The tipsae package: mapping indicators through space and time via small area estimation

Silvia De Nicolò Aldo Gardini

Department of Statistical Sciences, University of Bologna

email: `silvia.denicolo@unibo.it`

StanConnect, 31st of October 2022

Small Area Estimation

We want to estimate a generic indicator on the unit-interval
(e.g. poverty, health insurance coverage rates):

- in a specific sub-population (domains or areas)
  (e.g. districts, counties, sex-age-race groups)
- from survey data
- in domains not originally planned by the survey design
- resulting in small-sized sample

Survey estimation is unreliable due to the high variability.

Small Area Estimation

$\rightarrow$ We have to resort Small Area Estimation (SAE) techniques.

Area-level model class:

- hierarchical Bayes models with survey estimators as responses
- exploit domain-specific quantities as auxiliary information
- borrow strength across areas, producing estimates with a decreased and acceptable level of uncertainty.

Space-time models in SAE

- When historical data are available, it is also possible to borrow strength from time.
- Often domains are geographical regions: spatial correlation may be observed.
- Recent Bayes spatio-temporal SAE models have been used to measure:
  - relative risk of disease (Choi et al. 2011);
  - gender-based violence (Vicente et al. 2020);
  - patterns of crime (Law et al. 2014).

SAE and R packages

Area-level SAE for unit interval-defined measures:

- Gaussian Model (Fay-Herriot) with suitable transformations;
  (Esteban et al., 2012, 2020; Marhuenda et al., 2013, 2014)

- Mixed Beta-based models.

R implementations: Gaussian model with **arcsin transformation** in `emdi` package without any spatio-temporal specification.

What we did

The tipsae goals

Providing a friendly framework to deal with unit interval indicators:

- Bayesian inferential framework.
- Focus on Beta, inflated Beta and other Beta mixtures models.
- Possible dependence structures in the data: spatial and/or temporal random effects.
- Estimation via Hamiltonian MC (`Stan`) and customized parallel computing imported from `rstan`.
- Ad-hoc functions for small-area model diagnostics with plots and maps tools.
- A friendly Shiny app.

Some notation

We consider a finite population with size N partitioned into D domains. For each domain d, we define:

- The quantities of interest θ_d, with $\theta_d \in (0;1)$;
- The survey (or crude) estimator y_d of θ_d with large variance;
- An estimate of the sampling variance $\widehat{\mathbb{V}}[y_d]$;
- A vector $\boldsymbol{x}_d$ of auxiliary variables (recorded without error).

Standard Beta Small Area Model

When $y_d \in (0;1)$, the **sampling level** is

$$y_d|\theta_d \sim \text{Beta}\left(\theta_d\phi_d, (1-\theta_d)\phi_d\right) \quad \forall d,$$

where

- $\theta_d = \mathbb{E}[y_d|\theta_d]$,
- ϕ_d is a **dispersion parameter** usually assumed to be **known**.

The **linking level** models the target quantity

$$\text{logit}(\theta_d)|\beta, e_d = \boldsymbol{x}_d^T\beta + e_d.$$

- e_d is the **random effects** term which can incorporate several data dependency structures.

Standard Beta Small Area Model

When $y_d \in (0;1)$, the **sampling level** is

$$y_d|\theta_d \sim \text{Beta}\left(\theta_d\phi_d, (1-\theta_d)\phi_d\right) \quad \forall d,$$

where

- $\theta_d = \mathbb{E}[y_d|\theta_d]$,
- ϕ_d is a **dispersion parameter** usually assumed to be **known**.

The **linking level** models the target quantity

$$\text{logit}(\theta_d)|\boldsymbol{\beta}, e_d = \boldsymbol{x}_d^T\boldsymbol{\beta} + e_d.$$

- e_d is the **random effects** term which can incorporate several data dependency structures.

