Teledetección aplicada a la gestión de recursos naturales y medio litoral marino. VII Congreso Nacional de Teledetección.
Universidad de Santiago de Compostela - Asociación Española de Teledetección.
Santiago de Compostela, 26-28 de junio de 1997. pp. 285-288.
Editores: Carlos Hernández Sande y Juan E. Arias Rodríguez.

UN ESTIMADOR MÁS EFICIENTE QUE EL DE REGRESIÓN PARA LA ESTIMACIÓN DE LOS USOS DEL SUELO POR MUESTREO DEL TERRENO Y TELEDETECCIÓN, EN PEQUEÑAS ÁREAS. *

L. Ambrosio Flores [(1)], R. Escudero Barroso [(2)], J. Fernández Casals [(2)], L. Iglesias Martínez [(1)], A. Porcuna Fdez-Monasterio [(2)]

(1) Dpto. de Economía y Ciencias Sociales Agrarias. Escuela Técnica Superior de Ingenieros Agrónomos. Universidad Politécnica de Madrid. Ciudad Universitaria s/n. 28040 Madrid.
E-mail: Flores@eco.etsia.upm.es

(2) Tecnologías y Servicios Agrarios S.A. (TRAGSATEC). Avda. Ciudad de Barcelona, 118-124. 28007 Madrid.

1.- Introducción.

Cada vez más, se requieren estimaciones precisas de los usos del suelo en pequeñas demarcaciones territoriales tales como municipios o polígonos de riego. Ni el estimador de expansión directa, ni el clásico de regresión son eficientes en pequeñas áreas. En este trabajo se propone un estimador alternativo más eficiente.

2.- La población.

La zona de estudio se considera dividida en "m" pequeñas áreas. Sea N_i el número de unidades de muestreo (segmentos) en la i-ésima pequeña área (i=1, 2 ..., m). Sea y_{ij} la superficie ocupada por un determinado uso del suelo, en el j-ésimo segmento (j=1, 2... , N_i) de la i-ésima pequeña área (i=1, 2 ..., m).

Se pretende estimar el total y la media de la variable "y" en cada una de las "m" pequeñas áreas, a partir de una muestra sobre el terreno de $n = \sum_{i=1}^{m} n_i$ segmentos [siendo n_i el número de segmentos que caen en la i-ésima pequeña área (para algún "i" n_i puede ser nulo)] y de la Teledetección. El total de y_{ij} en la i-ésima pequeña área es:

$$Y_i = \sum_{j=1}^{N_i} y_{ij} = \sum_{j=1}^{n_i} y_{ij} + \sum_{j'=1(j'\neq j)}^{N_i-n_i} y_{ij'} =$$

$$= n_i \, \overline{y}_i + (N_i - n_i) \, \overline{Y}_i^*$$

donde:

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

$$\overline{Y}_i^* = \frac{1}{N_i - n_i} \sum_{j'=1(j'\neq j)}^{N_i-n_i} y_{ij'}$$

y la media es:

$$\overline{Y}_i = \frac{Y_i}{N_i} = f_i \, \overline{y}_i + (1 - f_i) \, \overline{Y}_i^*$$

donde: $f_i = \frac{n_i}{N_i}$

3.- La información disponible.

Se dispone de los pares de valores (y_{ij}, x_{ij}) para j= 1, 2 ... n_i e i= 1, 2 ... m. "x_{ij}" es la superficie clasificada por Teledetección como del uso en estudio, en el j-ésimo segmento de la muestra de la i-ésima pequeña área. Se conoce, además, la superficie total en cada pequeña área clasificada por Teledetección como del uso en estudio.

4.- El estimador.

