

THE POST ENUMERATION SURVEY OF THE 15th ITALIAN POPULATION CENSUS: FEATURES AND METHODS¹

Matteo Mazziotta, Monica Russo

1. Introduction

The Post Enumeration Survey (PES) is regulated by the European Commission (1151/2010). Its main goals is to estimate the number of individuals actually and habitually dwelling in the reference time of the 15th Population Census (October 9, 2011), the coverage rate and the over-coverage rate. The PES is based on a stratified two-stage sampling design. The first stage selects a sample of 252 municipalities as primary sampling unit, stratified by Regions and five classes of demographic dimension of municipalities. The second stage selects a sample of about 2.500 enumeration areas, stratified by demographic dimension of enumeration areas.

The survey is designed to ensure independence with the Census. Individuals habitually dwelling in housing in the selected enumeration areas, at the time of the PES are being carried out. The survey was conducted from April to July 2012, and involved about 1,200 enumerators selected among the "best" of the Census (of course, each of them is assigned a different area than the Census).

The PES has enumerated approximately 329.000 eligible individuals; an accurate Record Linkage phase and the application of a complex estimation model (Dual System approach derived by ONS) will construct a solid study towards the continuous census. The paper wants to describe the methodological aspects of the most important Istat quality survey.

2. Features

In this section we outline the adopted estimation methodology for the coverage survey of the 15th Census of Population and Housing, and we want to focus² on the

¹ The paper is the result of combined work of the authors: Matteo Mazziotta has written Section 1; Monica Russo has written Sects. 2 and 3.

research that has been conducted to develop improvements and innovations in the estimation model and adjusting for over-coverage.

The coverage survey – or *Post Enumeration Survey* (PES) – is a survey conducted in connection with the 15th Census of Population and Housing (the reference date of both the Census and the PES is October 9, 2011) and is designed to produce an assessment of the level of accuracy of the Census through the determination of the over-count and under-count errors happened during the Census enumeration.

According to these aims, the main population parameters of interest are:

- the *coverage rate*, defined as the ratio between the number of individuals found to Census (net of over-coverage) and the true amount of the population,
- the *under-coverage rate*, defined as the ratio between the number of individuals missed to the Census (net of over-coverage) and the true amount of the population,

where the denominator of both – that is, the true amount of the population – is unknown and, then, it has to be estimated by PES through a *Dual-System Estimator* (DSE), or capture-recapture model³.

The purpose of the survey is the achievement of the estimates of the aforementioned parameters with reference to the following domains: (i) the whole national territory, (ii) the geographical regions, (iii) the age classes: 0-14, 15-29, 30-49, 50-64, 65 and more, (iv) the nationality: italian, foreign.

The PES is a stratified two-stage areal sample selection of Census Enumeration Areas (EAs) that are independently re-enumerated. The first stage selected a sample of municipalities, stratified by geographical region and five classes of municipal demographic size. The second stage selected a sample of EAs, stratified by the tertiles of the EAs population distribution, from within each selected Municipality. All individuals who live in sample EAs are enumerated. The selection of the primary units is with unequal probability without replacement and the selection of secondary units is with equal probability without replacement.

The main types of counting errors that can occur in a Census are:

- 1) the *under-coverage error*, that happens when a person living in Italy on October 9, 2011, is not counted by the Census,
- 2) the *over-coverage error*, that can be identified in four forms:

Type 1 (*duplicate returns at the same location*): this is where two or more returns are made in the same small area (EA) by the same individual. For example, a person may return a paper form, and submit an internet

² For a complete description of the sampling strategy and of the estimates accuracy assessment of PES 2011, see Istat (2015).

³ Sekar, C.C., Deming W.E. (1949).

questionnaire,

Type 2 (*duplicate returns from different locations*): this is where a response is received for the same person but from a different small area, such as students being counted at both their term-time address (correctly) and their parents' address (incorrectly).

