REVSTAT – Statistical Journal
Volume 6, Number 3, November 2008, 231–252

POST ENUMERATION SURVEY OF THE 2001 PORTUGUESE POPULATION AND HOUSING CENSUSES

Authors: Pedro Simões Coelho
– ISEGI, Universidade Nova de Lisboa, Portugal
psc@isegi.unl.pt

Fernando Casimiro
– Instituto Nacional de Estatística, I.P., Portugal
fernando.casimiro@ine.pt

Received: December 2007 Revised: March 2008 Accepted: April 2008

Abstract:

- Within the framework of the quality control and evaluation program for the Portuguese 2001 Census, the Portuguese statistical office (INE) conducted a Post Enumeration Survey (PES) to measure quality. The main aims of the PES were to evaluate coverage errors and content errors for the target populations. The PES is a probabilistic sampling survey representative at NUTS II level. This paper describes the methodology for this survey. The paper includes a discussion of sample size and allocation resulting from the imposition of maximum coefficients of variation for a set of variables both at regional and national level. The methodology used to obtain predictions for resident populations and dwellings is also presented. These predictions are used in the definition of inclusion probabilities for the primary sampling units. The sampling design is finally compared with two alternative designs (with a smaller number of stages), concluding for the advantage of the proposed design in regard to the survey goals.

Key-Words:

- *Post Enumeration Survey; sampling design; census; population and housing census; census quality; census errors; coverage errors.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

1.1. Context

Quality is an increasingly important subject in the production of statistics. Customers tend to be increasingly demanding and critical about statistical data.

For population and housing censuses, quality evaluation is carried out in various ways, one of them being the Post Enumeration Survey (PES). Usually, when they exist, PES results are assumed to be the final quality indicators for these censuses. "The PES, a special kind of survey designed to measure census coverage and/or content error, has been used effectively in a wide range of countries in recent decades; (...) The final publication should include an estimate of coverage error, together with a full indication of the methods used for evaluating the completeness of the data..." ([20]).

Coverage and content evaluation of population census data are carried out with two main purposes:

(1) to provide customers with quality indicators;

(2) for internal use, to improve knowledge of the problems encountered, in order to improve the capacity to plan and conduct this type of statistical operation in the future.

1.2. Historical note

In the USA, evaluation of population census coverage and content began in 1950, while in Australia "the first PES was run in 1966, but the 1976 PES was the first to be used for population estimates" ([1]). In France, the first quality census evaluation with a PES was carried out in 1962 but the next assessment was not held until 1990 ([6]). Canada started measuring gross undercoverage in 1971, but estimates of net undercoverage are only available from 1991, as 1991 marked the first comprehensive measure of overcoverage following an experimental study in 1986 ([19]). The UK first used a PES to measure census quality in 1961, since when a PES has been conducted in every census ([14, 15, 16, 12]). In other countries the introduction of a PES began later, as in the case of New Zealand where the first one was conducted in 1996.

In fact, post enumeration surveys have been regularly used in countries where a population census is performed, but for 2000 and 2001 some countries

made an even heavier investment in the measurement of census data quality. This is the case of Canada, United Kingdom and the USA ([19, 12, 3, 4]).

The first attempt to conduct a PES in Portugal occurred with the 1981 census, though technical constraints and the lack of human resources did not allow the task to be completed. However, for internal purposes only, a comparative tabulation was made using the two equivalent samples of statistical units (census and PES), which were used to produce two independent but equivalent sets of data tables. For the 1991 census a new PES was designed with strict rules on each person in each selected household being re-interviewed: that is, each person had to be re-interviewed face to face and no other person in the household could be substituted. In Census 91 it was possible to produce gross and net indicators on coverage for each statistical unit but not on the content of census variables. The delay in the census fieldwork and consequently in the PES also made it very difficult to match and apply the automatic rules for the imputation of responses in the PES questionnaires in the same way as for the census responses.

Certain leaders of public opinion also expressed their doubts about the quality of coverage in the 1991 census (about 1% net undercoverage on population, measured by the PES), which led the Portuguese Statistical Office (INE) to decide that the estimates of quality for the 2001 Census should be clear and proven. In order to reach this target, a "special" programme on quality evaluation was designed in which PES was to provide the final quality measure for the 2001 Census.

Given the unexpected outcome from the 1991 Census (the count being approximately 5% below the population estimates made by National Statistical Institute (INE) itself before the first results of the 1991 Census became available), INE was convinced that the 2001 Census would be subject to very close scrutiny by its main users. In a way, despite the fact that ten years have passed since 1991, the 2001 census data would end up being an important evaluation factor of the 1991 census, given that no significant or unexpected demographic "accident" has taken place or was expected to take place in the country's demographic development ([5]).

1.3. The 2001 Portuguese post enumeration survey

The main goal of the census 2001 PES was to evaluate coverage and content errors, giving information to census users about the accuracy of the results, thus allowing to assess the risks involved in basing conclusions or decisions on census data. Coverage and content errors are evaluated for the following universes: buildings, dwellings, private households and resident population.

The evaluation of coverage errors includes three main causes:

(1) Statistical units of the target populations that have not been enumerated;

(2) Statistical units outside the target populations that have been wrongly enumerated;

(3) Statistical units that have been enumerated more than once.

The evaluation of content errors includes census flaws related to observing statistical unit characteristics that can affect the quality of census information about resident population and housing.

To assess coverage and content errors the census enumeration process was repeated in the selected sampling units. At statistical section level recounts of buildings, dwellings, private households and resident population were made. Also, the various types of questionnaires are again completed, for the different statistical units, regarding the characteristics that those units had on census day.