El estimador de la media $\overline{Y}_i$ que se propone es:

$$\hat{\overline{Y}}_i = f_i \, \overline{y}_i + (1 - f_i) \, \hat{\overline{Y}}_i^* \qquad [1]$$

donde $\hat{\overline{Y}}_i^*$ es el Predictor Lineal Insesgado y Óptimo (PLIO) de $\overline{Y}_i^*$. Este estimador se define a partir del modelo mixto [BATTESE et al (1988)]:

$$y_{ij} = \beta_1 + \beta_2 \, x_{ij} + v_i + \varepsilon_{ij} \qquad [2]$$

donde:

$\underline{\beta}^T = [\, \beta_1 \beta_2 \,]$ (efectos fijos),
v_i (efectos aleatorios) $\rightarrow N(0, \sigma_v^2)$,
ε_{ij} (perturb. aleatorias) $\rightarrow N(0, \sigma_e^2)$,
v_i y ε_{ij} independientes.

El PLIO de $\overline{Y}_i^*$ es:

$$\hat{\overline{Y}}_i^* = \hat{\beta}_1 + \hat{\beta}_2 \, \overline{X}_i^* + \hat{v}_i \qquad [3]$$

donde:

$$\overline{X}_i^* = \frac{1}{N_i - n_i} \sum_{j'=1(j'\neq j)}^{N_i-n_i} x_{ij'}$$

$$\underline{\hat{\beta}}^T = [\, \hat{\beta}_1 \; \hat{\beta}_2 \,] =$$

$$= [\, (\underline{X}^T \, \underline{V}^{-1} \, \underline{X})^{-1} (\underline{X}^T \, \underline{V}^{-1} \, \underline{Y}) \,]^T$$

$\underline{X} = [\, \underline{1} \; \underline{x} \,]$; siendo $\underline{1}$ un vector columna (nx1) de unos y $\underline{x}$ el vector columna (nx1) de los valores de x_{ij} observados en la muestra.

$\underline{V}^{-1}$ = diagonal $(\underline{V}_1^{-1}, \underline{V}_2^{-1} \ldots, \underline{V}_i^{-1} \ldots \underline{V}_m^{-1})$

$$\underline{V}_i^{-1} = \frac{1}{\sigma_e^2} \, \underline{I}_{(ni)} - \frac{g_i}{n_i \, \sigma_e^2} \, \underline{1}_{(ni)} \, \underline{1}_{(ni)}^T$$

$\hat{v}_i = g_i \, (\overline{y}_i - \overline{\underline{x}}_i \, \underline{\hat{\beta}})$ es el PLIO de v_i, siendo:

$\overline{\underline{x}}_i = [\, 1 \; \overline{x}_i \,]$ donde:

* Este trabajo ha sido parcialmente financiado por TRAGSATEC en el marco de un convenio de colaboración suscrito con la E.T.S.I. Agrónomos de Madrid. Los datos empleados pertenecen al "Estudio del uso actual de la tierra en la Cuenca del Duero y especial atención a las tierras de regadío", realizado por TRAGSATEC para la Dirección General de Desarrollo Rural y del Medio Natural del M.A.P.A.

$$\overline{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

$\underline{Y}$ es el vector (nx1) de los valores de y_{ij} observados en la muestra.

Por sustitución de $\overline{\hat{Y}}_i^*$ de [3] en [1] se tiene:

$$\overline{\hat{Y}}_i = (1 - g_i)\, \overline{\underline{X}}_i\, \hat{\underline{\beta}} + g_i [\, \overline{y}_i + (\, \overline{\underline{X}}_i - \overline{\underline{x}}_i)\hat{\underline{\beta}}\,] \qquad [4]$$

donde:

$$\overline{\underline{X}}_i = [1 \;\; \overline{\underline{X}}_i]$$

$$\overline{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}$$

$$g_i = \sigma_v^2 / (\, \sigma_v^2 + \sigma_e^2 \,/\, n_i)$$

Las componentes σ_v^2 y σ_e^2 de la varianza de y_{ij} se estiman por:

$$\hat{\sigma}_e^2 = \hat{\underline{e}}^T \hat{\underline{e}} \,/\, (n - m - 1)$$

$$\hat{\sigma}_v^2 = [\, \hat{\underline{u}}^T \hat{\underline{u}} - (n-2)\, \hat{\sigma}_e^2 \,] / n_*$$

$$n_* = n - traza\,(\, \underline{X}^T \underline{X} \,)^{-1} \sum_{i=1}^{m} n_i^2\, \overline{\underline{x}}_i^T\, \overline{\underline{x}}_i$$

donde $\hat{\underline{e}}^T \hat{\underline{e}}$ y $\hat{\underline{u}}^T \hat{\underline{u}}$ son, respectivamente, la suma de los cuadrados de los residuos del modelo [2] ajustado para $\upsilon_i = 0$ y υ_i fijo en lugar de aleatorio.