Type 3 (*counted in the wrong location*): this is a return where the Census only counts the 'wrong' half of a duplicate. This might happen by only counting the student at their parents' address or by counting a mover after Census reference date only at the new address,

Type 4 (*erroneous returns*): this is a return that is purely fictitious and should not appear in the Census at all at any location. This can be a joke return, a creation of the processing, a baby born just after Census reference date or individuals that died just prior to Census reference date.

Type 1 over-count should be rectified during data processing. Type 4 over-count is only identifiable by further field work. Types 2, 3 will remain in the Census population; therefore, we took in account these two types of error in the estimation procedure.

The aim of the estimation methodology is to produce 'net' adjustments⁴: (i) reducing the Dual-System estimates by the estimated level of over-count, (ii) imputing the fewest number of individuals, (iii) not removing duplicates.

The estimation process is structured as below:

- a) definition of a measure of the *over-coverage propensity*, based on the correct and erroneous Census counts, and study of the over-coverage propensity estimator;
- b) determination of the DSE adjusted by over-coverage, for age-sex groups separately, for each sample Municipality selected in each Hard to Count (HtC) index⁵ stratum, within each geographical region;
- c) computation of a population estimate at geographical region level separately for age-sex groups and HtC index modalities, by using a ratio estimator;
- d) application of the *Sample Balance Adjustment* procedure in all the HtC strata in which samples selected proved to be *outlier*, to make the estimates more reliable;
- e) finally, you can immediately obtain estimates of true (unknown) regional and

⁴ Separate adjustments for under-count and over-count will not be made.

⁵ As the accuracy of population estimates is dependent on response rate and its variation within estimation strata and as the under-count is disproportionately distributed across areas, a hard to count categorization has been produced and the municipalities within each geographical region are post-stratified according to an HtC index. This index attempts to capture the non-geographical variation in under-count in a Census. For defining homogeneous areas, we studied an index based on the predicted individual non-response rate, so as to classify municipalities from 1 to 3 according to their expected relative difficulty of enumeration.

national populations totals using estimates produced up to this point.

3. Estimation model⁶

The methodology adopted for estimating the over-coverage in PES 2011 is constituted of the following three steps: (1) estimate the number of duplicates using PES; (2) estimate people in wrong place using PES; (3) calibration of the estimates referred to point (1), using Census duplicates number, to improve the accuracy of the over-coverage estimates (PES sample is not designed for getting estimates of duplicates number)⁷.

The *over-coverage propensity*, $\gamma_{a_w,g}$, is the propensity for those individuals with characteristics defined⁸ by a_w age class within area g to be over-counted by the Census; in symbols

$$\gamma_{a_w,g} = X_{a_w,g} / Y_{a_w,g} = (Y_{a_w,g} + E_{a_w,g}) / Y_{a_w,g} , \quad (1)$$

in which: $X_{a_w,g}$, $Y_{a_w,g}$, $E_{a_w,g}$ are the total Census count, the *correct* Census count and the erroneous (over-count) Census count, for group a_w in area g , respectively. The plug-in estimator of (1) is given by

$$\hat{\gamma}_{a_w,g} = (\hat{Y}_{a_w,g} + \hat{E}_{a_w,g}) / \hat{Y}_{a_w,g} , \quad (2)$$

$$\text{with: } \hat{Y}_{a_w,g} = \sum_{j \in S_g} w_{jg} \sum_{i \in j} c_{a_w,i,jg} \quad \text{and} \quad \hat{E}_{a_w,g} = \sum_{j \in S_g} \frac{\tilde{P}}{\hat{D}} \sum_b \sum_{\substack{k \in S_b \\ k \neq j}} w_{kb} \sum_{i \in k} o_{a_w,i,kb,jg} ,$$

where $\tilde{P}$ is the Census duplicates number, $\hat{D} = \sum_g \sum_{j \in S_g} w_{jg} \sum_b \sum_{\substack{k \in S_b \\ k \neq j}} w_{kb} \sum_{i \in k} o_{i,kb,jg} c_{i,jg}$ is its estimate by PES.

