Small area model-based direct estimator for spatial data

C. Bocci, A. Petrucci* & E. Rocco***

*Department of Statistics, Informatics, Applications "G. Parenti" University of Florence

**IRPET - Regional Institute for Economic Planning of Tuscany

chiara.bocci@irpet.it - alessandra.petrucci@unifi.it - rocco@disia.unifi.it

Motivation

- Analysis of the regional spatial pattern of socio-economical processes has become a relevant area of economics.
- Spatial data types spatial relationships, and spatial autocorrelation are complex and make difficult extracting patterns from spatial data sets.
- Novel approach to identify the spatial patterns using a semiparametric models that allow incorporation of spatial location as an additional component.
- Socio-economic data characteristic: economic variables (income, consumption, etc) are usually skewed
- Use of a log-linear geoadditive small area estimation model, in this context a model-based direct estimator (MBDE).
- Application: estimate the district level mean of the household per-capita consumption expenditure for the Republic of Albania.

Log-linear geoadditive small area model

Suppose that there are T small areas for which we want to estimate a quantity of interest Y that assumes strictly positive values with a skewed distribution for a population of size N. Let y_{it} denote the value of the response variable for the ith unit, $i = 1, ..., n$, in small area t, $t = 1, ..., T$ and let $\mathbf{x}_{it}$ be a vector of p covariates associated with the same unit, that are related to the log-transformed value of Y, and suppose that both are measured at spatial location $\mathbf{s}_i$, $\mathbf{s} \in \mathbb{R}^2$. A geoadditive SAE model for $y_{it}^* = log(y_{it})$ it is a linear mixed model with two random effects components:

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{1}$$

with

$$\mathrm{E}\begin{bmatrix}\boldsymbol{\gamma}\\ \mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}, \qquad \mathrm{Cov}\begin{bmatrix}\boldsymbol{\gamma}\\ \mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\sigma_\gamma^2\mathbf{I}_K & 0 & 0\\ 0 & \sigma_u^2\mathbf{I}_T & 0\\ 0 & 0 & \sigma_\varepsilon^2\mathbf{I}_n\end{bmatrix}.$$

and where $\mathbf{y}^* = \left[y_{it}^*\right]$, $\mathbf{X} = \left[\mathbf{x}_{it}^T, \mathbf{s}_{it}^T\right]_{1\le i\le n}$ has $p+2$ columns, $\boldsymbol{\beta}$ is a vector of $p+2$ unknown coefficients, $\mathbf{u}$ are the random small area effects, $\boldsymbol{\gamma}$ are the thin plate spline coefficients (seen as random effects), $\boldsymbol{\varepsilon}$ are the individual level random errors, the matrix $\mathbf{D}$ is defined as $\mathbf{D} = [d_{it}]$ with

$$d_{it} = \begin{cases}1 & \text{if observation } i \text{ is in small area } t,\\ 0 & \text{otherwise}\end{cases}$$

and, finally, $\mathbf{Z}$ is the matrix of the thin plate spline basis functions

$$\mathbf{Z} = \left[C\left(\mathbf{s}_i - \boldsymbol{\kappa}_k\right)\right]_{1\le i\le n, 1\le k\le K}\left[C\left(\boldsymbol{\kappa}_h - \boldsymbol{\kappa}_k\right)\right]_{1\le h,k\le K}^{-1/2},$$

with K knots $\boldsymbol{\kappa}_k$ and $C(\mathbf{v}) = \|\mathbf{v}\|^2 \log\|\mathbf{v}\|$.
The unknown variance components are estimated via REML or ML estimators and are indicated with $\hat\sigma_\gamma^2$, $\hat\sigma_u^2$ and $\hat\sigma_\varepsilon^2$. The estimated covariance matrix of $\mathbf{y}^*$ is

$$\hat{\mathbf{V}} = \hat\sigma_\gamma^2\mathbf{Z}\mathbf{Z}^T + \hat\sigma_u^2\mathbf{D}\mathbf{D}^T + \hat\sigma_\varepsilon^2\mathbf{I}_n \tag{2}$$

and the EBLUP estimators of the model coefficients are

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}^*, \tag{3}$$

$$\hat{\boldsymbol{\gamma}} = \hat\sigma_\gamma^2\mathbf{Z}^T\hat{\mathbf{V}}^{-1}(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{4}$$

$$\hat{\mathbf{u}} = \hat\sigma_u^2\mathbf{D}^T\hat{\mathbf{V}}^{-1}(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{5}$$

