

# Estimation of poverty indicators: a comparison of small area methods at LAU1-2 level in Tuscany\*

Caterina Giusti<sup>1</sup>, Monica Pratesi<sup>2</sup>, Nicola Salvati<sup>3</sup>

Department of Statistics and Mathematics Applied to Economics, University of Pisa

<sup>1</sup>e-mail: [caterina.giusti@ec.unipi.it](mailto:caterina.giusti@ec.unipi.it)

<sup>2</sup>e-mail: [m.pratesi@ec.unipi.it](mailto:m.pratesi@ec.unipi.it)

<sup>3</sup>e-mail: [salvati@ec.unipi.it](mailto:salvati@ec.unipi.it)

## Abstract

The aim of this paper is to compare the results of analyses based on different small area estimation methods, including some estimators recently proposed. The target is the estimation of the *at-risk-of-poverty rate* (Head Count Ratio – HCR) and of the mean and the quantiles of the income distributions. These measures can be considered as a starting point for more in deep analyses, such as the estimation of the income cumulative distribution function.

To compute the poverty indicators at the small area level, we consider several different methods, which can be grouped into two main categories: parametric and nonparametric small area estimators based on linear mixed models (Opsomer et al., 2008); parametric and nonparametric models based on M-quantile regression (Chambers and Tzavidis, 2006; Pratesi *et al.*, 2008). In particular, the nonparametric extensions of the parametric models, based on penalized splines, relax the hypothesis of a linear relationship between the variable of interest and the covariates, avoiding bias and providing a more flexible estimation procedure.

The focus of our analyses is the application of the parametric and nonparametric small area estimation methods on the same case study, since this comparison is still lacking in the literature. In particular, we perform an application to poverty mapping using census data and data from the EU-SILC (European Union - Statistics on Income and Living Conditions) survey on Tuscany provinces (Local Administrative Units 1 level).

**Keywords:** poverty mapping; linear mixed models; M-quantile models.

## 1. Introduction

The *at-risk-of-poverty rate* (Head Count Ratio – HCR) and the quantiles of the income distribution are widely used poverty indicators. Their estimation at small area level can be performed by a variety of methods among which linear mixed models estimators and M-quantile estimators have a prominent position. In fact, predictors based on mixed linear models are nowadays the most diffused methods to perform small area estimation. They rely on the assumption of normality and independence of the random area effects; in addition, they are linear models. Avoiding the assumption of normality, M-quantile methods offer a different approach to small area estimation.

---

\* Work supported by the project SAMPLE “Small Area Methodology for Poverty and Living Condition Estimates” awarded by the European Commission in the 7thFP.

However, linearity of the effect of covariates on the quantiles of the distribution of the study variable is still a feature of the M-quantile models; a nonparametric version of these models can be able to handle possible nonlinearities in this relation.

The aim of this paper is twofold: firstly we compare the results of the application of the methods to the estimation of the HCR and mean household income in the provinces of Tuscany (LAU1 level); secondly we approach the estimation of the quantiles of the income distribution in each province. The structure of the paper is as follows. Section 2 describes the models and the estimators. In section 3 the study variable and the covariates are defined. Moreover, the main features of the EU-SILC sample design in the small areas of interest and of the data from the Census of the population in Tuscany are briefly described. The discussion of the results, the final remarks and the envisioning of the future research lines conclude the paper in section 4.

## 2. Theory

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. In our case study the focus is on (a) the mean and the distribution function of household income in each small area (b) the percentage of households below the poverty line (Head Count Ratio).

A popular approach for this purpose is to use mixed effects models with random area effects to model household income. Given the so-called unit level nested error regression model (Battese *et al.*, 1988), the Empirical Best Linear Unbiased Predictor (EBLUP) of the mean for small area  $i$  is:

$$\hat{m}_i^{EBLUP} = N_i^{-1} \left[ \sum_{j \in s_i} y_j + \sum_{j \in r_i} \hat{y}_j \right] \quad (2.1)$$

where  $\hat{y}_j = \mathbf{x}_j^T \hat{\boldsymbol{\beta}} + z_j \hat{u}_i$ ,  $s_i$  denotes the  $n_i$  sampled units in area  $i$ ,  $r_i$  denotes the remaining  $N_i - n_i$  units in the area and  $\hat{\boldsymbol{\beta}}$ ,  $\hat{u}_i$  are obtained by substituting an optimal estimate of the covariance matrix of the random effects into the best linear unbiased estimator of  $\boldsymbol{\beta}$  and the best linear unbiased predictor of  $u_i$  respectively. The Mean Squared Error (MSE) of (2.1) and its estimates are obtained following the results of Kackar and Harville (1984) and Prasad and Rao (1990). Details and formulas can be found in Rao (2003, Chapter 7).