Furthermore, the symbols showed in the formulas represent: i the individual; b the

⁶ I would like to thank the doctors Owen Abbott and Amy Large of the Office for National Statistics (ONS) for their useful comments, valuable suggestions and reports of inaccuracies that I took into account in writing this note.

⁷ For what just declared, it is clear that a key assumption of the over-coverage estimation procedure is the definition of the correct location of the individual on the Census reference date. We assume that the correct location is given by the individual's response to the question 1.5 of the PES questionnaire.

⁸ We set five age classes a_w for the over-coverage estimation procedure: 0-2 and 25-60, 3-17, 18-24, 61+, and fifteen areas g : five geographical distributions – North-Ovest, North-East, Centre, South, Islands – by HtC.

geographical distribution by HtC (b can be the same as g); j and k the EAs belonging to g and b respectively ($j \neq b$); s_g and s_b the number of sample EAs within g and b respectively; S_g and S_b the number of EAs within g and b respectively; w_{jg} and w_{kb} the PES sampling weights for j selected in g and for k selected in b respectively; $c_{a_w,i,jg}$ an indicator random variable that assumes value 1 if individual i is correctly counted by the Census in j within g and 0 otherwise; $o_{a_w,i,kb,jg}$ an indicator random variable that assumes value 1 if individual i from k within b ($k \neq j$ and b can be the same as g) is incorrectly counted by the Census in j within g and 0 otherwise.

The basis of the under-count estimation is the use of dual-system estimation; the model adopted for PES 2011 is a variant of *Petersen model*⁹, in which the Census total is adjusted for taking over-coverage into account.

Let us denote with a the age-sex groups¹⁰, p the generic modality of HtC, r the geographical region and c the generic Municipality and, for the moment, suppose that PES is in fact a complete enumeration of EAs. The DSE of ${}_r N_{apc}$, the true population total, is

$${}_r \tilde{N}_{apc} = {}_r N_{+1,apc} \bar{\bar{N}}_{1+,apc} / {}_r N_{11,apc} , \quad (3)$$

where ${}_r N_{+1,apc}$, $\bar{\bar{N}}_{1+,apc}$ and ${}_r N_{11,apc}$ is the PES count, the Census count of correct individuals and the matched count respectively. However, we observe ${}_r N_{1+,apc}$, the Census count including over-count; therefore, in order to correct a possible bias in the results and considering the fact that the over-coverage propensity is constant in g , then $E[({}_r N_{1+,apc} / \hat{\gamma}_{a_w,g}) | {}_r N_{1+,apc}] \cong \bar{\bar{N}}_{1+,apc}$, we can adjust the DSE to give

$${}_r \tilde{\tilde{N}}_{apc} = {}_r N_{+1,apc} ({}_r N_{1+,apc} / \hat{\gamma}_{a_w,g}) / {}_r N_{11,apc} , \quad (4)$$

⁹ Wolter, 1986.

¹⁰ For the under-coverage estimation procedure we defined thirty four age-sex groups formed by the age classes: 0-2, 3-7, 8-17, 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85+, separately for female and male.

as ${}^o\tilde{N}_{apc}$ is approximately unbiased for ${}_r\tilde{N}_{apc}$ given ${}_rN_{+1,apc}$, ${}_rN_{11,apc}$ and ${}_rN_{1+,apc}$.

The estimation methodology goes further and applies the *Chapman correction* for DSE's applications to small populations, which allows a better respect of the homogeneity hypothesis of the capture probability of Petersen model. With respect to (3), we have: ${}_r\tilde{N}_{apc}^C = ({}_rN_{+1,apc} + 1)({}_r\bar{\bar{N}}_{1+,apc} + 1)({}_rN_{11,apc} + 1)^{-1} - 1$, whose first-order approximation is given by ${}_r\tilde{N}_{apc}^C \cong {}_r\tilde{N}_{apc} (1 + {}_rN_{+1,apc}^{-1} + {}_r\bar{\bar{N}}_{1+,apc}^{-1}) (1 - {}_rN_{11,apc}^{-1}) - (1 - {}_rN_{11,apc}^{-1})$.

