Methods for multisource statistics perspective Discussion

David Haziza

Department of mathematics and statistics
University of Ottawa

3rd Workshop on Methodologies for Official Statistics
Rome, Italy

December 4, 2024

The three papers...

- **Paper 1**: *Optimization of surveys: The 'Integrated Census and Social Surveys System' project*
  C. De Vitiis, S. Falorsi, A. Guandalini, F. Inglese, S. Loriga, M. Mazziotta, F. Piersimoni, R. Ranaldi, M. Russo, M.D. Terribili, R. Benedetti

- **Paper 2**: *Combining survey data and mobile network operator data for commuting statistics*
  T. Tuoto, E. Cerasti, L. Di Consiglio, D. Filipponi, T. Pichiorri, L.-C. Zhang

- **Paper 3**: *Linearization approach for measuring the accuracy of multinomial outcomes from a statistical register*
  S. Falorsi, D. Chianella, R. Filippini, S. Toti, N. Deliu, P.D. Falorsi

Paper 1: Goal

- The initial design of the Integrated Census and Social Surveys System (SICIS) was based on a modular approach but was partially implemented (e.g., respondent burden)
- New SICIS: Explore alternative sampling designs
  - Two-phase sampling design
  - Spatially balanced sampling
- Two empirical studies were conducted.

Paper 1: Two-phase sampling

- Three different sampling designs were implemented using the Master Sample, LFS and AVQ.
- Three scenarios were considered: No integration (S1), Integration at the first stage (S2A), Integration in both stages (S2B).
- For each scenario, three estimators were considered: HT, CAL1 and CAL2.
- Results were presented for the total number of employed individuals in three Italian regions:
  - For a given estimator, results, in terms of estimated coefficient of variation, were very similar for the three sampling designs
  - For a given sampling design, the point estimator (CAL1 or CAL2) does not have much effect.
- Would these results still be observed at the domain level?

Paper 1: Some results

Table 2.1 – Estimated percentage coefficient of variation for the total number of employed individuals in the three regions

Survey	Integration scenario	Cal1 estimator	Cal2 estimator
MS	S1	0.191	//
LFS	S1	0.609	0.597
	S2A	0.650	0.649
	S2B	0.644	0.640
AVQ	S1	0.977	0.986
	S2A (ind. strat)	1.013	0.996
	S2B (ind. strat)	1.073	1.052
	S2A (dep. strat)	1.012	0.979
	S2B (dep. strat)	0.990	0.966

Paper 1: Some results

Table 2.2 – Estimated design effect, estimator effect, estimator effect due to MS estimates, for a proportion equal to 0.15

Survey	Integration scenario	Cal1 estimator	Cal2 estimator	DesEff	EstEff	MSEstEff
MS	S1	0.40	//	2.55	0.41	//
LFS	S1	1.41	0.98	1.46	0.54	0.28
	S2A	1.05	1.05	3.34	0.28	0.15
	S2B	1.05	1.05	3.40	0.28	0.15
AVQ	S1	2.50	1.62	1.31	0.56	0.24
	S2A (ind. strat)	2.63	1.67	2.23	0.37	0.22
	S2B (ind. strat)	2.65	1.70	2.24	0.37	0.22
	S2A (dep. strat)	2.50	1.62	1.36	0.54	0.31
	S2B (dep. strat)	2.45	1.59	1.27	0.56	0.32

Paper 1: Spatially balanced sampling

- Change of paradigm:
  - In previous design, the assumption is that municipalities with approximately same number of people have the same behavior
  - In a spatially balanced design, neighboring municipalities have the same behavior
- Use of the Local Pivotal Method $\longrightarrow$ the selection tends to favor municipalities that are spatially distant from each other.
- 6 types continuous indicators (e.g. income variables, demographic variables, etc.) $\longrightarrow$ Moran's index to measure spatial autocorrelation
- Wide range of Moran's index
- Three different spatially balanced sampling design were tested. They are different with respect to the stratification

Paper 1: Spatially balanced sampling

- Gains seem significant:
  - Income variables: 30%-50% gains,
  - Labor market participation variables: 25%-30% gains
  - Demographic and family variables: up to 25% gains
- It would be interesting to see the results for a calibration estimator (e.g., CAL2) instead of the HT estimator. Would the gains be as significant?
- If not, the use of spatially balanced sampling may generate some complexity in terms of variance estimation, small area estimation (?) $\longrightarrow$ is it worth it?
- Advantage of spatially balanced sampling: no longer necessary to perform stratification at the sub-provincial level $\longrightarrow$ but are these design robust?