5.- Error cuadrático medio del estimador.

El error Cuadrático Medio del Estimador viene dado por [PRASAD y RAO (1990)]:

$$\hat{ECM}(\overline{\hat{Y}}_i) = (1 - f_i)^2 [\, h_{1i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + h_{2i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2\, h_{3i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)] \qquad [5]$$

donde:

$$h_{1i}(\hat{\sigma}_v^2,\, \hat{\sigma}_e^2) = g_i \left(\frac{\hat{\sigma}_e^2}{n_i}\right) + (1 - f_i)^2 \frac{(N_i - n_i)}{N_i^2}$$

$$h_{2i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = \hat{\sigma}_e^2 (\, \overline{\underline{X}_i^*} - g_i \overline{\underline{x}_i}) \underline{A}^{-1} (\, \overline{\underline{X}_i^*} - g_i \overline{\underline{x}_i})^T$$

$$h_{3i}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = \frac{1}{n_i^2} \frac{1}{(\hat{\sigma}_v^2 + \frac{\hat{\sigma}_e^2}{n_i})^3} [(\hat{\sigma}_e^2)^2 Var(\hat{\sigma}_v^2) + (\hat{\sigma}_v^2)^2 Var(\hat{\sigma}_e^2) - 2\hat{\sigma}_e^2 \hat{\sigma}_v^2 Cov(\hat{\sigma}_e^2, \hat{\sigma}_v^2)]$$

donde:

$$A = \sum_{i=1}^{m} [\, \sum_{j=1}^{n_i} \underline{x}_{ij}^T \underline{x}_{ij} - g_i\, n_i\, \overline{\underline{x}}_i^T \overline{\underline{x}}_i \,]$$

$$\underline{x}_{ij} = [\, 1 \;\; x_{ij} \,]$$

$$Var(\hat{\sigma}_v^2) = \frac{2}{n_*^2} [\frac{1}{n-m-1} (m-1)(n-2)(\hat{\sigma}_e^2)^2 + 2\, n_* \hat{\sigma}_e^2 \sigma_v^2 + n_{**} (\hat{\sigma}_v^2)^2]$$

$$Var(\hat{\sigma}_e^2) = \frac{2 (\hat{\sigma}_e^2)^2}{n - m - 1}$$

$$Cov(\hat{\sigma}_e^2, \hat{\sigma}_v^2) = -\frac{1}{n_*} (m-1)\, Var(\hat{\sigma}_e^2)$$

donde:

$$n_* = n - traza\,(\, \underline{X}^T \underline{X} \,)^{-1} \sum_{i=1}^{m} n_i^2 \overline{\underline{x}}_i^T \overline{\underline{x}}_i =$$

$$= \sum_{i=1}^{m} n_i [\, 1 - n_i \overline{\underline{x}}_i \,(\, \underline{X}^T \underline{X} \,)^{-1} \overline{\underline{x}}_i^T \,]$$

$$n_{**} = \sum_{i=1}^{m} n_i^2 (\, 1 - n_i \overline{\underline{x}}_i \underline{A}_I^{-1} \overline{\underline{x}}_i^T \,) + traza[(\, \underline{A}_I^{-1} \sum_{i=1}^{m} n_i^2 \overline{\underline{x}}_i^T \overline{\underline{x}}_i)^2]$$

donde

$$A_I = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \underline{x}_{ij}^T \underline{x}_{ij}$$

por lo tanto:

$$n_{**} = \sum_{i=1}^{m} n_i^2 [\, 1 - \overline{\underline{x}}_i (\, \underline{X}^T \underline{X} \,)^{-1} \overline{\underline{x}}_i^T \,] = n_* - n + \sum_{i=1}^{m} n_i^2$$

6.- Precisión relativa respecto de los estimadores (i) Expansión Directa, (ii) Sintético y (iii) de Regresión.