Finally the predicted values of y_{it} is derived as

$$\hat y_{it} = \hat\lambda_{it}^{-1}\exp\left(\mathbf{x}_{it}^T\hat{\boldsymbol{\beta}} + \frac{\hat v_{iit}}{2}\right) \tag{6}$$

where $\hat v_{iit}$ is the i-th diagonal element of $\hat{\mathbf{V}}$, and $\hat\lambda_{it}$ is the bias adjustment factor for the log-back transformation: $\hat\lambda_{it} = 1 + 0.5[\hat a_{it} + 0.25\hat V(\hat v_{iit})]$ where $\hat a_{it} = \mathbf{x}_{it}^T\hat V(\hat{\boldsymbol{\beta}})\mathbf{x}_{it}$, $\hat V(\hat{\boldsymbol{\beta}})$ is the usual estimator of $Var(\hat{\boldsymbol{\beta}})$ and $\hat V(\hat v_{iit})$ is the estimated asymptotic variance of $\hat v_{iit}$.

Small area mean estimator

For a given small area t, we are interested in predicting the area mean $\overline{Y}_t = N_t^{-1}\sum_{i=1}^{N_t} y_{it}$.
The generic MBDE is defined as a weighted average of the sample values Y in area t,

$$\hat{\overline{y}}_t^{MBDE} = \sum_{i\in S_t} w_{it}y_{it}.$$

where the specific weights w_{it} formulation depends on the underlying SAE model.
Under the log-linear geoadditive SAE model the empirical model-based model-calibrated weights can be written as:

$$\mathbf{w}^{mbmc} = [w_{it}^{mbmc}] = \mathbf{1}_s + \hat{\mathbf{H}}_s^T(\mathbf{J}_U^T\mathbf{1}_U - \mathbf{J}_s^T\mathbf{1}_s) + (\mathbf{I}_s - \hat{\mathbf{H}}_s^T\mathbf{J}_s^T)\hat{\mathbf{O}}_{ss}^{-1}\hat{\mathbf{O}}_{sr}\mathbf{1}_r \tag{7}$$

where $\hat{\mathbf{H}}_s = (\mathbf{J}_s^T\hat{\mathbf{O}}_{ss}^{-1}\mathbf{J})^{-1}\mathbf{J}_s^T\hat{\mathbf{O}}_{ss}^{-1}$,

$$\mathbf{J}_U = [\mathbf{1}_U\ \hat{\mathbf{y}}_U] = \begin{bmatrix}\mathbf{J}_s\\ \mathbf{J}_r\end{bmatrix}, \qquad \hat{\mathbf{O}}_U = \left[\hat\omega_{jk}\right]_{1\le j,k\le N} = \begin{bmatrix}\hat{\mathbf{O}}_{ss} & \hat{\mathbf{O}}_{sr}\\ \hat{\mathbf{O}}_{rs} & \hat{\mathbf{O}}_{rr}\end{bmatrix}.$$

$\mathbf{1}_U$ is the unit vector of size N, $\hat{\mathbf{y}}_U$ is the vector of the predicted values (6), $\hat\omega_{jk} = \hat{\mathrm{cov}}(y_j, y_k|\hat y_j, \hat y_k) = e^{(\mathbf{x}_j^T\hat{\boldsymbol{\beta}} + \mathbf{x}_k^T\hat{\boldsymbol{\beta}}) + (\hat v_{jj} + \hat v_{kk})/2}(e^{\hat v_{jk}} - 1)$, $\hat v_{jk}$ is the element j,k of matrix $\hat{\mathbf{V}}$, and the matrices $\mathbf{J}_U$ and $\hat{\mathbf{O}}_U$ are partitioned according to sample (s) and non-sample (r) units.
Following [2], we suggest the use of the Horvitz-Thomspon type, that is

$$\hat{\overline{y}}_t^{HT-MBDE} = N_t^{-1}\sum_{i\in S_t} w_{it}^{mbmc}y_{it}. \tag{8}$$

For the estimation of the MSE of (8) we followed [1] and [2] and adapted standard methods for estimating the MSE of a weighted domain mean estimate.

