UNIVERSITÀ
DEGLI STUDI
FIRENZE

DiSIA
DIPARTIMENTO DI STATISTICA,
INFORMATICA, APPLICAZIONI
"GIUSEPPE PARENTI"

Small area estimation in the framework of multivariate models for sustainable development

E. Rocco, M.F. Marino, A. Petrucci
Dipartimento di Statistica, Informatica, Applicazioni "G.Parenti"

ERCIM 2918
11th International Conference of the ERCIM WG on Computational and Methodological Statistics

14-16 December – University of Pisa

Motivation

In studies concerning the equitable and sustainable development it is often desirable:

- to obtain estimates of population parameters in some sub-populations (*like small spatial areas or age domains*) that are not sampled or are under sampled in surveys;
- to obtain small area estimates of multidimensional characteristics;
- to obtain such estimates for variables of different nature that may do not satisfy the basic assumptions underlying the general linear model (e.g. binary, count).

Small area estimation

- Direct small area estimates may not provide acceptable precision when the small area sample size is small and are not applicable with zero sample size.
- In this framework, models represent a powerful tool since in addition to using auxiliary variables, they allow to borrow information across related areas.
- The most popular models for SAE are linear-mixed models that include independent random area effects to account for the variability between the areas exceeding that explained by auxiliary variables.
- Model-based SAE can be conducted based on either area level or unit level models on the basis of data availability.

Small area estimation models

- Several extensions to original SAE linear mixed models have been considered in the literature, including cases in which data follow various generalized linear models, or have more complicated random-effects structures.
- The aim of most of these extensions is to estimate a finite population mean of a **single** response variable for each small area.

Small area estimation of multiple characteristics

- In many small area problems, estimates for related multiple characteristics may be of interest.
- Moreover, in principle one can also improve SAE by making use of other survey outcomes that are related to the primary outcome of interest
- The multivariate SAE has not been studied so much
- For multivariate area level data Fay (1987) proposed the multivariate Fay-Herriot model and some extensions to its set-up have been considered in the literature.
- For multivariate unit level data the use of multivariate linear mixed models has been considered.

Small area estimation of multiple characteristics

- We assume to have *unit level data with multiple response variables of different nature* (continuous, semi-continuous, counting, dichotomous) and to point on the estimation of the finite population mean vector of such characteristics for small areas.
- When the aim is to estimate a finite population mean vector of multiple characteristics for each small area and the assumption of the multiple linear mixed model are not satisfied, the natural extension is to use multiple generalized linear mixed models.

Small area estimation of multiple characteristics

The use of multiple generalized linear mixed models allows to account for the multivariate dependence though the latent error term that is the specific small area random effect.

Basic setup, definitions and assumptions

- Let U be a finite population of N units, partitioned in m subsets (areas) of size N_i, with $\sum_{i=1}^{m} N_i = N$.
- $\mathbf{y}$ and $\mathbf{x}$ are a response vector (of size r) and an auxiliary variables vector (of size p) respectively.
- $\mathbf{y}_{ij}$ and $\mathbf{x}_{ij}$ denote the values of $\mathbf{y}$ and $\mathbf{x}$ respectively for the unit $j = 1, ..., N_i$ in small area $i = 1, ..., m$.

Basic setup, definitions and assumptions (II)

We assume that the following generalized linear mixed model relates the response variables to the auxiliary ones

$$\begin{cases} g_1(E[y_{ij1} \mid u_{i1}]) = & \mathbf{x}'_{ij1}\boldsymbol{\beta}_1 + u_{i1} \\ g_2(E[y_{ij2} \mid u_{i2}]) = & \mathbf{x}'_{ij2}\boldsymbol{\beta}_2 + u_{i2} \\ \vdots & \\ g_r(E[y_{ijr} \mid u_{ir}]) = & \mathbf{x}'_{ijr}\boldsymbol{\beta}_r + u_{ir} \end{cases} \tag{1}$$

where

- $g_k(\cdot), k = 1, \ldots, r$, are proper link functions
- $\mathbf{x}_{ijk}$ are subsets of x_{ij}, that is $\mathbf{x}_{ijk} \subseteq \mathbf{x}_{ij}$
- $\boldsymbol{\beta}_k$ is a vector of fixed unknown parameters describing the effect of covariates on the (transformed) response y_{ijk}
- u_{ik} is an area-specific effect which is meant to describe sources of unobserved heterogeneity not captured by $\mathbf{x}_{ijk}$

