sciendo

Journal of Official Statistics, Vol. 37, No. 3, 2021, pp. 655–671, http://dx.doi.org/10.2478/JOS-2021-0029

Optimal Sampling for the Population Coverage Survey of the New Italian Register Based Census

Paolo Righi[1], *Piero Demetrio Falorsi*[2], *Stefano Daddi*[1], *Epifania Fiorello*[1], *Pierpaolo Massoli*[1], *and Marco Dionisio Terribili*[1]

For the first time in 2018 the Italian Institute of Statistics (Istat) implemented the annual Permanent Population Census which relies on the Population Base Register (PBR) and the Population Coverage Survey (PCS). This article provides a general overview of the PCS sampling design, which makes use of the PBR to correct population counts with the extended dual system estimator (Nirel and Glickman 2009). The sample allocation, proven optimal under a set of precision constraints, is based on preliminary estimates of individual probabilities of over-coverage and under-coverage. It defines the expected sample size in terms of individuals, and it oversamples the sub-populations subject to the risk of under/over coverage. Finally, the article introduces a sample selection method, which to the greatest extent possible satisfies the planned allocation of persons in terms of socio-demographic characteristics. Under acceptable assumptions, the article also shows that the sampling strategy enhances the precision of the estimates.

Key words: Population census; balanced area sample; capture-recapture estimator; administrative data; sample allocation.

1. Introduction

The Italian National Statistical Institute is currently engaged in a modernization program involving a vital revision of the approach traditionally embraced for the production of official statistics. In accordance with the European Statistical System commitment to Vision 2020 and UNECE – High-level group on Modernization of official statistics (Istat 2016), the program aims to enhance the Italian Integrated System of Statistical Registers (ISSR). The ISSR is the result of the integration of administrative with survey data (Alleva 2017). An important part of the ISSR is the Population Base Register (PBR), which updates the list of administrative households and people each month.

In this new environment, Istat has started to implement the Permanent Census in order to count the people living in their usual places of residence (or living people) at the municipality level. The use of the PBR as a unique tool to implement the Permanent Census incorporates some errors in counting living people mainly for two reasons. First, the PBR is not able to correctly register the movements of people among the municipalities, the numbers of immigrated and emigrated people, births and deaths. Moreover, the archives, which feed the PBR, have non-statistical purposes. Thus, the

[1] Italian National Statistical Institute, (Istat), Via Balbo 16 Rome 00184, Italy. Emails: parighi@istat.it, daddi@istat.it, fiorello@istat.it, pimassol@istat.it, and terribili@istat.it
[2] Former Italian National Statistical Institute, (Istat) Email: piero.falorsi@gmail.com

registered place of residence could differ from the usual place of residence (for example, for reasons related to paying taxes). These above errors introduce under-coverage and over-coverage into the living population counts. To overcome these problems, the Census Population Coverage Survey (PCS) is carried out each year as a component of the Permanent Census Survey System. The PCS selects an area sample in order to estimate the coverage problems, which affect the PBR. It implements a two-step process. The objective of the first step is to estimate the under-coverage using a capture/recapture estimator (Wolter 1986) and the second step calculates the over-coverage by following up on the list of persons captured and recorded in the PBR, but not recaptured in the first step of the process (Nirel and Glickman 2009).

The Permanent Census incorporates the results of the PCS into the PBR by defining individual coverage weights. The sum of these weights over the register units yields unbiased estimates of the living population.

The Permanent Census also includes a list survey which observes some target variables for the census statistics. The two surveys together (the PCS and the list survey) represent the Master Sample of the social surveys system of Istat, and the Permanent Census' annual counts essentially become the basis for demographic statistics. This entire process thus facilitates both the complete integration and the convergence of three statistical information systems (census counts, demographic statistics and social surveys). Until now, these three systems were completely independent of one another, creating an additional risk of producing inconsistent statistical figures.

This article focuses on the planning of a sample design of the PCS, optimal for estimating the living population at municipality level under a given set of precision constraints. This sampling design is both challenging and innovative for different reasons. It uses the output of statistical models flexibly as auxiliary information for the sampling design definition (Section 2). Furthermore, the area sampling takes into account the indirect sampling of individuals for estimating the unknown population size (Section 3 and Section 4). Eventually, the selection of the areas uses the balanced sampling method to guarantee as far as possible that the planned optimal allocation in terms of persons by socio-demographic characteristics is satisfied (Section 5). Section 6 shows some evidence of the realized sampling process and presents the first results of collection from the field. Finally, Section 7 summarises the main findings and provides some conclusions.

2. Parameters of Interest

The target parameters of the PCS are different characteristics of the living population for different geographical domains and demographic sub-classes.

To expedite comprehension of this work, here below we will concentrate on the two primary target parameters to be estimated at the municipality level.

These are: the number of people living in a given municipality, denoted as N_L, and the proportion of living people in the municipality over those counted in the register, indicated in the following as $P_{L|R}$. Both parameters are highly relevant. N_L is the primary input for defining the legal population, which determines the number of parliamentarians for specific areas. $P_{L|R}$ represents a significant quality measure of the PBR.

2.1. The Formal Definition of the Target Parameters

Let $\mathcal{g}$ be a variable which divides the living and the registered population into G socio-demographic groups (or classes). The $\mathcal{g}$ variable identifies the individual demographic profile for defining the sub-populations or *strata* where the probabilities of over-coverage are homogeneous (Nirel and Glickman 2009) and where the assumptions of the capture/recapture model hold (Wolter 1986). We summarise the statistical process for arriving at the definition of this variable in Subsection 2.2. In the present section, we note that proper identification of the $\mathcal{g}$ variable is crucial for enhancing the quality of the overall statistical process and for improving the efficiency of the sampling strategy.