(i) El estimador de Expansión Directa es:

$$\overline{\hat{Y}}_i = \overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

Los estimadores (ii) Sintético de Regresión y (iii) Clásico de Regresión son de la forma [HARTER (1983)]:

$$\overline{\hat{Y}}_i\, (\delta_i) = \overline{\underline{X}}_i \hat{\underline{\beta}} + \delta_i\, (\, \overline{y}_i - \overline{\underline{x}}_i \hat{\underline{\beta}} \,)$$

Para $\delta_i = g_i$ se obtiene el Estimador Óptimo, para $\delta_i = 0$ se obtiene el Estimador Sintético y para $\delta_i = 1$ se obtiene el Estimador de Regresión.

Los Errores Cuadráticos Medios respectivos son:

$$ECM[\hat{\overline{Y}}_i\,(0)] = ECM(\hat{\overline{Y}}_i) + g_i^2\,[\sigma_v^2 + \sigma_e^2 / n_i - \underline{\overline{x}}_i (\underline{X}^T \underline{V}^{-1} \underline{X})^{-1} \underline{\overline{x}}_i^T]$$

$$ECM[\hat{\overline{Y}}_i\,(1)] = ECM(\hat{\overline{Y}}_i) + (1 - g_i^2)\,[\sigma_v^2 + \sigma_e^2 / n_i - \underline{\overline{x}}_i (\underline{X}^T \underline{V}^{-1} \underline{X})^{-1} \underline{\overline{x}}_i^T] \quad [6]$$

donde $ECM(\hat{\overline{Y}}_i)$ es el del óptimo definido en [5].

La varianza del estimador de expansión directa es:

$$V(\hat{\overline{Y}}_i)_{EXP} = (1 - f)\frac{S_w^2}{n_i} + Q_i \frac{S_w^2}{n_i^2}$$

donde el estimador de S_w^2 es:

$$\hat{s}_w^2 = \sum_{i=1}^{m} (n_i - 1)\frac{s_i^2}{(n - m)}$$

$$s_i^2 = \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$$

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

$$Q_i = 1 - (N_i / N)$$

7.- Eficiencia de la Teledetección.

Los niveles de precisión de las estimaciones que se alcanzan con la ayuda de la Teledetección pueden ser alcanzados sin ella, pero a costa de un mayor tamaño de la muestra de segmentos. El tamaño de muestra necesario para alcanzar el nivel de precisión actual sin ayuda de la Teledetección es:

$$n_i^* = n_i \, ER$$

donde ER se evalúa mediante la expresión

$$ER = V(\hat{\overline{Y}}_i)_{EXP} \,/\, ECM(\hat{\overline{Y}}_i)_{OPT}$$

en la que $ECM(\hat{\overline{Y}}_i)_{OPT}$ se define en [5] y $V(\hat{\overline{Y}}_i)_{EXP}$ es la varianza del estimador de expansión directa.

8.- Estimación en áreas no muestreadas.

Para una pequeña área no muestreada, esto es, para la que $n_i = 0$, la estimación se tiene haciendo en [4] $n_i = 0$ con lo que el estimador óptimo se reduce al Sintético $\hat{\overline{Y}}_i\,(0) = \underline{\overline{X}}_i \hat{\beta}$, y así mismo el Error Cuadrático Medio se reduce a:

$$\hat{ECM}(\hat{\overline{Y}}_i) = \underline{\overline{X}}_i (\underline{X}^T \underline{V}^{-1} \underline{X})^{-1} \underline{\overline{X}}_i^T + \hat{\sigma}_v^2$$

9.- Estudio de un caso.

a) La Población.

La población está constituida por 59 pequeñas áreas, que son las Zonas Regables del Estado incluidas en la Cuenca Hidrográfica del Duero.

b) La Información Disponible.

Es la observada en una muestra de n=187 segmentos cuadrados de 500 x 500 m^2. El número de segmentos que caen en una pequeña área varía de cero a 17, con una media de 4.

