Estimation strategies with different sources of information

Stefano Falorsi, Andrea Fasulo, Fabrizio Solari
Via Cesare Balbo, 16
Istat, DIRM and DIPS
Rome, Italy
stfalors@istat.it fasulo@istat.it solari@istat.it
DOI: 10.1481/icasVII.2016.f35c

ABSTRACT

Since 2011, Eurostat began a reorganization of EU social statistics. This project has evolved over time up to the final version presented at the meeting of Directors of Social Statistics, held in September 2014.

The model proposed by Eurostat is based on an approach in modules of target variables which, by construction, can be pooled and, where possible, can exploit the use of information measured at different surveys for the construction of the estimates.

Eurostat also presented a roadmap (Eurostat, 2013j) for the implementation of the project which contemplates short, medium and long term studies. The first study focuses on methods for pooling estimates to be made with the overlap of samples on which were recorded the same variables, regardless of the drawings below; in the medium term the study focuses on redesign of sample surveys aimed to optimize sample size and allocation and exploiting the new modular approach; in the long term a final study for the integrated micro-database for social statistics, powered by both surveys and the information from the statistical registers.

This paper presents a possible scenario for the integration of social surveys which arises from a specific strategy associated with a specific sampling design. The whole purpose is to achieve a complete integration of the system of social surveys and ensure maximum integration with the registries system present in National Statistical Institute.

A Montecarlo simulation study using Census 2011 data has been carried out. In the simulation 200 samples has been drawn for each of 4 very important Istat surveys, referring to two regions Trentino-Altoadige and Marche. In particular the surveys considered are the Labour Force survey, the Multipurpose survey, the Eusilc survey and the Consumer Expenditure survey.

Finally an empirical evaluation is performed on different estimators of the labour force characteristics (employed and unemployed), for different domains, computing the traditional Monte Carlo indicators base on the difference with respect to the census values in order to evaluate the empirical performances of estimator in terms of bias and variability.

Keywords: Integration, pooling, projection

1. Introduction

There are several alternative scenarios studied and proposed by Istat for the System for integration of Social Surveys (SINTESY). These scenarios are in line with the Eurostat project of modernization of social surveys, aimed to obtain a complete integration of the social surveys SINTESY could be exploited for the estimation of hypercube of the next permanent census. The methodologies studied – both at survey design stage and in the estimation phase- is aimed to limit the use of direct survey for the collection of data on socio-economic variables, focusing on a strategy based on the use of administrative sources and on the integration of social surveys. In the paper, paragraph 2 describes the classification of the target variables and auxiliary variables considered in the system. Paragraph 3 presents the integration scenario called *"One survey occasion with pooled sample"* used in the empirical study. Paragraph 4 discusses the estimation strategies with reference to the scenario proposed; paragraph 5 describes the simulation study carried out and the results obtained.

2. Classification of variables

The classification of variables proposed in the paper follows the Eurostat's approach and is aimed to group in homogeneous basic building blocks, called modules, both the variables of interest that the auxiliary variables. These groups of variables are to be kept together for analytical/ data collection reasons.

The modules considered are:

