Istat | Istituto Nazionale di Statistica

3rd Workshop on Methodologies for Official Statistics

Rome - December 4-5, 2024

Optimization of Surveys: The “Integrated Census and Social Surveys System” Project

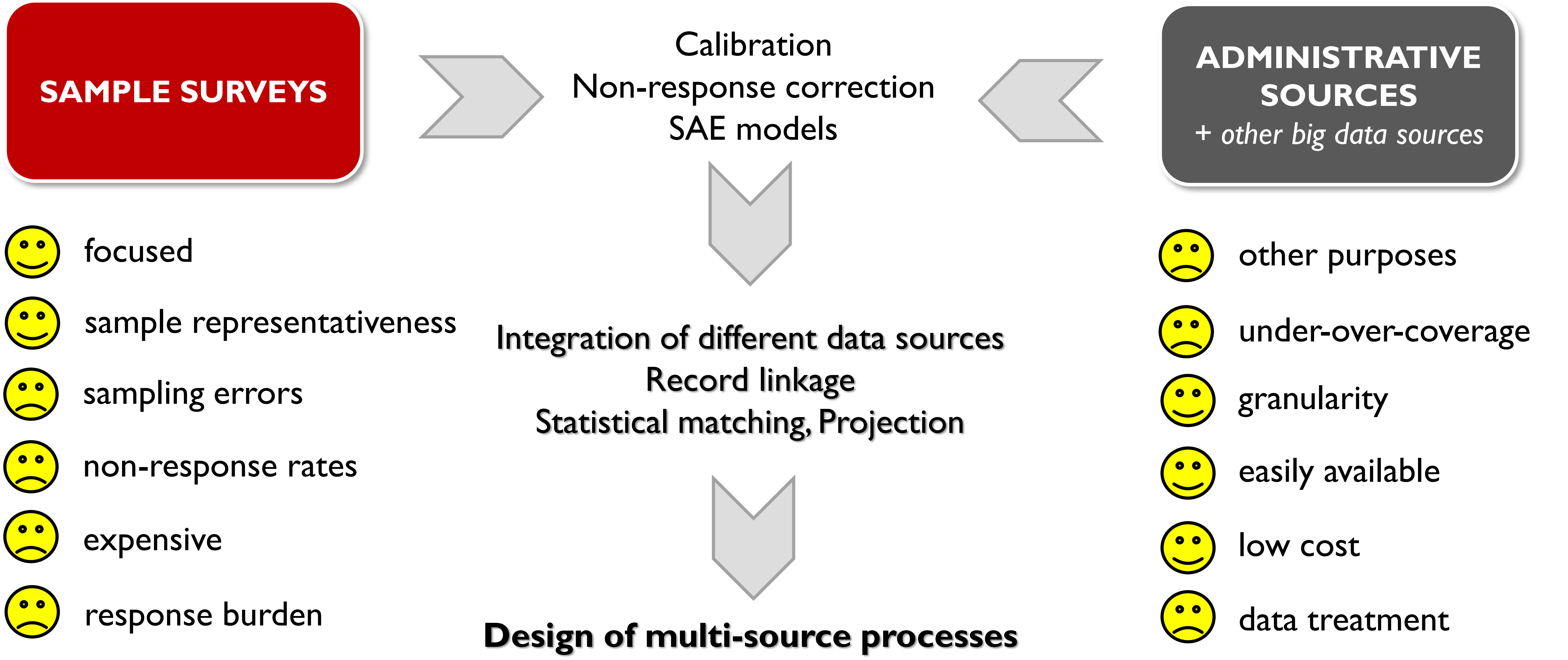
Silvia Loriga (siloriga@istat.it), Claudia De Vitiis, Stefano Falorsi, Alessio Guandalini, Francesca Inglese, Matteo Mazziotta, Federica Piersimoni, Rita Ranaldi, Monica Russo, Marco Dionisio Terribili (Istat)

Roberto Benedetti (Chieti-Pescara University)

Outline

- Context
- **The Integrated Census and Social Surveys System (SICIS)**
  - Aim of the project
  - The SICIS initial design
  - Advantages of a two-phases sampling scheme
  - Current implementation of SICIS
- Two experimental studies
  - **Two phases sampling scheme**
    - Simulated scenarios
    - Results
  - **Spatially balanced sample selection**
    - Results and further developments
- Final remarks

Context



SAMPLE SURVEYS

- ☺ focused
- ☺ sample representativeness
- ☹ sampling errors
- ☹ non-response rates
- ☹ expensive
- ☹ response burden

