

Geographically Weighted Regression for Small Area Estimation: An Agricultural Case Study

*La Regressione Geografica Pesata nella Stima per Piccole Aree:
Un'Applicazione in Ambito Agricolo*

Chiara Bocci, Alessandra Petrucci, Emilia Rocco

Dipartimento di Statistica "G. Parenti,
Università degli Studi di Firenze, viale Morgagni, 59 – 50134 Firenze
bocci@ds.unifi.it, alex@ds.unifi.it, rocco@ds.unifi.it

Riassunto: Il lavoro intende valutare se l'informazione sulla collocazione spaziale delle aziende agricole possa essere utilizzata in un modello di regressione per ottenere stime per domini sub-regionali. Questa necessità scaturisce dall'impossibilità di ottenere stime attendibili a livello sub-regionale con gli attuali disegni di campionamento utilizzati in ambito agricolo e dalla possibilità di collocare le aziende agricole sul territorio introdotta dal Censimento dell'Agricoltura del 2000 effettuato dal ISTAT. A tal fine nel lavoro, tenendo conto della presumibile non stazionarietà della maggior parte dei fenomeni agricoli, viene esaminata l'applicazione della regressione geografica pesata (Geographically Weighted Regression) per stimare a livello comunale la superficie coltivata a vite nella provincia di Firenze.

Keywords: Small area estimation, spatial data, stationary process, geographically weighted regression

1. Introduction

The Fifth Agricultural Census driven in year 2000 has given the opportunity to locate the farms on the territory introducing a new challenge for statistical analysis. The geographical location of each farm can constitute a particularly useful information for the analysis of many phenomena concerning the agricultural field (Lipizzi, 2004; Bocci *et al.*, 2005).

In particular, in this paper we underline the usefulness of the farms geographical location for the analysis of non stationary spatial phenomena.

"Global" dependence models, such as the classical regression model, assume the independence of the phenomenon from the data spatial location. For many agricultural phenomena the application of one of these models could bring to wrong conclusions and generate spatially autocorrelated residuals. Hence the need to apply statistical techniques able to take into account the spatial variability of the phenomenon.

Obviously the application of these techniques is subject to the availability of the knowledge of the spatial location of the population units under study.

The technique we considered in this paper is the Geographically Weighted Regression (GWR), a specific model which allows to represent non stationary local phenomena (Fotheringham *et al.*, 2002).

We applied the GWR methodology in order to estimate at municipal level the area allocated to grapevine cultivation using data from the current Farm Structure Survey.

The paper is organized as follows: section 2 summarises the GWR methodology. Section 3 discusses the results of its application to the estimation of the area allocated to grapevine and reports some final remarks.

2. Methodology

GWR extends the traditional regression framework by allowing local rather than global parameters to be estimated, so that the classical regression model is rewritten as

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^p \beta_k(u_i, v_i) x_{ik} + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

where (u_i, v_i) denotes the coordinates of the i -th point in space, $\beta_k(u_i, v_i)$ is a realisation of the continuous function $\beta_k(u, v)$ at point i , $x_{i1}, x_{i2}, \dots, x_{ip}$ are the explanatory variables at point i and ε_i are error terms.

For a given data set, local parameters $\beta_k(u_i, v_i)$ are estimated using the weighted least square procedure. The weights w_{ij} , for $j = 1, \dots, n$, at each location (u_i, v_i) are obtained as a continuous function of the distance between the point i and the other data points.

Let

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(u_1, v_1) & \beta_1(u_1, v_1) & \cdots & \beta_p(u_1, v_1) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0(u_n, v_n) & \beta_1(u_n, v_n) & \cdots & \beta_p(u_n, v_n) \end{bmatrix} \quad (2)$$

be the matrix of the local parameters. Each row is estimated by

$$\hat{\boldsymbol{\beta}}(i) = (\mathbf{X}^T \mathbf{W}(i) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(i) \mathbf{y}, \quad (3)$$

where $i = 1, \dots, p$ represents the row of the matrix (2), $\mathbf{X}$ is the matrix of explanatory variables, $\mathbf{y}$ is the dependent variable and $\mathbf{W}(i)$ is an n by n spatial weighting matrix of the form

$$\mathbf{W}(i) = \begin{bmatrix} w_{i1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & w_{in} \end{bmatrix}. \quad (4)$$

The estimator in (3) is a least square estimator but the weighting matrix is not constant, hence $\mathbf{W}(i)$ has to be computed for each point i and the w_{ij} indicate the proximity of each data point to the location of i : data points closer to i carrying more weight in the estimation of the $\boldsymbol{\beta}(i)$ parameters than those farther away.

We can note, however, that equation (3) can be estimated at any point, even at location

where no data have been observed.

The choice of the weighting scheme $\mathbf{W}(i)$ is a relevant step in GWR procedure. Several different weighting functions can be defined, the more common kernels being the Gaussian and the bi-square weighting functions. For a comprehensive overview about kernel functions in GWR analysis see Fotheringham *et al.* (2002).

A modified bi-square function taking into account only the N^{th} nearest neighbours is

$$w_{ij} = \begin{cases} \left[1 - (d_{ij}/b)^2\right]^2 & \text{if } j \text{ is one of the } N^{th} \text{ nearest neighbours of } i \text{ and} \\ & b \text{ is the distance to the } N^{th} \text{ nearest neighbour} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

This kernel function varies in space and presents an adaptive bandwidth depending on the data points density. Consequently the calibration of the model involves also the choice of N , the number of data point to be included in the estimation of local parameters.

