

SPATIAL M-QUANTILE MODELS FOR SMALL AREA ESTIMATION

Nicola Salvati¹, Monica Pratesi¹, Nikos Tzavidis², Ray Chambers³

ABSTRACT

In small area estimation direct survey estimates that rely only on area-specific data can exhibit large sampling variability due to small sample sizes at the small area level. Efficient small area estimates can be constructed using explicit linking models that borrow information from related areas. The most popular class of models for this purpose are models that include random area effects. Estimation for these models typically assumes that the random area effects are uncorrelated. In many situations, however, it is reasonable to assume that the effects of neighbouring areas are correlated. Models that extend conventional random effects models to account for spatial correlation between the small areas have been recently proposed in literature. A new semi-parametric approach to small area estimation is based on the use of M-quantile models. Unlike traditional random effects models, M-quantile models do not depend on strong distributional assumptions and are robust to the presence of outliers. In its current form, however, the M-quantile approach to small area estimation does not allow for spatially correlated area effects. The aim of this paper is to extend the M-quantile approach to account for such spatial correlation between small areas.

Key words: Quantile regression, Robust models, Spatial correlation, Weighted least squares

1. Introduction

In small area estimation direct survey estimates that rely only on area-specific data can exhibit large sampling variability due to small sample sizes. In order to increase the efficiency of small area estimates, it is common practice to construct small area estimates using explicit linking models that borrow information from

¹ Dipartimento di Statistica e Matematica Applicata all'Economia, Università di Pisa, via Ridolfi, 10–56124 Pisa, salvati@ec.unipi.it, m.pratesi@ec.unipi.it.

² Centre for Census and Survey Research, University of Manchester, Manchester M13 9PL, UK, nikos.tzavidis@manchester.ac.uk.

³ Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522, Australia, ray@uow.edu.au.

related areas. The most popular class of these are models that include random area effects to account for between area variation beyond that explained by the auxiliary variables in the model. Estimation for these models is typically carried out assuming that the random area effects are uncorrelated (Rao 2003). In many situations, however, it is reasonable to assume that the effects of neighbouring areas – where neighbourhood is often defined in terms of a contiguity criterion – are correlated with the correlation decaying to zero as between area distance increases (Petrucci *et al.* 2005). In such cases the assumption of spatial independence of the random area effects becomes questionable. The problem of accounting for spatial correlation between the small areas has been recently tackled by extending the model of Battese *et al.* (1988) using a Simultaneously Autoregressive (SAR) process (Salvati 2004; Petrucci and Salvati 2006).

Chambers and Tzavidis (2006) have proposed a new approach to small area estimation that is based on the use of M-quantile models. Unlike traditional random effects models, M-quantile models do not depend on strong distributional assumptions and are robust to the presence of outliers. In its current form, however, the M-quantile approach to small area estimation does not allow for spatially correlated area effects. The aim of this paper is to extend the M-quantile approach to account for spatial correlation between small areas.

The paper is organized as follows. In section 2 we review random effects models that allow for spatially correlated random effects. In section 3 we propose an extension of the M-quantile approach to account for spatial correlation between the small areas. In section 4 we demonstrate usefulness of this framework through Monte Carlo simulation studies. The main focus of the comparisons is between the Spatial M-quantile and M-quantile models. In section 5 we present an application of spatial M-quantile models for estimating the average and median production of olives at the level of Local Economy System (LES) in Tuscany. Finally, in section 6 we summarise our main findings.

2. Small Area Models with Spatially Correlated Random Effects

Let $\mathbf{x}_i$ be a known vector of p auxiliary variables for each population unit j in small area i and assume that information for the variable of interest y is available only on the sample. The target is to use these data to estimate various area specific quantities. The most popular models used for this purpose are mixed effects models, i.e. models with random area effects. A linear mixed effects model has the following form:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + z_{ij} u_i + \varepsilon_{ij}, \quad j = 1K \ n_i, \quad i = 1K \ m \quad (0.1)$$

where u_i denotes a random area effect that characterizes differences in the conditional distribution of y given $\mathbf{x}$ between the m small areas, z_{ij} is a constant whose value is known for all units in the population and ε_{ij} is the error term

associated with the j -th unit within the i -th area. Conventionally, u_i and ε_{ij} are assumed to be independent and normally distributed with mean zero and variances σ_u^2 and σ_ε^2 respectively (Battese *et al.* 1988).

Model (0.1) can be extended to allow for correlated area effects. Let the deviations $\mathbf{v}$ from the fixed part of the model $\mathbf{x}^T \boldsymbol{\beta}$ be the result of an autoregressive process with parameter ρ and proximity matrix $\mathbf{W}$ (Cressie 1993; Anselin 1992), i.e.

$$\mathbf{v} = \rho \mathbf{W} \mathbf{v} + \mathbf{u} \Rightarrow \mathbf{v} = (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{u} \quad (0.2)$$

where $\mathbf{I}$ is $m \times m$ identity matrix. The matrix $(\mathbf{I} - \rho \mathbf{W})$ needs to be strictly positive definite to ensure the existence of $(\mathbf{I} - \rho \mathbf{W})^{-1}$. This happens if $\rho \in \left(\frac{1}{\min(\lambda_i)}, \frac{1}{\max(\lambda_i)} \right)$ where λ_i 's are the eigenvalues of matrix $\mathbf{W}$.

