STATISTICAL GAP ALLOCATION IN AREA ESTIMATION - AN APPLICATION TO AN AGRICULTURAL SURVEY

REPARTIÇÃO DO GAP ESTATÍSTICO NA ESTIMAÇÃO EM DOMÍNIOS - UMA APLICAÇÃO AO INQUÉRITO À ESTRUTURA DAS EXPLORAÇÕES AGRÍCOLAS

Autor: Pedro Simões Coelho
- Professor Auxiliar no Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa

ABSTRACT:

- The *Inquérito à estrutura das explorações agrícolas* is a National Agricultural Survey conducted by the Portuguese Statistical Office (INE) and based on a random stratified sample. Inference is made using the Horvitz-Thompson estimator. Estimates of this survey are published at *Região Agrária* level, as it is considered the lowest level of aggregation for witch the estimates present adequate precision.

  This article results from a research project aiming to understand if inference can be made at a lower level of aggregation (as NUTIII), and to propose possible alternative estimators to Horvitz-Thompson.

  Several approximately unbiased estimators for area estimation are proposed. The proposed estimators are special forms of the regression estimator. They can combine sample data with auxiliary data arising from agricultural census or administrative sources. Approximate variances of the proposed estimators are deducted under the stratified sampling design and their estimates are produced using data from the 1993 survey. The aggregation of estimates is addressed and proposed transformations to the estimators that guaranty their consistency with the Horvitz-Thompson estimate at a higher level of aggregation. The proposed estimators are also compared in terms of precision and recommendation regarding their use for domain estimation proposes is made.

KEY-WORDS:

- *Area estimation, domain estimation, gap allocation, agricultural survey.*

RESUMO:

- O inquérito à estrutura das explorações agrícolas é um inquérito probabilístico longitudinal conduzido pelo Instituto Nacional de Estatística de Portugal (INE). A inferência tem sido realizada usando o estimador de Horvitz-Thompson. Os resultados deste inquérito são actualmente disponibilizados por Região Agrária, a qual corresponde ao menor nível de agregação para o qual as estimativas são consideradas fiáveis.

  O presente artigo resulta de um projecto de investigação que teve por principal objectivo compreender se é possível efectuar inferências a mais baixos níveis de agregação (como NUTIII) e propor estimadores alternativos ao de Horvitz-Thompson.

  São propostos diversos estimadores que podem ser considerados casos particulares do estimador pela regressão, apresentando em comum a propriedade de não enviesamento

aproximado. Estes combinem dados provenientes da amostra com dados auxiliares provenientes de recenseamentos agrícolas ou de fontes de natureza administrativa. São deduzidas as suas variâncias aproximadas e produzidas estimativas destas, com base nos dados do inquérito de 1993. É abordado o problema da aditividade das estimativas e propostas modificações aos estimadores que asseguram a sua consistência com as estimativas de Horvitz-Thompson a um mais elevado nível de agregação. Os estimadores propostos são comparados em termos de precisão, sendo efectuadas recomendações relativamente à sua aplicação no âmbito de estimação em domínios.

***PALAVRAS-CHAVE*:**

- *Domínio, área, estimação em domínios, gap allocation, agrícola.*

1. INTRODUCTION

The "Inquérito à estrutura das explorações agrícolas" is a National Agricultural Survey conducted by the Portuguese Statistical Office (INE) and based on a random stratified sample. Inference is made using the Horvitz-Thompson estimator. Estimates of this survey are published at "Região Agrária" level, as it is considered the lowest level of aggregation for witch the estimates present adequate precision.

The estimation of totals of subpopulations referring to lower levels of aggregation, as NUTIII, using the same sample, is of interest to INE. Nevertheless, the used sampling design and the random character of the sample sizes at this domain level result, in may cases, in a poor precision of the Horvitz-Thompson estimator.

This article results from a research project aiming to understand if inference can be made at a lower level of aggregation (as NUTIII), and to propose possible alternative estimators to Horvitz-Thompson. In particular we aim to combine sample data with auxiliary data from agricultural census.

Several approximately unbiased estimators are proposed. The proposed estimators are special forms of the regression estimator. Approximate variances of the proposed estimators are deducted under the stratified sample design and their estimates are produced using data from the 1993 survey. The aggregation of estimates is also addressed and proposed transformations to the estimators that guaranty their consistency with the Horvitz-Thompson estimate at a higher level of aggregation. The resulting estimators can then be seen as direct modified estimators since they use sample and auxiliary data exogenous to the domains of interest, yet maintaining approximately unbiased. The proposed estimators are also compared in terms of precision and recommendation regarding their use for domain estimation proposes is made.

2. DESCRIPTION OF THE SURVEY

The "Inquérito à estrutura das explorações agrícolas" survey is based on a random stratified sampling, using ***Região Agrária*** as the geographic level of stratification.

In each ***Região Agrária*** a new stratification is done, based on *SAU (Utilized Agricultural Area)* classes. In some ***Região Agrária*** other strata are defined using some variables considered badly correlated with SAU.

Tables 1 and 2 present the allocation of population and sample between the different Região Agrária and NUTIII.

Table 1. Population and sample by Região Agrária

Região Agrária	Population Size	Sample Size	Sample fraction
Entre Douro e Minho	111505	10545	0.09
Trás os Montes	80551	8569	0.11
Beira Litoral	125307	10556	0.08
Beira Interior	60386	7550	0.13
Ribatejo e Oeste	99938	9572	0.10
Alentejo	47049	7060	0.15
Algarve	26143	5022	0.19
R.A. Açores	24706	6020	0.24
R.A. Madeira	23157	6214	0.27

Table 2. Population and sample by NUTIII

NUT III	Population Size	Sample Size
MINHO-LIMA	28649	2455
CÁVADO	18039	1651
AVE	14540	1474
GRANDE PORTO	7963	1084
TÂMEGA	33413	3136
ENTRE DOURO E VOUGA	8901	745
DOURO	37694	3681
ALTO TRÁS-OS-MONTES	42857	4888
BAIXO VOUGA	26444	2089
BAIXO MONDEGO	28072	2233
PINHAL LITORAL	19416	1519
PINHAL INTERIOR NORTE	16143	1341
DÃO-LAFÕES	35232	3374
PINHAL INTERIOR SUL	11054	857
SERRA DA ESTRELA	5833	527
BEIRA INTERIOR NORTE	23263	3464
BEIRA INTERIOR SUL	11033	1562
COVA DA BEIRA	9203	1140
OESTE	39896	3658
GRANDE LISBOA	5862	626
PENÍNSULA DE SETÚBAL	9243	908
MÉDIO TEJO	22838	1747
LEZÍRIA DO TEJO	22099	2633
ALENTEJO LITORAL	8925	1380
ALTO ALENTEJO	12720	1645
ALENTEJO CENTRAL	12126	1758
BAIXO ALENTEJO	13278	2277
ALGARVE	26143	5022
REGIÃO AUTÓNOMA DOS AÇORES	24706	6020
REGIÃO AUTÓNOMA DA MADEIRA	23157	6214

3. METHODOLOGICAL ISSUES

3.1 INTRODUCTION

Several estimators for domain totals are proposed in the following sections. Their approximate variances are also obtained using 1st order approximations in Taylor series.