Calibration
Non-response correction
SAE models

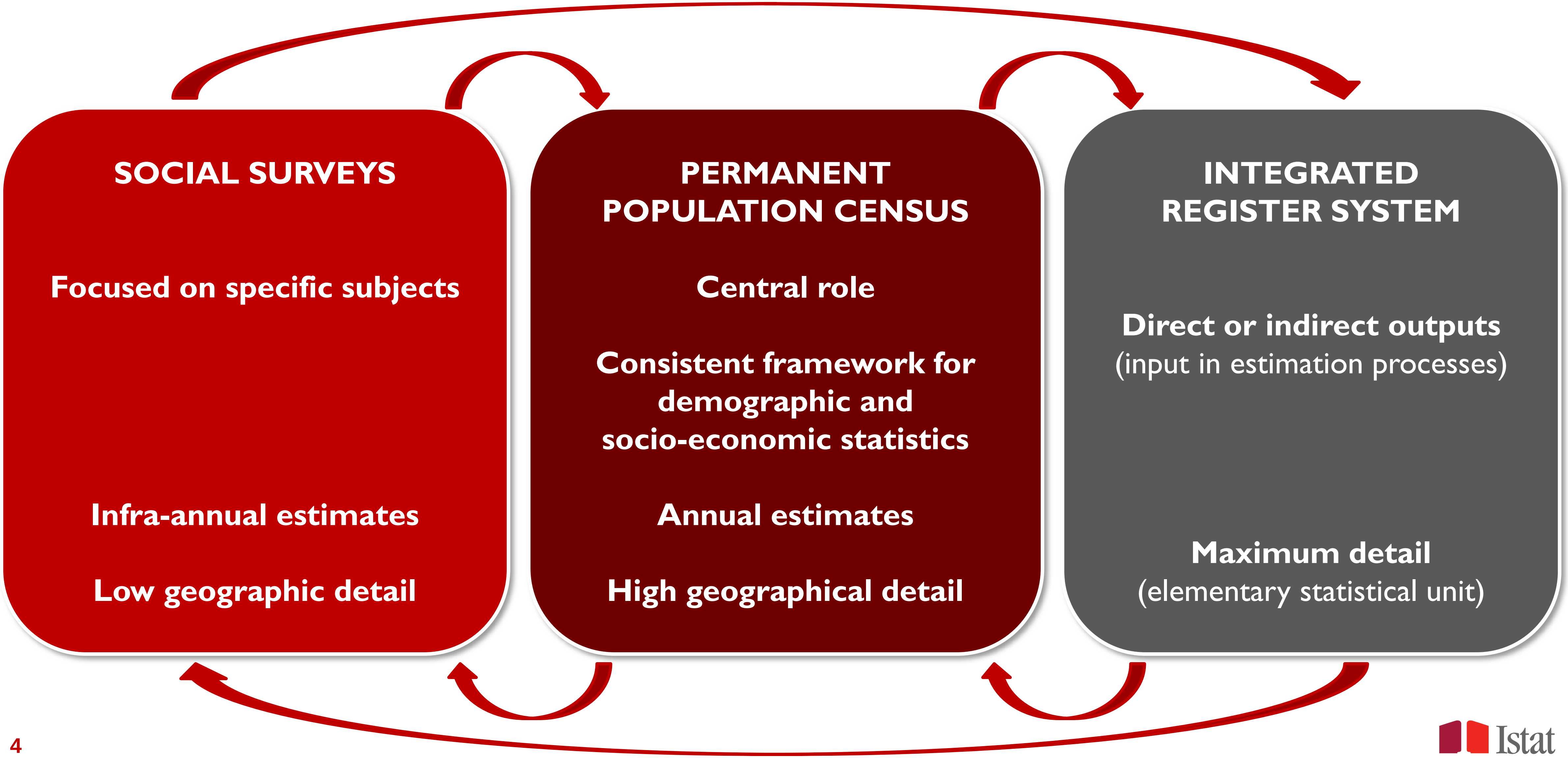
Integration of different data sources
Record linkage
Statistical matching, Projection

Design of multi-source processes

ADMINISTRATIVE SOURCES
+ other big data sources

- ☹ other purposes
- ☹ under-over-coverage
- ☺ granularity
- ☺ easily available
- ☺ low cost
- ☹ data treatment

The Integrated Census and Social Surveys System (SICIS)



SOCIAL SURVEYS

Focused on specific subjects

Infra-annual estimates

Low geographic detail

PERMANENT POPULATION CENSUS

Central role

Consistent framework for demographic and socio-economic statistics

Annual estimates

High geographical detail

INTEGRATED REGISTER SYSTEM

Direct or indirect outputs
(input in estimation processes)

Maximum detail
(elementary statistical unit)

The Integrated Census and Social Surveys System (SICIS)

The core idea is the **integrated design of sampling strategies for Social Surveys**

- sampling schemes
- direct and indirect estimators
- small-area estimators

aiming to

- optimize the accuracy
- ensuring the desired level of granularity
- improve the coherence

jointly considering the strategic choices for

- data collection (survey techniques, response burden)
- thematic objectives (harmonization of questionnaires)

Aim of the project

Why initiate a study to improve the efficiency of social surveys?

