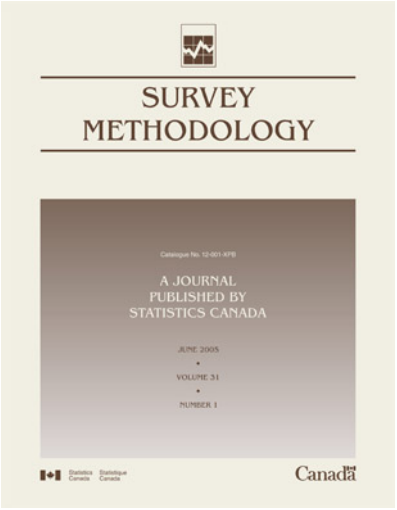
Catalogue no. 12-001-XIE

# Survey Methodology

December 2003




Statistics Canada Statistique Canada

Canada

## How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

| | |
|---|---|
| National inquiries line | 1 800 263-1136 |
| National telecommunications device for the hearing impaired | 1 800 363-7629 |
| Depository Services Program inquiries | 1 800 700-1033 |
| Fax line for Depository Services Program | 1 800 889-9734 |
| E-mail inquiries | infostats@statcan.ca |
| Website | www.statcan.ca |

## Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

# Survey Methodology

December 2003

Published by authority of the Minister responsible for Statistics Canada



February 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

---

**Note of appreciation**

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 29, No. 2, pp. 167-175
Statistics Canada, Catalogue No. 12-001

# A Hierarchical Model for the Anaylsis of Local Census Undercount in Italy

**D. Cocchi, E. Fabrizi and C. Trivisano** [1]

Abstract

Census counts are known to be inexact based on comparisons of Census and Post Enumberation Survey (PES) figures. In Italy, the role of municipal administrations is crucial for both Census and PES field operations. In this paper we analyze the impact of municipality on Italian Census undercount rates by modeling data from the PES as well as from other sources using Poisson regression trees and hierarchical Poisson models. The Poisson regression trees cluster municipalities into homogeneous groups. The hierarchical Poisson models can be considered as tools for Small Area estimation.

Key Words: Census undercount; Post enumeration survey; Bayesian hierarchical modeling; Gamma-Poisson regression models; Poisson regression trees.

## 1. Introduction

The Italian Population Census takes place every ten years and represents the most important institutional duty of the Italian National Institute of Statistics (ISTAT) (The work leading to this paper has been developed just before 2001 Italian Census and the subsequent PES. The results have been considered in performing the 2001 PES). In order to carry out the Census, ISTAT relies on municipal administrations who are responsible for all field operations (training of interviewers, planning of interviews, data gathering and basic data processing). During Census operations, each municipality works independently from the others under ISTAT supervision. The accuracy of the Census results therefore differ considerably from one municipality to another, even if contiguous. In Italy, the geographical area of a municipal borough is sub-divided into Census Enumeration Areas (EAs), which are assigned to a single interviewer during Census operation. The EAs differ in terms of shape, structure and difficulty of enumeration, as well as interviewer. It is likely that the undercount rate varies substantially among EAs within the same municipality.

After the 1991 Population Census, ISTAT conducted a Post Enumeration Survey (PES) to measure the phenomenon of undercount. Population Census counts are known to be generally incorrect because of missed, multiple and misplaced enumeration. Missed enumeration is the most important inaccuracy and typically yields a net population undercount that may vary geographically and between different social groups, and impacts the determination of the relatives sizes of sub-populations (Abbate, Masselli, Signore 1993). Field operations of the PES were carried out by the sampled municipalities themselves. The 1991 Italian PES data have been analyzed by Abbate, Masselli and Signore (1993), who estimate the overall national undercount rate by means of a Lincoln-Petersen model (see Wolter 1986) using post-strata of municipalities based on large geographical areas (North, Center, South). Working on the same data, Fortini (1994) estimates the overall national undercount by means of latent class models.

Instead of estimating the undercount rate for the whole country or smaller domains, we propose models designed to explain the variation in undercount rate at the municipal level. The availability of factors accounting for the size of the net undercount may be a basis for creating homogeneous groups of municipalities, for planning a more efficient stratification in future Post Enumeration Surveys. Moreover, knowledge of those flaws in municipal organization which significantly influence the undercount may provide guidelines for actions designed to reduce its size.

Contributions which use disaggregated PES data are present in the literature. Alho, Mulry, Wurdeman and Kim (1993) consider a logistic regression model for the individual (household) probability of being censused. In keeping with Moura and Holt (1999), their model could be extended to include municipality or other group effects. We are in fact aware that our choice of modelling municipal data is not the same as the analysis of household level records, since many features determining individual propensity to be caught by the Census average out when dealing with aggregated data. A comprehensive analysis based on individual records is not feasible in the Italian case, since there were very few questions for individuals included in the 1991 PES schedule. Similarly, the 1991 PES provides very little auxiliary information on the EAs, with the consequence that models based on EA undercounts cannot be proposed.

Our analysis is based on combining different data sources. The auxiliary information comes from the abovementioned 1991 PES, two studies on the statistical quality of municipalities conducted by ISTAT during the early 90s (Di Pietro 1998, 1999) and demographic and social indicators obtained from the 1991 official Census results.

