GIORNATE DELLA RICERCA IN ISTAT | 10-11 NOVEMBRE 2014

SESSIONE I
CAMPIONAMENTO E STIMA

L'utilizzo dell'indice *Hard To Count* nell'indagine di copertura del 15° Censimento della popolazione

Antonella Bernardini, Andrea Fasulo e Marco Dionisio Terribili

Istat

The use of the HTC index during the Post Enumeration Survey

Antonella Bernardini, Andrea Fasulo, Marco D. Terribili

Istat

anbernar@istat.it , fasulo@istat.it , terribili@istat.it

Sommario

L'indagine di copertura del censimento 2011(PES) ha il duplice obiettivo di stimare il numero corretto di individui residenti in Italia alla data di riferimento del censimento e di valutare gli errori di sotto e sovra copertura dell'indagine censuaria. Per stimare tali quantità è stato applicato un modello statistico basato essenzialmente sulle ipotesi adottate nel modello di Petersen; una delle ipotesi alla base del modello di stima impone che le probabilità di cattura del Censimento e dell'indagine PES siano costanti per ogni unità all'interno della sottopopolazione.

L'Hard To Count index (HTC) è una delle variabili utilizzata nella fase di post stratificazione, che ha contribuito ad individuare delle aree omogenee rispetto alla difficoltà di una popolazione ad essere conteggiata.

Nel paper viene descritta la costruzione dell'HTC; attraverso l'applicazione di un modello di regressione logistica multilivello si è ottenuta una categorizzazione dei comuni Italiani basata sulla propensione alla mancata risposta al Censimento del 2011.

Infine viene proposta una partizione alternativa del territorio nazionale Italiano e comparata con la distribuzione originaria del HTC.

Parole chiave: Mancata risposta, Categorizzazione, Comuni Italiani

Abstract

The Post Enumeration Survey (PES) 2011 has the goal of estimating the real number of the people living in Italy on 9th October 2011 and also the aim of evaluating the errors of overcoverage and undercoverage in the individuals count. In order to estimate the coverage rate a statistic model, based on the Petersen's model, assumption have been estimated; one of the basic hypothesis of this estimation model is the constant capture probabilities, for all the units belonging to the subpopulation.

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 paper describes the development of the HTC; by means of a multilevel logistic regression model a categorization of the Italian municipalities has been obtained, based on non-response propensity in the 2011 census,.

Finally an alternative partition of the Italian territory is presented and compared with the original HTC distribution.

Key words: Non-response propensity, categorization, Italian municipalities.

Introduction

The counting operations carried out during a population census can be affected by non-sampling errors.
The quality of such operation is represented by the precision which, in turn, is expressed as an inverse function of the statistical error. The Italian National Institute of Statistics (Istat) aims 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, like organizational aspects of the survey and the behavior of a plurality of individuals or Institutions.
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 the 8th of December 2010 implementing Regulation (EC) No. 763/2008 of the European Parliament and of the Council (Eurostat 2003). The Post Enumeration Survey (PES) has the goal of estimating the real number of the people living in Italy on the 9th of October 2011, which is the reference day of the 15th population and housing general census; it also has 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; under the assumption of not undercovering the population, it is calculated 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 PES (Grossi, Mazziotta 2012), is designed as a two-stages survey 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 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 prepared a statistical 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 among those used to quantify the Census coverage errors (Wolter 1986). One of the basic hypotheses of the estimation model used, is the constant capture probabilities at the census and at the PES, for all the units belonging to the sub-population.
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 subdomains. During the estimation phase, thanks to a greater number of auxiliary available variables, compared to the design and sampling phase, a post-stratification has been carried out.
One of the post-stratification variables that we used, 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 studied, on which the index has been designed, leads to the analysis of social, economical and demographical 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 (Abbott 2000, Office for National Statistics 2011), experience about the HTC applied during the population census of 2001 and 2011, an index has been studied to categorize Italian municipalities regarding a homogeneous, expected level of right enumeration of the individuals.