Spatial structure

- **Spatial structure:** $e_d = s_d + v_d$
  - v_d an unstructured area random effect;
  - s_d a spatial random effect;

where $\boldsymbol{s} = (s_1, \ldots, s_D)$ has prior

$$\boldsymbol{s}|\sigma_s \sim ICAR(\sigma_s^2 \tilde{\mathbf{K}}^-), \quad \sigma_s \sim \text{half-}\mathcal{N}\left(0, 2.5\right).$$

- $\tilde{\mathbf{K}}^-$ is the inverse of a singular precision matrix $\tilde{\mathbf{K}}$
- obtained from $\mathbf{K} = \mathbf{D} - \mathbf{W}$ with
  - $\mathbf{D}$ a diagonal matrix with the number of connections,
  - $\mathbf{W}$ the adjacency matrix.

Spatial structure

$\tilde{\mathbf{K}}$ results from $\mathbf{K}$ after a:

1. scaling procedure (Sørbye and Rue 2014):

The structure of $\mathbf{K}$ affects prior variability irrespective of the hyperprior set up on dispersion parameter.

2. contemplating the presence of **disconnected graphs in the model** e.g., islands (Freni-Sterrantino et al. 2018; Morris et al. 2019).
   - Independent scaling for **sub-blocks of $\mathbf{K}$ related to components with size > 1**.
   - Every components has its own intercept.
   - Constant prior replaced with standard Gaussian for **components of size 1** (singletons).

Temporal or spatio-temporal

- **Temporal effect:** $e_{dt} = u_{dt} + v_d$,
  each area d is repeatedly observed at times $t = 1, \ldots, T$.
  - v_d an unstructured area random effect;
  - u_{dt} a temporal random effect;

$$u_{dt} | u_{d,t-1}, \sigma_u \overset{ind}{\sim} \mathcal{N}\left(u_{d,t-1}, \sigma_u\right), \quad \sigma_u \sim \text{half-}\mathcal{N}\left(0, 2.5\right).$$

- **Spatio-Temporal effect:** $e_{dt} = u_{dt} + s_d$.

Marginal spatial and temporal components are **non-identifiable**!
SAE models have only predictive purposes.

Temporal or spatio-temporal

- **Temporal effect:** $e_{dt} = u_{dt} + v_d$,
  each area d is repeatedly observed at times $t = 1, \ldots, T$.
  - v_d an unstructured area random effect;
  - u_{dt} a temporal random effect;

$$u_{dt} | u_{d,t-1}, \sigma_u \overset{ind}{\sim} \mathcal{N}\left(u_{d,t-1}, \sigma_u\right), \quad \sigma_u \sim \text{half-}\mathcal{N}\left(0, 2.5\right).$$

- **Spatio-Temporal effect:** $e_{dt} = u_{dt} + s_d$.

Marginal spatial and temporal components are **non-identifiable**!
SAE models have only predictive purposes.

Alternative Likelihoods

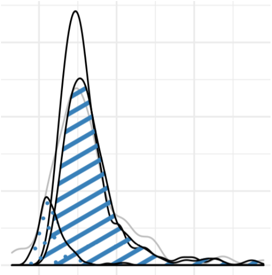
Let $f_B(y;\mu,\phi)$ be the Beta p.d.f., the following extensions are provided:

- **2-components Beta mixture:** for $y_d \in (0,1)$(De Nicolò et al. 2022).

$$y_d|\ldots \overset{ind}{\sim} p \cdot f_B(y_d;\lambda_{1d},\phi_d) + (1-p)\cdot f_B(y_d;\lambda_{2d},\phi_d).$$

- **Zero/one inflated Beta:** for $y_d \in [0,1]$

$$y_d|\ldots \overset{ind}{\sim} p_d^{(0,1)} f_B(y_d;\mu_d,\phi_d)\mathbb{1}\{0<Y_d<1\} + p_d^z \cdot \mathbb{1}\{Y_d=0\} + p_d^o \cdot \mathbb{1}\{Y_d=1\}.$$

