Census and Social Surveys Integrated System

Stefano Falorsi
stfalors@istat.it

1. Introduction

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, 2013) 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. The performances of the scenario integrating the social surveys is compared to the scenario in which the social surveys are pooled together.
A Montecarlo simulation study using Census 2011 data is carried out. In the simulation 200 samples are drawn for each of 4 very important Istat surveys, referring to two regions: Trentino-Alto Adige 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 in terms of bias and MSE is performed on different estimators of the labour force characteristics (employed and unemployed counts), for different domains, by means of standard Monte Carlo indicators based on the 2011 census values.

2. Pooled Sample and Master Sample

This scenario named pooled sample implies that the households included in each sample of the social surveys are interviewed in a single occasion during the year, in which all the variables of interest are collected at the same time, that is structural, harmonized and specific variables.
The pooled sample so constructed allows the use of the same information observed in different surveys/instruments.
The Census and Social Surveys Integrated System (CSSIS) is a complex statistical process exploiting and integrating the information arising from registers and surveys on socio-economic variables. It is designed as a two phases Master Sample (MS) design based on a set of balanced and

coordinated sampling surveys. It is planned for supporting the Istat Population Register (PR) in order to increase the amount of provided statistical information and to improve the level of coverage and quality.

The PR is the backbone of the system for the production of social statistics, with a row for each target unit referred to a *usual resident person* (living in households or in institutional households). For each target unit, the core information, coming from demographic sources, is extended to all the basic social variables (coming from administrative sources and/or social surveys) among which employment status, economic and health conditions.

For an optimal design of the CSSIS for supporting the PR, it is useful to classify the variables included as *totally*, *partially* or *not replaceable* ones. The first class contains the variables for which the administrative sources provide the correspondent *proxy* information. At the end of the statistical process, including editing and imputation for partial non response, these variables are considered *complete*, because they are available for all units in PR, and *accurate*, having a good level of coverage and quality. Administrative sources provide the correspondent *proxy* information, also, for partially replaceable variables, but these are considered complete and accurate only for a subset of the target population. For the remaining subset of the target population, this type of variables are unknown or cannot be considered accurate because of the failure of the synthetic model of imputation. For instance, this is the case of the "Regular employed in the Labor Market" variable. Finally, for not replaceable variables it is not directly available the correspondent proxy information coming from administrative registers. Then, for these variables, target parameters can be estimated by means of sample surveys and exploiting the auxiliary information coming from the PR. The set of estimates should meet the requirements of: (a) *reliability*, obtained by means of an approximately design-unbiased estimator, or by a model-based method in which the model used is plausible in some sense. In both cases the CV of the estimates should be kept lower than a chosen threshold; (b) *consistency*, that is the data obtained combining estimates in different ways must produce the same results.

The main scope of the CSSIS is filling the informative gap of the PR for the estimation of target parameters referred to partially replaceable and not replaceable variables on social and economic data. To this aim the MS design is planned for exploiting together (*pooling*) and in an efficient way all the common information (target and auxiliary variables) observed by the different sampling surveys belonging to the system. Furthermore, the MS estimation strategy uses all the complete auxiliary information of the PR. This strategy should be able to produce more efficient direct estimates than the estimates produced by adopting separate estimation strategy for each survey. Within this context the harmonization of the common variables – i.e. *core structural* variables (which are target variables for all surveys) and *harmonized* variables (which are target variables for more than one survey) - and the harmonization statistical production process are a crucial issue.

The permanent census process is integrated within this context. It is a register based census using both the information produced by the PR, for replaceable variables, and the CSSIS for the remaining variables. More precisely, the permanent census is aimed to produce, starting from 2018, both annual data for a subset of the target parameters (hypercubes), and multiannual data for the complete set of hypercubes traditionally produced every ten years by the population census, possibly pooling sampling data over a period of consecutive years. Of course, statistics for replaceable variables can be disseminated every year.

As regards the basic objectives of support to the permanent census, the first phase of MS design is based on two different component samples, namely A and L.