The focus was posed on obtaining approximately unbiased estimators. In fact, the sample sizes available at NUTSIII level seemed enough to avoid the use of biased synthetic or combined estimators.

In the following we use the definitions:

- **Direct estimators** : estimators that only use sample values of the interest variable in the domain and time period witch is being object of inference.
- **Direct modified estimators** : indirect estimators that can use sample data from outside the domain or time period of interest, which maintain certain design-based properties as being approximately unbiased.

3.2 ESTIMATORS

3.2.1 DIRECT ESTIMATORS

Three direct estimators that not use any auxiliary information (except for the population sizes) are considered and represent by the notation $\hat{t}_{d1}, \hat{t}_{d2}$ and $\hat{t}_{d3}$.

$\hat{t}_{d1}$ is the Horvitz-Thompson estimator, witch is presently used by INE to estimate at *região agrária* level. It has the usual form in the context of a stratification

$$\hat{t}_{d1} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_{hd}} y_i \qquad (1)$$

Its variance is given by

$$V(\hat{t}_{d1}) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} \left(S_{hd}^2 + \left(1 - \frac{N_{hd}}{N_h}\right) m_{hd}^2 \right) \qquad (2)$$

The variance presents a term involving the squares of the means of the interest variable at the subpopulations *hd*. This term may have an important contribution to the variance, specially in subpopulations of small size and where the variation coefficient of the interest variable is small.

An estimator of the variance may be obtained substituting the true population variance and mean S_{hd}^2 and m_{hd}, by their sample counterparts s_{hd}^2 and $\bar{y}_{hd}^2$.

$$\hat{V}(\hat{t}_{d1}) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h}\left(s_{hd}^2 + \left(1 - \frac{N_{hd}}{N_h}\right)\bar{y}_{hd}^2 \right). \quad (3)$$

$\hat{t}_{d2}$ is a pos-stratified estimator, for witch we assume that the population size at the domain level is known. It is given by

$$\hat{t}_{d2} = N_d \frac{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in S_{hd}} y_i}{\sum_{h=1}^{H} \frac{N_h}{n_h} n_{hd}} = N_d \frac{\hat{t}_{d1}}{\hat{N}_d} \quad (4)$$

The approximate variance of $\hat{t}_{d2}$ and its estimator is given, respectively, by the expressions

$$V(\hat{t}_{d2}) \approx V(\hat{t}_{d1} - \hat{N}_d m_d) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h}\left(S_{hd}^2 + \left(1 - \frac{N_{hd}}{N_h}\right)(m_{hd} - m_d)^2 \right) \quad (5)$$

and

$$\hat{V}(\hat{t}_{d2}) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h}\left(s_{hd}^2 + \left(1 - \frac{N_{hd}}{N_h}\right)\left(\bar{y}_{hd} - \frac{\hat{t}_{d2}}{N_d}\right)^2 \right). \quad (6)$$

$\hat{t}_{d2}$ will usually perform better then $\hat{t}_{d1}$, since the quantities $(m_{hd} - m_d)^2$ can be expected to be smaller then m_{hd}^2. In each stratum *h*, the variability of $\hat{t}_{d2}$ will only be bigger then the one associated to $\hat{t}_{d1}$ when $m_{hd} < \frac{1}{2} m_d$.

For the pos-stratified estimator $\hat{t}_{d3}$ it is assumed that population sizes at the level *hd* representing the intersection between domains and the strata are known. It has the form

$$\hat{t}_{d3} = \sum_{h=1}^{H} N_{hd} \frac{\hat{t}_{hd1}}{\hat{N}_{hd}} = \sum_{h=1}^{H} \frac{N_{hd}}{n_{hd}} \sum_{i \in s_{hd}} y_i \ . \quad (7)$$

The approximate variance of $\hat{t}_{d3}$ and its estimator are given by

$$V(\hat{t}_{d3}) \approx \sum_{h=1}^{H} V(\hat{t}_{hd1} - \hat{N}_{hd} m_{hd}) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} S_{hd}^2 \quad (8)$$

and

$$\hat{V}(\hat{t}_{d3}) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} s_{hd}^2 \,. \qquad (9)$$

With the used order of approximation in the Taylor series expansion, $\hat{t}_{d3}$ will always have lower variance than $\hat{t}_{d1}$ and $\hat{t}_{d2}$. This approximate variance is equal to the one that would be obtained for the Horvitz-Thompson estimator if we had selected from each subpopulation *hd* (*h*=1,…,*H*) a sample of controlled size $n_{hd}^0 = E(n_{hd}^s) = N_{hd} n_h / N_h$. However, $\hat{t}_{d3}$ is not defined unless all the sample sizes n_{hd} are strictly positive. Although asymptotically unbiased this estimator can also have a non-negligible bias as well as important contributes of order higher than one to the variance, when the subsample sizes n_{hd} are small. These facts are not of particular concern in the framework of this paper, since the expected samples sizes are always significantly bigger than zero.

3.2.2 DIRECT MODIFIED ESTIMATORS

3.2.2.1 ESTIMATION BASED ONLY ON SURVEY DATA

One of the main pitfalls associated with the estimators $\hat{t}_{d2}$ and $\hat{t}_{d3}$ is the lack of internal consistency of their estimates, i.e. the sum of total estimates for all the NUTIII in a certain *região agrária* is not necessarily equal to the estimate presently published at that higher level of aggregation. When one works with published data and specially with official statistics it is usually desired to have that kind of internal consistency for the used estimators.

In order to overcome this problem modifications were imposed to these two estimators, witch have resulted in the modified versions $\hat{t}_{d2\,\mathrm{mod}}$ and $\hat{t}_{d3\,\mathrm{mod}}$. These estimators, not only present that internal consistency, but also provide consistency of estimates of totals referring to different variables at any level of aggregation.

$\hat{t}_{d2\,\mathrm{mod}}$ results from the following modification to $\hat{t}_{d2}$

$$\begin{aligned} \hat{t}_{d2,\mathrm{mod}} &= \hat{t}_{d1} + (N_d - \hat{N}_d)\hat{m}_{RA} \\ &= N_d \hat{m}_{RA} + \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_{hd}} (y_i - \hat{m}_{RA}) \end{aligned} \qquad (10)$$

where $\hat{m}_{RA} = \sum_{h=1}^{H} \frac{N_h}{n_h N_{RA}} \sum_{i \in s_h} y_i$ is an estimator of population mean at *região agrária* level.