Basic setup, definitions and assumptions (III)

The proposed multivariate small area model is based on the following assumptions

- Conditional on u_{ik}, measures from different units in a given area i of the k-th response, area independent, with joint conditional density

$$f(\mathbf{y}_{ik} \mid u_{ik}) = f(y_{i1k}, \ldots, y_{iN_ik} \mid u_{ik}) = \prod_{j=1}^{N_i} f(y_{ijk} \mid u_{ik})$$

  where $f(y_{ijk} \mid u_{ik})$ denotes the EF density with canonical parameter $\theta_{ijk} = g_k^{-1}(E[y_{ijk} \mid u_{ik}])$

- Furthermore, conditional on the vector $\mathbf{u}_i = (u_{i1}, \ldots, u_{ik})'$, multiple responses from the same area i are independent, with joint conditional density

$$f(\mathbf{y}_i \mid \mathbf{u}_i) = f(\mathbf{y}_{i1}, \ldots, \mathbf{y}_{ir} \mid \mathbf{u}_i) = \prod_{k=1}^{r} f(\mathbf{y}_{ik} \mid u_{ik})$$

Basic setup, definitions and assumptions (IV)

- Finally, we assume that

$$\mathbf{u}_i = (u_{i1}, \ldots, u_{ir})' \sim N_r(\mathbf{0}, \boldsymbol{\Sigma}_u)$$

where $\boldsymbol{\Sigma}_u$ is a $r \times r$ dimensional covariance matrix

  - Diagonal elements of $\boldsymbol{\Sigma}_u$ denote the variances of the u_{ik}'s
  - Off-diagonal elements denote instead the covariance between couples $(u_{ik}, u_{ik'})$
  - These latter provide an (indirect) measure of dependence between the corresponding responses $(Y_{ijk}, Y_{ijk'})$

Small area estimation problem

- We are interested in predicting the vector of small area means $\bar{\mathbf{y}}_i = (\bar{y}_{i1}, \dots, \bar{y}_{ir})'$, where

$$\bar{y}_{ik} = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ijk}, \quad k = 1, \dots, r$$

- To this aim, a sample s of n units is selected from U according to a non-informative sampling design.
- s_i denotes the area-specific sample of size n_i, $(\bigcup_{i=1}^{m} s_i = s)$.
- Response y_{ijk} are observed for each unit in the sample.
- To predict $\bar{\mathbf{y}}_i$, we may first observe that each component, $\bar{y}_{ik}$, can be split into sampled and non-sampled elements

$$\bar{y}_{ik} = \frac{1}{N_i} \left[\sum_{j \in s_i} y_{ijk} + \sum_{j \notin s_i} y_{ijk} \right] \tag{2}$$

Small area estimation problem (II)

The non-sampled part of $\bar{y}_{ik}$, is derived as follows

- Estimate model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r, \boldsymbol{\psi})$, with $\boldsymbol{\psi}$ begin the vector of variance components, via a ML approach

$$\ell(\theta) = \sum_{i=1}^{m} \log \int f(\mathbf{y}_i \mid \mathbf{u}_i) \phi(\mathbf{u}_i) d\mathbf{u}_i$$

  where $\phi(\mathbf{u}_i)$ denotes the density of a r-variate Gaussian distribution
- For all $j \notin s_i$, compute the plug-in predictor

$$\hat{y}_{ijk} = g^{-1}(\mathbf{x}'_{ijk}\hat{\boldsymbol{\beta}}_k + \hat{u}_{ik})$$

  where $\hat{u}_{ik} = E(u_{ik} \mid \mathbf{y}_i)$
- Substitute $\hat{y}_{ijk}, j \notin s_i$ in equation (2)

Remarks

- In the presence of correlated responses, the proposed multivariate approach is expected to overcome the univariate counterparts
- This has to be interpreted in terms of efficiency, rather than bias
- Indeed, even when responses are correlated, the univariate approach returns unbiased estimates of the model's parameters
- However, as far as the efficiency is entailed, using the multivariate approach allows us to borrow strength not only from areas (as for the univariate approach), but also from multiple responses

Remarks (II)