However, with the DSE, the observed total is ${}_rN_{1+,apc}$, so the adjusted DSE with Chapman correction is given by

$$\begin{aligned} {}^o\tilde{N}_{apc}^C &= ({}_rN_{+1,apc} + 1) [({}_rN_{1+,apc} / \hat{\gamma}_{a_w,g}) + 1] ({}_rN_{11,apc} + 1)^{-1} - 1 \cong \\ &\cong {}_r\tilde{N}_{apc} (1 + {}_rN_{+1,apc}^{-1} + \hat{\gamma}_{a_w,g} {}_rN_{1+,apc}^{-1}) (1 - {}_rN_{11,apc}^{-1}) - (1 - {}_rN_{11,apc}^{-1}) \end{aligned} \quad (5)$$

and, therefore, ${}^o\tilde{N}_{apc}^C$ is an approximately unbiased estimator for ${}_r\tilde{N}_{apc}$, given the observed counts ${}_rN_{+1,apc}$, ${}_rN_{11,apc}$ and ${}_rN_{1+,apc}$. As PES is an enumeration of only a sample of EAs and not of all of them, the estimator of ${}_rN_{apc}$ is

$${}^oDSE_{apc} = {}^o\hat{N}_{apc}^C = ({}_r\hat{N}_{+1,apc} + 1) [({}_rN_{1+,apc} / \hat{\gamma}_{a_w,g}) + 1] ({}_r\hat{N}_{11,apc} + 1)^{-1} - 1, \quad (6)$$

where ${}_r\hat{N}_{+1,apc}$ and ${}_r\hat{N}_{11,apc}$ are estimated using PES.

The following stage in the estimation process is to generalize the DSEs to the non-sampled areas, computing the population estimates at geographical region level separately for age-sex groups and HtC index strata, by using a ratio estimator that estimates the relationship in the sample between the Census count and the dual-system estimate for each age-sex group within each HtC stratum

$${}^oDSE_{apc} = {}_r\beta_a {}_rX_{apc} + {}_r\varepsilon_{apc}, \quad (7)$$

in which: ${}_r\beta_a$ is the regression coefficient referred to group of individuals within r having sex and age defined by a , ${}_rX_{apc}$ is the number of persons enumerated by the

Census within the sample Municipality c selected in p within r and with sex and age defined by a and ${}_r\varepsilon_{apc}$ is an error term expressed as the difference between the variable and its conditional average.

Let us indicate with: ${}_r\hat{\beta}_a$ the least squares estimator of ${}_r\beta_a$, ${}_rX_{ap}$ the Census count of persons with sex and age defined by a from all the municipalities within p in r . We can use the relationship (7) to estimate the total region population for each a in each p by multiplying the Census count by the estimated slope of the line: ${}_r\ddot{N}_{ap} = {}_r\hat{\beta}_a {}_rX_{ap}$.

The efficiency of the estimates ${}_r\ddot{N}_{ap}$ can be improved by applying the methodology, showed below and known as *Sample Balance Adjustment* (SBA), in cases where the samples used for obtaining them were not sufficiently representative of the population being measured.

Under normal circumstances the samples, from the PES, were expected to have a good representation of Census non-response. However, for every sampling process there is a risk that a sample may be an *outlier* among all possible samples. In other words, our PES sample could have, by chance, drawn EAs in an area where the Census had managed to count everyone.

The aim of this methodology is to assess if the PES sample was sufficiently representative when compared to all other possible samples that could have been drawn from the underlying population. If it was not, then the coverage adjustment estimates would be skewed (either too large or too small), because the sample was unusual. Census *dummy* questionnaires were used to evaluate whether any of the PES samples were outlying (i.e. unbalanced). These data were believed to be the best possible proxy for coverage, because they represent households from which a return was not received. The procedure can be summarized as follow.

