Boletín de Estadística e Investigación Operativa
Vol. 25, No. 1, Febrero 2009, pp. 32-41

ESTADÍSTICA OFICIAL

Selective editing as a stochastic optimization problem

Ignacio Arbués, Margarita González and Pedro Revilla
Instituto Nacional de Estadística
✉ iarbues@ine.es, mgonzalez@ine.es, previlla@ine.es

Abstract

In this paper, we approach the problem of selective editing by proposing a formal definition of selection strategy. Using this definition, the search for an efficient selective editing method can be expressed as an optimization problem with stochastic constraints. We show how to solve this problem by duality methods. We also present the results of a practical application.

Keywords: Selective Editing; Score Function; Stochastic Programming.
AMS Subject classifications: 90C15, 90C46, 90C90.

1. Introducción

Disponer de métodos eficientes de depuración es fundamental para los organismos estadísticos. La depuración manual exhaustiva es considerada ineficiente, puesto que la mayor parte del trabajo de depuración no tiene consecuencias en el nivel agregado e incluso puede deteriorar la calidad de los datos (ver [6] y [3]).

Los métodos de depuración selectiva son estrategias para seleccionar un subconjunto de los cuestionarios recogidos en una encuesta para someterlos a una depuración minuciosa. Esta depuración puede hacerse de formas diversas, pero en general implica la intervención humana y por tanto, un gasto en personal que conviene reducir. Una razón por la que es conveniente seleccionar algunos cuestionarios es que depurando ciertas unidades es más probable que mejore la calidad que si se depuran otras. Esto puede ser debido a que unas son más sospechosas de contener un error o a que en caso de tenerlo, éste probablemente tenga más impacto en los agregados que se calculan con los datos de los cuestionarios. Por tanto, es razonable suponer que una buena estrategia de depuración selectiva puede permitir que se consigan hasta cierto punto dos objetivos: buena calidad de las estimaciones agregadas y reducido trabajo de depuración. A menudo esto se hace mediante una función *score* (FS) que se emplea para dar

prioridad a algunas unidades. Cuando para cada unidad se recogen varias variables, se pueden calcular diferentes FS *locales* y combinarlas en una *global*. Así, las unidades cuya FS excede un cierto umbral, son depuradas manualmente.

En consecuencia, es necesario decidir: (a) las FS locales; (b) cómo combinarlas en una global (suma, máximo, ...) y (c) el umbral. Hasta ahora, estas cuestiones han sido tratadas de manera empírica debido a la falta de una base teórica. En [5], [7] y [3] se proponen algunas directrices, pero en esencia se depende del criterio del experto.

En la sección 2, se plantea formalmente el problema, en las secciones 3 y 4 mostramos como resolver dos variantes del problema, en la sección 5 se propone un método para calcular ciertos momentos condicionales que se requieren para obtener la solución, en la sección 6 presentamos los resultados de una aplicación práctica y finalizamos en la sección 7 con unas conclusiones. Los detalles matemáticos están explicados en [1], que puede ser obtenido del primer autor.

2. El problema de la depuración selectiva

Introduzcamos un poco de notación,

- x_t^{ij} es el valor *verdadero* de la variable j en el cuestionario i en el periodo t, con $i = 1, \ldots, N$ y $j = 1, \ldots, q$.
- $\tilde{x}_t^{ij} = x_t^{ij} + \varepsilon_t^{ij}$ es el valor *observado*, siendo ε_t^{ij} el error de observación.
- $X_t^k = \sum \omega_{ij}^k x_t^{ij}$ es el k-ésimo estadístico calculado con los valores verdaderos, donde $k = 1, \ldots, p$.

Suponemos que x_t^{ij} y ε_t^{ij} son variables aleatorias con respecto al espacio probabilístico $(\Omega, \mathcal{F}, P)$. Puede haber otras variables relevantes, como $\tilde{x}_t^{ij}$, x_s^{ij} para $s < t$ o incluso variables de otras encuestas. Denotaremos por $\mathcal{G}_t$ la $\sigma-$álgebra generada por toda la información disponible hasta t. Para simplificar la notación, omitimos el subíndice t cuando no hay riesgo de ambigüedad.