The component A - based on an area sample of Enumeration Areas (EA) or selected by an Integrated Address File (IAF) - is designed to satisfy the needs of estimating under-coverage (S_U) and over-coverage (S_O) rates of the PR at national and local level for different sub-population profiles like sex, age classes, nationality. These rates should be applied to the PR for obtaining weighted population counts corrected for coverage errors. The estimated population counts are obtained using the Extended Dual System Estimator (EDSE), taking into account both under-coverage and over-coverage.
The component L - based on a list sample - is designed with the purpose of: (T_I) thematic integration, that is estimating the hypercubes which cannot be obtained using the replaceable information coming from registers. Furthermore, in order to pool the information coming from the two components, component L, could be planned to provide reliable information on spatial variability of over-coverage indicators (S_{OI}) of the PR. On the other hand, the component A, could be designed to meet, also, the target T_I. In turn, the component L could also be modified to improve the estimation process with the focus of estimating via indirect sampling some aspects of Undercoverage S_U.

More in general, the first phase survey should be focused on the following aims:
(a1) obtaining sampling information on partial and not replaceable core structural variables useful for the PR;
(b1) establishing a first contact with the sample households, a subsample of which will be re-interviewed in the second phase the following year for the second phase. The first contact could be managed in order to reduce potential second-phase non-response;
(c1) obtaining updated contact information on telephone numbers and email addresses. This contact information, which is not available on the sampling frame, may allow to carry out less expensive interview techniques (CAWI or CATI) in the second phase.

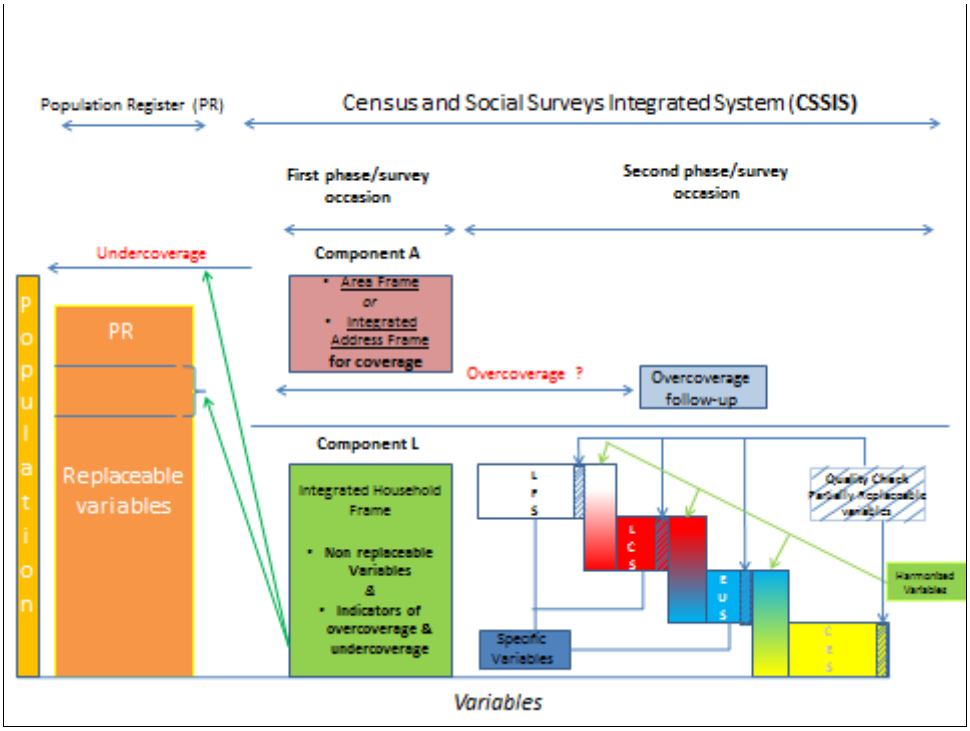
From the first phase sample a set of negatively coordinated samples of households can be selected for the second phase surveys, aimed:
(a2) to provide information on harmonized and specific socio-economic variables currently observed by Labour Force (LFS), Living Conditions (LCS); EuSilc (EUS) and Consumer Expenditure (CES) surveys;
(b2) to confirm the common structural variables already surveyed in first phase interview. These surveys are currently based on stratified two stage sampling designs (municipalities-households), and they are planned, selected and realized separately. For this reasons, it may happen to observe important differences among the estimates related to the same variables observed from the different social surveys even if the definitions and the wording of the related questions are the same. Then, in order to be able in the future to pool the same information coming from the different two phase surveys, a strategic issue will be to improve "harmonization" between the social surveys. As a matter of fact, one of the main purposes of the system described above is to reduce potential systematic differences among the surveys via harmonization of survey designs.

