Small Area Estimation for Skewed Georeferenced Data

E. Dreassi - A. Petrucci - E. Rocco

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

SAE 2013

THE FIRST ASIAN ISI SATELLITE MEETING ON SMALL AREA ESTIMATION (SAE)
September 1-4 2013, Bangkok, Thailand

UNIVERSITÀ DEGLI STUDI FIRENZE
DiSIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

Outline

1. Introduction
   - Motivation of the work
   - Data
   - Reference framework
2. Model and small area means estimation
   - Basic setup, definition and assumption
   - The two-part model
     - Part one
     - Part two
     - Random effects distributions
   - Bayesian formulation of the model
   - Model selection
   - Small area mean predictions
3. Real data example
   - Grape wine production analysis
4. Conclusions and ongoing research

UNIVERSITÀ DEGLI STUDI FIRENZE
DiSIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

Introduction

- Surveys with response variable that may have a continuous skewed distribution with a large number of values clustered at zero.
- We are interested in reliable small area estimates of some parameters (i.e. mean or total).
- Variables with often spatial distribution and small area are geographical domains.
- The small sample sizes in sampled areas requires the use of model based estimation methods.
- Standard small area estimation methods are not suitable for this kind of data.

Issues

- Operational small area model with:
- excess of zero values,
- skewed distribution of the nonzero values,
- spatial structure (related to georeferenced units).

Data description

- The Italian Statistical Institute (ISTAT) drives an Agricultural Census ten-yearly and a sample Farm Structure Survey (FSS) two-yearly.
- Both in the Census and in the FSS, the unit of observation is the farm and the data of the surface areas allocated to different crops are registered for each farm.
- In the FSS, until 2005, the productions of each crop were also observed.
- The FSS survey is designed to obtain estimates only at regional level.
- We are interest in producing the mean estimation of grapevine production for the 52 Agrarian Regions in which Tuscany region is partitioned.

UNIVERSITÀ DEGLI STUDI FIRENZE
DiSIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