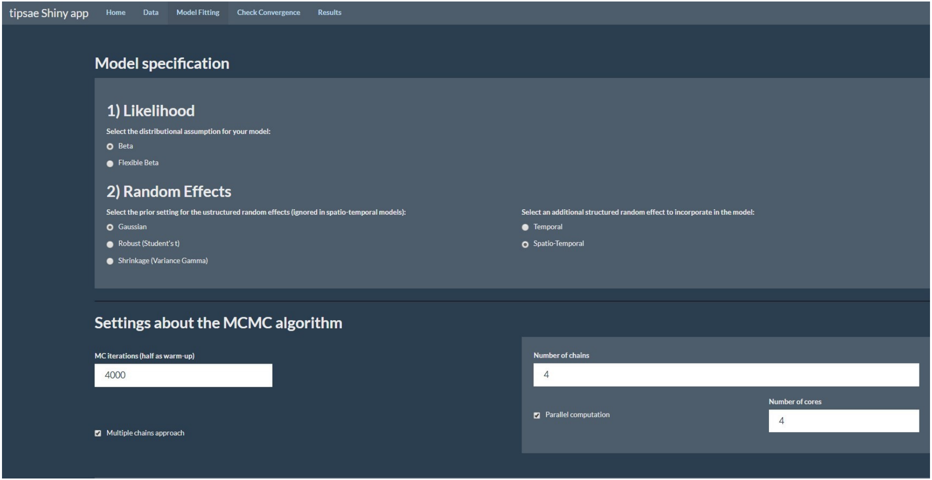


A Case Study

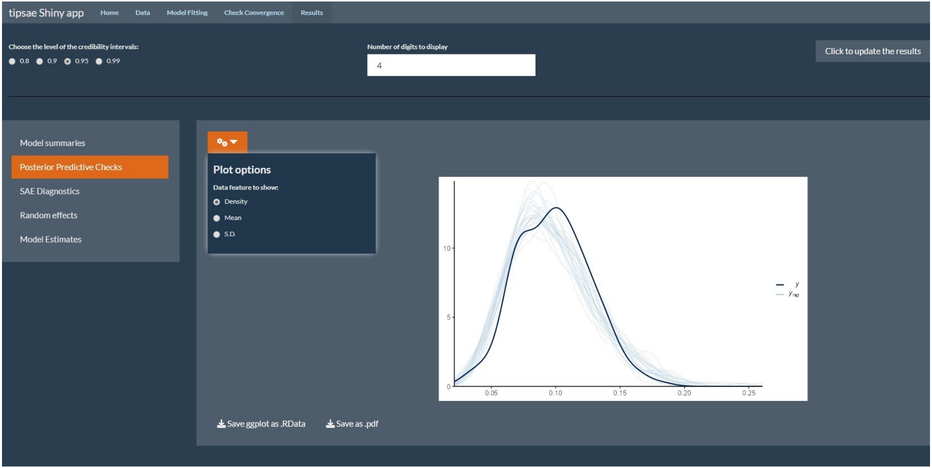
A Case Study

- **Poverty rates** in 38 health districts within the Emilia-Romagna region.
- Starting from unreliable survey estimates of the **Head-Count Ratio indicator**,
- recorded annually from 2014 to 2018,
- and generated covariates.