Furthermore, the first phase sample can be stratified or balanced, using variables in the PR, to identify areal or structural subpopulations supposed to be problematic for coverage or subjected to structure and characteristics changes in the short time period. Similarly, the second phase sample can be balanced on the set of harmonized specific variables, and to observe directly variables correlated with target variables (for instance the self-declared employment condition) useful to be used in the estimation process.

The component L of the first phase sample should be based on a yearly sample size of about 2000 municipalities out of 8100 and around 300.000 households. The first phase sample size should be at least large enough to cover the 140.000 households sample size needed for the second phase.

For an overview of CSSIS, based on the two phase MS design, see figure 1 below.

Figure 1 – An overview of the CSSIS.



The administrative records support mainly the development of the Census Population Frame (CPF) from which the component L is selected. This integrates the PR with other sources related to labour and educational archives, tax returns. A further goal is the correction of individual addresses in order to obtain the correct geographic population. These corrections are made on individual records and, therefore, all sources of information have to be linked by using a unique identification code. In the CPF records are need to be associated to their dwelling unit, via the centroid of their building. The component L is selected from CPF and the Final Sampling Units (FSU) are households or addresses belonging to the CPF.
The component A is based on an sample design, in which the FSUs are census EAs or the addresses of an Integrated Addresses Frame IAF. The IAF is obtained integrating the addresses belonging to CPF with addresses related to new buildings.
The main difference between the components L and A sampling schemes with addresses as FSUs is that the latter must be "blind" with respect to the information and the units belonging to the CPR. In this way the hypotheses below the DSE are completely satisfied.

Referring to similar international experiences, for the definition of a general master sample design for social surveys, analogous designs have been proposed by Eurostat considering a modular approach for the design of integrated social surveys. Furthermore, the ABS is designing an integrated system of investigations very similar to what described here. In this case, this survey system, called Australian Population Survey, does not replace the census.

The design with two components supporting the register census is similar to what ONS has been studying for the register-based census supposed to start in 2023 after the 2021 census run (ONS, 2016). In particular, in 2021 the ONS will conduct a traditional census and, at the same time, will carry on a parallel census run based on the construction of an integrated population registry using several administrative sources and two investigations with characteristics similar to those of the components L and A of the Italian strategy. It is worthwhile to mention that every year since 2015,

and until 2023 the ONS will produce an assessment to evaluate how much they are away from the model to be.

Another international experience showing similarities with what is planned in Italy is the Israelian rolling integrated census. They use an integrated register which is adjusted by means of weights computed by means of an EDSE (Pfeffermann, 2015).

3. 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 using the master and pooled sample;
2. Projection from master/pooled sample to register: obtained by evaluating the predicted values on the population register data based on the model fitted on the master/pooled sample data.

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

3. Synthetic unit level estimator: obtained computing the predicted values using the population totals of the auxiliary variables included in the working-model fitted on the master and pooled sample.

4. 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 both master and pooled samples. In particular, we consider four types of sub-regional territorial domains: provinces, aggregation of Labour Market Areas (macro-LMAs), Labour Market Areas (LMAs) and municipalities.

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, using the master sample approach and the pooled sample approach. The sample size for the two regions corresponds to the sample size needed to cover the samples of the four social surveys. 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 provinces level. The auxiliary variables used in the models are: marital status, educational level, citizenship, cross classification gender-age. The models are also enriched with information from the administrative register ARCHIMEDE, which is the result of the integration nineteen different registers and contains useful micro-data for socio-economic phenomena. Specifically, for each individual an indicator of the presence of signal in at least one administrative source related to the employment world. Then, the auxiliary information used in the model specification is: marital status, citizenship, cross classification gender-age, ARCHIMEDE variable.

Once model selection and fitting is completed, the prediction properties of the different estimates, obtained on the basis of the selected models, are evaluated. All the estimators are 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). Furthermore, we consider the values to compare the goodness of fit of each model and to evaluate the explanatory power of the different external variables considered in the application.

The evaluation indicators are formulated as follows:

$$MARE = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{R}\sum_{r=1}^{200}\hat{y}_{rd} - Y_d\right|, \quad ARRMSE = \frac{1}{D}\frac{1}{R}\sum_{d=1}^{D}\sum_{r=1}^{R}\frac{\sqrt{(\hat{y}_{rd}-Y_d)^2}}{Y_d},$$

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

Table 1 and Table 2 show the results for the variable employed respectively for master and pooled sample. Table 3 and Table 4 display the analogous outputs for the unemployed counts. MARE and ARRMSE indicators are computed for the four types of domains described above.