Paper 2: Goal

- Paper is exploring the use of MNO data $\longrightarrow$ first experimentation with MNO data
- Do MNO data help in building better estimators/predictors?
- Potential issues with MNO data:
  - Measurement errors: e.g., location errors.
  - Missing Data: Device may be inactive or unable to transmit data due to network issues or privacy settings.
  - Undercoverage: MNO data may represent only subscribers from a particular operator
- Do we know if MNO data in Italy suffer from these errors and if so, do we have an idea of the magnitude of these errors ?

Paper 2: Estimation/Prediction

- If $n_{ij} > 0$ then use the EBLUP (based on the Fay-Herriot model)
- If $n_{ij} = 0$ (empty sample domains), two options:
  - Use the customary synthetic estimator;
  - Use an alternative using "transfer learning" based on a proxy Y_{ij}^* available from an independent source.
  - In the application, it looks like (Figure 1) the regression estimator (synthetic estimator) was based on 2011 Census data, admin data for work and study and MNO data. Since the proxy Y_{ij}^* is available for all the domains, what effect would its inclusion would have on the performance of the regression estimator (i.e., the synthetic estimator based also on Y_{ij}^*)?
  - Similarly, from Table 1, it seems that the proxy Y_{ij}^* was not used in the Fay-Herriot model. Can it be included and if so, what's its contribution in improving the model?

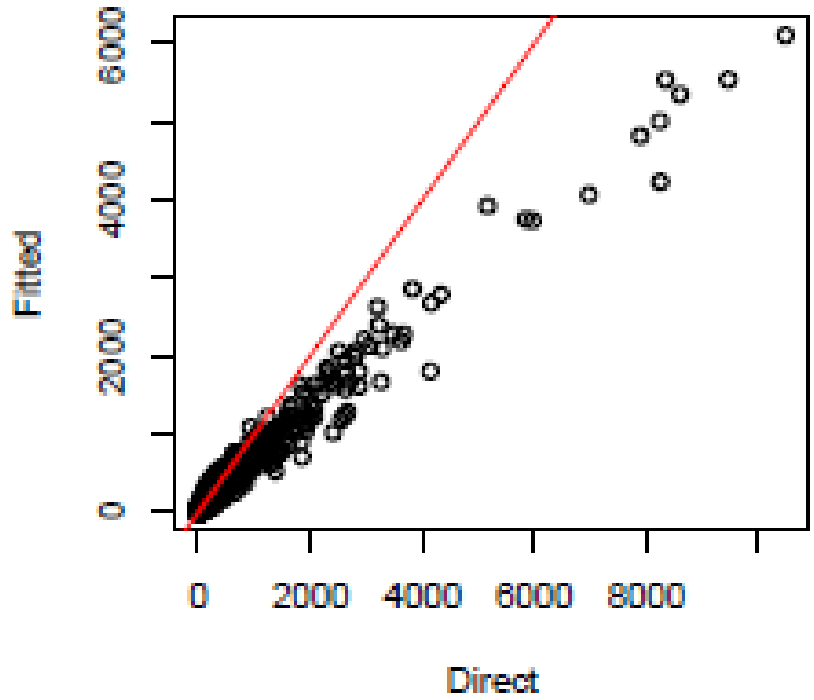
Paper 2: MNO data

- From the results (Table 1), it looks that in the presence of 2011 Census data, Admin data on work and study, MNO data "do not bring much" (BIC of 8567 vs. 8507)
- Is this due to potential measurement errors in the MNO data?
- For most domains with $n_{ij} > 0$, it looks that the EBLUP was essentially equal to the synthetic estimator $\longrightarrow$ This suggest that the number of domains is large...