Based on the partition induced by $\mathcal{g}$, we express N_L as:

$$N_L = \sum_{g=1}^{G} N_{g,L} = \sum_{g=1}^{G} N_{g,R} \frac{N_{g,RL}/N_{g,R}}{N_{g,RL}/N_{g,L}} = \sum_{g=1}^{G} N_{g,R} \frac{P_{g,L|R}}{P_{g,R|L}} = \sum_{g=1}^{G} N_{g,R}\, d_g,$$

where, regarding the class g ($g = 1 \ldots, G$) of $\mathcal{g}$: $N_{g,L}$ is the number of living people; $N_{g,R}$ denotes the total number of people in the PBR; $N_{g,RL}$ indicates the total number of living people who are correctly included in the register; $P_{g,L|R} = N_{g,RL}/N_{g,R}$ is the proportion of living people in PBR; $P_{gR|L} = N_{g,RL}/N_{g,L}$ is the proportion of living people in PBR with respect to all the living people; $d_g = P_{gR|L}/P_{gR|L}$ are the coverage weights. Note that $\left(1 - P_{g,L|R}\right)$ is the over-coverage proportion of PBR and $(1 - P_{g,R|L})$ is the under-coverage proportion of PBR.

In the same way, we have

$$P_{L|R} = \frac{N_{RL}}{N_R} = \frac{1}{N_R} \sum_{g=1}^{G} N_{g,R} P_{g,L|R},$$

being $N_R = \sum_g N_{g,R}$.

2.2. Definition of the $\mathcal{g}$ Variable

The definition of the $\mathcal{g}$ variable is a complex process in which both results from statistical models as well as practical issues related to the feasibility of the PCS sampling design are taken into consideration. At the end of this process, we can state that definition of the $\mathcal{g}$ variable has been driven by four *guiding principles*, herein listed as:

1. the $\mathcal{g}$ variable should be somehow *predictive* of both phenomena of *over-coverage* and *under-coverage* affecting the quality of the register,
2. The register data should permit the calculus of the specific $\mathcal{g}$ value for each unit in the register,
3. The partitioning into G classes must be parsimonious, and
4. The partitioning criteria ought to be robust as a result of studying the coverage of different observational data sets with various statistical models.

The rationale of these principles is the following: the word "somehow" in the first principle means that the definition of $\mathcal{g}$ is guided not only by statistical model considerations. The models represent *working* tools, useful for choosing the variables to be involved in the calculus of $\mathcal{g}$. The basic implication of the second principle is that only register variables should be used for defining $\mathcal{g}$. The third principle states that partitioning

into G classes should not define population subgroups that are so small that they would impose constraints on the sample sizes, making them too challenging to handle in the phase of sampling design (see Section 5). The fourth principle affirms that the efficacy of the definition of the g variable should be verified with more than one experiment.

The process of building the g variable starts from examining the results of two logistic models, one for under-coverage and the other for over-coverage. The models use the *ad hoc* 2015 survey data based on a sample of about 145,000 people in 800 enumeration areas (areas defined at sub-municipality level) located in more than 80 municipalities (Mancini and Ronconi 2017). The data collection for this survey adopts the two-step process. In the first step, the capture is the 2015 administrative register listing the people in the sample areas, and the recapture occurs in the field survey. The second step for identifying over-coverage is a follow-up of the sub-set of non-recaptured people in the first step. The people not recaptured in the follow-up represent cases of over-coverage.

The models utilize the auxiliary variables available within the register. These are both at the individual and the area level. The variables at the individual level are *Gender*, *Age-class*, *Citizenship*, *Type of household* (one-component/more than one-component). Those at the area level are *Type of census enumeration area* (urban/rural), *Size of the municipality*, *Macro Region*.

Table 1 presents the salient results of using the two models in terms of the conditional odds ratio related to the control class.

An odds ratio of 1 indicates that the event being studied (under-coverage or over-coverage) is equally likely to occur in both classes; that is, the interest class and the control class. An odds ratio greater than 1 indicates that the under-coverage or over-coverage is more likely to occur in the interest class. An odds ratio lower than 1 indicates that the under-coverage or over-coverage is more likely to occur in the control class. Analysis of the odds ratio reveals that the larger positive values are for people living in one-component households and for foreigners. Also, the *Age-class* variable shows a significant difference between each class and the control class (50–64). Values larger than 1 for younger people signal a higher probability of both under-covered and over-covered. On the other hand, the register covers people in the 65 + class better. Further results, not presented in the article, show that the effect of each age class on coverage is statistically significant (at the 5% level). The *Gender* variable here has only a minor impact on coverage. Specifically, the odds ratio value is not statistically significant for under-coverage. The odds ratio for the *Municipality by population size* variable is significantly smaller than one, which means that larger cities ($>50{,}000$) have greater under-coverage and over-coverage. However, if we consider the odds for the municipalities with less than 50,000 inhabitants, coverage is similar. Finally, the *Type of enumeration area* and *Macro-region* variables do not reveal a statistically significant impact. We used the data of the 2011 Census Post Enumeration Survey (PES) to validate the robustness of these models. The PES estimated the under-coverage and identified the *Age-class* and *Citizenship* as the variables affecting the coverage. The two models treated above include *Household type* as an explicative variable.

For reasons of the fourth principle (that of parsimony), we shall not consider other variables. Furthermore, we implement a further step of aggregation of age classes from five to three since the former number generates an overly detailed stratification, and

Table 1. Odds ratio of logistic models for under-coverage and over-coverage prediction according to the 2015 survey.

