i a més es mostra la contribució de cada subnúvol a la inèrcia intra recollida per cada factor.

Taula 2

Índexs de contribució a la inèrcia. Percentatges d'inèrcia inter-partits i intra-partits a l'anàlisi dels perfils dels barris

	ESPAI	F1	F2	F3	F4
INTER	0.894	0.991	0.039	0.978	0.914
INTRA	0.106	0.009	0.961	0.022	0.086
ABS	0.030	——	0.303	0.002	0.024
CiU	0.037	0.003	0.371	0.003	0.002
PSC	0.017	0.003	0.138	0.004	0.004
ERC	0.006	——	0.041	0.009	——
PP	0.010	0.002	0.078	0.003	——
IC	0.006	0.001	0.031	——	0.056

Com es pot veure els percentatges d'inèrcia inter dels factors 1, 3 i 4 d'aquesta anàlisi són més grans que un 90%, mentre que el factor 2 té un percentatge d'inèrcia intra superior a un 95%, evidentment, aquesta descomposició entre les dues inèrcies no sempre és tan clara. Per tant la major part de la inèrcia intra és ortogonal a la inèrcia inter i es concentra al segon eix.

❷ ***Qualitat de la representació dels diferents subnúvols***

És el quocient entre la quantitat d'inèrcia intra d'un subnúvol projectada sobre un factor, dividit per la inèrcia intra total del subnúvol. Es mostren a la taula 3:

Taula 3

Índexs de qualitat de la representació del núvol dels baricentres i dels 6 subnúvols en l'anàlisi dels perfils dels barris

	F1	F2	F3	F4
BARICENTRES	0.905	0.004	0.048	0.033
ABS	____	0.940	0.004	0.026
CiU	0.059	0.920	0.004	0.002
PSC	0.135	0.753	0.010	0.007
ERC	0.009	0.684	0.071	____
PP	0.162	0.735	0.013	0.001
IC	0.152	0.441	0.001	0.280

Cal destacar la alta qualitat de les representacions de tots els subnúvols respecte del segon eix, la qual cosa confirma el que s'ha afirmat anteriorment.

Finalment a les figures 2, 3, 4 i 5 es presenten alguns gràfices corresponents a plans factorials d'aquesta anàlisi, no es presenta interpretació, ja que no és l'objectiu del present article, només analitzarem el segon eix.

