United Nations

ECE/CES/GE.41/2018/20

Economic and Social Council

Distr.: General
19 July 2018

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twentieth Meeting
Geneva, 26–28 September 2018
Item 2 of the provisional agenda
Methodology, new data sources including big data

Preliminary Experimental: Results on the Italian Population and Housing Census Estimation Methods

Note by the Italian National Statistical Institute (Istat)[1]

Summary

Starting from October 2018, the population Census in Italy will abandon the traditional decennial ‘door-to-door’ enumeration for a ‘combined’ approach which will integrate administrative data and sample surveys. The goal of the ‘permanent’ Census is to produce annual data - replacing the previous decennial cycle - using information from administrative sources integrated with sample surveys information. The new Census strategy will allow a significant reduction of the cost of the census, of respondents’ burden, and of the organizational impact on municipalities.

[1] This document was submitted late due to the late submission of the paper from the National Institute of Statistics.

GE.18-11953(E)

1811953

I. Introduction

1. Starting from October 2018, the population Census in Italy will abandon the traditional decennial ‘door-to-door’ enumeration for a ‘combined’ approach which will integrate administrative data and sample surveys. In fact, in 2012, the so-called ‘permanent’ Census of Population and Housing (in Italian Censimento permanente della popolazione e delle abitazioni) was introduced in Italian legislation (Article 3 of Legislative Decree 179/2012, converted with amendments into Law 221/2012). The goal of the ‘permanent’ Census is to produce annual data - replacing the previous decennial cycle - using information from administrative sources integrated with sample surveys information. This will be done within the frame of Istat’s (the Italian National Institute of Statistics) modernization program, whose focus is to integrate administrative data, create statistical registers and conduct supporting statistical surveys, in line with the new organizational, technological and methodological model aimed at fully exploiting data already available. The new Census strategy will allow a significant reduction of the cost of the census, of respondents’ burden, and of the organizational impact on municipalities (that had traditionally been responsible for the census field-work).

2. Since 2015, Istat has initiated several projects to explore the use of administrative sources for statistical purposes. To manage the increasing number of administrative data sets and to maximize the benefit, Istat has built an integrated system of available administrative sources, called ISM (Integrated System of Microdata).

3. The population Census should provide a set of hypercubes currently required by Eurostat on socio-economic variables (employment status, educational level, migrant status, etc.), as well as a set of hypercubes to fulfill national requirements (especially at municipality level).

4. 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 Australian Bureau of Statistics, 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.

5. A similar census design is under study by the UK Office for National Statistics, ONS, 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.

6. Another international experience showing similarities with what is planned in Italy is the Israeli rolling integrated census (see Pfeffermann, 2015, for more details).

II. Italian Permanent Census

7. The new population Census is a complex statistical process exploiting and integrating the information derived from registers and collected in surveys on socio-economic variables. It is designed as a two-phase design based on a Master Sample (MS) and on a set of balanced and coordinated sample surveys. It is planned for supporting the Population Register (PR) in order to increase the amount of statistical information provided and to improve the level of coverage and quality.

8. The PR is the backbone of the system for the production of social statistics, with a row for each target unit i.e. a usually resident person (whether living in private or institutional households). For each target unit, the core information (taken from demographic sources) is extended to all the basic social variables (obtained from administrative sources and/or social surveys) among which employment status and health conditions.

9. It is useful to classify the variables included in the PR as totally, partially or not replaceable. The first category encompasses those variables for which the administrative sources provide the corresponding proxy information and which, at the end of the statistical process - including editing and imputation for partial non-response, are considered to be complete (because they are available for all units in the PR), and accurate (i.e. having a good level of coverage and quality). For instance, sex and age are variables which are known for all the individuals in the PR and, therefore, they are considered totally replaceable variables.

10. Administrative sources also provide the corresponding proxy information for partially replaceable variables, which are considered complete and accurate only for a sub-set of the target population. For the remaining population, these variables are either unknown or cannot be considered accurate because of the failure of the synthetic model of imputation. For instance, this is the case for the ‘employed’ variable, which is completely replaceable only with respect to the sub-population of the ‘regularly employed’.

11. Finally, for not replaceable variables the corresponding proxy information coming from administrative registers is not directly available. For these variables, target parameters can be estimated by means of sample surveys and exploiting the auxiliary information coming from the PR, that is the set of variables contained in the register which are supposed to be predictive for the non-replaceable variables under study. 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 coefficients of variation of the estimates should be kept lower than a chosen threshold;

(b) consistency in that the data obtained by combining estimates in different ways must produce the same results.

12. With regard to the basic objectives of the ‘permanent’ census, the first phase of the MS design is based on two different component samples, namely A and L.

13. The component A sample - based on a sample of Enumeration Areas (EA) or of addresses selected from an Integrated Address File (IAF) - is designed to satisfy the needs of estimating under-coverage and over-coverage rates of the PR both at national and local level for different sub-population profiles such as several different combinations of sex, age and 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), i.e. taking into account both under-coverage and over-coverage.

14. The component L sample - based on a list of households - is designed with the purpose of: 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 of the PR. On the other hand, the component A sample could be designed to also meet the thematic integration target. In turn, the component L sample could also be modified to improve the

estimation process with the aim of estimating via indirect sampling some aspects of under-coverage.