Combining (0.1) and (0.2), with $\boldsymbol{\varepsilon}$ independent of $\mathbf{v}$, the model with spatially correlated errors can be expressed as

$$\mathbf{y} = \mathbf{x}^T \boldsymbol{\beta} + \mathbf{Z}(\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{u} + \boldsymbol{\varepsilon}. \quad (0.3)$$

The error term $\mathbf{v}$ then has the $m \times m$ Simultaneously Autoregressive (SAR) dispersion matrix:

$$\mathbf{G} = \sigma_u^2 [(\mathbf{I} - \rho \mathbf{W}^T)(\mathbf{I} - \rho \mathbf{W})]^{-1}. \quad (0.4)$$

The $\mathbf{W}$ matrix describes the neighbourhood structure of the small areas whereas ρ defines the strength of the spatial relationship among the random effects associated with neighbouring areas. For ease of interpretation, the general spatial weight matrix is defined in row standardized form in which case ρ is referred to as the spatial autocorrelation parameter (Banerjee *et al.* 2004). Under (0.3), the Spatial Best Linear Unbiased Predictor (Spatial BLUP) estimator of the small area parameters and its empirical version (SEBLUP) are obtained following Henderson (1975). The SEBLUP estimator of the mean for small area i , $\bar{y}_i$, is

$$\begin{aligned} \hat{\bar{y}}_i = & \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}} + \mathbf{b}_i^T \hat{\sigma}_u^2 [(\mathbf{I} - \hat{\rho} \mathbf{W}^T)(\mathbf{I} - \hat{\rho} \mathbf{W})]^{-1} \mathbf{Z}^T \\ & \left\{ \hat{\sigma}_\varepsilon^2 \mathbf{I}_n + \mathbf{Z} \hat{\sigma}_u^2 [(\mathbf{I} - \hat{\rho} \mathbf{W}^T)(\mathbf{I} - \hat{\rho} \mathbf{W})]^{-1} \mathbf{Z}^T \right\}^{-1} (\mathbf{y}_s - \mathbf{x}^T \hat{\boldsymbol{\beta}}) \end{aligned} \quad (0.5)$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{x}^T \mathbf{V}^{-1} \mathbf{x})^{-1} \mathbf{x}^T \mathbf{V}^{-1} \mathbf{y}_s$, $\bar{\mathbf{x}}_i^T$ denotes a known area specific vector of population means for the auxiliary variables, $\mathbf{y}_s$ is a $n \times 1$ vector of the sampled

observations, $\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2, \hat{\rho}$ are asymptotically consistent estimators of the parameters obtained by Maximum Likelihood (ML) or Restricted ML (REML) method; $\mathbf{b}_i^T$ is a $1 \times m$ vector $(0, 0, \dots, 0, 1, \dots, 0)$ with value 1 in the i -th position.

The mean squared error (MSE) of SEBLUP and its estimator are obtained following the results of Kackar and Harville (1984), Prasad and Rao (1990) and Datta and Lahiri (2000). More specifically, the MSE estimator consists of three components called g_1 , g_2 and g_3 . These are due to the variability associated with the estimation of the random effects (g_1), the estimation of β (g_2) and the estimation of $\sigma_u^2, \sigma_\varepsilon^2, \rho$ (g_3). The corresponding equations are in Singh *et al.* (2005) and Pratesi and Salvati (2008). Note that due to the introduction of the additional parameter ρ , the component g_3 of the MSE is not the same as in the case of the traditional EBLUP estimator (Saei and Chambers 2003; Singh *et al.* 2005; Petrucci and Salvati 2006, Pratesi and Salvati 2008).

3. Spatial M-quantile Models for Small Area Estimation

Chambers and Tzavidis (2006) have proposed a new approach to small area estimation that is based on modelling the M-quantiles of the conditional distribution of the study variable (y) given the covariates (Breckling and Chambers, 1988). Unlike mixed effects models, which assume that the variability associated with the conditional distribution of y given $\mathbf{x}$ can be at least partially explained by a pre-specified hierarchical structure, such as the small areas of interest, M-quantile regression does not depend on a hierarchical structure. Instead, we characterise the conditional variability across the population of interest by the so-called M-quantile coefficients of the population units. The corresponding M-quantile coefficients, $\{q_j; j \in s\}$, of the units in the sample are then estimated using a grid-based interpolation procedure. In particular, a fine grid on the $(0,1)$ interval is first defined and, using the sample data, M-quantile regression lines are fitted at each value q on this grid using an iteratively reweighted least squares procedure (see Chambers and Tzavidis 2006 for details). If a data point lies exactly on a fitted M-quantile regression line, then the estimated M-quantile coefficient of the corresponding sample unit is set equal to q . Otherwise, if a data point lies between two fitted M-quantile regression lines, then the estimated 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, we expect units within clusters defined by this hierarchy to have similar M-quantile coefficients. Let $\bar{q}_i$ denote the average M-quantile coefficient for the population units in area i . An estimate of $\bar{q}_i$ is obtained by the corresponding

average value of the sample M-quantile coefficients of units j in area i , i.e. $\hat{q}_i = \sum_{j \in s_i} q_j$. An estimator of the corresponding small area mean, $\hat{y}_i$ is then

$$\hat{y}_i = \frac{1}{N_i} \left(\sum_{j \in s_i} y_j + \sum_{j \in r_i} \mathbf{x}_j^T \hat{\boldsymbol{\beta}}(\hat{q}_i) \right) \quad (0.6)$$

where $\hat{\boldsymbol{\beta}}(\hat{q}_i)$ denotes the slope coefficient of the fitted M-quantile regression line at $\hat{q}_i$, 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 . Note that (0.6) is equivalent to predicting the unobserved value y_j for population unit $j \in r_i$ using $\mathbf{x}_j^T \hat{\boldsymbol{\beta}}(\hat{q}_i)$.