Variable	Classes*	Under-coverage model conditional odds ratio**	Over-coverage model conditional odds ratio**
Gender	Male	0.959	**0.880**
	Female (control)	–	–
Age-class	0–14	**2,056**	**2,038**
	15–29	**3,348**	**2,539**
	30–49	**2,910**	**2,915**
	50–64	–	–
	65+	**0,459**	**0,401**
Household type	One-component	**4,681**	**3,055**
	More than one-component (control)	–	–
Citizenship	Foreigner	**2,269**	**4,586**
	Italian	–	–
Enumeration area type	Urban	0,644	0,754
	Not-Urban (control)	–	–
Municipality by population size	$<1{,}000$	0,470	0,438
	1,000–4,999	**0,359**	**0,322**
	5,000–9,999	0,637	**0,359**
	10,000–49,999	**0,337**	0,389
	$>50{,}000$ (control)	–	–
Macro region	Islands	1,366	0,596
	South	1,166	0,712
	Centre (control)	–	–
	North-East	1.105	0.738
	North-west	0.653	0.817

*Corner point Parametrization: control case. Gender: Female; Age class: 50–64; Household type: more than one component; Citizenship: Italian; Municipality population: >50,000; Macro Region: Center.
**Bold values: P value minor than significance level of 5%

sample allocation should not be unduly complicated. We combine the first three classes characterized by higher under-coverage and over-coverage and define the 0–49 class. This class is vast. It covers people in very different stages of their life cycle, and the reasons that affect the coverage of these groups are different. Nevertheless, they have a roughly similar level of coverage rate, and this represents an essential property for the scope of our research. At the end of this process, we identify $G = 12$ classes obtained by the cross-classification of the *Age-class* (0–49; 50–64; + 65), *Citizenship* (Italian/Foreigner), and *Type of household* (one-component /more than one-component). These are then the strata of the sampling design, and we note these strata overlap the post-strata used in the estimation phase.

3. General Description of the PCS Sampling Design and Estimators

The first cycle of the Permanent Census considers the years ranging from 2018 to 2021. The PCS and the List Survey (LS) serve to support the Permanent Census. The PCS in that it selects a sample of areas to correct the coverage problems of the PBR. The LS in that it draws a sample of households from the list of households in the PBR and collects the census target variables which are not recorded in the PBR.

The two surveys are conducted each year according to a two-stage sampling design, in which the Primary Sampling Units (PSUs) are Italian municipalities. For each year of the census cycle, both the PCS and the LS share the same sample PSUs. Nevertheless, the samples of households in the two surveys are negatively coordinated one to the other and for different survey occasions.

3.1. The Sampling Design

From this point onwards, we limit ourselves exclusively to a description of the sampling design of the PCS. The sampling design splits the PSUs into:

1. Self-Representative PSUs selected each year including 600 municipalities with a population of more than 18,000 inhabitants and about 545 municipalities drawn from the Labor Force Survey;
2. Non-Self-Representative PSUs, roughly 6,840 municipalities. Within each province, these are stratified with reference to homogeneous population size strata. Each stratum has four PSUs, and one PSU is drawn from each stratum each year using simple random sampling so that, within a four-year period of time, all the Italian municipalities have been observed.

As a general rule, the Secondary Sampling Units (SSUs) are street address numbers (hereinafter addresses) in urban areas and Enumeration Area (EAs), which are clusters of addresses, in non-urban areas. The EAs are the SSUs also in urban areas for about 15% of small municipalities, where a significant part of the addresses in the territory are poorly identified. It is worth noting that the EAs constitute the most detailed partitioning of the national territory and that an EA cannot overlap into two or more municipalities. Two primary considerations have driven the definition of two different types of SSUs. The first is that there is a more precise definition and localization of addresses within densely populated urban areas and higher risk of address errors in non-urban areas. The second has to do with the reduction of the cost of an interview: it is significantly more cost-effective to collect data for more than one address in cases where the interviewer has to arrive at remote non-urban areas.

The generic PSU is indicated with the subscript i. The PCS sampling frame lists the SSU codes with associated geographical references (region, municipality, and so on) and the registered number of people in the G classes $N_{i,g,R}$, being $\sum_i N_{g,i,R} = N_{g,R}$. In the following, the vector $\boldsymbol{N}_{i,R} = (N_{i,1,R}, \ldots, N_{i,g,R}, \ldots, N_{i,G,R})'$ is specified as the SSU *profile*.

The sampling frame integrates the Geographical Base Register – listing the Italian addresses and any geographical references such as municipality and EA codes – with the PBR. In this integration, a data cleaning process is carried out to delete all the non-residential addresses such as markets, shops, garages, and so on.

In the 2018 PCS round, 2,745 municipalities (containing approximately 43 million people, 72% of the country's population) are sampled using the area sampling design while the PCS considers all the areas in 107 small-scale municipalities. In 2,745 municipalities, the sample selection includes approximately 166,000 addresses and 3,450 EAs. The expected sample size, given by the available PBR information is around 935,000 people (about 2.2% of the sampling fraction).

The sample selection of SSUs from each municipality is realized by means of a three-step process, with the aim of obtaining the optimal sampling of individuals:

1. *Definition of the optimal sample sizes in terms of individuals, $n_{g,R}$, for each subclass g* ($g = 1, \ldots, G$). This is described in Section 4,
2. *Definition of the inclusion probabilities of SSUs* by a calibration process ensuring that the expected sample sizes for each subclass g are equal to the optimal $n_{g,R}$. This is described in Subsection 5.1, and
3. *Selection of a balanced sample* of SSUs utilizing the cube algorithm (Tillé 2006), ensuring that the number of registered people in the selected sample of SSUs equals the optimal sample sizes. This is described in Subsection 5.2.