Aren’t the sampling designs and estimators in use already chosen to maximize efficiency?

Why should a reconsideration of the survey techniques be necessary?

Overcoming the stove-pipe approach towards a systemic perspective

- Between surveys
- Between phases of a survey

Evaluate potential improvements for sampling design, estimation methodology, and survey techniques

- Considering these aspects jointly
- To improve efficiency
- To correct biases

Exploitation of auxiliary information

- Coordinating Social Surveys with each other and with Population Census
- Leveraging information from the Integrated Register System

Analysis and solution of some identified issues on data collection

- Undercoverage of telephone lists
- Increase in total non-responses

The SICIS initial design

The initial design of SICIS consisted in a modular approach, based on the two-phases sampling scheme:

1st phase: A general module to collect the target variables of Population Census, primarily demographic and social variables

The Population Census (L survey) based on a large yearly sample (two-stages: municipalities-households) referred to as the Master Sample

2nd phase: Specific modules to observe the target variables of other Social Surveys

The other Social Surveys based on sub-samples of municipalities and households selected from the Master Sample

Other pieces in this integrated system:

- Integration with registers
- Projection-type estimators

Advantages of a two-phases sampling scheme

Leverage of information collected in the 1st phase as auxiliary variables in the 2nd phase

- balanced sampling
- calibration
- non-response correction
- small area estimators

Leverage of the repeated observation of the same variable over the same units in both the phases

- reconciliation techniques
- measurement error models
- improving coherence

Leverage of the home/mobile phone number or email collected in the 1st phase

- to conduct CATI or CAWI in the 2° phase

Leverage of structural variables collected in the 1st phase

- lightening the questionnaire in the 2nd phase (asking for confirmation)

main disadvantage: response burden

Current implementation of SICIS

The initial design of SICIS has been partially applied:

Labour Force Survey (LFS)

overlapping with MS only at the first selection stage (municipalities, except for the smallest ones)

Aspects of Daily Life Survey (AVQ)

- 2018-2022 - selected as sub-sample of the MS, both for the first-stage (municipalities) and for the second-stage units (households)
- Since 2023 - overlapping with MS only at the first selection stage (for purely operational reasons: the use of the same tablets)

EuSilc

selected as sub-sample of the MS

(to exploit telephone contact from MS for adopting Cati technique)

- No exploitation of the overlap during the estimation phase

need to evaluate the current implementation and rethink the entire system

Two experimental studies

1. Two phases sampling scheme

to compare the efficiency of various scenarios for integrating the 1st and 2nd phase samples

Three surveys:

- The Population Census (L survey: Master Sample) → 1st phase
- The Labour Force Survey (LFS)
- The Aspects of Daily Life survey (AVQ)

2nd phase

2. Spatially balanced sample selection

to evaluate the efficiency gain arising from a spatially balanced sampling at the first stage (municipalities)

- A two stages sampling design (similar to LFS)
- spatially balanced sample = balanced with respect to the available auxiliary variables
  maximally spatially distributed

Experiment on Two-Phases Sampling Design

- 3 scenarios have been simulated

  Scenario S1: No Integration

  Scenario S2A: Integrated Designs only at the 1st stage (municipalities)

  Scenario S2B: Integrated Designs at the 1st and 2nd stages (municipalities and households)

- 3 estimators:

  Horvitz-Thompson

  Calibration 1 – only demographic auxiliary information (Cal1)

  Calibration 2 – for LFS and AVQ demographic vars + education level from MS (Cal2)