1. D. Cocchi, E. Fabrizi and C. Trivisano, Dipartimento di Scienze Statistiche “P. Fortunati”, Università di Bologna, Italy.

We face the problem of how to make efficient use of the information obtained from the various data sources. We have in fact a large number of variables, most of which are categorical or polychotomous. Instead of using a variable selection algorithm, we have chosen to build homogeneous groups of municipalities which are then introduced into the model by means of a design matrix for the random effects. These groups are constructed using Poisson regression trees (Therneau and Atkinson 1997). This hierarchical usage of information provides a natural basis for the design of strata of geographically non-contiguous municipalities.

Few EAs are re-censused within each sampled municipality in the PES; the average EAs sampling rate is 0.001. This is a typical Small Area setting where direct estimates of the municipal undercount rate are unreliable and ought to be replaced by synthetic or composite estimates based on a suitable model. The phenomenon of undercount is rare. Our data consist of counts and may show a large overdispersion with respect to a Poisson distributional assumption. We suggest the use of hierarchical Poisson regression models to manage overdispersion.

The hierarchical models here adopted manage explicitly overdispersion due to municipal heterogeneity. A further extra Poisson variability source is due to heterogeneity within municipalities, because of clustering of missed enumeration within EAs, or of clustering due to missed enumerations of individuals in the same family. This kind of overdispersion is not explicitly treated in the models.

We adopt a full Bayesian approach for specification and estimation purposes and base the solution of the models on Markov chain Monte Carlo simulation methods. Within this hierarchical framework, we deal with overdispersion by imposing a Gamma distribution on the rate of the first level Poisson distribution, thus marginally obtaining a Negative Binomial. Moreover, conditionally on the hyperparameters, the proposed model features posterior linearity and the corresponding posterior means for the municipal undercount rates are linear composite estimators. Thus, the amount of smoothing depends on how much information is provided by each municipal sample in the PES.

Our results show that the municipality stratification employed in designing the 1991 PES (based on geographical area and population size) can be improved, since the undercount rate is shown to be largely independent of geographical area. On the contrary, variables describing the statistical efficiency of local administrations are useful in discriminating between the different degrees of undercount among municipalities of similar size and demographic structure. Whilst leaving the design of the PES unchanged, our results may provide useful guidance when performing data analysis.

The present paper is organized as follows. Section 2 describes the basic features of the PES and of the other data sources we have taken into consideration. Section 3 looks at the Poisson regression trees used to build homogeneous groups of municipalities. In section 4 we introduce the hierarchical Poisson regression models, while empirical results and model comparisons are discussed in section 5.

## 2. The PES Data and Auxiliary Information

### 2.1 The Italian Post Enumeration Survey

The 1991 Italian Population Census took place on October 20$^{th}$. The subsequent Post Enumeration Survey, based on a two stage stratified sampling design, was carried out a few weeks later. Municipalities constitute the primary units, whereas the secondary ones are represented by the Census EAs. An EA is the smallest area into which the municipal territory is partitioned for Census operations; each EA is assigned to a single interviewer.

The primary sampling units were stratified according to geographical area (North-West, North-East, Center, South, Islands) and demographic size (7 classes for the municipalities below 350,000 inhabitants), producing 35 strata. Within each stratum the sampled municipalities were selected without replacement and with probability proportional to their demographic size. The 10 municipalities with more than 350,000 inhabitants have been included in the sample as self-representative units. The secondary sampling units were selected with equal probabilities by systematic sampling. The final PES sample contains 85 municipalities and 638 EAs (out of a national total of 8,095 municipalities and 64,000 EAs) with a national design based estimate of 1.24% (Abbate, Masselli and Signore 1993).

The PES forms were filled out during face to face interviews and contained just a few simple questions. The characteristics of the sampled households are limited to the number and gender of household members. Other PES questions were designed to facilitate record linkage with the Census result, and therefore to reduce both misplaced enumeration and other non sampling errors in the evaluation of undercount (see Fortini 1994 for details).

### 2.2 The Surveys of the Statistical Quality of Municipalities

A data set on the statistical quality of Italian municipalities was constructed by ISTAT (see Di Pietro 1998, 1999). It integrates different sources: information from 1991 Census performance records, municipal population registers and Interior Ministry data. This data set contains also the results of three administrative surveys, conducted during the 90s, carried out in order to evaluate the performance of municipalities with regard to their commitments to ISTAT. The first survey is about the computerization of municipal Statistics Bureaus. The second survey, known with the acronym POSAS, is a post-Census survey of the demographic registers of the resident population, classified by year of birth, age and civil status. The third survey, known with the acronym ISCAN, regards the

appropriateness of registrations on the municipal population registers list. These surveys provide data for all Italian municipalities.

From this data set we selected a subset of variables related to the municipal activity at the time of the 1991 Census:

a) the percentage of noncoded fields of the Census household forms which had to be filled out, after the interview of the households, by the municipal Statistics Bureaus (PERCOD);

b) the ratio of the population temporarily abroad to the population present at the 1991 Census (PERCEST);

c) the ratio of the difference between the 1991 Census and population registers counts to the 1991 Census counts (PERDIFF);

d) the time needed to update municipal demographic registers on the basis of 1991 Census results (IND01);

e) delay in street name updating (IND11).