15. From the first phase a set of negatively coordinated samples of households can be selected so that they can be used as samples for the main social surveys carried out by Istat. This represents the second phase of the MS.

16. The component A and L are based on a yearly sample size of about 400,000 households and 1,000,000 households respectively, drawn from 2,850 municipalities out of 7,950.

III. Estimation methods

17. The scenario presented in the 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 method proposed in this paper is the projection estimator of Kim and Rao (2012).

18. This approach involves the identification of a working model linking the dependent variable and the auxiliary variables observed in the MS and the information in the register. Fitting the model on the data collected in the MS it is possible to project the variable of interest, by means of the estimated model parameters and the auxiliary variables 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 and in terms of detail of the information that can be produced. In this study the linking functions proposed are: the linear regression and logistic regression.

IV. Simulation Study

19. The simulation study aims to evaluate the quality of the projection estimator previously presented using different regression models.

20. The study focused on the hypercube measuring the occupational status at municipality level cross-classified by gender using the MS described in the previous section.

21. Municipality and gender can be considered as replaceable variables because they are included in the PR. These variables define the domain as an exhaustive partition of the target population.

22. The goal of the simulation based on a Monte Carlo experiment is to compare the empirical properties of the estimates in terms of bias and mean squared error (MSE). To this aim 200 samples have been drawn from the 2011 Italian Population Census for one Italian region, Piedmont, using the MS sampling design. In this region there are 1201 municipalities and 359 of those are included in the sample. The target variables are the population counts for the five modes of the variable occupational status (employed, unemployed, retired, student, in other condition) in this region. Therefore, the overall number of cells in the hypercube is 12010.

23. Linear and not linear models for the projection estimator have been fitted, with a fixed intercept at NUTS3 level. Different auxiliary variables have been used in the simulation study: demographic variables (cross-classification of gender and age, marital status, educational level, citizenship), latitude, longitude and the target variables referring to

the 2001 Census data. These variables are grouped in 4 different set of covariates summarized in the table below.

Table 1
Set of auxiliary variables

Set	*Auxiliary Variables*
A	cross-classification of gender and age (28), marital status(6), educational level (12), citizenship (2)
B	cross-classification of gender and age (28), marital status(6), educational level (12), citizenship (2), latitude, longitude
C	cross-classification of gender and age (28), marital status(6), educational level (12), citizenship (2), census counts 2001
D	cross-classification of gender and age (28), marital status(6), educational level (12), citizenship (2), latitude, longitude, census count 2001

24. Furthermore, a composite estimator is considered, that is the values used for the individuals included in the MS are the true values instead of the synthetic projected values as for the standard projection estimator. For the sake of the simplicity, the former estimator will be denoted by composite projection estimator, while the latter by synthetic projection estimator.

25. Once model selection and fitting is completed, the prediction properties of the different estimates are evaluated. All the estimators are compared by means of the standard indicators of accuracy of prediction: the Average Absolute Relative Bias (AARB) and Average Relative Root Mean Squared Error (ARRMSE). The evaluation indicators are formulated as follows:

$$AARB = \frac{1}{D}\sum_{d=1}^{D}\frac{\left|\frac{1}{R}\sum_{r=1}^{R}\hat{Y}_{rd} - Y_d\right|}{Y_d},$$

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

where $\hat{Y}_{rd}$ and Y_d are, respectively, the predicted value and the correspondent true value of the target variable in area *d*.

V. Results

26. The results for the five modes of the variable occupational status are shown in the tables below. In particular, Tables 2 and 3 show the AARB results for the LM and GLM respectively, while Tables 4 and 5 show results for the ARRMSE using LM and GLM respectively. These indicators have been carried out for the overall hypercube and, therefore, also considering in-sample and out-of-sample domains.

27. The results for the AARB and ARRMSE indicators referring to the LM are shown in Tables 2 and 3, in which the composite estimator outperform almost always the projection results. The unemployed counts estimates present the worst results in terms of AARB and ARRMSE for all the estimators and for all the sets of auxiliary information. Set D shows

the best performances among the other sets of auxiliary variables. As expected, the indicators computed in the in-sample domains show better results than the out-of-sample counterpart.

28. Tables 4 and 5 report the results for the AARB and ARRMSE indicators using GLM. The same considerations made for LM can be applied at the GLM figures. For this model, sets B and D are very similar.

29. Comparing LM and GLM results, the best combination between type of model and set of auxiliary variables is always the composite estimator using set D and GLM model specification.

30. A last evaluation was made considering the number of negative estimates produced by using the LM. The average number of negative cell estimates is 5.4 for the synthetic projection and 8.2 for the composite projection out of 12010 cells.

