

The new role of sample surveys in official statistics

Giorgio Alleva

Presidente dell'Istituto Nazionale di Statistica

1. BACKGROUND

Challenges in the new eco-system of statistical information

The Istat Modernization Programme

The Integrated System of Statistical register (ISSR)

2. THE NEW ROLE OF SURVEYS AND THE METHODOLOGICAL CHALLENGES

The new challenges of the traditional role

The new role for the Integrated System of Statistical Registers (ISSR)

Supporting the statistical process using new data sources

3. THE NEW CENSUS APPROACHES

4. CONCLUSIONS

1. Background

Challenges in the new eco-system of statistical information

- ❑ Change in the demand for statistical information
- ❑ New phenomena (e.g. globalization) hard to capture with traditional surveys
- ❑ Hard to reach populations (mobile populations, immigrations)
- ❑ Wealth of information, including unstructured information
- ❑ Availability of new methodological and technological tools
- ❑ International best practices
- ❑ Crisis of traditional data collection systems (high costs, response burden, lower response rates)
- ❑ Presence of competitors on the market

The outside world is changing rapidly

Summarizing

Crisis of traditional data collection systems

New **opportunities** from data and digital technologies

Integration as the response to these challenges

“Official statistical offices need to move from the probability sample survey paradigm of the past 75 years to a mixed data source paradigm for the future”

C. Citro (2014)

Istat's Modernization Programme

In order to deal with the new information environment, Istat has launched its Modernization Programme.

Second half of 2014:

Start of Istat's Modernization Programme, in accordance with:

- ❑ **UNECE** - High-level Group on the Modernisation of Official Statistics
- ❑ **European Statistical System** commitment to Vision 2020

January, 2016:

Official approval of Istat's Modernization Programme by the Governing Board

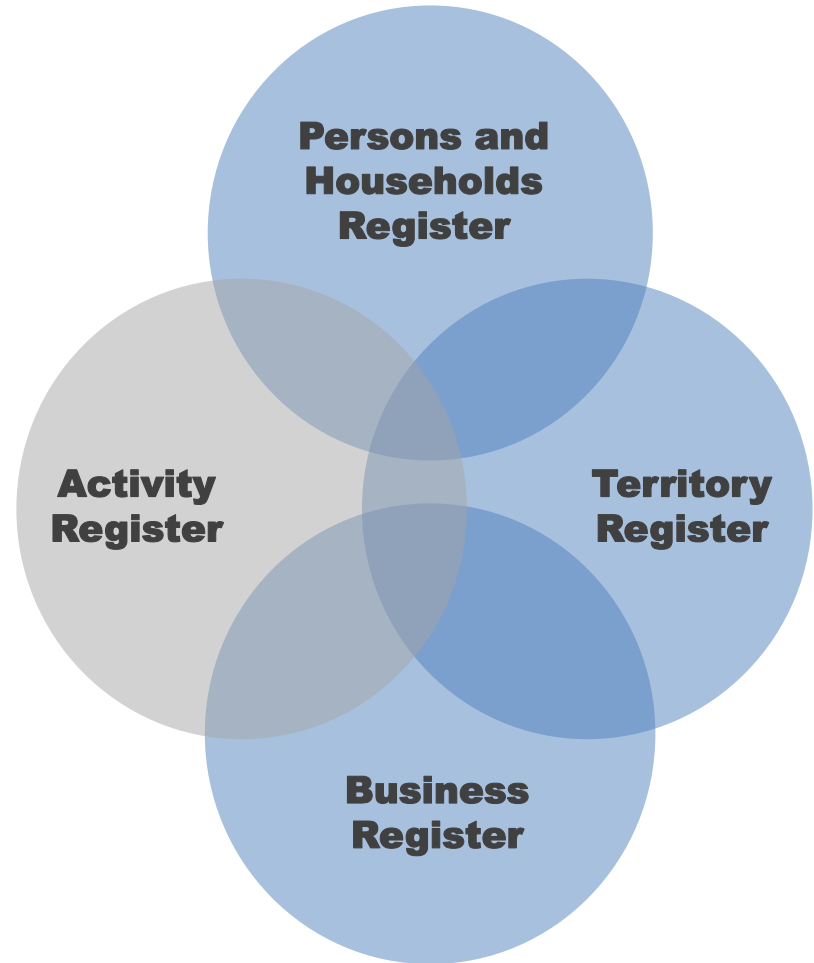


Integrated System of Statistical Registers

Single logical environment to support the consistency of statistical production processes and improve outputs for users