It should be noted that the Portuguese PES presents a number of specificities when compared to other post enumeration surveys, namely: not only aimed to measure coverage errors but also content errors; all measures were obtained using only one sample (though data were obtained from different sampling stages); it was designed within a framework where no sampling frames besides administrative division of the country and auxiliary information regarding population and dwelling estimates were available; it used a three-stage design with selection probabilities proportional to size in the first two sampling stages; it was designed to avoid the selection of sampling units being dependent on the conclusion of census fieldwork in order to reduce the time between the census date and the implementation of the post enumeration survey; it is meant to use information on the geographical coordinates of sampling units in the sampling design; sample size and allocation are obtained by means of an optimization problem that tries to minimize the overall survey cost.

This paper discusses the methodology for this survey. The first section introduces the problem and its context. The sampling design is presented in the second section. Also, the methodology used for defining sample size and allocation between strata, resulting from the imposition of maximum coefficients of variation (CVs) both at regional and national level, for a set of variables, is presented. The following section presents the methodology used in producing predictions for the resident population and dwellings at the time of the census. These predictions are used as auxiliary information in the definition of inclusion probabilities for the primary sampling units. The paper finishes with some final remarks and a discussion about the sampling design (compared with two alternative designs with a smaller number of stages).

2. SAMPLE DESIGN

2.1. Introduction

The quality survey for the 2001 census is a probabilistic sampling survey. It covers the whole of the national territory and aims to be representative at NUTS II[1] level for the variables *dwellings, private households, resident population, active population, employed population, resident population aged 18 years or more and population by decennial age group between 20 and 80 years of age.* Figure 1 shows the partition of Portugal into the 7 NUTS II.

Figure 1: NUTS II division.

A sample of statistical sections[2] is used to evaluate coverage errors for buildings and dwellings, while a sample of dwellings is used to assess coverage errors for private households and resident population and content errors.

[1]NUTS (Nomenclature of Territorial Units for Statistical Purposes) II is an administrative division that divides the country into seven regions (*Norte*, *Centro*, *Lisboa e Vale do Tejo*, *Alentejo*, *Algarve*, *Região Autónoma dos Açores* and *Região Autónoma da Madeira*).

[2]The statistical section is a statistical division corresponding to an area belonging to a single parish (*freguesia*) with approximately 300 dwellings.

The sample is previously stratified by NUTS II. In each stratum a sample of *freguesias*[3], statistical sections and dwellings is obtained. The approach includes the selection, in each stratum, of a multi-stage self-weighted sample through systematic selection with probability proportional to size (pps) at the first and second stage. The primary sampling units are *freguesias*, the secondary sampling units statistical sections and the tertiary sampling units dwellings. The sampling design assures equal probability of selection for dwellings within strata. See [17, pp. 144–150] for general theory about multi-stage designs.

At the first sampling stage inclusion probabilities are defined through the use of auxiliary information based on resident population and estimates of dwelling totals, at census time. An exception is made in the stratum of the *Algarve* where the selection of the sub-sample of *freguesias* is based on resident population estimates in each *freguesia*.

Auxiliary information resulting from preliminary counts from the questionnaire delivery phase of the census is used to define inclusion probabilities for the statistical sections (secondary sampling units). This is due to the impossibility of producing reliable estimates (for population or dwellings) at statistical section level. With this approach it is possible to incorporate updated and high quality auxiliary information in the selection process for statistical sections, which contributes to a more efficient sampling design.

It should be noted that primary unit selection is carried out a priori, i.e. before the census date, using estimates for the number of dwellings and the resident population by *freguesia*. On the other hand, statistical sections are selected as soon as the counts from the questionnaire delivery phase of the census are obtained for each of the previously selected *freguesias*. Given the multistage nature of the sampling design, the selection of statistical sections is not dependent on the conclusion of all counts in the questionnaire delivery phase, but only those referring to *freguesias* selected at the first stage. Such dependence would be undesirable, given the obvious interest in reducing the time between the census date and the quality survey.

Finally, at the third stage, the dwelling samples are extracted, through systematic selection and with equal probabilities as soon as the dwelling recounts are completed for statistical sections selected at the second sampling stage.

A more detailed description is given in the following sections.

[3] *Freguesia* (NUTS V) is an administrative division corresponding to one or more Statistical Sections. At the census day there were 4,208 freguesias in Portugal.

2.2. Selection of *freguesias* (primary units)

At the first sampling stage, *freguesias* are selected in each region (stratum) with probability proportional to the estimated number of dwellings. For the *Algarve* the selection probability is proportional to the estimated resident population. The use of a different approach is motivated by the weak correlation (observed in the simulations using data from the 1991 census) between resident population and dwellings in that region. Therefore, the choice of an alternative sampling design for this region contributes to a significant reduction in sampling effort (cf. Section 2.5).

Freguesias are sorted beforehand using the geographical coordinates of their centroids[4]. In each stratum *freguesias* are ordered by ascending order of their Euclidean distance from the origin. The goal is to assure that the sample is geographically dispersed while still allowing a probability of selection proportional to its size. Finally, *freguesias* are selected through systematic sampling.

The selection probability for freguesia i of stratum h was defined as

$$\pi_{hi} = \begin{cases} \dfrac{A_{hi}}{I_h} & \text{if } A_{hi} < I_h \,, \\ 1 & \text{otherwise} \,, \end{cases} \tag{2.1}$$

where A_{hi} is the estimated number of dwellings (population for the *Algarve*) of *freguesia* i of stratum h.

The selection interval for *freguesias* at stratum h, I_h, is

$$I_h = \frac{A_h}{m_h} \tag{2.2}$$

where m_h is the number of statistical sections to be selected for the sample in stratum h and A_h is the estimated number of dwellings (population for the *Algarve*) in stratum h.

Note that the selection interval I_h is inversely proportional to the number of secondary sampling units, m_h. As it will be explained in more detail in the next section, the reasoning behind this choice is to support the selection of only one secondary sampling unit (*statistical section*) at each primary sampling unit (*freguesia*) with selection probability lower than one.

[4]The point of origin of these coordinates is situated in the Atlantic Ocean to the southwest of Portugal.

2.3. Selection of statistical sections (secondary units)

Lists of statistical sections are formed in *freguesias* selected at the first sampling stage. These sections are sorted using their geographical coordinates (distance from the centroid to the origin).