Table 2
AARB for the variable occupational status with LM model

Auxiliary Variables	Target Variable	Synthetic Projection			Composite Projection		
		ARRB	ARRB IN	ARRB OUT	ARRB	ARRB IN	ARRB OUT
A	Employed	5.6	5.3	5.7	4.9	3.1	5.7
	Unemployed	32.4	29.3	33.8	29.7	20.0	33.8
	Retired	7.0	6.6	7.2	6.2	4.1	7.2
	Student	16.3	14.4	17.1	14.6	8.9	17.1
	Other	11.0	10.3	11.3	9.8	6.1	11.3
B	Employed	5.3	5.0	5.4	4.7	2.9	5.4
	Unemployed	33.2	29.5	34.7	30.1	19.1	34.7
	Retired	6.7	6.4	6.9	6.0	3.8	6.9
	Student	16.3	14.6	17.0	14.5	8.7	17.0
	Other	10.9	10.1	11.2	9.6	6.0	11.2
C	Employed	5.6	5.3	5.7	5.0	3.1	5.7
	Unemployed	31.9	29.1	33.1	29.2	20.0	33.1
	Retired	7.0	6.6	7.2	6.3	4.1	7.2
	Student	16.4	14.6	17.1	14.8	9.2	17.1
	Other	11.0	10.2	11.3	9.8	6.1	11.3
D	Employed	5.3	5.0	5.4	4.7	2.9	5.4
	Unemployed	32.3	29.3	33.6	29.2	19.0	33.6
	Retired	6.7	6.4	6.9	6.0	3.8	6.9
	Student	16.1	14.6	16.8	14.4	9.0	16.8
	Other	10.9	10.1	11.2	9.6	5.9	11.2

Table 3

ARRMSE for the variable occupational status with LM model

Auxiliary Variables	Target Variable	Synthetic Projection			Composite Projection		
		ARRMSE	ARRMSE IN	ARRMSE OUT	ARRMSE	ARRMSE IN	ARRMSE OUT
A	Employed	5.6	5.4	5.8	5.2	3.9	5.8
	Unemployed	33.1	30.0	34.4	31.7	25.4	34.4
	Retired	7.1	6.7	7.2	6.5	4.8	7.2
	Student	16.6	14.6	17.4	15.6	11.4	17.4
	Other	11.1	10.4	11.4	10.2	7.4	11.4
B	Employed	5.4	5.1	5.5	5.0	3.8	5.5
	Unemployed	34.0	30.4	35.6	32.4	24.9	35.6
	Retired	6.8	6.5	7.0	6.3	4.6	7.0
	Student	16.9	15.2	17.7	16.3	12.9	17.7
	Other	11.0	10.3	11.4	10.1	7.2	11.4
C	Employed	5.7	5.4	5.8	5.0	3.8	5.5
	Unemployed	32.4	29.7	33.6	32.4	24.9	35.6
	Retired	7.1	6.7	7.2	6.3	4.6	7.0
	Student	16.7	15.0	17.4	16.3	12.9	17.7
	Other	11.1	10.3	11.4	10.1	7.2	11.4
D	Employed	5.4	5.1	5.5	5.0	3.8	5.5
	Unemployed	32.9	30.0	34.1	31.3	24.8	34.1
	Retired	6.8	6.4	6.9	6.2	4.6	6.9
	Student	16.6	15.5	17.0	16.1	14.1	17.0
	Other	11.0	10.2	11.3	10.1	7.2	11.3

Table 4

AARB for the variable occupational status with GLM model

Auxiliary Variables	Target Variable	Synthetic Projection			Composite Projection		
		ARRB	ARRB IN	ARRB OUT	ARRB	ARRB IN	ARRB OUT
A	Employed	5.4	5.2	5.5	4.8	3.0	5.5
	Unemployed	31.2	28.6	32.3	28.6	19.8	32.3
	Retired	6.9	6.5	7.1	6.1	3.9	7.1
	Student	16.3	14.2	17.2	14.7	9.0	17.2
	Other	10.1	9.5	10.3	8.9	5.7	10.3
B	Employed	5.2	4.9	5.3	4.6	2.9	5.3
	Unemployed	30.9	28.2	32.0	28.1	18.9	32.0
	Retired	6.8	6.3	7.0	6.0	3.8	7.0
	Student	16.2	14.1	17.1	14.6	8.8	17.1
	Other	10.0	9.4	10.3	8.8	5.4	10.3
C	Employed	5.4	5.2	5.6	4.8	3.0	5.6
	Unemployed	31.1	28.5	32.3	28.5	19.8	32.3
	Retired	6.9	6.5	7.1	6.2	3.9	7.1
	Student	16.3	14.3	17.2	14.7	9.0	17.2
	Other	10.1	9.6	10.4	9.0	5.7	10.4
D	Employed	5.4	5.2	5.6	4.6	2.9	5.3
	Unemployed	31.1	28.5	32.3	28.0	18.8	32.0
	Retired	6.9	6.5	7.1	6.0	3.8	7.0
	Student	16.3	14.3	17.2	14.6	8.8	17.1
	Other	10.1	9.6	10.4	8.9	5.5	10.3

Table 5
ARRMSE for the variable occupational status with GLM model

Auxiliary Variables	Target Variable	Synthetic Projection			Composite Projection		
		ARRMSE	ARRMSE IN	ARRMSE OUT	ARRMSE	ARRMSE IN	ARRMSE OUT
A	Employed	5.5	5.2	5.6	5.1	3.8	5.6
	Unemployed	31.8	29.3	32.9	30.5	25.0	32.9
	Retired	7.0	6.5	7.1	6.4	4.6	7.1
	Student	16.5	14.5	17.4	16.0	12.6	17.4
	Other	10.2	9.6	10.4	9.4	6.9	10.4
B	Employed	5.2	5.0	5.3	4.9	3.7	5.3
	Unemployed	31.6	29.0	32.7	30.2	24.5	32.7
	Retired	6.8	6.4	7.0	6.3	4.5	7.0
	Student	16.5	14.4	17.3	15.9	12.5	17.3
	Other	10.1	9.5	10.4	9.3	6.7	10.4
C	Employed	5.5	5.2	5.6	5.1	3.8	5.6
	Unemployed	31.6	29.0	32.7	30.3	24.9	32.7
	Retired	7.0	6.5	7.2	6.4	4.6	7.2
	Student	16.5	14.4	17.3	15.9	12.6	17.3
	Other	10.2	9.6	10.5	9.4	6.9	10.5
D	Employed	5.5	5.2	5.6	4.9	3.7	5.3
	Unemployed	31.6	29.0	32.7	30.0	24.4	32.4
	Retired	7.0	6.5	7.2	6.3	4.5	7.0
	Student	16.5	14.4	17.3	15.8	12.5	17.3
	Other	10.2	9.6	10.5	9.3	6.7	10.4