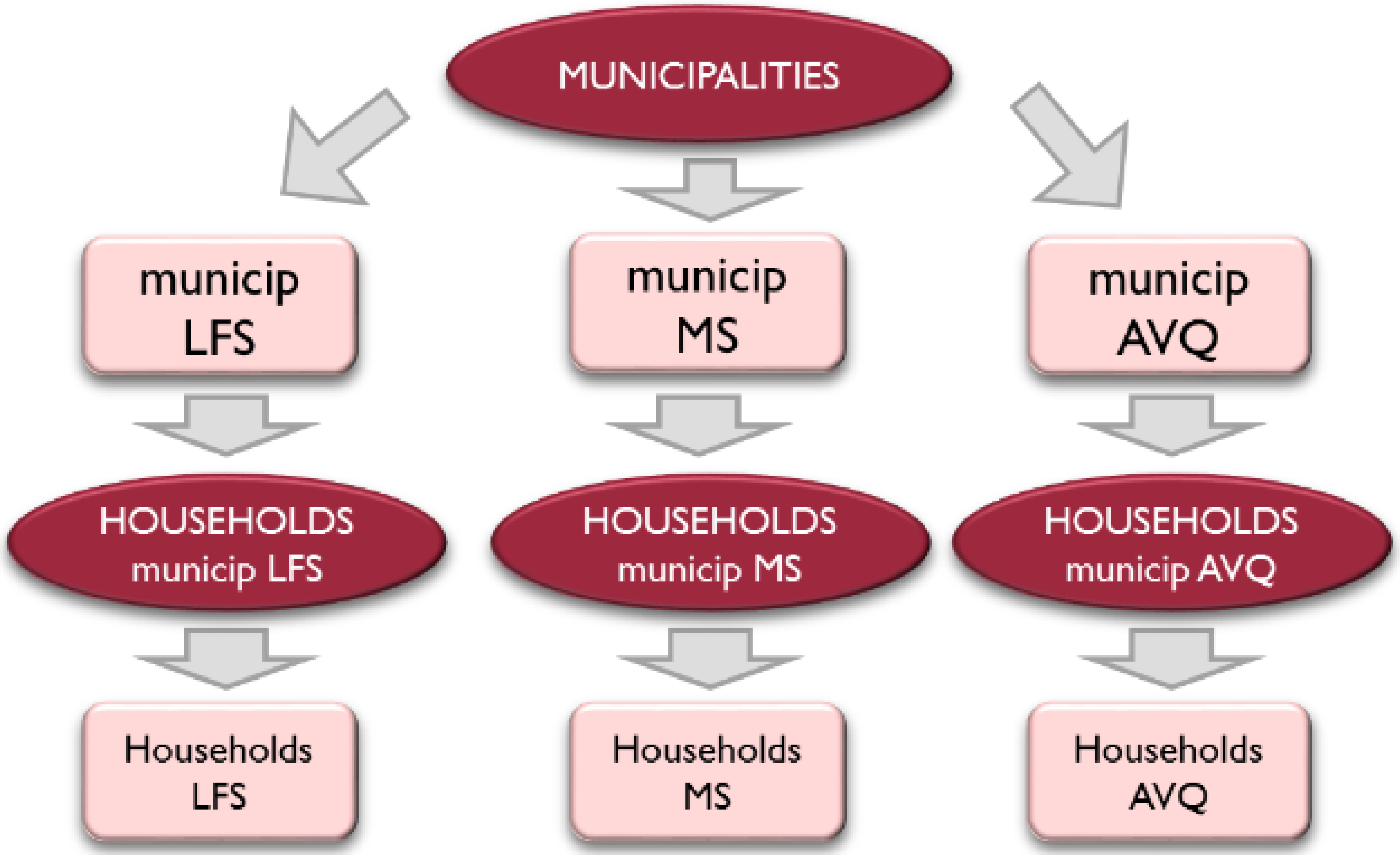
  The target variable for estimation is the employment status

  Hps: Full response

  No measurement errors

This experimental study has been conducted by Alessio Guandalini, Marco Dionisio Terribili