Nuestro objetivo es encontrar una adecuada estrategia de selección, que debería indicar para cada i, si el cuestionario i es depurado o no, empleando la información disponible. En realidad, vamos a permitir que se determine sólo la probabilidad de depuración de una unidad, dejando una cierta indeterminación

Definición 2.1. *Una estrategia de selección (ES) con respecto a $\mathcal{G}_t$ es un vector aleatorio $\mathcal{G}_t$-medible, $r = (r_1, \ldots, r_N)^T$ tal que $r_i \in [0, 1]$.*

Denotamos por $S(\mathcal{G}_t)$ el conjunto de todas las ES con respecto a $\mathcal{G}_t$. La interpretación de r es que el cuestionario i es depurado con probabilidad $1 - r_i$. El permitir $0 \leq r_i \leq 1$ en lugar de obligar a $r_i \in \{0, 1\}$ es conveniente tanto desde el punto de vista teórico como del práctico porque así, el conjunto de estrategias es convexo y podemos emplear técnicas de optimización más sencillas

que las de la programación entera. Si para un cierto i, $r_i \in (0,1)$, la unidad es depurada si $\chi_t^i < r_i$, donde χ_t^i es una variable aleatoria uniforme en $[0,1]$, e independiente de cualquier otra de las que se consideren. Denotamos por $\tilde{r}_i$ la variable indicatriz del suceso $\chi_t^i < r_i$ y $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_N)^T$. Si una ES cumple que $r_i \in \{0,1\}$ c.s., $\tilde{r} = r$ c.s. y decimos que r es entera. En nuestra aplicación, las soluciones son aproximadamente enteras.

También es conveniente tener una definición del concepto de función *score*.

Definición 2.2. *Sea r una ES, $\delta = (\delta_1, \ldots, \delta_N)^T$ un vector aleatorio y $\Theta \in \mathbb{R}$, tal que $r_i = 1$ si y solo si $\delta_i \leq \Theta$. Entonces, decimos que δ es una función* score *que genera r con el umbral Θ.*

Para plantear formalmente nuestro problema, suponemos que tras la depuración manual, se obtienen los valores verdaderos. Así, solo tenemos que considerar los valores observados y verdaderos. Definimos el estadístico depurado como el calculado con los valores obtenidos tras depurar los que hayan sido seleccionados y lo denotamos por $X^k(r) = \sum \omega_{ij}^k (x_t^{ij} + \tilde{r}_i \varepsilon_t^{ij})$.

La calidad de $X^k(r)$ debe ser medida con arreglo a alguna función de pérdida. En este trabajo, solo consideramos el error cuadrático, $(X^k(r) - X^k)^2$, quedando la adaptación a otras funciones para futuros desarrollos. El valor de la función de pérdida se puede escribir como

$$(X^k(r) - X^k)^2 = \sum_{i,i'} \epsilon_i^k \epsilon_{i'}^k \tilde{r}_i \tilde{r}_{i'}, \tag{2.1}$$

donde $\epsilon_i^k = \sum_j \omega_{ij}^k \varepsilon_t^{ij}$. O en forma matricial como $(X^k(r) - X^k)^2 = \tilde{r}' E^k \tilde{r}$, con $E^k = \{E_{i,i'}^k\}_{i,i'}$ y $E_{i,i'}^k = \epsilon_i^k \epsilon_{i'}^k$.

Ahora podemos plantear el problema de la selección como un problema de optimización:

$$\begin{array}{lll} [P_Q] & \text{máx}_r & \mathbf{E}[1^T \tilde{r}] \\ & \text{sujeto a} & r \in S(\mathcal{G}_t) \\ & & \mathbf{E}[\tilde{r}^T E^k \tilde{r}] \leq e_k^2, k = 1, \ldots, p \end{array}$$

Éste es un problema de optimización lineal con restricciones cuadráticas, que tiene la dificultad de que se busca la solución en un espacio de dimensión infinita. En la sección 4 mostramos como resolverlo.

