# Official Statistics: Bayesian Small Area Estimation of Origin-Destination Matrix

Loredana Di Consiglio  Fabrizio Solari  Tiziana Pichiorri
Emanuela Scavalli  Massimo Armenise  Lorenzo Asti
Carolina Ciccaglioni  Isabella Corazziari  Luca Faustini

# Origin-Destination Matrix

- Origin-Destination Matrix:

$$\mathcal{M} = \{M_{ij}\}_{i,j=1,\dots,D}$$

- $M_{ij}$ = commuting flow between origin location $i$ and destination location $j$
- $D$ = number of locations

Problem:

- $\#(M_{ij} > 0) \ll D^2 \implies$ the origin-destination matrix is sparse

# Small Area Estimation Problem

## Small Areas

- The small areas are given by the pairs $(i, j)$ (origin province–destination province)

## Target Parameter

for all $(i, j) \in D \times D$:

- work commuting rate $m_{ij} = M_{ij}/N_i$, $N_i$ employed counts for province $i$
- $0 \leq m_{ij} \leq 1$

Definition: the definition of commuter adopted in Italy is:

*any worker traveling daily back and forth from home and his workplace at least three times per week*

# Fay-Herriot Model (Fay & Herriot, 1979)

- sampling model:

$$\widehat{m}_{ij} = m_{ij} + e_{ij}$$

  - $\widehat{m}_{ij}$ = direct estimate of $m_{ij}$
  - $e_{ij} \overset{\text{ind}}{\sim} N(0, V_{ij})$, $V_{ij}$ supposed to be known

- linking model:

$$m_{ij} = x_{ij}^T \beta + u_{ij}$$

  - hp standard: $u_{ij} \mid \sigma_u^2 \overset{\text{ind}}{\sim} N(0, \sigma_u^2)$

# Comments

- The O-D matrix is a sparse matrix $\Longrightarrow$
- Fixed effects and random effects must be defined in a way that avoids inflating commuting flows corresponding to cells that are empty or have very low commuting flows.
- fixed effects: $x_{ij} = (z_{ij}, s_i, t_j)$:
  $z_{ij}$: variables referring to the origin-destination pair $(i, j)$
  $s_i$: variables referring to the origin $i$
  $t_j$: variables referring to the destination $j$

  Attention must be paid to the impact of variables of type $s$ and $t$ on the final estimates
- random effects: It is reasonable to expect small or null values of $u_{ij}$ for some small areas and higher values for others $\Longrightarrow$ the standard assumption of constant variance $\sigma_u^2$ for $u_{ij}$ might not be the optimal solution

# Fay-Herriot Model Generalizations: Random Effects

- spike-and-slab (Datta & Mandal, 2015):

$$u_{ij} \mid \sigma_u^2, \delta_{ij} \overset{\text{ind}}{\sim} N(0, \delta_{ij}\,\sigma_u^2), \quad \delta_{ij} = 0, 1$$

- The random effect $u_{ij}$ is included or excluded based on the value of the indicator variable $\delta_{ij}$ (on/off switch)

- global-local (Tang, Ghosh, Ha & Sedransk, 2018):

$$u_{ij} \mid \sigma_u^2, \lambda_{ij}^2 \overset{\text{ind}}{\sim} N(0, \lambda_{ij}^2\,\sigma_u^2), \quad \lambda_{ij}^2 > 0$$

- The intensity of the random effect $u_{ij}$ is adjusted based on the value of the parameter $\lambda_{ij}^2$ (dimmer switch)

# Weakly Informative Priors

Fay Herriot, Spike-and-Slab, Global-Local Models:

- $\pi(\beta) \propto 1$
- $\pi(\sigma_u^2) \sim IG(a, b)$
  - Different values for a and have been used: $a = b = 10^{-3}$, $a = b = 10^{-6}$, $a = b = 10^{-9}$ and we did not find any relevant sensitivity to hyperparameters choice
  - Alternatively Half-Cauchy or Half-Student T distributions can be used (Gelman, 2006)

Spike-and-Slab Model: $p = \text{Prob}(\delta_{ij} = 1)$

- $\pi(p) \sim Beta(c, d)$
  - $c = 1$, $d = 4$ as in Datta & Mandal (2015)
  - $c = 1$, $d = 1$ uniform in (0,1)
  - $c = 4$, $d = 1$ symmetric to Datta & Mandal (2015)

# Global-Local Model: Local Priors

| Name | $\pi_{\lambda_{ij}^2}(x)$ |
|---|---|
| Horseshoe | $x^{-1/2}(1+x)^{-1}$ |
| Strawderman-Berger | $(1+x)^{-3/2}$ |
| Normal Exponential Gamma | $(1+x)^{-1-b}$ |
| Laplace | $\exp(-x)$ |
| Normal Gamma | $x^{a-1}\exp(-x)$ |

- The first column gives the names of the priors of $u_{ij}$ marginalized over $\lambda_{ij}^2$
- The second column reports the corresponding expression for each prior
- LA is a special case of NG setting $a = 1$
- HS, SB and SB priors are special case of the Three Parameter Beta Normal distribution

# Auxiliary Variables

- 2011 O-D flows
- 2021 O-D trajectories from administrative data
- Distances (km or travel times) between origin and destination

## Observations

- Administrative trajectories are potential commuting journeys between origin provinces and destination provinces
- The ability of the administrative trajectories to detect true commuters is a function of the distance between origin and destination provinces (*attrition*)
- Administrative trajectories can be modeled using a decay function, such as an inverse distance weighting, i.e. $x_{ij}/dist_{ij}^{q}$, where $q$ controls the rate of decay ($q = 0, 1, 2$ has been used)
- Alternatively, spline functions can be used to model $x_{ij} = /dist_{ij}^{q}$ in order to adjust for not optimal choices of $q$.

