

Census and social surveys integrated system

Michele D'Alò, Stefano Falorsi

5 December 2022

- 1 Overview
- 2 Census and Social Surveys Integrated System (CSSIS)
- 3 CSSIS framework
- 4 Second phase sampling strategy
- 5 Conclusions

The new production strategy

In 2016, a deep methodological review of the official statistical production system began, which moved from a traditional functional scheme system to a new one based on a *multi-source approach*.

- The new strategy underlying the new *Population Census System (PCS)* aims to integrate the information stored into *registers* with those specifically collected by *census*’ *master sample survey*.
- The *Integrated System of Registers (ISR)* is the backbone of the framework. It is built at single record level, mainly through the massive integration of administrative data, but also from surveys, when the sub-populations of interest are not covered by administrative sources.

Population Register and Master Sample data

First two pillars of framework ...

- The set of *Register Variables* - for which the **primary input sources are from administrative data** referring to large part of the population while secondary input sources are from sample data - are included in the *Population Register (PR)*. These allow to obtain just some of the census counts of interest.
- The set of *Census Survey Variables* - for which **primary input source is from sample data** collected with *Master Sample (MS)*. It is planned to provide the remainder of the Census counts of interest characterized by high granularity in terms of territorial and structural detail. MS process exploits administrative data through PR as secondary input source in terms of auxiliary information.

The third pillar if the framework ...

- A subset of *Social Surveys Variables* - for which the primary input source is from sample data while **secondary input sources are from MS and PR data** in terms of auxiliary information - related to census complete the overall framework.
- these extend census information producing estimates of counts for some fundamental variables arising from annual and sub-annual Social Surveys related to census.

Census and Social Surveys Integrated System

The overall framework is a general informative and methodological infrastructure named ...

- *Census and Social Survey Integrated System (CSSIS)* whose aim is an integrated statistical production process based on a integrated data-base containing, at different levels of granularity, information gathered from the census and social surveys
- For this system, integrated estimation strategies can be used to combine different sources of information (surveys and registers) in order to *increase the accuracy* of the estimates, ensuring at the same time that all the set of estimates computed within this framework *are among them consistent* .

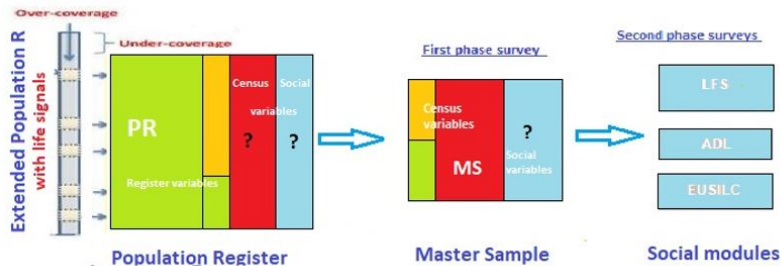
Sampling scheme of CSSIS

Given the goal of CSSIS to *increase accuracy and consistency* of estimates a *Two phase sampling scheme* is adopted:

- The *first-phase sample* of MS provides information needed for Census figures and auxiliary information for the second phase round, in which the variables of interest for the social modules are collected
- The *second-phase surveys* are a set of negatively coordinated *nested*, *not-nested* or *partially nested two stage samples* (Municipality-Household) drawn from MS aimed at extending and integrating information on social statistics within *Labour Force Survey (LFS)*, *Aspects of Daily Life (ADL)* and *(EUSILC)*.

General overview of CSSIS

This scheme gives a general representation of CSSIS showing integration of data among PR, MS and SS



Extended Population Register

In the previous figure ...

the *number of person belonging to the usual resident population* has been initially computed using, as *primary source* the information collected with the MS, but since 2020 it is identified removing and adding records in the *Extended Population R* (including not resident workers and students too) on the basis of:

- *Administrative life signals*
- *Hypothesis of absence of under coverage of the Extended PR .*

Register Variables

The set of *Register Variable* such as ...

- *gender*, *age*, *marital status* are primarily derived from demographic sources;
- *educational level*, for which the administrative information provides a good approximation for the large part of population records, except for some hard to get sub-populations a micro-level model-based prediction step is performed using MS data
- *employment status* (employed/not-employed) for which the administrative data are strongly correlated with the target variables is involved in a micro prediction stage of data using Admindata, MS and LFS.

Census Survey Variables 1

The set of *Census Survey Variables* such as *not-employed status*, *commuting* are primarily derived from MS; for which the Register Variables and other administrative sources provide the set of available auxiliary information

- *Indirect unit-level small area estimators* are applied due to the extreme granularity of the domains of interest, which makes it *impossible to produce reliable direct estimates for all domains of interest* . In addition, some of them, such as municipalities, are not structurally covered by the sample because not all smaller municipalities are surveyed each year, but only one-fifth of them.