Grapevine production I

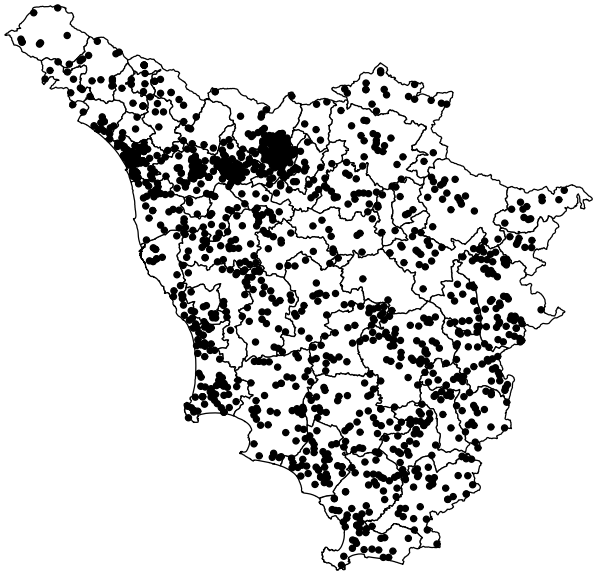

Figure : Spatial pattern of farms with zero grape wine production

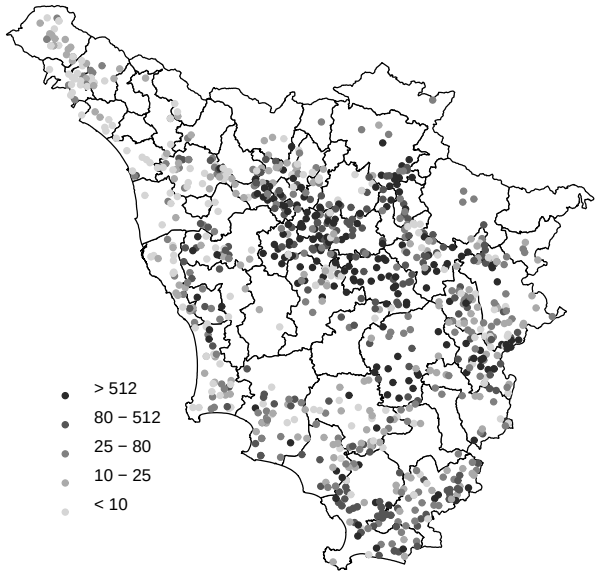

Figure : Spatial pattern of farms with positive grape wine production

Grapevine production II

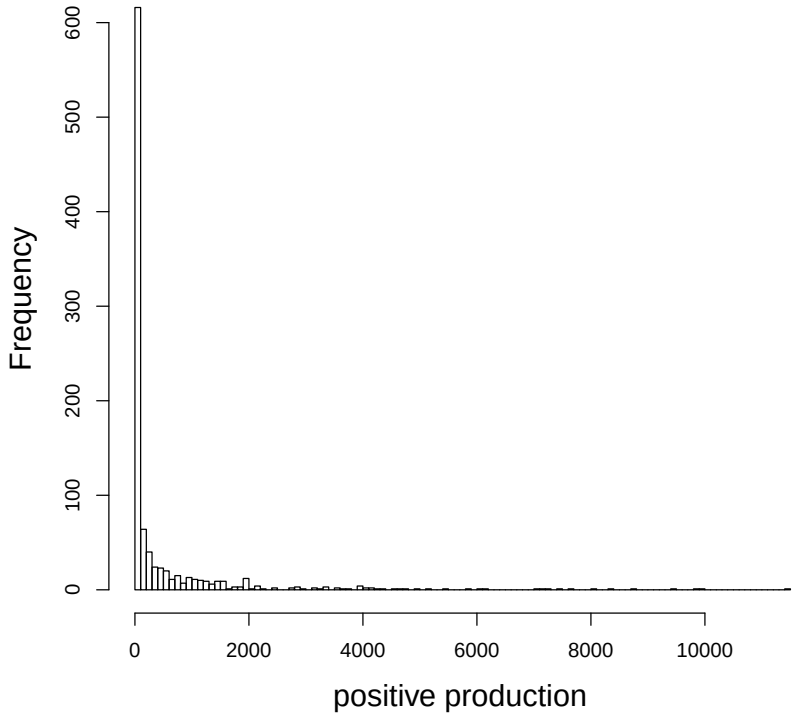

Figure : Histogram of positive production (Tuscany region)

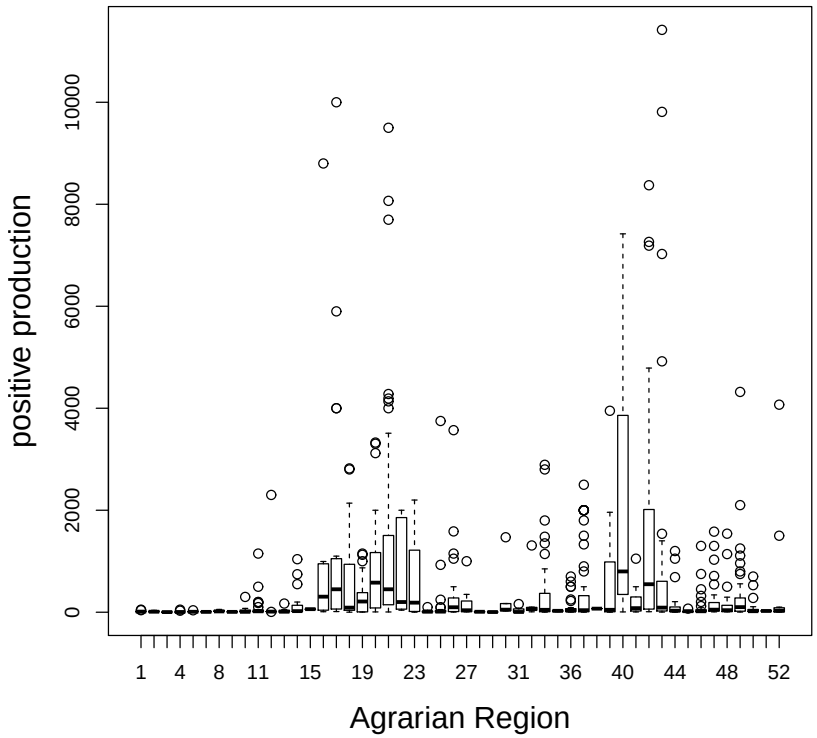

Figure : Box-plot for each ARs

Grapevine production III

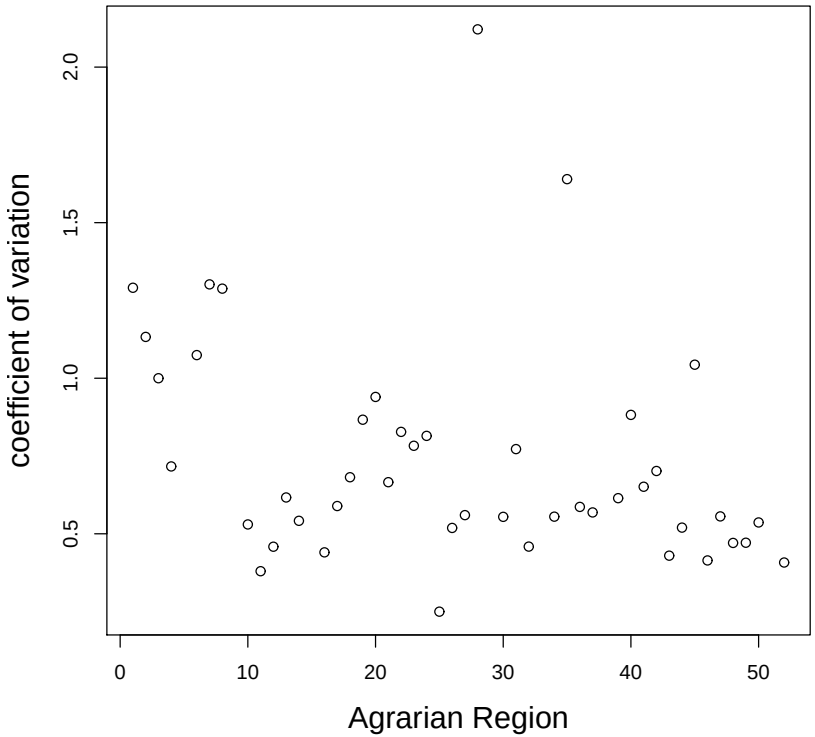

Figure : The positive grape wine production from sampled farms:the coefficient of variation for each AR