Analicemos ahora la expresión (2.1). Podemos descomponerla como

$$(X^k(r) - X^k)^2 = \sum_i (\epsilon_i^k)^2 \tilde{r}_i + \sum_{i \neq i'} \epsilon_i^k \epsilon_{i'}^k \tilde{r}_i \tilde{r}_{i'}. \tag{2.2}$$

El primer término de la derecha de (2.2) mide el impacto individual de cada error, independientemente de su signo. En el segundo término, los sumandos

son negativos cuando los factores tienen signos opuestos. Por tanto, para reducir el error total, sería conveniente dejar sin depurar parejas de unidades con distinto signo. La no-linealidad del segundo término complica los cálculos, así que estudiaremos también la versión despreciando esa parte. Si definimos $D^k = (D_1^k, \ldots, D_N^k)^T$, con $D_i^k = (\epsilon_i^k)^2$, podemos plantear

$$\begin{array}{ll} [P_L] \quad \text{máx}_r & \mathbf{E}[1^T \tilde{r}] \\ \text{s.a.} & r \in S(\mathcal{G}_t), \mathbf{E}[D^k \tilde{r}] \leq e_k^2, k = 1, \ldots, p. \end{array}$$

En la sección 3 veremos que la solución viene dada por una cierta FS. Puesto que no hay justificación teórica para despreciar la parte cuadrática, el problema lineal debe justificarse empíricamente mediante su eficacia en la práctica.

3. Caso lineal

Vamos a resolver el problema $[P_L]$ mediante un método de dualidad, pero antes debemos expresar la función objetivo como $\mathbf{E}[1^T r]$ y la restricción como $\mathbf{E}[\Delta^k r] \leq e_k^2$, donde $\Delta^k = \mathbf{E}[D^k | \mathcal{G}_t]$. Definamos ahora la lagrangiana $\mathcal{L}(r, \lambda) = \mathbf{E}[1^T r] - \lambda^T \mathbf{E}[\Delta^k r - e_k^2]$. En [1], se demuestra que bajo hipótesis no muy restrictivas, si $\bar{\lambda}$ es una solución del problema dual,

$$[D] \quad \text{mín}_{\lambda \geq 0} \quad \varphi(\lambda)$$

con $\varphi(\lambda) = \text{máx}_r \mathcal{L}(r, \lambda)$, entonces $r = \arg\text{máx}\, \mathcal{L}(\cdot, \bar{\lambda})$ es una solución del problema primal $[P_L]$. Como Δ^k es conocido, la maximización de $\mathcal{L}$ respecto a r se reduce al problema determinista

$$\begin{array}{ll} \underset{r}{\text{máx}} & 1^T r - \sum_k \lambda_k (\Delta^k r - e_k^2) \qquad (3.1) \\ \text{s.a.} & r_i \in [0, 1] \qquad (3.2) \end{array}$$

Aplicando las condiciones de Karush-Kuhn-Tucker (ver [2]) tenemos que,

$$r_i = \begin{cases} 1 & \text{if} \quad \lambda^T \Delta_i < 1 \\ 0 & \text{if} \quad \lambda^T \Delta_i > 1 \end{cases} \qquad (3.3)$$

donde $\Delta_i = (\Delta_i^1, \ldots, \Delta_i^p)^T$. El caso $\lambda^T \Delta_i = 1$ se puede despreciar como un suceso de probabilidad cero, ya que para datos cuantitativos, la distribución de Δ_i es continua. Por tanto, la solución a $[P_L]$ es la ES generada por la FS $\delta_i = \lambda^T \Delta_i$ con el umbral 1.

En la sección 5 describimos como usar un modelo para el cálculo de Δ^k. Para estimar $\varphi(\lambda)$ cambiamos la esperanza por la media muestral, donde la muestra puede ser simulada u obtenida a partir de datos reales si se dispone de ellos. Si

tenemos una muestra de M valores, usamos la estimación

$$\hat{\varphi}(\lambda) = \frac{1}{M}\sum_{l=1}^{M} L(\lambda, r^l(\lambda)).$$

La minimización de $\hat{\varphi}$ se hará por métodos numéricos.

