Rivista Italiana di Economia Demografia e Statistica Volume LXVIII n. 3/4 Luglio-Dicembre 2014

A MODEL BASED CATEGORISATION OF THE ITALIAN MUNICIPALITIES BASED ON NON-RESPONSE PROPENSITY IN THE 2011 CENSUS

Antonella Bernardini, Andrea Fasulo, Marco D. Terribili

1. Introduction

The counting operations carried out during a population census can be afflicted by non-sampling errors.

The quality takes on the meaning of precision that is expressed as an inverse function of the statistical error. The aim of the Istat is to provide accurate estimates of the main non-sampling errors, particularly in complex investigations like the Census. The non-sampling error is a function of many factors: organizational aspects of the survey, the behaviour of a plurality of individuals or Institutions.

The Italian National Institute of Statistics (Istat) certifies the quality of the 15th Population and housing census through a sample survey of coverage assessment, as required by Commission Regulation (EU) No 1151/2010 of 8th December 2010 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council. The Post Enumeration Survey (PES) has the goal of estimating the real number of the people living in Italy on 9 October 2011, at the reference day of the 15^{th} population and housing general census; it has also the aim of evaluating the errors of overcoverage and undercoverage in the individuals count.

The main indicators to evaluate the accuracy, is the coverage rate, which is calculated (under the assumption of to not undercover the population) as the ratio between the number of the enumerated units during the census and the real population dimension, denoted by N and obviously unknown.

The survey design of the PES is a two stages with stratification of the primary sample units (252 municipalities) and of the secondary units (about 2500 enumeration areas). The collection of data has been planned to guarantee the independence between the two surveys. The interest of the survey is focused on the families and on the individuals habitually living in the enumeration areas selected for the sample of the PES.

In order to estimate the coverage rate we have estimated a statistic model based on the Petersen's model assumption; this model is part of a models class, called dual-system (or capture-recapture methods) and it represents one of the most common model between those used to quantify the Census coverage errors (Wolter, 1986). One of the basic hypothesis of the estimation model used is the constant capture probabilities at the census and at the PES, for all the units belonging to the soubpopulation.

We need to fit the estimation model to small domains in which the capture probability is the same and then to calculate the estimate in wider domains, given by aggregation of sub-domains. In estimation phase, thanks to a greater number of auxiliary available variables, regarding design and sampling phase, a post-stratification has been carried out.

One of the used post-stratification variables is the Hard To Count index (HTC), which contributes to detect homogeneous areas relatively to the difficulty of a subpopulation to be enumerated. The model study, on which the index has been designed, leads to analyse social, economic and demographic characteristics, significantly influential on the individual probability to be censused. These characteristics point out some differences, relatively to local non-response levels.

Following the important ONS experience about the HTC applied during the population census of 2001 and 2011, an index has been studied to categorize Italian municipalities regarding an homogeneous expected level of right enumeration of the individuals.

2. Predictive models for right enumeration

To study the propensity of the individuals to be correctly numbered during the Population Census, data coherence with the Post Enumeration Survey (PES) has been taken into account. With the aim of output the individual estimated probability of right enumeration, a predictive model has been fitted; this model assumes a link function between several auxiliary variables, collected during the PES or available from other sources, and the dependent variable. The latter is a binary variable that points out the missing record linkage between the individuals listed during the Post Enumeration Survey and those ones listed during the Population Census. So the variable modalities are:

$$Y = \begin{cases} 1 & \text{unsuccessfull record linkage} \\ 0 & \text{successful record linkage} \end{cases}$$

Being the dependent variable a binary one, the implemented models are fixed effects logistic ones, they can be expressed in the following way:

$$\text{Logit } P(Y_i = 1 | X_i) = \text{Log } \frac{P(Y_i = 1 | X_i)}{1 - P(Y_i = 1 | X_i)} = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$$

As an alternative to fixed effects models, Random-Effect Logit Models are implemented too, to take into account the enumeration areas (territorial division in which Municipalities are divided) with the intercept γ_d :

$$\text{Logit } P(Y_{id} = 1 | X_{id}) = \text{Log } \frac{P(Y_{id} = 1 | X_{id})}{1 - P(Y_{id} = 1 | X_{id})} = \alpha + \beta_1 X_{1id} + \beta_2 X_{2id} + ... + \beta_k X_{kid} + \gamma_d$$

Auxiliary variables, available for the statistical units reached by the Post Enumeration Survey, describe socio-demographic characteristics of the individual and of the municipalities/provinces of which they belong to. Post-stratification allows to exploit the data richness of the Post Enumeration Survey, its updated individual information, and to integrate it with other local variable, available from archive.