In this paper we propose an extension to the above approach to account for spatial correlation between the small areas. In particular, assuming that the target population is made up of m small areas, we propose modelling the sample M-quantile coefficients using the model

$$\mathbf{logQ} = \mathbf{1}_s^T \boldsymbol{\beta} + \mathbf{Z}_s (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{u} + \boldsymbol{\varepsilon} \quad (0.7)$$

where $\mathbf{logQ}$ is the $n \times 1$ vector of the *logit* transformation of q_j , $\log[q_j / (1 - q_j)]$, $\mathbf{1}_s^T$ is vector of ones with dimension n and $\mathbf{Z}_s$ is the $n \times m$ incidence matrix for the random effects vector. In expression (0.7) we could have employed alternative link functions such as the *probit* link function. However, we expect that the choice of the link function will have little impact upon the small area estimates. Under model (0.7), the Spatial Empirical Best Predictor of $\bar{q}_i$ is

$$\hat{q}_i = \frac{\exp(\hat{p}_i)}{1 + \exp(\hat{p}_i)} \quad (0.8)$$

where

$$\hat{p}_i = \hat{\beta}_0 + \mathbf{b}_i^T \hat{\sigma}_u^2 \mathbf{D}_s^{-1} \mathbf{Z}_s^T \left\{ \hat{\sigma}_\varepsilon^2 \mathbf{I}_n + \mathbf{Z}_s \hat{\sigma}_u^2 \mathbf{D}_s^{-1} \mathbf{Z}_s^T \right\}^{-1} (\hat{\mathbf{q}}_s - \mathbf{1}_s \hat{\beta}_0) \quad (0.9)$$

the coefficient $\hat{\beta}_0$ is the estimated value of the intercept of the mixed model and $\mathbf{D}_s = [(\mathbf{I} - \hat{\rho} \mathbf{W}^T)(\mathbf{I} - \hat{\rho} \mathbf{W})]$. Here, $\hat{\mathbf{q}}_s$ is the $n \times 1$ vector of estimated M-quantile coefficients for the sample units q_j . An M-quantile estimator of the mean for area i that accounts for spatial correlation is then given by (0.6), but with $\hat{q}_i$ now given by (0.8). A drawback of this specification is that although we use

the M-quantile approach in order to avoid using a parametric model in small area estimation, we still use the parametric model (0.7) 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, developing a fully non-parametric approach is beyond the scope of this paper.

As Tzavidis and Chambers (2006) note, the M-quantile estimator (0.6) can be biased particularly when small areas contain outliers. These authors have therefore proposed a bias-adjusted M-quantile estimator of the mean that is based on representing this estimator as a functional of the small area distribution function. More specifically, it is straightforward to see that (0.6) is derived by appropriately integrating the empirical distribution function

$$\hat{F}_i(t) = \frac{1}{N_i} \left(\sum_{j \in s_i} I(y_{ij} \leq t) + \sum_{j \in r_i} I(\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\hat{q}_i) \leq t) \right). \quad (0.10)$$

Instead of using the empirical distribution function, the proposal of Tzavidis and Chambers (2007) is based on using the Chambers-Dunstan (1986) -hereafter denoted by a subscript CD- estimator of the small area distribution function

$$\hat{F}_{CD,i}(t) = \frac{1}{N_i} \left(\sum_{j \in s_i} I(y_{ij} \leq t) + \frac{1}{n_i} \sum_{j \in r_i} \sum_{k \in s_i} I \left\{ \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\hat{q}_i) + (y_{ik} - \mathbf{x}_{ik}^T \hat{\boldsymbol{\beta}}(\hat{q}_i)) \right\} \leq t \right\}. \quad (0.11)$$

The corresponding bias-adjusted mean estimator for small area i is then

$$\hat{y}_i = \int t d\hat{F}_{CD,i}(t) = \frac{1}{N_i} \left(\sum_{j \in s_{n_i}} y_{ij} + \sum_{j \in r_i} \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\hat{q}_i) + \frac{N_i - n_i}{n_i} \sum_{j \in s_{n_i}} (y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}(\hat{q}_i)) \right). \quad (0.12)$$

Estimates of other quantiles of the distribution of y in small area i can be obtained by appropriately integrating (0.11) (see also Tzavidis and Chambers 2007).

Tzavidis and Chambers (2006) and Chambers *et al.* (2007) proposed an estimator of the mean squared error of (0.12). The main limitation of this estimator is that it does not account for the variability introduced in estimating the area specific q 's. Empirical evaluations (Tzavidis *et al.* 2006), however, indicate that this mean squared error estimator provides a good approximation to the true mean squared error. In this paper, however, we focus our attention on the performance of M-quantile estimator (0.12), obtained with and without spatial information in the presence or not of outliers.