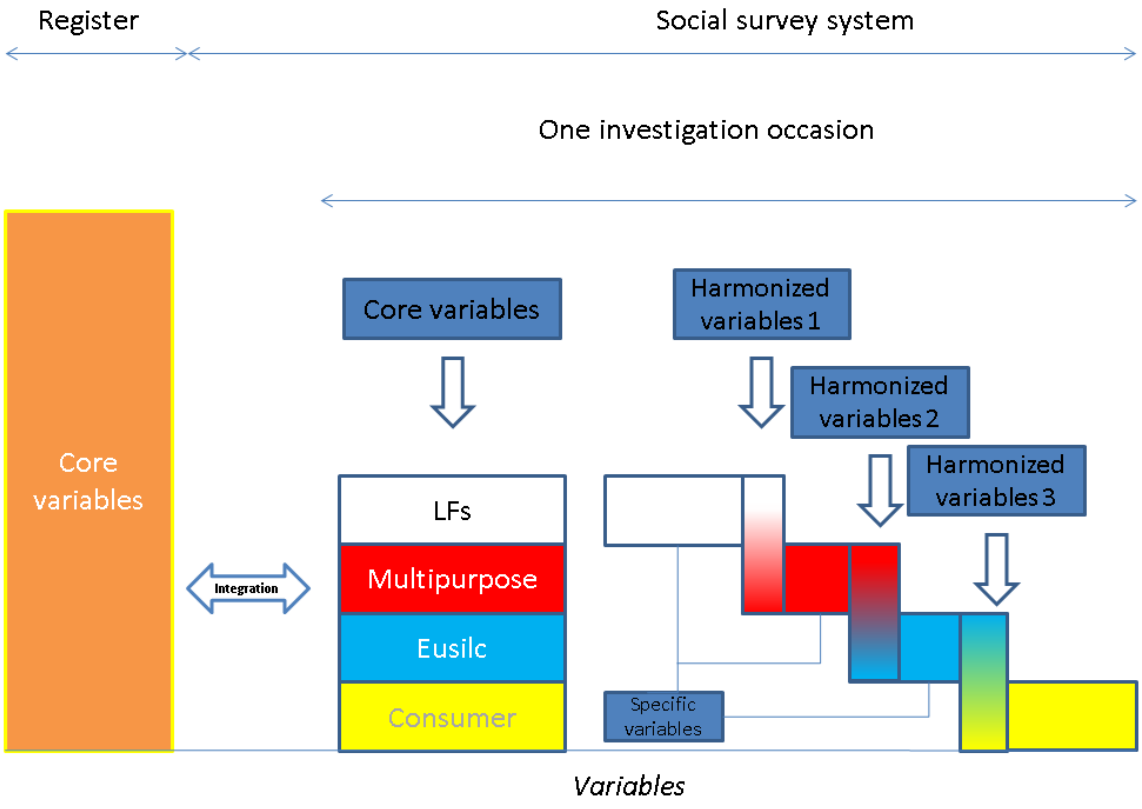
- *Core modules*: in this module are included core variables available in all data collections (samples and register). Furthermore, we consider core variables also those variables available from administrative register whose quality level is not currently considered sufficient for the production of estimates by aggregation of individual administrative register data;
- *Specific modules*: inside this module there are variables observed in only one survey. For example, are specific variables the people looking for job from the Labour Force survey and the income by type collected by the Eusilc survey;
- *Harmonised modules*: this module includes variables observed by more than one survey but often with different statistical domains. For example, fall into this module the income class currently collected by different social surveys.

3. One survey occasion with pooled sample

This scenario is based on a survey instrument design that provides that the households included in a given sub-sample, relative to a specific survey, are interviewed in a single occasion during the year, in which are collected at the same time all the variables of interest, both the structural variables, the harmonized variables and the specialized variables specific to that instrument.

Figure 1 shows the scenario taken into account in the paper. Each sub-sample is composed by different households. In this scenario, all the variables of interest are measured together in a single wave. The pooled sample so constructed allows the use of the same information observed in different surveys/instruments. The ARCHIMEDE register (Integrated archive of economic and demographic micro data) has been integrated with the sample data. This register has the aim of creating bases of useful micro-data for the study of socio-economic phenomena through the integration of variables extracted from 19 different administrative registers. In paragraph 5 will be descripted the ARCHIMEDE's variable considered in the simulation study.

Figure 1: *One survey occasion with pooled sample*



4. Estimation methods

The scenario presented in previous paragraph, thanks to the collection of both specific and auxiliary variables, offers the possibility of pooling information using model based or model

assisted estimation techniques methodologies. In particular, the variables can be pooled with model assisted (Kim and Rao, 2012) or model based (Battese et al. 1988) projection estimators.

This approach involves the identification of a working-model linking the dependent variable and the auxiliary variables observed in the different sub-samples and presents in the register. Fitting the model on the data collected in the specific survey it is possible to project the variable of interest, by means the parameters of the estimated model and the auxiliary variables, both on the pooled sample or on the register. This method requires a high level of quality of the auxiliary variables and a high goodness-of-fit of the working-models to provide considerable advantages both in terms of statistical properties of the estimators that in terms of detail of the information that can be produced.