Consistency in **identification** and **estimation** of units and variables for the system as a whole

New analyses starting from populations in registers



2. The new role of surveys and methodological challenges

The role of surveys in Istat

Traditional role

New challenges of the traditional role

- Observe elusive or hard to reach populations not captured in the ISSRs
- Integrated survey framework

Enhancing quality and contents of the ISSRs

- Units
- Variables
- Census

Supporting the statistical process using new data sources

New challenges of the traditional role

Elusive or hard to reach populations:

- ❑ Homeless,
- ❑ Immigrants,
- ❑ Nomadic livestock's (especially in development countries).

Many of this situations can be described as follows

Critical situation	Methodological Challenge : But
1. There is no unique frame for the population	But there are many frames which jointly could cover the target population
2. The units in the frame are not the units of the target populations	But the frame units are related to the target units by a link

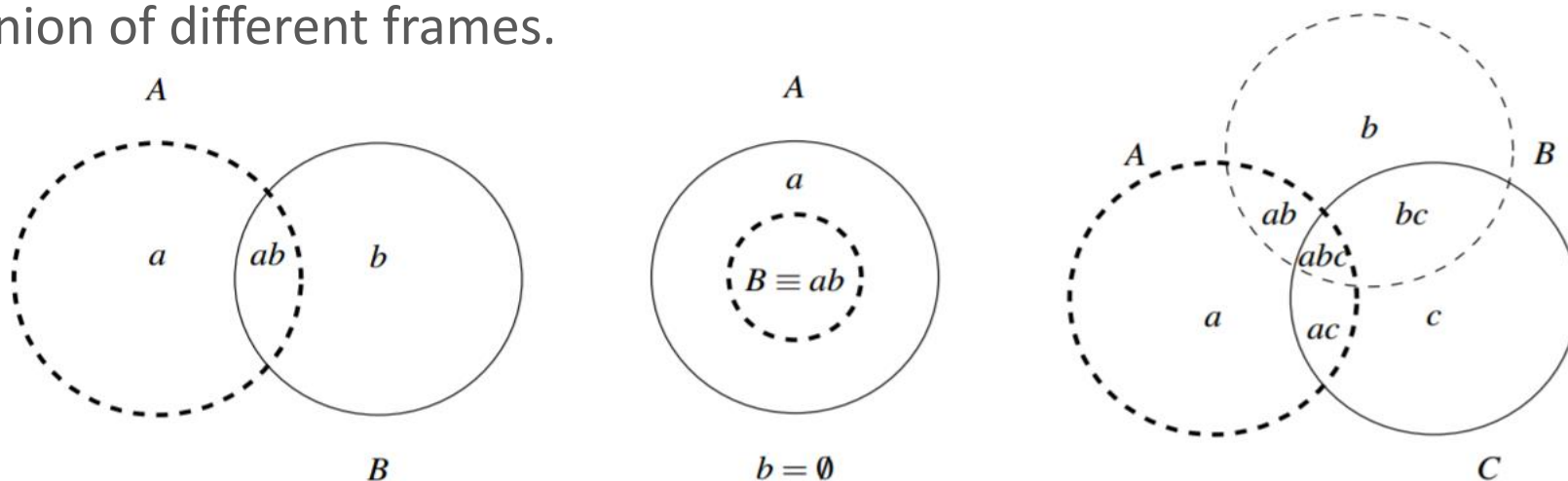
Both challenges are connected with the treatment of multiplicity in surveys

Possible solutions:

- ❑ **Multiple frames** (Singh and Mecatti, 2014)
- ❑ **Indirect sampling** (Lavallée 2007)
- ❑ **Other even non probabilistic techniques** (Statistics Canada, 2014)

Multiple frames

Multiple frames (Mecatti and Singh, 2014): the population is covered by the union of different frames.



“Lists may be complete or incomplete, and may be overlapping with unknown amounts of overlap. MF surveys are often suggested for improving coverage of surveys about difficult-to-sample populations such as elusive and hidden populations as well as rare populations for which a single frame might even be non-existent...”