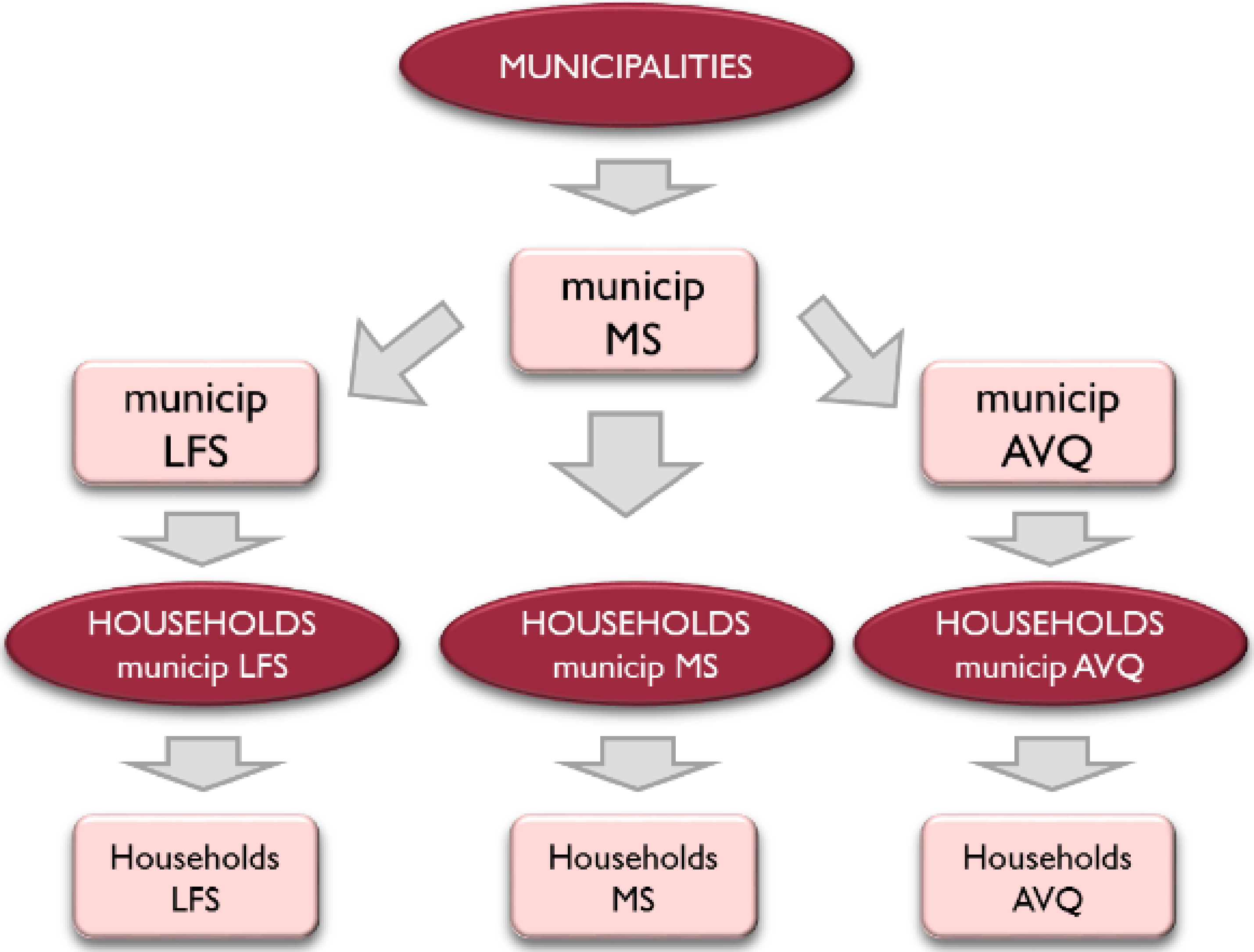
Scenario S1



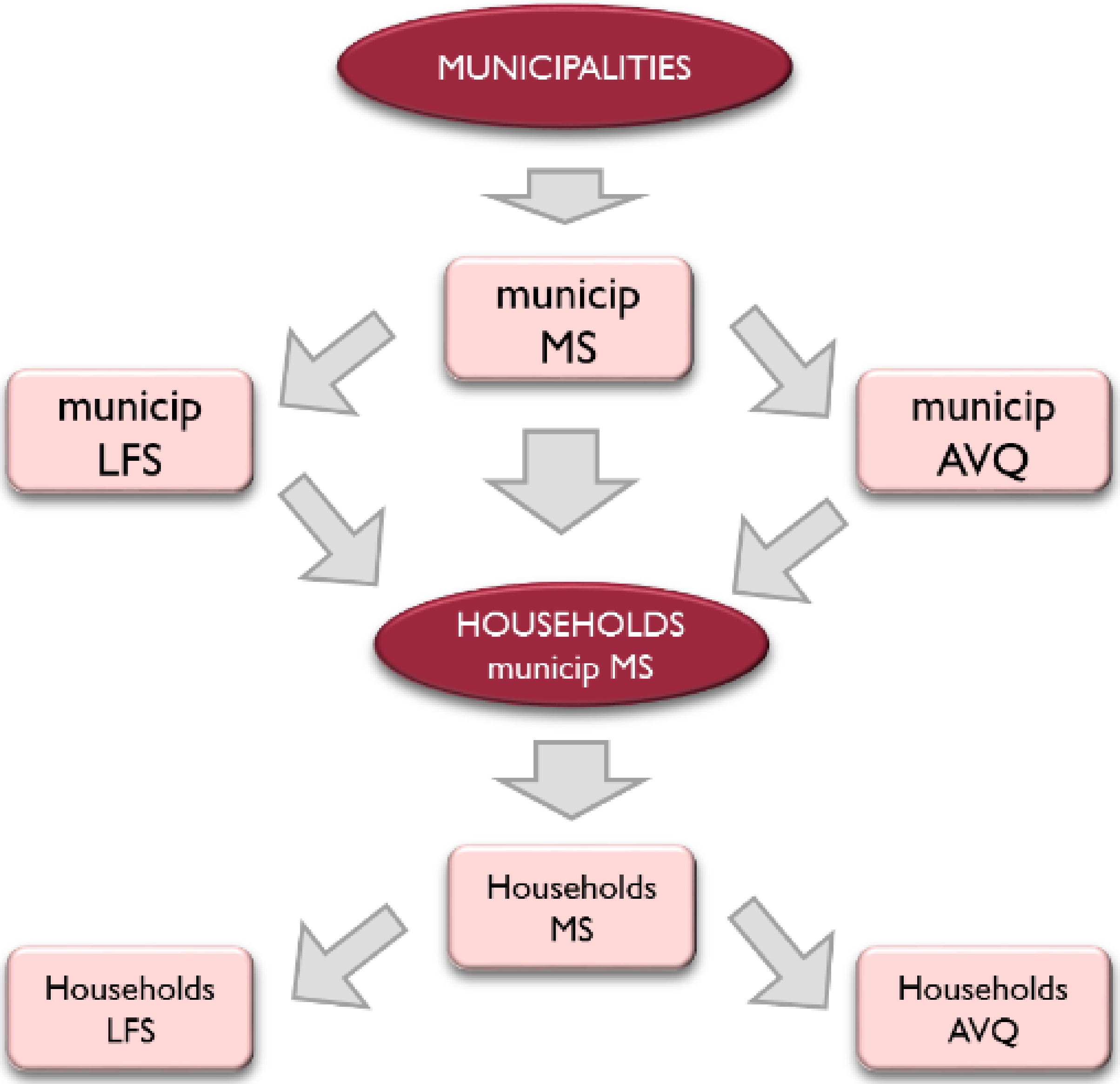
For the three surveys:
two-stages sample
(municipalities - households)
stratification of municipalities
by demographic size