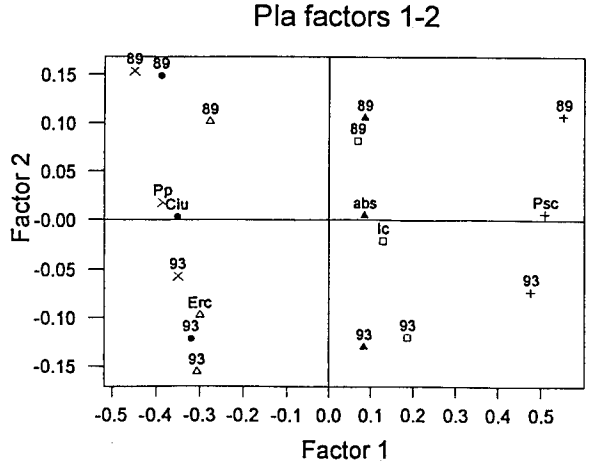


Figura 2. ***AFC-Taula juxtaposada per files (Proj. column. suplementàries)***

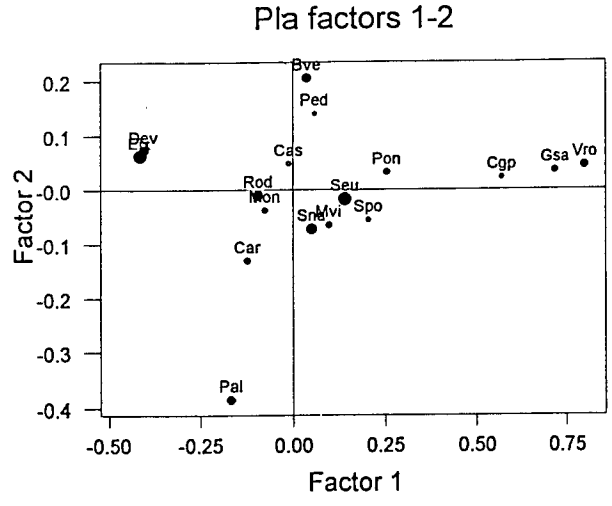


Figura 3. *AFC-Taula juxtaposada per files*

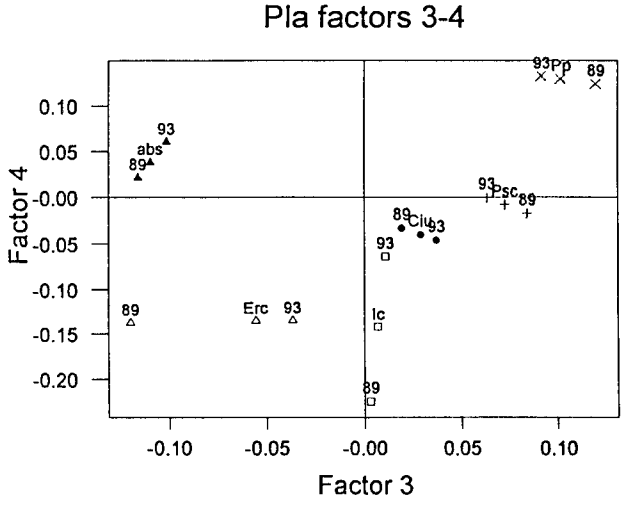


Figura 4. *AFC-Taula juxtaposada per files (Proj. colum. suplementàries)*

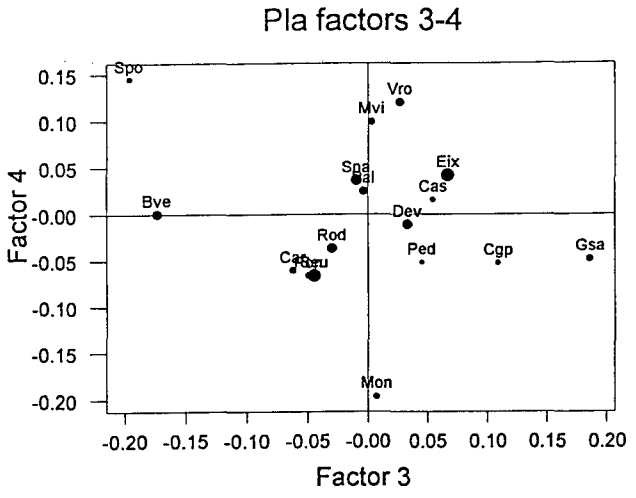


Figura 5. ***AFC-Taula juxtaposada per files***

El segon eix:

Recull un 9.3% de la inèrcia total de la qual un 96% és inèrcia intra i, tenint en compte que la inèrcia intra és un 10.6% del total, recull més del 84% de la inèrcia intra.

Oposa les modalitats dels partits del 89 a les del 93. Pel que fa als barris oposa per una banda PALAU, barri perifèric i residencial de fort creixement demogràfic, i per l'altra banda el BARRI VELL, que perd població, és a dir, oposa els barris on ha crescut el cens amb els que ha baixat, per això el podem anomenar «l'eix del creixement de cens». És lògic que la distribució percentual dels votants de cada partit entre els barris, augmenti d'una elecció a l'altra en els barris on ha augmentat el cens.

Per tal de corroborar aquesta manera particular de descomposició de la inèrcia mostrarem alguns resultats corresponents a l'ANÀLISI INTER i l'ANÀLISI INTRA.

4. ANÀLISI INTER

AFC de la taula suma i projecció de les columnes de les dues taules com a suplementàries.

Aquest tipus d'anàlisi permet posar de manifest les tendències comunes de les dues taules. El que es fa és sumar les dues taules i aplicar a la taula suma una

anàlisi de correspondències. El núvol de punts columna representa les posicions dels baricentres dels perfils de les columnes que corresponen al mateix partit a les dues taules; per exemple, la columna corresponent al PSC és el baricentre de les columnes PSC-89 i PSC-93.

Després és possible projectar les columnes de les taules de 1989 i 1993 com a columnes suplementàries, la qual cosa permetrà comparar-les a partir d'un referencial comú, que són els eixos d'inèrcia del núvol dels seus baricentres, així es poden veure les desviacions dels perfils fila (o columna) de cadascuna de les taules respecte del perfil mitjà.

Cal tenir en compte que les diferències poden no ser visibles, si aquestes desviacions són ortogonals als eixos dels baricentres, com passa en aquesta aplicació o si són molt petites respecte de les desviacions dels perfils mitjans, és a dir, si la diferència entre PSC-89 i PSC-93 és molt petita respecte a la diferència entre els baricentres de PSC i CiU no ès farà evident, per això, tot i que aquest és l'anàlisi que es realitza en primer lloc i potser és el que ens aporta més en el coneixement de la taula ternària, cal completar-lo amb les altres dues anàlisis INTER + INTRA i INTRA.

Anem a veure la inèrcia total i la seva descomposició segons els factors:

TAULA SUMA
INÈRCIA TOTAL: 0.12958

Factor	1	2	3
Inèrcia Total	0.117	0.0063	0.0045
Inèrcia Percentual	90.5	4.9	3.5

La resta de factors es poden obviar.

Un primer fet que corrobora el que s'ha dit a l'anàlisi inter + intra és la gran coincidència entre les inèrcies dels eixos 1, 2 i 3 d'aquesta anàlisi i les dels eixos 1, 3 i 4 de la taula juxtaposada per files, cosa que fa sospitar una semblança entre ells.

Un segon fet és veure que les figures 6 i 7, que corresponen al pla dels factors 2-3 de l'anàlisi INTER, són molt semblants a les figures 4 i 5 respectivament amb l'eix horitzontal invertit, aquestes figures corresponen al pla dels factors 3-4 de l'anàlisi INTER + INTRA.

