Ordenamiento del Territorio y Medio Marino. V Reunión Científica de la Asociación Española de Teledetección.
Universidad de Las Palmas de Gran Canaria - Servicio de Publicaciones.
Las Palmas de Gran Canaria, 10, 11 y 12 de noviembre de 1993. pp. 601-608.
Editor: Manuel Cantón Garbín.

Estimación de superficies cultivadas en La Comunidad Autónoma de Cataluña mediante estimadores de regresión múltiple

F. González-Alonso[1], *R. Arbiol*[2], *J.M. Cuevas*[1], *J. Romeu*[1]
[1] Laboratorio de Teledetección
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria
[2] Servicio de Teledetección
Instituto Cartográfico de Cataluña

RESUMEN

Se ha realizado la estimación de superficies cultivadas en la provincia de Lérida, utilizando un muestreo de marco de área e imágenes Landsat TM del año 1991. Junto al tradicional estimador de regresión, se ha probado con un estimador de regresión múltiple, con el que se han obtenido mejores resultados de estimación, especialmente en el caso de los cultivos poco representados.

ABSTRACT

The estimation of crop areas in the province of Lérida, using an area sample frame and spectral Landsat TM information are presented. Toghether with the usual regression estimator, a multiple regression estimator has been tested. The obtained results have been better with the last, specially in the case of crops with small presence.

Introducción

La estimación objetiva de superficies cultivadas mediante la integración de muestras de campo, basadas en un marco de área, con los resultados de imágenes de satélite clasificadas, ha demostrado que es una metodología operacional en el contexto europeo (Delince, 1990).

La técnica más usual para integrar los resultados de una muestra de campo con imágenes de satélite clasificadas es el estimador de regresión simple (Cochram, 1977; Ozga *et al.*, 1977; González-Alonso *et al.*, 1991; González-Alonso *et al.*, 1993).

Para realizar la estimación de superficies cultivadas por el método de regresión simple es necesario ajustar el modelo de regresión:

$$Y_i = b_0 + b_1 X_i \quad i = 1, n$$

Siendo:

n = Tamaño de la muestra o número de segmentos visitados y encuestados en el campo.
Y_i = Porcentaje, determinado por digitalización, que ocupa el cultivo de interés en el segmento i.
X_i = Porcentaje que ocupa el cultivo de interés en el segmento i, calculado a par- tir de la clasificación de los píxeles que componen dicho segmento.
b_0, b_i = Coeficientes de regresión simple.

Una medida de la calidad del ajuste lineal de regresión existente entre Y_i y X_i viene dada por el coeficiente de determinación simple:

$$r^2 = \frac{\left\{ \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \right\}^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Siendo $\bar{Y}$ y $\bar{X}$ las medias muestrales de Y_i y X_i respectivamente.

En el método de expansión directa (utilizando sólo datos de campo) la superficie total ocupada por el cultivo de interés será: $T = D\ \bar{Y}$, siendo D la superficie en hectáreas del área en estudio. La varianza de esta estimación será: $V(T) = D^2\ V(\bar{Y})$.

En el método de regresión simple, la superficie total ocupada por el cultivo de interés, será:

$$T_{reg} = D\ \bar{Y}_{reg}, \text{ siendo } \bar{Y}_{reg} = \bar{Y} + b_i (\bar{X}_{pob} - \bar{X})$$

Donde b_i es el coeficiente de regresión simple y $\bar{X}_{pob}$ la proporción poblacional obtenida para el cultivo de interés a partir de la clasificación de todos los píxeles de la imagen de satélite comprendidos dentro del área en estudio. En estas condiciones, $V(T_{reg}) = V(T)(1 - r^2)$.

El interés del estimador de regresión simple consiste en corregir la estimación de la media poblacional de la variable Y, conocida a través de una muestra de tamaño n, mediante el empleo de una variable auxiliar X (deducida de las imágenes de satélite) que es conocida para el conjunto de los N elementos que componen la población (la imagen entera) y que está correlacionada linealmente con la variable de interés Y (superficie ocupada por el cultivo en cuestión).

Hay que recordar que la variable auxiliar X puede ser "cualquier variable", aunque su uso estará tanto más justificado cuanto mayor sea la correlación lineal existente entre las variables Y y X.

El ratio ER = V (T) / V (T_{reg}) se denomina eficiencia relativa. En el caso de la regresión simple, ER = 1 / (1 – r^2).

Cuanto mayor sea el valor de ER, mayor será el interés de emplear el estimador de regresión simple y más justificado estará el empleo de las imágenes de satélite en la estimación objetiva de las superficies cultivadas.

Material y métodos

Metodología utilizada

Un método alternativo al anteriormente expuesto que puede mejorar la estimación de las superficies cultivadas, cuando dicha estimación no se considera que posee la precisión suficiente, consiste en el empleo de estimadores de regresión múltiple (Cárdenas *et al.*, 1978).