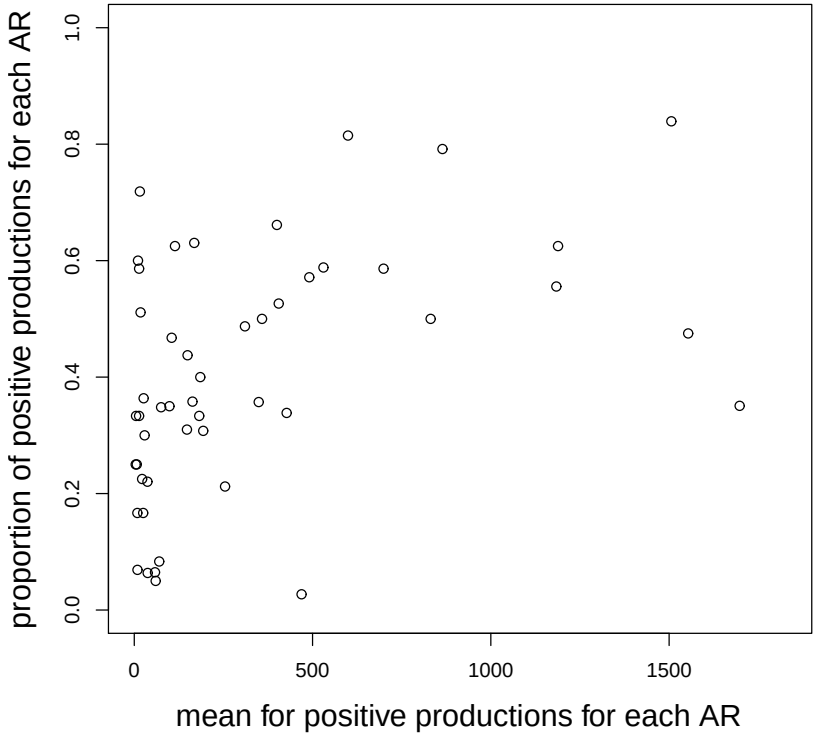

Figure : The proportion of farms with positive grape wine production *vs* the mean of positive grape wine production for each ARs

Semicontinous data

How the "excess" zeros are treated in the literature?

- In classical regression literatures,mixture models are widely used to account for "excess" zeros.
- This is realized considering a pair of regression models:
  - a logit or probit model, for the probability of nonzero responses;
  - a conditional regression model for the nonzero values.
- These models has been originally developed to analyze count data and in this context are referred as zero-inflated (ZI) models (example include ZIP, ZINB and ZIB models).
- In the context of continuous data the mixture models are referred as two-part modes and have been used essentially for the analysis of longitudinal data.

Two part models and small area estimation

- Pfefferman et al. (2008) described problem of zero-inflated data for SAE under a two part random effects model using a bayesian approach.
- Chandra and Sud (2012) consider the same framework of Pfeffermann adopting a frequentist approach.
- but both consider a **not skewed** distribution for non zero responses.

Skewed data

- When the data are skewed can be used a transformed scale *e.g.* the logaritm scale.
- The use of transformations in inference has a long history (Carroll and Ruppert, 1998, chapter 4; Chen and Chen, 1996; Kaiberg 2000).
- Under the log trasformed models, there are alternative approaches to obtain better indirect predictors for small area mean (Slud and Maiti, 2006; Chambers and Chandra, 2011)
- Zero inflated lognormal models have been applied for the analysis of longitudinal data (Holsen and Shafer 2001, Gosh and Albert 2009).

UNIVERSITÀ DEGLI STUDI FIRENZE
DiSIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

Spatially structured data

- The spatial distribution of the study variable with possible linear or non-linear covariate effects (geoadditive models, Kammann and Wand (2003))
- The area-specific effects and the spatial effects (Opsomer et al., 2008).

Basic setup, definition and assumption
of a SAE problem

- A finite population U of N units, partitioned in m subsets (areas) of size N_i, with $\sum_{i=1}^{m} N_i = N$, is considered.
- A sample r of n units is selected from U according to a non-informative sampling design.
- r may be decomposed as $r = \bigcup_{i=1}^{m} r_i$ where r_i is the area specific sample of size n_i.
- A response variable y is observed for each unit in the sample.
- y_{ij} denotes the value of y for the unit $j = 1, \ldots, N_i$ in small area $i = 1, \ldots, m$.
- Estimation of the area means $\overline{y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$: $\overline{y}_i = N_i^{-1}(\sum_{j \in r_i} y_{ij} + \sum_{j \in q_i} y_{ij})$ where q_i is the complement of the area specific sample r_i to the area population (of size $N_i - n_i$).
- It is assumed that the sample area sizes n_i are too small to calculate reliable direct estimates.
- The values of some covariates are available at area and/or unit level for $j \in r_i$ and for $j \in q_i$.
- For each unit j in small area i two vectors $\mathbf{t}_{ij}$ and $\mathbf{t}^*_{ij}$ of covariates and the spatial location $\mathbf{s}_{ij}$ ($\mathbf{s} \in R^2$) are known.

The two-part model

- In our model, the response variable y_{ij} is

$$y_{ij} = \delta_{ij} \; z_{ij}, \quad i = 1, \ldots, m; \; j = 1, \ldots, N_i,$$

  δ_{ij} is an indicator independent of the random variable $z_{ij} > 0$.
- z_{ij} has a Gamma distribution with mean μ_{ij} and coefficient of variation $1/\sqrt{\nu}$.
- the distribution function F_{ij} of y_{ij} can be written as

$$F_{ij} = \pi_{ij} \, G_{ij} + (1 - \pi_{ij}) \, F_0$$

  where $\pi_{ij} = P(\delta_{ij} = 1)$, G_{ij} and F_0 are the distribution functions of z_{ij} and δ_{ij}, respectively.
- π_{ij} and μ_{ij} of the Gamma distribution are modelled depending on some covariates.
- The two-part model is specified conditionally on two sets of covariates ($\mathbf{t}_{ij}$ and $\mathbf{t}^*_{ij}$), the geographical coordinates ($\mathbf{s}_{ij}$) and two sets of area random effects $\{u_1, \ldots, u_m\}$ and $\{u^*_1, \ldots, u^*_m\}$.