4. Simulation Experiments

Monte Carlo simulation experiments were designed for assessing the performance of the spatial M-quantile model described in the previous section. In particular the aim of these simulation exercises is to examine the usefulness of this framework for capturing the spatial structure of the data used for small area estimation and to investigate the performance of the small area methods in the presence of outliers. We illustrate the performance of the standard M-quantile estimator (0.6) and the CD form of the M-quantile estimator (0.12), with $\hat{q}_i$ determined by simple averaging over area i , and the corresponding Spatial M-quantile versions of (0.6) and (0.12) with $\hat{q}_i$ determined by (0.8). For reasons of completeness, we also considered other widely used methods for small area estimation such as the EBLUP estimator, and the Spatial EBLUP estimator (SEBLUP).

Synthetic population data are generated for the small areas using a spatial nested error regression model with random area effects distributed according to a SAR dispersion matrix with fixed spatial autoregressive coefficient given by

$$y_{ij} = 5 + 2x_{ij} + v_i + e_{ij}$$

for $i = 1K$ m and $j = 1K$ N_i , with $m = 42$ and $N_i = 100$. The values x_{ij} of the auxiliary variable were drawn from a uniform distribution between 0 and 10, the vector $\mathbf{v} = [v_1, \dots, v_m]^T$ of the random area specific effects was generated from $MVN(\mathbf{0}, \sigma_u^2 [(\mathbf{I} - \rho \mathbf{W}^T)(\mathbf{I} - \rho \mathbf{W})]^{-1})$ with $\sigma_u^2 = 1$. The SAR dispersion matrix was generated with ρ equal to $0, \pm 0.25, \pm 0.50, \pm 0.75$ and the neighbourhood structure ($\mathbf{W}$) was defined by randomly assigning neighbours for each area as follows: The value 1 was assigned to if the value drawn from a uniform $[0, 1]$ distribution was greater than 0.5, otherwise it was set to 0. The maximum number of neighbours for each area was 5, and the $\mathbf{W}$ matrix was standardized by row, i.e. the row elements summed to one. We can therefore refer to ρ as an autocorrelation parameter. The $\mathbf{W}$ matrix was kept fixed for all simulations.

The experiment is designed to investigate also the performance of the small area methods in the presence of outliers. For this reason the individual error e_{ij} were generated from a contamination distribution. This means that in ten areas, randomly selected, a $(1 - \delta)$ proportion of the e_{ij} 's were generated from the underlying 'true' distribution $N(0, 4)$ and the remaining δ proportion of the e_{ij} 's were generated from the contaminated distribution $N(0, 100)$. The choice

$\delta = 0$ corresponds to uncontaminated data. For contaminated data we fixed $\delta = 0.1$.

We conducted a total of $T = 200$ independent simulations, consisting of generating population and sample data as described above. For each sample drawn of size 420 with $n_i = 10 \forall i$, the small area mean was estimated using (a) the EBLUP estimator, (b) the SEBLUP estimator, (c) the M-quantile estimator and (d) the Spatial M-quantile estimator.

For each estimator we computed the Average Absolute Relative Bias ($\overline{ARB}$), the Average Relative Root MSE ($\overline{RRMSE}$) and the Percentage Relative Bias $\overline{PRB}$ defined as follows:

$$\begin{aligned}\overline{ARB} &= \frac{1}{m} \sum_{i=1}^m \frac{1}{T} \sum_{t=1}^T |\hat{Y}_{it} / Y_i - 1| \\ \overline{RRMSE} &= \frac{1}{m} \sum_{i=1}^m \frac{[MSE(\hat{Y}_i)^{1/2}]}{Y_i} \\ \overline{PRB} &= \frac{1}{m} \sum_{i=1}^m \frac{B_i^2}{MSE(\hat{Y}_i)}\end{aligned}$$

where

$$\overline{MSE}(\hat{Y}_i) = \frac{1}{m} \sum_{i=1}^m \frac{1}{T} \sum_{t=1}^T (\hat{Y}_{it} - Y_i)^2 \text{ and } B_i^2 = \frac{1}{T} \sum_{t=1}^T (\hat{Y}_{it} - Y_i)^2.$$

On the $T = 200$ simulations the coverage rate (CR) of the nominal 95% confidence interval has been obtained by using the analytical MSE estimators recalled in sections 2 and 3 for each of the presented estimators.

Tables 1 and 2 summarise the results from the simulation study for $\rho = \pm 0.25, \pm 0.75$. Results obtained with other ρ values are available from the authors.

Table 1 allows us to examine the usefulness of Spatial M-quantile modelling for capturing the spatial structure of the data used for small area estimation. When spatial correlation is high and positive the SEBLUP and Spatial M-quantile outperform the other estimators in term of $\overline{RRMSE}$. The estimators are biased and the Spatial M-quantile is a good competitor of SEBLUP only when the CD correction is adopted. In the case of negative spatial correlation the results are confirmed, even if the gain of spatial modelling in terms of $\overline{RRMSE}$ and bias reduction slightly decreases. If the spatial correlation is negligible, the traditional estimators achieve results that are similar to those of the spatial models.