31. Other results are shown, analyzing the AARB and ARRMSE values for the in-sample and out-of-sample municipalities considering seven different population dimensional sizes.

32. In particular, for the in-sample municipalities the following classes have been considered: up to 500 individuals (151 municipalities), 501-1000 (65 municipalities), 1001-2000 (51 municipalities), 2001-5000 (34 municipalities), 5001-10000 (22 municipalities), 10001-30000 (27 municipalities), more than 30001 individuals (4 municipalities).

33. For the out-sample municipalities the classes taken into consideration are: up to 50 individuals (27 municipalities), 51-100 (52 municipalities), 101-200 (116 municipalities), 201-500 (247 municipalities), 501-1000 (191 municipalities), 1001-3000 (166 municipalities), more than 3001 individuals (43 municipalities).

34. Tables 6 and 7 show the AARB indicator results for the in-sample municipalities for LM and GLM respectively. The composite estimator outperforms the projection results especially in the first population class. The bias for the unemployment status decreases up to three times from the smallest demographic size to the greatest one. For the employment, retired and in other condition variables, the values of the AARB indicator in the class ‘up to 500’ are smaller than the following class ‘500-1000’. The ARRMSE results are presented in Tables 8 and 9 for the LM and GLM respectively. The usage of the composite estimator produces an improvement in terms of MSE smaller than the corresponding improvement in terms of bias.

35. Tables 10 and 11 show results for AARB and ARRMSE in the out-of-sample municipalities. The bias and MSE indicators show very poor results for the four smallest classes, especially for the variables 'unemployment status' and 'student'.

36. A direct comparison between in-sample and out-of-sample results is possible for the population size '500-1000'. The overall gain in terms of mse is about 20%, in details 14% for the unemployed counts and 26% for the retired counts estimates.

Table 6

AARB for the in-sample domains by demographic size with LM model

Auxiliary Variables	Target Variable[2]	Synthetic Projection					Composite Projection				
		Employed	Unemployed	Retired	Student	Other	Employed	Unemployed	Retired	Student	Other
A	up to 500 (151)	8.9	54.4	10.1	31.1	18.0	3.0	29.3	3.6	15.8	6.6
	500-1000 (65)	6.0	30.3	6.9	14.5	10.5	4.0	22.4	4.8	10.1	7.1
	1000-2000 (51)	4.4	26.2	6.7	10.9	8.3	3.0	19.7	4.5	7.4	5.7
	2000-5000 (34)	4.0	18.4	5.4	8.4	7.5	3.1	15.4	4.4	6.8	6.1
	5000-10000 (22)	3.1	12.0	4.3	6.9	6.4	2.6	10.1	3.6	5.8	5.3
	10000-30000 (27)	2.8	14.8	3.5	5.3	5.0	2.6	13.9	3.3	5.0	4.7
	more than 30000 (4)	2.4	12.1	3.7	5.2	5.5	2.3	11.1	3.5	4.8	5.2
B	up to 500 (151)	9.0	55.0	10.1	30.7	18.4	3.0	29.1	3.5	14.5	6.8
	500-1000 (65)	6.0	30.1	6.9	14.1	10.4	3.8	21.7	4.6	9.7	7.3
	1000-2000 (51)	4.3	26.5	6.6	10.7	8.3	2.8	18.3	4.3	7.2	5.5
	2000-5000 (34)	3.9	18.4	5.3	8.6	7.4	2.9	14.1	4.0	6.6	5.8
	5000-10000 (22)	2.9	12.1	4.2	6.9	6.2	2.2	8.5	3.3	6.0	4.6
	10000-30000 (27)	2.8	15.0	3.5	5.4	5.3	2.5	13.3	2.9	4.2	4.4
	more than 30000 (4)	2.4	12.2	3.7	5.1	5.5	2.3	11.1	2.6	3.3	3.9
C	up to 500 (151)	8.8	58.9	10.0	32.3	18.6	3.0	29.3	3.6	15.8	6.6
	500-1000 (65)	5.7	30.2	6.7	14.2	10.7	4.0	22.4	4.8	10.1	7.1
	1000-2000 (51)	4.0	25.0	6.5	10.9	8.1	3.0	19.7	4.5	7.4	5.7
	2000-5000 (34)	3.6	17.1	4.9	8.2	7.1	3.1	15.4	4.4	6.8	6.1
	5000-10000 (22)	2.7	10.2	3.9	7.0	5.7	2.6	10.1	3.6	5.8	5.3
	10000-30000 (27)	2.7	14.2	3.1	4.5	4.6	2.6	13.9	3.3	5.0	4.7
	more than 30000 (4)	2.4	11.9	2.8	3.5	4.2	2.3	11.1	3.5	4.8	5.2
D	up to 500 (151)	8.7	58.4	10.0	32.1	18.5	2.9	28.8	3.5	15.6	6.7
	500-1000 (65)	5.7	30.4	6.7	14.5	10.7	3.8	21.7	4.6	9.9	7.3
	1000-2000 (51)	4.0	24.7	6.5	11.0	8.2	2.8	18.0	4.3	7.4	5.6
	2000-5000 (34)	3.7	17.0	4.9	8.1	7.2	2.9	14.1	4.0	6.5	5.9
	5000-10000 (22)	2.8	10.2	3.9	6.8	5.7	2.3	8.5	3.3	5.8	4.6
	10000-30000 (27)	2.7	14.1	3.1	4.5	4.4	2.5	13.3	2.9	4.2	4.2
	more than 30000 (4)	2.4	11.8	2.8	3.8	4.2	2.3	10.9	2.6	3.5	3.9