### 2.3 Demographic Variables

We also consider a set of demographic ratios from the 1991 Census results. In particular, we use the percentages of "single member" and "more than one family" households, and sex ratios (males/females) in the municipality. The municipal resident population – resulting from the uncorrected 1991 Census counts – is also a very important variable. The number of EAs sampled in each municipality for the PES is a further signal of the municipality importance.

## 3. Poisson Regression Trees

The available data sources provide us with a large number of auxiliary variables, many of which are categorical or polychotomous. Before we fit the hierarchical models, we group municipalities with homogeneous household undercount rates using Poisson binary regression trees. Groups based on trees are included as factors in the models described in the next section. Our principal aim is to check the effectiveness of traditional stratifications, improving them ex post by hierarchical models with suitable covariates and to verify how they differ from comparable results based on optimal groupings.

The conditional regression models are based on the canonical logarithmic link. The splitting criterion is based on the usual deviance statistic (Therneau and Atkinson 1997):

$$\text{Deviance}_{\text{parent}} - (\text{Deviance}_{\text{child, left}} + \text{Deviance}_{\text{child, right}})$$

The basic idea for building a tree is to begin with a large tree $T_0$ constructed using a naive and mild stopping rule (as the minimum number of observations in the final nodes of the tree) and then to select the right-sized tree among the sub-trees of $T_0$ by pruning. The established methodology for pruning trees is cost-complexity pruning, first introduced by Breiman, Friedman, Olshen and Stone (1984). Let $D_T$ be the deviance of a subtree $T$ of $T_0$, $size(T)$ the number of terminal nodes of $T$ and $\alpha > 0$ a cost-complexity parameter for defining the cost-complexity measure:

$$D_T(\alpha) = D_T + \alpha \text{size}(T) \qquad (1)$$

For a specified $\alpha$ the tree $T(\alpha)$ that minimizes (1) can be found. It can be shown (Breiman *et al.* 1984) that a nested family of subtrees $\{T_0, T_1, ..., T_k, ..., T_{\text{root}}\}$ of $T_0$ exists such that each tree is optimal for a range of values of $\alpha$.

The problem is now reduced to selecting one of these subtrees. The selection is carried out in order to minimize the prediction error defined as the deviance contribution for a new observation. To estimate the prediction error, the availability of an independent sample would be in principle the best option, but since it is advisable to use all data to "instruct" the tree in the best possible way, a cross-validation method is used. Usually, the tree $T_{k_0}$ with the minimum estimated prediction error is selected. Here we use a more severe pruning rule which consists in selecting the smallest tree with an estimated prediction error not larger than the estimated prediction error of $T_{k_0}$ plus its standard error. This pruning rule, known as the "1 SE rule" (Breiman *et al*. 1984), is adopted in order to avoid model overfitting.

Since the cross-validation of Poisson regression trees may give, in some nodes, infinite values for the deviance statistic, we use Bayesian shrinkage estimators of the true rates, based on a simple Poisson-Gamma model, as suggested in Thernau and Atkinson (1997).

We built three different trees based on different starting subsets of auxiliary variables.

Tree 1(shown in Figure 1) is based on demographic variables only. The first split separates municipalities with population less than 100,100 from those with more than 100,100. This splitting value is almost coincident with the 100,000 demarcation value used in the stratification of municipalities for the 1991 PES. The second split isolates a sub-sample of small municipalities for which less than 4 EAs were sampled in the PES. A further split is made on the basis of the sex ratio.

Tree 2 (Figure 2) is based exclusively on variables concerning the quality of the statistical performance of municipalities. The first split is based on the timing in correcting demographic registers (IND01): those municipalities that were quickest in performing this activity have the lowest undercount rates. Lower level splits highlight the problem of people temporarily abroad (PERCEST) which in areas characterized by massive emigration may lead to serious undercounting of the municipal population and errors in the book-keeping of demographic registers (PERDIFF). In this tree, one half of the sample is classified in a single node which is likely to contain residual heterogeneity.

Tree 3 (Figure 3) is based on both demographic and quality variables. The first split is based on the municipal population exactly as was the case in Tree 1. Subsequently, the subset of municipalities with less than 100,100 inhabitants is split into small and middle sized municipalities at a threshold of 13,200. The quality variable included in this tree consists of timing in correcting demographic registers (IND01).

