

# M-quantile regression with spatial information in Small Area Estimation

*La Regressione M-quantilica con informazione spaziale  
nella Stima per Piccole Aree*

Ray Chambers

Division of Social Statistics, School of Social Sciences, University of Southampton,  
SO17 1BJ, Southampton UK, [rc6@soton.ac.uk](mailto:rc6@soton.ac.uk)

Monica Pratesi, Nicola Salvati

Dipartimento di Statistica e Matematica Applicata all'Economia,  
Università di Pisa, via Ridolfi, 10 – 56124 Pisa,  
[m.pratesi@ec.unipi.it](mailto:m.pratesi@ec.unipi.it) [salvati@ec.unipi.it](mailto:salvati@ec.unipi.it)

N. Tzavidis

Institute of Education, University of London, 20  
Bedford Way London WC1H 0AL, UK, [n.tzavidis@ioe.ac.uk](mailto:n.tzavidis@ioe.ac.uk)

**Riassunto:** In questo lavoro si inserisce l'informazione spaziale nell'approccio basato sulla regressione M-quantilica per la stima per piccole aree, proponendo un'alternativa semiparametrica allo stimatore di tipo SEBLUP (Spatial Empirical Best Linear Unbiased Predictor). Il metodo è applicato alla stima della produzione media e mediana di olive per azienda in Toscana con dati provenienti dall'indagine campionaria sulla Struttura e Produzione delle aziende agricole (2003).

**Keywords:** Small area estimation, Spatial EBLUP, M-quantile regression.

## 1. Introduction

A new approach to small area estimation based on quantile-like parameters of the conditional distribution of the study variable given the covariates has been recently proposed by Chambers and Tzavidis (2006a). This technique does not depend on strong distributional assumption like the small area models that use both covariates and random effects, and it is robust against outlying area values. Moreover the approach allows for the estimation of the distribution function in each small area and overcomes an important problem in small area estimation that is the impact on the estimates of changing small area geographies.

The aim of this work is to include in the M-quantile approach the information about the spatial distribution of the study variable in the small areas. What we define here is essentially a semiparametric alternative to the Spatial EBLUP developed in Pratesi and Salvati (2005a); Petrucci and Salvati (2005); Saei and Chambers (2003). The proposed methodology is applied to the Farm Structure Survey (2003) data on the production of olives in the 42 Local Economic Systems (LESS) of Tuscany. The paper is organized as follows: section 2 recalls the main steps of the proposed procedure. Section 3 shows some results of the application to the estimation of the mean and median production of olives per farm at LES level and reports some final remarks.

## 2. M-quantile Models with spatial information

The first target in M-quantile models for small estimation is to estimate the M-quantile coefficients,  $\{q_j; j \in s\}$ , of the units in the sample without reference to the small areas of interest. To do this we define a fine grid on the (0,1) interval and, using the sample data, we fit the M-quantile regression line at each value  $q$  on this grid. Let us define by  $\Psi$  an influence function and by  $\mathbf{x}$  the matrix of covariates  $n \times p$ , with  $n = \sum_{i=1}^m n_i$  and  $m$  the number of small areas. Estimation of the regression parameters of the linear M-quantile regression model

$$Q_q(\mathbf{x}, \Psi) = \mathbf{x}^T \boldsymbol{\beta}_\Psi(q) \quad (1)$$

is achieved by solving the estimating equations below using iteratively reweighted least squares (IRLS)

$$\sum_{j=1}^n \Psi_q(r_{jq\Psi}) \mathbf{x}_j = 0 \quad (2)$$

where  $r_{jq\Psi} = y_j - \mathbf{x}_j^T \boldsymbol{\beta}_\Psi(q)$ ,  $\Psi_q(r_{jq\Psi}) = 2\Psi(s_r^{-1} r_{jq\Psi}) \{qI(r_{jq\Psi} > 0) + (1-q)I(r_{jq\Psi} \leq 0)\}$  and  $s_r = MAD / 0.675$  is a robust measure of spread that is employed in preference to the standard deviation of the residuals and the influence function is the Huber Proposal 2 (Chambers and Tzavidis, 2006a). If a data point lies exactly on a fitted M-quantile regression line, then the M-quantile coefficient of the corresponding sample unit is equal to  $q$ . Otherwise, if a data point lies between two fitted M-quantile regression lines, the M-quantile coefficient of the corresponding sample unit is derived by linear interpolation. If a hierarchical structure does explain part of the variability in the population data, then we expect units within clusters defined by this hierarchy to have similar M-quantile coefficients. Therefore, we can compute the average value of the sample M-quantile coefficients of units  $j$  in area  $i$ ,  $\hat{q}_i = \sum_{j \in i} q_{js}$ . A small area estimator of the mean,  $\hat{y}_i$ , is then derived by fitting an M-quantile regression model at  $\hat{q}_i$