The appropriate bandwidth, or the appropriate value of N , can be obtained by a least square approach using the *cross-validation* criteria

$$CV = \sum_{i=1}^n [y_i - \hat{y}_{\neq i}(b)]^2, \quad (6)$$

where $\hat{y}_{\neq i}(b)$ is the fitted value of y_i with the observations for point i omitted from the calibration process.

3. Application and results

As stated above, the objective of this work is to investigate how the information on the spatial location of the farms gathered by the Fifth Italian Agricultural Census can be used in a GWR model in order to predict, at municipal level, average (or total) agricultural surface area or production for a specific cultivation.

In this paper we restrict our attention to the estimation of agricultural surface area since for the productions no real data or other estimates, sufficiently stable to represent a term of comparison for evaluating the efficacy of the proposed method, are available. Furthermore, we focus on the 44 municipalities of the Florence province and we refer only to the estimation of the agricultural area allocated to grapevines.

The implemented procedure considers a simple random sample of 300 farms drawn from the census list. The sample size has been fixed roughly equal to the size of the sample used in the Farm Structure Survey (FFS, 2003).

The available spatial information consists in the geographical coordinates of each farm's administrative centre. The Euclidean distance between them is used as the distance between the corresponding data points.

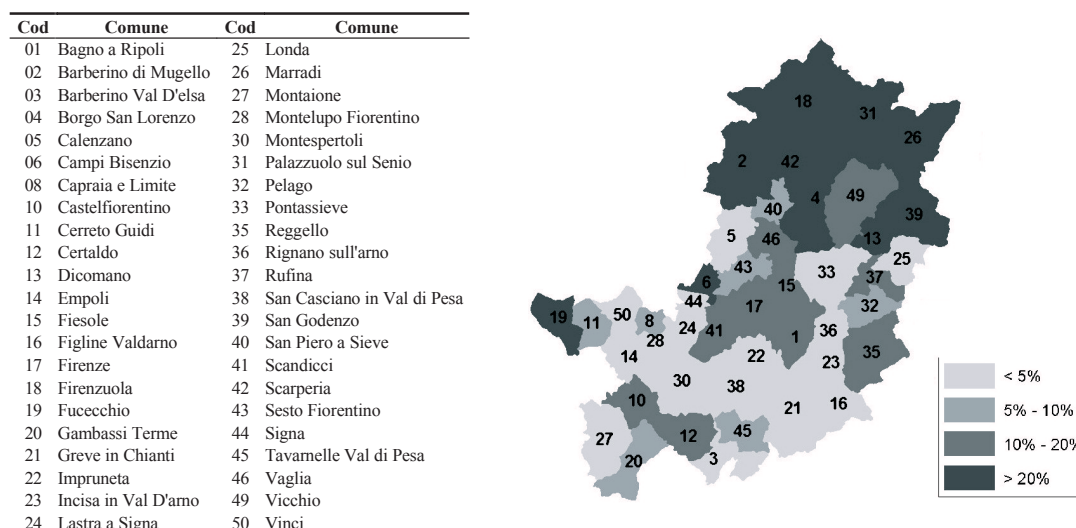
With regard to the geographically weighted regression model we adopt the bi-square weighting function with adaptive bandwidth, shown in (5). The number of nearest neighbour data points to be included within the calibration of the local model is determined from the cross-validation criteria, shown in (6).

Among the available explicative variables collected by the Agricultural Census, only the utilised agricultural surface (SAU) and the European size unit (UDE) are statistically significant and are considered in the model. In order to verify the spatial non-stationarity of the data, a traditional regression model is applied. The estimated parameters of this model are significant but the distribution of the residuals from the model appears non-random. Supported by this result, for each farm of the province of Florence a local regression model is derived and the total area allocated to grapevines is estimated. Finally, the total municipal area allocated to grapevines is calculated by the sum over all the farms belonging to the same municipality.

The obtained results are encouraging: the error is less than 10% for more than the 50% of them, and less than 20% for the 77% of the municipalities (figure 1).

Although these results, based on a single “pseudo-application”, do not allow any definitive conclusions about the proposed method, we believe that they suggest the opportunity of a thorough examination.

Figure 1: Error of the estimates of the total area allocated to grapevines for the 44 municipalities of the province of Florence (with ISTAT codes).



References

- Bocci C. Petrucci A., Rocco E. (2005), L'uso di informazioni di tipo spaziale in ambito agricolo: un'applicazione relativa al comune di Fiesole, *Atti del convegno "Metodi d'Indagine e di Analisi per le Politiche Agricole"*, Università di Pisa, 21-22 ottobre 2004.
- Fotheringham A.S., Brunsdon C., Charlton M. (2002), *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, John Wiley & Sons, Chichester, UK.
- Lipizzi, F. (2004), L'integrazione dei disegni territoriali del Censimento della popolazione e degli edifici e del Censimento dell'agricoltura, *Atti del Convegno "L'Informazione Statistica e le Politiche Agricole (ISPA2004)"*, Università di Cassino, 6 maggio 2004.