Finalment presentarem una última evidència per veure la coincidència entre aquests eixos:

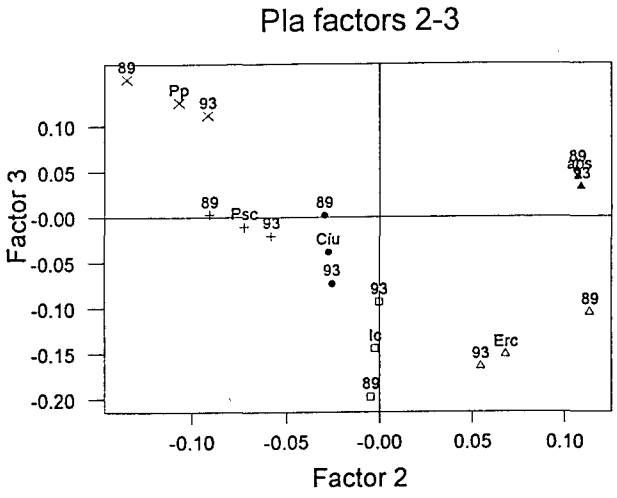


Figura 6. ***AFC-Taula suma (Proj. colum. suplementàries)***

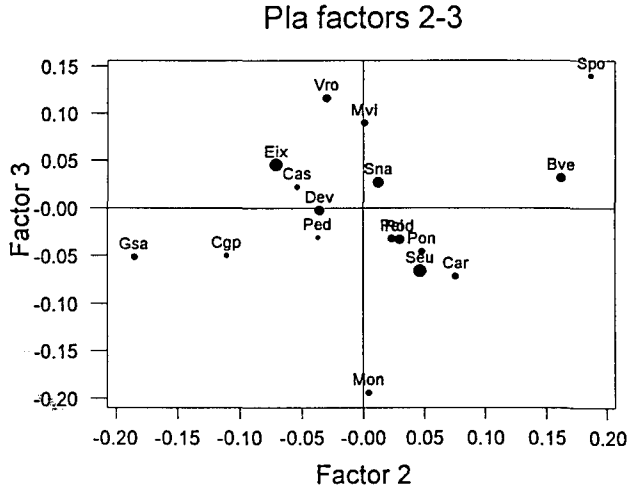


Figura 7. ***AFC-Taula suma***

Comparació dels factors de les diferents anàlisis

Els factors de l'anàlisi INTER + INTRA proporcionen un marc adequat per comparar els factors de l'anàlisi de la taula suma i els de les anàlisis de les taules del 89 i del 93 per separat. Com que aquests poden considerar-se com a variables, és possible calcular-ne correlacions i disposar-les en una taula, com es mostra a la taula 4.

Cal tenir en compte que la marginal de les modalitats dels barris és diferent a les dues taules separades i també a la taula juxtaposada i a la taula suma, encara que en aquestes dues últimes és igual, per això abans de fer les correlacions cal fer unes correccions en els factors de les taules separades per igualar les marginals, un cop fetes aquestes correccions ja es poden fer les correlacions que es posen de manifest a la taula 4.

Taula 4

Correlacions entre els factors de l'anàlisi de la taula juxtaposada i els de les altres anàlisis

		Factor 1	Factor 2	Factor 3	Factor 4
TAULA	Factor 1	0.999	0.172	0.251	0.136
SUMA	Factor 2	−0.277	−0.116	−0.996	0.215
	Factor 3	0.157	0.129	−0.280	0.974
	Factor 4	0.015	0.209	−0.109	−0.110
TAULA	Factor 1	0.997	0.119	0.274	0.117
1989	Factor 2	0.241	0.172	0.981	−0.229
	Factor 3	−0.095	−0.036	0.265	−0.919
	Factor 4	−0.026	0.135	−0.164	−0.124
TAULA	Factor 1	0.997	0.185	0.229	0.162
1993	Factor 2	0.387	−0.372	−0.834	−0.115
	Factor 3	0.049	−0.042	−0.504	0.940
	Factor 4	0.036	−0.080	−0.160	0,088

Com es pot veure hi ha una altra correlació, superior a 0.97, entre els factor 1, 2 i 3 de la taula suma i els 1, 3 i 4 de la taula juxtaposada respectivament. Cal tenir en compte que alguna d'aquestes correlacions és negativa perquè els eixos estan invertits.

5. ANÀLISI INTRA

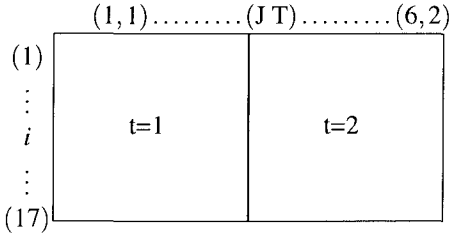
Anàlisi intra dels partits, relacions condicionals

A l'anàlisi INTER + INTRA, s'ha estudiat la matriu juxtaposada per files (una taula de 17 files i 12 columnes), que, com hem explicat, es podia considerar formada per 6 subnúvols, cadascun d'ells amb les dues modalitats corresponents a un mateix partit. Hem vist que la inèrcia total de la taula es descomponia en inèrcia inter, del núvol dels baricentres dels subnúvols, més la suma de les inèrcies intra, de les modalitats de cada subnúvol respecte del seu baricentre.