Part one

I

- The mixing proportion π_{ij} is modelled as

$$\eta_{ij} = \log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_{0t} + \mathbf{t}_{ij}^T \boldsymbol{\beta}_t + h(\mathbf{s}_{ij}) + u_i \qquad (1)$$

  where $h(\cdot)$ is some bivariate smooth function depending on geographical unit coordinates $\mathbf{s}_{ij}$.

- To estimate $h(\cdot)$, we use a penalized spline [Eilers and Marx(1996)]; [Ruppert et al. (2003)]):

$$h(\mathbf{s}_{ij}) = \beta_{0s} + \mathbf{s}_{ij}^T \ \boldsymbol{\beta}_s + \sum_{k=1}^{K} \gamma_k \ b(\mathbf{s}_{ij}, \kappa_k)$$

- We follow Ruppert for the choice of the basis, the number and location of knots. Therefore, we use a transformed *radial basis*, defined as

$$\mathbf{B} = \{b(\mathbf{s}_{ij}, \kappa_k)\} = \left\{ [C(\mathbf{s}_{ij} - \kappa_k)]_{\substack{1 \le j \le N_i, 1 \le i \le m \\ 1 \le k \le K}} \quad [C(\kappa_h - \kappa_k)]^{-1/2}_{\substack{1 \le h \le K \\ 1 \le k \le K}} \right\}$$

  where $C(\mathbf{r}) = \|\mathbf{r}\|^2 \log \|\mathbf{r}\|$.

Part one
II

- With this representation for $h(\cdot)$, the model can be written as a mixed model ([Kammann and Wand(2003)];

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u} \tag{3}$$

where: $\mathbf{X}$ is the fixed effects matrix with rows $[1, \mathbf{t}_{ij}^T, \mathbf{s}_{ij}^T]$; $\mathbf{B}$ is the $N \times K$ matrix of the thin plate spline basis functions defined in (2); $\mathbf{D}$ is the $N \times m$ area-specific random effects matrix with rows $\mathbf{d}_{ij}$ containing indicators taking value 1 if observation j is in area i and 0 otherwise; $\boldsymbol{\beta} = (\beta_{0t}, \beta_{0s}, \boldsymbol{\beta}_t^T, \boldsymbol{\beta}_s^T)^T$ is a vector of unknown coefficients; $\mathbf{u}$ is the vector of the m area specific random effects; $\boldsymbol{\gamma}$ is the vector of the K thin plate spline coefficients, treated as random effects.

Part two

- The density function of z_{ij} is

$$f(z_{ij}) = \frac{(\nu/\mu_{ij})(\nu z_{ij}/\mu_{ij})^{\nu-1}\exp(-\nu z_{ij}/\mu_{ij})}{\Gamma(\nu)}, \quad z_{ij} > 0,$$

  where $\nu > 0$ and $\mu_{ij} > 0$. Here, μ_{ij} is the mean and $1/\sqrt{\nu}$ the coefficient of variation of z_{ij}.
- The mean μ_{ij} is modelled through a log-link function as

$$\log \mu_{ij} = \beta^*_{0t} + \mathbf{t}_{ij}^{T*}\beta^*_t + h^*(\mathbf{s}_{ij}) + u^*_i. \tag{4}$$

- Representing $h^*(\cdot)$ with a low rank thin plate spline with K knots, as we did for $h(\cdot)$, the model (4) becomes

$$\log \boldsymbol{\mu} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{B}^*\boldsymbol{\gamma}^* + \mathbf{D}^*\mathbf{u}^* \tag{5}$$

  where all terms $(\mathbf{X}^*, \beta^*, \mathbf{B}^*, \gamma^*, \mathbf{D}^*, \mathbf{u}^*)$ have the same meaning as those indicated by the same symbol without an asterisk in model (3).

Random effects distributions

Assumptions on the distribution

- The area effects and the spline random effects are jointly normal and correlated,

$$(u_i, u_i^*)^T \sim N\left(\mathbf{0}, \Sigma_u = \begin{bmatrix} \sigma_u^2 & \sigma_{uu^*} \\ \sigma_{uu^*} & \sigma_{u^*}^2 \end{bmatrix}\right) \quad \text{and}$$

$$(\gamma_k, \gamma_k^*)^T \sim N\left(\mathbf{0}, \Sigma_\gamma = \begin{bmatrix} \sigma_\gamma^2 & \sigma_{\gamma\gamma^*} \\ \sigma_{\gamma\gamma^*} & \sigma_{\gamma^*}^2 \end{bmatrix}\right).$$

  This assumption defines a 'full two-part model'.

- The random effects in the two parts of the model are independent or the two parts of the model are fitted separately:

$$(u_i, u_i^*)^T \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_{u^*}^2 \end{bmatrix}\right) \quad \text{and} \quad (\gamma_k, \gamma_k^*)^T \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_\gamma^2 & 0 \\ 0 & \sigma_{\gamma^*}^2 \end{bmatrix}\right).$$

  This assumption defines a 'separate two-part model'.