Its approximate variance is given by

$$V\left(\hat{t}_{d\,2,\mathrm{mod}}\right) \approx V\left(\hat{t}_{d\,1} - \hat{N}_d m_{RA}\right)$$
$$= \sum_{h=1}^{H} \frac{N_{hd}\left(N_h - n_h\right)}{n_h}\left(S_{hd}^2 + \left(1 - \frac{N_{hd}}{N_h}\right)\left(m_{hd} - m_{RA}\right)^2 \right) \qquad (11)$$

and a possible estimator by

$$\hat{V}\left(\hat{t}_{d\,2,\mathrm{mod}}\right) = \sum_{h=1}^{H} \frac{N_{hd}\left(N_h - n_h\right)}{n_h}\left(s_{hd}^2 + \left(1 - \frac{N_{hd}}{N_h}\right)\left(\bar{y}_{hd} - \hat{m}_{RA}\right)^2 \right). \qquad (12)$$

$\hat{t}_{d\,3\,\mathrm{mod}}$ results from the following modification to $\hat{t}_{d\,3}$

$$\hat{t}_{d\,3,\mathrm{mod}} = \hat{t}_{d\,1} + \sum\nolimits_h \left(N_{hd} - \hat{N}_{hd}\right)\hat{m}_h$$
$$= \sum_{h=1}^{H} N_{hd}\bar{y}_h + \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_{hd}} \left(y_i - \bar{y}_h\right) \qquad (13)$$

Its approximate variance is given by

$$V\left(\hat{t}_{d\,3,\mathrm{mod}}\right) \approx \sum\nolimits_h V\left(\hat{t}_{d\,1} - \hat{N}_{hd} m_h\right)$$
$$= \sum_{h=1}^{H} \frac{N_{hd}\left(N_h - n_h\right)}{n_h}\left(S_{hd}^2 + \left(1 - \frac{N_{hd}}{N_h}\right)\left(m_{hd} - m_h\right)^2 \right) \qquad (14)$$

and the respective estimator by

$$\hat{V}\left(\hat{t}_{d\,3,\mathrm{mod}}\right) = \sum_{h=1}^{H} \frac{N_{hd}\left(N_h - n_h\right)}{n_h}\left(s_{hd}^2 + \left(1 - \frac{N_{hd}}{N_h}\right)\left(\bar{y}_{hd} - \bar{y}_h\right)^2 \right). \qquad (15)$$

Both estimators show a variance with two terms. The first term equals the approximate variance of $\hat{t}_{d\,3}$. The second term of $\hat{t}_{d\,2,\mathrm{mod}}$ variance involves the squares of the differences between population mean at the subpopulation level *hd* and the population mean at *região agrária* level. This term will be small if the inter-domain variance in each *região agrária* is also small, i.e. if the mean of the interest variable is relatively constant from one NUTIII to another. The same comment applies to $\hat{t}_{d\,3,\mathrm{mod}}$, but now only is necessary the homogeneity of means to be verified in each stratum.

It can be expected that $\hat{t}_{d\,3,\mathrm{mod}}$ will usually be more precise than $\hat{t}_{d\,2,\mathrm{mod}}$, as we can, often, presume that the condition of homogeneity will more easily be verified in each stratum than in each *região agrária*. Nevertheless, in situations where the variance of $\hat{t}_{d\,3,\mathrm{mod}}$ doesn't show much improvement over the one associated with $\hat{t}_{d\,2,\mathrm{mod}}$ and the domains of interest are

small, then this later estimator may be preferred given it's bigger simplicity and the smaller risk of bias for small samples.

When we compare these modified estimators with their original versions $\hat{t}_{d2}$ and $\hat{t}_{d3}$, we can presume some loss of efficiency. In fact, the variance of $\hat{t}_{d2}$ involves the terms $(m_{hd} - m_d)^2$, $h = 1, \ldots, H$, while the variance of $\hat{t}_{d2,\mathrm{mod}}$ involves $(m_{hd} - m_{RA})^2$, $h = 1, \ldots, H$. We can again often expect a bigger homogeneity of the means in the domain rather than in the *região agrária*. Also, the approximate variance of $\hat{t}_{d3,\mathrm{mod}}$ will always be bigger than the one associated with $\hat{t}_{d3}$. The term $\sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)(N_h - N_{hd})}{N_h n_h}(m_{hd} - m_h)^2$ associated with $\hat{t}_{d3,\mathrm{mod}}$ variance represents a penalty that one pays in order to guarantee the internal consistency of the estimates. Nevertheless, these penalties may be small if the used stratification criterion is good.

The propriety of internal consistency can be easily verified since

$$\sum_{d \in RA} \hat{t}_{d2\,\mathrm{mod}} = \sum_{d \in RA} \hat{t}_{d3\,\mathrm{mod}} = \hat{t}_{RA} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_h} y_i \ .$$

Also the consistency of the estimates referring to different variables can be easily verified as

$$\hat{t}_{d2\,\mathrm{mod},y} + \hat{t}_{d2\,\mathrm{mod},x} = \hat{t}_{d2\,\mathrm{mod},(y+x)}$$

and

$$\hat{t}_{d3\mathrm{mod},y} + \hat{t}_{d3\mathrm{mod},x} = \hat{t}_{d3\mathrm{mod},(y+x)} \ .$$

3.2.2.2 ESTIMATION WITH AUXILIARY INFORMATION

It is proposed a direct modified ratio estimator

$$\begin{aligned} \hat{t}_{dQ} &= \hat{t}_{d1} + \sum_{h=1}^{H} (t_{xhd} - \hat{t}_{xhd}) \hat{R}_h \\ &= \sum_{h=1}^{H} \left(t_{xhd} \hat{R}_h + \frac{N_h}{n_h} \sum_{i \in s_{hd}} \left(y_i - \hat{R}_h x_i \right) \right) \end{aligned} \qquad (16)$$

where $\hat{R}_h = \dfrac{\sum_{i \in s_h} y_i}{\sum_{i \in s_h} x_i}$ is an estimator of $R_h = \dfrac{m_{yh}}{m_{xh}}$, $\boldsymbol{t}_{xhd}$ is the total of an auxiliary variable in the subpopulation *hd* and x_i is the value of an auxiliary variable for the agricultural establishment *i*.

In order to obtain estimates it is assumed that the values of the auxiliary variable are observed over the sample units, $x_i, i \in s_d$ and that the true population totals $\boldsymbol{t}_{xhd}, h = 1, \ldots, H$ are known.

Its approximate variance is given by:

$$\begin{aligned} V\left(\hat{\boldsymbol{t}}_{dQ}\right) &\approx \sum\nolimits_h V\left(\hat{\boldsymbol{t}}_{hd} - \hat{\boldsymbol{t}}_{xhd} R_h\right) \\ &= \sum_{h=1}^{H} \frac{N_h\left(N_h - n_h\right)}{n_h}\left(\frac{N_{hd}}{N_h}\left(S_{yhd}^2 + R_h^2 S_{xhd}^2 - 2R_h S_{x,yhd}\right)\right) \\ &+ \sum_{h=1}^{H} \frac{N_{hd}\left(N_h - n_h\right)}{n_h}\left(1 - \frac{N_{hd}}{N_h}\right)\left(m_{yhd} - R_h m_{xhd}\right)^2 \end{aligned} \tag{17}$$

We can easily see that this variance will be zero when, in each stratum, the variable of interest *y* is exactly proportional to the auxiliary variable *x*. When there is an approximate proportionality between these two variables one can expect to have smaller variance than the ones associated with the Horvitz-Thompson or the pos-stratified estimators.

A variance estimator is given by

$$\begin{aligned} \hat{V}\left(\hat{t}_{dQ}\right) &= \sum_{h=1}^{H} \frac{N_h\left(N_h - n_h\right)}{n_h}\left(\frac{N_{hd}}{N_h}\left(s_{yhd}^2 + \hat{R}_h^2 s_{xhd}^2 - 2\hat{R}_h s_{x,yhd}\right)\right) \\ &+ \sum_{h=1}^{H} \frac{N_{hd}\left(N_h - n_h\right)}{n_h}\left(1 - \frac{N_{hd}}{N_h}\right)\left(\overline{y}_{hd} - \hat{R}_h \overline{x}_{hd}\right)^2 \end{aligned} \tag{18}$$

The internal consistency of the estimator is also guaranteed, i.e. the sum of domain estimates in a *região agrária* equals an estimate of *região agrária* total produced by a direct ratio estimator defined at that level of aggregation.