El que es pretén ara és eliminar d'aquesta taula la inèrcia inter, per estudiar només la inèrcia intra-partits, per aconseguir-ho el que es fa és traslladar tots els subnúvols, de manera que tots els baricentres vagin a parar a l'origen de coordenades. Així, en aquest nou núvol, les coordenades del punt de la modalitat PSC-93 mostren la diferència entre el perfil dels barris del PSC-93 i el perfil dels barris del PSC si s'acumulen les dues eleccions.

Per fer això es construeix primer una nova taula de 17 files i 12 columnes que se'n diu TAULA MODEL. En aquesta taula els perfils de les dues columnes corresponents a les dues modalitats del mateix partit (se'n diuen columnes homòlogues), seran iguals al perfil mitjà del partit en les dues eleccions, a més a més, és necessari que les marginals d'aquesta nova taula coincideixin amb les de la taula juxtaposada per files, ja que al final hem de restar les dues taules. Amb aquestes restriccions queden completament definits els elements de la taula model que seran igual al perfil baricèntric del partit multiplicat per la marginal de la columna corresponent, així, per exemple, a la columna CiU-89 hi figurarà el perfil baricèntric de CiU multiplicat per la marginal de la columna CiU-89 de la taula juxtaposada.

Siguin f_{ijt} $i = 1,\ldots,17$ (barris) $j = 1,\ldots,6$ (partits) $t = 1, 2$ (eleccions), els elements de la taula juxtaposada per files.



Les seves marginals són: Marginal de les files: $\sum_{t=1}^{2}\sum_{j=1}^{6} f_{ijt} = f_i$

Marginal de les columnes: $\sum_{i=1}^{17} f_{ijt} = f_{jt}$

Si anomenem $f_j = \sum_{t=1}^{2}\sum_{i=1}^{17} f_{ijt}$

El perfil mitjà dels partits a les dues eleccions serà: $\frac{f_{ij}}{f_j}$ i els elements de la taula model seran:

$$m_{ijt} = \frac{f_{ij}}{f_j} \cdot f_{jt}$$

La seva marginal de les files: $\sum_{t=1}^{2}\sum_{j=1}^{6} m_{ijt} = \sum_{j=1}^{6}\sum_{t=1}^{2} \frac{f_{ij}}{f_j} \cdot f_{jt} = \sum_{j=1}^{6} f_{ij} = f_i$

La seva marginal de les columnes: $\sum_{i=1}^{17} \frac{f_{ij}}{f_j} \cdot f_{jt} = f_{jt}$

que com es veu coincideix amb els de la taula juxtaposada tal com s'havia proposat.

Una vegada construïda aquesta matriu es resta de la matriu juxtaposada, així al fer els perfils columna s'obtindrà la diferència entre el perfil de la modalitat corresponent i el del seu baricentre.

Finalment, com que un AFC analitza les desviacions d'una taula de la hipòtesis d'independència s'ha de sumar a cada element el producte de les marginals, d'aquesta manera al fer un AFC s'analitzaran les desviacions entre la taula juxtaposada i la taula model. Amb totes aquestes operacions s'aconsegueix produir l'efecte següent: els perfils columna de la taula que s'analitza menys el perfil del baricentre de totes les columnes del núvol és igual a la diferència entre els perfils de la modalitat corresponent de la taula juxtaposada menys el perfil del baricentre del partit corresponent, això equival a traslladar tots els baricentres dels subnúvols a l'origen de coordenades.

D'aquesta manera el que s'analitza és la desviació a la hipòtesi d'independència dels barris i les eleccions condicionats als partits, és a dir, la independència de les variables I i T condicionades a J.

Anem a veure ja els resultats d'aquest AFC. Començarem per la inèrcia total de la taula i la seva descomposició en factors:

TAULA INÈRCIA INTRA
INÈRCIA TOTAL: 0.015368

Factor	1	2	3	4
Inèrcia Total	0.0131	0.0019	0.0005	0.0004
Inèrcia Percentual	85.4	7.7	3.2	2.4

Com es pot veure la inèrcia total recollida pel primer factor, és molt semblant a la inèrcia recollida pel segon factor de l'anàlisi INTER + INTRA 0.0134, a més, el percentatge d'inèrcia intra recollida 85,4% coincideix força amb el del segon eix de l'altra anàlisi, més d'un 84%.

D'aquest AFC es mostren les figures 8 i 9 que corresponen al pla dels factors 1-2, aquestes figures es poden comparar amb les figures 2 i 3 respectivament per buscar coincidències entre el primer eix de l'anàlisi INTRA i el segon eix de l'anàlisi INTER + INTRA.

Finalment per facilitar-ne la comparació es mostra una taula amb les coordenades de totes les modalitats respecte d'aquests dos factors.