At the second sampling stage statistical sections are selected through systematic sampling with probability proportional to the number of dwellings obtained in the preliminary counts from the questionnaire delivery phase of the census.

The selection probability for section j of *freguesia* i of stratum h conditioned to the selection of *freguesia* hi is defined as

$$\pi_{hij|hi} = \begin{cases} \dfrac{N_{hij}}{N_{hi}} & \text{if } A_{hi} < I_h \,, \\ \dfrac{A_{hi}}{I_h}\,\dfrac{N_{hij}}{N_{hi}} & \text{otherwise} \,, \end{cases} \tag{2.3}$$

where N_{hij} is the number of dwellings in section j of *freguesia* i of stratum h (data from the preliminary counts from the questionnaire delivery phase of the census), N_{hi} is the number of dwellings at *freguesia* i of stratum h (data from the preliminary counts from the questionnaire delivery phase of the census).

The unconditional selection probability for section j of *freguesia* i of stratum h is consequently

$$\pi_{hij} = \frac{A_{hi}}{I_h}\,\frac{N_{hij}}{N_{hi}} \,. \tag{2.4}$$

To guarantee this selection probability, the selection interval in *freguesia* i of stratum h is defined as

$$\mathbb{I}_{hi} = \begin{cases} N_{hi} & \text{if } A_{hi} < I_h \,, \\ \dfrac{N_{hi}}{A_{hi}}\, I_h = \dfrac{N_{hi}\, A_h}{A_{hi}\, m_h} & \text{otherwise} \,. \end{cases} \tag{2.5}$$

The number of statistical sections being re-enumerated in each *freguesia* selected at the first sampling stage is equal to one for *freguesias* with selection probability lower than one, allowing therefore a strong dispersion of sampled sections in a high number of *freguesias*. The sample size at the second stage will only be higher than one for *freguesias* with selection probability equal to one in order to assure that the unconditional selection probability at the second stage remains proportional to $A_{hi}\, N_{hij}/N_{hi}$.

The number of statistical sections being re-enumerated in *freguesia* i will then be

$$m_{hi} = \begin{cases} 1 & \text{if } A_{hi} < I_h \,, \\ m_h\,\dfrac{A_{hi}}{A_h} & \text{otherwise} \,. \end{cases} \tag{2.6}$$

2.4. Selection of dwellings (tertiary units)

The estimation of coverage errors relative to buildings and dwellings is achieved through the sample of secondary units. For that purpose, each statistical section in the sample should be exhaustively re-enumerated in order to obtain the "true" totals for buildings and dwellings. After obtaining these recounts, a list of dwellings is formed in each statistical section. These lists are used to select the samples of dwellings (tertiary units) to be re-enumerated.

At this third sampling stage dwellings are selected through systematic sampling, with equal probabilities, in order to obtain a self-weighted sample in each stratum.

The selection probability for dwelling k, of section hij, conditioned to the selection of the section to which it belongs, is defined as

$$\pi_{hijk|hij} = \frac{n_{hij}}{N'_{hij}} \tag{2.7}$$

where n_{hij} is the number of sampled dwellings in section hij and N'_{hij} is the number of dwellings in section hij, (obtained from the second stage of the PES).

The unconditional selection probability for dwelling k of section hij is therefore belongs, is defined as

$$\pi_{hijk} = \pi_{hij} \cdot \pi_{hijk|hij} = \frac{A_{hi}}{I_h N_{hi}} \frac{N_{hij}}{N'_{hij}} n_{hij} \ . \tag{2.8}$$

In each section, the sample size for tertiary units is obtained in order to get a self-weighted sample of dwellings in each stratum. For that a constant selection probability is defined in each stratum, equal to the overall sampling rate $f_h = n_h/N'_h$.

The sample size at section j of *freguesia* i in stratum h is consequently given by

$$n_{hij} = \frac{f_h A_h N_{hi} N'_{hij}}{m_h A_{hi} N_{hij}} \ . \tag{2.9}$$

To guarantee the defined selection probabilities, the sampling interval in section hij is defined as

$$I_{hij} = \frac{m_h A_{hi} N_{hij}}{m_h A_h N_{hi}} \ . \tag{2.10}$$

Therefore, the resulting selection probability

$$\pi_{hijk} = \frac{A_{hi}}{I_h N_{hi}} \frac{N_{hij}}{N'_{hij}} n_{hij} = \frac{A_{hi}}{I_h N_{hi}} \cdot \frac{N_{hij}}{N'_{hij}} \cdot \frac{f_h A_h N_{hi} N'_{hij}}{m_h A_{hi} N_{hij}} = f_h$$

will be constant in each stratum.

2.5. Sample size and allocation

It should be remembered that the sample is previously stratified by NUTS II, resulting in seven strata. Also the survey is intended to be representative not only at national level but also at NUTS II level. The overall sample size and its allocation by each stratum was obtained as the solution to the following optimization problem[5]:

$$
\begin{aligned}
(2.11) \qquad & \min \left(C = \sum_{i=1}^{H} c_{1h} m_h + c_{2h} n_h \right) \\
& \text{s.t.} \\
& CV(\hat{\tau}_{k,h}) \leq d_{k,h} \qquad (k = 1, ..., K; \quad h = 1, ..., H) \\
& CV(\hat{\tau}_k) \leq d_k \qquad (k = 1, ..., K) \\
& m_h \leq M_h, \quad n_h \leq N'_h \qquad (h = 1, ..., H) \\
& m_h \geq 0, \quad n_h \geq 0 \qquad (h = 1, ..., H)
\end{aligned}
$$