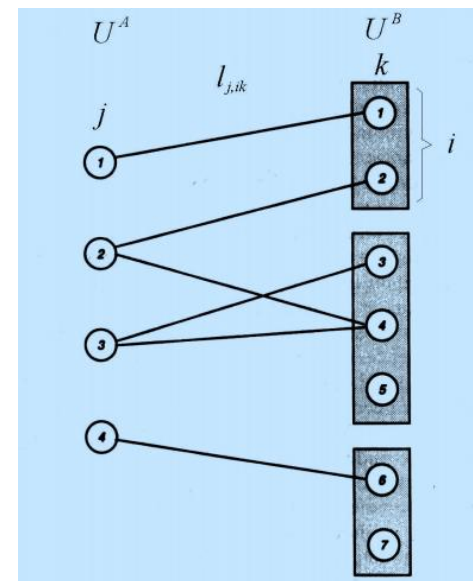
MF surveys can be cost effective even if a complete frame exists

For instance, in agricultural surveys

New challenges of the traditional role

Elusive or hard to reach populations: Indirect sampling represents a way to treat uniformly different sampling techniques introduced in literature

- ❑ **Fair share method**, for the reconstruction of the longitudinal households in the context of longitudinal surveys (Ernst, Hubble, Judkins, 1984; Ernst, 1989).
- ❑ **Network sampling**, for social surveys, particularly useful in defining populations that are rare or difficult to identify. The notion of network often corresponds to a range or set of contacts (Sanders, Kalsbeek, 1990).
- ❑ **Adaptive cluster sampling** (Thompson, 2002), discussed sampling methods to use for populations that are difficult to reach because there is no sampling frame or because these populations are migratory or elusive.
- ❑ **Snowball sampling** (Goodman, 1961).



New challenges of the traditional role

An innovative proposal is the **Integrated Survey Framework**

This deepens the problem of defining a survey strategy in the multivariate case (more than one parameter to be estimated) where the variables of interest are observed in different target populations related to each other by:

- ❑ formal rules,
- ❑ contingent dependencies,
- ❑ relationships created for the pursuit of common purposes.

In order to get insights to given phenomena, the **observation** has to be carried out in **an integrated way**, implying that units of a given population have to be **observed jointly** with the related units of the other population.

This background is typical in agricultural surveys where statistical units refer to **rural households, farms and land parcels** that are related to each other.

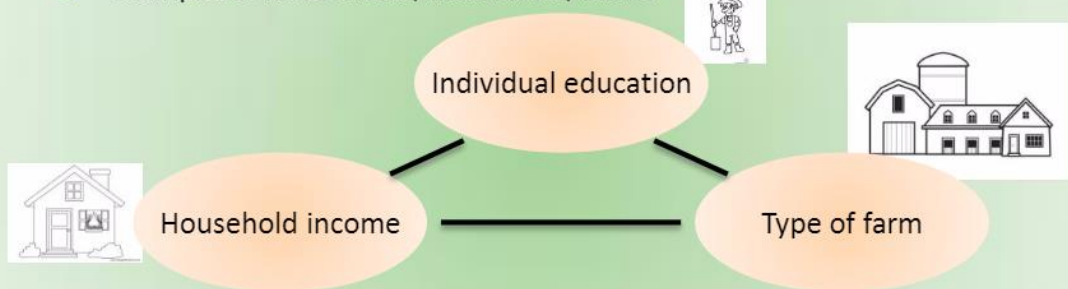
Joint use of multiple frames and indirect sampling (Fao, 2014)

An example from the agriculture world

Farm Holders ↔ Rural households

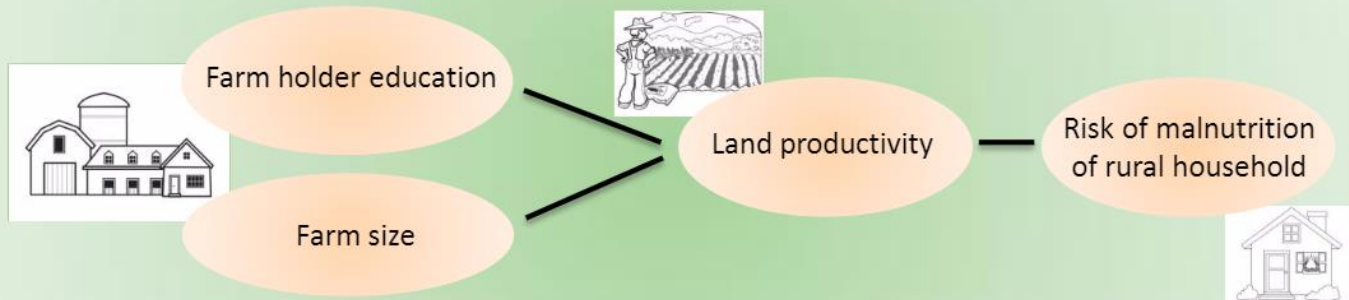
○ Examples1: Households, individuals, farms