First of all, we analyze the correlation between the *non-response rates* (that is, the ratio between the number of Census *dummy* questionnaires and the number of all Census questionnaires) and their PES estimates. If there is correlation, then the Census *dummy* variable is a good proxy for coverage and it can be used if the sample is found not representative.

The second step consists in comparing the number of Census *dummy* questionnaires referred to r and p , ${}_rY_p$, and its PES ratio estimate, ${}_r\hat{Y}_p$: if the two are significantly different then the sample related to r and p was unbalanced, or an outlier.

However, we went further and a statistical test has been also used. Being known the true distribution of the Census *dummy* questionnaires across all the municipalities within r and p , we determine the variance (and not its estimate) of

${}_r\hat{Y}_p$. Then a 95% confidence interval has been constructed around the true number of

Census	<i>dummy</i>	questionnaires
--------	--------------	----------------

 ${}_rY_p$:
 ${}_rY_p - 2\sqrt{\text{Var}({}_r\hat{Y}_p)} \leq {}_rY_p \leq {}_rY_p + 2\sqrt{\text{Var}({}_r\hat{Y}_p)}.$

If the sample estimate was within the 95% confidence interval then there was no evidence to suggest the sample was extreme. If there was evidence to suggest the sample was extreme, a multiplicative adjustment was made to all the ratio estimators by age and sex in that HtC stratum using the ratio of the population level *dummy* questionnaire implied coverage to the equivalent for the sample

$${}_rF_p = ({}_rY_p + {}_rX_p) {}_rX_p^{-1} \left(\frac{{}_rM_p}{{}_rm_p} \sum_{c=1}^{{}_rm_p} {}_rX_{pc} \right) \left[\frac{{}_rM_p}{{}_rm_p} \sum_{c=1}^{{}_rm_p} ({}_rY_{pc} + {}_rX_{pc}) \right]^{-1}. \quad (8)$$

This adjustment factor was: 1, when the sample was exactly balanced; >1, when the sample was underestimating; <1, when overestimating. ${}_rF_p$ effectively uprated (or deflated) the estimated Census coverage rate.

Given ${}_rF_p$, we can obtain the balanced estimate of ${}_rN_{ap}$, through the relationship: ${}_r\ddot{N}_{ap} = {}_rF_p {}_r\ddot{N}_{ap} = {}_rF_p {}_r\hat{\beta}_a {}_rX_{ap}$.

At this point, we can easily determine all the other estimates that are the primary aim of PES 2011. By summing with respect to HtC, you have the estimates for each sex-age group within region r : ${}_r\ddot{N}_a = \sum_p {}_r\ddot{N}_{ap}$; by summing with respect

to age-sex groups, you have the estimates for each geographical region: ${}_r\ddot{N} = \sum_a {}_r\ddot{N}_a = \sum_a \sum_p {}_r\ddot{N}_{ap}$; the national estimate is, then, provided by the relationship: $\ddot{N} = \sum_r {}_r\ddot{N}$.

References

- ISTAT. 2015. L'indagine di controllo della copertura censuaria in *Atti del 15° Censimento della popolazione*, Volume No. 6.
- SEKAR C. C., W.E. DEMING. 1949. On a Method of Estimating Birth and Death Rates and the Extent of Registration, *Journal of the American Statistical Association*, Vol. 44, No. 245, pp. 101-115.
- WOLTER K.M. 1986. Some Coverage Error Models for Census Data, *Journal of the American Statistical Association*, Vol. 81, No. 394, pp. 338-346.

SUMMARY

The Post Enumeration Survey of the 15th Italian Population Census: Features and Methods

The Post Enumeration Survey (PES) has the main goals to estimate the number of individuals actually and habitually dwelling in the reference time of the 15th Population Census (October 9, 2011), the coverage rate and the over-coverage rate. In order to achieve this objective, a complex estimation model, derived by ONS approach, has been implemented. The paper presents the main features of the survey and the steps of the estimation process.