A Case Study

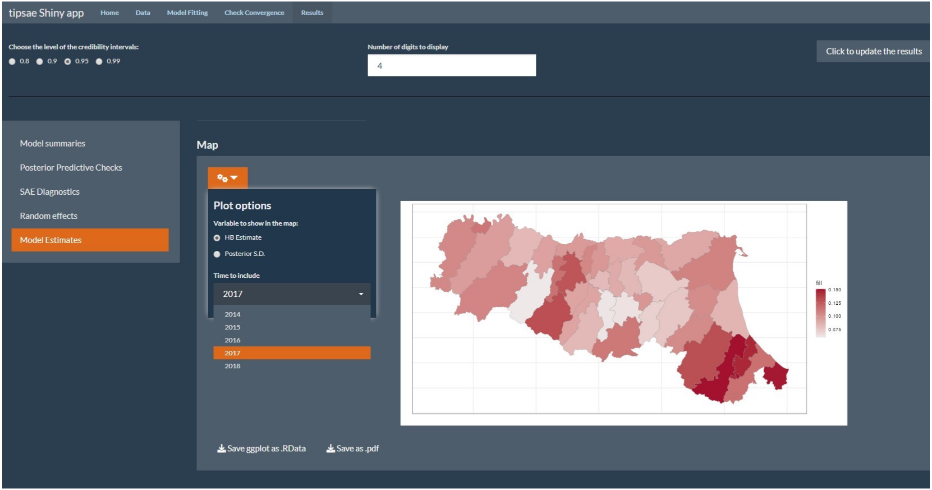

tipsae Shiny app
Home
Data
Model Fitting
Check Convergence
Results
Model specification
1) Likelihood
Select the distributional assumption for your model:
Beta
Flexible Beta
2) Random Effects
Select the prior setting for the ustructured random effects (ignored in spatio-temporal models):
Gaussian
Robust (Student's t)
Shrinkage (Variance Gamma)
Select an additional structured random effect to incorporate in the model:
Temporal
Spatio-Temporal
Settings about the MCMC algorithm
MC iterations (half as warm-up)
4000
Multiple chains approach
Number of chains
4
Parallel computation
Number of cores
4

A Case Study



tipsae Shiny app
Home
Data
Model Fitting
Check Convergence
Results
Choose the level of the credibility intervals:
0.8
0.9
0.95
0.99
Number of digits to display
4
Click to update the results
Model summaries
Posterior Predictive Checks
SAE Diagnostics
Random effects
Model Estimates
Plot options
Data feature to show:
Density
Mean
S.D.
Save ggplot as .RData
Save as .pdf

A Case Study



tipsae Shiny app
Home
Data
Model Fitting
Check Convergence
Results
Choose the level of the credibility intervals:
0.8
0.9
0.95
0.99
Number of digits to display
4
Click to update the results
Model summaries
Posterior Predictive Checks
SAE Diagnostics
Random effects
Model Estimates
Map
Plot options
Variable to show in the map:
HB Estimate
Posterior S.D.
Time to include
2017
2014
2015
2016
2017
2018
fill
0.150
0.125
0.100
0.075
Save ggplot as .RData
Save as .pdf

Conclusions

- R package available on CRAN
- fitting Beta-based small area models with spatio-temporal dependency structures.

Additional tools under developments:

- Regularizing prior for regression coefficients (**HorseShoe**).
- Alternative 0/1 inflated Beta models that incorporate **survey information**.

References

Choi, Jungsoon et al. (2011). “Evaluation of Bayesian spatiotemporal latent models in small area health data”. In: *Environmetrics* 22.8, pp. 1008–1022.

De Nicolò, Silvia, Maria Rosaria Ferrante, and Silvia Pacei (2022). “Small Area Estimation of Inequality Measures using Mixtures of Betas”. In: *arXiv preprint arXiv:2209.01985*.

Freni-Sterrantino, Anna, Massimo Ventrucci, and Håvard Rue (2018). “A note on intrinsic conditional autoregressive models for disconnected graphs”. In: *Spatial and spatio-temporal epidemiology* 26, pp. 25–34.

Law, Jane, Matthew Quick, and Ping Chan (2014). “Bayesian spatio-temporal modeling for analysing local patterns of crime over time at the small-area level”. In: *Journal of quantitative criminology* 30.1, pp. 57–78.

Morris, Mitzi et al. (2019). “Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan”. In: *Spatial and spatio-temporal epidemiology* 31, p. 100301.

Sørbye, Sigrunn Holbek and Håvard Rue (2014). “Scaling intrinsic Gaussian Markov random field priors in spatial modelling”. In: *Spatial Statistics* 8, pp. 39–51.

Vicente, G, T Goicoa, and MD Ugarte (2020). “Bayesian inference in multivariate spatio-temporal areal models using INLA: analysis of gender-based violence in small areas”. In: *Stochastic Environmental Research and Risk Assessment* 34.10, pp. 1421–1440.