[2] The number of municipalities falling in each demographic size class are reported in brackets.

Table 7
AARB for the in-sample domains by demographic size with GLM model

Auxiliary Variables	Target Variable[3]	Synthetic Projection					Composite Projection				
		Employed	Unemployed	Retired	Student	Other	Employed	Unemployed	Retired	Student	Other
A	up to 500 (151)	8.8	53.0	10.2	31.4	17.5	3.0	29.2	3.6	16.0	6.5
	500-1000 (65)	5.9	29.8	6.7	13.3	9.8	3.9	22.2	4.6	9.3	6.7
	1000-2000 (51)	4.2	26.4	6.5	10.2	7.6	2.9	19.8	4.3	6.9	5.3
	2000-5000 (34)	3.9	18.5	5.1	8.4	6.7	3.0	15.5	4.1	6.8	5.4
	5000-10000 (22)	2.9	11.8	4.0	6.4	5.7	2.4	10.0	3.3	5.4	4.6
	10000-30000 (27)	2.6	14.4	3.2	4.8	4.7	2.4	13.5	3.0	4.6	4.4
	more than 30000 (4)	2.3	10.8	3.3	5.9	4.0	2.2	10.0	3.1	5.6	3.8
B	up to 500 (151)	9.0	55.0	10.1	30.7	18.4	3.0	28.7	3.6	15.9	6.5
	500-1000 (65)	8.6	54.6	10.3	31.8	17.9	3.7	21.2	4.6	9.2	6.7
	1000-2000 (51)	5.6	29.2	6.7	13.3	9.8	2.7	18.5	4.3	6.9	5.1
	2000-5000 (34)	4.0	25.2	6.4	10.3	7.3	2.8	14.4	3.9	6.5	5.1
	5000-10000 (22)	3.6	17.2	4.8	8.0	6.3	2.2	8.1	3.1	5.4	4.1
	10000-30000 (27)	2.6	9.7	3.8	6.3	5.1	2.4	12.9	2.8	4.0	3.8
	more than 30000 (4)	2.5	13.7	3.0	4.3	4.0	2.2	10.3	2.5	4.0	3.2
C	up to 500 (151)	8.7	52.6	10.2	31.2	17.4	3.0	29.1	3.6	15.9	6.5
	500-1000 (65)	5.9	29.8	6.8	13.4	9.9	3.9	22.2	4.6	9.4	6.8
	1000-2000 (51)	4.3	26.1	6.6	10.3	7.7	2.9	19.7	4.4	7.0	5.4
	2000-5000 (34)	3.9	18.6	5.1	8.4	6.8	3.1	15.7	4.1	6.8	5.5
	5000-10000 (22)	3.0	11.8	4.0	6.4	5.9	2.5	10.0	3.4	5.4	4.8
	10000-30000 (27)	2.6	14.2	3.1	4.8	4.6	2.4	13.4	2.9	4.5	4.3
	more than 30000 (4)	2.3	10.7	3.3	6.0	4.1	2.2	9.8	3.1	5.6	3.9
D	up to 500 (151)	8.7	52.6	10.2	31.2	17.4	2.9	28.6	3.6	15.9	6.5
	500-1000 (65)	5.9	29.8	6.8	13.4	9.9	3.7	21.2	4.6	9.2	6.7
	1000-2000 (51)	4.3	26.1	6.6	10.3	7.7	2.8	18.3	4.3	7.0	5.2
	2000-5000 (34)	3.9	18.6	5.1	8.4	6.8	2.8	14.5	3.9	6.5	5.2
	5000-10000 (22)	3.0	11.8	4.0	6.4	5.9	2.2	8.2	3.1	5.3	4.2
	10000-30000 (27)	2.6	14.2	3.1	4.8	4.6	2.4	12.8	2.8	4.1	3.7
	more than 30000 (4)	2.3	10.7	3.3	6.0	4.1	2.2	10.1	2.5	4.1	3.2

[3] The number of municipalities falling in each demographic size class are reported in brackets