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 has been taken into account. With the aim of outputting 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 listed during the Population Census. So the variable modalities are:

$$Y = \begin{cases} 1 & \text{successfull record linkage} \\ 0 & \text{unsuccessful record linkage} \end{cases}$$

Being the dependent variable a binary one, the implemented models are fixed effects logistic ones (Agresti 2007), they can be expressed in the following way:

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

As an alternative to the fixed effects models, Random-Effect Logit Models are implemented too (Agresti 2007), to take into account the enumeration areas (territorial division in which Municipalities are divided) with the intercept γd :

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

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. 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 the interactions between some of the variables paired. In the following table 1, we have indicated the complete list of auxiliary variables, categorized by degree of detail.

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)
Provincial	Foreigners rate
Interactions	Unemployment rate
	Foreigners * Foreigners rate
	One unit family * Age class 10÷29
	University city * Proxy student

Hard To Count model

The multi-level modeling involves the prediction of the variance at different levels, so it often starts with an analysis to determine what are the levels this variation that 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 τ_{00} is the between-group or Intercept variance, and σ^2 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. The latter estimates that 3.2% of the total variance in the study of non-response probability, is a function of the enumeration area of the person.
The study of the model was also performed in different phases. The selected model was generated 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, have been used. The best model was the one using 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 city”, “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, the combined effects of different variables were considered, 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. For the complete model also the area under the Receiver Operating Characteristic (ROC) curve has been computed to have a measure of predictive power. The area under the ROC curve is 0.9, suggesting an excellent discriminatory power. 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

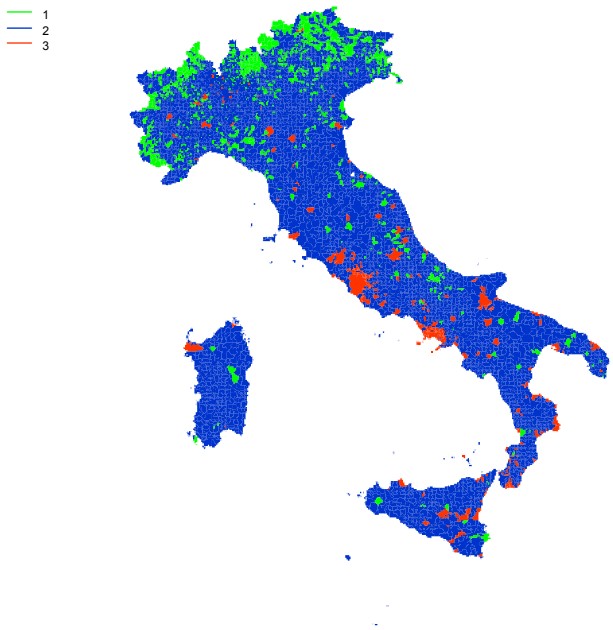
Once calculated, the probability of being been counted or not at the census were averaged at the municipal level, so as to return to the spatial detail of interest.
The ordered distribution of the predicted values, relative to the 252 municipalities of the sample, was divided on the basis of percentiles in 3 modes following the distribution 40%, 40% and 20%. Thus, the virtuous municipalities, where counting people accurately is relatively easy, 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 the probability of the municipalities outside the sample, predicted by using only the synthetic part of the multilevel logistic regression model described above.

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 green, are those that are distributed along the Alps and, in a lesser way, along the Apennines. These municipalities have a good census coverage being very small municipalities.
Municipalities with an intermediate index, colored blue, are the majority and they cover almost the entire territory. One of the future developments will consider a better division of the HTC index to create areas as homogeneous as possible also in term of numerousness.
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. Among these municipalities, we identify several locations of important universities in Italy, highlighting also the issue of the correct enumeration of non-resident students.