Covariates	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
MNO	0.094	0.109	0.119	0.299	0.458	0.992
MNO + 2011 census	0.009	0.010	0.011	0.078	0.068	0.914
MNO + 2011 census + Admin W&S	0.004	0.005	0.005	0.046	0.033	0.833

Table 2: Summaries of γ values of the FH models

Paper 2: Some results



The information displayed on this figure suggest that if we had done a plot of residuals vs. the predicted values, we would not have an average equal to 0!

Paper 2: Some results

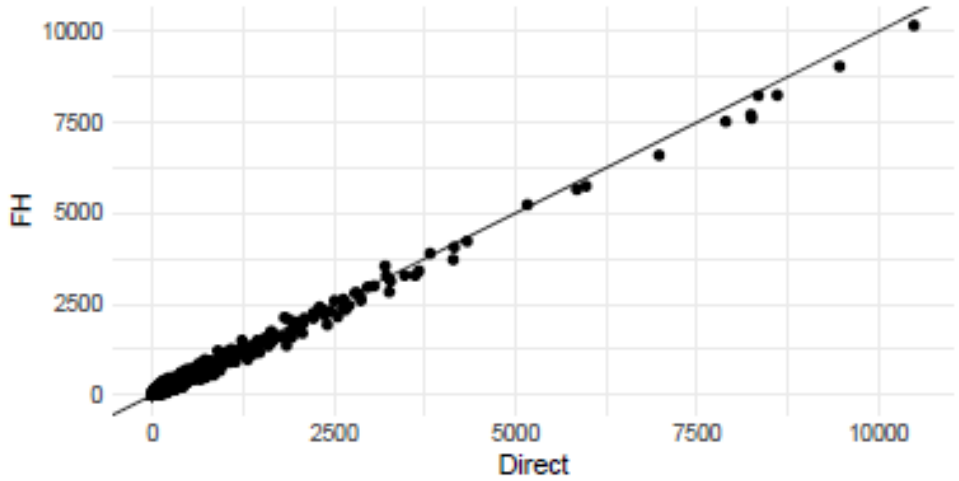


Figure 2: Direct estimates of the flows vs FH model estimates with MNO, 2011-census, work and study admin data as covariates

Paper 2: Some results

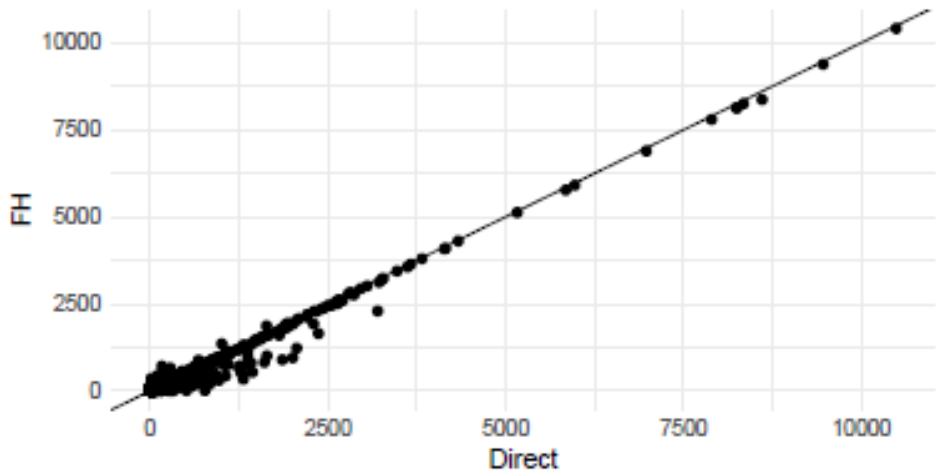


Figure 4: Direct estimates of the flows vs FH model estimates with only MNO data as covariates

Paper 2: Some final remarks

- It looks that that some log-transformation was applied to the data $\longrightarrow$ the normal assumption for the errors of the FH model is not tenable with flows? Presence of influential domains? Application of robust methods?
- MNO data may be complex $\longrightarrow$ may be worth to explore the use of ML methods such as random forests in the context of the Fay-Herriot model; e.g., Bosa et al. (2024).
- Estimation of the mean square error in the case of spatial model: Use of bootstrap? $\longrightarrow$ a good starting point may be Molina, Salvati and Pratesi (2008).