Bayesian formulation of the model

I

- We assume noninformative priors for the parameters of the whole model. Each element of β and β^* is given a normal distribution with zero mean and large variance (i.e., 1.0e+8).
- The shape parameter ν of the Gamma distribution is the squared reciprocal of the standard deviation-like parameter. Following [Marley and Wand(2010)], $\nu^{-1/2}$ is given a half-Cauchy distribution (with scale parameter 25).
- Under the assumption of correlated random effects u_i and u_i^* and/or γ_k and γ_k^* between the two parts of the model, the inverse variance-covariance matrices Σ_u^{-1} and Σ_γ^{-1} are given a Wishart distribution with scale matrix diag(0.001,0.001) and 2 degrees of freedom.
- When the random effects between the two parts of the model are assumed to be uncorrelated, a half-Cauchy distribution (with parameter 25) is given to each standard deviation parameter σ_u, σ_u^*, σ_γ and σ_γ^*.

UNIVERSITÀ DEGLI STUDI FIRENZE
DiSIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

Bayesian formulation of the model
II

- We use the half-Cauchy distribution to achieve non-informativeness for variance parameters ([Gelman(2006)] and [Polson and Scott(2012)]).
- Following [Marley and Wand(2010)], we can sampling from a half-Cauchy distribution defined as the distribution of U/V, where U and V are independent with $U \sim N(0, \sigma_1^2)$ and $V \sim N(0, \sigma_2^2)$.

Model selection

- We have defined a class of models: including or not area and/or spline random effects.
- To select a suitable model the Deviance Information Criterion (DIC) (see [Spiegelhalter et al. (2002)]) is used:
  $DIC = \bar{D} + pD$, where $\bar{D}$ is the posterior expectation of the deviance and pD represents the 'effective number' of parameters and reflects the complexity of the model.

Small area mean predictions

- For each MCMC algorithm iteration $l = 1, \ldots, L$, the empirical predictive distribution is

$$\hat{y}_{ij}^{(l)} = \hat{\pi}_{ij}^{(l)} \hat{z}_{ij}^{(l)}$$

for $i = 1, \ldots, m$ and $j \in q_i$

$$\hat{\pi}_{ij}^{(l)} = \frac{\exp(\mathbf{x}_{ij}^T \hat{\beta}^{(l)} + \mathbf{b}_{ij}^T \hat{\gamma}^{(l)} + \hat{u}_i^{(l)})}{1 + \exp(\mathbf{x}_{ij}^T \hat{\beta}^{(l)} + \mathbf{b}_{ij}^T \hat{\gamma}^{(l)} + \hat{u}_i^{(l)})} \quad \text{and} \quad \hat{z}_{ij}^{(l)} = \exp\left(\mathbf{x}_{ij}^{*T} \hat{\beta}^{*(l)} + \mathbf{b}_{ij}^{*T} \hat{\gamma}^{*(l)} + \hat{u}_i^{*(l)}\right)$$

where $\mathbf{b}_{ij}$ and $\mathbf{b}_{ij}^*$ represent, respectively, the ij-row of the matrix $\mathbf{B}$ and $\mathbf{B}^*$ of the thin plate spline basis.

Small area mean predictions
II

- The empirical predictive distribution for the mean of small area i is

$$\hat{\bar{y}}_i^{(l)} = N_i^{-1} \left(\sum_{j \in r_i} y_{ij} + \sum_{j \in q_i} \hat{y}_{ij}^{(l)} \right).$$

- As a measure of precision, to each estimate is associated the corresponding credibility interval, that is, the interval between the $\alpha/2$ and $1 - \alpha/2$ quantiles of the empirical predictive distribution.

Grape wine production analysis
Data description

- The goal is to estimate the per-farm average grape wine production in Tuscany at AR level.
- y_{ij} is the grape wine production of farm j in area i and the mixing proportion π_{ij} should be viewed as the probability that farm j in area i has a strictly positive production.
- For the logistic model, two auxiliary variables are considered: the surface allocated to grape wines in logarithmic scale and a dummy variable that indicates the selling of grape wine related products, both at census2000.
- The same two variables are included in the log-linear model for the Gamma distribution, together with the number of days worked by farm family members in census2000.
- Part one and part two of the model include individual level covariates, random effects representing the spatial structure of the data, and area random effects: model denoted by SA-SA

Grape wine production analysis

The four compared models

Table : DIC and pD values for the 'plausible' full two-part models: S stands for "spline" random effects and A for "area" random effects

Model	first part	second part	DIC	pD
SA-SA	$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u}$	$\log \boldsymbol{\mu} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{B}^*\boldsymbol{\gamma}^* + \mathbf{D}^*\mathbf{u}^*$	12791.2	125.60
SA-A	$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u}$	$\log \boldsymbol{\mu} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{D}^*\mathbf{u}^*$	12904.3	87.16
S-SA	$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}\boldsymbol{\gamma}$	$\log \boldsymbol{\mu} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{B}^*\boldsymbol{\gamma}^* + \mathbf{D}^*\mathbf{u}^*$	12812.4	121.20
A-SA	$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{u}$	$\log \boldsymbol{\mu} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{B}^*\boldsymbol{\gamma}^* + \mathbf{D}^*\mathbf{u}^*$	12784.8	118.40