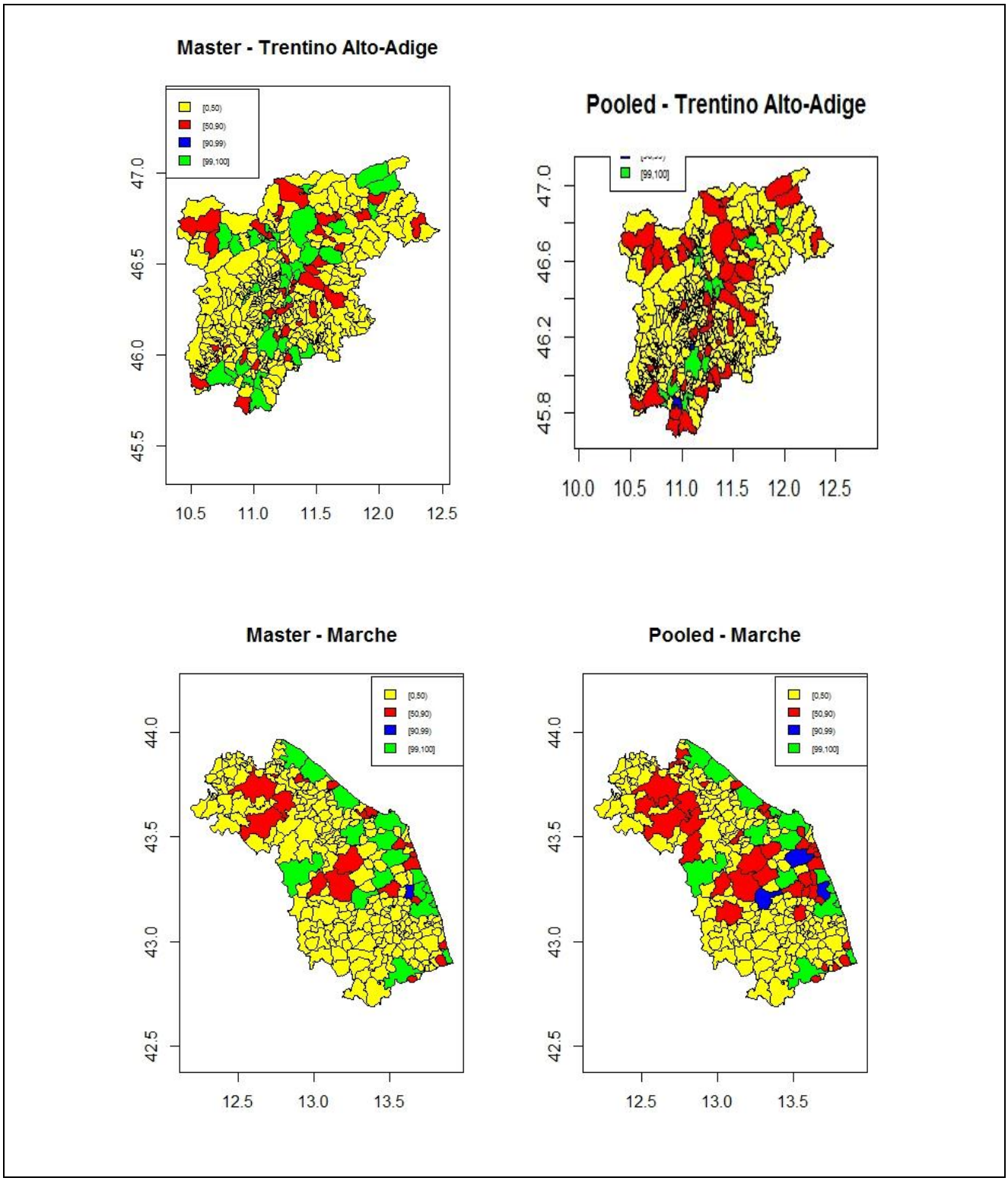
The R^2 in Tables 1 and 2 shows very good performance of the models. From the comparison of the MARE and ARRMSE values, it results that the Projection estimator outperforms the other two methods.

In Table 1, at LMA and municipality level the estimator based on the master sample shows very poor results with respect to those referred to macro-LMAs. This is due to the fact that out of 54 LMAs in the regions only 30 are always included in the 200 simulation samples, while for the municipalities out of 572 areas only 49 are always included in the simulated samples. For this reason, the Projection and the SAE estimators outperform the GREG estimator. If we take into account only the areas with rate of inclusion in the simulation study equal to 100%, 90% and 50%, the GREG estimator displays good results both in terms of bias and MSE.