Census Survey Variables 2

In particular the following small area estimation process is applied ...

- *multinomial small area models with fixed effects* are adopted as working models for the estimation phase which are similar to a *mass imputation process*
- the corresponding direct estimates at a more aggregated level are not introduced as benchmark in the process but are only an element of consultation and reference

Adopt. methods for Register/Census variables

<u>Registers/Unit-level databases</u>	<u>Estimated variables</u>	<u>Adopted statistical methods</u>
<u>Population Register (PR)</u>	Living population indicator for <u>years 2020 and 2021</u> under the assumption of 10 undercoverage of the extended population register	Latent class models on contact/noncontact outcomes of MS individuals crossed with life signals to support deterministic correction based on profiling of administrative source life signals
<u>Population Register (PR)</u>	Correction weight for under and over coverage of the register for <u>years 2011 and 2019</u>	Area-level small area estimators borrowing strength from <u>Geographic Regions</u> with <u>municipality random effects</u> for estimating undercoverage and <u>overcoverage</u> based on administrative life signals and demographic information as auxiliary variables
<u>Population Register (PR)</u>	<u>Educational level</u>	Imputation of missing records and subpopulations using a log-linear model based on 2011 census data and administrative data of individuals with educational attainment from 2011 onward
<u>Population census individual level data file (with PR records as backbone)</u>	<u>Occupational status (Employed/Unemployed)</u>	<u>Latent class hidden markov models</u> which integrate administrative data source on regular employment with data from the Labor Force Survey (multiple survey occasions) together with data from the MS
<u>Population census individual level data file (with PR records as backbone)</u>	(1) <u>Unemployed people by condition</u> – i.e. housewife student withdrawn from work,... - (2) commuting people;	Unit level small area estimators borrowing strength from <u>Geographic Regions</u> based on <u>multinomial fixed effect models</u> using data from MS exploiting <u>PR variables as auxiliaries</u>
<u>Population census housing and buildings data file</u>	Residential dwellings for some structural characteristics	Composite design-based small area estimator using MS data and 2011 Census and Cadastre variables as auxiliaries

ADL Social Survey Variables

In the figure...

- *ADL module* e is a *nested second phase sample* drawn from MS , *LFS* and *EUSILC* are instead *partially nested second phase sample*.
- *ADL* has been *already selected as nested module of MS*: both the municipalities and households are selected among the sampled municipalities and households of the MS.
- *The complexity of this design* is given by the *different stratifications* that can be used in the two surveys and, moreover, *the use of stratified two-stage sample design* for both surveys

LFS and EUSILC Social Survey Variables

- *LFS* and *EUSILC* are *still independent surveys* even if, for LFS in the first stage a sub-sample of the municipalities has been selected for the MS, while the households selected for EUSILC has been partially included in the MS
- The implementation of an integrated *LFS* and *EUSILC* sampling strategies should also take account of the *complexity caused by their household rotation groups schemes*
- It *still under study how to better integrate all these surveys* with the Census MS, taking into account the estimation goals of each survey, under the constrain due to the statistical burden.

CSSIS framework 1

An integrated sampling strategy can allow to generate an *Integrated System* in which several different blocks, defined by subsets of units and available variables, can be linked and properly treated in the estimation phase.

- *First block* is defined considering *all units in the PR and the register variables*: aggregated values are obtained summing up all the register information at the desired level of granularity
- *Second block* is given by *intersection between MS and PR*. A subset of units of PR for which besides the register information are available also survey variables collected by MS. The *register variables can be used as auxiliary information*. Estimating via calibration methods or model based estimators for MS target variables.

CSSIS framework 2

- *Other blocks* can be defined considering the *intersection between MS and PR and each of second phase sample (module)* drawn from the MS.
- Each of them is given by a *subset of units belonging* to *PR* for which besides the register information also survey variables collected by *MS* and in each single module of *social survey's* variables are available.
- Also in this case, estimating via *calibration methods* or *model-based estimators* of target variables collected by MS and each single module can be used.
- Moreover *design and model based projection estimators* or estimators based on micro and macro integration.

Conclusions 1

The estimates of interest for each table, derived from the above blocks, can be computed by means of

- *design-based method* as long as the *sample is sufficiently large* to yield reliable estimates.
- *small area methods* should be applied when the *sample size associated with a very detailed table is not sufficiently large* to produce reliable direct estimates

Conclusions 2

In this case it is necessary to better investigate how small area methods can be used within this framework considering that:

- an overall *consistency* among the different *detailed* and *marginal* tables that need to be produced for census and social surveys must be guaranteed.
- *multivariate estimation methods* can better integrate the available information, *modelling the correlations among targets variables*.
- *multiple random effects* in *multivariate projection type estimators* can allow to consider marginal effects into the estimation process.
- the random effects can be used to *model time and/spatial correlations* useful for *out of sample domains* (NSR municipalities).