We have then,

$$\sum_{d \in RA} \hat{\boldsymbol{t}}_{dQ} = \hat{\boldsymbol{t}}_{RAQ} = \sum_{h=1}^{H} \boldsymbol{t}_{xh} \frac{\hat{\boldsymbol{t}}_{yh}}{\hat{\boldsymbol{t}}_{xh}}. \tag{19}$$

Two regression estimators denoted by $\hat{\boldsymbol{t}}_{dR1}$ and $\hat{\boldsymbol{t}}_{dR2}$ are also proposed. For $\hat{\boldsymbol{t}}_{dR1}$, one assumes an approximate linear relation between *y* variable and the auxiliary variable *x*. It is now assumed a possible intercept in the regression model. This relation is supported by an independent model in each *região agrária* as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i \in RA \qquad (20)$$
$$E(\varepsilon_i) = 0 \quad E(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}$$

$\hat{t}_{dR1}$ is then given by

$$\begin{aligned} \hat{t}_{dR1} &= \hat{t}_{d1} + \left(\hat{\boldsymbol{\theta}}_{xd} - \hat{\boldsymbol{\theta}}_{xd}\right)' \hat{\boldsymbol{\alpha}}_{RA} \\ &= N_d \hat{\beta}_0 + t_{xd} \hat{\beta}_1 + \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in shd} \left(y_i - \hat{\beta}_0 - x_i \hat{\beta}_1\right) \end{aligned} \qquad (21)$$

where $\hat{\boldsymbol{\alpha}}_{RA} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \left[\sum_{h \in RA} \sum_{i \in s_h} \frac{N_h}{n_h} \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \sum_{h \in RA} \sum_{i \in s_h} \frac{N_h}{n_h} \mathbf{x}_i y_i$ e $\mathbf{x}_i = \begin{pmatrix} 1 & x_i \end{pmatrix}'$.

Its approximate variance is

$$\begin{aligned} V(\hat{t}_{dR1}) &\approx V\left(\hat{t}_{d1} - \hat{\boldsymbol{\theta}}_{xd}' \mathbf{B}\right) \\ &= \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} \left(S_{yhd}^2 + B_1^2 S_{xhd}^2 - 2B_1 S_{x,yhd}\right) \\ &+ \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} \left(1 - \frac{N_{hd}}{N_h}\right) \left(\mu_{yhd} - B_0 - B_1 \mu_{xhd}\right)^2 \end{aligned} \qquad (22)$$

where $\mathbf{B} = \begin{bmatrix} B_0 \\ B_1 \end{bmatrix} = \left[\sum_{h \in RA} \sum_{i \in U_h} \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \sum_{h \in RA} \sum_{i \in U_h} \mathbf{x}_i y_i$ are the OLS estimates of $\boldsymbol{\alpha}$ that we would obtain from an hypothetical fitting of model (20) to the subpopulation *RA*.

The variance of $\hat{t}_{dR1}$ will be zero if the postulated regression model holds exactly in the population. In practical terms, if the model holds approximately, $\hat{t}_{dR1}$ can have a small variance and represent a gain of efficiency relatively to the other estimators.

A possible estimator of this approximate variance is

$$\begin{aligned} \hat{V}(\hat{t}_{dR1}) &= \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} \left(s_{yhd}^2 + \hat{\beta}_1^2 s_{xhd}^2 - 2\hat{\beta}_1 s_{x,yhd}\right) \\ &+ \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} \left(1 - \frac{N_{hd}}{N_h}\right) \left(\bar{y}_{hd} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_{hd}\right)^2 \end{aligned} \qquad (23)$$

$\hat{t}_{dR2}$, is also a regression estimator based on the assumption of linear relation between y variable and the auxiliary variable x. For this estimator it is now considered a set of design variables representing the stratum to where each observation belongs.

For each região agrária the assumed relation results in a regression model

$$y_i = \beta_1 x_i + \sum_{h=1}^{H} \beta_{2h} E_{hi} + \varepsilon_i,\ i \in RA, \qquad (24)$$
$$E(\varepsilon_i) = 0 \quad E(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}$$

where E_{hi} is a dummy variable having the value 1 if the establishment i belongs to stratum h and the value 0 otherwise.

$\hat{\tau}_{dR2}$ is then given by

$$\hat{\tau}_{R2} = \tau_{xd} \hat{\beta}_1 + \sum_{h=1}^{H} N_{hd} \hat{\beta}_{2h} + \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in shd} \left(y_i - x_i \hat{\beta}_1 - \hat{\beta}_{2h} \right). \qquad (25)$$

Its approximate variance is now

$$V(\hat{\tau}_{dR2}) \approx \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} \left(S^2_{yhd} + B_1^2 S^2_{xhd} - 2B_1 S_{x,yhd} \right) + \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} \left(1 - \frac{N_{hd}}{N_h} \right) \left(\mu_{yhd} - B_{2h} - B_1 \mu_{xhd} \right)^2 \qquad (26)$$

and its estimator

$$\hat{V}(\hat{\tau}_{dR2}) = \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} \left(s^2_{yhd} + \hat{\beta}_1^2 s^2_{xhd} - 2\hat{\beta}_1 s_{x,yhd} \right) + \sum_{h=1}^{H} \frac{N_{hd}(N_h - n_h)}{n_h} \left(1 - \frac{N_{hd}}{N_h} \right) \left(\bar{y}_{hd} - \hat{\beta}_{2h} - \hat{\beta}_1 \bar{x}_{hd} \right)^2 \qquad (27)$$

The reasoning that supports this second regression estimator is to allow the intercept to change from one stratum to another. If the stratification criterion represents an important homogeneity factor, and the second term in the variance expression is of some importance this may contribute to obtain some gains in precision.

Once again the internal consistency of estimate is guaranteed. The sum of these domain estimates in a certain *região agrária* will equal the estimate produce by a direct regression estimator of the same type at that level of aggregation.

We have then,

$$\sum_{d \in RA} \hat{\tau}_{dR1} = \hat{\tau}_{RAR1} = N_{RA} \hat{\beta}_0 + \tau_{xRA} \hat{\beta}_1 + \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in sh} \left(y_i - \hat{\beta}_0 - x_i \hat{\beta}_1 \right) \qquad (28)$$

and

$$\sum_{d\in RA}\hat{t}_{dR2}=\hat{t}_{RAR2}=t_{xRA}\hat{\beta}_1+\sum_{h=1}^{H}N_h\hat{\beta}_{2h}+\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i\in sh}\left(y_i-x_i\hat{\beta}_1-\hat{\beta}_{2h}\right).$$

3.3 OTHER ISSUES ON STATISTICAL GAP ALLOCATION

The application of any of the estimators using auxiliary information results in a gap between the estimate produced at *região agrária* level, and the published Horvitz-Thompson estimate at that level. Even the proposed direct modified ratio and regression estimators present this characteristic, as its internal consistency is only achieved relatively to estimators of the same type (cf. equations 19 and 28).