Recently, Chambers and Tzavidis (2006) have developed another approach to small area estimation based on the quantiles of the conditional distribution of the variable of study ( $y$ ) given the covariates (Breckling and Chambers, 1988). The  $q^{th}$  M-quantile  $Q_q(x; \boldsymbol{\psi})$  of the conditional distribution of  $y$  given  $x$  satisfies:

$$Q_q(\mathbf{x}_{ij}; \boldsymbol{\psi}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\psi(q) \quad (2.2)$$

where  $\boldsymbol{\psi}$  denotes the influence function associated with the M-quantile. For specified  $q$  and continuous  $\boldsymbol{\psi}$ , an estimate  $\hat{\boldsymbol{\beta}}_\psi(q)$  of  $\boldsymbol{\beta}_\psi(q)$  is obtained via an iterative weighted least squares algorithm. When (2.2) holds the bias adjusted M-quantile predictor of  $m_j$  is:

$$\hat{m}_j^{MQ/CD} = N_i^{-1} \left[ \sum_{j \in s_i} y_j + \sum_{j \in r_i} \mathbf{x}_j^T \hat{\beta}_\psi(\hat{\theta}_i) + \frac{N_i - n_i}{n_i} \sum_{j \in s_i} (y_j - \hat{y}_j) \right] \quad (2.3)$$

where  $\hat{y}_j = \mathbf{x}_j^T \hat{\beta}_\psi(\hat{\theta}_i)$  is a linear combination of the auxiliary variables and  $\hat{\theta}_i$  is an estimate of the average value of the M-quantile coefficients of the units in area  $i$  (Tzavidis and Chambers, 2007). The MSE of the estimator (2.3) can be estimated analytically as suggested in Chambers *et al.* (2007).

When the functional form of the relationship between the response variable and the covariates is unknown or has a complicated functional form, an approach based on use of a nonparametric regression model can offer significant advantages compared with one based on a linear model. A technique of nonparametric regression modelling is by using penalized splines or p-splines, see Eilers and Marx (1996). By expressing the spline coefficients in the model as random effects, Ruppert *et al.* (2003) show how fitting a p-spline model is equivalent to fitting a linear mixed model. On the basis of this property, Opsomer *et al.* (2008) have recently proposed a new approach to small area estimation that extends the unit level nested error regression model (Battese *et al.*, 1988) by combining small area random effects with a p-spline regression model. Opsomer *et al.* (2008) studied the theoretical properties of the mean squared error (MSE) of the small area mean estimator and proposed a bootstrap estimator for it. This performs reasonably well, but is computationally intensive and it is not considered here.

By using penalized splines, Pratesi *et al.* (2008) have extended the M-quantile regression to nonparametric regression, in the sense that the M-quantile regression functions do not have to be assumed to have a certain parametric form, but can be left undefined and estimated from the data. The authors have applied the nonparametric M-quantile regression to the small area estimation framework. Mean squared error (MSE) estimation of M-quantile based small area mean estimators relies on the approach described in Chambers *et al.* (2007). A nonparametric bootstrap technique for estimating the MSE of the biased adjusted estimator of the small area distribution function (2.4) and its confidence interval was proposed in Pratesi *et al.* (2008).

While there are many alternative estimators of the small area mean, the estimators of the distribution function have not yet been developed at small area level. A useful starting point is the so-called Chambers and Dunstan biased adjusted estimator of the small area distribution function in the presence of outliers (Tzavidis *et al.*, 2008a).

This is defined as:

$$\hat{F}_i^{CD}(t) = N_i^{-1} \left[ \sum_{j \in s_i} I(y_j \leq t) + n_i^{-1} \sum_{j \in s_i} \sum_{k \in r_j} I\{\hat{y}_k + (y_j - \hat{y}_j) \leq t\} \right]. \quad (2.4)$$

The  $p^{th}$  quantile  $m_{pi}$  of the distribution of  $y$  in area  $i$  can be estimated by the solution to

$$\int_{-\infty}^{m_{pi}} d\hat{F}_i(t) = p \quad (2.5)$$

We can note that by substituting properly  $\hat{y}_j$  in (2.3) one can define parametric or nonparametric M-quantile or mixed model versions of the CD-based mean estimator as well as corresponding estimators of the within area quantiles of  $y$ .