# Fixed Effects

Models description:

| model | intercept | auxiliary variables |
|---|---|---|
| $M_1$ | yes | 2011 commuting rate + 2021 commuting trajectory rate |
| $M_2$ | yes | 2011 commuting rate + 2021 commuting trajectory rate / dist |
| $M_3$ | yes | 2011 commuting rate + 2021 commuting trajectory rate / $dist^2$ |
| $M_4$ | no | 2011 commuting rate + 2021 commuting trajectory rate |
| $M_5$ | no | 2011 commuting rate + 2021 commuting trajectory rate / dist |
| $M_6$ | no | 2011 commuting rate + 2021 commuting trajectory rate / $dist^2$ |

- target parameter: work commuting flows from provinces in South of Italy (38 provinces) to all the Italian provinces
- 323 in-sample small areas, 3,743 out-of-sample areas

# Model Diagnostics

Model Fitting Diagnostics:

- Deviance Information Criterion (DIC)
- Watanabe Akaike Information Criterion (WAIC)

Model Bias Diagnostics:

- Bayesian p-value (You & Rao, 2002; Fabrizi *et al.*, 2011)
  - It evaluates the probability of the posterior means to be larger than the direct estimates. In absence of systematic bias, the expected value of the Bayesian p-value is 0.5
- Regression line fitting model-based estimates versus direct estimates (Brown *et al., 2001*)
  - The regression line should be close the 0-1 line

# MCMC

- All the posteriors are proper (Datta & Mandal, 2015; Tang *et al.*, 2018)
- It is possible to sample from all the conditional posterior distributions $\Longrightarrow$ Gibbs Sampling

- number of chains = 4
- chain length = 100,000
- burn-in = 50,000
- thin = 10

# Results

Model fitting was divided in two separate process:

- intra-provincial flows
- inter-provincial flows

Intra-provincial flows estimation is dominated by the direct estimates

$\Longrightarrow$ focus on the estimation of inter-provincial flows

- The models with an intercept term slightly outperform the corresponding models without the intercept term in terms of DIC, WAIC and Bp
- We prefer to adopt a model without an intercept term because an intercept term would assign a non-zero commuting mass to all pairs of provinces, even for:
  - provinces very far from each other
  - provinces for which both the 2011 commuting rates and the 2021 administrative trajectories rates are 0

# Results

$M_5$: 2011 commuting rate + 2021 commuting trajectory rate / dist (no intercept)

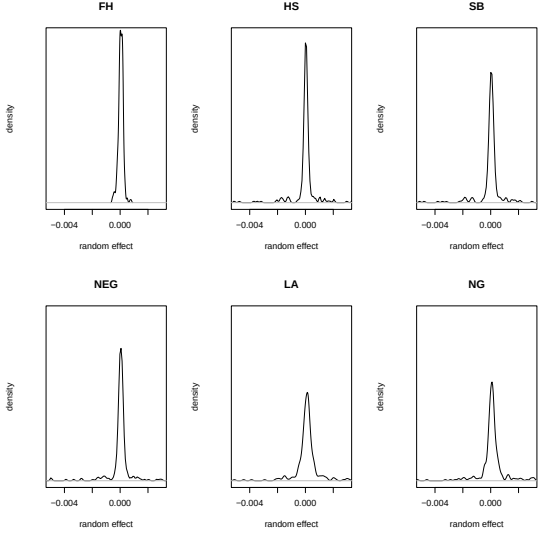
**Model $M_5$ diagnostics**.

| model | DIC | WAIC | Bp | a | b |
|---|---|---|---|---|---|
| SYNTH | -2764.37 | -5137.62 | 0.3360 | 5.2E-05 | 0.9947 |
| FH | -2872.55 | -5544.98 | 0.4252 | 2.9E-05 | 1.0011 |
| SS | -3292.44 | -6379.66 | 0.4028 | 4.3E-05 | 1.0015 |
| GL-HS | -3236.90 | -6391.17 | 0.4347 | 2.4E-05 | 1.0045 |
| GL-SB | -3229.06 | -6405.15 | 0.4427 | 2.0E-05 | 1.0050 |
| GL-NEG | -3206.05 | -6346.88 | 0.4372 | 2.2E-05 | 1.0052 |
| GL-LA | -3093.73 | -6148.28 | 0.4464 | 1.7E-05 | 1.0073 |
| GL-NG | -3170.79 | -6306.13 | 0.4361 | 2.3E-05 | 1.0069 |