Correlations



○ Example 2: Farms, Land parcels, Rural households

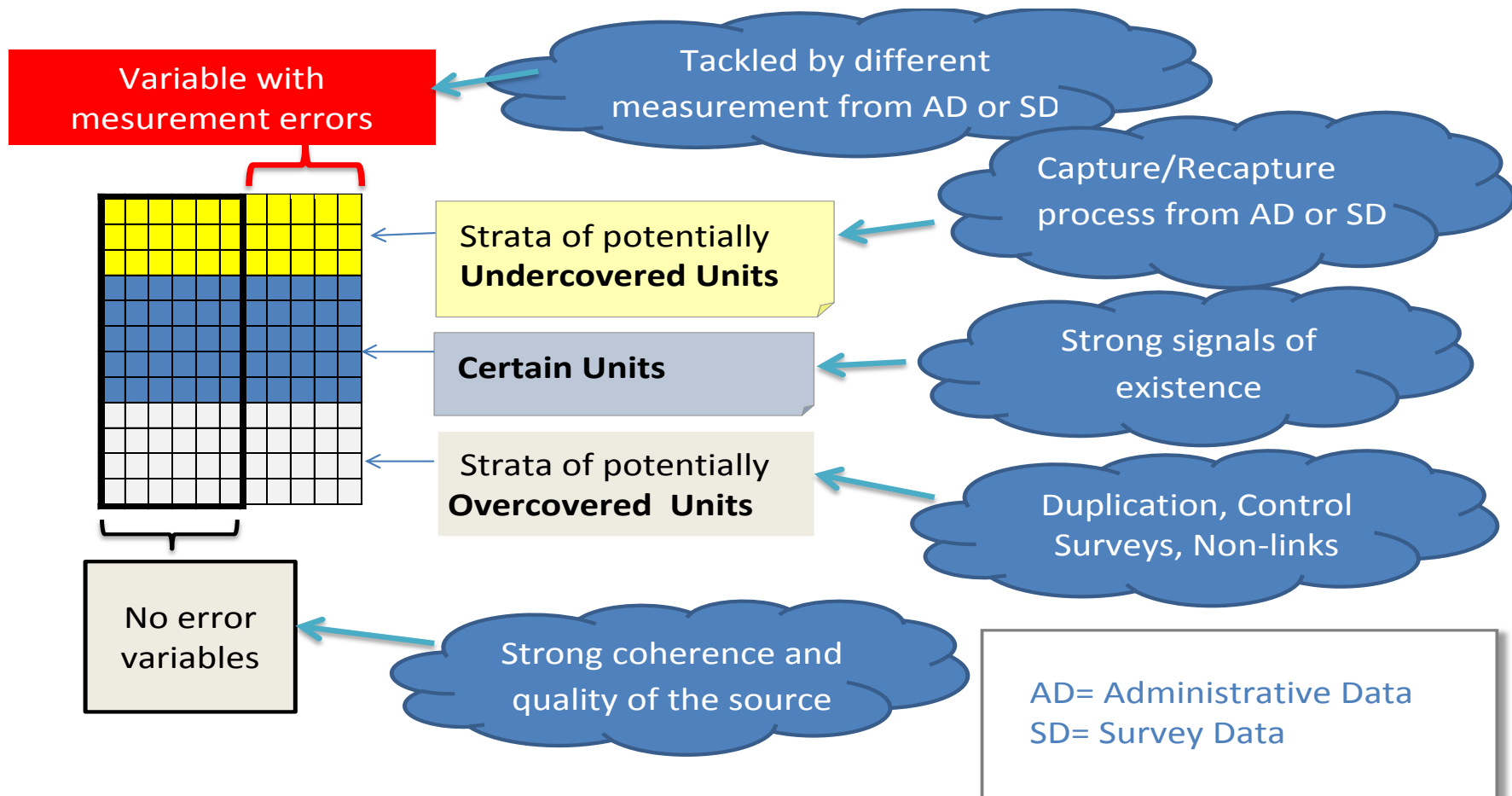
Conditional distributions



New role for the ISSRs

Coverage/undercoverage of units in the registers

Typical situation of a statistical register



How to deal with uncertainty

A strategic issue for NSIs (responsibility and transparency)