The considered design-based estimators are:

1. Generalized regression (GREG) estimators: built separately for each sub-sample;
2. GREG estimators applied to the pooled sample: obtained by applying direct estimator to the pooled sample;
3. Projection from LFS to Pooled sample: obtained computing the predicted values on the pooled sample based on the working-model fitted on the LF sample data;
4. Projection from LFS to register: determined by defining the predicted values on the population register data based on the model fitted on the LF sample data.
5. Projection from Pooled sample to register: obtained by evaluating the predicted values on the population register data based on the model fitted on the pooled sample data.

Instead, within the case of model-based estimators is considered:

6. EBLUP unit level estimator: obtained computing the predicted values using the population totals of the auxiliary variables included in the working-model fitted on the LF sample data.

5. Simulation study

The simulation study aims to evaluate the quality of the estimators previously presented for different sub-regional domains obtainable either by design-based methods (projection estimators) and model-based estimators (Small Area Estimators, SAE) using the potential of the pooled sample. In particular, we consider three types of sub-regional territorial domains: provinces, aggregation of Local Labour Market Areas (macro-LLMA) and Local Labour Market Areas (LLMA). Only for the variables employed also the municipality estimates have been carried out.

The simulation based on a Monte Carlo experiment is aimed to compare the empirical properties of the estimates in terms of bias and mean square error. 200 samples have been drawn from the 2011 Italian population census, for two Italian regions, Trentino-Alto-Adige and Marche. The target variables are the total of persons employed and unemployed in these two regions. Linear model for the projection estimator have been fitted, with a fixed intercept at macro-LLMA level. The auxiliary variables used in the models are: marital status, educational level, citizenship, not in labour force, cross classification gender-age. The models were also enriched with information from the ARCHIMEDE register, which were linked with the 2011 census and so available for all individuals. Specifically, the variable used is a binary variable that indicates for every individual if he has a signal or not in at least one administrative source related to the employment world.

The working-models studied are summarized in the table below.

Table 1: *Working-models details*

Projection on pooled sample	*Variables*
Full model	*Marital status, educational level, citizenship, not in labor force, cross classification gender-age, ARCHIMEDE variable by regions*
Reduced model	*Marital status, educational level, citizenship, cross classification gender-age, ARCHIMEDE variable by regions*
Projection on register	
Minimal model	*Marital status, citizenship, cross classification gender-age, ARCHIMEDE variable by regions*

Once model selection and fitting has been completed, the prediction properties of the different estimates, obtained on the basis of the selected models, have been evaluated. All the estimators were compared by means of the standard indicators of accuracy of prediction: the Mean Absolute Relative Error (MARE) and Average Relative Root Mean Squared Error (ARRMSE). We further considered the $\boldsymbol{R^2}$ values to compare the goodness of fit of each model and so to evaluate the explanatory power of the different external variables considered in the application.

The indicators are formulated as follows:

$$MARE = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{R}\sum_{r=1}^{200}\hat{y}_{rd} - Y_d\right|$$

$$ARRMSE = \frac{1}{D}\frac{1}{R}\sum_{d=1}^{D}\sum_{r=1}^{R}\frac{\sqrt{(\hat{y}_{rd} - Y_d)^2}}{Y_d}$$

Where $\hat{y}_{rd}$ and Y_d are respectively the predicted value and the correspondent true value of the target variable.

The results for the variable *employed* are shown in Table 2, in which the $\boldsymbol{R^2}$ shows very high values for all the models. The MARE and the ARRMSE indicators are computed for the four type of domains described above. At province level the best results are obtained by the Projection estimator using the register, but also good performances are obtained for direct GREG estimator. At Macro-LLMA level the GREG estimator loose its good properties showed at provinces level, presenting a huge increase of variability (ARRMSE 21.36%). The Pooled estimators presents still good results on this level, very closed to the Projection estimator using the register. At LLMA and at municipality level the estimators both based on the LF data or on the pooled sample show very poor results with respect to those referred to macro-LLM. This is due to the fact that on 54 LLM areas included in the regions only 26 are always present in the 200 simulations, while for the municipalities on 572 areas only 27 are always included in the simulations. For this reason, the synthetic estimators (projection on register and SAE estimator) show similar results in terms of bias and variability as well.

