Session 4
New data sources for small area estimation

Discussion

M. Giovanna Ranalli[1]

[1]Dip. Scienze Politiche, Università degli Studi di Perugia

Third Workshop on Methodologies for Official Statistics
ISTAT — Rome, 4-5 December 2024

Outline

1. Overview of the two papers
2. Official statistics: Bayesian Small Area Estimation of Origin-Destination Matrix
   - Loredana Di Consiglio, Fabrizio Solari, Emanuela Scavalli Massimo Armenise, Carolina Ciccaglioni, Isabella Corazziari, Tiziana Pichiorri, Lorenzo Asti, Luca Faustini — Istat, Italy
3. Controlling selection bias in non-probability sample using small area estimation: an application to official statistics
   - Francesco Schirripa Spagnolo, Gaia Bertarelli, Nicola Salvati, Stefano Marchetti — Università di Pisa, Italy; Monica Pratesi — Istat and Università di Pisa, Italy; Donato Summa — Istat, Italy; Monica Scannapieco — Agenzia per la cybersecurity, Italy.

The two papers

Similarities

- Use Small Area Estimation methods for "non-standard" SAE inferential problems
- Small Areas, but Large number of Authors

Differences

- Work in progress vs. published paper
- Bayesian vs Frequentist approach

Outline

1. Overview of the two papers
2. Official statistics: Bayesian Small Area Estimation of Origin-Destination Matrix
   - Loredana Di Consiglio, Fabrizio Solari, Emanuela Scavalli Massimo Armenise, Carolina Ciccaglioni, Isabella Corazziari, Tiziana Pichiorri, Lorenzo Asti, Luca Faustini — Istat, Italy
3. Controlling selection bias in non-probability sample using small area estimation: an application to official statistics
   - Francesco Schirripa Spagnolo, Gaia Bertarelli, Nicola Salvati, Stefano Marchetti — Università di Pisa, Italy; Monica Pratesi — Istat and Università di Pisa, Italy; Donato Summa — Istat, Italy; Monica Scannapieco — Agenzia per la cybersecurity, Italy.

Overview of the paper

- Inferential target: origin-destination (O-D) Matrix

$$M_{ij} = N_i m_{ij},$$

  commuting rates m_{ij} are the parameters of interest
  N_i is the number of employed individuals in location i

- Sparse problem $\rightarrow$ small areas: cells ij – Italian Provinces (NUTS3) 107×107 (38×107 in the application)
- Survey data from the Permanent Census
- Auxiliary data
  - Q_{ij} is the number of potential commuting journeys from Tax registers $\rightarrow q_{ij} = Q_{ij}/N_i$
  - Previous m_{ij}'s from 2011 Census.

Overview of the paper

- Inferential target: origin-destination (O-D) Matrix

$$M_{ij} = N_i m_{ij},$$

  commuting rates m_{ij} are the parameters of interest
  N_i is the number of employed individuals in location i
- Sparse problem $\rightarrow$ small areas: cells ij – Italian Provinces (NUTS3) 107×107 (38×107 in the application)
- Survey data from the Permanent Census
- Auxiliary data
  - Q_{ij} is the number of potential commuting journeys from Tax registers $\rightarrow q_{ij} = Q_{ij}/N_i$
  - Previous m_{ij}'s from 2011 Census.

Modeling approaches

- Fay-Herriot area-level model (FH)
- The assumption of a common variance σ_u^2 should be relaxed
  - Spike-n-slab (SS): $\delta_{ij}\sigma_u^2$, where δ_{ij} is a Bernoulli rv with probability θ
  - Global-Local (GL): $\lambda_{ij}^2\sigma_u^2$ with many choices of prior distributions for the local parameter
- The ability of q_{ij} to detect true commuters is a function of the distance between origin and destination locations
  - $z_{ij} = q_{ij}/d_{ij}^p$, $p = 0, 1, 2$
  - the effect of z_{ij} could be modeled by splines (in the paper but not pursued in the application)

Modeling approaches

- Fay-Herriot area-level model (FH)
- The assumption of a common variance σ_u^2 should be relaxed
  - Spike-n-slab (SS): $\delta_{ij}\sigma_u^2$, where δ_{ij} is a Bernoulli rv with probability θ
  - Global-Local (GL): $\lambda_{ij}^2\sigma_u^2$ with many choices of prior distributions for the local parameter
- The ability of q_{ij} to detect true commuters is a function of the distance between origin and destination locations
  - $z_{ij} = q_{ij}/d_{ij}^p$, $p = 0, 1, 2$
  - the effect of z_{ij} could be modeled by splines (in the paper but not pursued in the application)

Suggestions/ideas

- Gaussian assumption for commuting rates, which can be small: test for transformations
- What about modelling counts M_{ij}'s directly accounting for constraints $\sum_j M_{ij} = N_i$?
- In the need of setting many random effects to zero $\rightarrow$ Lasso or SSL?
- In SS, is it possible to have $P(\delta_{ij} = 1)$ to depend on d_{ij}?
- To relax common σ_u^2 (and also common β), a mixture model can be considered in which

$$m_{ij} | \boldsymbol{x}_{ij}, C_{ij} = c \sim N(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_c, \sigma_c^2)$$

  for $c = 1, \ldots, k$, and k is the number of groups/latent classes. The probability $P(C_{ij} = c)$ can depend on covariates.
- In case you pursue the use of p-splines, Demmler-Reinsch orthogonalization for the matrix $[\boldsymbol{Z}_1 \boldsymbol{Z}_2]$ should be used to help proper mixing of chains