where K is the number of variables considered for sample dimensioning, H the number of strata, m_h is the sample size of statistical sections in stratum h, n_h is the sample size of dwellings in stratum h, M_h is the number of sections in stratum h, c_{1h} the cost of observing one section in stratum h, c_{2h} the cost of observing one dwelling in stratum h, $CV(\hat{\tau}_{k,h}) = \dfrac{\sqrt{V(\hat{\tau}_{k,h})}}{\tau_{k,h}}$, $CV(\hat{\tau}_k) = \dfrac{\sqrt{V(\hat{\tau}_k)}}{\tau_k}$, $\tau_{k,h}$ is the population total of variable k in stratum h and $\hat{\tau}_{k,h}$ the Horvitz–Thomson estimator for the same parameter. Also, $\tau_k = \sum_{i=1}^{H} \tau_{k,h}$ is the population total of variable k, $\hat{\tau}_k$ is its estimator and $V(\hat{\tau}_k) = \sum_h V(\hat{\tau}_{k,h})$. As proved in Appendix 1, the variance $V(\hat{\tau}_{k,h})$ can be approximated by the expression

$$
\begin{aligned}
V(\hat{\tau}_{k,h}) \;\approx\; & \frac{1}{m_h} \left[\sum_{i \in U_{1h}^I} \frac{A_h A_{hi}}{A_{1h}^2} \left(\frac{A_{1h} \tau_{k,hi}}{A_{hi}} - \tau_{k,1h} \right)^2 \right. \\
& \left. + \sum_{i \in U_{1h}^I} \sum_{j \in U_{hi}^{II}} \frac{A_h N_{hij}}{A_{hi} N_{hi}} \left(\frac{N_{hi} \tau_{k,hij}}{N_{hij}} - \tau_{k,hi} \right)^2 \right] + \frac{N_h'^2}{n_h} \sigma_{k,h,intra}^2 \;,
\end{aligned}
$$

where $\sum_{i \in U_h^I}$ is the summation over all *freguesias* of stratum h, $\sum_{i \in U_{1h}^I}$ is the summation over the *freguesias* of stratum h where $\tau_{a,hi} < I_h$, $\sum_{j \in U_{hi}^{II}}$ is the summation over all the sections of *freguesia* i of stratum h, $\tau_{a,1h}$ is the estimate of the total of dwellings (residents in the *Algarve*) in population U_{1h}^I, $\tau_{k,1h}$ is the total of variable k in the same population and $\sigma_{k,h,intra}^2 = \sum_{i \in U_h^I} \sum_{j \in U_{hi}^{II}} \dfrac{N'_{hij}}{N'_h} \sigma_{k,hij}^2$ is the intra-section variance for variable k in stratum h. All other parameters are as defined in the previous sections.

[5]The problem was solved through Generalized Reduced Gradient Nonlinear Optimization.

The aim of the strategy adopted is to minimize the total sampling cost, C, with the application of maximum limits for the variation coefficients in estimating totals for the K selected variables, at regional level (the stratum corresponding to NUT II) and national level.

The maximum variation coefficients at NUT II level (d_{kh}) were established at 5% for the variables *dwellings, private households, resident population, active population, employed population, and resident population of 18 years of age or more* and at 7% for *resident population by decennial age group between 20 and 80 years of age*. Also, the maximum variation coefficients for estimating national totals (d_k) were set at 3% for the variables *dwellings, private households, resident population, active population, employed population, and resident population of 18 years of age or more* and at 3.5% for *resident population by decennial age group between 20 and 80 years of age*.

Figure 2 shows the geographical location of sections in the sample. Attention should be paid to the location of sample sections associated with high population concentrations in coastal areas and urban centers. The calculations used to determine sample sizes were based on data from the 1991 census[6].

It should be remembered that a specific sampling design was adopted in the Algarve, since the selection probabilities for primary sampling units in that NUTS II were defined as proportional to the estimated resident population. If the sampling design in the *Algarve* were the same as that adopted in other NUTS II, using the number of dwellings to define selection probabilities, the necessary sample size to guarantee the achieved variation coefficients would be equal to 122 statistical sections. This result clearly demonstrates the advantage of the procedure adopted, which is justified by the low correlation between resident population and dwellings in that region (cf. Table 1). In fact, one should remember that the *Algarve* is a tourist region where many people keep a second home.

It should also be noted that the high sample size obtained in *Lisboa e Vale do Tejo* is essentially justified by the significant variance that some of the variables show at *freguesia* level, as well as by the low correlation between resident population and dwellings at section level. This sample could be downsized using a design similar to that adopted in the *Algarve*[7]. The decision to keep the primary

[6]Since the simulation used 1991 census data summarized at *freguesia* level, the variances ($k = 1, ..., K;\ h = 1, ..., H$) were replaced by estimates obtained from other surveys conducted by INE. In practice, c_{1h} and c_{2h} were considered non invariant with h, i.e. $c_{1h} = c_1$; $c_{2h} = c_2$, $h = 1, ..., H$. A new set of restrictions was also imposed, $f_h = f$, $h = 1, ..., H$, in order to achieve a constant selection probability for all dwellings in the country. The overall sampling rate f came out approximately equal to 0.00275. Simulations showed that the impact of this procedure on the overall sampling cost was moderate.

[7]In the simulations using data from the 1991 census the reduction in sample size (for the same level of precision) would be from 110 to 97 statistical sections.

units proportional to the number of dwellings in that region resulted from the observation that the reduction in sample size would be less significant than in the *Algarve*. In addition, the strategy adopted shows certain advantages in field procedures, associated with the greater stability of the sample size at the last sampling stage.