All the living households are observed in each SSU, and every individual who is a member of the household is interviewed. In the first step of the survey, the interviewer does not have access to the list of the registered population of the SSU. In the over-coverage follow-up, interviewers for a given SSU investigate whether the registered households which they did not find in the first step are living in the SSU or not.

3.2. Basic Elements of the Estimator

Considering the PBR as the first capture and the PCS as the second capture (or recapture), N_L can be estimated by applying the Extended Dual System Estimator (Nirel and Glickman 2009) given by

$$\hat{N}_L = \sum_{g=1}^{G} \hat{N}_{g,L} = \sum_{g=1}^{G} N_{g,R} \frac{\hat{P}_{g,L|R}}{\hat{P}_{g,R|L}} = \sum_{g=1}^{G} N_{g,R}\, \hat{d}_g, \tag{1}$$

where $\hat{P}_{gL|R}$ and $\hat{P}_{g,R|L}$ are the sample estimates of the proportions $P_{gL|R}$ and $P_{g,R|L}$ being that $\hat{d}_g = \hat{P}_{g,L|R}/\hat{P}_{g,R|L}$. Note that in the PBR each individual of a given municipality in a g class has the same weight, $\hat{d}_g$, which is to be included in the PBR at the individual level. This weight is *unique* for that individual. All the estimates of living populations for different domains or sub-classes will be produced directly from PBR using these weights. The use of a unique weight ensures internal coherence of the estimates of the population size of different population sub-groups.

Comparable to Equation (1), we can also estimate $\hat{P}_{L|R}$ as:

$$\hat{P}_{L|R} = \frac{1}{N_R} \sum_{g=1}^{G} N_{g,R}\, \hat{P}_{g,L|R}.$$

The PCS allows for the building of unbiased estimates of both $\hat{P}_{g,R|L}$ and $\hat{P}_{g,L|R}$ ratios.

In its first step, the PCS is carried out using the usual capture/recapture technique. To illustrate this step with regard to sub-group g, let

$$S_{g,(1)RL} = S_{g,L} \cap S_{g,R},$$

denote the sampling intersection set, where $S_{g,L}$ identifies the set of individuals detected as living and belonging to the addresses and EAs selected in the sample, and $S_{g,R}$ identifies the set of individuals recorded in the register as belonging to the addresses and EAs selected in the sample. We compute $\hat{P}_{g,R|L}$ as

$$\hat{P}_{g,R|L} = \frac{\hat{N}_{g,(1)RL}}{\hat{N}_{g,(1)L}},$$

where: $\hat{N}_{g,(1)RL}$ is the estimate of $N_{g,RL}$ obtained by summing the PCS sampling weights over the units in $S_{g,(1)RL}$, and $\hat{N}_{g,(1)L}$ is the rough estimate of $N_{g,L}$ obtained by summing the PCS sampling weights over the units in $S_{g,L}$.

The ***second step*** of the PCS is a follow-up on the sampling set $S_{g,(2)R} = S_{g,R} \backslash S_{g,(1)RL}$, which includes the people belonging to the register and not found as belonging to the living population in the first step. These individuals are observed in the second step with different data collection techniques. At the end of this process, we identify a subset of them,

$$S_{g,(2)RL} = S_{g,(2)R} \cap S_{g,L},$$

as belonging to the living population. We then estimate the probability $\hat{P}_{g,L|R}$ as

$$\hat{P}_{g,L|R} = \frac{\hat{N}_{g,(1)RL} + \hat{N}_{g,(2)RL}}{\hat{N}_{g,R}},$$

where: $\hat{N}_{g,R}$ is the estimate of $N_{g,R}$ obtained by summing up the PCS sampling weights over people of the sample set $S_{g,R}$; $\hat{N}_{g,(2)RL}$ is obtained by summing the PCS sampling weights over the units of the set $S_{g,(2)RL}$. $\hat{N}_{g,(2)RL}$ represents an additional estimate of the living population included in the register obtained from the follow-up operations of the second step.

4. Sample Size Allocation of Individual Sub-Groups at the Municipality Level

The sample allocation problem at the municipality level centers on the definition of the vector $\boldsymbol{n}_R = \{ n_{g,R} : g = 1, \ldots, G \}$ of expected sample sizes for population sub-groups. This vector is the most influential factor in calculating the cost of doing surveys and it determines the payments for interviewers. Moreover, as shown below, it has a direct impact on the expected precision of the estimates.

As shown in Section 5, the numbers of sampled addresses and EAs are a by-product derived from $\boldsymbol{n}_R$.

To define $\boldsymbol{n}_R$, we solve the Optimization Problem (OP) illustrated in Falorsi and Righi (2015). The OP has the objective function of minimizing the number of interviews in a municipality, which is equivalent to cutting survey costs. Moreover, the OP defines as constraints the requirement that the sampling variances of the target estimates ($\hat{N}_L$ and $\hat{P}_{L|R}$) are lower than the prefixed thresholds. Defining the sampling variances $V(\hat{N}_L)$ and $V(\hat{P}_{L|R})$ as a function of the vector $\boldsymbol{n}_R$ is the first step to making the OP operational. For $V(\hat{N}_L)$ we consider the approximated design variance using the Taylor linearization method (Pfeffermann 2015):

$$V(\hat{N}_L) = \sum\nolimits_g N_{g,R}^2 \left[\frac{V(\hat{P}_{g,L|R})}{\left[E(\hat{P}_{g,R|L})\right]^2} + \frac{\left[E(\hat{P}_{g,L|R})\right]^2}{\left[E(\hat{P}_{g,R|L})\right]^4} V(\hat{P}_{g,R/L}) \right]. \quad (2)$$