4. Caso cuadrático

El problema cuadrático plantea dificultades particulares. Si intentamos aplicar el método de dualidad como en la sección anterior, expresamos la restricción como $\mathbf{E}[r^T\Gamma^k r + (\Delta^k)^T r] \leq e_k^2$, donde $\Gamma^k = \{\Gamma_{ij}^k\}_{ij}$ y

$$\Gamma_{ij}^k = \begin{cases} \mathbf{E}[\epsilon_i^k \epsilon_j^k | \mathcal{G}_t] & \text{if} \quad i \neq j \\ 0 & \text{if} \quad i = j \end{cases},$$

pero por desgracia, las matrices Γ^k son indefinidas, luego la restricción no es convexa. Podemos superar esta dificultad reemplazando la restricción por otra convexa y que bajo ciertas hipótesis es equivalente a la original.

Consideremos la restricción $\mathbf{E}[r^T M^k r + (v^k)^T r] \leq e_k^2$, donde $M_{ij}^k = m_i^k m_j^k$, $m_i^k = \mathbf{E}[\epsilon_i^k | \mathcal{G}_t]$, $v_i^k = \mathbf{V}[\epsilon_i^k | \mathcal{G}_t]$. Se puede demostrar (ver [1]) que si r es una solución entera al problema con la nueva restricción y $\mathbf{E}[\epsilon_i^k \epsilon_j^k | \mathcal{G}_t] = m_i^k m_j^k$ para $i \neq j$, entonces r es una solución al problema original. Hemos comprobado que la condición de integridad se cumple aproximadamente en nuestra aplicación. La hipótesis sobre la esperanza condicionada es restrictiva y su relajación puede ser estudiada en futuros desarrollos. En la sección 5 describimos un método para obtener Σ^k y v^k. En [1] se describe como resolver con poco coste computacional el problema de programación cuadrática que resulta.

5. Momentos condicionales basados en un modelo

La aplicación práctica de los resultados de las secciones anteriores requiere un método para calcular los momentos condicionales del error que aparecen. En esta sección omitimos el índice j para simplificar la notación, pero los resultados son aplicables para el caso en el que hay varias variables por cuestionario.

Sea $\mathcal{H}_t$ la $\sigma-$álgebra generada por toda la información disponible en t a excepción de $\tilde{x}_t^i$. Así, $\mathcal{G}_t = \sigma(\tilde{x}_t^i, \mathcal{H}_t)$. Sea $\hat{x}_t^i$ un predictor de x_t^i calculado empleando la información de $\mathcal{H}_t$, es decir, $\mathcal{H}_t$-medible. Denotamos el error de predicción por $\xi_t^i = x_t^i - \hat{x}_t^i$. Supongamos que,

(a) ξ_t^i y η_t^i están distribuidas como normales independientes con media cero y varianzas ν_i^2 y σ_i^2 respectivamente.

(b) $\varepsilon_t^i = \eta_t^i e_t^i$, donde e_t^i tiene una distribución de Bernoulli tomando los valores

1 y 0 con probabilidades p y $1-p$ y es independiente de ξ_t^i y η_t^i.

(c) ξ_t^i, η_t^i y e_t^i son conjuntamente independientes de $\mathcal{H}_t$.