Table 8

AARMSE for the in-sample domains by demographic size with LM model

Auxiliary Variables	Target Variable[4]	Synthetic Projection					Composite Projection				
		Employed	Unemployed	Retired	Student	Other	Employed	Unemployed	Retired	Student	Other
A	up to 500 (151)	9.0	55.7	10.2	30.0	18.6	5.2	43.2	5.3	19.2	10.5
	500-1000 (65)	6.1	30.7	7.0	14.5	10.5	4.8	27.6	5.5	13.1	8.1
	1000-2000 (51)	4.4	27.1	6.7	11.1	8.4	3.5	23.2	4.9	9.4	6.3
	2000-5000 (34)	4.0	19.1	5.4	9.0	7.5	3.3	17.4	4.5	8.2	6.4
	5000-10000 (22)	3.0	13.1	4.4	7.4	6.4	2.6	11.9	3.8	6.7	5.3
	10000-30000 (27)	2.9	15.7	3.6	5.9	5.5	2.7	15.0	3.4	5.7	5.2
	more than 30000 (4)	2.5	13.1	3.8	5.6	5.6	2.4	12.2	3.7	5.3	5.3
B	up to 500 (151)	8.9	59.7	10.1	32.8	18.8	5.1	43.5	5.3	26.4	10.5
	500-1000 (65)	5.8	31.0	6.8	14.7	10.8	4.6	27.4	5.4	13.1	8.3
	1000-2000 (51)	4.1	25.8	6.6	11.4	8.3	3.3	22.2	4.8	9.5	6.2
	2000-5000 (34)	3.7	18.0	5.0	8.7	7.3	3.1	16.4	4.2	7.9	6.2
	5000-10000 (22)	2.8	11.5	4.1	7.5	5.9	2.4	10.7	3.5	6.8	4.9
	10000-30000 (27)	2.8	15.1	3.3	5.1	4.8	2.6	14.4	3.1	5.0	4.6
	more than 30000 (4)	2.6	13.4	3.1	4.6	4.5	2.5	12.7	2.9	4.4	4.3
C	up to 500 (151)	9.0	54.8	10.1	31.6	18.1	5.2	43.1	5.3	27.9	10.3
	500-1000 (65)	6.0	30.8	7.0	14.7	10.5	4.8	27.6	5.5	13.2	8.2
	1000-2000 (51)	4.4	26.7	6.7	11.2	8.4	3.5	23.0	5.0	9.5	6.4
	2000-5000 (34)	4.0	19.0	5.4	8.8	7.6	3.3	17.3	4.6	8.1	6.4
	5000-10000 (22)	3.2	12.8	4.4	7.3	6.5	2.7	11.6	3.8	6.5	5.4
	10000-30000 (27)	2.8	15.5	3.6	5.7	5.2	2.7	14.7	3.4	5.5	4.9
	more than 30000 (4)	2.5	12.9	3.8	5.6	5.6	2.4	12.0	3.6	5.3	5.3
D	up to 500 (151)	8.8	59.0	10.1	34.7	18.6	5.1	43.4	5.2	32.1	10.5
	500-1000 (65)	5.7	30.9	6.8	14.8	10.8	4.6	27.4	5.4	13.1	8.3
	1000-2000 (51)	4.1	25.3	6.5	11.4	8.3	3.3	21.9	4.8	9.5	6.2
	2000-5000 (34)	3.7	17.8	5.0	8.5	7.3	3.1	16.3	4.2	7.8	6.2
	5000-10000 (22)	2.9	11.2	4.0	7.2	5.8	2.5	10.3	3.5	6.6	4.8
	10000-30000 (27)	2.8	14.9	3.3	5.1	4.6	2.6	14.2	3.1	4.9	4.4
	more than 30000 (4)	2.6	13.2	3.0	4.7	4.5	2.5	12.4	2.9	4.5	4.3

[4] The number of municipalities falling in each demographic size class are reported in brackets