Se consideran los siguientes usos en regadío: girasol, maíz y remolacha. Se dispone del dato relativo a la superficie de estos usos clasificada por Teledetección en cada zona de riego.

c) Los resultados.

En el **Cuadro 1** se recogen las estimaciones obtenidas para la superficie de remolacha en regadío mediante el PLIO [4] y se comparan los errores típicos (raíces cuadradas de los Errores Cuadráticos Medios o Varianzas) de este estimador con los de los otros estimadores considerados.

Se observa cómo el estimador propuesto es el más preciso, esto es, el de menor error típico. El de expansión directa es el menos preciso. Las precisiones de los estimadores Sintéticos y de Regresión dependen de g_i: la del sintético se acerca a la del óptimo cuando g_i tiende a 0 y cuando g_i tiende a 1 es la del de Regresión la que se aproxima a la del Óptimo.

La eficiencia de la Teledetección medida a través del tamaño de la muestra de segmentos necesaria para alcanzar los niveles de precisión actuales sin ayuda de la Teledetección, no varía substancialmente de una a otra pequeña área. Por término medio se requiere una muestra de 31 segmentos en cada pequeña área, frente a los 4 utilizados con la ayuda de la Teledetección.

En el **Cuadro 2** se recogen las estimaciones obtenidas para las zonas sin muestra.

d) Eficiencia Relativa de los estimadores PLIO y Clásico de Regresión.

Para zonas con tamaños de muestra muy reducidos ($1 <= n_i <= 10$), la Eficiencia Relativa del PLIO con respecto al estimador Clásico de Regresión (cociente entre el Error Cuadrático Medio del Clásico de Regresión y el del PLIO) es muy alta, hasta 12.75.

Para zonas con tamaños de muestra más altos, como en la zona del Páramo (n_i = 17), la Eficiencia Relativa baja a valores cercanos a 1.

Estos resultados se explican a partir de las expresiones [4] y [6]. Cuando n_i disminuye (aumenta) también disminuye (aumenta) g_i, por lo que en [6] el Error Cuadrático Medio del estimador Clásico de Regresión aumenta (disminuye), y la Eficiencia Relativa del PLIO con respecto a él también aumenta (disminuye).

Con tamaños de muestra altos en las pequeñas áreas, el empleo del estimador Clásico de Regresión no supone pérdidas notables de precisión respecto al PLIO y supone menos cálculos, por lo que puede ser preferible. En cambio, para tamaños de muestra pequeños, las pérdidas de precisión respecto del PLIO pueden ser notables.

REFERENCIAS.

BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error-component model for prediction of county crop areas using surveys and satellite data. Journal of the American Statistical Asociation, 83, 401, 28-36.

HARTER, R..M. (1983) Small area estimation using nested error models and auxiliary data. Ph. D. Iowa State University.

PRASAD, N. G. N., RAO, J.N.K. (1990): The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Asociation, 85, 409, 163-171.

Cuadro 1.-Estimaciones obtenidas para el cultivo de remolacha en regadío, errores típicos del estimador PLIO, del Sintético de Regresión, Clásico de Regresión y de Expansión Directa, Eficiencia Relativa de la Teledetección, en zonas muestreadas.