Suggestions/ideas

- Gaussian assumption for commuting rates, which can be small: test for transformations
- What about modelling counts M_{ij}'s directly accounting for constraints $\sum_j M_{ij} = N_i$?
- In the need of setting many random effects to zero $\rightarrow$ Lasso or SSL?
- In SS, is it possible to have $P(\delta_{ij} = 1)$ to depend on d_{ij}?
- To relax common σ_u^2 (and also common $\boldsymbol{\beta}$), a mixture model can be considered in which

$$m_{ij} | \boldsymbol{x}_{ij}, C_{ij} = c \sim N(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_c, \sigma_c^2)$$

  for $c = 1, \ldots, k$, and k is the number of groups/latent classes. The probability $P(C_{ij} = c)$ can depend on covariates.
- In case you pursue the use of p-splines, Demmler-Reinsch orthogonalization for the matrix $[Z_1 Z_2]$ should be used to help proper mixing of chains

Suggestions/ideas

- Gaussian assumption for commuting rates, which can be small: test for transformations
- What about modelling counts M_{ij}'s directly accounting for constraints $\sum_j M_{ij} = N_i$?
- In the need of setting many random effects to zero $\rightarrow$ Lasso or SSL?
- In SS, is it possible to have $P(\delta_{ij} = 1)$ to depend on d_{ij}?
- To relax common σ_u^2 (and also common $\boldsymbol{\beta}$), a mixture model can be considered in which

$$m_{ij} | \boldsymbol{x}_{ij}, C_{ij} = c \sim N(\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_c, \sigma_c^2)$$

  for $c = 1, \ldots, k$, and k is the number of groups/latent classes. The probability $P(C_{ij} = c)$ can depend on covariates.
- In case you pursue the use of p-splines, Demmler-Reinsch orthogonalization for the matrix $[\boldsymbol{Z}_1 \boldsymbol{Z}_2]$ should be used to help proper mixing of chains

Outline

1. Overview of the two papers
2. Official statistics: Bayesian Small Area Estimation of Origin-Destination Matrix
   - Loredana Di Consiglio, Fabrizio Solari, Emanuela Scavalli Massimo Armenise, Carolina Ciccaglioni, Isabella Corazziari, Tiziana Pichiorri, Lorenzo Asti, Luca Faustini — Istat, Italy
3. Controlling selection bias in non-probability sample using small area estimation: an application to official statistics
   - Francesco Schirripa Spagnolo, Gaia Bertarelli, Nicola Salvati, Stefano Marchetti — Università di Pisa, Italy; Monica Pratesi — Istat and Università di Pisa, Italy; Donato Summa — Istat, Italy; Monica Scannapieco — Agenzia per la cybersecurity, Italy.

Overview of the paper

- Estimating the proportion of Italian Enterprises sensitive of SDGs at provincial (NUTS3) level.
- Bias Correction of the estimates from a non-probability sample B for small domains using a probability sample A.
- Target variable comes only from Big Data source B
- The domains are small in the probability sample A
- Doubly robust approach: IPW + Mass imputation
- The selection mechanism of the big data sample is ignorable
- SAE "flavour" in both models

Overview of the paper

- Estimating the proportion of Italian Enterprises sensitive of SDGs at provincial (NUTS3) level.
- Bias Correction of the estimates from a non-probability sample B for small domains using a probability sample A.
- Target variable comes only from Big Data source B
- The domains are small in the probability sample A
- Doubly robust approach: IPW + Mass imputation
- The selection mechanism of the big data sample is ignorable
- SAE “flavour” in both models

Things I would have asked if I were a Referee

- Binary target variable: why the naive approach is used instead of EBP?
- Which is the predictive power of the predictors for both models? Is it comparable to that in the simulations?
- Small n_{Bi}: IPW for areas that are small also in the Big Data source: can this be an issue in terms of extra variability?
- Simulations: it would be interesting to disentangle the role of IPW and Mass Imputation $\rightarrow$ performance of Mass Imputation only and of IPW only?

Things I would have asked if I were a Referee

- Binary target variable: why the naive approach is used instead of EBP?
- Which is the predictive power of the predictors for both models? Is it comparable to that in the simulations?
- Small n_{Bi}: IPW for areas that are small also in the Big Data source: can this be an issue in terms of extra variability?
- Simulations: it would be interesting to disentangle the role of IPW and Mass Imputation $\rightarrow$ performance of Mass Imputation only and of IPW only?

Bootstrap

- Joint Model-based / Design-based variance

?? Step 1: Extract a sample of size n_A from sample A using the inclusion probabilities π_{ij}

- Step 1 needs to follow standard practice in survey sampling on creating replication weights for design-based variance estimation, e.g. using Rao-Wu or other FP bootstrap methods. How was it performed?

?? Step 2: Why SRS? Is Step 3 enough to account for IPW?

- The proposal is DR and does not rely on Mass Imputation only: explore and exploit all the literature on non-response IPW (two-phase approach using Poisson).

Bootstrap

- Joint Model-based / Design-based variance

?? Step 1: Extract a sample of size n_A from sample A using the inclusion probabilities π_{ij}

- Step 1 needs to follow standard practice in survey sampling on creating replication weights for design-based variance estimation, e.g. using Rao-Wu or other FP bootstrap methods. How was it performed?

?? Step 2: Why SRS? Is Step 3 enough to account for IPW?

- The proposal is DR and does not rely on Mass Imputation only: explore and exploit all the literature on non-response IPW (two-phase approach using Poisson).

Title of Session 4: a little twist

New data sources for small area estimation

$\Downarrow$

Small area estimation for new data sources

Title of Session 4: a little twist

New data sources for small area estimation

$$\Downarrow$$

Small area estimation for new data sources