Nevertheless, these gaps can be eliminated using a smoothing procedure. Having a set of estimators $\hat{t}_d$ where $\sum_{d\in RA}\hat{t}_d\neq\hat{t}_{RA,p}$ the gap at *região agrária* level can be allocated to the domains, obtaining new estimators

$$\tilde{t}_d=\hat{t}_d+\lambda_d\left(\hat{t}_{RA,p}-\sum_{d\in RA}\hat{t}_d\right) \tag{29}$$

where $\hat{t}_d$ is the allocation parameter, that should be defined in such a way that $\sum_{d\in RA}\lambda_d=1$.

The variance of this new estimator is given by

$$V(\tilde{t}_d)=V(\hat{t}_d)+\lambda_d^2V\left(\hat{t}_{RA,p}-\sum_{i\in RA}\hat{t}_i\right)+2\lambda_dCov\left[\hat{t}_d,\left(\hat{t}_{RA,p}-\sum_{i\in RA}\hat{t}_i\right)\right]. \tag{30}$$

If $\hat{t}_{RA,p}$ was a constant then the minimum of this polynomial would be obtained when $\lambda_d^*=Cov\left(\hat{t}_d,\sum_{i\in RA}\hat{t}_i\right)\Big/V\left(\sum_{i\in RA}\hat{t}_i\right)$. It follows immediately that the constraint $\sum_{d\in RA}\lambda_d=1$ is met. When $\hat{t}_{RA,p}$ is an estimator, as in the present context, the same principle can be applied to the choice of $\hat{t}_d$. Although λ_d^* will no longer be the value that minimizes $V(\tilde{t}_d)$, it may still be a good choice as it takes into account the covariance between each estimator $\hat{t}_d$ and the estimator of the total at the aggregated level $\sum_{i\in RA}\hat{t}_i$. With this choice the variance of $\tilde{t}_d$ tends to the its minimum as the variance of $\hat{t}_{RA,p}$ tends to zero.

The variance of $\tilde{t}_d$ is then given by

$$\begin{aligned}V(\tilde{t}_d)&=V(\hat{t}_d)+\left[\frac{Cov\left(\hat{t}_d,\sum_{i\in RA}\hat{t}_i\right)}{V\left(\sum_{i\in RA}\hat{t}_i\right)}\right]^2V\left(\hat{t}_{RA,p}-\sum_{i\in RA}\hat{t}_i\right)+2\frac{Cov\left(\hat{t}_d,\sum_{i\in RA}\hat{t}_i\right)}{V\left(\sum_{i\in RA}\hat{t}_i\right)}Cov\left[\hat{t}_d,\left(\hat{t}_{RA,p}-\sum_{i\in RA}\hat{t}_i\right)\right] \\ &=V(\hat{t}_d)+\left[\lambda_d^*\right]^2\left[V\left(\hat{t}_{RA,p}\right)+\left(2\frac{Cov\left(\hat{t}_{RA,p},\hat{t}_d\right)}{Cov\left(\hat{t}_d,\sum_{i\in RA}\hat{t}_i\right)}-1\right)V\left(\sum_{i\in RA}\hat{t}_i\right)-2Cov\left(\hat{t}_{RA,p},\sum_{i\in RA}\hat{t}_i\right)\right]\end{aligned} \tag{31}$$

Note that $\tilde{t}_d$ may have a higher variance than $\hat{t}_d$, mainly if $V(\hat{t}_{RA,p})$ is high and the covariance of the two estimators at the aggregated level, $Cov(\hat{t}_{RA,p}, \sum_{i \in RA} \hat{t}_i)$, is not very important. This is a plausible situation in practice since the estimators $\hat{t}_d$ are chosen with the purpose to be more accurate then the Horvitz-Thompson estimator. Nevertheless, the loss of efficiency may be small, mainly if the two levels of aggregation *d* and *RA* are very different. In fact, in that framework is plausible that $V(\hat{t}_{RA,p}) << V(\hat{t}_d)$ and l_d^* to be small, resulting in a small difference between $V(\tilde{t}_d)$ and $V(\hat{t}_d)$.

Applying illustratively this principle to the gap allocation when using the direct modified ratio estimator $\hat{t}_{dQ}$ we obtain a new estimator

$$\tilde{t}_{dQ} = \hat{t}_{dQ} + l_d^* (\hat{t}_{RA,p} - \hat{t}_{RA,Q}) \tag{32}$$

where $\hat{t}_{RA,Q} = t_{x,RA} \frac{\hat{t}_{y,RA,p}}{\hat{t}_{x,RA,p}}$ and $l_d^* = Cov(\hat{t}_{dQ}, \hat{t}_{RAQ}) / V(\hat{t}_{RAQ})$

$$\begin{aligned} V(\hat{t}_{RAQ}) &\approx \sum_h V(\hat{t}_h - \hat{t}_{xh} R_h) \\ &= \sum_{h=1}^{H} \frac{N_h (N_h - n_h)}{n_h} (S_{yh}^2 + R_h^2 S_{xh}^2 - 2 R_h S_{x,yh}) \end{aligned} \tag{33}$$

$$\begin{aligned} &Cov(\hat{t}_{dQ}, \hat{t}_{RAQ}) \approx \sum_h Cov[(\hat{t}_{yhd} - \hat{t}_{xhd} R_h), (\hat{t}_{yh} - \hat{t}_{xh} R_h)] \\ &= \sum_{h=1}^{H} \frac{N_h (N_h - n_h)}{n_h} \left(\frac{N_{hd}}{N_h} (S_{yhd}^2 + R_h^2 S_{xhd}^2 - 2 R_h S_{x,yhd}) \right) \\ &+ \sum_{h=1}^{H} \frac{N_{hd} (N_h - n_h)}{n_h} [m_{yhd} (m_{yhd} - m_{yh}) + R_h^2 m_{xhd} (m_{xhd} - m_{xh}) - R_h m_{xhd} (m_{yhd} - m_{yh}) - R_h m_{yhd} (m_{xhd} - m_{xh})] \end{aligned} \tag{34}$$

$$V(\hat{t}_{RAp}) = \sum_{h=1}^{H} \frac{N_h (N_h - n_h)}{n_h} S_{yh}^2 \tag{35}$$

$$\begin{aligned} Cov(\hat{t}_{dQ}, \hat{t}_{RAp}) &\approx \sum_h Cov[(\hat{t}_{yhd} - \hat{t}_{xhd} R_h), \hat{t}_{yh}] \\ &= \sum_{h=1}^{H} \frac{N_h (N_h - n_h)}{n_h} \left(\frac{N_{hd}}{N_h} (S_{yhd}^2 - R_h S_{x,yhd}) \right) \\ &+ \sum_{h=1}^{H} \frac{N_{hd} (N_h - n_h)}{n_h} [m_{yhd} (m_{yhd} - m_{yh}) - R_h m_{xhd} (m_{yhd} - m_{yh})] \end{aligned} \tag{36}$$

$$Cov\left(\hat{t}_{RAp}, \hat{t}_{RA,Q}\right) \approx \sum_{h=1}^{H} V\left(\hat{t}_{yh}\right) - R_h Cov\left(\hat{t}_{yh}, \hat{t}_{xh}\right)$$
$$= \sum_{h=1}^{H} \frac{N_h\left(N_h - n_h\right)}{n_h}\left(S_{yh}^2 - R_h S_{x,yhd}\right) \qquad (37)$$

4. APPLICATION TO THE 1993 SURVEY OF THE "INQUÉRITO ÀS ESTRUTURAS AGRÍCOLAS"

4.1 INTRODUCTION

The above proposed estimators were applied to data resulting from the 1993 survey of the "inquérito às estruturas agrícolas". For illustration purposes two variables were selected: *total of cereals* and *total of fresh fruits*. The first variable is known to be well correlated with the stratification criteria and particularly with SAU, while the second one presents a higher variability and smaller correlation with those criteria.[1]

When estimating at Região Agrária level, the simple average of the variation coefficients of Hortvitz-Thompson estimator is 2,6% regarding the variable *total of cereals* and 7,4% for the variable *total of fresh fruits*. These should be taken as reference values that allow a better assessment of the precision achieved when doing inference at NUTIII level.