Bajo estas hipótesis, los momentos condicionales de ε_t^i con respecto a $\mathcal{G}_t$ dependen solo de $u_t^i = \hat{x}_t^i - \tilde{x}_t^i$, es decir, de la diferencia entre el valor predicho y el observado. En las fórmulas siguientes, también omitimos i y t por simplicidad.

$$\mathbf{E}[\varepsilon|\mathcal{G}] = \frac{\sigma^2}{\sigma^2+\nu^2}u\zeta \tag{5.1}$$

$$\mathbf{E}[\varepsilon^2|\mathcal{G}] = \left[\frac{\sigma^2\nu^2}{\sigma^2+\nu^2} + \left(\frac{\sigma^2}{\sigma^2+\nu^2}\right)^2 u^2\right]\zeta \tag{5.2}$$

donde,

$$\zeta = \frac{1}{1+\frac{1-p}{p}\left(\frac{\nu^2}{\sigma^2+\nu^2}\right)^{-1/2}\exp\{-\frac{u^2(\sigma^2)}{2\nu^2(\sigma^2+\nu^2)}\}}. \tag{5.3}$$

6. Caso práctico

En esta sección presentamos los resultados de la aplicación de los métodos descritos a los datos de la encuesta de Cifras de Negocios y Entradas de Pedidos en la industria que lleva a cabo el Instituto Nacional de Estadística (datos disponibles en http://www.ine.es/inebmenu/mnu_industria.htm). En el momento del estudio, se disponía de datos mensuales desde enero de 2002 hasta septiembre de 2006 ($t = 1\ldots,57$) recogidos para una muestra de alrededor de $N = 13,500$ unidades. Solo dos de las variables recogidas fueron consideradas: importe neto de la cifra de negocios (x_t^{i1}) y nuevos pedidos recibidos (x_t^{i2}). Estas dos variables son agregadas separadamente para obtener los dos indicadores, así que $p = q = 2$ y $\omega_{i2}^1 = \omega_{i1}^2 = 0$.

Para aplicar (5.1)-(5.3) es necesario ajustar un modelo a los datos, pero antes empleamos la transformación $y_t^{ij} = \log(x_t^{ij} + m)$, donde m es una constante positiva ajustada por máxima verosimilitud ($m \approx 10^5$€). Para recuperar los momentos condicionales de la variable original, se pueden usar las propiedades de la log-normal o un desarrollo de Taylor de primer orden, que resulta $\mathbf{E}[(\tilde{x}_t^{ij} - x_t^{ij})^2|\mathcal{G}_t] \approx (\tilde{x}_t^{ij} - m)^2\mathbf{E}[(\tilde{y}_t^{ij} - y_t^{ij})^2|\mathcal{G}_t]$. En nuestra aplicación hemos empleado la aproximación. También hemos observado que si $\tilde{x}_t^{ij} - m$ es sustituido por una media de los últimos 12 valores de $\tilde{x}_t^{ij}$ el resultado es más robusto frente a valores pequeños de $\tilde{x}_t^{ij} - m$.

El modelo que hemos planteado para las variables transformadas es muy simple. Suponemos que x_t^{ij} y $x_t^{i'j'}$ son independientes si $(i,j) \neq (i',j')$. Para

cada par (i,j) seleccionamos para la serie x_t^{ij} uno de los siguiente modelos

$$(1-B)y_t^{ij} = a_t \tag{6.1}$$
$$(1-B^{12})y_t^{ij} = a_t \tag{6.2}$$
$$(1-B^{12})(1-B)y_t^{ij} = a_t \tag{6.3}$$

donde B es el operador de retardos $Bu_t = u_{t-1}$ y a_t es un proceso de ruido blanco. El criterio de selección es el de mínimo error residual cuadrático. Una vez seleccionado el modelo, lo empleamos para calcular la predicción $\hat{y}_t^{ij}$ y su desviación típica ν_{ij}. La desviación típica *a priori* del error de observación y de la probabilidad de error se suponen constantes (la transformación logarítmica hace que tenga sentido suponer que los errores de observación tienen la misma d.t.), los denotamos por σ_j y p_j con $j = 1,2$ y son estimados usando datos históricos de la encuesta. Para ello, nos aprovechamos de que tanto la primera versión recibida de cada dato como las que posteriormente resulten de posibles correcciones quedan almacenadas en una base de datos. Para este estudio, la primera versión es considerada como dato *observado* y la definitiva como dato *verdadero*.

Una vez que hemos calculado σ_j, p_j, ν_{ij} y u_t^{ij}, usamos (5.1)-(5.3) para calcular los momentos condicionales, Δ^k, Σ^k y v^k.