- at provincial level for MS and LFS
- for AVQ, it is at regional level by the type of municipality (six types).

Scenario S2A



Scenario S2B



Simulation plan

3 Italian regions (Piemonte, Lazio, and Basilicata)

500 replications for each scenario

R software

Results 1

Estimated percentage coefficient of variation (for the total number of employed individuals in the 3 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

Modest overall impact of the integration with the MS (for both LFS and AVQ)

Results 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	MS.EstEff
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

Design effect: >3 for LFS; >2 for AVQ (ind. strat); ~1.3 for AVQ (dep. strat) similar to S1; note: ~2.5 for MS

Cal1 is able to compensate the Design effect

Cal2 further improves

Further results and future developments

Further results

○ Increasing the overlap between MS and social surveys municipalities improve the efficiency

○ Statistical burden has been taken into account in terms of overlapping of municipalities and households

Should be considered that

○ Different data collection techniques and questionnaires may have an impact in terms of measurement errors

Future developments already planned

○ Assessing the impact of non-response

- 3 non-response processes based on the indicators observed in the MS, LFS, and AVQ surveys
- auxiliary variables from the Integrated Register System (IRS)

Experiment on Spatially Balanced Sampling Design

Taking into account the spatial dependence of statistical units in sampling design and estimation improves the accuracy of the estimates

Maximizing spatial distribution to capture the spatial heterogeneity of the population of interest

Municipality indicators from

- Archimede database “Socio-economic Conditions of Households”

  Income variables

  Labor market precariousness

- Population Census estimates

  Demographic variables and family structure

  Distribution by levels of educational attainment

  Percentage of employed, unemployed, and inactive individuals.

The parameter being estimated is the mean of the variable of interest (Horvitz Thompson estimator)

This experimental study has been conducted by Roberto Benedetti, Federica Piersimoni, Monica Russo

Experiment on Spatially Balanced Sampling Design

The following sampling design were compared:

- STR1: Self Representative (SR) municipalities

  Non Self Representative (NSR) municipalities stratified at sub-provincial level – pps selection

- LPM_STR1 adopting for NSR the same stratification as in STR1
- LPM_PROV adopting for NSR a stratification by provinces
- LPM_REG adopting for NSR a stratification by regions

} spatially balanced designs, implemented using the Local Pivotal Method (LPM)