Figures 1 and 2 present respectively the simple averages of design effects and variation coefficients for the analysed estimators over the several NUTIII. Detailed results can be found in annexe in tables A1 to A4.

The used notations are the same presented at section 3.2. Note that two ratio estimators denoted by $\hat{t}_{Q1}$ and $\hat{t}_{Q2}$ have been used. They only differ in the used auxiliary variable. While for $\hat{t}_{Q1}$ the auxiliary variable corresponds to the interest variable (*total of cereals* or *total of fresh fruits*) referenced to the year of the latest available agricultural census (1989), for $\hat{t}_{Q2}$ the auxiliary variable is SAU referenced to the same year. The advantage of $\hat{t}_{Q2}$ would be the possibility of using the same auxiliary variable to estimating totals of different interest variables. This can represent an important operational characteristic. Moreover, it would guarantee the consistency of total estimates for different variables at domain level.

In the regression estimators the used auxiliary variable is also *total of cereals* or *total of fresh fruits* referenced to the year of the census.

4.2 RESULTS AND CONCLUSIONS

As it could be expected, it is possible to observe that the regression estimators $\hat{t}_{dR1}$ and $\hat{t}_{dR2}$, as well as the ratio estimator $\hat{t}_{dQ1}$, are the ones that globally present better precision.

[1] Complete results of this study can be found in the report "Estimação em domínios no inquérito à estrutura da exploração agrícola. Relatório final".

Nevertheless, the gains of precision associated with the regression estimators, when compared to the ratio estimator $\hat{t}_{Q1}$ are usually small and some times inexistent. The ratio between the averages of estimated variances for $\hat{t}_{R1}$ and $\hat{t}_{Q1}$ is 0.97, for the variable *total of cereals*. For the variable *total of fresh fruits* it is even possible to observe that the best of the regression estimators presents estimated variances slightly higher than the ones associated with $\hat{t}_{Q1}$. In fact, one should remember that the only auxiliary information used in the regression estimators that is not accounted in the ratio estimator is the population sizes N_d and N_{hd}. The contribute of this information seems to be little expressive. Having also into account the bigger complexity of the regression estimators, the ratio estimator seems to be one of the most interesting ones for domain estimation at this level.

Also, the comparison between the two used regression estimators revels that the inclusion of a stratum varying intercept don't allow for any expressive precision improvement. The ratios between the averages of the variances of $\hat{t}_{dR2}$ and $\hat{t}_{R1}$ are only 0.96 for three variable total of cereals and 0.99 for *total of fresh fruits*. This later result seems then to confirm the weaker correlation between the *total of fresh fruits* and the stratification criteria.

The average design-effects of $\hat{t}_{Q1}$ regarding $\hat{t}_{\mathrm{d1}}$ over the several NUTSIII is 0.79 for the variable total of cereals and 0,62 to total of fresh fruits. These results represent very important improvements of the precision of the estimators for both variables. Nevertheless, the variances of $\hat{t}_{dQ2}$ don't allow these important reductions of variance. The ratios between the averages of the approximate variances of $\hat{t}_{dQ2}$ and $\hat{t}_{dQ1}$ over the several NUTSIII are 1,09 for the variable total of cereals and 1,6 for the variable total of fresh fruits. These results seem to show a week correlation between the two interest variables and SAU when compared to the ones observed with the auxiliary variables associated to $\hat{t}_{dQ1}$ and give a clear indication that SAU can not be used as the simple auxiliary variable for all the survey variables.

In spit of guaranteeing important internal consistency properties the sum of the estimates produced by $\hat{t}_{Q1}$ at NUTIII level is not equal to the estimate presently produced at *região agrária* level. In fact, the internal consistency of this estimator only guarantees that the sum of NUTIII estimates will be equal to the one produced by a direct ratio estimator at região agrária level using the same auxiliary variable (cf. eq. 19)[2].

This consistency with the Horvitz-Thompson estimate can nevertheless, be guaranteed by the pos-stratified estimators $\hat{t}_{d\,2\,\mathrm{mod}}$ and $\hat{t}_{d\,3\,\mathrm{mod}}$. Particularly $\hat{t}_{d\,3\,\mathrm{mod}}$, although presenting a worse performance than $\hat{t}_{Q1}$, still permit to achieve important gains of precision regarding the Horvitz-Thompson estimator. Its design-effect relatively to $\hat{t}_{\mathrm{d1}}$ is 0.85 for the variable total of cereals. For total of fresh fruits the gains of precision associated with $\hat{t}_{d\,3\,\mathrm{mod}}$ are negligible. This result could be anticipated and is associated with the weak correlation between that variable and the stratification criteria. While the ratio estimator seem to provide important gains in precision even for variable less correlated with the stratification criteria, $\hat{t}_{d\,3\,\mathrm{mod}}$ will only result in important gains for the ones well correlated with those criteria (one should remember that this estimator doesn't use any auxiliary information besides the population sizes N_{hd}).

[2] The analysis of transformations as the proposed in section 3.3 were out of the scope of this study.

It is also possible to conclude that the modifications introduced to the estimators $\hat{t}_{d2}$ and $\hat{t}_{d3}$, in order to obtain an internal consistency of estimates, produce some reduction of precision. Nevertheless, this degradation is usually very small; the ratio between the approximate variance of $\hat{t}_{d2\,\mathrm{mod}}$ and $\hat{t}_{d2}$ or between $\hat{t}_{d3\,\mathrm{mod}}$ and $\hat{t}_{d3}$ never exceeds the value 1.06.