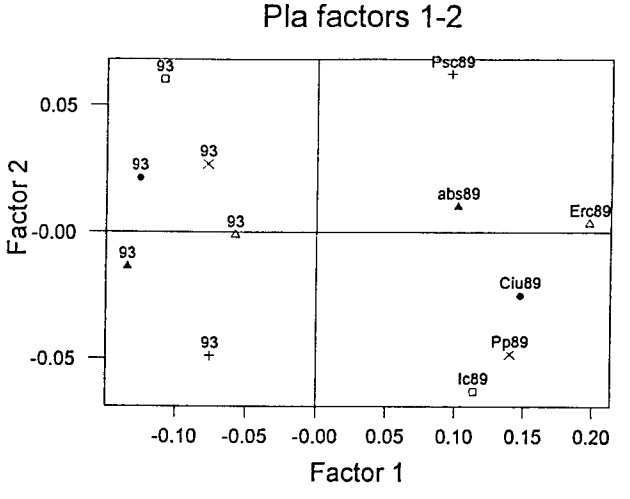


Figura 8. ***Anàlisis inèrcia intra-barris***

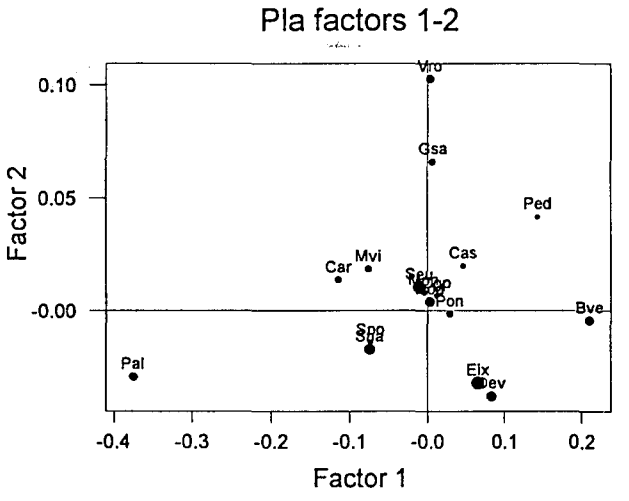


Figura 9. *Anàlisis inèrcia intra-partits*

Taula 5

Coordenades de les modalitats dels partits i dels barris respecte del segon factor de l'anàlisi inter + intra i del primer factor de l'anàlisi intra partits

PARTITS	F2 in + in	F1 intra	BARRIS	F2 in + in	F1 intra
			CASERNES	0.049	0.046
ABS-89	0.106	0.102	PALAU	−0.385	−0.376
CiU-89	0.149	0.147	PEDRET	0.142	0.143
PSC-89	0.107	0.097	ST. PONÇ	−0.055	−0.073
ERC-89	0.101	0.198	ST. NARCÍS	−0.071	−0.074
PP-89	0.153	0.140	EIXAMPLE	0.063	0.065
IC-89	0.081	0.114	DEVESA	0.074	0.083
			CARME	−0.130	−0.115
ABS-93	−0.130	−0.135	LA RODONA	−0.010	0.003
CiU-93	−0.122	−0.126	BARRI VELL	0.206	0.210
PSC-93	−0.073	−0.075	CAN GIBERT	0.023	0.012
ERC-93	−0.156	−0.058	STA. EUGÈNIA	−0.017	−0.011
PP-93	−0.057	−0.078	MONTILIVI	−0.064	−0.076
IC-93	−0.120	−0.109	VILA-ROJA	0.044	0.003
			MONTJUÏC	−0.037	−0.005
			PONT MAJOR	0.033	0.028
			GER. SABAT	0.035	0.006

Com es pot veure hi ha una certa coincidència amb les coordenades, sobretot pel que fa als partits més votats, per tant, amb més pes i també a la majoria dels barris.

6. CONCLUSIÓ

Vistes totes aquestes evidències podem concloure que la inèrcia inter i la inèrcia intra-partits són en un alt percentatge ortogonals i que els factors 1, 3 i 4 de l'anàlisi INTER + INTRA són molt coincidents amb els factors 1, 2 i 3 de l'anàlisi INTER, mentre que el factor 2 de l'anàlisi de la taula juxtaposada és semblant al factor 1 de l'anàlisi INTRA. Aquests fets són els que fan particularment interessant l'exemple presentat.

7. AGRAÏMENT

Agraeixo l'ajuda i el suport que el Dr. Tomàs Aluja Banet m'ha donat durant la realització d'aquest treball.

REFERÈNCIES BIBLIOGRÀFIQUES

[1] **Aluja, T.** i **Nonell, R.** (1993). *Multivariate Analysis: Future directions 2.* C.M. Cuadras, C.R. Rao. North Holand, 233–244.

[2] **Escofier, B.** i **Pagès, J.** (1992.) *Análisis factoriales simples y múltiples.* Universidad del País Vasco.

[3] **Lebart, L., Morineau, A.** i **Warwich, K.** (1984). *Multivariate Descriptive Statistical Analysis.* Wiley.

[4] **Van der Heijden, P.** i **Van Leeuw, J.** (1989). «Correspondence Analysis with Special Attention to the Analysis of Panel Data and Event History Data». *Sociological Methodology*, **19**, 43–87.

ENGLISH SUMMARY:

ELECTIONS IN GIRONA: AN EXAMPLE OF A STUDY OF A TERNARY TABLE

F. Borrell Thió

One way to analyse ternary tables (those tables resulting from the crossing of three qualitative variables, I, J and T) is the way proposed by Escofier, B. and Pages, J. in 1992. They consider the data as a sequence of tables of binary frequency resulting from the crossing of two variables I and J, indexed by the other variable T, which is playing a different role.