Table 9

AARMSE for the in-sample domains by demographic size with GLM model

Auxiliary Variables	Target Variable[5]	Synthetic Projection					Composite Projection				
		Employed	Unemployed	Retired	Student	Other	Employed	Unemployed	Retired	Student	Other
A	up to 500 (151)	8.8	53.5	10.2	31.6	17.6	5.1	42.5	5.3	26.5	10.2
	500-1000 (65)	5.9	30.3	6.8	13.5	9.9	4.7	27.3	5.4	12.4	7.8
	1000-2000 (51)	4.3	26.9	6.6	10.4	7.7	3.4	23.1	4.9	9.0	6.0
	2000-5000 (34)	3.9	19.1	5.1	8.6	6.8	3.3	17.4	4.3	7.8	5.7
	5000-10000 (22)	3.0	12.8	4.0	6.7	5.8	2.6	11.7	3.5	6.1	4.8
	10000-30000 (27)	2.7	15.1	3.3	5.2	4.8	2.5	14.3	3.1	5.0	4.5
	more than 30000 (4)	2.4	11.7	3.3	6.2	4.2	2.3	11.0	3.2	5.8	4.0
B	up to 500 (151)	8.7	55.2	10.3	32.0	18.1	5.1	42.5	5.3	26.5	10.2
	500-1000 (65)	5.6	29.9	6.8	13.5	9.9	4.5	27.0	5.4	12.4	7.8
	1000-2000 (51)	4.0	25.9	6.5	10.5	7.5	3.3	22.2	4.8	9.1	5.8
	2000-5000 (34)	3.7	18.1	4.9	8.3	6.5	3.1	16.5	4.1	7.6	5.4
	5000-10000 (22)	2.8	11.0	3.8	6.7	5.2	2.4	10.3	3.3	6.1	4.3
	10000-30000 (27)	2.6	14.6	3.1	4.7	4.2	2.5	13.9	2.9	4.6	4.0
	more than 30000 (4)	2.6	12.5	2.8	4.8	3.7	2.4	11.8	2.7	4.6	3.5
C	up to 500 (151)	8.8	52.9	10.2	31.4	17.4	5.1	42.4	5.3	26.4	10.1
	500-1000 (65)	5.9	30.2	6.8	13.6	10.0	4.7	27.2	5.4	12.5	7.8
	1000-2000 (51)	4.3	26.5	6.6	10.4	7.8	3.4	22.9	4.9	9.0	6.1
	2000-5000 (34)	3.9	19.1	5.2	8.6	6.9	3.3	17.4	4.4	7.8	5.8
	5000-10000 (22)	3.1	12.6	4.1	6.6	6.0	2.6	11.4	3.5	6.0	5.0
	10000-30000 (27)	2.7	14.8	3.2	5.1	4.7	2.5	14.1	3.0	4.9	4.5
	more than 30000 (4)	2.4	11.5	3.3	6.1	4.2	2.3	10.7	3.2	5.8	4.0
D	up to 500 (151)	8.8	52.9	10.2	31.4	17.4	5.1	42.4	5.3	26.4	10.2
	500-1000 (65)	5.9	30.2	6.8	13.6	10.0	4.5	26.9	5.4	12.4	7.8
	1000-2000 (51)	4.3	26.5	6.6	10.4	7.8	3.3	22.0	4.8	9.1	5.8
	2000-5000 (34)	3.9	19.1	5.2	8.6	6.9	3.1	16.5	4.1	7.6	5.5
	5000-10000 (22)	3.1	12.6	4.1	6.6	6.0	2.4	10.0	3.3	6.0	4.4
	10000-30000 (27)	2.7	14.8	3.2	5.1	4.7	2.5	13.7	2.9	4.6	3.9
	more than 30000 (4)	2.4	11.5	3.3	6.1	4.2	2.4	11.6	2.7	4.6	3.5

[5] The number of municipalities falling in each demographic size class are reported in brackets

Table 10

AARB for the out-of-sample domains by demographic size with LM and GLM models

Auxiliary Variables	Target Variable[6]	LM					GLM				
		Employed	Unemployed	Retired	Student	Other	Employed	Unemployed	Retired	Student	Other
A	up to 500 (151)	11.2	100.0	17.1	55.0	73.3	12.3	101.8	16.2	80.4	59.4
	500-1000 (65)	10.2	120.9	8.8	57.1	27.1	10.2	82.0	8.5	64.1	22.0
	1000-2000 (51)	8.9	58.7	11.2	39.3	22.9	8.5	55.5	11.3	42.2	21.4
	2000-5000 (34)	7.1	43.3	8.2	23.2	14.1	6.8	42.4	8.2	22.5	13.3
	5000-10000 (22)	5.8	31.1	7.3	15.2	10.2	5.7	31.1	7.3	14.7	9.4
	10000-30000 (27)	4.5	23.9	6.3	10.4	8.2	4.3	23.8	6.2	10.1	7.4
	more than 30000 (4)	3.9	17.9	4.9	8.3	6.3	3.8	18.0	4.8	7.9	5.7
B	up to 500 (151)	10.6	101.1	16.4	55.7	75.9	12.0	102.2	15.6	80.4	60.6
	500-1000 (65)	10.4	158.2	8.3	59.4	25.4	10.4	85.1	8.1	65.1	22.6
	1000-2000 (51)	8.5	63.4	11.1	38.9	22.7	8.3	56.9	11.3	42.6	21.6
	2000-5000 (34)	6.9	44.8	8.0	23.5	14.1	6.7	43.2	8.2	22.7	13.3
	5000-10000 (22)	5.7	30.4	7.0	14.9	10.3	5.5	30.3	7.1	14.6	9.5
	10000-30000 (27)	4.1	23.2	6.1	10.2	8.2	3.9	23.1	6.1	9.9	7.2
	more than 30000 (4)	3.5	16.0	4.5	7.7	5.9	3.5	16.4	4.6	7.5	5.3
C	up to 500 (151)	10.4	99.7	17.4	54.2	74.6	11.6	101.6	16.6	80.4	57.8
	500-1000 (65)	10.5	101.4	8.6	54.4	25.1	10.4	81.5	8.4	63.3	22.0
	1000-2000 (51)	8.8	58.2	11.1	40.0	22.6	8.5	55.5	11.3	42.3	21.3
	2000-5000 (34)	7.2	43.1	8.2	23.4	14.1	6.9	42.4	8.3	22.5	13.4
	5000-10000 (22)	5.9	30.8	7.4	15.2	10.3	5.7	30.9	7.3	14.8	9.6
	10000-30000 (27)	4.5	23.9	6.4	10.4	8.4	4.3	23.8	6.3	10.1	7.5
	more than 30000 (4)	3.9	18.0	4.9	8.4	6.3	3.8	18.1	4.9	7.9	5.8
D	up to 500 (151)	10.3	101.1	16.4	54.8	78.7	11.6	101.6	16.6	80.4	57.8
	500-1000 (65)	10.4	118.7	8.3	47.7	24.9	10.4	81.5	8.4	63.3	22.0
	1000-2000 (51)	8.5	62.9	11.1	41.5	22.8	8.5	55.5	11.3	42.3	21.3
	2000-5000 (34)	6.9	44.8	8.0	23.5	14.0	6.9	42.4	8.3	22.5	13.4
	5000-10000 (22)	5.7	30.2	7.0	15.0	10.4	5.7	30.9	7.3	14.8	9.6
	10000-30000 (27)	4.1	23.1	6.1	10.1	8.1	4.3	23.8	6.3	10.1	7.5
	more than 30000 (4)	3.5	16.0	4.5	7.6	6.0	3.8	18.1	4.9	7.9	5.8

