

# Discussion: Methodologies for the new censuses

David Haziza

Department of mathematics and statistics  
University of Ottawa

Workshop on Methodologies for Official Statistics  
Rome, Italy

December 5, 2022

## Two papers

- S. Falorsi: *Census and social surveys integrated systems*
  - Combining administrative and survey data → Permanent Census
  - Software package MIND
- M. Di Zio and D. Filippini: *Multi-source statistics in the Italian permanent census*
  - Combining administrative and survey data → Imputation of ALE and OCC
  - Variance estimation

## Census and social surveys integrated systems

- Italian permanent census: combine administrative and survey data
- Reasons:
  - Significant reduction of the costs;
  - Reduction of the respondent's burden;
- ISTAT developed a new data methodological/statistical framework by integrating 3 components:
  - Integrated Register System (IRS): Integrates data from administrative sources and surveys at the individual level → the missing data are imputed
  - Permanent Population Census (PPC): produces set of estimates (via small area techniques) that cannot be obtained through administrative sources
  - Census and Social Surveys Integrated System (CSSIS): Other set of estimates (SAE based on unit level models)

# Census and social surveys integrated systems

- PPC: Goal is to produce a set of values (observed or predicted) at the individual level → micro-level approach
- Requirements for the set of estimates produced by the PPC:
  - Accuracy (Validity): this requires the specification of a model → Model diagnostics become central

$$\text{MSE}(\hat{\theta}) = \mathbb{V}(\hat{\theta}) + \text{Bias}^2(\hat{\theta})$$

- Efficiency: small variance
- Consistency: internal and external consistency → Specify a set of calibration constraints

## Census and social surveys integrated systems

- In PPC/CSSIS, estimation procedures involve SAE methods;
- Impressive *R*-package Mind (Multivariate model-based INference for Domains)
- Can handle multiple survey variables but with a common/unique set of covariates (Limitation?)
- Allows for different correlation structures (including spatial correlations between levels of each random effect)
- Estimation of MSE

## Census and social surveys integrated systems

- Since with data integration methods, we make many assumptions, a lot of efforts should be placed on model diagnostics to detect departures from model assumptions
- For instance, the macro GEST developed at Statistics Canada includes a SAE component. GEST offers several useful diagnostics for the Fay-Herriot model (only the combined Fay-Herriot model can be validated):
  - Plot of residuals vs. set of predictors, predicted values, etc.; If the assumptions do not seem to be satisfied, then we may consider adding polynomial term of higher order or consider piecewise linear regression;
  - Plot of standardized square residuals vs. set of predictors, predicted values; If the assumptions do not seem to be satisfied, then try to determine the right amount of heteroscedasticity (may not be easy);
  - Normality of the standardized errors;
  - Before validating the Fay-Herriot model, important to validate the smoothing model (that was used to smooth the sampling variances)

## Census and social surveys integrated systems

- Does MIND propose (or will propose) a set of diagnostics in the context of multivariate models? May be useful to users.
- Other important issue in SAE: Outliers
  - Outlier detection (influential domains or influential units);
  - Important issue in unit level models;
  - May be also an issue in area level models;
  - Vast literature on robust SAE for unit level models (e.g., Bertarelli et al., 2022, Favre-Martinoz, 2015, Dongmo Jiongo et al., 2013, Sinha and Rao, 2009) but not much has been done for area level models;
  - What about robust multivariate SAE models → More research needed?

## Census and social surveys integrated systems

- Mass imputation/propensity score weighting are common in the data integration context;
- The resulting estimators are vulnerable to model misspecification → nonparametric/machine learning methods may bring some robustness
- Often we make a MAR-type assumption. What if it's not satisfied?
  - We can have recourse to multiply robust procedures (e.g., Chen and Haziza, 2022): In the case of NMAR, these methods tend to lead to a better estimator (although inconsistent) than the one that would have been obtained under a single misspecified model;
  - Multiply robust propensity score weighting assuming NMAR (Kim and Cho, 2022) → the price to pay is to have an independent validation sample with the same measurements ( $y$  and  $x$ ) → There is no free lunch!