Further developments

With the aim of reducing the variability of dependent linkage variable in the three HTC levels, several alternative classifications have been studied.
Below the result of an experimentation are shown. The ordered coverage probability distribution has been divided in five levels, delimited by the 20th, 40th, 60th and 90th percentiles.
This partition, studied just for exploratory purpose analysis, was not used during PES post-stratification phase, cause the HTC was crossed with other post-strata variables and it made really important the numerousness of the statistical units in these cross-cells.
However the experimentation of a further classification should be interesting: cutting the

ordered coverage probability distribution in a different way, we can study the variability decrease in the levels, without looking after to the number of the units in the groups.

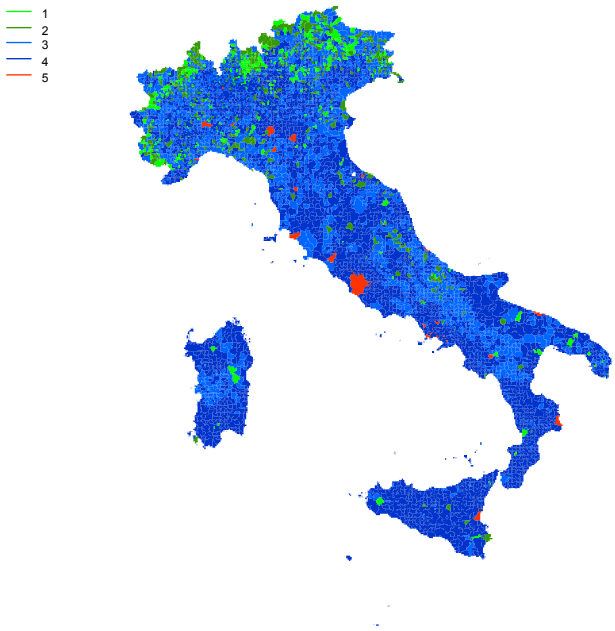
Table 3 – Coefficient of variation (%) for the different levels in the two HTC index.

Index		HTC 1		HTC 2	HTC 3
Original HTC		37		32	43
	HTC 1	HTC 2	HTC 3	HTC 4	HTC 5
New HTC	28	14	19	28	30

The table 3 shows the dependent variable coefficients of variation calculated within the levels, obtained with the two alternative classifications compared. Splitting the modalities of the original HTC, a large decrease of the CV values can be observed.

Figure 2 − HTC distribution in the Italian Municipalities.

HTC level 1 in fluo green, HTC level 2 in dark green, HTC level 3 in light blue, HTC level 4 in dark blue, HTC level 5 in red.



From the figure 2 we can observe that the most virtuous municipalities, colored fluo green, are those that are strewn along the Alps. The level 2 municipalities, colored dark green, are again distributed on the North, but along the Apennines too. The level 3, consisting of the municipalities in the most intermediate situations, colors (light blue) most of the Po valley in a scattered way, and South Apennines and about half Sardinia in a concentrated way. Several municipalities of Lazio, Tuscany, Sicily, Campania, Calabria and Puglia (especially coastal towns) belong to the 4th level, colored dark blue. In red we can observe the 31 municipalities consisting of the 5th index level, the most problematic cities, among which

the biggest city centers (such as Rome, Milan, Florence, and Bari).
Relating to original classification, dividing in two levels the green level we have a more detailed view of the most virtuous municipalities; the same procedure on the colored blue municipalities shows two different situations for the intermediate situation, solving the issues related to the large number of municipalities in the second level of the original HTC index. Finally has been created a last level of the new HTC taken into account only municipalities with a undercoverage error bigger than the 90° percentiles of the ordered coverage probability distribution. In this way a more homogeneous level has been carried out.

Bibliography

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

Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, Wiley

Eurostat (2003), *Assessment of Quality in Statistics: Glossary,* Working Group, Luxembourg

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

Office for National Statistics (2011), 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

Wolter, K. M. (1986), *Some Coverage Error Models for Census Data*, Journal of the American Statistical Association, Vol. 81, No. 394 (Jun., 1986), pp. 338-346