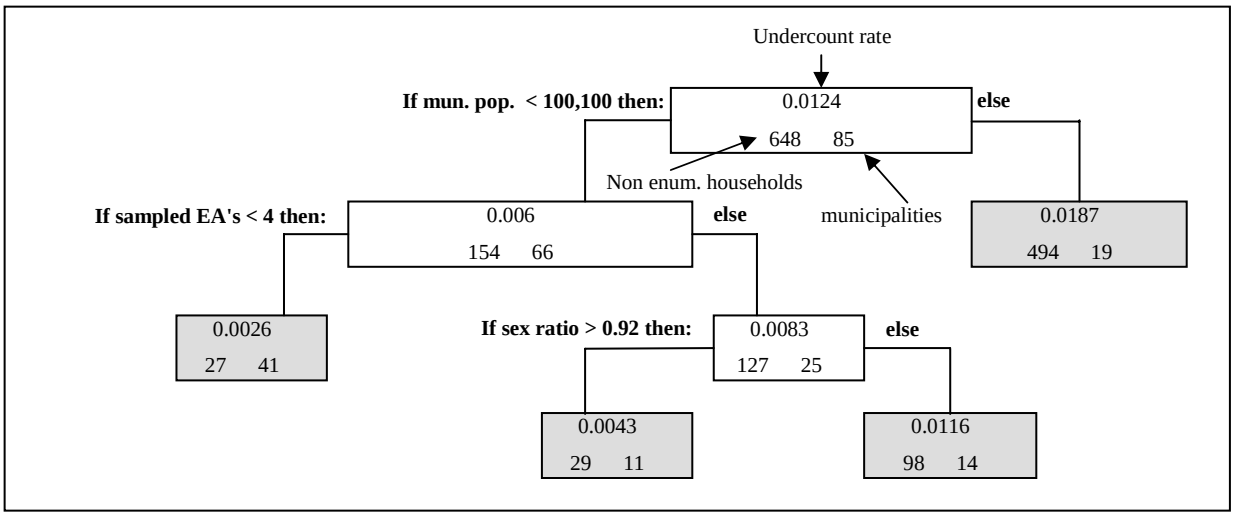



**Figure 1.** Tree 1 based on demographic variables.

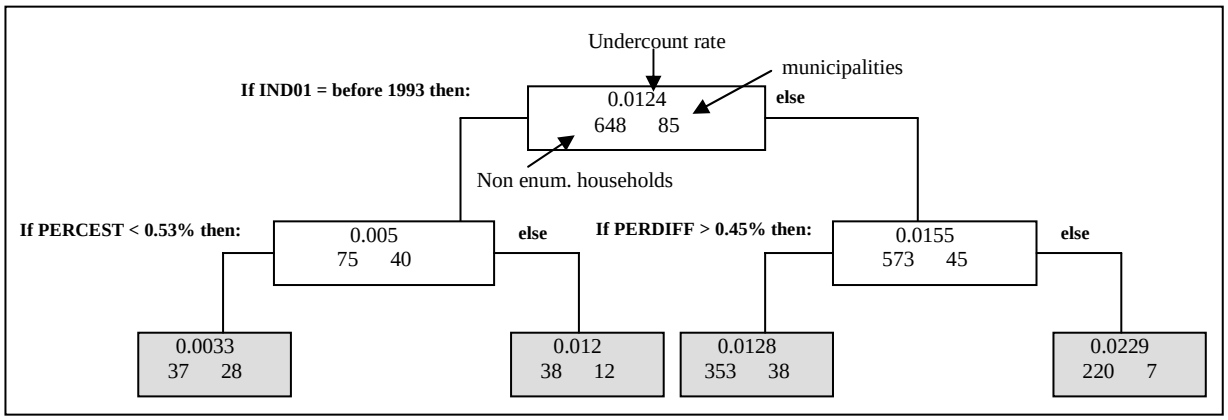



**Figure 2.** Tree 2 based on municipal statistical quality variables.

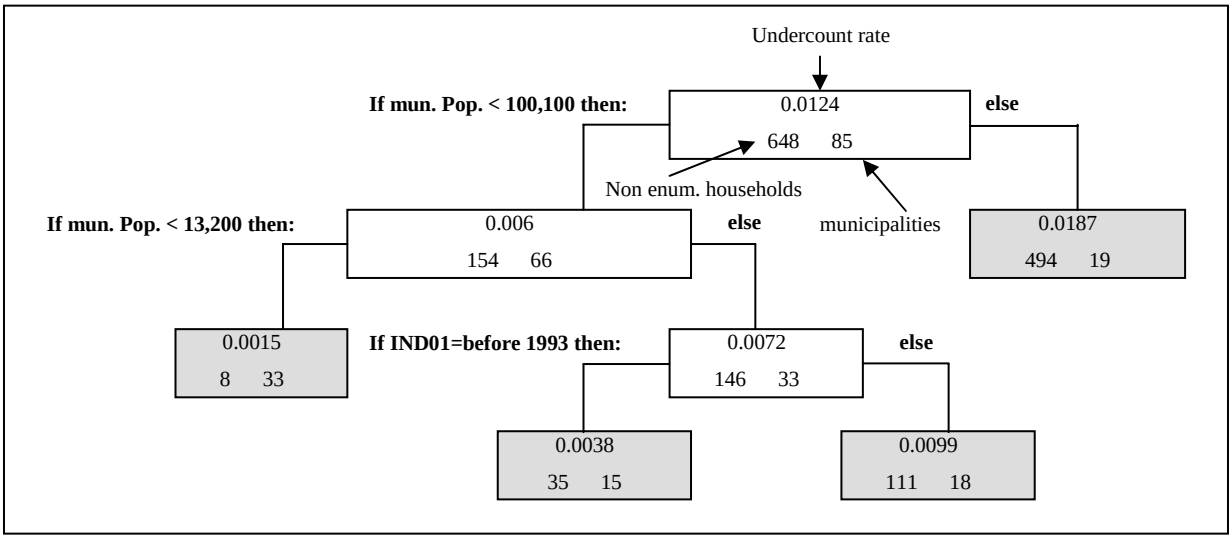



**Figure 3.** Tree 3 based on demographic and quality variables.

## 4. Hierarchical Poisson-Gamma Models

We denote the observed number of not enumerated households in each municipal sample with $y_i\,(i = 1, ..., 85)$. As an initial approximation, these counts can be modeled using a Poisson distribution:

$$y_i \mid \delta_i,\ e_i \sim \text{Pois}(\delta_i e_i) \tag{2}$$

where $\delta_i$ represents the rate of undercount to be estimated and $e_i$ is given by the number of households in the sampled EAs within the municipality. Dependency on a set of explanatory variables is expressed by means of a canonical log-linear link:

$$\ln(\delta_i e_i) = X_i \beta + Z_i \xi \tag{3}$$

where $Z_i$ is the $i^{\text{th}}$ row of a categorical design matrix introduced for modelling group effects. Each $X_i$ is a $p-$vector of explanatory variables associated with the $i^{\text{th}}$ municipality and $\beta$ and $\xi$ are the regression parameters.

The occurrence of failure to enumerate is relatively rare when compared to the number of observed households. For this reason, the data may show strong overdispersion. Overdispersion can be managed by hierarchically modelling the parameters $\delta_i$ in (2). If the $\delta_i$ are Gamma $(\alpha,\ \nu)$ distributed, the Negative Binomial distribution is marginally obtained for $y_i$ by integrating out the parameters $\delta_i$: *i.e.*, $y_i \mid \alpha,\ \nu, e_i \sim \text{NegBin}\ (\alpha, \nu/(\nu + e_i))$ with moments:

$$E(y_i \mid \alpha,\ e_i,\ \nu) = \frac{\alpha e_i}{\nu},\ V(y_i \mid \alpha,\ e_i,\ \nu) = \frac{\alpha e_i\,(\nu + e_i)}{\nu^2}$$

(see Lawless 1987).

Instead of the parameterization above, we adopt the parameterization of the Gamma distribution at the second level of the hierarchy according to the proposal made by Christiansen and Morris (1997). When assuming

$$\delta_i \mid \lambda_i,\ \zeta \sim \text{Gamma}(\zeta,\ \zeta/\lambda_i) \tag{4}$$

with moments $E(\delta_i \mid \lambda_i,\ \zeta) = \lambda_i$ and $V(\delta_i \mid \lambda_i,\ \zeta) = \lambda_i^2/\zeta$, we have

$$y_i \mid e_i,\ \lambda_i,\ \zeta \sim \text{NegBin}\left(\zeta,\ \frac{\zeta/\lambda_i}{\zeta/\lambda_i + e_i}\right),$$

where $V(y_i \mid e_i,\ \lambda_i,\ \zeta) - E(y_i \mid e_i,\ \lambda_i,\ \zeta) = e_i^2\ \lambda_i^2/\zeta$. As $\zeta$ moves towards infinity, the variance of the Negative Binomial converges towards that of the Poisson (the variance of the Gamma in (4) tends towards 0), while small values of $\zeta$ point to high overdispersion.

From (4) it is immediate to see that:

$$E(\delta_i\ e_i \mid e_i,\ \lambda_i,\ \zeta) = \lambda_i\ e_i;$$

therefore the dependence assumption (3) is re-stated in terms of $\lambda_i\ e_i$:

$$\ln(\lambda_i\ e_i) = X_i\ \beta + Z_i\ \xi.$$

The prior (4) is conjugate to the likelihood defined by (2). Consequently one obtains

$$\delta_i \mid y_i,\ e_i,\ \lambda_i,\ \zeta \sim \text{Gamma}(y_i + \zeta,\ e_i + \zeta/\lambda_i)$$

from which it follows that

$$E(\delta_i \mid y_i,\ e_i,\ \lambda_i,\ \zeta) = (1 - B_i)\,r_i + B_i\ \lambda_i \tag{5}$$

where $r_i = y_i/e_i$ and $B_i = \zeta/(\zeta + e_i\lambda_i)$.

Each posterior mean (5) can be seen as a composite Small Area estimator where both the direct and the synthetic components are weighted according to the information available from the sample.

From (5) we note that the posterior mean of the distribution of the rate parameters $\delta_i$ is a linear combination of the observed undercount rate $r_i$ and the prior mean $\lambda_i$. In other words, the model features posterior linearity. The two terms in (5) are weighted according to $B_i$, which varies between 0 and 1. The larger the $B_i$, the more the prior means $\lambda_i$ (synthetic estimators) receive weight and the model estimates gain in importance compared with the observed rates. We note that each $B_i$ is inversely proportional to the $e_i\ \lambda_i$, expressing the amount of information provided by the sample of each domain.

To complete the full Bayesian specification of the model we assign a distribution to the third level parameters $\zeta$, $\beta$, $\xi$. According to an approximate non-informative criterion, we introduce proper, but flat, prior distributions. In particular we assume that:

$$\beta_j \overset{\text{iid}}{\sim} N(0,\ 100), \qquad j = 1, ..., p \tag{6}$$

$$\xi_k \overset{\text{ind}}{\sim} N\left(k\bar{u}_k,\ \frac{1}{\tau\,\bar{n}_k}\right), \quad k = 1, ..., q \tag{7}$$

where $\bar{u}_k$ is the average undercount in the $k^{\text{th}}$ group and $\bar{n}_k$ is the average number of sampled households in the municipalities of the same group. Priors (7), associated to group effects, are therefore centered on groups means and their precision is proportional to the group size. They are built to be weakly informative for improving the stability and convergence properties of the model. Priors for regression coefficients (6) associated to the remaining regressors are centered in 0. For the overdispersion parameter $\zeta$ we select the prior

$$\zeta \sim 1{,}000^{*}\ \text{Gamma}\,(0.001,\ 1) \tag{8}$$

following the suggestion given by Christiansen and Morris (1997). Note that the first two prior moments of (8) are $E(\zeta) = 1$ and $V(\zeta) = 1{,}000$; thus the prior is very diffuse and characterized by high positive skewness.