1. **Simply ignore** (traditional solution): simple but risk of severe bias.
2. **Evaluate** the sources of errors in order to inform the users (Eg. PES): lack of consistency of different production lines, 2 lines of production.
3. **Identify the improvements** in the process for building the registers: continuous improvement/ the identified bias is still present.
4. **Correct the bias** in the main estimates (External Benchmarks) without modifying the register: lack of consistency of different production lines, 2 lines of production.
5. **Modify units and variables in the register to correct the bias** in the main estimates: consistency of different outputs, relevant computable efforts, some outputs may be inaccurate (transfer the uncertainty to the microdata level).

Pfefferman equation

Let consider a subgroup g in the register with a specific behaviour with respect to **undercoverage/overcoverage**

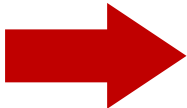
N_g true number of people living in sub-group g

R_g number of people registered to sub-group g

$P_{g,L|R}$ proportion of people living in sub-group g among those registered in the sub-group

$P_{g,R|L}$ proportion of people registered to sub-group g among those living in the sub-group

$$N_g \times P_{g,R|L} = R_g \times P_{g,L|R} \Leftrightarrow N_g = R_g \times \frac{P_{g,L|R}}{P_{g,R|L}}$$


$$\hat{N}_g = R_g \times \frac{\hat{P}_{g,L|R}}{\hat{P}_{g,R|L}}$$

Pfefferman D., 2015

Methodological issues and challenges in the production of Official Statistics JSSM

Estimates

$$\hat{P}_{g,R|L} \leftarrow$$

- ❖ A capture/recapture process of the different administrative archives
- ❖ Capture/recapture surveys for estimating undercoverage

$$\hat{P}_{g,L|R} \leftarrow$$

- ❖ Survey on over-coverage as in the Israelian Census (Pfefferman, 2015)
- ❖ The current survey data (which in the contact phase may establish the eligibility of the units) as studied for the *Permanent Italian Census*

Solution

$$\text{Unit weight: } d_k = \frac{\hat{P}_{g,L|R}}{\hat{P}_{g,R|L}} \quad \text{for } k = 1, \dots, R_g$$

d_k is a statistic with a computable variance

$$\hat{N}_g = \sum_{k=1}^{R_g} d_k$$

$$\text{Var}(\hat{N}_g | R_g) = R_g^2 \text{Var}(d_k) = R_g^2 \left[\frac{\text{Var}(\hat{P}_{g,L|R})}{E(\hat{P}_{g,R|L})^2} + \frac{E(\hat{P}_{g,L|R})^2}{E(\hat{P}_{g,R|L})^4} \times \text{Var}(\hat{P}_{g,L|R}) \right]$$

The same reasoning can be applied to compute the estimates and the related variance for any domain (different from g).

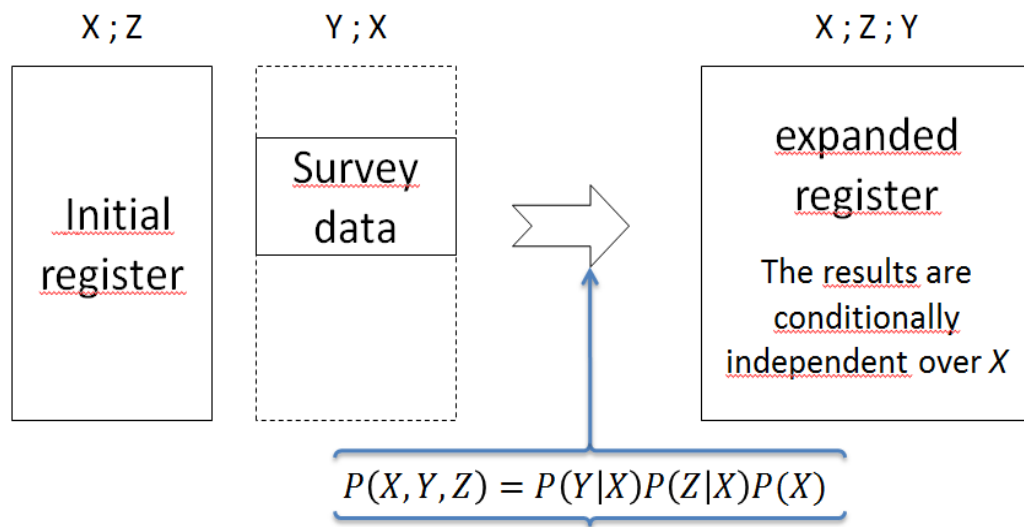
Pfefferman D., 2015

Methodological issues and challenges in the production of Official Statistics JSSM