The MSE of (2.4) is under study and some preliminary results are to appear (Tzavidis *et al.*, 2008b).

### 3. Application

In Italy, the European Survey on Income and Living Conditions (EU-SILC) is conducted yearly by ISTAT to produce estimates of the living conditions of the population at national and regional (NUTS3 - LAU1) levels.

The regional sample in Tuscany is based on a stratified two stage sample design: in each province the municipalities are the primary sampling units (PSUs), while the households are the Secondary Sampling Units (SSUs). The PSUs are divided into strata according to their dimension in terms of population size; the SSUs are selected by means of systematic sampling in each PSU. All members of each sample household are interviewed through an individual questionnaire, and one individual in each household (usually, the household head) is interviewed through an household questionnaire.

Tuscany Region is a planned domain for which EU-SILC estimates are published, while the provinces are unplanned domains. These are 10 administrative areas constituted by a different number of municipalities (NUTS4 – LAU2 level) and whose boundaries do not cut across the municipalities themselves. It is useful to note that some Provinces – generally the smaller ones – may have very few sampled municipalities; furthermore, many municipalities are not even included in the sample at all (233 out of 287 municipalities had a zero sample size in the 2004 survey). Direct estimates may therefore have large errors at province level or they may not even be computable at municipalities level, thereby requiring resort to small area estimation techniques.

Data sources for the present application are the 2004 EU-SILC survey (for the  $n_i$  sampled units in area  $i$ ) and the 2001 Population Census of Italy, with a total of 1388260 individuals in Tuscany (for the not sampled  $N_i - n_i$  units in area  $i$ ).

In 2004 the EU-SILC regional sample size in Tuscany was of 1751 households; 54 municipalities were included in the sample. The small areas of interest are the 10 Tuscany provinces, with sample sizes  $n_i$  ranging from 70 (Province of Grosseto) to 545 (Province of Firenze). Due to the large sample size in the Province of Firenze and to the differences characterising that territory, we consider the Municipality of Florence, with 178 units out of 545, as a stand-alone small area.

The characteristic of interest  $y$  is the household disposable income, equivalised according to the Eurostat guidelines (European Commission, 2006). We are interested in estimating the small area mean of  $y$  and the Head Count Ratio in the small area. In each area, the HCR is computed both as the percentage of households below the poverty line and as the percentage of individuals below the poverty line (Foster *et al.*, 1984). The poverty line, equal to 9188,16 Euros, is computed as the 60% of the median of the household equivalent income.

Log-transformation of household income has not been considered at this stage of the work, to avoid the possible bias and the complications of the back-transformation on the MSE estimation of the small area estimators (Chambers and Dorfman, 2003).

The following auxiliary variables are known for each unit in the population and have resulted significant in the models for income:

- size of the household in terms of the number of components of the household  $j$  in the small area  $i$  (integer value);
- age of the head of the household  $j$  in the small area  $i$  (integer value);
- years in education of the head of the household  $j$  in the small area  $i$  (integer value);

- working position of the head of the household  $j$  in the small area  $i$  (employed/unemployed in the previous week);
- tenure status of household  $j$  in the small area  $i$  (owner/tenant).

The results of the application of the estimators described in section 2 are shown in Tables 1-4. The estimators of the mean and of the HCR under the linear mixed model are denoted as EBLUP, while the one under the M-quantile model as MQ CD. The nonparametric versions of the previous estimators are denoted, respectively, as NEBLUP and as NPMQ.

**Table 1.** Head Count Ratio (computed as the number of individuals below the poverty line) under all the estimators.

Province	MQ CD	EBLUP	NPMQ	NEBLUP
Massa	0,23	0,25	0,21	0,24
Lucca	0,17	0,20	0,16	0,18
Pistoia	0,12	0,15	0,12	0,15
Province of Firenze	0,13	0,15	0,12	0,14
Livorno	0,15	0,18	0,15	0,18
Pisa	0,16	0,19	0,15	0,18
Arezzo	0,11	0,14	0,10	0,14
Siena	0,12	0,14	0,12	0,14
Grosseto	0,16	0,19	0,15	0,18
Prato	0,15	0,17	0,15	0,16
Municipality of Firenze	0,13	0,15	0,13	0,16