Table 1. Comparison of small area estimators, uncontaminated data

	Estimator	$\overline{ARB}$ (%)	$\overline{RRMSE}$ (%)	PRB(%)	CR(%)
$\rho = 0.75$	SEBLUP	0.83	7.85	19.09	95.4
	EBLUP	0.84	8.02	17.31	95.3
	M-quantile	1.21	8.37	32.95	93.3
	M-quantile CD	0.20	8.67	0.38	92.2
	Spatial M-quantile	1.12	7.85	16.16	94.0
	Spatial M-quantile CD	0.20	8.67	0.38	91.9
$\rho = 0.25$	SEBLUP	0.99	8.58	21.06	95.4
	EBLUP	0.97	8.52	20.38	95.2
	M-quantile	1.05	8.61	28.06	94.4
	M-quantile CD	0.21	9.53	0.37	92.3
	Spatial M-quantile	1.27	8.64	18.17	93.9
	Spatial M-quantile CD	0.21	9.53	0.37	92.1
$\rho = -0.25$	SEBLUP	1.05	8.81	20.53	95.4
	EBLUP	1.03	8.75	19.78	95.2
	M-quantile	1.08	8.85	27.60	94.4
	M-quantile CD	0.22	9.77	0.37	92.2
	Spatial M-quantile	1.29	8.89	17.89	93.9
	Spatial M-quantile CD	0.22	9.77	0.38	92.1
$\rho = -0.75$	SEBLUP	1.14	9.09	18.33	95.3
	EBLUP	1.14	9.17	17.02	95.1
	M-quantile	1.37	9.45	28.71	93.9
	M-quantile CD	0.23	9.97	0.38	92.3
	Spatial M-quantile	1.45	9.21	16.27	93.7
	Spatial M-quantile CD	0.23	9.97	0.38	92.1

The exam of Table 2 is useful to investigate the performance of the small area methods in the presence of outliers.

Generally, both for high and low values of $|\rho|$, the M-quantile estimators have smaller bias in comparison with EBLUP and SEBLUP. This is an expected result as the peculiarity of the M-quantile modelling is to correct the influence of outliers on bias. Particularly, the Spatial M-quantile estimators present the lowest value of $\overline{ARB}$. This seems to be due to the combination of the control of outliers with the advantage of considering the spatial structure of the data.

This ability is also responsible for the good performance of Spatial M-quantile in terms of $\overline{RRMSE}$. When the spatial correlation is high and positive, the estimator achieves the lowest level of $\overline{RRMSE}$ in Table 2.

The coverage rate of the empirical confidence interval is always close to its 95% nominal value for all the investigated estimators under each level of spatial correlation and contamination scenario.

Table 2. Comparison of small area estimators, contaminated data, thirty-two uncontaminated areas

	Estimator	$\overline{ARB}$ (%)	$\overline{RRMSE}$ (%)	PRB(%)	CR(%)	AwCIs
$\rho = 0.75$	SEBLUP	1.88	8.98	40.56	98.5	2.88
	EBLUP	2.51	9.02	7.87	98.8	2.92
	M-quantile	1.88	8.80	31.44	93.4	2.31
	M-quantile CD	-0.01	9.21	0.53	92.3	2.30
	Spatial M-quantile	1.20	8.50	31.95	94.0	2.31
	Spatial M-quantile CD	-0.02	9.21	0.53	92.6	2.30
$\rho = 0.25$	SEBLUP	2.03	9.82	44.66	97.6	2.81
	EBLUP	2.03	9.56	42.88	97.5	2.87
	M-quantile	1.13	9.01	28.06	94.3	2.29
	M-quantile CD	-0.03	9.92	0.53	92.4	2.29
	Spatial M-quantile	1.31	9.19	34.46	94.0	2.30
	Spatial M-quantile CD	-0.04	9.93	0.53	92.5	2.30
$\rho = -0.25$	SEBLUP	2.01	10.12	43.89	97.4	2.82
	EBLUP	1.87	9.83	42.24	97.0	2.77
	M-quantile	1.00	9.26	27.99	94.0	2.29
	M-quantile CD	-0.04	10.11	0.54	92.3	2.29
	Spatial M-quantile	1.27	9.41	32.88	95.1	2.30
	Spatial M-quantile CD	-0.04	10.11	0.54	92.9	2.30
$\rho = -0.75$	SEBLUP	1.87	10.14	42.22	97.9	2.87
	EBLUP	2.05	10.14	44.80	98.1	2.88
	M-quantile	1.27	9.91	29.60	92.6	2.30
	M-quantile CD	-0.03	10.30	0.54	92.5	2.30
	Spatial M-quantile	1.62	9.87	31.68	93.4	2.31
	Spatial M-quantile CD	-0.03	10.30	0.54	92.7	2.30