De forma análoga al caso de regresión simple:

$$\bar{Y}_1 \text{ (reg. mul.)} = \bar{Y}_1 + b_1 (\bar{x}_1 - \bar{X}_1) + b_2 (\bar{X}_2 - \bar{X}_2) + b_3 (\bar{X}_3 - \bar{X}_3) + \ldots$$

Siendo:

Y_1 = (reg. mul.) = Media de regresión múltiple del cultivo de interés.
Y_1 = Media muestral del cultivo de interés obtenida mediante la digitalización de la encuesta de campo.
X_i = Medias poblacionales de los diferentes cultivos obtenidas mediante clasificación de la imagen de satélite.
X_i = Medias muestrales de los diferentes cultivos obtenidas mediante clasificación de la imagen de satélite.
b_i = Coeficiente de regresión múltiple.

La estimación de la varianza será:

$$V (\bar{Y}_1 \text{ (reg. mul.)}) = V (\bar{Y}_1) (1 - R^2)$$

Siendo R el coeficiente de correlación múltiple.

Un problema que puede surgir en la utilización del estimador de regresión múltiple es que exista multicolinearidad entre las diferentes variables independientes, en

este caso entre las superficies que ocupan los diferentes cultivos considerados al clasificar la imagen de satélite.

La detección de este problema se realiza analizando el último autovalor de la matriz de correlaciones entre los diferentes cultivos. Si el último autovalor es muy pequeño (inferior a 0,4 para cuatro variables o a 0,20 para una veintena) se puede afirmar que existe un problema de colinearidad. Una forma de evitar este problema puede consistir en realizar un procedimiento de selección de variables, de manera que al final de este proceso dispongamos de un modelo de regresión en el que sólo estén incluidas variables poco correlacionadas entre sí.

El área en estudio, inventario realizado e imágenes utilizadas

El área en estudio comprende los estratos 1 (herbáceo no irrigado) y 2 (herbáceo irrigado) de la provincia de Lérida correspondientes a la campaña agrícola de 1991. Dichos estratos suponen una superficie de 340.541 ha y el número de segmentos cuadrados de 49 ha investigados en los mismos ha sido 103.

Los trabajos de campo se realizaron por parte de personal contratado por el Instituto Cartográfico de Cataluña (ICC) durante los meses de abril, mayo y junio de 1991.

Se han utilizado imágenes Landsat-5 TM captadas el 29 de mayo y el 8 de junio de 1991. El proceso de estas imágenes se ha realizado con el *software* específico desarrollado por el ICC.

En la Tabla 1 se presentan las proporciones poblacionales, expresadas en tantos por uno, que se han obtenido mediante la clasificación multitemporal de las imágenes indicadas para el área en estudio.

Maíz	0,0492
Cebada	0,3950
Legumbres	0,0033
Alfalfa	0,0401
Otros forrajes	0,0203
Barbecho	0,0203
Forestal	0,1567
Urbano	0,0292

Tabla 1
Proporciones poblacionales por clasificación de los diferentes cultivos

Para analizar la influencia del empleo del estimador de regresión múltiple en comparación con el estimador de regresión simple y el estimador de expansión directa, se han seleccionado un cultivo muy representado en el área en estudio, como es la cebada (39,5% de la imagen clasificada), y otro muy poco representado, como son las legumbres (0,33% de la imagen clasificada).

Resultados y discusión

En el caso de la cebada, después de realizar un proceso de selección de variables por el método de todas la regresiones posibles, el modelo de regresión múltiple seleccionado ha sido:

$$DIGCEB = b_0 + b_1\ CEB + b_2\ FOR$$

Siendo:

DIGCEB = Tanto por uno de cebada determinado en cada segmento de campo mediante digitalización.
CEB = Tanto por uno de cebada determinado en cada segmento mediante la clasificación de las imágenes de satélite.
FOR = Tanto por uno de área forestal determinado en cada segmento mediante la clasificación de las imágenes de satélite.

El cuadrado del coeficiente de correlación múltiple con el modelo considerado ha sido 0,6854.

En la Tabla 2 se presentan los resultados de estimación obtenidos en el caso de la cebada empleando los tres métodos comparados. Se comprueba que la eficiencia relativa del método de regresión múltiple es superior a la eficiencia relativa obtenida mediante el método de regresión simple, tomando como referencia la estimación por expansión directa.

	Superficie	**Error stand.**	**Coef. var.**	**E.R.**
Expan. direc.	112.764,7	11.206,2	9,93%	
Regr. multip.	107.214,9	6.285,4	5,86%	3,17
Regr. simple	109.992,6	6.473,4	5,88%	2,99

Tabla 2
Resultados comparativos de los tres métodos de estimación de superficies en el caso de la cebada

Si calculamos la eficiencia relativa entre ambos métodos de regresión, tendremos:

$$ER\ (Cebada) = \hat{V}(\hat{T}_{reg.}) / \hat{V}(\hat{T}_{reg.\ mul.}) = 1{,}06$$

Es decir, el efecto de utilizar el estimador de regresión múltiple en lugar del estimador de regresión simple equivale a que se hubiera incrementado en un 6% el tamaño de la muestra, pero sin haber tenido que realizar ningún gasto adicional.