- Furthermore, the proposed multivariate small area model directly nests the corresponding univariate ones
- When responses are uncorrelated, the covariance matrix for the area-specific effects reduces to

  $$\boldsymbol{\Sigma}_u = \boldsymbol{\sigma}_u \mathbb{I}_r$$

  where $\boldsymbol{\sigma}_u = (\sigma_{u_1}, \ldots, \sigma_{u_r})'$
- Last, but not least, it can be the case that analyzing the association structure between multiple responses is itself of interest
- Alternatively, one may be interested in predicting a function of multiple responses
- In these circumstances, knowing the covariance between variables is essential for estimating the variability of a such a transform

Model based simulation study

- In order to investigate the performance of the proposed multivariate SAE approach a large scale model-based simulation experiment is performed.
- Different scenarios are investigated.
- For each scenario, $T = 1000$ replicates are considered.
- Bivariate population data are generated under some model assumptions and sample data are selected from the simulated population.
- In this analysis, we only consider responses of the same type; the extension to mixed type responses is still ongoing work
- For each scenario, the multivariate small area estimates obtained though the multivariate GLMM are compared with the estimates obtained through the corresponding univariate models.

Model based simulation study (II)

- The population and the sample sizes are constant across areas and are fixed to $N_i = 100$ and $n_i = 10$, respectively.
- A varying number of areas is considered: $i = 1, \ldots, m$, with $m = 50, 100, 200$.
- In all scenarios, a unique auxiliary variable is considered for each unit j in small area i, that is

$$x_{ij} \sim Unif(1, i/b),$$

where $b = 4, 8, 16$, for $m = 50, 100, 200$, respectively.

Model based simulation study

Data are generated according to the following bivariate GLMM

$$\begin{cases} g_1(E[Y_{ij1} \mid u_{i1}]) = \beta_0 + x_{ij}\beta_{11} + u_{i1} \\ g_2(E[Y_{ij2} \mid u_{i2}]) = \beta_0 + x_{ij}\beta_{12} + u_{i2}, \end{cases}$$

- $g_k(\cdot), k = 1, 2,$ denotes a proper link function
- area-specific effects $\mathbf{u}_i = (u_{i1}, u_{i2})'$ are simulated from a bivariate Gaussian distribution

$$\mathbf{u}_i \sim N_2(\mathbf{0}, \mathbf{\Sigma}_u)$$

with two different specification for the covariance matrix

$$\mathbf{\Sigma}_u^{(high)} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix} \qquad \mathbf{\Sigma}_u^{(low)} = \begin{bmatrix} 1 & 0.32 \\ 0.32 & 1 \end{bmatrix}$$

Model based simulation study

Different types of responses are considered

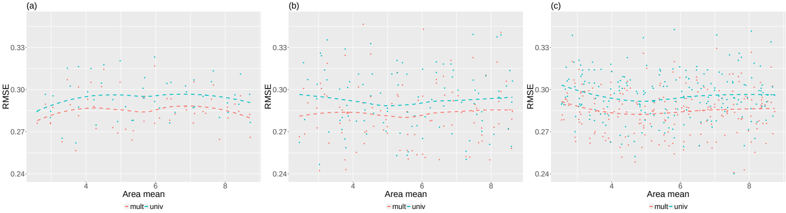
1. Gaussian data: $Y_{ij1} \mid u_{i1} \sim N(\cdot, \sigma^2_{e_1})$ and $Y_{ij2} \mid u_{i1} \sim N(\cdot, \sigma^2_{e_2})$
   - $g_1(\cdot) = g_2(\cdot) = 1$
   - $\sigma^2_{e_1} = 1.1$, $\sigma^2_{e_2} = 0.9$, and $\sigma_{e_1 e_2} = 0$
   - $\beta_0 = 3, \beta_{11} = 1$, and $\beta_{12} = 0.8$
2. Poisson data: $Y_{ij1} \mid u_{i1} \sim Pois(\cdot)$ and $Y_{ij2} \mid u_{i1} \sim Pois(\cdot)$
   - $g_1(\cdot) = g_2(\cdot) = \log(\cdot)$
   - $\beta_0 = 0.7, \beta_{11} = -0.1$, and $\beta_{12} = -0.2$
3. Bernoulli data: $Y_{ij1} \mid u_{i1} \sim Bern(\cdot)$ and $Y_{ij2} \mid u_{i1} \sim Bern(\cdot)$
   - $g_1(\cdot) = g_2(\cdot) = \text{logit}(\cdot)$
   - $\beta_0 = 0.5, \beta_{11} = -0.4$, and $\beta_{12} = -0.6$