The variances $V(\hat{P}_{g,L|R})$and $V(\hat{P}_{g,R|L})$ in Equation (2) are those of a two-stage cluster sampling design. We may obtain their values by starting from the sampling variances of a Simple Random Sampling (SRS) design – respectively of $n_{g,R}$ registered units and $n_{g,L}$ living units. We then multiply the latter variances by the *design effect*, *deff*, (Kish 1965; Cochran 1977) which is the ratio of the sampling variance of a complex sampling design over that of an SRS. Thus, we have:

$$V(\hat{P}_{g,L|R}) = \left[\frac{P_{g,L|R}(1-P_{g,L|R})}{n_{g,R}}\right] deff, \tag{3}$$

$$V(\hat{P}_{g,R|L}) = \left[\frac{P_{g,R|L}(1-P_{g,R|L})}{n_{g,L}}\right] deff \cong \left[\frac{P_{g,R|L}(1-P_{g,R|L})}{n_{g,R}} \frac{E(\hat{P}_{g,R|L})}{E(\hat{P}_{g,L|R})}\right] deff, \tag{4}$$

being $n_{g,L} \cong n_{g,R}\left[E(\hat{P}_{g,L|R})/E(\hat{P}_{g,R|L})\right]$.

Plugging the Equations (3) and (4) into (1), we obtain:

$$V(\hat{N}_L) = \sum\nolimits_g N_{g,R}^2 \frac{1}{n_{g,R}} \left[\frac{P_{g,L|R}(1-P_{g,L|R})}{\left[E(\hat{P}_{g,R|L})\right]^2} + \frac{E(\hat{P}_{g,L|R})}{\left[E(\hat{P}_{g,R|L})\right]^3} P_{g,R|L}(1-P_{g,R|L})\right] deff, \tag{5}$$

In order to make Equation (4) computable for the OP, we considered the estimates $\hat{P}_{g,L|R}$, $\hat{P}_{g,R|L}$ and $\widehat{deff}$ respectively of $P_{g,L|R}$, $P_{g,R|L}$ and *deff*.

The use of the parameters from the logistic models (Subsections 2.2) gives us the estimates of $\hat{P}_{g,L|R}$ and $\hat{P}_{g,R|L}$. We compute the $\widehat{deff}$ by simulating the actual sampling design on the data of the previous Population Census. We obtain the plug-in estimator of $V(\hat{N}_L)$ as

$$\hat{V}(\hat{N}_L) = \sum\nolimits_g N_{g,R}^2 \frac{1}{n_{g,R}} \left[\frac{\hat{P}_{g,L|R}(1-\hat{P}_{g,L|R})}{(\hat{P}_{g,R|L})^2} + \frac{\hat{P}_{g,L|R}}{(\hat{P}_{g,R|L})^3} \hat{P}_{g,R|L}(1-\hat{P}_{g,R|L})\right] \widehat{deff}. \tag{6}$$

The computable expression of $\hat{V}(\hat{P}_{L|R})$, used as input in the OP, is

$$\hat{V}(\hat{P}_{R|L}) = \frac{1}{N_R^2} = \sum\nolimits_g N_{g,R}^2 \frac{1}{n_{g,R}} \hat{P}_{g,R|L}(1-\hat{P}_{g,R|L}) \widehat{deff}. \tag{7}$$

Thus, based on Equations (6) and (7), the OP for defining the n_R is given by:

$$\begin{cases} min\left(n_R = \sum_{g=1}^{G} n_{g,R}\right) \\ \sum_{g=1}^{G} N_{g,R}^2 \frac{1}{n_{g,R}} \left[\frac{\hat{P}_{g,L|R}(1-\hat{P}_{g,L|R})}{(\hat{P}_{g,R|L})^2} + \frac{\hat{P}_{g,L|R}}{(\hat{P}_{g,R|L})^3} \times \hat{P}_{g,R|L}(1-\hat{P}_{g,R|L})\right] \widehat{deff} \le V_1^*, \\ \frac{1}{N_R^2} \sum\nolimits_g N_{g,R}^2 \frac{1}{n_{g,R}} \hat{P}_{g,R|L}(1-\hat{P}_{g,R|L}) \widehat{deff} \le V_2^* \\ n_{g,R} \le N_{g,R}, \quad g = (1, \ldots, G) \end{cases} \tag{8}$$

Table 2. Coefficient of variation thresholds for implementing the allocation at municipality level.

Municipality by number of residents	<1,000	1,000–4,999.	5,000–19,999	20,00–49,999	50,000–799,999	>799,999
Estimates						
$\hat{N}_L$	0.20%	0.20%	0.20%	0.20%	0.09%	0.05%
$\hat{P}_{L\|R}$	0.40%	0.33%	0.25%	0.16%	0.09%	0.06%

where V_1^* and V_2^* are the pre-defined variance-thresholds and $N_{g,R}$ is the resulting total number of people (from the register) in subgroup g of the municipality. These thresholds, expressed in terms of coefficient of variations (Table 2), vary according to municipality size.

Further variance constraints have been set up at the regional level and national level. These influence the final solution only marginally. The final solution is mainly affected by constraints defined at the municipality level on $\hat{V}(\hat{N}_L)$. This is the main reason why we decided to focus this section on describing the constraints at the municipality level only.

Practical implementation of the optimal allocation algorithm has been carried out employing the open-source software MAUSS-R (Buglielli et al. 2013). This software automatically guarantees the respect of the constraints $n_{g,R} \leq N_{g,R}\, g = (1, \ldots, G)$ of Equation (8) reiterating the algorithm until the constraints are respected.