Analogous consideration can be drawn also for Table 2. At first sight, comparing Table 1 and 2, it would seem that better results are achieved with the pooled sample with respect to the master sample. Indeed, the better figures in Table 2 are due to the number of in-sample LMAs or municipalities in the pooled sample are smaller than the analogous numbers in the master sample. For instance in the pooled sample 26 LMAs and 27 municipalities are always included in the samples, while for the master sample the corresponding numbers are respectively 30 and 49. Similar

consideration can be done also for the other in-sample rates and for the overall LMAs and municipalities included in the samples. This means that for the master sample the sampling households are spread over a larger number of municipalities and LMAs, and, therefore, the sampling size for each municipality is smaller than the pooled sample. In the Figure 1 the municipality's in-sample rate are showed.

Figure 1: Municipalities in-sample rate: 0%-50% yellow; 50%-90% red, 90%-99% blue, 99%-100% green



At LMA level, the results for the areas with 50% in-sample rate for master and pooled simulation are directly comparable. The results in terms of MSE obtained for the master sample are better than the corresponding results for the pooled results.

Table 1: Master sample - MARE and ARRMSE for the variable employed

Mean Absolute Relative Error			
	GREG	Projection Master to Register	SAE
R^2	-	0.89	-
Provinces (7)	0.3	0.07	-
Macro LMA (20)	1,4	0,5	2,3
LMA (54)	95,9	1,2	2,9
100% in-sample LMA (30)	*1,9*	*0,6*	*2,6*
90% in-sample LMA (38)	*2,6*	*0,6*	*2,5*
50% in-sample LMA (50)	*10,0*	*0,9*	*2,7*
Municipalities (572)	841,3	2,0	3,6
100% in-sample municipalities (49)	*2,6*	*0,9*	*3,0*
90% in-sample municipalities (50)	2,7	0,9	3,0
50% in-sample municipalities (99)	*34,5*	*1,0*	*3,1*
Average Relative Rooot Mean Squared Error			
	GREG	Projection Master to Register	SAE
Provinces (7)	2.6	1	-
Macro LMA (20)	18,0	1,1	3,5
LMA (54)	123,7	1,7	4,0
100% in-sample LMA (30)	*24,0*	*1,1*	*3,8*
90% in-sample LMA (38)	*28,5*	*1,1*	*3,7*
50% in-sample LMA (50)	*38,7*	*1,4*	*3,9*
Municipalities (572)	848,2	2,3	4,6
100% in-sample municipalities (49)	*9,7*	*1,3*	*4,2*
90% in-sample municipalities (50)	9,8	1,3	4,1
50% in-sample municipalities (99)	*39,5*	*1,4*	*4,2*

Table 2: Pooled sample - MARE and ARRMSE for the variable employed

Mean Absolute Relative Error			
	GREG	Projection Pooled to Register	SAE
R^2	-	0.89	-
Provinces (7)	0,3	0,5	-
Macro LMA (14)	1,2	0,1	2,4
LMA (54)	72,5	1,1	2,7
100% in-sample LMA (26)	*3,0*	*0,4*	*2,5*
90% in-sample LMA (41)	*3,8*	*0,6*	*2,4*
50% in-sample LMA (50)	*7,4*	*0,8*	*2,6*
Municipalities (572)	563,6	2,0	3,5
100% in-sample municipalities (27)	*2,7*	*0,6*	*2,1*
90% in-sample municipalities (32)	3,2	0,7	2,1

50% in-sample municipalities (113)	*41,9*	*0,9*	*2,8*
Average Relative Rooot Mean Squared Error			
	GREG	Projection Pooled to Register	SAE
Provinces (7)	4,1	2,3	-
Macro LMA (14)	13,0	1,3	3,7
LMA (54)	110,2	1,9	4,0
100% in-sample LMA (26)	*25,2*	*1,3*	*3,8*
90% in-sample LMA (41)	*34,3*	*1,4*	*3,8*
50% in-sample LMA (50)	*41,1*	*1,6*	*3,9*
Municipalities (572)	621,0	2,5	4,6
100% in-sample municipalities (27)	*20,9*	*1,5*	*3,7*
90% in-sample municipalities (32)	25,6	1,5	3,6
50% in-sample municipalities (113)	*76,0*	*1,6*	*4,0*