Grape wine production analysis

The coefficient estimates

Table : Results from full (A-SA) and separate (A⊥SA) two-part models: coefficient estimates with their 95% credibility intervals (CI95%)

		full two-part model A-SA		separate two-part model A⊥SA	
	parameter	estimate	CI95%	estimate	CI%95
first	constant	-1.405	-1.719 - -1.119	-1.373	-1.677 - -1.081
	log surface allocated to grape wines	1.912	1.548 - 2.279	1.901	1.533 - 2.287
	selling of grape wines products	1.099	0.726 - 1.463	1.099	0.735 - 1.458
	$\hat{\sigma}_u$	0.873	0.664 - 1.132	0.878	0.666 - 1.149
second	constant	0.022	-0.007 - 0.051	0.447	0.347 - 0.539
	x coordinate	-0.054	-0.079 - -0.010	0.158	0.045 - 0.281
	y coordinate	0.059	0.049 - 0.069	0.053	0.027 - 0.077
	log surface allocated to grape wines	1.083	1.034 - 1.144	1.230	1.166 - 1.289
	selling of grape wines products	0.787	0.733 - 0.829	0.142	0.050 - 0.261
	number of days worked	0.0004	0.0002 - 0.0006	0.0004	0.0002 - 0.0006
	$\hat{\nu}$	1.528	1.402 - 1.683	1.499	1.378 - 1.631
	$\hat{\sigma}^*_u$	0.401	0.291 - 0.540	0.292	0.175 - 0.428
	$\hat{\sigma}^*_\gamma$	2.395	1.883 - 2.960	2.091	1.687 - 2.591
	$\hat{\sigma}_{uu*}$	-0.0086	-0.183 - 0.163		

Grape wine production analysis

The estimates of the parameters for A⊥SA model

Larger credibility intervals correspond to higher production values, possibly because the model assumes a constant coefficient of variation.

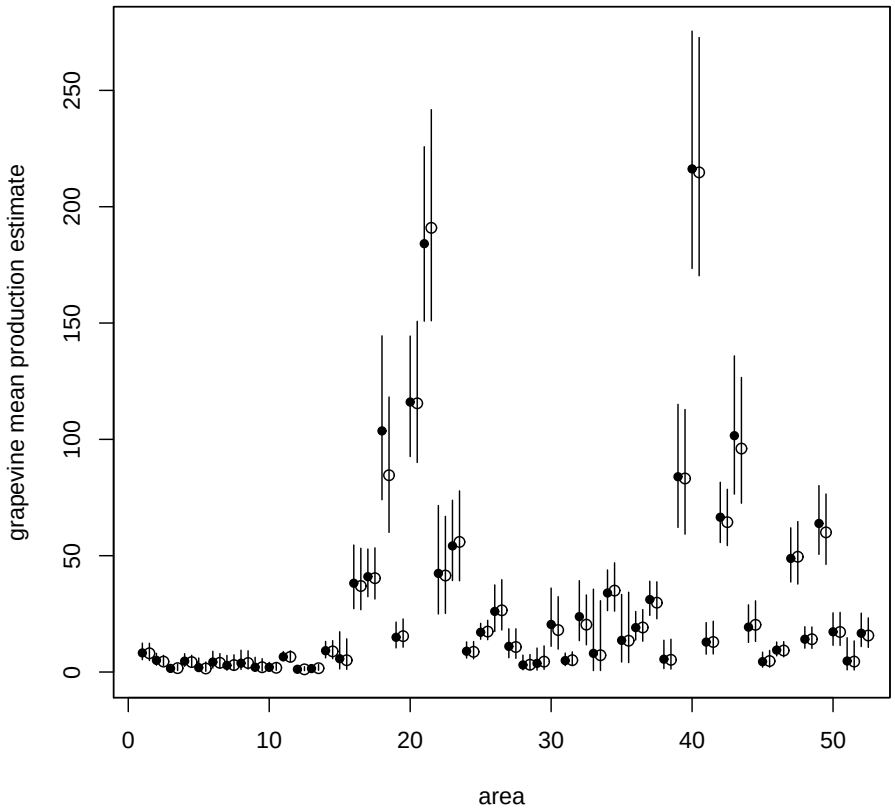

Figure : Estimates of the means of the grape wine production for each agrarian region and their 95% credibility intervals from full (• for A-SA) and separate (○ for A⊥SA) two-part models

Grape wine production analysis

The estimates of the mean production from model A-SA

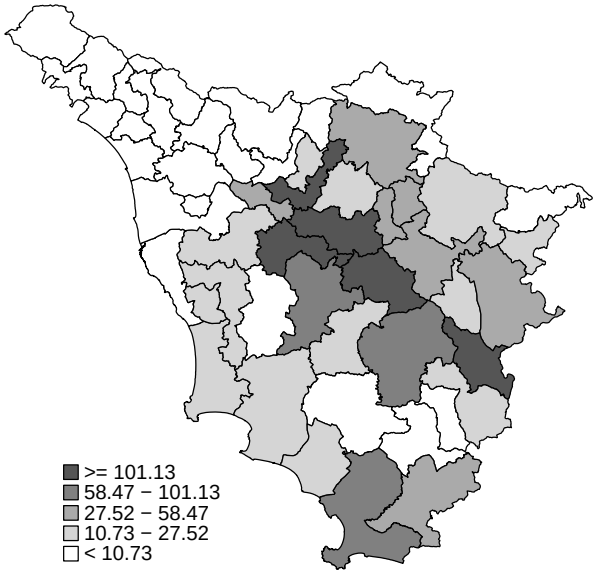

Figure : Estimates from the suggested model plotted against the corresponding direct estimates