New role for the ISSRs (continued)

Variables: Integrate variables partially included in the ISSRs

Survey data can be exploited via **model prediction** or with **projection estimator** to expand the variables in the register



This operation is convenient but could be **problematic** with respect to the inference on $P(Y, Z)$:
Risk of introducing synthetic relationships

Kim and Rao, 2012; Righi 2014; Fao, 2014

Survey data can be also exploited via **spree** (design or model based) **estimator** to obtain a timely model based version of the register

More effective sampling designs

The richness of auxiliary information deriving from registers may be exploited at the design phase to obtain more effective designs via two powerful unifying methodologies:

1. **Balanced samples** (Deville and Tillé, 2005)

- ❑ They allow to select sample collections which reproduce the known distribution of auxiliary variables from registers.
- ❑ More efficient sampling designs and reduction of model bias.
- ❑ Several customary sampling designs may be considered as special cases of balanced sampling.

2. **Unified framework for optimal sampling** (Falorsi and Righi, 2015)

- ❑ It allows defining the optimal inclusion probabilities for a variety of survey contexts in which disseminating survey estimates of pre-established accuracy for a multiplicity of both variables and domains of interest is required.
- ❑ The framework can define standard stratified or incomplete stratified sampling designs.
- ❑ The target variables are unknown, but can be predicted with suitable super-population models. The algorithm takes properly into account this model uncertainty.

Supporting the statistical process using new data sources

Traditional data (survey, census, administrative data) seems to be fundamental for setting up a framework for:

- ❑ Evaluating the quality of the Big Data statistics.
- ❑ Improving the quality of the official statistics.
- ❑ Selectivity and representativeness can be evaluated by reconciling data from the two independent sources, Big Data and sample surveys.
- ❑ Predicting micro data in the new source. Observing the target variable with a sample survey, a prediction based on supervised model approach can be carried out.

3. The new census approaches

The Census and Social Surveys Integrated System (CSSIS)

The **CSSIS** consists in two phases:

1. the phase supporting the aim of the Population Census;
2. the phase supporting the objectives of the Social Surveys.

The first phase

It's planned to be held, yearly, in Autumn (starting from 2018), with the aims of:

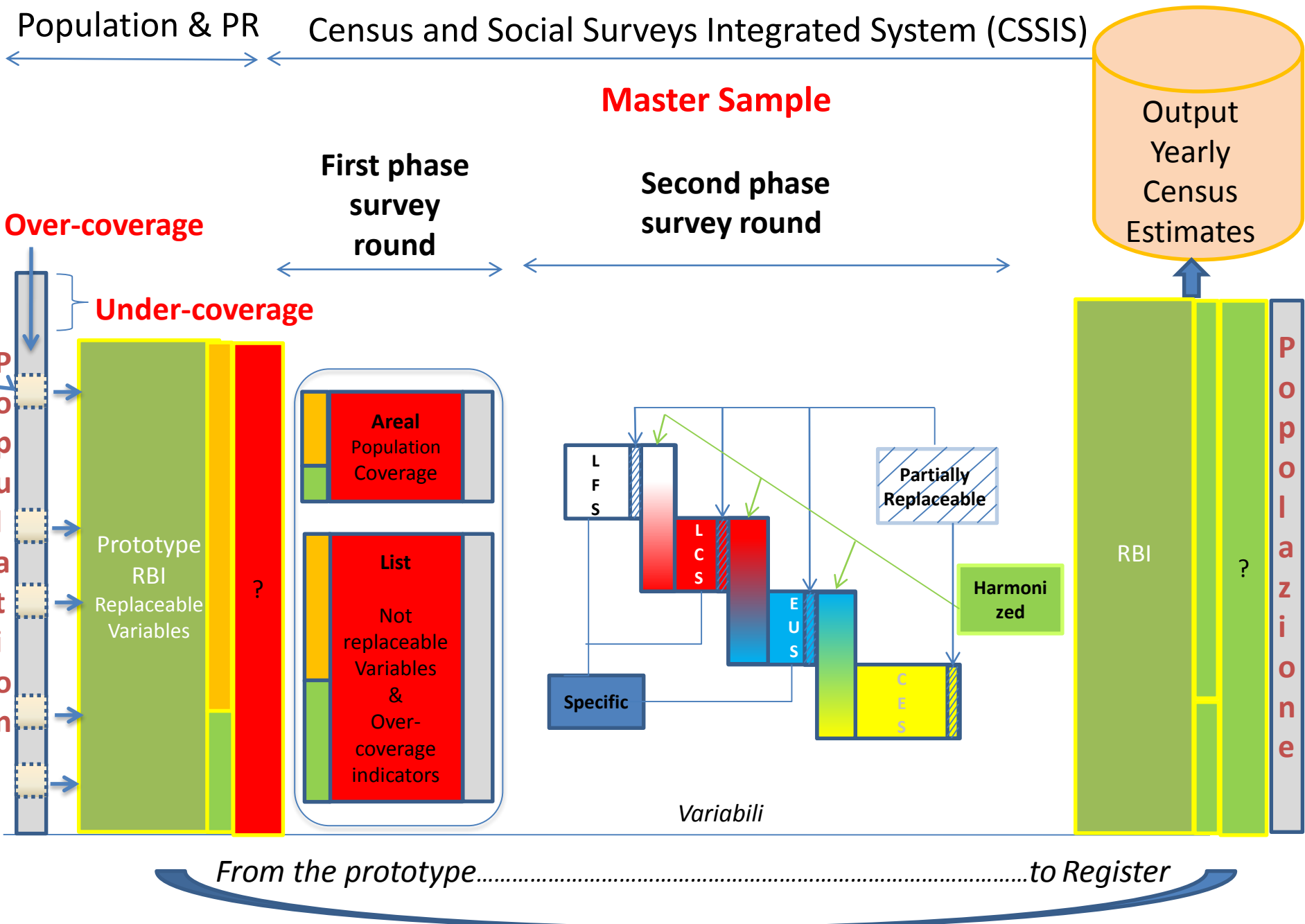
- ❑ **correcting** for under and over coverage the Base Register of individuals improving the quality of the population totals produced;
- ❑ **collecting** the information for not replaceable variables by means of an ad hoc sample survey (Master Sample)

The first phase can be carried out following two different schemes: the component based on an area sample (A) and the component based on a list sample (L).

The second phase

It takes place throughout the year following that of the first phase (i.e. from January 2019) sample households are selected as a sub-sample of those already involved in the first phase sample

The Census and Social Surveys Integrated System



The Census and Social Surveys Integrated System (CSSIS)

This scheme may ensure

1. a framework of coherence between census production annual statistics and that, of the same type, produced annually by social surveys exploiting the availability (at unit-level) between the core variables observed in the first phase and the same variables observed in the second phase
2. the progressive use of less expensive CAWI and CATI techniques, favored by the increased availability of contact information (email and telephone) required for all the first-phase respondents
3. possibly more efficient estimates than those produced with pre-existing estimation processes
 - ❑ through the exploitation of the observed variables (not available from registers) on the first phase MS as post-stratification variables
 - ❑ clustering effect can be reduced for the increase of municipalities potentially involved with the increase of CAWI and CATI
4. this will allow cost reduction of social surveys

4. Conclusions

Conclusions

Relevant methodological advances allow to deal with new challenges through unified and generalised approaches

- **Multiple frames**
- **Indirect sampling**
- **Balanced samples**
- **Unified framework for optimal sampling**

Conclusions

Surveys **change their role**, being mainly focused on feeding the ISSR.

- ❑ Estimates of **error components**:
 - Coverage
 - Specification
 - Measurement.
- ❑ Estimation of **non-measured** or non-measurable **variables from administrative sources**.
- ❑ Estimates of **unmeasurable associations** from administrative sources.
- ❑ **Smaller surveys**, but more complex . A portion of the savings must be **reinvested in quality**.
- ❑ Strong need to **involve the scientific community** for studying together this big change
- ❑ New role for experimental statistics

The new role of sample surveys in official statistics

Giorgio Alleva

Presidente dell'Istituto Nazionale di Statistica