The biggest problem raised by this working method is that it does not analyse the ternary table in a direct way. Instead it studies binary tables derived from the ternary one. However, the newest characteristic introduced by this method is the separation of the total inertia of the table into INTER inertia and INTRA inertia. This separation helps us understand the relationships among the three variables.

The basic technique used is the FCA (Factorial Correspondence Analysis) because different kinds of binary tables of contingency are involved. Basically three analysis are done:

1. **GLOBAL ANALYSIS** or analysis of INTER + INTRA inertia. The FCA is applied to the table formed by the juxtaposition of the rows (or the columns) in the sequence of tables. For example, in the case of the juxtaposition of the rows, the number of rows of the analysed table will be the same as the number of items in the variable I, and the number of columns will be the same as the number of items in the variable J multiplied by the number of tables in the sequence.

   Moreover, the analysis is completed, in the case of the juxtaposition of the rows, with the projection as supplementary columns of the columns in the table that is the addition of all the tables in the sequence. These columns correspond to the barycentric positions of the items in the variable J. This analysis is difficult to interpret because the two kinds of inertia, INTER and INTRA, are mixed.

2. **INTER ANALYSIS**. The FCA is applied to the table that is the addition of all the tables. It is completed with the projection of the columns (or rows) of the tables in the sequence as supplementary columns (or rows). It is useful to

analyse the Inter inertia, that is, the inertia among the barycentric positions of the different items in the variable I and J. This analysis allows to reveal the common tendencies in the tables of the sequence.

3. **INTRA ANALYSIS**. An FCA is applied to a table resulting from the transformation of a juxtaposed table in the GLOBAL ANALYSIS. This transformation consist of placing all the barycentric positions of the items in the columns (or in the rows) at the beginning of the coordinates. Therefore, what is analysed is the differences among all the columns corresponding to one item of the variable J in relation to its barycentre. In other words, the INTRA inertia is analysed.

This article presents an example of how to apply these techniques to a real table taken from the election results in the town of Girona. In this table the variables are: I = «neighbourhoods», J = «parties» and T = «years». What makes this study interesting is that the INTER and the INTRA inertia are orthogonal, and this facilitates the interpretation very much. This is the reason I think this study is quite valid as a didactic application with the purpose of helping to understand this methodology better.

SOLUCIONS ALS PROBLEMES PROPOSATS AL VOLUM 20, N° 1

PROBLEMA N° 57

Consideremos una población finita de tamaño N, $U = \{1,2,\ldots,i,\ldots,N\}$. Sean «$y$» y «$x$» la variable de interés y auxiliar respectivamente. Tratamos de estimar la media poblacional de interés

$$\bar{y} = \sum_{i \in U} y_i / N,$$

haciendo uso del estadístico «media-de-productos»

$$\bar{p}_s = \frac{1}{n} \sum_{i \in s} y_1 x_i.$$

Para ello nos basamos en el trabajo de Murthy (1964). De (3.3) y (3.4), deducimos que bajo muestreo aleatorio simple sin reemplazamiento, tenemos los sesgos (siendo $\bar{y}_s$ y $\bar{x}_s$ las medias muestrales de las variables «y» y «x»),

$$B(\bar{y}_s\, \bar{x}_s) = E(\bar{y}_s\, \bar{x}_s) - \overline{yx}$$

y

$$B(\bar{p}_s) = E(\bar{p}_s) - \overline{yx}$$

donde $\bar{x}$ es la media poblacional de la variable auxiliar «x». Como $B(\bar{p}_s) = nB(\bar{y}_s\, \bar{x}_s)$, deducimos que

$$E(\bar{p}_s) - \overline{yx} = nB(\bar{y}_s\, \bar{x}_s) - n\overline{yx},$$

por lo que

$$(n-1)\overline{yx} + E(\bar{p}_s) = nE(\bar{y}_s\, \bar{x}_s)$$

y por tanto

$$z = \bar{y}_s[(n-1)\bar{x} - n\bar{x}_s] + \bar{p}_s$$

es una variable aleatoria de esperanza nula. La clase de estimadores insesgados para $\bar{y}$ basados en el estadístico $\bar{p}_s$ será

$$t = \bar{y}_s + kz.$$

Esta metodología es análoga a la vista por Ruiz y Santos (1989), de manera que los últimos comentarios realizados en el párrafo final de estos autores, son válidos también en la clase aquí propuesta.

Referencias

Murthy, M.N. (1964). «Product method of estimation». *Sankhyā.* **Ser. A, 26**, 69–74.

Ruiz, M. y Santos, J. (1989). «Unbiased mean-of-the-ratios estimators». *Statistica.* **49**, 617–622.