To explain more clearly how the System (1) works, it is useful to consider the univariate single domain-case in which the only target parameter is $\hat{N}_L$ in a given municipality. Let us reformulate the Equation (6), in the following way:

$$\hat{V}(\hat{N}_L) = \sum\nolimits_g N_{g,R}^2 \frac{1}{n_{g,R}} \hat{\sigma}_{g,\hat{N}_L}^2,$$

being

$$\hat{\sigma}_{g,\hat{N}_L}^2 = \left[\frac{\hat{P}_{g,L|R}\left(1 - \hat{P}_{g,L|R}\right)}{\left(\hat{P}_{g,R|L}\right)^2} + \frac{\hat{P}_{g,L|R}}{\left(\hat{P}_{g,R|L}\right)^3} \hat{P}_{g,R|L}\left(1 - \hat{P}_{g,R|L}\right) \right] \widehat{deff}.$$

Thus, the univariate OP is defined as

$$\begin{cases} min\left(n_R = \sum\nolimits_{g=1}^{G} n_{g,R} \right) \\ \sum\nolimits_g N_{g,R}^2 \frac{1}{n_{g,R}} \hat{\sigma}_{g,\hat{N}_L}^2 = V_1^* \\ n_{g,R} \leq N_{g,R},\ g = (1, \ldots, G) \end{cases}.$$

Solving the above OP, we obtain the classical *Neyman* solution:

$$n_R = \frac{\left(\sum\nolimits_{g=1}^{G} N_{g,R} \hat{\sigma}_{g,\hat{N}_L} \right)^2}{V_1^*}, \tag{9}$$

and

$$n_{g,R} = n_R \frac{N_{g,R}\hat{\sigma}_{g,\hat{N}_L}}{\sum_{g=1}^{G} N_{g,R}\hat{\sigma}_{g,\hat{N}_L}} \tag{10}$$

The actual sample sizes defined at municipality level are very close to those defined by Equations (9) and (10) since, as anticipated before, the constraints defined at the municipality level on $\hat{V}(\hat{N}_L)$ have the greatest impact on the final solution of the OP.

Finally, we notice that the actual final allocation is slightly different from the optimal allocation as described above since further operative constraints have been applied. First, we could have increased the sample sizes for the municipalities with less than 1,000 inhabitants since at least 100 households need to be interviewed. Then, we have carried out a traditional census in municipalities with a population of less than 300 inhabitants. Lastly, additional efforts can be made to smooth the solution towards proportional allocation wherever the optimal allocation is much too different from the proportional one.

5. Sample Selection at the Municipality Level

5.1. Definition of the Inclusion Probabilities of the SSU

The inclusion probability, π_i $(i = 1, \ldots, M)$, of the ith SSU in a given municipality is defined by the following calibration system ensuring that the expected sample sizes for each subclass g are equal to the optimal $n_{g,R}$:

$$\begin{cases} min\left(\sum_{i=1}^{M} D(\pi_i, \bar{\pi})\right) \\ \sum_{i=1}^{M} \pi_i N_{i,g,R} = n_{g,R}, \ \ (g = 1, \ldots, G), \\ L\bar{\pi} \leq \pi_i \leq U\bar{\pi} \text{ for } i = 1, \ldots, M \end{cases} \tag{11}$$

where: M is the number of SSUs in the municipality, $D(\pi_i, \bar{\pi})$ is the *truncated logarithmic distance function* (Singh and Mohl 1996) between π_i and $\bar{\pi} = n_R/N_R$ which is the initial inclusion probability to ensure that the expected sample size for the municipality is equal to the *optimal* one; $0 \leq L \leq 1$; and $U \geq 1$. Furthermore, the truncated logarithmic distance function ensures that the final inclusion probabilities are bounded in the interval $(L\bar{\pi}, U\bar{\pi})$. The problem of Equation (11) is solved with the software Regenesees (Zardetto 2020).

We note that the system (11) finds the final inclusion probabilities starting from an initial inclusion probability, $\bar{\pi}$, which ensures that each person in the gth class is selected in the sample with the same probability, irrespective of the SSU in which they are registered. This solution seems optimal since it minimizes the variance of the final weights (Kish 1965, sec. 11.7). However, the initial probability does not ensure the respect of the optimal $n_{g,R}$, $(g = 1, \ldots, G)$ sample sizes, since once that particular SSU i is selected in the sample, the whole vector $\boldsymbol{N}_{i,R} = (N_{i,1,R}, \ldots, N_{i,g,R}, \ldots, N_{i,G,R})'$ must be observed.

5.2. Selection of the SSUs

In all municipalities, the SSUs are selected with the cube algorithm (Tillé 2006), guaranteeing approximate respect of the following balancing equations

$$\sum_{i \in s} \boldsymbol{N}_{i,R} = (n_{1,R}, \ldots, n_{g,R}, \ldots, n_{G,R})', \tag{12}$$

s being the sample of the SSUs.

In practice, selecting a sample with this level of detail might be unfeasible, especially for smaller municipalities. Therefore, in these situations, starting from the constraints (12), we can define a new system with fewer equations. In order to accomplish this, we aggregate the $n_{g,R}$ optimal sample sizes into Γ classes (with $\Gamma < 12$), where the size of each class γ ($\gamma = 1, \ldots, \Gamma$) of the new classification is obtained by summing the sampling dimensions of the original grouping into G classes. Table 3 illustrates this aggregation scheme. We see that in municipalities with over 100,000 registered people, the original sample allocation with $G = 12$ is implemented; in municipalities with less than 100,000 registered people and sample size < 50 people, no allocation into classes is executed. The balancing equations control the sample size according to *Age-class* variable and separately by *Household type* variable in municipalities with a sample size between 50 and 150). Eventually, with a sample size of over 150 people and less than 100,000 registered people, allocation considers *Age-class*, *Household type* and *Citizenship* separately.