- SS and GL performs better than FH and SYNTH
- All the estimators provide good results in term of the bias diagnostic in Brown *et al.* (2001)

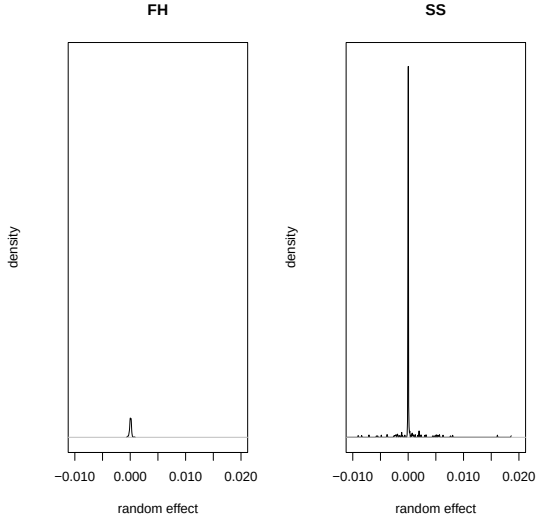
# Results

**Model $M_5$ random effects: FH, HS, SB, NEG, LA and NG.**



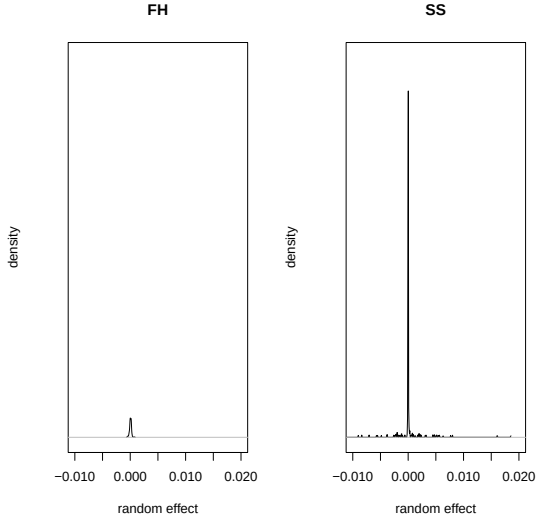
# Results

**Model $M_5$ random effects: FH, DM ($\pi(p) \sim Beta(1, 4)$.**



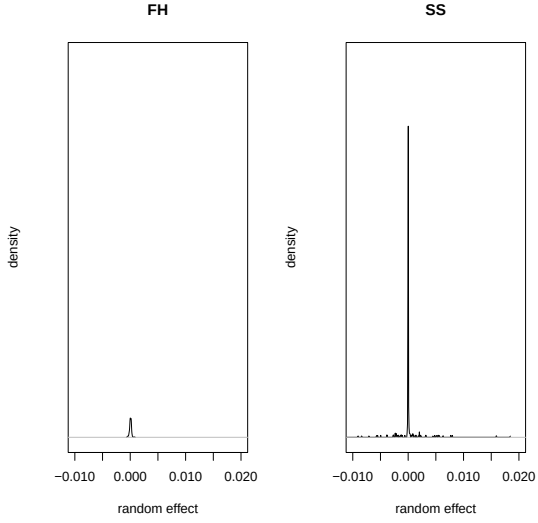
# Results

**Model $M_5$ random effects: FH, DM ($\pi(p) \sim Beta(1, 1)$.**



# Results

**Model $M_5$ random effects: FH, DM ($\pi(p) \sim Beta(4, 1)$.**



# Final Comments

- 2011 Census information much more predictive than 2021 administrative information
- SS model activates random effects only when it is considered necessary $\Longrightarrow$ it is the closest model to the SYNTH model
- Given the relatively small number of activated random effects in the SS model, the SYNTH estimates can be considered to be good estimates in most of small areas
- Methods can be grouped in classes according to the usage of random effects: SYNTH, SS, FH, GL (Three Parameter Beta Normal local priors), GL (Normal Gamma local priors)

## Future works

- administrative data adjustment through the use of spline functions
- Unmatched linking model
- Alternative sampling distribution: beta distribution instead of normal distribution

# References

Brown, G., Chambers, R., Heady, P. & Heasman, D.J. (2001). Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS". *Proceedings of the 2001 Statistics Canada International Symposium*.

Fabrizi, E., Ferrante, M.R., Pacei, S. & Trivisano, C. (2011). Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. *Computational Statistics & Data Analysis*, 55(4), 1736–1747.

Datta, G.S., & Mandal, A. (2015). Small area estimation with uncertain random effects. *J. Am. Stat. Assoc.*, 110, 1735–1744.

Fay, R.E., & Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *J. Am. Stat. Assoc.*, 74(366a), 269–277.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (with discussion). *Bayesian Analysis*, 1(3), 515-–534.

Tang, X., Ghosh, M., Ha, N.S., & Sedransk, J. (2018). Modeling random effects using global-local shrinkage priors in small area estimation. *J. Am. Stat. Assoc.*, 113(1), 1476–1489.

You Y., & Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, 30(1), 3–15.

THANKS

FOR

YOUR

ATTENTION!!!