$$\hat{y}_i = \int t d\hat{F}_{CD,i}(t) = \frac{1}{N_i} \left( \sum_{j \in s_i} y_{ij} + \sum_{j \in r_i} \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_\Psi(\hat{q}_i) + \frac{N_i - n_i}{n_i} \sum_{j \in s_i} (y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_\Psi(\hat{q}_i)) \right) \quad (3)$$

where  $\hat{F}_{CD,i}(t)$  is the estimated cumulative distribution function for each small area (Chambers and Tzavidis, 2006b) and  $s_i$  and  $r_i$  respectively denote the sampled and non sampled units in area  $i$  and  $N_i$  is the number of population units in area  $i$ .

A model that includes spatial relationship could have significant advantages compared to the linear M-quantile regression model employed in Chambers and Tzavidis (2006), when the variable of interest is spatially correlated (Pratesi and Salvati, 2005a). To account for spatial correlation in the M-quantile approach, we propose modelling the M-quantile coefficients, derived from the first step of the Chambers and Tzavidis (2006a) approach using a model of the following form

$$\log \left( \frac{\mathbf{q}_s}{I - \mathbf{q}_s} \right) = \mathbf{x}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{I} - \rho\mathbf{W})^{-1} \mathbf{u} + \boldsymbol{\varepsilon} \quad (4)$$

where  $\mathbf{x}$  be the matrix of covariates  $n \times p$ ,  $\mathbf{Z}$  is the incidence matrix  $n \times m$  for the random effects vector, the deviations from the fixed part of the model  $\mathbf{x}\beta$  are the result of an autoregressive process with parameter  $\rho$  (spatial autoregressive coefficient) and proximity matrix  $\mathbf{W}$   $m \times m$  and  $\varepsilon$  the error vectors.  $\mathbf{I}$  is an identity matrix  $m \times m$ . Under the model, the Spatial EBLUP estimator of  $\hat{q}_i$  is:

$$\hat{q}_i(\hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\rho}) = \frac{\exp(p_i)}{1 + \exp(p_i)} \quad (5)$$

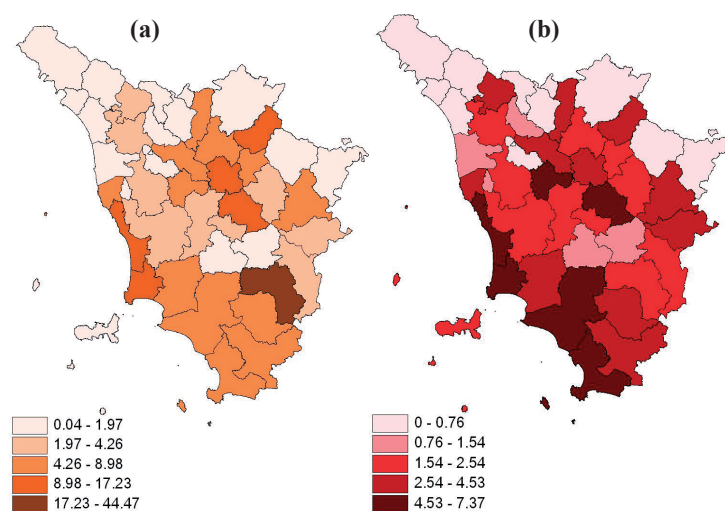
where  $p_i = \bar{\mathbf{x}}_i \hat{\beta} + \mathbf{b}_i^T \hat{\sigma}_e^2 [(\mathbf{I} - \hat{\rho} \mathbf{W})(\mathbf{I} - \hat{\rho} \mathbf{W}^T)]^{-1} \mathbf{Z}^T \times [\hat{\sigma}_e^2 \mathbf{I}_n + \mathbf{Z} \hat{\sigma}_u^2 [(\mathbf{I} - \hat{\rho} \mathbf{W})(\mathbf{I} - \hat{\rho} \mathbf{W}^T)]^{-1} \mathbf{Z}^T]^{-1} (\mathbf{q}_i - \mathbf{x}_i \hat{\beta})$ ,  $\bar{\mathbf{x}}_i$  are the known population mean,  $\mathbf{q}_i$  is the  $n \times 1$  vector of  $q_{js}$ ,  $\hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\rho}$  are asymptotically consistent estimators of the parameters and  $\mathbf{b}_i^T$  is a  $1 \times m$  vector  $(0, 0, \dots, 0, 1, \dots, 0)$  with value 1 in the  $i$ -th position. Similarly to (3), an M-quantile small area estimator of the mean, that accounts for spatial correlation, is then defined by fitting an M-quantile model at  $\hat{q}_i(\hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\rho})$ . The drawback of the previous approach is that although we use the M-quantile approach in order to avoid specifying a parametric model, we still use parametric model (4) to account for spatial correlation in the M-quantile coefficients. Ideally, we would like to employ a non-parametric approach to account for spatial correlation in the M-quantile coefficients. However, this is beyond the scope of this paper.