Paper 3: Goals and setup

Goals:

- Provide a register with predicted values of an outcome variable (Here, ALE).
- Provide a measure of quality (GMSE) based on a first-order Taylor expansion

Setup:

- U: Finite population of size N
- S: sample of size n selected according to a probability sampling design
- Assumption: sampling is non-informative (i.e., the design variables are incorporated in the model, if appropriate)
- Y: survey variable (Here, categorical with K categories)
- Goal: estimate the domain totals: $\theta_k^{(d)} = \sum_{i \in U} \gamma_i^{(d)} Y_{ik}$

Paper 3: Estimation and accuracy

Data:

- Y: available for all $i \in S$
- $\mathbf{x}$: available for all $i \in U$.

Proposed estimator of $\theta_k^{(d)}$:

$$\widehat{\theta}_k^{(d)} = \sum_{i \in U} \gamma_i^{(d)} \widehat{Y}_{ik} = \sum_{i \in U} \gamma_i^{(d)} f(\mathbf{x}_i; \widehat{\beta}) = \sum_{i \in U_d} f(\mathbf{x}_i; \widehat{\beta})$$

- Nice features: simple and (approximately) model-unbiased if the first moment of the model is correctly specified.
- Drawback: vulnerable to bias if model is misspecified.

Paper 3: Remarks and questions

- Question: Can we consider a design consistent estimator instead?
  $\longrightarrow$ Double robustness

$$\widehat{\theta}_k^{Reg,(d)} = \sum_{i \in U_d} f(\mathbf{x}_i; \widehat{\beta}) + \sum_{i \in S_d} \frac{1}{\pi_i} \{y_i - f(\mathbf{x}_i; \widehat{\beta})\}$$

- Issue: $\widehat{\theta}_k^{Reg,(d)}$ does not have a convenient form
- Possible remedy: Calibrated imputation; i.e., determine final predicted values $\widehat{Y}_{ik,F}$ as close as possible to to the initial values $\widehat{Y}_{ik} = f(\mathbf{x}_i; \widehat{\beta})$ subject to

$$\sum_{i \in U_d} \widehat{Y}_{ik,F} = \sum_{i \in U_d} f(\mathbf{x}_i; \widehat{\beta}) + \sum_{i \in S_d} \frac{1}{\pi_i} \{y_i - f(\mathbf{x}_i; \widehat{\beta})\}$$

Paper 3: Remarks and questions

- Measure of accuracy

$$\text{GMSE}(\widehat{\theta}_k^{(d)}, \theta_k^{(d)}) \approx \mathbb{E}_{\boldsymbol{\lambda}} \mathbb{V}_{\mathbf{Y}}(\widehat{\theta}_k^{(d)})$$

  if the sampling fraction n/N is negligible

- Issue: The validity of $\widehat{\text{GMSE}}(\widehat{\theta}_k^{(d)}, \theta_k^{(d)})$ generally requires the second moment of the model to be correctly specified, and in the case of a binary variable, it requires the first moment of the model to be correctly specified.

- If we use a regression type estimator, then we could use the estimated design variance $\longrightarrow$ brings some robustness

Paper 3: Alternative prediction methods

- The paper investigates the use of a parametric model $\longrightarrow$ vulnerable to model misspecification (functional form, omission of interactions, etc.)
- An alternative method:
  - Obtain preliminary estimated $\widehat{f}(\mathbf{x}_k)$ using a possibly ML procedure
  - Use a clustering algorithm to form C cells with respect to both $\widehat{f}(\mathbf{x}_k)$
  - Perform random-hot-deck imputation within each cell
- Variance estimation: would assume that the classes are fixed $\longrightarrow$ may lead to some underestimation.