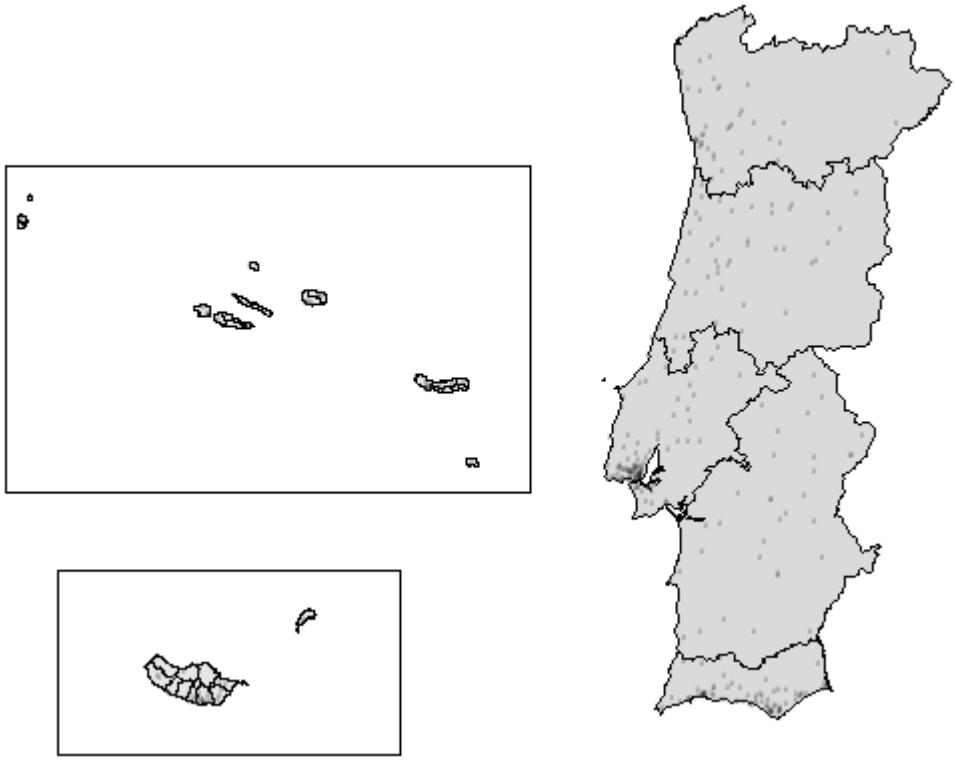

Figure 2: Statistical section in the sample.

Table 1: Standard deviations and correlations between dwellings and resident population (data from the 1991 census).

Stratum (NUTS II)	Freguesia			Statistical Section		
	Correlation	Standard deviation (dwellings)	Standard deviation (population)	Correlation	Standard deviation (dwellings)	Standard deviation (population)
Norte	0.99	1,286	3,532	0.84	135	422
Centro	0.97	992	2,283	0.75	99	278
LVT	0.98	3,936	9,608	0.57	84	249
Alentejo	0.99	795	1,849	0.75	82	212
Algarve	0.87	3,432	5,273	0.43	115	279
Açores	0.96	435	1,477	0.68	82	320
Madeira	0.98	1,350	5,131	0.61	73	318

3. DWELLING AND RESIDENT POPULATION ESTIMATES

It has already been mentioned that the selection probabilities at the first sampling stage use estimates of the number of dwellings and the resident population by *freguesia*, in relation to the census date.

The production of these estimates is addressed below, particularly as regards the data sources and methodology used. The proposed estimators were tested with the exhaustive observation of 107 *freguesias* in April 2000. This observation followed a similar approach to the one adopted in the 2001 census. Estimators for *freguesia* totals (obtained using the methodology presented in the following sections) showed a mean absolute relative error of 11% in estimating the resident population of *freguesias* in the test and of 8% in estimating the number of dwellings. Note that these errors are compatible with the approach used for determining sample size, since that approach was based on data from the 1991 census where the mean absolute relative error for the estimated number of dwellings was about 12%. A more detailed presentation of the methodology used can be found in [7].

3.1. Estimation of the number of dwellings by *freguesia*

Some of the data sources available for the number of dwellings were: the results of the 1991 census, statistics from the INE construction survey and EDP household registers (Portuguese electricity company data on domestic consumption locations).

The estimate for the number of dwellings in freguesia i on the census date is obtained as

$$A_{hi} = 0.5 \times C^{hi} + 0.5 \times D^{hi}_{2001} \ , \tag{3.1}$$

where, A_{hi} is the estimate of the number of dwellings in *freguesia* hi, at the census date, C^{hi} is the simple count of the number of domestic electricity consumption contracts in the EDP registers, by *freguesia*, D^{hi}_{2001} is an estimate derived from accumulating the annual balance of newly constructed dwellings and demolished dwellings (data from the construction survey), on the basis of the number of dwellings obtained in the 1991 census (corrected from the coverage error).

This last estimate is given by[8]:

$$\begin{aligned} D^{hi}_t &= D^{hi}_{t-1} + NewD^{hi}_t - Dem^{hi}_t \\ D^{hi}_{91} &= D^{hi}_{C91} \end{aligned} \tag{3.2}$$

[8]This methodology presented various limitations, namely: the absence of information about demolitions prior to 1993; the fact that data prior to 1994 is only available at *concelho* level

where D_t^{hi} is the estimate of the number of dwellings in *freguesia hi* at year t, $NewD_t^{hi}$ is the number of dwellings constructed in *freguesia hi*, at year t (data from the construction survey), Dem_t^{hi} is the number of demolitions in *freguesia hi*, at year t (data from the construction survey), D_{C91}^{hi} is the number of dwellings in *freguesia hi* counted in the 1991 census.

The decision to use an average of the two estimates is based on the observation (using the 107 test *freguesias*) that EDP data tends to systematically underestimate the true number of dwellings, while the estimate obtained from the construction survey shows a tendency of to overestimate them. In the absence of other information on the precision of each of the alternative estimators, a natural choice in producing a weighted average of both estimates is to use equal weighting.

3.2. Estimation of resident population by *freguesia*

Some of the data sources available on resident population were: the results of the 1991 census, official data on births and deaths in the decade, electoral roll databases, and the legalization of immigrants and residence cancellations (data from the Immigration Service).

The estimates produced result from the accumulation of the balance between births and deaths, the balance of transfers in the electoral rolls, and net immigration (taken as the difference between legalization requests and cancellations in the Immigration Service). The resident population enumerated in the 1991 census (corrected from the coverage error) was taken as a base for the calculation.