Table 2: *MARE and ARRMSE for the variable employed*

Mean Absolute Relative Error - Employed							
	GREG LFS	Projection LFS-Pooled *Reduced model*	Projection LFS-Pooled *Full model*	Projection LSF-Register *Minimal model*	Projection Pooled-Register *Minimal model*	Pooled	Eblup
R^2	-	**89**	**95**	**89**	**89**	-	-
PROV (7)	0,33	0,55	0,56	0,12	0,08	0,6	-
Macro LLMA (14)	1,97	0,47	0,49	0,14	0,09	0,44	2,42
LLMA (54)	232,44	72,94	74,04	1,15	1,11	71,96	2,74
MUNIC. (527)	1779	550	550	1.97	1.96	551	3.48
Average Relative Root Mean Squared Error - Employed							
	GREG LFS	Projection LFS-Pooled *Reduced model*	Projection LFS-Pooled *Full model*	Projection LSF-Register *Minimal model*	Projection Pooled-Register *Minimal model*	Pooled	Eblup
PROV (7)	4,12	5,5	5,49	1,51	0,95	5,44	-
Macro LLMA (14)	21,36	3,15	2,96	2,07	1,31	2,73	3,74
LLMA (54)	264	109	110	2.6	1.9	108	4
MUNIC. (527)	1791	608	608	3	2.5	610	4.62

The results for the variable un*employed* are shown in Table 3. For this variable the R^2 value is similar using the *reduced* and the *minimal* model (14-15%) while goes up to the 33% using the *full* model. As well as for the employed, at provinces level and at Macro-LLMA good results are obtained from the GREG estimator and from the Pooled estimator, especially in terms of bias. At LLMA level only the projection on register estimator show good performance both with the MARE and the ARRMSE indicators below the threshold of the 13% and the 30%. The table 3 shows that considering only the 26 LLMAs always sampled in the 200 simulations, the bias estimates goes down up to the 5%.

Table 3: *MARE and ARRMSE for the variable unemployed*

Mean Absolute Relative Error - Unemployed							
	GREG LFS	Projection LFS-Pooled *Reduced model*	Projection LFS-Pooled *Full model*	Projection LSF-Register *Minimal model*	Projection Pooled-Register *Minimal model*	Pooled	Eblup
R^2	-	**15**	**33**	**15**	**14**	-	-

PROV (7)	0,88	1,04	2,12	0,96	0,46	0,6	-
Macro LLMA (14)	2,53	1,25	1,4	1,36	1,05	0,98	34,39
LLMA (54)	242,8	68,1	44,45	12,46	12,26	82,22	48,77
LLMA in-samples (26)	8,02	7,26	6,02	5,34	5,22	2,78	35,76

Average Relative Root Mean Squared Error - Unemployed

	GREG LFS	Projection LFS-Pooled *Reduced model*	Projection LFS-Pooled *Full model*	Projection LSF-Register *Minimal model*	Projection Pooled-Register *Minimal model*	Pooled	Eblup
PROV (7)	16,13	15,46	15,21	14,25	9,38	11,59	-
Macro LLMA (14)	29,95	22,41	21,71	21,92	14,3	15,17	42,2
LLMA (54)	312	111	99	29	21	136	57
LLMA in-samples (26)	54	34	34	22	15	33	44

REFERENCES

Battese G. E., Harter R. M., Fuller W. A. (1988) An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83 (401):28–36.

Chipperfield J. O., Steel D. G. (2009) Design and Estimation for Split Questionnaire Surveys, *Journal of Official Statistics*, Vol. 25, No. 2, pp. 227–244.

EUROSTAT 2013j. D12. (2013) *Roadmap for the integration of European social survey*s, http://ec.europa.eu/eurostat/cros/sites/crosportal/files/D12_Roadmap.pdf

Kim J.K., Rao J.N.K. (2012) Combining data from two independent surveys: a model-assisted approach, *Biometrika*, Vol. 99, No.1, pp. 85-100.

Lavallée P., Rivest L.P. (2012) Capture–Recapture Sampling and Indirect Sampling, *Journal of Official Statistics*, Vol.28, No.1, pp. 1–27.