At the fourth level of the hierarchy we specify the following priors:

$$k \sim N(0,\ 100) \tag{9}$$

$$\tau \sim \text{Gamma}\,(0.001,\ 0.001). \qquad (10)$$

which are both designed to have a very mild impact on posterior inferences.

We compute the posterior distributions of $(\delta_i \mid y_i,\ e_i)$ by using Markov chain Monte Carlo (McMC) sampling algorithms. For these calculations we use the software BUGS (Spiegelhalter, Thomas, Best and Gilks 1995), which is based on Gibbs sampling. Since the solution of models involving discrete distributions is computationally very demanding, we specify the prior distributions (6) – (10), by selecting simple well know functional forms, as Normal and Gamma, that facilitate fast computations. We examined the sensitivity of the posterior means in (6) – (10), and we did not find any substantial changes in the posterior means. Hence, these priors can be considered noninformative. For the convergence assessment we consider the multiple chain approach suggested by Gelman and Rubin (1992), running three different chains with well separated starting points for each model. The visual inspection of the chains path and the modified Gelman and Rubin statistic (Brooks and Gelman 1998) are considered as basic convergence assessment tools. We run 10,000 iterations for each chain, discarding on average a conservative "burn in" if 3,000, thus yielding an approximate 20,000 draws from the posterior of each model.

## 5. Model Comparison and Discussion of Empirical Results

We estimated a variety of models for different definitions of the matrixes of regressors *X* and *Z*. As regards the design matrix $Z$ we consider seven different cases, in which municipalities are grouped using either traditional stratification criteria (geographical area and demographic size) or the results of the partitioning techniques discussed in section 3. They are: a) geographical area (North, Center, South and Islands), b) demographic size classes only, c) demographic classes by geographical area, d) demographic size classes and geographical areas, e) Tree 1 (based on demographic variables), f) Tree 2 (based on quality variables), tree 3 (based on both quality and demographic variables). Two kinds of variables may be proposed in matrix $X$: the quality variables of section 2.2 and the demographic variables of section 2.3. Matrix $X$ has therefore three different possible compositions: I) quality variables only, II) demographic variables only, III) both quality and demographic variables. By matching the different definitions of $X$ and $Z$, twenty-eight different models have been estimated. In this way we do not perform variable selection procedures, rather we introduce alternative blocks of variables.

The quantity commonly used for comparing models within the Bayesian framework is the Bayes factor (*BF*). A large sample approximation of –2ln (*BF*) is given by

$$\Delta\text{BIC} = -2\ln\left[\frac{\sup_{M_0} f(y \mid \theta_0)}{\sup_{M_k} f(y \mid \theta_k)}\right] - (p_k - p_0)1\text{n}\,n \qquad (11)$$

(see Schwarz 1978) which, moreover, makes no reference to the prior assumptions. We note that in (11) the $M_k$ $(k = 1, ..., K)$ index the set of competing models and $\theta_k$ is the $p_k$ dimensional parameter indexing the likelihood associated to each model. The null model against which all the others are compared is the one with the only intercept, and is denoted by $M_0$. Positive and large values of (11) support model $M_k$.

The complexity penalization in (11) depends on the size of the subset of third level parameters; that is, all models are compared as if they were non hierarchical. Since they share a similar hierarchical structure, this operational modification of the standard *BIC* criterion does not alter the results of model comparison summarized in Table 1.

We note that those models where group effects are based on geographical area perform very poorly (row 1), and the same happens when the geographical area is combined with the demographic size of the municipalities (rows 3 and 4). This is rather surprising, since geographical areas are employed in designing the stratification of the PES sample, and the efficiency of administrations, together with other social and economic indicators, are currently supposed to be clustered with respect to Italy's large geographical subdivisions (North, Center, South). This outcome may be ascribed to the predominant role that the specific organization of each municipality plays in determining the efficiency of Census operations within its territory.

Models with tree-based group effects (rows 5 – 7) clearly perform better than models with group effects based on ISTAT traditional stratification criteria (rows 1 – 4). The only exception to this behavior are those models relying on Tree 2 (row 5), which perform rather poorly when demographic size and other demographic variables are not included. In fact, the municipal population can be thought of as a proxy of municipal organizational complexity. It seems that quality variables are powerful in discriminating the level of undercount among municipalities with similar demographic features, but have little relevance when the effect of a different degree of organizational complexity is not accounted for by introducing a variable of demographic size. We point out that adding a design matrix *Z* based on Poisson regression trees grouping of municipalities allows us to model non linear relations between the undercount and the predictors.