The estimate of the resident population in *freguesia hi*, at year t, is

$$\begin{aligned} Pop_t^{hi} &= Pop_{t-1}^{hi} + BB_t^{hi} + TrB_t^{hi} + FmB_t^{hi} \\ Pop_{91}^{hi} &= Pop_{C91}^{hi} \end{aligned} \tag{3.3}$$

where

- $BB_t^{hi} = Births_t^{hi} - Deaths_t^{hi}$;
- $Births_t^{hi}$ is the number of births in *freguesia hi*, at year t;
- $Deaths_t^{hi}$ is the number of deaths in *freguesia hi*, at year t;

(aggregation of *freguesias*); and the unavailability of data posterior to 1999. These limitations were overcomed by carrying out, respectively, an estimation of demolitions for 1991 and 1992 based on information about the following years, allocation of data prior to 1994 by *freguesia* using the average structure of each *concelho* in the period 1994–1999 and the prediction of the series value at the census day by adjustment of a linear regression model.

- $TrB_t^{hi} = T_t^{hi} - E_t^{hi}$;
  - T_t^{hi} is the number of transfers to *freguesia hi*, at year t, in the electoral roll;
  - E_t^{hi} is the number of cancellations due to transfer from *freguesia hi*, at year t, in the electoral roll;

- $FmB_t^{hc} = LR_t^{hc} - C_t^{hc}$, $FmB_t^{hi} = FmB_t^{hc} \frac{FPop_{C91}^{hi}}{FPop_{C91}^{hc}}$;
  - FmB_t^{hc} is the net migration of foreigners to *concelho* (municipality) *hc*, at year t;
  - LR_t^{hc} is the number of legalization requests from foreigners in *concelho hc*, at year t;
  - C_t^{hc} is the number of residence permit cancellations for foreigners in *concelho hc*, at year t;
  - FmB_t^{hi} is the net migration of foreigners to *freguesia hi*, at year t;
  - $FPop_{C91}^{hi}$ is the resident population of foreigners in *freguesia hi*, recorded in the 1991 census;
  - $FPop_{C91}^{hc}$ is the resident population of foreigners in *concelho hc*, recorded in 1991 census.

The estimated resident population for the census date is then

$$P_{hi} = Pop_{2001}^{hi} \ . \tag{3.4}$$

Thus, the estimator includes information about births and deaths, internal migrations and foreign immigration. It may be presumed that internal migration and foreign immigration had a great impact on the resident population at *freguesia* level, since there were less than 90,000 persons as natural increase at national level between 1991 and 2001.

It should also be noted that, although data for births and deaths are considered to be totally reliable, this is not the case with migration, both internal and international. For this reason, transfers in the electoral rolls were taken as a proxy for internal migration at *freguesia* level. In fact, the impossibility of producing reliable estimates for internal migrations at this aggregation level motivated the search for a variable that could be considered a proxy for internal migration as it was reliable at *freguesia* level. The main limitation is the fact that migrations of people under 18 years of age are not included. In addition, calculation of the estimates demanded the allocation of the net foreign migration (only available at *concelho* level) to *freguesias*, using the structure observed in the 1991 census.

4. DISCUSSION

The PES was designed to evaluate coverage errors and content errors in the main statistical units: buildings, dwellings, private households and resident population. For this purpose the enumeration process was repeated in the selected sampling units. The various questionnaires were completed again, for the different statistical units, with the characteristics that those units had at the time of the census day.

This paper discussed the methodology used for this survey, which is based on a three-stage sample. Primary and secondary sampling units were selected with probability proportional to the size. For this purpose, at the first sampling stage inclusion probabilities were defined through the use of auxiliary information based on the estimated resident population and dwelling totals, at the time of the census. Auxiliary information, resulting from preliminary counts obtained at the questionnaire delivery phase in the census, was used in the definition of inclusion probabilities for secondary sampling units. With this approach it was possible to incorporate updated and high-quality auxiliary information in the selection process for the statistical sections, contributing to a more efficient sampling design. In the first two sampling stages auxiliary information regarding the geographical coordinates of area units (statistical section and *freguesia*) was used in order to obtain an implicit stratification. Particularly with regard to the number of stages, the design took operational restrictions into account. The goal was to make use of available auxiliary information to determine the appropriate selection probabilities and to avoid the sample selection being totally dependent on the conclusion of all counts in the questionnaire delivery phase. In fact, the lowest level of aggregation for which information used to define inclusion probabilities is available is the *freguesia* level. So, the selection of statistical sections as primary sampling units would make it impossible to use reliable auxiliary information in defining such probabilities.

Sample size and allocation between strata were obtained as the solution to an optimization problem that minimizes the total survey cost, with maximum limits for the coefficients of variation in estimating totals for a number of variables, both at regional level (the stratum corresponding to NUTS II) and at national level. This approach resulted in a sample of 367 statistical sections.

Nevertheless, alternative sampling designs, with a smaller number of stages, could have been conceived, in order to try to reduce a possible design effect. This alternative approach could then be based on a two-stage sampling, successively selecting statistical sections and dwellings.

One possible approach would then be to select statistical sections with equal probability. A two-stage sampling with selection of statistical sections with equal probability at the first stage was simulated using data from the 1991 census.

To guarantee the same precision, that design would have to be based on the sample sizes shown in the last column of Table 2 [9]. It should be noted that with the exception of *Lisboa e Vale do Tejo*, the sample sizes necessary to achieve the same variation coefficients would be substantially higher than the ones used in the proposed design, resulting in an impressive increase in the overall sample size from 267 sections to 743 sections.

Table 2: Sample sizes for three alternative sampling desig.

Stratum (NUTS II)	**Sample Size**		
	3 stages	**2 stages Selection of sections (with probability proportional to the number of dwellings)**	**2 stages Selection of sections (with equal probability)**
Norte	42	52	159
Centro	46	63	119
Lisboa e Vale do Tejo	110	87	97
Alentejo	33	55	93
Algarve	73	122	145
Açores	32	40	74
Madeira	31	38	56
Total	367	457	743

Another approach would be to select statistical sections, at the first sampling stage, with probability proportional to preliminary counts obtained from the questionnaire delivery phase in the 2001 census. With such an approach the selection of sections would be dependent on the conclusion of all the counts from the questionnaire delivery. This dependency is undesirable given the goal of reducing the period between the census date and the implementation of the post enumeration survey. Moreover, certain operational restrictions recommend that the geographical distribution of the sample should be known before the selection of the sample is possible.