6.1. Restricciones del error

Queremos comprobar si las restricciones al error en $[P_L]$ y $[P_Q]$ se cumplen. Para esto, resolvemos ambos problemas para las cotas de error

$$e_{1l}^2 = e_{2l}^2 = e_l^2 = [s_0^{((l-1)/(b-1))} s_1^{((b-l)/(b-1))}]^2$$

con $l = 2,4,\ldots,b$, $b = 20$, $s_0 = 0{,}025$ y $s_1 = 1$.

La función dual φ es estimada usando una muestra de longitud h obtenida de los datos reales. Para cada periodo desde octubre de 2005 hasta septiembre de 2006 y para cada l, se obtiene una solución $r(t,l)$. La media a lo largo de t de los errores cuadráticos que quedan tras la depuración se calcula como

$$\hat{e}_{kl}^2 = \frac{1}{12}\sum_{t=t_0}^{t_0+11} r(t,l)^T E^k r(t,l)$$

Repetimos estos cálculos para $h = 1,3,6$ y 12 usando las versiones lineal y cuadrática. Los resultados para $h = 6$ se muestran en la tabla 1 (el resto se pueden ver en [1]). Para cada l, presentamos la cota fijada, los cocientes $\hat{e}_{kl}/e_{kl}$ con $k = 1,2$, y el número de unidades depuradas en promedio, todo ello tanto para la versión lineal como la cuadrática.

Se observa una tendencia a subestimar el error cuando las cotas son muy

pequeñas y a sobreestimarlo cuando las cotas son grandes. Con el método cuadrático se obtienen errores más pequeños, pero a costa de editar más unidades.

Tabla 1. Comparación entre las versiones lineal y cuadrática.

	Lineal			Cuadrática		
e_l	$\frac{\hat{e}_{1l}}{e_l}$	$\frac{\hat{e}_{2l}}{e_l}$	n	$\frac{\hat{e}_{1l}}{e_l}$	$\frac{\hat{e}_{2l}}{e_l}$	n
.0250	1.89	3.20	414.8	1.86	2.50	578.4
.0369	2.04	2.43	257.6	1.42	1.82	401.9
.0544	1.58	1.87	157.7	0.96	1.17	390.5
.0801	1.10	1.25	95.9	1.31	1.30	283.4
.1182	0.96	1.18	56.2	1.32	1.22	140.3
.1742	0.97	1.31	30.5	1.24	0.92	77.8
.2569	0.71	1.09	15.0	0.92	0.67	35.4
.3788	0.74	0.77	6.8	0.64	0.66	28.0
.5585	0.57	0.57	2.7	0.56	0.57	6.8
.8235	0.42	0.38	1.2	0.39	0.38	2.9

6.2. Comparación de funciones *score*

Pretendemos ahora comparar la eficacia de nuestro método con la FS descrita en [4], $\delta_i^0 = \omega_i|\tilde{x}^i - \hat{x}^i|$, donde $\hat{x}_i$ es una predicción de x_i. En [4] se propone usar como predicción el último valor de la misma variable en periodos anteriores. Hemos considerado una modificación de esta función, δ^1, simplemente cambiando la predicción por la que se obtiene mediante los modelos (6.1)-(6.3). Finalmente, δ^2 es la FS calculada usando (5.1)-(5.3). En esta comparación, la combinación de FS locales en una global se ha hecho simplemente sumándolas. Medimos la efectividad de una FS mediante $E^j = \sum_n E^j(n)$, con $E^j(n) = [\sum_{i\geq n}^{N} \omega_i^j(\tilde{x}^{ij} - x^{ij})]^2$, donde consideramos las unidades ordenadas descendentemente según la correspondiente FS. La cantidad $E^j(n)$ puede ser considerada como una estimación del error que queda tras editar n unidades. Los resultados incluidos en la tabla 2 muestran que δ^2 mejora a δ^1, que a su vez es más eficaz que δ^0.

Tabla 2. Comparación de FS.