# Multi-source statistics in the Italian permanent census

- Goal: Mass impute the variable Attained Level of Education (ALE) and Occupational Status (OCC) in the Italian Base Register of Individuals
- Variables used to mass impute come from:
  - Ministry of Education Universities and Research (MIUR): administrative data;
  - 2011 Census Information;
  - Sample Survey (collected since 2018);
- Mass imputation is justified by the high amount of detailed information → Rich imputation model

# Multi-source statistics in the Italian permanent census

- Different patterns of missing data involve different set of covariates
- Imputation procedure:
  - First, estimate  $P(ALE^t | x)$  using a log-linear model applied to the contingency table obtained by cross-classifying the variables  $ALE^t$  and  $x \rightarrow \hat{P}(ALE^t | x)$
  - Randomly generate a ALE status with probability  $\hat{P}(ALE^t | x)$
  - If we use a saturated model, then equivalent to random hot-deck imputation within cells
  - If some cells are empty  $\rightarrow$  use a subset of covariates selected through a cross-validation procedure

# Patterns of missing data

$X_{BRI}$				$X_{miur}$				<b>Sample</b>	<b>Prediction</b>	<b>Group</b>
<b>G</b>	<b>E</b>	<b>P</b>	<b>Ct</b>	$L(t)$	$ALE^{t-12}$	$F(t)$	$ALE^{apr}$	$ALE'_s$	$ALE^t$	
										A
										22%
										B
										73%
										C
										5%

## Variance estimation: Numerical results (Di Zio, Filippini, Toti, 2022)

ALE	Analytical $\hat{V}$	Margin of error $1.96\sqrt{\hat{V}}$
Illiterate	$1.42 \times 10^{-8}$	0.000233
Literate but no attainment	$6.62 \times 10^{-8}$	0.000504
Primary education	$1.82 \times 10^{-7}$	0.000836
Lower secondary	$3.78 \times 10^{-7}$	0.001250
Upper secondary	$3.69 \times 10^{-7}$	0.001190
BSc	$7.49 \times 10^{-8}$	0.000536
MSc	$9.81 \times 10^{-8}$	0.000613
PhD	$1.16 \times 10^{-8}$	0.000211

Table 1: Variance estimates and associated margin of errors

# Multi-source statistics in the Italian permanent census

- Extremely small margins of error!
- May be due to:
  - Very large sample sizes;
  - Very powerful covariates → Imputation model highly predictive

- Point estimator:

$$\widehat{Y} = \frac{1}{N} \left\{ \sum_{i \in S} y_i + \sum_{i \in S^c} \widetilde{y}_i \right\} = \frac{n}{N} \bar{y}_s + \left( 1 - \frac{n}{N} \right) \bar{\widetilde{y}} \approx \bar{\widetilde{y}}$$

- With such small variances, a small bias may lead to invalid inferences → coverage probability of normal-based confidence intervals may be much lower than 95% even if the bias is small → Importance of validating the model as much as possible

# Multi-source statistics in the Italian permanent census

- Imputation of OCC is more complex as all the data sources may suffer from measurement errors → Use of mixture Markov models
- Variance estimation for OCC:
  - Analytical approach too complicated;
  - Use of Multiple Imputation (MI) → MILC procedure (Boeschoten et al., 2020)
- Multiple Imputation variance estimator (Rubin, 1978):

$$\widehat{V}_{tot} = \overline{W}_M + \left(1 + \frac{1}{M}\right) B_M,$$

where

$$\overline{W}_M = \frac{1}{M} \sum_{m=1}^M W^{(m)}, \quad B_M = \frac{1}{M-1} \sum_{m=1}^M \left(\widehat{\theta}_I^{(m)} - \widehat{\theta}_{I,M}\right)^2$$

- Validity of  $\widehat{V}_{tot}$  relies on the fact that the imputation procedure is proper.