Rome. 5-6 December 2022
WORKSHOP ON METHODOLOGIES IN OFFICIAL STATISTICS

Multi-source statistics in the Italian permanent census

Marco Di Zio, Danila Filipponi

Istat | DIRECTORATE FOR METHODOLOGY AND STATISTICAL PROCESS DESIGN

Outline

- Context
- Multi-source estimates of attained level of education
- Multi-source estimates of employment status
- Conclusions

Introduction

- Istat Census of Population is based on the integration of survey data, the base register of individuals (BRI), administrative data.
- Census sample surveys are conducted for dealing with errors and lack of information in administrative sources (coverage errors, variables not in register,…).
- After integration, for some core variables statistics are obtained by a direct computation on BRI data (register-based statistics).

Variables for multi-source register-based statistics

- *Gender*, *place* and *date of birth*, *citizenship* are obtained by integrating only admin data and can be thought free of errors (negligible errors).
- *Attained level of education* (ALE) and *employment status* (OCC) need to be estimated by integrating admin data and sample surveys.
- For ALE and OCC, statistical models are necessary and consequently are affected by a *natural degree of uncertainty* that should be taken into account in their usage.

Estimation of attained level of education (ALE)

- **SOURCES:** administrative, 2011 Italian census, sample surveys.
- Data can be divided in 3 groups:
- *Group A.* People with admin information on ALE at time t-2. Characterized with longitudinal information on course attendance in (t-2,t-1), for people entering a study program after 2011 to t-2 (≈ 22% of pop older than 9 yrs).
- *Group B.* People not enrolled in any school course included in admin data from 2011 to t-2 with information from 2011 Census (≈ 73% of pop older than 9 yrs).
- *Group C.* People neither in admin nor in 2011 Census data. No direct information on ALE. Mainly adults not Italian (≈ 5% of pop older than 9 yrs).

Estimation of attained level of education (ALE)

- **REMARK**

Admin info is affected by some gaps: some qualifications are not included, time-lag with respect to reference time.

- Census sample survey data are used to fill the gaps

Estimation of attained level of education (ALE)

- ALE prediction procedure at time t is based on log-linear imputation.
- Conditional probabilities of ALE at time t (ALE_t) given a set of covariates X, $\Pr(ALE_t|X)$ is estimated.
- $\Pr(ALE_t|X)$ estimated through log-linear model. It is applied to the contingency table of $(ALE_t|X)$ and $\Pr(ALE_t|X)$ is obtained from their estimated expected counts.
- ALE_t is predicted on BRI units by a random draw from $\Pr(ALE_t|X)$.

Estimation of attained level of education (ALE)

- Group A. Pr(ALEt |X) is estimated by using only administrative data focusing on data shifted by two years earlier:

Pr(ALEt-2| ALEt-4, age, citizenship, school attendance, …)

- For groups B and C, Pr(ALEt |X) estimated by using ALE_t of the census sample as a target variable.

Group B. Pr(ALEt| ALE_{2011}, age, citizenship, prov. residence, gender)

Group C. Pr(ALEt | age, citizenship, gender, apr, sirea).

Estimation of employment status (OCC)

- **SOURCES:** LFS, Admin, census sample survey.
- True employment status at time t for unit k is modelled as a binary latent variable $L(t,k)$ (employed or not).
- L is analyzed at times $t = 1, \dots, T$ and $L(1{:}T)$ denotes the r.v. $L_1, \dots, L_T$ and each time t corresponds to a specific month of the year.
- Census survey, LFS and administrative sources are treated as imperfect measures of the target process.
- $Y_{1:T}^{i}$ with $i = 1{,}2$ binary vectors of (possibly missing according to the sampling design) of the employment status at times $1, \dots T$ resulting from the two surveys
- $Y_{1:T}^{3}$ is 1 if unit appears in at least one of the administrative sources (0 otherwise).
- Covariates X. Sex, age class, income class, ALE, two binary flags associated with retired and student status.
- Covariate S account for different quality levels of the different administrative sources.

Estimation of emplyment status (OCC)

Model requires the definition of two parts:

- Latent model describes the distribution of the latent variables

- Measurement model describes the conditional distribution of the observed variables given the latent variables.

Estimation of emplyment status (OCC)

- **Latent model** is a mixture of Markov models.
- Heterogeneity in employment activities modelled through a latent variable G that are: *never working people*, *individuals with stable employment*, *people who are likely to change frequently their employment status*.
- Distribution of G is $P(G = g|X = x, S = s),\ g =, 1,2,3.$
- Employment dynamics L for each sub-population g is a 1st order Markov chain with initial probabilities $\tau_{jg} = P(L_j = j|G = g)$ and transition matrix M^g whose typical element $\left\{M_{jk}^g\right\}$ is $P\left(L_t = k\middle|L_{(t-1)} = j, G = g\right) \quad j, k = 0,1$

Estimation of employment status (OCC)

- MEASUREMENT MODEL, i.e., the probability distribution of $Y^{g}_{(1:T)}$ given the latent process and covariates.
- Assumptions. i) Measurement processes are independent; ii) measures in LFS, admin and Census at time t are independent with the corresponding measures at different times.
- Parameters of observational model: $\psi^{g}(j|i) = P(Y_t^g = j|L_t = i)$ for $g = 1,2$ and $\psi^{3}(j|i) = P(Y_t^3 = j\big|L_t = i, S = s)$ for admin, $t = 1, \dots, 12, (i,j) = \{0,1\}, s = \{1,2,3,4\}$.
- Constrain. $\psi^{1}(j|i) = 0$ which means that "no false positive" data are in LFS.
- OCC is scored trhough a random draw from the estimated conditional distributions of latent variables given an observed configuration of sources and covariates.

Multi-source statistics: two different strategies

- Statistical procedures for the prediction of ALE and OCC are representative of two approaches when dealing with multisource data.
- **ALE**. One of the data sources can be taken as a reference, i.e., one source provides a correct measures of target variable (*supervised approach*).
- **OCC**. All data sources are affected by errors. To overcome the deficiencies of sources, they are considered multiple measures of the true target variable. The target (non-observed) variable is considered as a latent variable and a prediction conditionally on the observed values of the data sources is obtained using latent variable models (unsupervised approach).

Multi-source statistics of ALE OCC: common features

- Procedures aim at estimating a value of OCC and ALE for each unit in BRI through a random draw from the estimated probability distribution.
- This approach naturally increases the variability of the data and estimates but has the advantage of better preserving the probability distribution of the variables, thus ensuring greater flexibility in their usage.
- On the other hand, this advantage may transform to a risk, because users can be tempted to use micro-data without any limitation.
- For this reason, it is of fundamental importance to provide a flexible tool for measuring uncertainty of estimates at various unplanned level of aggregation.

A key aspect: Quality assessment.

- Since the aim is providing a set of micro-data on which a user may easily compute estimates, a flexible tool for computing accuracy is desirable.

- For ALE, an *analytical approximation* is studied, while *multiple imputation* (MI) is considered for OCC.

A key aspect: Quality assessment.

- Analytical approach is strictly connected to the method, moreover it resorts to approximations that may fail when domains of estimates become small.
- MI is designed to allow the user to evaluate uncertainty of unplanned estimates through the release of multiple micro-data but its use in a NSI is still a problem especially for managing and even more for accepting the idea of producing ‘more’ potential registers (multiple registers).
- More research is needed

Thank you

MARCO DI ZIO | dizio@istat.it
DANILA FILIPPONI | dafilipp@istat.it

Istat | Istituto Nazionale di Statistica