**Table 2.** Mean of the equivalent household income under the EBLUP estimators.

Province	EBLUP	CV%	NEBLUP	CV%
Massa	15643,99	5,55	15678,08	2,52
Lucca	15805,37	5,50	16035,05	2,43
Pistoia	16467,34	5,16	16464,34	2,39
Province of Firenze	16326,55	3,85	16413,94	2,35
Livorno	17110,78	4,92	16874,11	2,34
Pisa	15950,31	5,18	16068,30	2,44
Arezzo	17327,64	4,71	17155,62	2,24
Siena	16660,39	5,11	16595,37	2,33
Grosseto	16592,89	5,65	16522,92	2,34
Prato	16963,66	4,86	16909,10	2,30
Municipality of Firenze	18138,89	4,45	17809,70	2,33

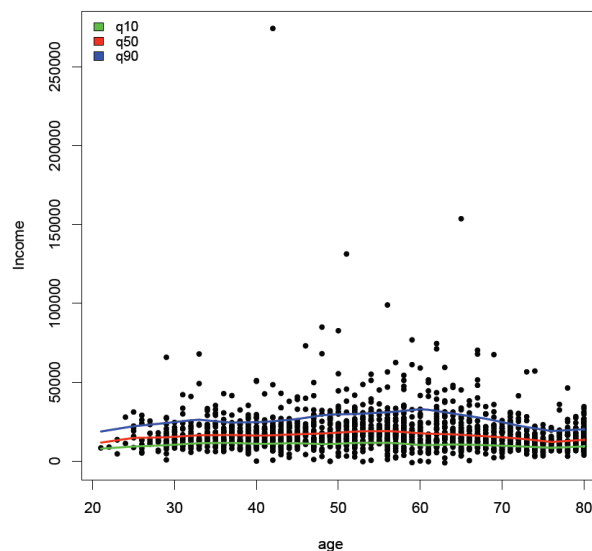
Two main results stand from the tables: firstly, all the estimators give the same indication about the monetary poverty in the small areas. Massa province has the highest percentage of poor individuals (HCR more than 20% under all the models) while Arezzo has the lowest one (HCR in the 10-14% range under all the models). The same conclusions follow from the HCR computed in each small area as the number households below the poverty line (results not reported here). In the next step of the analysis the MSE of the HCR estimates will be also computed, using a nonparametric bootstrap.

The estimates of the household mean income confirm that the Province of Massa is one of the poorest areas, while the Province of Arezzo, together with the Municipality of Firenze, is one of the richest.

**Table 3.** Mean of the equivalent household income under the M-quantile estimators.

Province	MQ CD	CV%	NPMQ	CV%
Massa	15374,38	8,36	15517,73	8,21
Lucca	15213,17	4,49	15607,18	4,38
Pistoia	16717,31	5,42	17046,68	5,27
Province of Firenze	16286,02	2,50	16499,26	2,43
Livorno	18297,97	10,82	18312,56	10,82
Pisa	15731,30	4,21	16001,32	4,04
Arezzo	18078,32	5,68	18212,06	5,59
Siena	16711,60	3,66	16822,36	3,61
Grosseto	17624,68	8,27	17729,03	8,07
Prato	17201,37	4,91	17231,80	4,89
Municipality of Firenze	19679,34	4,91	17231,80	4,85

A second significant conclusion is that the main differences among the results are in the variability of the estimates and in their precision. As it is well known, the EBLUP estimator tends to over-shrink the distribution of the small area estimates, underestimating the small areas between variability (Rao, 2003). In fact, the range of the EBLUP estimates of the mean is less wide than that of the MQ CD estimates. For what concerns the precision of the estimates, the Prasad and Rao (1990) method seems to smooth the precision of the estimates among the areas in the EBLUP case. The precision of the MQ estimator has a wider range: the percentage coefficient of variation goes from about the 3% of the Municipality of Firenze to the 11% of the Province of Livorno, following approximately the distribution of the area sample sizes.

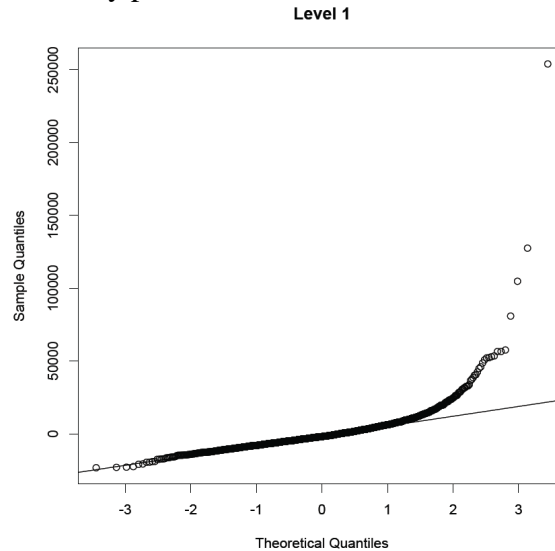
**Figure 1.** Nonparametric part of the nonparametric M-quantile small area model.