5. Application

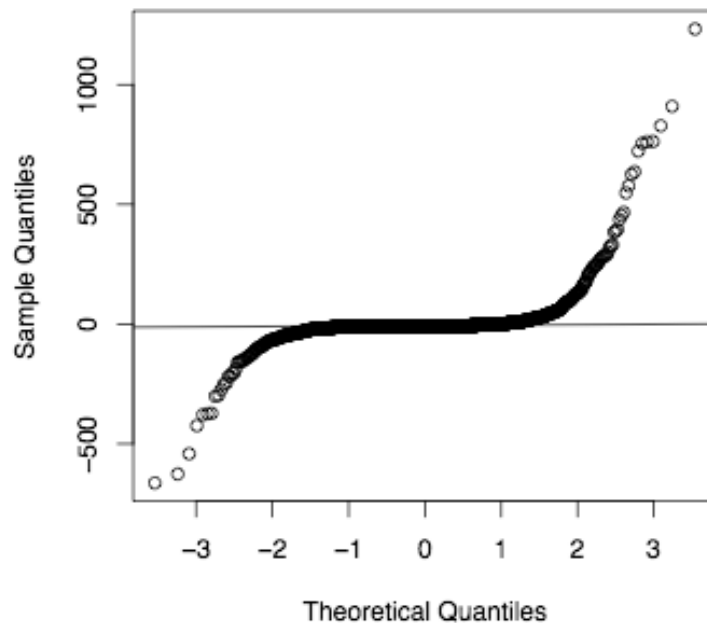
In the context of Italian agricultural surveys it is often of interest to produce accurate estimates of the average or of the total of farm production at local geographical areas, such as municipalities. However, such estimates can be difficult to produce due to the sparsity of the available survey data at this level of geography. As a result, previous work has focused on producing estimates at

higher geographical levels such as Italian provinces (Benedetti *et al.* 2004). Accurate estimates at sub-regional level require either the enlargement of the sample size or the application of small area estimation methods.

In this application we employ data from the Farm Structure Survey (FSS - ISTAT 2003) that is carried out once every two years and collects information on farm land by type of cultivation, amount of animal production and structure and amount of farm employment from 55,030 farms. The target of inference is the average production of olives per farm in quintal units for each of the (small) areas making up the Local Economy System (LES) in the Tuscany region. However, as our exploratory analysis will show, the presence of outliers in the data suggests that it may be also useful to produce estimates of median olive production in each of the LES areas.

The Atlas of Coverage of the Tuscany Region maintained by the Geographical Information System of the Regione Toscana provided information on coordinates, surface area and positions of the small areas of interest (UTM system). The centroid of each area is the spatial reference for all the units residing in the same small area. The auxiliary variable we employ in our models is the surface area used for olive production.

Exploratory analysis was first used to test for the presence of spatial dependence in the data. Essential to this is the neighbourhood structure $\mathbf{W}$ that is defined as follows: the 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. In order to detect the spatial pattern (spatial association and spatial autocorrelation) of the average production of olives per farm, two standard global spatial statistics have been calculated: Moran's I and Geary's C (Cliff and Ord, 1981). The spatial dependence in the target variable is weak, but the value of Moran's I is statistically significant. This is consistent with the estimated value for Geary's C .

Figure 1. Normal probability plot of the linear regression model residuals

Using Restricted Maximum Likelihood estimation the value of spatial autoregressive coefficient, $\hat{\rho}$, is estimated to be 0.441 ($s.e. = 0.183$), which suggests a moderate spatial relationship. In addition, as part of our exploratory analysis we also used a regression model for investigating the relationship between the production of olives in quintal units and the surface area used for olive production. A normal probability plot of the model residuals shows a skewed distribution of the residuals and hence evidence of outlying observations (Figure 1). Given the spatial correlation in the data and the presence of outliers, instead of using the Skew-normal and Skew-t distribution (Genton, 2004) for dealing with possible anomalous features of the data, we decided to perform small area estimation using a small area model that appears robust to outliers estimation method. A model of this type is the proposed spatial M-quantile model.

Small area estimates of olive production per farm at LES level are therefore produced under this model using the Spatial M-quantile CD estimator. The choice of the Spatial M-quantile CD estimator is justified (i) because of the presence of outliers in the data; as Tzavidis and Chambers (2007) suggest, when outliers are present in the data, the M-quantile CD estimator of the small area average is more efficient than the corresponding M-quantile estimator, and (ii) because one of the targets of our analysis is to estimate the small area medians. In order to obtain consistent estimators of small area medians (and other quantiles), it is necessary to base these estimators on a consistent estimator of the small area distribution

such as the Chambers-Dunstan estimator. Finally, in order to complete our comparisons we also present small area mean estimates using the EBLUP and SEBLUP estimators.

The maps in Figures 2 and 3 depict small area model estimates. Figure 2 shows the predicted values of the (a) average (Figure 2a) and (b) median (Figure 2b) of olive production per farm for each of the 42 LES areas in the Tuscany region under the Spatial M-quantile model. Figure 3 presents the small area estimates of the average of olive production per farm under EBLUP (Figure 3a) and SEBLUP (Figure 3b) estimators. We can note that EBLUP and SEBLUP estimates are very similar and they differ from the estimates obtained under the Spatial M-quantile model. The spatial distribution of M-quantile-based estimates appears to be less variable than that obtained with the traditional EBLUP and SEBLUP approaches. At this point we should also mention that in two LES areas the EBLUP and SEBLUP estimates of the small area means were negative. This can happen when there are outliers in the data which invalidate the assumptions of the linear mixed model. For these two LES areas we therefore decided to replace the negative model-based estimates (EBLUP and SEBLUP) with the corresponding direct estimates. We should also mention that we did not encounter negative small area estimates when using the M-quantile model. This is explained by the robust estimation method employed for fitting the M-quantile models.