Model based simulation study (II)

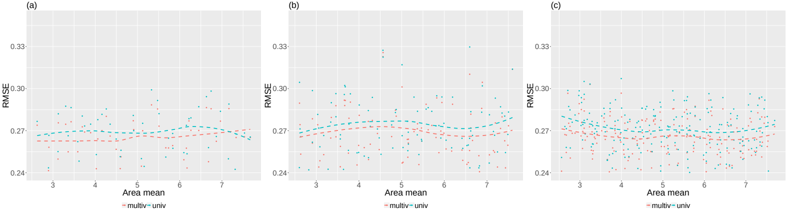
The performance of the small area estimators were evaluated by computing, for each area $i = 1, ..., m$, the Root Mean Squared Error (RMSE), defined as follows:

$$RMSE_i = \sqrt{T^{-1} \sum_{t=1}^{m} (\hat{\bar{y}}_{it}^{Model} - \bar{y}_{it})^2}$$

Gaussian data - high correlation

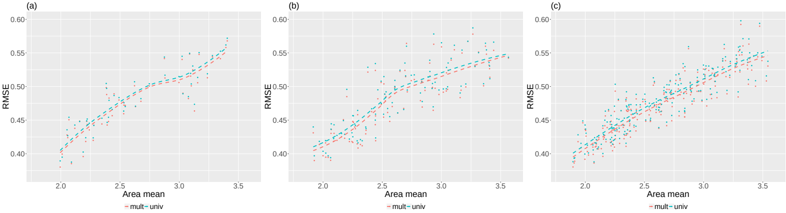


RMSE for response variable 1 with (a) m=50, (b) m=100, (c) m=200

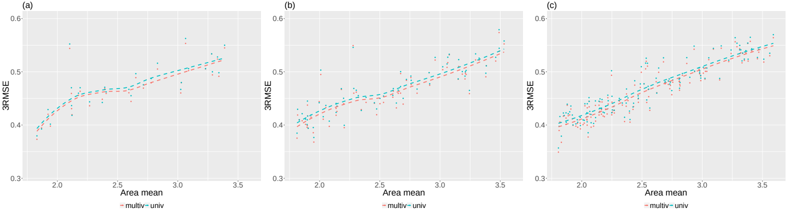


RMSE for response variable 2 with (a) m=50, (b) m=100, (c) m=200

Poisson data – high correlation

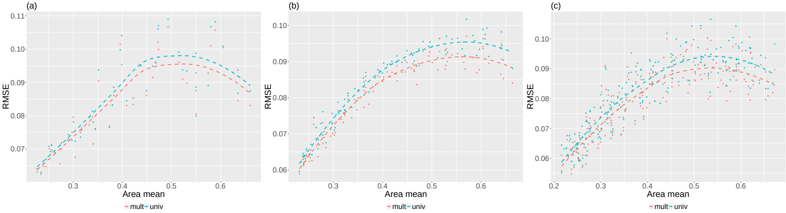


RMSE for response variable 1 with (a) m=50, (b) m=100, (c) m=200

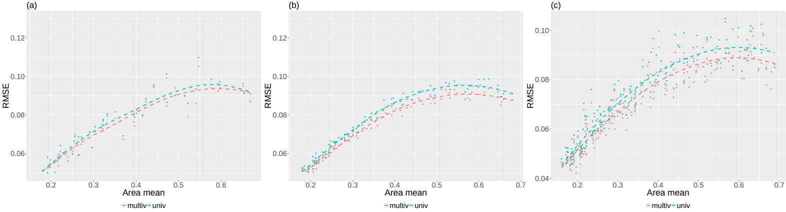


RMSE for response variable 2 with (a) m=50, (b) m=100, (c) m=200

Bernoulli data – high correlation

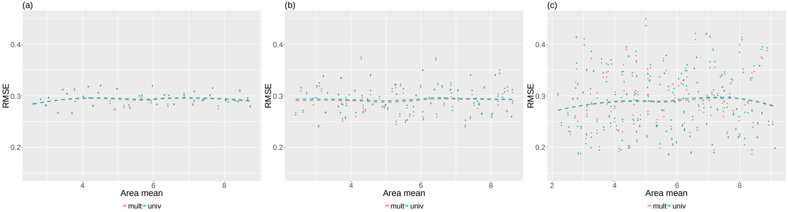


RMSE for response variable 1 with (a) m=50, (b) m=100, (c) m=200

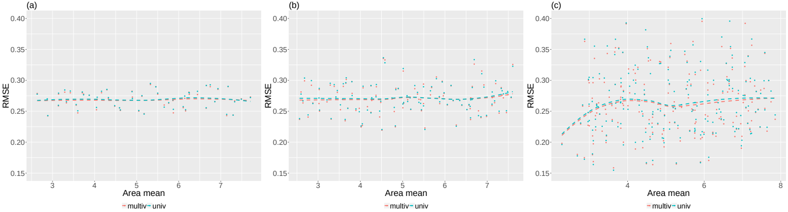


RMSE for response variable 2 with (a) m=50, (b) m=100, (c) m=200

Gaussian data – low correlation

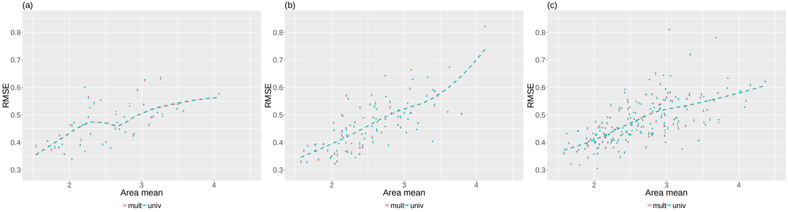


RMSE for response variable 1 with (a) m=50, (b) m=100, (c) m=200

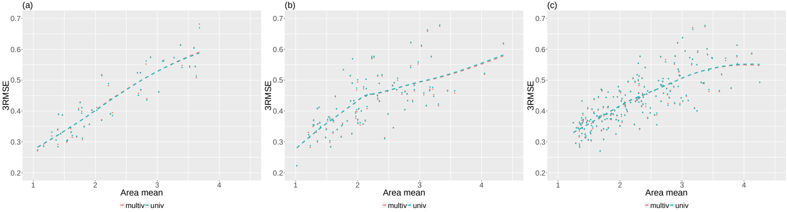


RMSE for response variable 2 with (a) m=50, (b) m=100, (c) m=200