It is worth nothing that, in this estimation context, linear models seem able to handle the slight non-linearity in the relationship between the variable of interest and the age of the head of the household (see Figure 1).

The results obtained by NEBLUP and NPMQ are very close to those of their parametric versions. Note that the coefficients of variations under the NEBLUP models could be underestimated, since the Prasad and Rao (1990) specification of the

MSE does not take into account some sources of variability. Moreover, the Gaussian assumptions of the mixed models seem not to be met. Figure 2 shows the normal probability plots of level 1 residuals obtained by fitting a two-level mixed model to the sample data (where level 1 are the households and level 2 the Provinces).

**Figure 2.** Normal probability plots of level 1 residuals.



The plot denotes also the presence of outliers. Hence, the use of a model that relaxes these assumptions, such as an M-quantile model with a bounded influence function, seems more reasonable for these data.

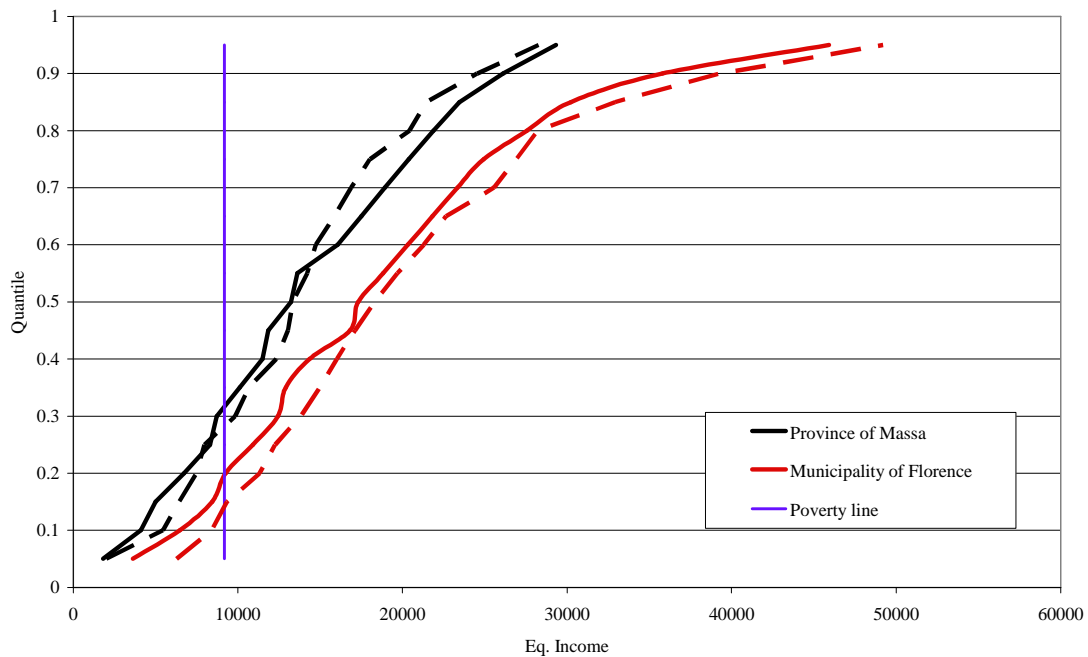
Figure 3 show the Chambers and Dunstan biased adjusted estimates of the small area distribution function.

The  $y$  values for not sampled units have been predicted using the MQ CD estimator. In particular we show the estimates for the Municipality of Florence (red solid line) and the Province of Massa (black solid lines). The direct estimates of the small area distribution function are in dashed lines, they are obtained simply applying the Horvitz and Thompson estimator to the  $n_i$  sampled units in area  $i$  (Särndal *et al.*, 1992).

The estimation of the quantiles of the cdf allows following the behaviour of the income distribution across the areas. The HCR of the province of Massa is more than 20% and it comes from a cumulative distribution that rapidly approaches the value 1. It is useful to note that it is always steeper than the analogous distribution of the Municipality of Firenze. In particular, the amount of people immediately after the poverty line (in blue) is bigger under the black line than under the red one.

The direct estimates (dashed lines) are consistent with the model based ones. The red lines, solid and dashed, are closer than the black ones reasonably because the sample size at each quantiles is larger in Florence than in Massa ( $n_i=178$  households versus 126).