In a given municipality with less than 100,000 registered people, the cube algorithm selects the sample of SSUs, respecting the following balancing equations approximately,

$$\sum_{i \in s} \boldsymbol{N}_{i,\Gamma,R} = \boldsymbol{n}_{\Gamma,R},$$

in which $\boldsymbol{N}_{i,\Gamma R} = \left(N_{i,1,R}, \ldots, N_{i,\gamma,R}, \ldots N_{i,\Gamma,R}\right)'$ and $\boldsymbol{n}_{\Gamma,R} = \left(n_{1,R}, \ldots, n_{\gamma,R}, \ldots n_{\Gamma,R}\right)'$ are the population vector of the ith SSU and the sample allocation vector defined according to Table 3.

6. Some Findings from the Sampling Design of the 2018 PCS

We now analyze the sample, having considered the expected sample distributions and a rough index of accuracy related to the allocation (Subsection 6.1). Moreover, we make preliminary comparisons between the expected and the observed sample distributions (Subsection 6.2).

Table 3. Sample allocation classes by expected sample size $\boldsymbol{n_R}$ and population size $\boldsymbol{N_R}$.

Expected sample size n_R	Population size N_R	Sample allocation classes
<50		Expected overall sample size
50–150	$<100{,}000$	Age class + household type
>150		Age class + household type + citizenship
	100,000 and more	Age class + household type + citizenship

6.1. The Expected Sample Distributions and a Rough Index of the Accuracy

The proposed sample design aims to select a sample for controlling the number of people according to specific socio-demographic characteristics (Section 5) in order to minimize the sampling variance of the population count estimation. The optimal allocation reveals important differences with the proportional allocation. Table 4 gives the sampling distribution of the optimal allocation and the proportional-to-municipality population allocation by municipality size. The evidence indicates that the optimal allocation oversamples in the smallest municipalities due to the precision thresholds and the operative constraints applied to about 600 cities with less than 1,000 inhabitants. Samples allocated in the largest municipalities are relatively much smaller than those of proportional allocation since the sample size needed to respect precision constraints is somehow independent of the population size. This is the counterpart of the phenomenon observed in the smallest municipalities, where optimal allocation gives a sample more than four times greater than the proportional allocation.

Table 5 offers insights into the demographic distribution of the sample. Note that the proportional allocation would be the expected allocation if we selected a sample without demographic constraints. Comparison of the two allocations underlines that the optimal allocation oversamples, particularly people living in a one-component household, foreigners and the 0–49 age class.

Table 6 shows a *Raw Efficiency Index* (*REI*) calculated as

$$REI = Mean\left[\frac{\hat{V}(\hat{N}_L)|\,optimal\,allocation}{\hat{V}(\hat{N}_L)|\,proportional\,allocation}\right], \tag{13}$$

the average of the municipality ratios between the variances in Equation (5) computed with the optimal allocation and the proportional allocation. We denote this as "raw" since the index takes the parameters in the variance Equation (5) as true values. This is a strong assumption and we consider the result as a qualitative index.

The findings here highlight the improvement in efficiency of the proposed sampling strategy.

Table 4. Optimal population sample allocation versus proportional allocation by municipality size according to the PBR.

Municipality by size	Optimal expected sample allocation	Proportional expected sample allocation
<5,000	26.6%	6.6%
5,001–2,0000	24.2%	19.5%
2,0001–5,0000	12.3%	26.3%
5,0001–250,000	31.5%	26.5%
+250,000	5.4%	21.1%
Total	100.0%	100.0%
Sample size	**935,396**	**935,396**

Table 5. Optimal population sample allocation versus proportional allocation by demographic characteristics according to the PBR.

Characteristics	Optimal expected sample allocation	Proportional expected sample allocation
People living in a 1-component household	25.0%	15.5%
People living in a household with 2 or more components	75.0%	84.5%
Italian	87.2%	91.8%
Foreigners	12.8%	8.2%
0–49 age class	56.1%	42.0%
50–64 age class	20.4%	42.4%
+65 age class	23.5%	15.6%
Sample size	**935,396**	**935,396**

Table 6. Raw efficiency index (ratio of expected variance with optimal allocation and proportional-to-population size allocation) by municipality population size.

Municipality by size	REI*
$<10{,}000$	50.6%
10,000–100,000	48.4%
100,000–250,000	53.0%
$>250{,}000$	67.8%

*see Equation (13)

6.2. Evidence from the PBR and Field Data Collection

We compare the planned sample with the realized sample in PCS, highlighting discrepancies in the socio-demographic sample distributions. We recommend considering the results below with a degree of caution since there are some aspects not thoroughly addressed in this article. These include (1) the occurrence of non-responses in the PCS; (2) register addresses that cannot be identified in the territory, thereby reducing the number of sampled persons; (3) the difference between the reference time of the PBR used for the sample selection and the reference time of the PCS.

The oversampled categories such as people living in a one-component household and foreigners display significant differences between realized and expected sample size. The *Relative Percentage Difference Index* (RPDI), given by

$$RPDI = 100 \times \frac{\left(realized\ sample\ size - expected\ sample\ size\right)}{expected\ sample\ size}, \tag{14}$$

is, respectively, equal to -25.3% and 35.4% for the two classes (corresponding to approximately 61,000 and 23,000 persons). Although no conclusion can be drawn on the reasons for these results (PBR over-coverage, obsolete sampling frame, non-response in the PCS), the results are a wake-up call, and larger samples for these sub-populations

Table 7. Expected sample size by demographic distributions of the expected and observed sample sizes.