Figure 2. Small area estimates of the (a) average and (b) median olive production per farm in quintal units for each of the 42 LES in the Tuscany region under Spatial M-quantile model

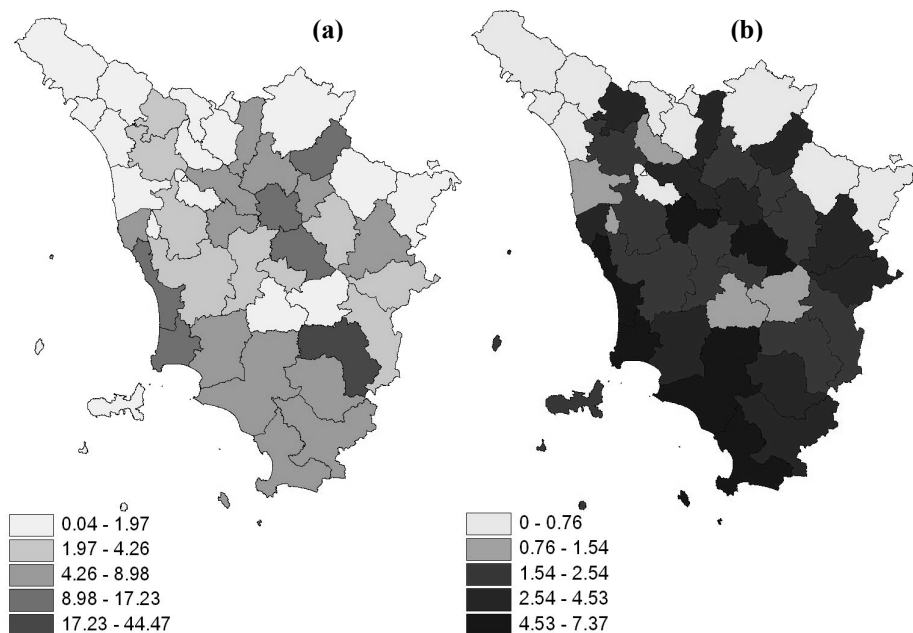
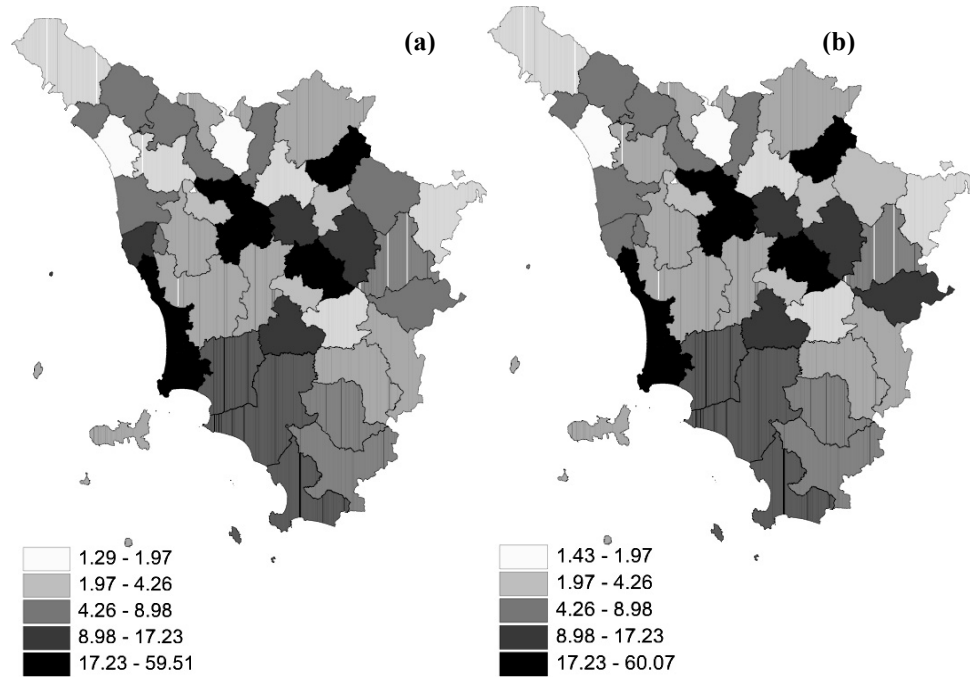


Figure 3. Small area estimates of the average olive production per farm in quintal units for each of the 42 LES in the Tuscany region under (a) EBLUP and (b) SEBLUP estimators.



Estimates of the small area median production of olives are also obtained under the Spatial M-quantile model. The median appears to be insensitive to the presence of few big farms that raise the average level of production. The spatial distribution of the median production also appears to be more homogenous than the corresponding spatial distribution of the small area means (see Figure 2). This emphasizes the importance of producing maps that represent not only the spatial distribution of the mean but also of other quantiles of the cumulative distribution function within each small area. The information contained in such maps is valuable both for agricultural policy interventions and for data users.

6. Conclusions

In this paper we propose an extension to the Chambers and Tzavidis (2006) small area M-quantile approach. The Spatial Empirical Best Predictor of $\bar{q}_i$ is obtained under a mixed model with spatially correlated random effects. Small area estimates are then obtained by fitting an M-quantile model at the average area specific M-quantile coefficient predicted under this parametric model. The proposed estimator of the small area mean, called Spatial M-quantile estimator, captures the spatial structure of the data and is robust in the presence of outliers.

Results from a simulation study show that this approach works well in comparison to the conventional M-quantile estimator. The main findings of our simulation study are: (i) when the data are not affected by outliers, in the case of strong spatial correlation the Spatial M-quantile estimator performs better than the M-quantile estimators; (ii) when the outliers contaminate the data, the properties of M-quantile estimators are confirmed, and in the case of high spatial correlation the Spatial M-quantile estimator emerges as the best choice in the small area mean estimation.