Real data example

- Republic of Albania is divided in 3 geographical levels: there are 12 prefectures, 36 districts and 374 communes.
- Main sources of statistical information available in Albania: 2001 Population and Housing Census (PHC) and the 2002 Living Standard Measurement Study (LSMS), both conducted in Albania by the INSTAT (Albanian Institute of Statistics).
- The household per-capita consumption expenditure in each district is modeled using a geoadditive SAE log-transformed model (1).
- The estimates of the average per-capita consumption expenditure in each of the 36 district areas are derived using the MBDE model-calibrated estimator (8).
- The covariates selected to fit the model are chosen following prior studies on poverty assessment in Albania and through the application of a stepwise procedure.
- All the variables are available both in LSMS and PHC surveys and the geographical location of each household is available for the LSMS data. For the non sampled data, the geographical location is approximated with the centroid coordinates of the commune where the household is located.
- The model has been fitted by REML using the `lme` function in the `R` package `nlme` and almost all the parameters are highly significant. The random effects' parameters are showed in Table 1.

Parameter	Estimate	Confidence Interval	p-value
σ_γ	0.4096	(0.2700 ; 0.6214)	< 0.001
σ_u	0.1756	(0.1290 ; 0.2389)	< 0.001
σ_e	0.3285	(0.3208 ; 0.3363)	< 0.001

Table 1: Estimated random effects' parameters of the geoadditive SAE log-transformed model for the household per-capita consumption expenditure at district level.

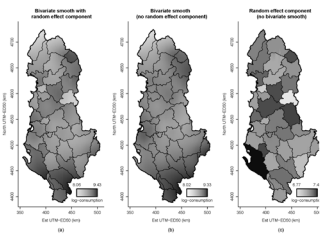

Figure 1: Spatial smoothing and district random effects of the household log per-capita consumption expenditure. Map (a) shows the smoothing obtained with the geoadditive sae model as sum of two components: the bivariate smoothing, map (b), and the small area random effects, map (c). Linear covariates are set at their mean values.

- The geoadditive SAE log-transformed model produces a graphic representation of the spatial pattern of the consumption expenditure in Albania (Figure 1) which shows the presence of both a spatial dynamic and a district level effect in the Albanian consumption expenditure.
- In order to map the estimated spatial smoothing of the log per-capita consumption expenditure we set the linear covariates at their mean values and we predict the consumption expenditure in a specific location as sum of three components: the constant linear part, the bivariate spline smoother and the small area effect.
- The district level estimates are showed in Figure 2 which exhibits a clear geographical pattern, with the higher values of the average household per-capita consumption expenditure in the south and south-west of the country (coastal area) and the lower value in the mountainous area (north and north-east of the country).

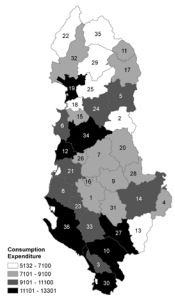

Figure 2: District level estimates of the average household per-capita consumption expenditure.

Remarks and Work in progress

- Empirical evidence suggests the spatial location is an important component to understand the distribution of the consumption expenditure.
- The geoadditive SAE model produces not only the map of estimated mean values, but also a spatial interpolation of all the observation.
- The location of all population units at the point level must be known. Otherwise, the classic approach is to refer the data with respect to the centroids of the areas to which they belong.
- An aspect to be explored is the use of a more precise spatial location data: an imputation approach which considers a more realistic hypothesis on spatial distribution. A possibility may be the imputation of the unknown sample locations based on the population density map overlayed on the map of the communes (Figure 3) through a probability proportion to size (PPS) imputation method. Further investigations will be done in this direction.

Figure 3: Population density (GEOSTAT 2011 grid dataset-EUROSTAT).

References

[1] R. Chambers, H. Chandra, and N. Tzavidis. On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, 37:153–170, 2011.

[2] H. Chandra and R. Chambers. Small area estimation under transformation to linearity. *Survey Methodology*, 37:39–51, 2011.