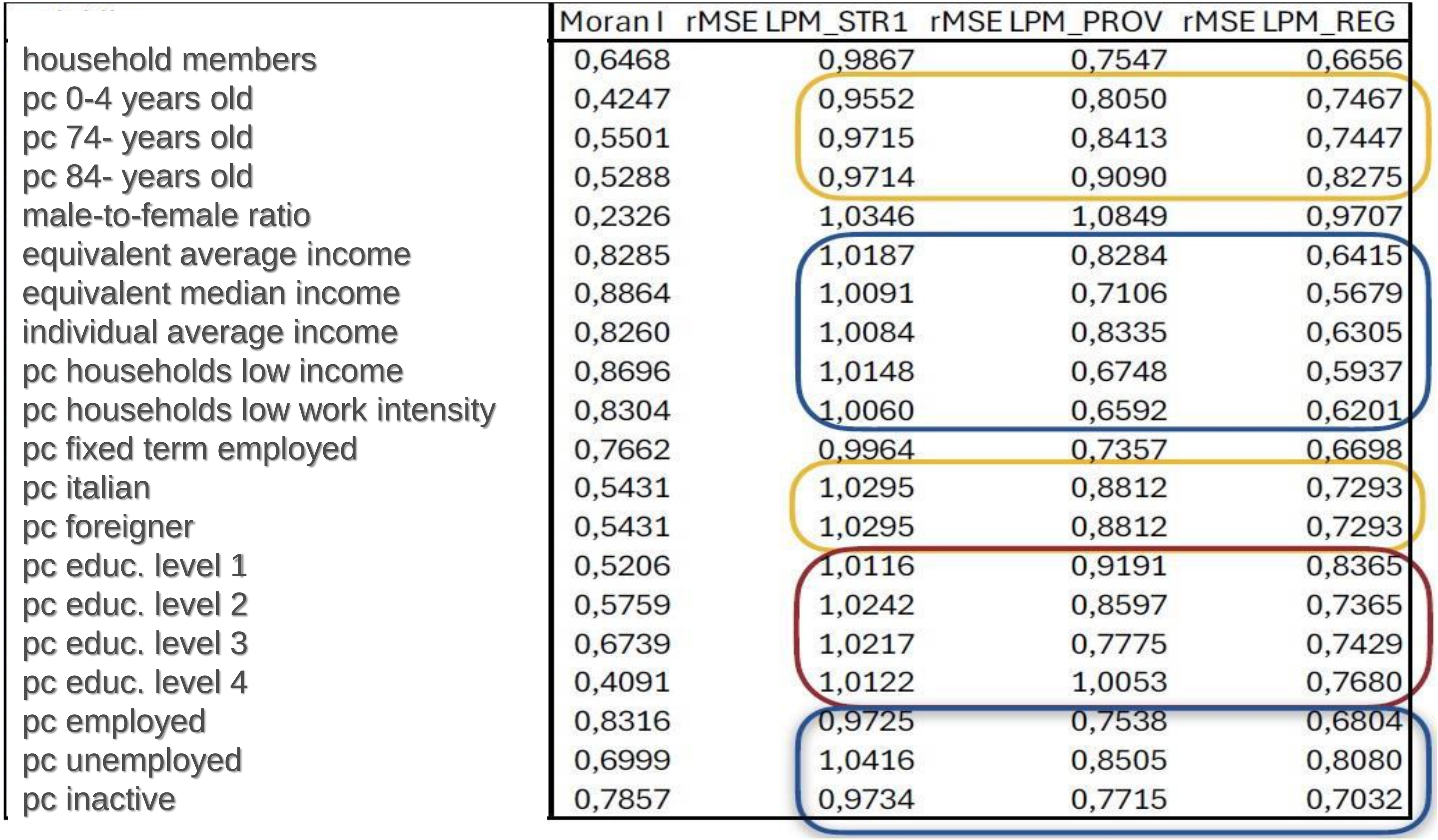
Simulation plan

Entire national territory

10,000 replications for each design

R software + BalancedSampling package for LPM

Results



	Moran I	rMSE LPM_STR1	rMSE LPM_PROV	rMSE LPM_REG
household members	0,6468	0,9867	0,7547	0,6656
pc 0-4 years old	0,4247	0,9552	0,8050	0,7467
pc 74- years old	0,5501	0,9715	0,8413	0,7447
pc 84- years old	0,5288	0,9714	0,9090	0,8275
male-to-female ratio	0,2326	1,0346	1,0849	0,9707
equivalent average income	0,8285	1,0187	0,8284	0,6415
equivalent median income	0,8864	1,0091	0,7106	0,5679
individual average income	0,8260	1,0084	0,8335	0,6305
pc households low income	0,8696	1,0148	0,6748	0,5937
pc households low work intensity	0,8304	1,0060	0,6592	0,6201
pc fixed term employed	0,7662	0,9964	0,7357	0,6698
pc italian	0,5431	1,0295	0,8812	0,7293
pc foreigner	0,5431	1,0295	0,8812	0,7293
pc educ. level 1	0,5206	1,0116	0,9191	0,8365
pc educ. level 2	0,5759	1,0242	0,8597	0,7365
pc educ. level 3	0,6739	1,0217	0,7775	0,7429
pc educ. level 4	0,4091	1,0122	1,0053	0,7680
pc employed	0,8316	0,9725	0,7538	0,6804
pc unemployed	0,6999	1,0416	0,8505	0,8080
pc inactive	0,7857	0,9734	0,7715	0,7032

Results and future developments

Spatially balanced sampling is more efficient (at the 1st stage) compared to the stratified design.