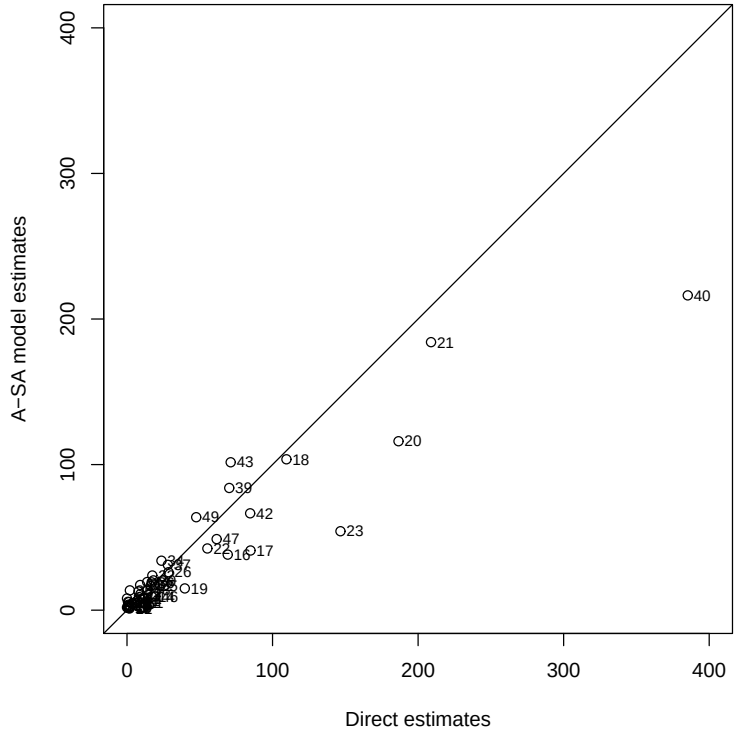

Figure : Estimates from the suggested model plotted against the correspondent direct estimates

Grape wine production analysis

The mean of the posterior distributions for splines and random effects

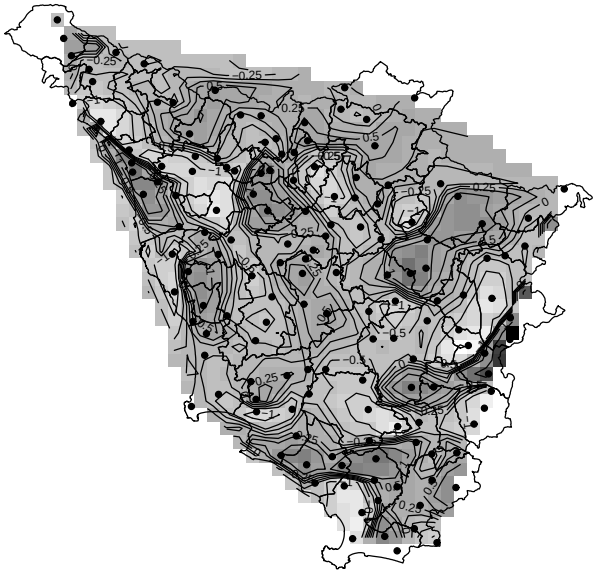

Figure : Mean values of the posterior distributions for spline random effects γ_k^*

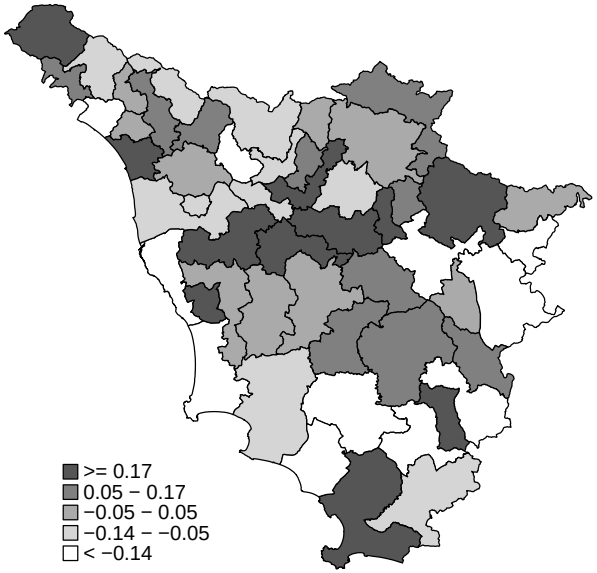

Figure : Mean values of the posterior distributions for area random effects u_i^*

It should be noted that bivariate splines overlap the borders of the municipalities included in the AR.

Conclusions

- Taking into account the presence of zeros in the data is crucial.
- Another inefficient approach to investigate zero inflated data consists in taking only non zero data into account.
- It is fundamental to consider the highly skewed distribution of the positive responses.
- For target variable which shows a spatial trend, the use of geographical information allows more accurate SAE.

UNIVERSITÀ DEGLI STUDI FIRENZE
DiSIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

Ongoing research

- An accurate evaluation of the conditions about the choice of the model.
- A frequentist perspective could be developed.
- Consider informative design.
- Other parameters (total, median, etc.).
- Other areas.
- Other spatial covariate (satellite imagines, etc).

Main References

Chandra, H., and Sud, U.C. (2012). Small Area Estimation for Zero-Inflated Data. *Communications in Statistics - Simulation and Computation* **41**, 632–643.