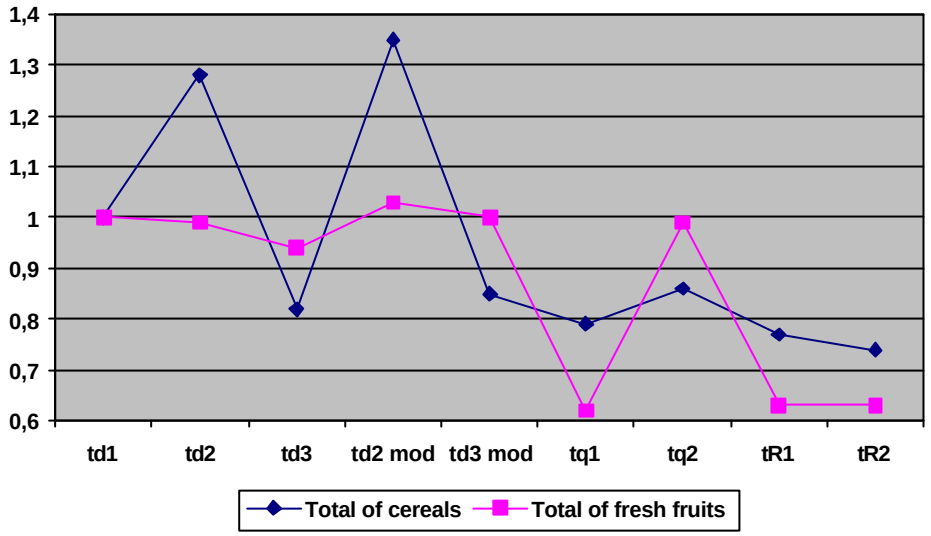

Figure 1 – Average design-effects

Choosing $\hat{t}_{dQ1}$ as a reference and observing its empirical variation coefficients when applied to the estimation of domain (NUTIII) totals for the two interest variables, one can conclude that the achieved precision seem to be enough to allow the use of its estimates.

The variation coefficients of this estimator vary, for the variable *total of cereals*, between a minimum of 1.6% and a maximum of 12.2%. The simple average of these coefficients over the several NUTIII is 4.5%. Even when considering the variable *total of fresh fruits* one can observe that the variation coefficients still maintain moderate (although bigger) values. They vary now from 3.0% to 31.0%, being their simple average equal to 13.4%.

To stress the quality of the produced estimates at NUTIII level, one should take into account that these variation coefficients are only 70% to 80% bigger than the ones presently obtained at *região agrária* level, for the same variables and smaller than the ones associated with some of the survey variables at that same level.

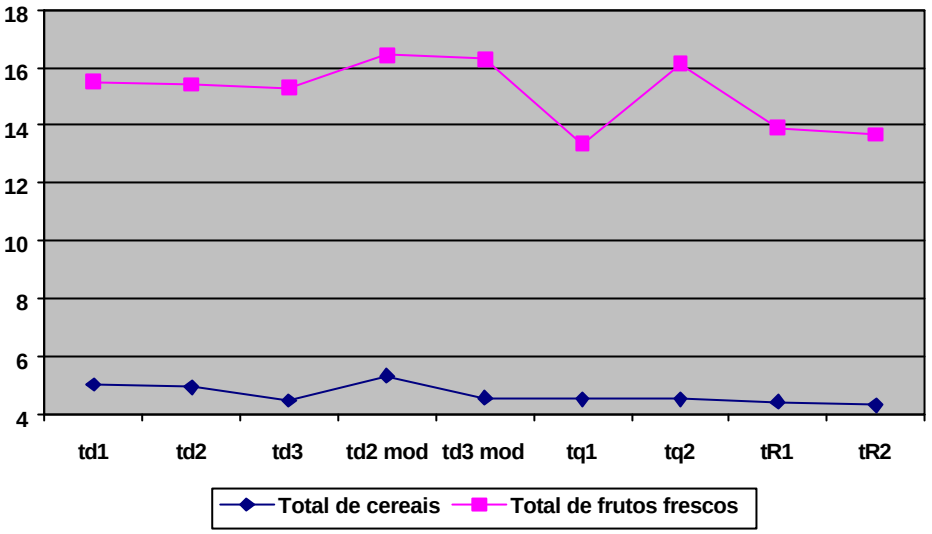

Figure 2 – Average of variation coefficients over NUTIII

5. ANNEXES

NUT III	t_{d1}	t_{d2}	t_{d3}	$t_{d2\ mod}$	$t_{d3\ mod}$	t_{q1}	t_{q2}	t_{R1}	t_{R2}
MINHO-LIMA	63632.08	50816.37	48685.30	50904.68	49150.86	49144.21	48936.26	47694.29	47610.11
CAVADO	51715.84	42580.89	40125.87	42516.49	40176.66	39202.57	39966.99	38148.39	37959.61
AVE	50996.61	42621.49	40108.46	42509.95	40159.13	39713.59	40077.67	38787.78	38536.98
GRANDE PORTO	26811.24	23832.31	21753.68	24799.73	22613.78	19866.33	22966.08	19691.70	19459.29
TAMEGA	60594.38	53065.48	49890.82	53383.36	50170.45	48804.68	49579.80	47774.61	47614.31
ENTRE DOURO E VOUGA	24756.04	19848.75	18154.28	21101.17	18363.05	20206.27	18121.58	18415.03	18079.35
DOURO	54136.66	51744.77	50593.22	63455.54	60208.80	49649.79	60093.55	46171.11	46695.35
ALTO TRAS-OS-MONTES	156699.47	155435.87	143730.20	150224.50	144289.52	137817.29	142376.20	135562.48	135177.30
BAIXO VOUGA	49947.18	45223.47	43456.65	46372.48	43705.77	42082.09	43376.62	41077.32	40915.95
BAIXO MONDEGO	72320.39	68950.28	59611.99	66698.41	61680.83	54861.38	61114.70	55107.55	54750.03
PINHAL LITORAL	33725.58	30575.63	29614.06	33887.66	30492.03	29582.14	30686.55	28399.29	28272.71
PINHAL INTERIOR NORTE	29720.81	24963.37	24161.47	26895.86	25402.34	25668.88	25398.42	24216.88	24234.31
DAO-LAFOES	64629.84	55519.38	53617.93	55013.60	54302.23	54855.27	53953.31	52981.69	53087.46
PINHAL INTERIOR SUL	18511.92	16677.69	15135.18	29155.15	15680.87	17833.08	15480.76	19561.45	14695.01
SERRA DA ESTRELA	31446.07	30179.21	29088.57	33811.05	29485.29	30654.22	29024.75	30344.84	29067.04
BEIRA INTERIOR NORTE	73371.38	74750.68	68682.36	72341.74	68742.98	70416.79	68670.28	67664.23	67095.76
BEIRA INTERIOR SUL	47044.64	51354.29	43981.43	48779.66	44357.78	38989.54	38972.59	40717.02	40399.29
COVA DA BEIRA	71914.42	72518.00	66016.86	70506.55	67120.07	67332.11	66659.39	67180.39	66349.81
OESTE	67267.52	66687.65	64041.09	69989.63	64423.26	66942.68	64396.77	62788.92	62209.21
GRANDE LISBOA	47899.07	48295.70	45806.73	47536.72	46627.26	42266.71	45538.82	42109.36	42354.00
PENINSULA DE SETUBAL	63499.72	64720.55	60172.21	65208.27	60962.64	62324.89	61542.94	62365.40	61474.17
MEDIO TEJO	38942.17	38926.27	36241.63	45546.52	37053.12	34262.36	37395.50	33061.15	32341.57
LEZIRIA DO TEJO	170251.18	177963.92	166054.02	170389.65	166640.02	158275.16	164483.35	164788.73	164048.16