Regarding the unemployment counts, Tables 3 and 4 show very poor performances of the models in terms of R^2. Furthermore, for this variable the comparison of MARE and ARRMSE values show that the Projection estimator outperforms the other two methods. The comparison of Tables 3 and 4 for the unemployment counts lead to the same considerations made before about the employment counts: a direct comparison between pooled and master results is not possible as the number of in-sample LMAs or municipalities in the pooled sample are smaller than the analogous numbers in the master sample. Only the results at LMA level for the areas with 50% in-sample rate are directly comparable between master and pooled simulation (in both cases there are 50 areas). Again, the best results in terms of MSE are obtained when master sample is applied.

Table 3: Master sample - MARE and ARRMSE for the variable unemployed

Mean Absolute Relative Error			
	GREG	Projection Master to Register	SAE
R^2	-	0.15	-
Provinces (7)	0.3	0.5	-
Macro LMA (20)	2,3	10,3	37,0
LMA (54)	91,3	15,9	45,7
100% in-sample LMA (30)	*2,5*	*10,3*	*36,3*
90% in-sample LMA (38)	*3,3*	*13,7*	*42,5*
50% in-sample LMA (50)	*10,8*	*15,1*	*44,9*
Municipalities (572)	835,5	34,8	69,5
100% in-sample municipalities (49)	*4,0*	*16,1*	*38,7*
90% in-sample municipalities (50)	4,2	15,8	38,6
50% in-sample municipalities (99)	*36,5*	*22,1*	*49,1*
Average Relative Rooot Mean Squared Error			
	GREG	Projection Master to Register	SAE
Provinces (7)	11	10	-
Macro LMA (20)	26,1	16,0	44,9

LMA (54)	143,9	20,3	53,8
100% in-sample LMA (30)	*36,0*	*15,9*	*44,7*
90% in-sample LMA (38)	*41,8*	*18,3*	*50,7*
50% in-sample LMA (50)	*54,8*	*19,7*	*53,0*
Municipalities (572)	985,8	38,5	78,7
100% in-sample municipalities (49)	*53,0*	*20,4*	*48,7*
90% in-sample municipalities (50)	52,9	20,3	48,4
50% in-sample municipalities (99)	*95,3*	*26,3*	*58,9*

Table 4: Pooled sample - MARE and ARRMSE for the variable unemployed

Mean Absolute Relative Error			
	GREG	Projection Pooled to Register	SAE
R^2	-	0.15	-
Provinces (7)	0,4	0,5	-
Macro LMA (14)	1,1	1,0	34,4
LMA (54)	83,0	12,3	48,8
100% in-sample LMA (26)	*2,8*	*5,2*	*35,8*
90% in-sample LMA (41)	*3,9*	*9,7*	*40,0*
50% in-sample LMA (50)	*7,5*	*11,3*	*48,0*
Municipalities (572)	559,2	33,4	73,2
100% in-sample municipalities (27)	*3,1*	*9,9*	*30,2*
90% in-sample municipalities (32)	3,4	8,8	31,1
50% in-sample municipalities (113)	*41,8*	*17,1*	*48,3*
Average Relative Rooot Mean Squared Error			
	GREG	Projection Pooled to Register	SAE
Provinces (7)	10,4	9,4	-
Macro LMA (14)	18,9	14,3	42,2
LMA (54)	137,2	21,7	57,3
100% in-sample LMA (26)	*35,1*	*15,2*	*44,5*
90% in-sample LMA (41)	*47,1*	*19,2*	*54,6*
50% in-sample LMA (50)	*56,2*	*20,9*	*56,7*
Municipalities (572)	768,3	40,9	83,0
100% in-sample municipalities (27)	*39,3*	*17,7*	*39,4*
90% in-sample municipalities (32)	44,1	16,9	39,8
50% in-sample municipalities (113)	*115,7*	*25,2*	*58,0*

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.

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

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

Kim J.K., Rao J.N.K. (2012) Combining data from two independent surveys: a model-assisted approach, *Biometrika*, Vol. 99(1), pp. 85-100.

ONS (2016). Annual assessment of ONS's progress towards an Administrative Data Census post-2021, downloadable at *https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments*.

Pfeffermann, D. (2015). Methodological Issues and Challenges in the Production of Official Statistics, *Journal of Survey Statistics and Methodology*, 3, 425–483.