FS	Cifras de Negocios	Nuevos Pedidos
δ^0	0.44	1.33
δ^1	0.38	0.45
δ^2	0.26	0.37

7. Conclusiones

Hemos descrito un marco teórico para tratar el problema de la depuración selectiva, definiendo el concepto de estrategia de selección. La búsqueda de una

adecuada ES se presenta como un problema de optimización lineal con restricciones cuadráticas. Consideramos también una versión modificada con restricciones lineales. Mostramos un método práctico para resolver ambos problemas.

La FS correspondiente a la versión lineal mejora la de referencia. Ambas versiones producen resultados en los que se satisfacen aproximadamente las restricciones, salvo para valores muy pequeños de las cotas. El método cuadrático parece más conservador y así, con él las cotas se satisfacen mejor, pero hay que depurar más unidades. Por otra parte, la implementación del método lineal es más fácil y computacionalmente menos costosa.

Referencias

[1] I. Arbués, M. González, P. Revilla (2008), A Class of Stochastic Optimization Problems with Application to Selective Data Editing, documento de trabajo, Instituto Nacional de Estadística.

[2] M. S. Bazaraa, H. D. Sherali y C. M. Shetty (1993), *Nonlinear Programming: Theory and Algorithms*, Nueva York: Wiley.

[3] L. Granquist (1997), The new view on editing, *International Statistics Review*, **65**, 381–387.

[4] D. Hedlin (2003), Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics, *Journal of Official Statistics*, **19**, 177–199.

[5] M. Latouche y J. Berthelot (1992), Use of a score function to prioritize and limit recontacts in editing business surveys, *Journal of Official Statistics*, **8**, 389–400.

[6] J. Berthelot y M. Latouche (1993), Improving the efficiency of Data Collection: A Generic Respondent Follow-Up Strategy for Economic Surveys, *Journal of Business and Economic Statistics*, **11**, 417–424.

[7] D. Lawrence y R. McKenzie (2000), The general application of Significance Editing, *Journal of Official Statistics*, **16**, 243–253.

Acerca de los autores

Ignacio Arbués Lombardía es licenciado en Matemáticas (especialidad en Matemática Aplicada y Computación) por la Universidad de Oviedo, donde fue profesor asociado. En el sector privado, ha trabajado en predicción de la demanda de energía mediante modelos de series temporales (Grupo Apex) y en el desarrollo de métodos de estimación para control de sistemas de navegación por satélite (GMV). Desde 2001 pertenece al Cuerpo Superior de Estadísticos del Estado, con destino en la Subdirección de Estadísticas Industriales y Agrarias

del INE. Ha publicado en *Journal of Time Series Analysis* y próximamente en *Journal of Multivariate Analysis*.

Margarita González Villa es licenciada en Ciencias Económicas (especialidad de Economía Cuantitativa) por la Universidad Autónoma de Madrid. Pertenece al Cuerpo de Diplomados de Estadística del Estado (1985-1986) y desde junio de 1986, al Cuerpo Superior de Estadísticos del Estado. Desde 1986 su trabajo siempre ha estado relacionado con la implantación, control, edición, obtención, difusión y análisis de distintos indicadores coyunturales de la industria tales como los Índices de Precios Industriales y los Índices de Producción Industrial. Actualmente, es Jefa de Área, de los Índices de Cifras de Negocios y de Entradas de Pedidos.

Pedro Revilla Novella es licenciado en Ciencias Económicas por la Universidad Autónoma de Madrid y máster en Series Temporales y Modelos Econométricos y Dinámicos (Centro de Formación del Banco de España). Es Estadístico Superior del Estado desde 1983 y Subdirector General de Estadísticas Industriales y Agrarias en el INE desde 1987. También es profesor asociado en la Universidad Carlos III de Madrid (desde 1996) y anteriormente en la Universidad de Salamanca. Miembro electo del International Statistical Institute, pertenece al *Steering Committee* del grupo de trabajo de depuración e imputación y al de la publicación *Statistical Data Editing, Vol. 3* (ONU).