Zona de riego	Tamaño de la muestra	Estimación ha/seg. (PLIO)	Error Típico				Eficiencia relativa del PLIO con respecto al estimador Clásico de Regresión	Eficiencia (Tamaño de la muestra sin Teledetección) n_i
			PLIO	Sintético de Regresión	Clásico de Regresión	Expansión Directa		
Agueda	1	1.16	0,75	0,77	2,35	4,70	9.82	39
Alba de Tormes	1	3.17	0,72	0,75	2,34	4,71	10.56	42
Almar	4	2.90	0,68	0,77	1,20	1,85	3.11	29
Almazan	5	1.44	0,69	0,79	1,09	1,62	2.50	27
Aranda	2	4.49	0,74	0,79	1,66	2,87	5.03	30
Arlanzon	6	4.41	0,64	0,76	0,98	1,46	2.34	30
Arriola	4	1,76	0,71	0,79	1,21	1,85	2.90	27
Babilafuente	3	3,61	0,72	0,79	1,38	2,21	3.67	28
Bajo Carrion	2	2,76	0,73	0,78	1,68	2,86	5.30	30
Campillo	1	0,92	0,65	0,68	2,32	4,71	12.74	51
Carrion-Saldaña	13	2,77	0,56	0,75	0,70	0,95	1.56	37
Carrizo	2	1,22	0,74	0,79	1,68	2,87	5.15	30
Castañon	4	4,15	0,69	0,78	1,20	1,85	3.02	28
Castilla Norte	5	2,72	0,69	0,79	1,10	1,62	2.54	27
Castilla Ramal de Campos	10	3,49	0,60	0,77	0,80	1,09	1.78	32
Castilla Sur	2	4,25	0,74	0,79	1,68	2,87	5.15	30
Esla	9	2.28	0,62	0,78	0,84	1,16	1.84	31
Florida	2	3,02	0,72	0,77	1,67	2,87	5.38	32
Geria-Villamarciel	1	6,58	0,78	0,81	2,36	4,71	9.15	36
Ines	3	7,45	0,77	0,84	1,41	2,21	3.35	24
La Maya	2	1,65	0,73	0,78	1,68	2,87	5.30	31
La Retencion	2	3,37	0,74	0,78	1,68	2,87	5.15	30
La Vid	2	4,64	0,67	0,73	1,65	2,88	6.06	36
Macias-Picaveas	1	2,86	0,74	0,77	2,35	4,70	10.08	39
Manganeses	8	2,37	0,63	0,77	0,88	1,24	1.95	31
Nava de Campos	4	1,99	0,70	0,79	1,21	1,85	2.99	27
Palencia	4	3,32	0,70	0,78	1,20	1,85	2.94	28
Paramo	17	3,75	0,52	0,75	0,62	0,82	1.42	42
Pisuerga	8	4,56	0,64	0,78	0,88	1,24	1.89	30
Pollos	1	5,06	0,76	0,79	2,35	4,71	9.56	38
Porma Margen Izda	7	2,21	0,65	0,78	0,94	1,33	2.09	29
Presa Tierra	2	2,90	0,73	0,78	1,67	2,87	5.23	30
Riaza	4	5,39	0,72	0,81	1,22	1,85	2.87	26
San José	7	3,60	0,65	0,78	0,94	1,33	2.09	28
San Román y San Justo	2	1,15	0,65	0,71	1,64	2,88	6.37	38
Tera	4	2,00	0,71	0,79	1,21	1,85	2.90	27
Tordesillas	4	5,60	0,73	0,82	1,23	1,85	2.84	25
Toro-Zamora	2	4,32	0,74	0,79	1,68	2,86	5.15	30
Tramo Hidroeléctrico	3	0,95	0,67	0,74	1,35	2,21	4.06	33
Velilla	2	1,10	0,73	0,78	1,67	2,87	5.23	30
Villadangos	4	2,07	0,70	0,78	1,21	1,85	2.99	28
Villagonzalo	3	4,49	0,73	0,80	1,39	2,21	3.63	27
Villalaco	5	4,82	0,71	0,81	1,11	1,62	2.44	25
Villalazán	2	3,28	0,72	0,77	1,67	2,87	5.38	31
Villares	4	2,52	0,71	0,79	1,21	1,85	2.90	26
Villoria	2	4,93	0,73	0,78	1,61	2,87	4.86	30
Zuzones	1	3,52	0,73	0,76	2,35	4,71	10.36	41

Cuadro 2.-Estimaciones obtenidas para el cultivo del remolacha en regadío y error típico del estimador, para zonas no muestreadas.

Zona de riego	Estimación ha/seg.	Error Típico
Aguilar	1.26	0.74
Aldearregada	3.19	0.73
Campo de Ledesma	1.26	0.74
Castronuño	1.47	0.74
Cervera Arbejal	1.26	0.74
Ejeme Galizancho	2.08	0.73
Guma	7.21	0.80
Olmillos	8.16	0.83
Ruesga	1.26	0.74
Villamayor	3.66	0.74
Zorita	1.76	0.74