M. Ruiz Espejo

UNED

PROBLEMA N° 58

a) La demostración de la condición suficiente, dada por Ruiz (1980, 1985), es básicamente la propuesta por Glasser (1962) hasta la ecuación (16). Pero en lugar de (16), señalamos que $m < n$, y entonces podemos deducir directamente que

$$k^2 > \frac{N}{n} - \frac{Nm-n}{n(N-1)} > \frac{N}{n} - 1,$$

y de aquí se deriva la nueva cota, ya que $k > 0$,

$$k > \sqrt{\frac{N}{n} - 1},$$

o en otras palabras

$$x > \mu + \sqrt{\frac{N}{n} - 1} \cdot \sigma.$$

b) Considerar una población finita de tamaño $N = 10$ cuyos valores de la variable de interés ordenados en orden creciente es: 1, 2, 2, 4, 4, 5, 7, 9, 12 y 14. Para una muestra de tamaño $n = 5$ tenemos la siguiente tabla de varianzas para varios procedimientos de muestreo

L = número de estratos	Diseño muestral	Varianza
1	Muestreo aleatorio simple sin reemplazo.	1.96
2	Estratificación especial con la cota de Glasser o Ruiz ($m = 2$).	0.98
2	Estratificación especial con $m = 3$.	0.75
2	Estratificación especial óptima ($m = 4$).	0.72

Este ejemplo garantiza que la cota inferior para el punto de estratificación especial con dos estratos no sea también cota inferior del punto de estratificación especial óptimo. En el ejemplo propuesto $\mu = 6$ y $\sigma^2 = 17.6$.

Referencias

Glasser, G.J. (1962). «On the complete coverage of large units in a statistical study». *Rev. Internat. Statist. Inst.* **30**, 28–32.

Ruiz, M. (1980). *Construcción de Estratos en el Diseño de Muestreo Estratificado Aleatorio.* Tesina de Licenciatura. Universidad Complutense de Madrid.

Ruiz, M. (1985). «Equiprecisional allocation and optimum stratification». *Statistics.* **16**, 559–562.

M. Ruiz Espejo

UNED

PROBLEMA N° 59

Prendrem la comparació del bucle `while` com a instrucció crítica. Així, podem observar que aquesta instrucció s'executa com a mínim 1 vegada i com a màxim i vegades en cada iteració del bucle `repeat`. D'altra banda, aquest bucle s'executarà de nou sempre que el bucle `while` s'hagi executat menys de i vegades (i, per tant, $j < i$).

Pel que fa al millor cas, doncs, podem veure fàcilment que correspon a l'execució del bucle `while` el màxim de vegades. Així, el bucle `repeat` s'executarà només una vegada. El cost en el millor cas serà, doncs:

$$T^m(n) = \sum_{i=1}^{n} i = \frac{n}{2}(n+1) = O(n^2).$$

Quant al pitjor cas, no es pot calcular estrictament ja que, per a un n donat, existeix una certa probabilitat que el cost sigui arbitràriament gran.

Si els valors aleatoris generats en cada iteració són independents, aleshores podem calcular el cost en el cas mitjà per a cada iteració separadament i després sumar per a totes les iteracions:

$$T(n) = \sum_{i=1}^{n} t(i) \tag{1}$$

Suposem que en la iteració i han estat generats $i-1$ valors no repetits que es troben en les $i-1$ primeres posicions del vector. És clar que el cost en la iteració i, $t(i)$, valdrà i amb probabilitat $\frac{n-i+1}{n}$ (la probabilitat de generar un valor diferent dels $i-1$ primers). Aquest valor és el mínim possible, i correspon a una única execució del bucle `repeat`.

Podem suposar, sense pèrdua de generalitat, que l'obtenció d'un dels $i-1$ primers valors en la iteració i és equiprobable. Per tant, tots els casos en què el bucle `repeat` s'execute dues vegades els podem resumir dient que el cost és igual a $\frac{i}{2}$ (cost mitjà si es genera qualsevol dels $i-1$ primers valors) més i (cost de l'última iteració) amb probabilitat $\frac{i-1}{n}\frac{n-i+1}{n}$ (probabilitat de generar un dels valors repetits per la probabilitat de generar-ne un no repetit).

De la mateixa manera, si el bucle `repetir` s'executa $j+1$ vegades (j valors repetits i un de no repetit) el cost corresponent seria $i + j\frac{i}{2}$ amb probabilitat $\left(\frac{i-1}{n}\right)^j \frac{n-i+1}{n}$. Per tant, el valor esperat del cost en cada iteració i el podem calcular com a

$$t(i) = \sum_{j=0}^{\infty} (i + j\frac{i}{2}) \left(\frac{i-1}{n}\right)^j \frac{n-i+1}{n} = i\frac{n-i+1}{n}\left[\sum_{j=0}^{\infty}\left(\frac{i-1}{n}\right)^j + \frac{1}{2}\sum_{j=0}^{\infty} j\left(\frac{i-1}{n}\right)^j\right].$$

Les dues sèries que apareixen són convergents ja que $\frac{i-1}{n} < 1$.

A partir dels resultats $\sum_{j=0}^{\infty} r^j = \frac{1}{1-r}$, i $\sum_{j=0}^{\infty} jr^j = \frac{r}{(1-r)^2}$, obtenim