Furthermore, this alternative design would not lead to a more precise estimation. Results from the simulation with data from the 1991 census (cf. Table 2), corresponding to a two stage sampling with selection of statistical sections with probability proportional to the preliminary counts from the questionnaire delivery at the first stage, show that the design adopted is generally more efficient than this alternative design[10]. In fact, it can be seen that the sample size necessary

[9]For this propose, in each stratum, the total estimator variance was approximated by the expression $V(\hat{\tau}_{k,h}) \approx \frac{M_h^2}{m_h(m_h-1)} \sum_{i \in U_h^I} \sum_{j \in U_{hi}^{II}} \left(\tau_{k,hij} - \frac{\tau_{k,h}}{M_h}\right)^2$, where M_h represents the number of sections in stratum h. This implies that for this benchmark, sampling error due to the last sampling stage is ignored.

[10]In the simulation the total estimator variance, in each stratum, was approximated by the expression $V(\hat{\tau}_{k,h}) \approx \frac{1}{m_h} \sum_{i \in U_h^I} \sum_{j \in U_{hi}^{II}} \frac{N_{hij}}{N_h} \left(\frac{N_h \tau_{k,hij}}{N_{hij}} - \tau_{k,h}\right)^2$.

to achieve the same variation coefficients would in general be greater than that obtained in the design adopted. It should be noted that this increase in sampling effort would be particularly significant in the *Algarve* because it would no longer be possible to use estimates for resident population in the definition of inclusion probabilities. In order to achieve the same precision this alternative design would cause an increase in the overall sample size from 267 sections to 457 sections[11].

Assuming that the cost of observing one statistical section is not significantly different among the three designs considered, it can be concluded that the survey cost (associated with the first sampling stage) would increase by 25% for the two-stage design with probability proportional to the number of dwellings and about 100% for the two stage design with equal probabilities. Only in *Lisboa e Vale do Tejo* would the two-stage designs lead to a reduction in sample size. This is due to the fact that the variance in *freguesia* totals as regards dwellings and population is higher in this stratum, while the same does not hold at section level. In addition, this was the region where dwelling estimates showed the poorest precision, affecting the quality of selection probabilities at the first stage of the three-stage design.

In fact, in the proposed design, the number of sections selected in each *freguesia* selected for the sample at the first stage is usually equal to one. Only in some *freguesias*, with inclusion probabilities equal to one, will more than one statistical section be selected for the sample in order to keep the unconditional selection probability at the second stage proportional to size. In this way it is possible to avoid a high concentration of sampled sections within a small number of *freguesias* and the typically associated design effect. Moreover it can be observed from Table 1 that the correlation between dwellings and resident population is significantly higher at *freguesia* level than at section level. This means that selection probabilities are more closely correlated with resident population totals at the first sampling stage than at the second, which can be considered as an indication of the superiority of the three-stage design.

Furthermore, a methodology for producing predictions for the resident population and dwellings at the time of the census was presented. This was achieved by combining demographic equations with information from other sources (data from other national surveys, data from the Portuguese electricity company on domestic consumption locations, data from the electoral rolls and data from the Immigration Service on the legalization of immigrants). These predictions were used as auxiliary information for defining inclusion probabilities for the primary sampling units. From a test carried out in 2000, it was concluded that their precision (a mean absolute relative error of 11% in estimating the resident population of *freguesias* in the test and of 8% in estimating the number of dwellings) was compatible with the aims of the survey design.

[11]The simulations were based on the assumption that the dwelling counts obtained from the questionnaire delivery phase of the census are free of error. In the (probable) situation where this assumption does not hold, the alternative two-stage design could erode precision even further.

APPENDIX

A. DERIVATION FOR THE APPROXIMATE VARIANCE OF $\hat{\tau}_{k,h}$

It should be noted that is obtained with a three-stage sampling design. Its approximate variance (Särndal, *et al.* 1992, pp. 148–149) can be written as

$$\begin{aligned} V(\hat{\tau}_k, h) \;\approx\; & \sum_{i\in U_h^I} \sum_{i'\in U_h^I} (\pi_{i,i'} - \pi_i \pi_{i'}) \, \frac{\tau_{k,hi}}{\pi_i} \, \frac{\tau_{k,hi'}}{\pi_{i'}} \\ & + \sum_{i\in U_h^I} \sum_{j\in U_{hi}^{II}} \sum_{j\in U_{hi}^{II}} \frac{\pi_{ij,ij'|i} - \pi_{ij|i}\,\pi_{ij'|i}}{\pi_i} \; \frac{\tau_{k,hij}}{\pi_{ij|i}} \; \frac{\tau_{k,hij'}}{\pi_{ij'|i}} \\ & + \sum_{i\in U_h^I} \sum_{j\in U_{hi}^{II}} \frac{V(\hat{\tau}_{k,hij})}{\pi_{ij}} \end{aligned}$$

where U^I and U^{II} represent respectively the population of primary units and secondary units.

Consider in each stratum h, two subpopulations: U_{1h}^I represents primary units of stratum h, such as $A_{hi} < I_h$, and U_{2h}^I the population formed by primary units in stratum h, where $A_{hi} \geq I_h$.