Table 1 − *Auxiliary variables, regarding informative level*

Level	Auxiliary variable
Individual	Age
	Age classes
	Sex
	One unit family
	Extended family (more than 7 individuals)
	Foreigners
	Singles (Separated, divorce, widow)
	Proxy student (19≤age≤30, educational qualification at least diploma
Municipal	University city
	Coastal city
	Altimetric zones (in 5 modalities)
	Population density (pop. Per km^2)
	Foreigners rate
Provincial	Unemployment rate
Interactions	Foreigners * Foreigners rate
	One unit family * Age class 10÷29
	University city * Proxy student

In the model study phase three alternative models have been proposed: the first one fits only individuals variables, the second model fits area variables in addition to the individual ones and the third fits also interactions between some variables paired. In the following table 1, the complete list of auxiliary variables, distinct for degree of detail and other.

3. Hard To Count model

The multi-level modeling involves the prediction of the variance at different levels, so often it start with an analysis to determine what levels this variation can be considered significant. In the first step two random intercepts were tested, one at the municipal level and one at enumeration area level, because it is useful to assess how much of the total variance is explained between the different groups. This can be accomplished by calculating the Intraclass Correlation Coefficient (ICC) using the formula:

$$ICC = \tau_{00}/(\tau_{00} + \sigma^2) \quad (1)$$

where is the between-group or Intercept variance, and the within-group or residual variance. The estimated ICC, at the municipal level, is .009, while at the enumeration area level is .032, a value that makes us lean towards that level of detail. In the second and last step, the significance of the Intercept variance was evaluated through a likelihood ratio test. In order to do this we compare the values of -2 log likelihood of the null model with random intercept with the likelihood of the null model without random intercept. The value of - 2 log likelihood for the model without the random intercept is -579.870. The same indicator for the model with the random intercept is -584.294. The difference of 4.423 is significant for a chi-square distribution with one degree of freedom. These results suggest that a random intercept of enumeration area produces a significant improvement of the model. It has been estimated that 3.2% of the total variance in the study of non-response probability, is a function of the enumeration area of the person.

Even the study of the model was performed in different phases. The model selected was made through the use of commonly used criteria for the choice of models that are the log-likelihood, the AIC and BIC indicators. In the first phase, the variables of the questionnaire, available for each person, have been used. The best was the model with the variables age classes, sex and citizenship, with AIC, BIC and log-likelihood respectively equal to 29.381, and 29.466 -14.682. Afterwards, area level covariates were added and the best model was the one with the variable rate of unemployment, university common flag, population density and rate of foreigners. Adding area level covariates, led an improvement of all 3 indicators, which amounted to 29.196 AIC, 29.324 BIC and -14.586 log-likelihood.

Finally, were considered the combined effects of different variables, but the only significant interaction, which improved the model was between citizenship and the rate of foreign residents in the municipality. Also adding this effect the AIC, BIC, and the log-likelihood are equal to 29.174, 29.313 and -14.574.

Table 2 shows the regression coefficients for the three models described above.

Table 2 − *Regression coefficients of the models.*
In grey the coefficients not significant

Auxiliary variables	Individual variables model	Individual + area level variables model	Complete model
Intercept	-5,711	-6,905	-7,067
Age class 10-29	0,075	0,074	0,072
Age class 30-49	0,048	0,046	0,041
Age class 50-74	-0,555	-0,555	-0,564
Age class ≥75	-0,481	-0,480	-0,488
Sex (female)	-0,164	-0,166	-0,168
Foreigners	2,395	2,395	2,848
Unemployment rate		10,411	10,489
University city		0,826	0,826
Population density		9,505e-05	9,178e-05
Foreigners rate		4,594	6,817
Foreigner * Foreigners rate			-5,795

Once calculated the probability of being been counted or not at the census, these were averaged at the municipal level, so as to return to the spatial detail of interest.