$$t(i) = \frac{i(2n-i+1)}{2(n-i+1)}.$$

Podem calcular ara el cost mitjà de l'algorisme fent servir l'Equació (1) com a

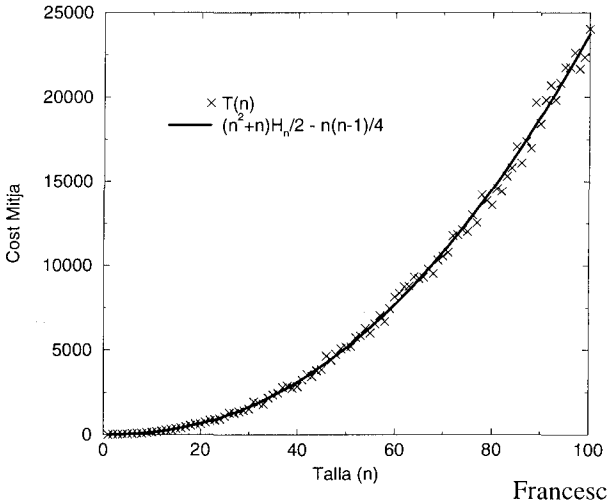
$$T(n) = \sum_{i=1}^{n} \frac{i(2n-i+1)}{2(n-i+1)} = \frac{1}{2}\sum_{k=1}^{n} \frac{(n-k+1)(n+k)}{k} = \frac{n^2}{2}\sum_{k=1}^{n}\frac{1}{k} - \frac{1}{2}\sum_{k=1}^{n}(k-1) + \frac{n}{2}\sum_{k=1}^{n}\frac{1}{k}.$$

Aleshores, el nombre mitjà de vegades que s'executa la instrucció crítica es pot escriure com a

$$T(n) = \frac{n^2+n}{2}H_n - \frac{n}{4}(n-1) = O(n^2 \log n),$$

on H_n és el nombre harmònic n-èsim.

Per tal de comprovar l'exactitud del càlcul que hem fet, en la següent figura es pot veure la funció $T(n)$ junt amb el cost mitjà mesurat (comptant les iteracions del bucle) després d'executar el corresponent programa 50 vegades per a valors de n entre 1 i 100. Encara que en la figura s'observa una relativa dispersió en els valors empíricament mesurats, es comprova que el càlcul del valor esperat de $T(n)$ és correcte.



Francesc F. Ferri

Universitat de València

COMENTARI DE LLIBRE

Peter J. Diggle

Times Series. A Biostatistical Introduction

Oxford University Press, N.Y. (1990), 254 pp.

Este libro es una obra muy didáctica, especialmente dirigida a investigadores de la Medicina y la Biología. Las principales características que diferencian este libro, de otros textos que tratan el mismo tema, es que, está enfocado a la resolución de problemas biológicos, con una metodología matemática sencilla y además presenta un listado completo de los datos que corresponden a los problemas que propone; esto último permite reproducir al lector los resultados que aparecen en el texto.

A pesar de ser un libro especialmente enfocado a la Bioestadística, puede ser muy útil a otros profesionales relacionados con la economía y la ingeniería que deseen consultar algún aspecto específico del tema.

El libro consta de 8 capítulos y apéndices. En los capítulos 1 y 2 se introducen los conceptos básicos y los métodos descriptivos del análisis de las series temporales. En los capítulos 3 y 4 se desarrolla la metodología referente a los procesos estacionarios y al análisis espectral. El capítulo 5 estudia las series de datos de medidas repetidas. Los capítulos 6 y 7 están dedicados a estudiar las series temporales ARIMA con el enfoque Box-Jenkins. En el capítulo 8 se hace una breve pero interesante introducción a las series temporales bivariantes. Finalmente, en los apéndices se da un listado completo de los datos que se utilizan en los ejemplos expuestos en el texto y una introducción al modelo lineal general.

En resumen, el libro resulta idóneo para médicos y en general para investigadores relacionados con temas biológicos y epidemiológicos aunque no tenga una sólida base matemática, ya que da una visión muy amplia y sencilla de las series temporales.

M. Ríos

Butlleta de subscripció a la revista **Qüestiió**

Nom i cognoms _____

_____

Empresa/Institució _____

_____

Adreça _____

_____

Codi postal _____ ciutat _____

Tel. _____ Fax _____

Data d'expedició _____

Signatura: DNI o NIF _____

Desitjo subscriure'm a **Qüestiió** per a l'any 1995.
El preu de la subscripció és de 2.700 PTA.

Forma de pagament

☐ Transferència al compte de la Caixa de Catalunya número 6985/77, Agència 100, Comte d'Urgell 162, 08036 Barcelona

☐ Domiciliació bancària

☐ Taló nominatiu a l'Institut d'Estadística de Catalunya

☐ Gir postal

☐ En efectiu

Retornar aquesta butlleta (o una fotocòpia) a:

Qüestiió:
Institut d'Estadística de Catalunya
Via Laietana, 58
08003 Barcelona

Autorització de domiciliació bancària per al pagament de les subscripcions anuals de la revista **Qüestiió**

El sotasignat _____

autoritza el Banc/Caixa _____

Agència núm. _____ adreça _____

Codi postal _____ ciutat _____

a abonar a la revista **Qüestiió** amb càrrec al meu compte número _____

_____, les subscripcions a **Qüestiió**.

_____ , a _____ d _____ de 19 _____

(Signatura)

El sotasignat _____

autoritza la revista **Qüestiió** a carregar al meu compte número _____

_____ , al Banc/Caixa _____

_____

Agència núm. _____ adreça _____

Codi postal _____ ciutat _____

l'import de les tarifes vigents de les subscripcions a la revista **Qüestiió.**

_____ , a _____ d _____ de 19 _____

(Signatura)