The sampling design is also such that $m_{hi} = 1$, $\forall\, i \in U_{1h}^I$ and $E(m_{hi}) = \frac{A_{hi}}{I_h}$, $\forall\, i \in U_{2h}^I$, with m_{hi} being the size of the sub-sample of sections corresponding to *freguesia* i of stratum h. Using an approximation through a sampling design with replacement we have

$$V(\hat{\tau}_{k,h}) \;\approx$$

$$\begin{aligned} \approx\; & \frac{1}{m_{1h}} \left[\sum_{i\in U_{1h}^I} \frac{A_{hi}}{A_{1h}} \left(\frac{A_{1h}\,\tau_{k,hi}}{A_{hi}} - \tau_{k,1h} \right)^2 + \sum_{i\in U_{1h}^I} \sum_{j\in U_{1hi}^{II}} \frac{A_{1h}\,N_{hij}}{A_{hi}\,N_{hi}} \left(\frac{N_{hi}\,\tau_{k,hij}}{N_{hij}} - \tau_{k,hi} \right)^2 \right] \\ & + \sum_{i\in U_{1h}^I} \sum_{j\in U_{1hi}^{II}} \frac{A_{1h}\,N_{hi}}{m_{1h}\,A_{hi}\,N_{hij}} V(\hat{\tau}_{k,hij}) + \frac{1}{m_h} \sum_{i\in U_{2h}^I} \sum_{j\in U_{2hi}^{II}} \frac{A_h\,N_{hij}}{A_{hi}\,N_{hi}} \left(\frac{N_{hi}\,\tau_{k,hij}}{N_{hij}} - \tau_{k,hi} \right)^2 \\ & + \sum_{i\in U_{2h}^I} \sum_{j\in U_{2hi}^{II}} \frac{A_h\,N_{hi}}{m_h\,A_{hi}\,N_{hij}} V(\hat{\tau}_{k,hij}) \end{aligned}$$

$$\begin{aligned} =\; & \frac{1}{m_h} \left[\sum_{i\in U_{1h}^I} \frac{A_h\,A_{hi}}{A_{1h}^2} \left(\frac{A_{1h}\,\tau_{k,hi}}{A_{hi}} - \tau_{k,1h} \right)^2 + \sum_{i\in U_h^I} \sum_{j\in U_{hi}^{II}} \frac{A_h\,N_{hij}}{A_{hi}\,N_{hi}} \left(\frac{N_{hi}\,\tau_{k,hij}}{N_{hij}} - \tau_{k,hi} \right)^2 \right] \\ & + \frac{N_h'^2}{n_h} \, \sigma_{k,h,intra}^2 \end{aligned}$$

where m_h is the size of the sample of sections corresponding to stratum h, $m_{1h} = m_h \dfrac{A_{1h}}{A_h}$, $V(\hat{\tau}_{k,hij}) = \dfrac{N'_{hij}\, m_h\, A_{hi}\, N_{hij}}{f_h\, A_h\, N_{hi}}\, \sigma^2_{k,hij}$, $\sigma^2_{k,h,intra} = \displaystyle\sum_{i \in U_h^I} \sum_{j \in U_{hi}^{II}} \frac{N'_{hij}}{N'_h}\, \sigma^2_{k,hij}$ is the intra-section variance in stratum h for variable k, and $\sigma^2_{k,hij}$ is the population variance of dwelling totals for variable k in section hij.

REFERENCES

[1] Australian Bureau of Statistics (1995). *Demographic Estimates and Projections: Concepts, Sources and Methods, Chapter 4 (Population Estimates: Data Sources – Census Undercount)*, Australian Bureau of Statistics, Australia.

[2] Australian Bureau of Statistics (1999). *Measuring Census Undercount in Australia and New Zealand (Working Paper No. 99/04)*, Australian Bureau of Statistics, Australia.

[3] Bureau of the Census (2000a). *Census 2000, A.C.E. Methodology*, Vol. 1, Bureau of the Census, USA.

[4] Bureau of the Census (2000b). *The Design of The Census 2000 Accuracy and Coverage Evaluation*, Bureau of the Census, USA.

[5] Casimiro, F. (1998). A Avaliação da Qualidade nos Recenseamentos da População e Habitação em Portugal, *Revista de Estatística*, **2**, 103–107.

[6] Coeffic, N. (1993). L'Enquête Post-Censitaire de 1990, Une mesure de l'exhaustivité du recensement, *Population*, **6**, 1665–1682.

[7] Coelho, P.S. (2001). Modelos de Estimação para os Censos 2001, *Actas do Seminário Censos 2001*, Instituto Nacional de Estatística, Lisbon.

[8] Dixie, J. (2000). *Planning for The 2001 Census Coverage Survey in England and Wales*, Office for National Statistics, UK.

[9] Gabinete dos Censos 2001 (2001). *Censos 2001 Programa Global*, Instituto Nacional de Estatística, Lisbon.

[10] Hansen, M.; Urvitz, W. and Madow, W. (1953). *Sample Survey Methods and Theory*, Vol. I and II, Wiley, New York.

[11] Instituto Nacional de Estatística (1991). *Censos 91 – Inquérito de Qualidade / Manual de Instruções*, Instituto Nacional de Estatística, Lisbon.

[12] Office for National Statistics (2000). *A One Number Census, Census Consultation Paper*, Government Statistical Service, UK.

[13] Office for National Statistics (2001). *Census 2001, A Guide to The One Number Census*, Government Statistical Service, UK.

[14] Office of Population Censuses and Surveys (1968). *Census 1961 General Report*, Office of Population Censuses and Surveys, Great-Britain.

[15] Office of Population Censuses and Surveys (1983). *Census 1971 General Report*, Office of Population Censuses and Surveys, Great-Britain.

[16] Office of Population Censuses and Surveys (1990). *Census 1981 General Report*, Office of Population Censuses and Surveys, Great-Britain.

[17] Särndal, C.E.; Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

[18] Sharp, J.; Kindelberger, J.C. and Bushery, J. (2000). *Measuring Response Variance in Census 2000*, Bureau of the Census, USA.

[19] Statistics Canada (1999). *1996 Census Technical Reports – Coverage (Catalogue No. 92-370-XIE)*, Ministry of Industry, Ottawa.

[20] United Nations (1999). *Principles and Recommendations for Population and Housing Censuses*, Revision 1, United Nations.