Actually, the models based on Tree 3 provide the best performance. A number of comments about the model with maximum ΔBIC follow. This model uses demographic and quality variables as regressors. The adequacy of the selected model is assessed by means of posterior predictive checks. In particular the general purpose goodness-of-fit discrepancy measure proposed by Brooks, Catchpole and Morgan (2000) as a suitable tool for rare occurrencies as census undercounts:

$$D(y;\theta)=\sum_i \left(\sqrt{y_i}-\sqrt{\mathrm{Exp}_i}\right)^2, \qquad (12)$$

where $\mathrm{Exp}_i = e_i\, E(\delta_i \mid y_i,\, e_i)$, is adopted. The associated 0.46 tail area probability highlights a good fit for the selected model.

The set of models has been estimated again after eliminating the greatest municipality, which is potentially an influential case. Again, the model based on Tree 3 with demographic and quality variables as regressors has been selected using the criterion (11). This model shows a good fit (the Bayesian $p-$value associated to the discrepancy measure (12) is equal to 0.51). Moreover, composite estimates do not change much when compared with those obtained with the whole sample.

In order to check model fitting, in Figure 4, composite estimates against direct estimates of the number of not enumerated household in each municipality are plotted (the values of the largest 10 municipalities are reported with a different scale). The composite estimates are $w_i\, e_i\, E(\delta_i \mid y_i,\, e_i)$, while the direct estimates are $w_i\, y_i$, $w_i$ being the expansion factor due to EA sampling in each municipality. Composite estimates are posterior expectations of first level parameters and, conditionally on the hyperparameters, are composite estimates in which the model predictions represented by the $\lambda_i$ receive little weight when there is sound sampling evidence. From (5) we know that this weighting process is ruled by the municipal shrinkage factors $B_i$. They weight the direct estimates $y_i / e_i$ in proportion to $e_i\, \lambda_i$, *i.e.*, the number of not enumerated households within the municipal sample predicted by the model.

**Table 1**
$\Delta$BIC of the Estimated Models Compared with the Reference Model $M_0$

| | | Variables in the models | | | |
|---|---|---|---|---|---|
| | | Only group effects | Group eff. + quality vars | Group eff. + demographic vars | Group eff. + quality and demographic vars |
| Group Effects | Area | −4.22 | −0.39 | 18.52 | 23.32 |
| | Classes of Mun. Pop. | 15.34 | 17.87 | 17.32 | 20.09 |
| | Area* Mun. Pop. Classes | 2.08 | 6.13 | 4.91 | 8.45 |
| | Area + Mun. Pop. Classes | 9.68 | 13.20 | 13.74 | 17.83 |
| | Tree2 (quality vars) | 11.81 | 8.34 | 23.48 | 26.15 |
| | Tree1 (demographic vars) | 35.14 | 35.37 | 32.28 | 35.53 |
| | Tree3 (quality + demographic vars) | 38.89 | 35.76 | 41.12 | 41.45 |

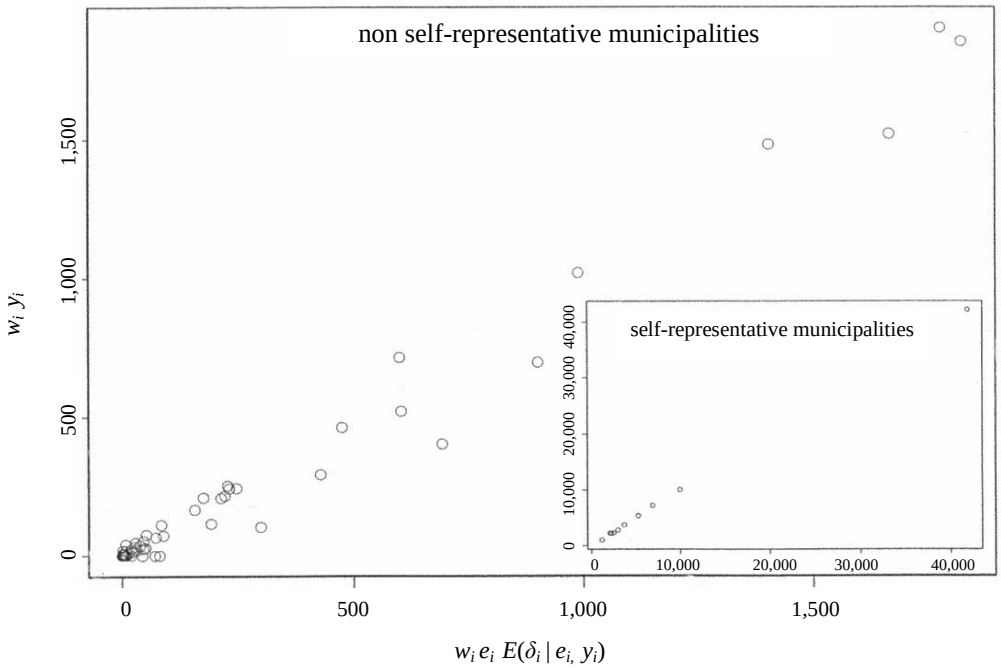



**Figure 4.** Composite estimates against direct estimates of the number of not enumerated households in each municipality.

For municipalities with resident population of up to 10,000 (this value is relatively close to the splitting value 13,200 of Tree 3) in almost all cases we have $B_i$ values that are very close to 1; this means that, for small municipalities, the role of the model component in the determination of the composite estimate is overwhelming. In Figure 5 composite estimates (and their 95% credibility intervals) are plotted against direct estimates.