Characteristics	Expected sample size	Realised sample size	*RPDI***
People living in a 1 component household	25.0%	20.7%	−25.3%
People living in a household with 2 or more components	75.0%	79.3%	−4.8%
Italian	87.2%	90.7%	−6.2%
Foreigner	12.8%	9.2%	−35.4%
0–49 age class	56.1%	55.2%	−11.4%
50–64 age class	20.4%	20.8%	−8.1%
+65 age class	23.5%	24.0%	−7.9%
Sample size	**935,396**	**842,539***	**−9.9%**

* Preliminary value; ** see Equation (14)

facilitate the capture/recapture process and offer greater accuracy in the estimates. The proposed sample allocation achieves this condition.

Finally, Table 7 shows a slight change in the distribution by age class between expected and realized samples. Consequently, the sample size reduction has been distributed across the classes more uniformly, and the *RPDI* is less significant.

7. Conclusions

The article describes the sampling design of the 2018 Population Coverage Survey (PCS) of the new Italian Permanent register-based Population Census that will produce census counts every year. The primary source of the Permanent Census is the Population Base Register (PBR). The PBR is the final result of the integration of multiple administrative archives implemented by the Istat. The PBR lists the population according to administrative place of residence, whereas the census target population is the living population at the municipality level. The reasons why the two populations differ can depend on: administrative errors, non-updated information in the PBR, the personal interests of particular persons which have them registered in places that are not their living area. That said, the direct use of the PBR for census counts produces biased statistics, and the PCS is needed to provide unbiased estimates of population sizes. The sampling design of the PCS introduces both relevant and reliable innovations for dealing with the census coverage errors. We briefly indicate what these are: (1) the PCS selects areas with a different level of detail such as addresses and EAs. The choice depends on cost and operative constraints (addresses in urban areas and EAs in not urban areas) and on the need to control expected sample size. (2) The use of previous data (the 2011 Population Census and pilot studies) to set up predictive models of the individual probabilities to be under/over-covered and useful for defining optimal sample sizes for specific sub-populations. (3) The calculation of the inclusion probabilities of the Secondary Stage Units varies according to their socio-demographic profiles in such a way that the sample design

oversamples specific sub-populations subjected to the risk of under/over coverage. In particular, we verify that the allocation oversamples people living in a one-component household and foreigners. (4) The application of a sample selection method for Secondary Stage Units is based on balanced sampling, which ensures respect of the planned sample sizes for socio-demographic sub-groups.

When compared with a sample design that ignores registered demographic information at the area level, we have seen that, if the assumptions of the super-population models for coverage hold, the sample design we have described does improve the precision of sampling estimates.

While making the comparison between the realized sample sizes with the expected ones, we notice that the PCS has major issues with trying to recapture those persons that are oversampled by the sampling design. Therefore, the opportunity to have a larger sample size for these sub-populations improves the accuracy of the final census count estimates.

Finally, fieldwork operations confirm that the factors which have the greatest relative importance in explaining the structural differences between the registered and the usual place of residence within the country are *Household type* and *Citizenship*. In particular, people living in a one-component household and foreigners are the sub-populations most affected by coverage problems. The reasons for these structural differences are partially self-evident. Still, there is room for further empirical analysis to investigate whether or not there are interactions with other factors, such as size of the municipality or geographical area. We would use such findings to refine the proposed methodology.

8. References

Alleva, G. 2017. "The new role of sample surveys in official statistics." Paper presented at *5th Italian Conference on Survey Methodology* – Itacosm 2017, June 14–17, 2017, Bologna. Available at: https://www.istat.it/it/files//2015/10/Alleva_ITACOSM_14062017.pdf (accessed September 2019).

Buglielli, T., De Vitiis C., Barcaroli G. 2013. *MAUSS-R*. R package version 1.1. Available at: https://www.istat.it/en/methods-and-tools/methods-and-it-tools/design/design-tools/mauss-r (accessed June 2021).

Cochran, W.G. 1977. *Sampling Techniques*. Wiley. New York.

Falorsi, P.D., and P. Righi. 2015. "Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys." *Survey Methodology* 41: 215–236. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14149-eng.pdf (accessed June 2021).

Istat. 2016. *Istat's Modernisation Programme.* Available at: https://www.istat.it/en/files/2011/04/IstatsModernistionProgramme_EN.pdf (accessed June 2020).

Kish, L. 1965. *Survey Sampling*. New York: Wiley.

Mancini, L., and A. Ronconi. 2017. "La rilevazione sperimentale 2015 per l'indagine C-sample del censimento permanente della popolazione: copertura delle liste anagrafiche comunali e confronto con i risultati del 2011 (with abstract in English)." *Istat Working Paper 2017.* DOI: https://doi.org./10.13140/RG.2.2.28529.79202.

Nirel, R., and H. Glickman. 2009. "Sample Surveys and Censuses." In *Handbook of Statistics* ed. D. Pfeffermann and C.R. Rao. 539–565. Elsevier.

Pfeffermann, D. 2015. "Methodological Issues and Challenges in the Production of Official Statistics." *Journal of Survey Statistics and Methodology*, 3: 425–483. DOI: https://doi.org/10.1093/jssam/smv035.

Singh, A.C., and C.A. Mohl. 1996. "Understanding Calibration Estimators in Survey Sampling." *Survey Methodology* 22: 107–115. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1996002/article/2973-eng.pdf?st=8C3F8n32 (accessed June 2021).

Tillé, Y. 2006. *Sampling Algorithm*. New York: Springer.

Wolter, K.M. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association*, 81: 338–346. DOI: https://doi.org/10.1080/01621459.1986.10478277.

Zardetto, D. 2020. Package ReGenesees R Evolved Generalized Software for Sampling Estimates and Errors in Surveys. R package version 2.0. Available at: https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/regenesees (accessed June 2021).

Received September 2019
Revised April 2020
Accepted November 2020