**Figure 3.** The biased adjusted estimates of the small area distribution function.



#### 4. Conclusions and future perspectives

The aim of this paper is to compare the results of the poverty indicators at the small area level, considering several different parametric and nonparametric methods. The mean squared error of the estimators has been also evaluated. The target of the estimation are the *at-risk-of-poverty rate* (HCR) and of the quantiles of the income distributions.

It is useful summarizing the main results standing from our analyses.

Firstly, all the methods give the same indication about the poverty in the small areas. Thus, the ranking of the provinces by HCR and by mean income is coherent.

Secondly, the methods seem to have different accuracy. EBLUP estimators seem more accurate, but their precision is too homogeneous across the areas. M-quantile estimators seem to better track the differences in precision across the areas.

Thirdly, the M-quantile estimators seem to out-perform the other methods when the data do not follow the Gaussian distribution. Indeed, this technique does not depend on strong distributional assumption and is robust against outlying area values. Moreover the approach overcomes an important problem in small area estimation that is the impact of changing small area geographies on the estimates, but this aspect has not been explored here.

Finally, as regards the nonparametric methods in this estimation context, linear models seem able to handle the slight non-linearity in the relationship between the variable of interest and the covariates.

The work done has some limitations, which suggest directions of additional research.

The analyzed measures of poverty can be considered as only a starting point for more in deep analyses. First of all, analyses should be done using also non-monetary indicators in order to give a more complete picture of poverty and deprivation (Cheli and Lemmi, 1995).

The variable of interest has a typical log-normal asymmetric distribution. The log transformation of income is a step to do in order to enhance the fitting of the income model and to exploit the significance of the covariates. More work has to be done on the back-transformation to apply appropriate formulas for the mean squared errors.

Finally, the estimator of the income cumulative distribution function is a very useful tool to follow the behavior of the distribution of income at the small area level. This tool should be refined studying a method to estimate its mean squared error in order to track a confidence interval around the cumulative distribution function line.

## References

- Battese, G., Harter, R. and Fuller, W. (1988). An Error-Components Model for Prediction of County Crop Areas using Survey and Satellite Data. *Journal of the American Statistical Association*, 83, 28-36.
- Breckling J. and Chambers R. (1988). M-quantiles. *Biometrika*, 75, 761-71.
- Chambers R., Tzavidis N. (2006) M-quantile models for small area estimation, *Biometrika*, 93, 255-268.
- Chambers, R., Chandra, H. and Tzavidis, N. (2007). On Robust Mean Squared Error Estimation for Linear Predictors for Domains. CCSR Working paper 2007-10. Cathie Marsh Centre for Census and Survey Research, University of Manchester.
- Chambers R., Dorfman A.H. (2003). Transformed Variables in Survey Sampling. S3RI Methodology Working Papers, M03/21, Southampton Statistical Sciences Research Institute, University of Southampton, UK.
- Cheli, B. and Lemmi, A. (1995). A Totally Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty. *Economic Notes*, 24, 115-134.
- Eliers, P. and Marx, B. (1996). Flexible Smoothing using B-splines and Penalized Likelihood (with comments and rejoinder). *Statistical Science*, 11, 1200-1224.
- European Commission (2006). *Description of SILC Database Variables: Cross-sectional and Longitudinal*. Version 2004.1 from 25-05-06. European Commission – Eurostat.
- Foster J., Greer J., Thorbecke E. (1984). A class of decomposable poverty measures. *Econometrica*, 52, 761-766.
- 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.
- Opsomer J.D., Claeskens G., Ranalli M.G., Kauermann G., Breidt F. J. (2008) Nonparametric small area estimation using penalized spline regression, *Journal of the Royal Statistical Society: Series B*, 70, 265-286.
- Prasad, N.G.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 M., Ranalli M.G., Salvati N. (2008) Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US, *Environmetrics*, 19, 687-701.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: Wiley.
- Ruppert, D., Wand, M.P. and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, New York.
- Särndal C.E., Swensson B., Wretman J.H. (1992). *Model Assisted Survey Sampling*. New York, Springer-Verlag.

- Tzavidis N., Salvati N., Pratesi M., Chambers R. (2008a). M-quantile Models with Application to Poverty Mapping. *Statistical Methods & Applications*, 17, 393-411.
- Tzavidis N., Marchetti S., Chambers R. (2008b). Robust Estimation of Small Area Means and Quantiles [*Paper submitted for publication*].