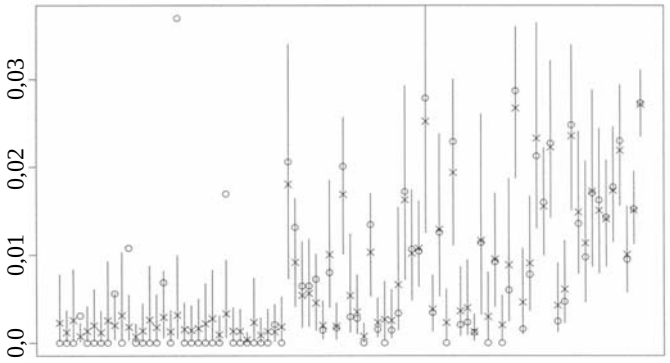



**Figure 5.** Composite estimates (×) and their 95% credibility intervals; (◦) direct estimates. Municipalities are sorted by demographic size.

The width of the credibility intervals depends on the undercount level and, as should be expected, is large when the size of the sample within the municipality is small. Composite estimates associated with large credibility intervals are also characterized by large shrinkage factors, as a consequence of the scarce sample information. Large intervals for some middle-sized municipalities can be justified with the fact that they are under-sampled with respect to their size.

In small municipalities, where Census is conducted more easily, the undercount is generally very small. The undercount estimate is difficult since very few EAs are currently sampled from each of the small municipalities, often providing no evidence of undercount. In such cases, the composite estimate essentially consists in the model based component. Therefore, for the next PES, given the overall sample size, our suggestion is not to insist in sampling a great number of small municipalities, but to redirect sampling towards middle-sized municipalities, which are more etherogeneous. Moreover, the number of EAs to sample in the selected small municipalities ought to be increased.

The results of this work, which considers for the first time a criterion for grouping together municipalities according to their performance in statistical operations, confirm that an improvement may be reached for future similar surveys by modifying the stratified sampling design and by modelling undercount by means of the covariates mimicking the difficulties of the municipality behaviour in conducting censuses.

## Acknowledgements

We would like to thank Angela Ferruzza, Marco Fortini, Aldo Orasi and Fernanda Panizon of the ISTAT team working on the 2001 Census and PES, together with Mariella Dimitri and Ersilia Di Pietro, of the ISTAT group working on surveys of statistical performance of municipalities, for their useful suggestions and continuous assistance.

The work has been partially funded by the (1999 – 2000) "Quality of total and partial surveys" Research Project grant from the University of Bologna (60%).

The PES data set and the archives containing the data on municipalities have been made available thanks to a special agreement between ISTAT and the Department of Statistics of the University of Bologna.

We would like to thank Francesca Bruno and Loredana Di Consiglio for their invaluable contribution in preparing the basic data sets, and Meri Raggi for her constant support and her discussion of the subjects of this research.

We thank the Editor, an Associate Editor and two anonymous referees for comments and suggestions which helped us in revising and improving the manuscript.

## References

Abbate, C., Masselli, M. and Signore M. (1993). A combined post-enumeration survey for the 1991 Italian population and industrial censuses. *Bulletin of the International Statistical Institute, Firenze, 48th Session*, Tome LV, 2, 159-173.

Alho, J.M., Mulry, M.H., Whurdeman, K. and Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, California.

Brooks, S.P., Catchpole, E.A. and Morgan, B.J.T. (2000). Bayesian animal survival estimation. *Statistical Science*, 15, 357-376.

Brooks, S.P., and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulation. *Journal of Computational and Graphical Statistics*, 7, 434-455.

Christiansen, C.L., and Morris, C. (1997). Hierarchical Poisson regression models. *Journal of the American Statistical Association*, 92, 618-632.

Di Pietro, E. (1998). Anagrafi comunali: funzione statistica e livello di informatizzazione. *Atti Della Quarta Conferenza Nazionale di Statistica*. Tomo 1 – Sessioni Plenarie, Workshop: II progetto anagrafi. Roma 11-13 novembre.

Di Pietro, E. (1999). Anagrafe informatizzata e Censimenti demografici : dal censimento tradizionale al censimento basato sugli Archivi. *Società Italiana di Statistica* : *Atti Del Convegno* : *"Verso i Censimenti del 2000*. Udine 7-9 giugno. 169-182.

Fortini, M. (1994). Un'applicazione de modello a classi latenti per l'analisi dell'errore di copertura del XIII censimento della popolazione. *Atti della XXXVII Riunione Scientifica della Società Italiana di Statistica*. San Remo 6-8 Aprile. 2, 423-430.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequence. *Statistical Science*, 7, 457-72.

Lawless, J.F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15, 209-225.

Moura, F.A.S., and Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology*, 25, 73-80.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Spiegelhalter, D.J., Thomas, A. , Best, N. and Gilks, W.R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50*. Technical Report, Medical Research Council biostatistics Unit, Institute of Public Health, Cambrdige University.

Therneau, T.M., and Atkinson, E.J. (1997). *An Introduction to Recursive Partitioning Using the RPART Routines*. Technical report, Mayo Foundation.

Wolter, K. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81, 338-346.