[6] The number of municipalities falling in each demographic size class are reported in brackets

Table 11
AARMSE for the out-of-sample domains by demographic size with LM model

Auxiliary Variables	Target Variable[7]	LM					GLM				
		Employed	Unemployed	Retired	Student	Other	Employed	Unemployed	Retired	Student	Other
A	up to 500 (151)	11.2	100.4	17.1	36.0	73.4	12.3	101.9	16.2	80.7	59.5
	500-1000 (65)	10.3	121.5	8.9	57.6	27.4	10.3	82.4	8.5	64.4	22.2
	1000-2000 (51)	9.0	59.4	11.3	40.2	23.0	8.6	56.0	11.4	42.4	21.5
	2000-5000 (34)	7.1	43.9	8.3	23.7	14.2	6.9	42.9	8.3	22.7	13.4
	5000-10000 (22)	5.9	31.8	7.4	15.6	10.3	5.7	31.6	7.3	14.9	9.5
	10000-30000 (27)	4.5	24.5	6.4	10.8	8.4	4.4	24.4	6.3	10.3	7.5
	more than 30000 (4)	4.0	18.6	5.0	8.7	6.4	3.8	18.6	4.9	8.2	5.8
B	up to 500 (151)	10.7	102.1	16.4	64.8	76.3	12.0	102.4	15.8	80.7	60.8
	500-1000 (65)	10.5	159.3	8.4	69.0	25.8	10.4	85.6	8.2	65.3	22.8
	1000-2000 (51)	8.6	64.2	11.2	38.2	22.8	8.4	57.5	11.4	42.8	21.7
	2000-5000 (34)	7.0	45.6	8.1	24.0	14.2	6.7	43.8	8.2	22.9	13.4
	5000-10000 (22)	5.7	31.3	7.1	15.4	10.5	5.6	31.0	7.2	14.8	9.6
	10000-30000 (27)	4.2	24.0	6.2	10.7	8.3	4.0	23.8	6.2	10.2	7.3
	more than 30000 (4)	3.6	16.9	4.6	8.3	6.1	3.5	17.3	4.6	7.8	5.4
C	up to 500 (151)	10.4	99.8	17.4	49.4	74.7	11.6	101.7	16.6	80.6	57.9
	500-1000 (65)	10.5	101.7	8.6	55.6	25.2	10.4	81.7	8.4	63.4	22.1
	1000-2000 (51)	8.9	58.6	11.1	40.4	22.7	8.6	55.8	11.3	42.4	21.4
	2000-5000 (34)	7.2	43.5	8.3	23.6	14.2	7.0	42.7	8.3	22.7	13.4
	5000-10000 (22)	6.0	31.3	7.4	15.5	10.4	5.8	31.3	7.3	14.9	9.6
	10000-30000 (27)	4.6	24.4	6.5	10.7	8.4	4.4	24.2	6.3	10.3	7.6
	more than 30000 (4)	4.0	18.5	5.0	8.8	6.4	3.8	18.6	5.0	8.1	5.8
D	up to 500 (151)	10.4	101.5	16.4	54.4	78.9	11.6	101.7	16.6	80.6	57.9
	500-1000 (65)	10.5	119.1	8.3	42.4	25.1	10.4	81.7	8.4	63.4	22.1
	1000-2000 (51)	8.6	63.4	11.1	42.2	22.9	8.6	55.8	11.3	42.4	21.4
	2000-5000 (34)	7.0	45.3	8.0	23.8	14.1	7.0	42.7	8.3	22.7	13.4
	5000-10000 (22)	5.8	30.8	7.0	15.3	10.5	5.8	31.3	7.3	14.9	9.6
	10000-30000 (27)	4.2	23.7	6.2	10.5	8.2	4.4	24.2	6.3	10.3	7.6
	more than 30000 (4)	3.6	16.7	4.6	8.0	6.1	3.8	18.6	5.0	8.1	5.8

[7] The number of municipalities falling in each demographic size class are reported in brackets

VI. Conclusions

37. The projection estimator is a very promising approach both in the synthetic and in the composite formulation, even if better results are obtained by the composite estimator. GLM outperforms LM results in terms of AARB and ARRMSE. The auxiliary information specified by set D brings to more favourable results for both in-sample domains and out-of-sample.

38. Most of the results seem to be good enough to be published, since the mse values reported in the previous tables are below the threshold of 33% commonly used to decide if disseminate or not estimates of labour market variables.

39. Further developments will focus on the evaluation of more complicated models in order to improve the quality of the estimation process. In particular, the models that will be investigated are multiway linear models, linear and generalized linear mixed models in which specific domain effect are included in the model specification. To this aim new simulation studies will be carried out, considering hypercubes not considered in this work.

References

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