### 3 Results and final remarks

From the cumulative distribution function  $\hat{F}_{CD,j}(t)$  the mean, median and other moments can be obtained for each small area. Their Mean Squared Errors (MSE) have been approximated by Chambers and Tzavidis (2006a) and empirically evaluated by Pratesi and Salvati (2005b). For details on MSE and a simulation study where the performance of the proposed estimator is evaluated as a function of the parameter  $\rho$  and its performance is compared with that Spatial EBLUP see Chambers *et al.* (2005). In the economy of this work we present only a sample of the results obtained for the estimated spatial distribution of the mean and median production of olives per farm per LES (Figure 1). The data are from the Farm Structure Survey (2003) (see Ballin and Salvi (2004) for details on sampling design). The proposed method is implemented on this data using the expressions (4) and (5). The major determinant of the average production of olives per farm at area level is utilized surface area for olive production (ha) ( $\mathbf{x}$  variable in eq. 4),  $\mathbf{Z}$  is the incidence matrix of dimension 2508 farms per 42 LESs. The neighbourhood structure  $\mathbf{W}$  is defined as follows: spatial weight,  $w_{ij}$ , is 1 if area  $i$  shares an edge with area  $j$  and 0 otherwise. For an easier interpretation, the general spatial weight matrix is defined in row standardized form, in which the row elements sum to one. The value of the estimated spatial autoregressive coefficient  $\hat{\rho}$  is 0.441 (*s.e.* = 0.183) with REML (Restricted Maximum Likelihood) procedure and suggests a spatial relationship of a medium strength. The maps in Figure 1 show two different aspects of the spatial distribution of the production per farm across the small areas (LESs): the mean production per farm (Figure 1a) highlights the LESs with relevant levels of production per farm (Val d' Orcia and the Chianti). The median (Figure 1b) is insensitive to the presence of few big farms that raise the medium level of production

and, as a consequence, the spatial distribution of the median production seems to be more homogenous. This fact stresses the importance of maps that represent not only the spatial distribution of mean but also of other quantiles, emphasizing the importance of the estimation of the cumulative distribution function in each small area area. This makes it possible to know the concentration of the agricultural production at local level, giving a crucial information both for Government intervention on agricultural policies and for guidance activities for the final users of the data, the research institutes, the citizens.

**Figure 1.** (a) *Mean and (b) median production of olives per farm per LES*



## References

- Ballin, M., Salvi, S. (2004) *Nota metodologica sul piano di campionamento adottato per l'indagine "Struttura e produzione delle aziende agricole 2003"*, Istat.
- Chambers, R., Tzavidis, N. (2006a) M-quantile Models for Small Area Estimation, forthcoming in *Biometrika*.
- Chambers, R., Tzavidis, N. (2006b) Bias adjusted distribution estimation for small areas, forthcoming in *Statistics in Transition*.
- Chambers, R., Pratesi, M., Salvati, N., Tzavidis, N. (2005) Small Area Estimation: spatial information and M-quantile regression to estimate the average production of olive per farm, *Wp 279, Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa*.
- Pratesi, M., Salvati, N. (2005a) Small Area Estimation: the EBLUP estimator with autoregressive random area effects, *Working Paper 261, Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa*.
- Pratesi, M., Salvati, N. (2005b) Regressione M-quantilica nella stima per piccole aree. Il caso della produzione di olive in Toscana, relazione invitata ed pubblicata negli atti del Convegno "AGRI@STAT", 30-31 Maggio 2005, Firenze.
- Petrucchi A., Salvati N. (2005) Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment, forthcoming in *Journal of Agricultural, Biological, and Environmental Statistics*.
- Saei, A., Chambers, R. (2003), Small Area Estimation under linear and generalized linear mixed models with time and area effects, *Southampton Statistical Sciences Research Institute, WP M03/15*, Southampton.