Eilers, P.H.C., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89-121.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515–533.

Gelman, A., and Rubin, D.R. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457–511.

Kammann, E.E., and Wand, M.P. (2003). Geoadditive Models. *Applied Statistics* **52**, 1–18.

Kass, R.E., Raftery, A.E. (1995), Bayes Factors. *Journal of the America Statistical Association* **90**, 773–795.

Kaufman, L., and Rousseeuw, P.J. (1990). *Finding Groups in Data: An introduction to cluster Analysis*, Wiley, New York.

Marley, J.K., and Wand, M.P. (2010) Non-Standard Semiparametric Regression via BRugs. *Journal of Statistical Software* **37**, 1–30.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., and Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B* **70**, 265–286.

Pfeffermann, D., Terryn, B., and Moura, F.A.S. (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Survey Methodology* **34**, 235–249.

Polson, N.G., and Scott, J.G. (2012). On the half-Cauchy Prior for a Global Scale Parameter. *Bayesian Analysis* **7**, 887–902.

Ruppert, D., and Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression*.Cambridge University Press, Cambridge.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde,A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society. Series B* **64**, 583–639.

UNIVERSITÀ DEGLI STUDI FIRENZE
DiSIA
DIPARTIMENTO DI STATISTICA, INFORMATICA, APPLICAZIONI "GIUSEPPE PARENTI"

Simulation Experiment

- A model-based simulation experiment has been performed to evaluate the proposed SAE approach. More precisely, the results obtained from the full (A-SA) and from the separate (A⫫SA) two-part models are compared with those obtained from the full two-part model suggested in Pfeffermann *et al.*(2008).
- This latter model, referred to as 'linear A-SA' in the sequel, is a two-parts model similar to the proposed one, but considers a linear model on the second part under the assumption of a normal distribution for z_{ij}. We decided not to compare our method with others that do not take the excess of zeros into account and/or are not suitable for a Bayesian approach.
- To carry out the simulation experiment, 200 populations have been generated, using the estimated A-SA model on the real application.
- The mean production for each AR in the simulated populations has been taken to be the 'true' mean production, thus allowing assessment of the performances of various models. Such performances have been evaluated through the following criteria: the relative bias (RB), the absolute relative bias (ABSRB) and the relative root mean squared error (RRMSE).

Simulation Experiment
Results

Area-specific values of RB and RRMSE under the compared models

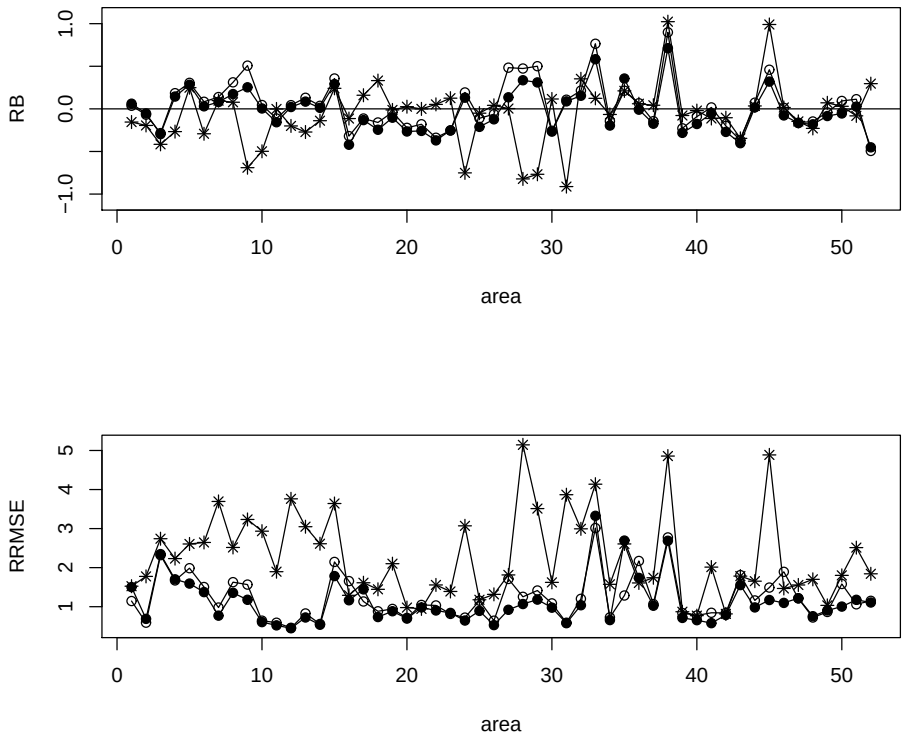

Figure : Estimates from the suggested model plotted against the corresponding direct estimates

Simulation Experiment
Results

- The averages of $RB_i\%$, $ABSRB_i\%$ and $RRMSE_i\%$ over areas i define AvRB%, AvABSRB% and AvRRMSE% respectively.
- For linear A-SA model, we have obtained -5.71, 24.07 and 226.80 values; for the A-SA model, respectively -2.16, 19.89 and 112.37; for the A⫫SA model 4.45, 21.93 and 122.85.
- The AvABSRB% and AvRB% criteria confirm our expectations: the Gamma model is to be preferred to the normal one.
- The full model seems to present some advantages with respect to the separate one.
- The AvRRMSE% criterion confirms the claims present in the literature about the inefficiency of using normal distribution for heavy-tailed data.