The orderly distribution of the predicted values, relative to the 252 municipalities of the sample, was divided on the basis of percentiles in 3 modes following the distribution 40% - 40% - 20%. Thus the virtuous municipalities, with a low problem with counting the person, will be categorized with the HTC level 1, the municipalities in an intermediate situation, will have the HTC level 2, and the most problematic municipalities from the point of view of the correct enumeration will have the HTC level 3. This categorization has also been applied to probability

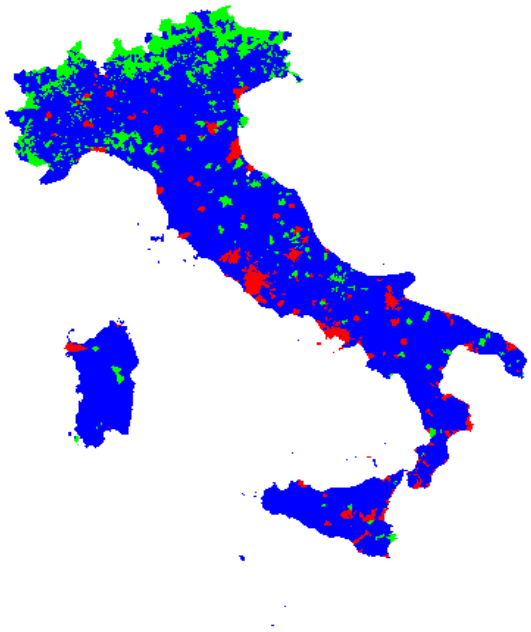
of the municipalities outside the sample, predicted by using only the synthetic part of the multilevel logistic regression model described above.

4. Results

The available wealth of information has allowed a detailed study on the hardest individuals to count in the census.

Figure 1 shows the distribution of HTC among Italian municipalities.

Figure 1 − *HTC distribution in the Italian Municipalities.*
HTC level 1 in green, HTC level 2 in blue, HTC level 3 in red.



The most virtuous municipalities, colored in green, are those that are distributed along the Alps and Apennines, show small municipalities. Municipalities with an intermediate index, colored in blue, are the majority and they cover almost the entire territory. Finally, the most problematic areas are colored red and representing large municipalities, focusing long the Italian coast highlighting the issues related to the second home or holiday house and movements for seasonal work.

References

Abbott O. 2000. 2001 Hard to Count Index, *One number census steering committee*. http://www.ons.gov.uk/ons/guide-method/census/census-2001/design-and-conduct/the-one-number-census/methodology/steering-committee/key-papers/hard-to-count-index.pdf

Grossi P., Mazziotta M. 2012. *Qualità del 15° Censimento generale della popolazione e delle abitazioni attraverso una indagine di controllo che misuri il livello di copertura*. Istat Working Papers n. 16/2012

Office for National Statistics. 2011. Office for National Statistics, London

Predicting patterns of household non response in the 2011 Census. http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/statistical-methodology/predicting-patterns-of-household-non-response-in-the-2011-census.pdf

SUMMARY

A model based categorisation of the Italian municipalities based on non-response propensity in the 2011 Census

The Italian National Statistical Institute had certified the quality of the 15th Italian population and housing census thanks to a Post Enumeration Survey (PES) taken throughout the months immediately after the Census. The aim of the PES is to produce total estimates adjusted for under coverage and, for the first time, over coverage.

The model underlying the under and over coverage estimation, takes into account the differences between individual probabilities of responding to the Census. For this aim a regression unit-level model was applied; in order to study the individual probability to be censused on the basis of which the Hard to Count Index (HTC) of Italian municipalitiesit was created. In the model were used variables derived from the PES questionnaire and additional area-level variables from other sources.

HTC categorises the 8092 Italian municipalities in 3 different levels, partitioning the distribution of municipal non-response propensities, based on percentiles.

This paper describes in detail the multilevel logistic regression model used to study non-response probability, the development of the HTC, the methods and the analysis carried out to evaluate the goodness of index, regarding the census coverage.

Antonella BERNARDINI, Italian National Institute of Statistics, anbernar@istat.it
Andrea FASULO, Italian National Institute of Statistics, fasulo@istat.it
Marco D. TERRIBILI, Italian National Institute of Statistics, terribili@istat.it