The empirical results confirm that the proposed semi-parametric approach offers one way of incorporating the spatial information in the M-quantile small area model preserving robustness to the presence of outliers.

A drawback of our approach is that we still need to specify a fully parametric model for the unit-specific M-quantile coefficients. We are currently investigating the use of non-parametric methods to incorporate spatial information into the M-quantile approach.

Acknowledgements

The work reported here has been developed under the support of the project PRIN *Metodologie di stima e problemi non campionari nelle indagini in campo agricolo-ambientale* awarded by the Italian Government to the Universities of Florence, Cassino, Pisa and Perugia. The authors are grateful to two anonymous Referees for their comments and suggestions that greatly improved the paper over earlier versions.

REFERENCES

- ANSELIN, L. (1992) *Spatial Econometrics: Method and Models*, Kluwer Academic Publishers, Boston.
- BATTESE, G.E., HARTER, R.M. and FULLER, W.A. (1988) An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, 83, 401, 28–36.
- BENEDETTI, R., ESPA, G. and PIERSIMONI, F. (2004) Esperienze di stime provinciali indirette di variabili aziendali, *Atti del Convegno ISPA 2004*, 6 maggio 2004, Università degli Studi di Cassino, 65–81.
- BNERJEE, S., CARLIN, B.P. and GELFAND, A.E. (2004) *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall, New York.

- BRECKLING, J., CHAMBERS, R. (1988) M-quantiles, *Biometrika*, 75, 4, 761–771.
- CHAMBERS, R., DUNSTAN, R. (1986) Estimating distribution function from survey data. *Biometrika*, 73, 597–604.
- CHAMBERS, R. and TZAVIDIS, N. (2006) M-quantile Models for Small Area Estimation, *Biometrika*, 93, pp. 255–268.
- CHAMBERS, R., CHANDRA, H. and TZAVIDIS, N. (2007) On robust mean squared error estimation for linear predictors for domains. [Paper submitted for publication. A copy is available upon request].
- CLIFF, A.D. and ORD, J.K. (1981) *Spatial Processes. Models & Applications*, Pion Limited, London.
- CRESSIE, N. (1993) *Statistics for spatial data*, John Wiley & Sons, New York.
- DATTA, G.S. and LAHIRI, P. (2000) A Unified Measure of Uncertainty of Estimates for Best Linear Unbiased Predictors in Small Area Estimation Problem, *Statistica Sinica*, 10, 613–627.
- GENTON, M. G. (2004) *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Edited Volume, Chapman & Hall / CRC, Boca Raton, FL, 416 pp.
- HENDERSON C. (1975) Best linear unbiased estimation and prediction under a selection model, *Biometrics*, 31, 423–447.
- KACKAR, R.N. and HARVILLE, D.A. (1984) Approximations for standard errors of estimators for fixed and random effects in mixed models, *Journal of the American Statistical Association*, 79, 853–862.
- PETRUCCI, A. and SALVATI, N. (2005) “Small Area Estimation: the Spatial EBLUP at area and at unit level”. Atti del Convegno “Metodi per l’integrazione di dati da più fonti”, Roma.
- PETRUCCI, A., PRATESI, M. and SALVATI, N. (2005) Geographic Information in Small Area Estimation: Small Area Models and Spatially Correlated Random Area Effects, *Statistics in Transition*, 7, 3, 609–623.
- PETRUCCI, A. and SALVATI, N. (2006) Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment, *Journal of Agricultural, Biological and Environmental Statistics*, 11, 2, 169–182.
- PFEFFERMANN, D. (2002) Small Area Estimation - New Developments and Directions, *International Statistical Review*, 70, 1, 125–143.
- PRASAD, N. and RAO, J. N. K. (1990), The estimation of the mean squared error of small-area estimators, *Journal of the American Statistical Association*, 85, 163–171.

- PRATESI MONICA, SALVATI NICOLA (2008) Small Area Estimation: the EBLUP estimator based on spatially correlated random area effects, *Statistical Methods & Applications*, 17, 1, 114–131.
- RAO, J.N.K. (2003) *Small area estimation*, John Wiley & Sons, New York.
- SAEI, A. and CHAMBERS, R. (2003) Small Area Estimation Under Linear and Generalized Linear Model With Time and Area Effects, *Working Paper M03/15*, Southampton Statistical Sciences Research Institute, University of Southampton.
- SALVATI, N. (2004) Small Area Estimation by Spatial Models: the Spatial Empirical Best Linear Unbiased Prediction (Spatial EBLUP), *Working Paper n 2004/04*, “G. Parenti” Department of Statistics, University of Florence.
- SINGH, B.B., SHUKLA, G.K. and KUNDU, D. (2005) Spatio-Temporal Models in Small Area Estimation, *Survey Methodology*, 31, 2, 183–195.
- TZAVIDIS, N. and CHAMBERS, R. (2006) Bias adjusted estimation for small areas with outlying values. *Southampton Statistical Sciences Research Institute, Working Paper M06/09*, Southampton.
- TZAVIDIS, N. and CHAMBERS, R. (2007) Robust prediction of small area means and distributions. Submitted for publication.