En el caso de las legumbres, después de realizar el proceso de selección de variables, el modelo retenido fue:

$$DIGLEG = b_0 + b_1\ LEG + b_2\ MAI + b_3\ CEB + b_4\ OFO + b_5\ FOR$$

Siendo:

DIGLEG = Tanto por uno de legumbres determinado en cada segmento mediante digitalización.
LEG = Tanto por uno de legumbres determinado en cada segmento mediante clasificación de las imágenes de satélite.
MAI = Tanto por uno de maíz determinado en cada segmento mediante clasificación de las imágenes de satélite.
CEB = Tanto por uno de cebada determinado en cada segmento mediante clasificación de las imágenes de satélite.
OFO = Tanto por uno de otros forrajes determinado en cada segmento mediante clasificación de las imágenes de satélite.
FOR = Tanto por uno de área forestal determinado en cada segmento mediante clasificación de las imágenes de satélite.

El cuadrado del coeficiente de correlación múltiple con el modelo considerado ha sido 0,31416.

En la Tabla 3 se presentan los resultados de las estimaciones obtenidas en el caso de las legumbres mediante los tres métodos de estimación empleados.

	Superficie	Error stand.	Coef. var.	E.R.
Expan. direc.	2.768,7	1.250,6	45,17%	
Regr. multip.	1.125,5	1.014,8	90,16%	1,51
Regr. simple	2.213,1	1.103,7	49,87%	1,28

Tabla 3
Resultados comparativos de los tres métodos de estimación de superficies en el caso de las legumbres

De la observación de dicha tabla se deduce que la eficiencia relativa del método de regresión múltiple es ligeramente superior a la obtenida en el caso del estimador de regresión simple.

Si calculamos la eficiencia relativa entre ambos métodos de regresión se obtiene:

$$ER\ (Legumbres) = \hat{V}(\hat{T}_{reg.}) / \hat{V}(\hat{T}_{reg.\ mul.}) = 1{,}1828$$

Así pues, el efecto de utilizar el estimador de regresión múltiple, en lugar del estimador de regresión simple, en el caso de las legumbres equivale a que se hubiera incrementado el tamaño de la muestra en un 18%. La reducción de la varianza obtenida para la estimación de la superficie ocupada por las legumbres se debe exclusivamente al empleo del estimador de regresión múltiple sin gasto adicional alguno.

Conclusiones

Del desarrollo del presente trabajo se pueden obtener las siguientes conclusiones:

- La estimación de las superficies cultivadas a partir de encuestas de campo e imágenes de satélite empleando el estimador de regresión múltiple ha proporcionado, en la provincia de Lérida en el año 1991, unas estimaciones con una varianza menor que la obtenida mediante el empleo del estimador por regresión simple. Esto se ha producido tanto en el caso de cultivos muy representados (cebada) como en el caso de cultivos poco representados (legumbres).
- El empleo del estimador de regresión múltiple ha resultado especialmente interesante en el caso de cultivos poco representados como las legumbres, en comparación con los resultados obtenidos en el caso de cultivos muy representados, como la cebada.
- La obtención de estas estimaciones más precisas se ha conseguido sin que sean necesarias inversiones adicionales a las imprescindibles para poder emplear el estimador de regresión simple (encuestas de campo e imágenes de satélite clasificadas).

– Debido a la existencia del problema de colinearidad se recomienda la realización de un procedimiento de selección de variables regresoras con objeto de salvar esta circunstancia y no comprometer la robustez de las estimaciones.

Bibliografía

Cárdenas, M. and Hanuschak, G.A.: 1978. *Multiple regression estimations using classified Landsat data.* Report NASS-USDA. Whasington.

Delince, J.: 1990. Un premier bilan de l'action 1 "Inventaires Régionaux" du project agriculture aprés deux annés d'activité. *Proceedings of the Application of Remote Sensing to Agricultural Statistics.* Ispra. pp. 53-58. Commission of the European Communities. Joint Research Centre. Ispra.

González-Alonso, F. and Cuevas, J.M.: 1993. Remote sensing and agricultural statistics: crop area estimation through regression estimators and confusion matrices. *International Journal of Remote Sensing.* 14:1215-1219.

González-Alonso, F., López, S. and Cuevas J.M.: 1991. Comparing two methodologies for crop area estimation in Spain using Landsat TM images and ground gathered data. *Remote Sensing of Environment.* 35:29-35.

Ozga, M., Donovan, W. and Gleason, C.: 1977. An interactive system for agricultural acreage estimates using Landsat data. *Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data.* West Lafayette, Indiana. pp. 113-123.