The efficiency gain is greater:

- For variables with a higher spatial autocorrelation (Moran index)
- For designs adopting a less "fine" stratification

In particular, regarding the estimates produced at the national level:

- Income variables show a very high spatial autocorrelation

  For the median equivalent income, the efficiency gain is ~ 30% of the variance with LPM_PROV and ~ 50% with LPM_REG

- Labor market participation variables show fairly high spatial autocorrelation

  For employed and inactive, variance reductions of ~ 25% with LPM_PROV and ~ 30% with LPM_REG

- Demographic and family variables show the lowest Moran index

  Even for these variables, efficiency gains of up to 25% with LPM_REG

An additional advantage of spatially balanced sampling is the coverage of unplanned territorial domains

Future developments regard the evaluation with Calibration estimator (beyond Horvitz Thompson)

Final remarks

The next developments of the project will focus on the main issues emerged in data collection, mainly:

- The increase in non-response rates (lack of representativeness of the samples; bias)
- The need to rethink data collection techniques (ex. crisis of CATI, due to the lack of reliable phone numbers)

The work approach:

- It is believed that methodological and operational strategies have to be studied jointly
- A systemic perspective (for the different surveys) must be adopted
- Currently an important added value of the project comes from the collaboration between the methodological and data collection teams of Istat
- A further step towards adopting a fully systemic approach must also involve collaboration with the teams responsible for data production

References

Australian Bureau of Statistics. (2012). Household Expenditure Survey and Survey of Income and Housing, User Guide. Australia, 2009-10

BEAUMONT, J. F. (2019). Are probability surveys bound to disappear for the production of official statistics? Survey Methodology, June 2020, Vol. 46, No. 1, pp. 1-28.

Benedetti, R., Piersimoni, F. (2015). Sampling spatial units for agricultural surveys. Advances in Spatial Science Series. Springer, Berlin, Heidelberg.

Benedetti, R., Piersimoni, F. (2017). A spatially balanced design with probability function proportional to the within sample distance. Biometrical Journal, 59(5), 1067-1084.

Cuppen, M.D.J., van der Laan, P., & van Nunspeet, W. (2013). Reengineering Dutch Social Surveys: From Single-Purpose Surveys to an Integrated Design. Statistical Journal of the International Association for Official Statistics, 29, 21-29.

D'Alò, M., & Falorsi, S. (2023). Census and social survey integrated system. In Workshop on Methodologies for Official Statistics Proceedings, Session 1, Methodologies for the new censuses, Roma, Italy, 5th – 6th December, pp. 21-30.

Falorsi, S., Falorsi, P.D., Nardelli, V., Righi P. (2025) Defining ad-hoc sampling designs for small area estimation. Journal of Official Statistics, forthcoming.

Falorsi, S., Loriga, S., & Di Zio, M. (2023). Overview of the Istat activities and open problems. In Workshop on Methodologies for Official Statistics Proceedings, Session 2, Methodologies for the new censuses, pp. 51-68.

Grafström, A., & Lisic, J. (2016). BalancedSampling: balanced and spatially balanced sampling. R package version 1.5.2.

GROVES, R. M. (2011). Three eras of survey research. Public Opinion Quarterly, 75, 861-871. (Special 75th Anniversary Issue).

HORVITZ, D. G., & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.

Ioannidis, E., Merkouris, T., Zhang, L.-C., Karlberg, M., Petrakos, M., Reis, F., & Stavropoulos, P. (2016). On a Modular Approach to the Design of Integrated Social Surveys. JOS, 32(2), 259–286.

Jae Kwang Kim, & Rao, J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach. Biometrika, 99(1), 85–100.

Lahiri, P. (2024). Preface to the special issue for papers presented at the 29th Morris Hansen Lecture on the use of nonprobability samples. Survey Methodology, June 2024, Vol. 50, No. 1, pp. 1-2.

LOHR, S. L., & RAGHUNATHAN, T. E. (2017). Combining survey data with other data sources. Statistical Science, 32, 293-312.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. Annals of Applied Statistics, 12, 685-72.

Moran, P.A.P. (1948). The Interpretation of Statistical Maps. Journal of the Royal Statistical Society B, 10, 243–251.

Moran, P.A.P. (1950). Notes on Continuous Stochastic Phenomena. Biometrika, 37(1/2), 17–23.

NEYMAN, J. (1934). On the two different approaches of the representative method: The method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, 97, 558-606.

Rao, J. N. K. (2021). On Making Valid Inferences by Integrating Data from Surveys and Other Sources. Sankhya B83, 242–272.

Reis, F. (2013). Links Between Centralisation of Data Collection and Survey Integration in the Context of the Industrialisation of Statistical Production. Working paper presented at the UNECE Seminar on Statistical Data Collection, 2013.

Smith, P. (2009). Survey Harmonization in Official Household Surveys in the United Kingdom. In Proceedings of the ISI World Statistical Congresses, 16–22 August 2009, Durban, South Africa.

This work is the result of the collaboration of many colleagues

Thanks to all of them

Thank you for the attention!