Poisson data – low correlation

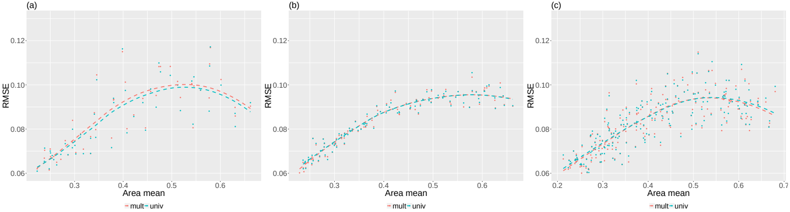


RMSE for response variable 1 with (a) m=50, (b) m=100, (c) m=200

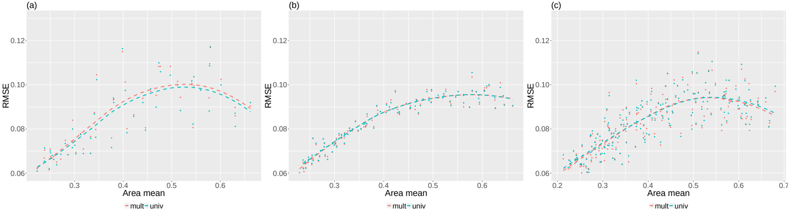


RMSE for response variable 2 with (a) m=50, (b) m=100, (c) m=200

Bernoulli data – lower correlation



RMSE for response variable 1 with (a) m=50, (b) m=100, (c) m=200



RMSE for response variable 2 with (a) m=50, (b) m=100, (c) m=200

Conclusion

- First results
  - From the empirical results it is evident that when there are highly correlated responses, the multivariate modelling is preferable to the univariate counterparts whatever is the nature (the distribution form) of the data
  - When the correlation is low, may be opportune to evaluate for each case (considering the nature of the data, the number of the areas, the size of the sample, the aims of the study) the trade of between the capability of the multivariate approach to exploit the relation among the two variables and the more complexity of the model itself
- WORK IN PROGRESS
  - We are still working in the evaluation of the performance of the multivariate approach on real data application
  - We have considered here only pairs of related variables of the same nature - the extension to mixed type responses is still ongoing work

Thank you!