ALENTEJO LITORAL	152835.63	174858.23	142796.54	250668.60	145617.49	151238.63	150130.66	143283.07	137712.30
ALTO ALENTEJO	125066.23	209159.52	116094.15	288423.74	122418.10	112680.40	147693.33	121876.58	108712.64
ALENTEJO CENTRAL	214238.93	296346.97	197136.01	312852.21	198169.04	197758.16	199734.28	196287.39	191393.81
BAIXO ALENTEJO	324652.48	493376.15	293911.01	371611.00	297131.51	270812.98	293410.69	274335.99	270930.52
ALGARVE	54008.89	54008.89	54008.89	54008.89	54008.89	52995.79	53599.39	51392.14	50994.09
R. AUTONOMA ACORES	5798.49	5798.49	5798.49	5798.49	5798.49	5186.66	5804.38	4960.72	4966.50
R. AUTONOMA MADEIRA	2783.84	2783.84	2783.84	2783.84	2783.84	2818.55	2774.79	2741.10	2745.80
DESVIO PADRÃO MÉDIO	74973.96	84786.14	67708.43	87239.17	68924.54	66474.77	69398.68	65982.89	64662.75
EFEITO DE SONDAGEM	1.00	1.28	0.82	1.35	0.85	0.79	0.86	0.77	0.74

*Table A1 –**Total of cereals:*** **Standard deviations by NUTIII**

NUT III	t_{d1}	t_{d2}	t_{d3}	$t_{d2\ mod}$	$t_{d3\ mod}$	t_{q1}	t_{q2}	t_{R1}	t_{R2}
MINHO-LIMA	10029.11	10030.46	9752.97	10081.84	9964.23	6286.17	9941.07	7306.22	7290.32
CAVADO	11439.20	11422.47	11348.49	11414.75	11366.31	8801.09	11332.92	8904.71	8902.21
AVE	5323.38	5302.66	5255.10	5306.20	5275.76	4568.98	5248.85	4510.94	4489.88
GRANDE PORTO	3120.78	3083.29	3062.30	3082.85	3138.64	3662.73	3145.19	2901.25	2908.73
TAMEGA	15191.34	15123.46	15058.93	15114.65	15086.79	13616.69	15074.09	12942.83	12936.49
ENTRE DOURO E VOUGA	3096.55	3085.60	3083.83	3130.56	3104.98	2991.37	3115.48	2954.88	2955.33
DOURO	54169.64	53469.29	52862.56	53535.13	53257.67	38179.95	53222.09	37388.43	37357.43
ALTO TRAS-OS-MONTES	24150.47	24051.42	23959.82	24236.43	24112.98	25324.45	24136.86	22072.63	22060.41
BAIXO VOUGA	3530.02	3507.44	3501.04	3858.71	3729.98	4296.88	3738.61	4024.58	4018.85
BAIXO MONDEGO	6516.68	6469.79	6435.20	6576.69	6524.36	5761.20	6555.50	5363.32	5358.73
PINHAL LITORAL	19359.68	19158.83	18804.13	19147.35	19090.57	14519.16	19058.32	14685.68	14677.30
PINHAL INTERIOR NORTE	7815.00	7809.10	7671.23	7848.33	7717.74	6045.44	7702.11	5913.33	5913.27
DAO-LAFOES	16720.42	16745.42	16521.95	16721.64	16572.24	11883.04	16415.01	11665.35	11665.48
PINHAL INTERIOR SUL	7070.29	7042.58	6620.23	8720.84	7177.59	7502.42	7236.64	6848.83	5809.78
SERRA DA ESTRELA	5965.57	5957.81	5872.00	7154.44	6159.83	3799.23	6146.81	6075.79	5460.85
BEIRA INTERIOR NORTE	55666.88	55643.00	55533.34	55695.68	55581.55	55287.67	55600.32	54818.82	54798.31
BEIRA INTERIOR SUL	10531.91	10506.63	10445.13	11387.55	10856.81	12709.83	11658.45	10691.53	10409.47
COVA DA BEIRA	42910.00	43024.10	40136.14	42106.54	41743.39	31280.58	41716.44	37607.60	37598.71
OESTE	109845.99	108317.21	102185.12	107225.96	104142.40	76465.66	103210.92	78416.76	78419.38
GRANDE LISBOA	7596.38	7436.92	7291.11	11337.06	12248.89	8452.32	12329.92	7850.40	7913.11
PENINSULA DE SETUBAL	22652.57	22428.86	21862.02	23855.65	23125.29	19413.48	23647.89	18520.64	18573.64
MEDIO TEJO	35496.06	34937.37	33424.16	36969.06	33743.30	30650.01	33542.15	30133.16	30001.08
LEZIRIA DO TEJO	27183.98	27218.49	26507.74	30327.14	29704.69	25397.25	30274.32	23713.43	23768.30
ALENTEJO LITORAL	3414.39	3393.27	3378.25	4653.17	4271.98	4850.48	4363.94	6501.60	6476.36
ALTO ALENTEJO	66950.36	67087.29	66312.04	66918.71	66635.85	33827.86	66626.79	39162.86	39152.25
ALENTEJO CENTRAL	14154.35	14092.30	14008.09	14211.07	14224.59	9777.42	14227.27	9436.52	9456.71
BAIXO ALENTEJO	27331.52	27430.88	27158.51	27428.03	27248.04	18897.87	27260.92	19328.93	19332.43
ALGARVE	16389.02	16389.02	16389.02	16389.02	16389.02	14797.89	15188.07	13939.14	13941.41

R. AUTONOMA ACORES	2254.16	2254.16	2254.16	2254.16	2254.16	1633.99	2271.37	1712.03	1713.41
R. AUTONOMA MADEIRA	2655.18	2655.18	2655.18	2655.18	2655.18	2653.96	2603.48	2538.88	2536.12
DESVIO PADRÃO MÉDIO	**21284.36**	**21169.14**	**20644.99**	**21644.81**	**21236.83**	**16777.84**	**21219.73**	**16931.04**	**16863.19**
EFEITO DE SONDAGEM	**1.00**	**0.99**	**0.94**	**1.03**	**1.00**	**0.62**	**0.99**	**0.63**	**0.63**

Table A2 **–*Total of fresh fruits:* Standard deviations by NUTIII**

NUT III	t_{d1}	t_{d2}	t_{d3}	$t_{d2\ mod}$	$t_{d3\ mod}$	t_{q1}	t_{q2}	t_{R1}	t_{R2}
MINHO-LIMA	2.88	2.31	2.21	2.31	2.23	2.27	2.22	2.19	2.18
CAVADO	3.90	3.14	2.93	3.14	2.94	2.91	2.92	2.83	2.81
AVE	4.15	3.47	3.28	3.46	3.28	3.21	3.26	3.15	3.14
GRANDE PORTO	7.18	6.32	5.68	6.55	5.86	5.64	5.98	5.44	5.33
TAMEGA	2.86	2.52	2.38	2.54	2.40	2.33	2.38	2.27	2.27
ENTRE DOURO E VOUGA	5.79	4.69	4.32	5.01	4.38	5.02	4.35	4.48	4.39
DOURO	4.34	4.11	4.01	4.92	4.66	3.98	4.60	3.67	3.69
ALTO TRAS-OS-MONTES	1.83	1.83	1.69	1.76	1.69	1.61	1.67	1.58	1.58
BAIXO VOUGA	4.23	3.84	3.69	3.94	3.73	3.50	3.69	3.43	3.43
BAIXO MONDEGO	3.28	3.10	2.68	3.00	2.78	2.42	2.75	2.44	2.43
PINHAL LITORAL	5.34	4.85	4.68	5.38	4.81	4.54	4.82	4.40	4.38
PINHAL INTERIOR NORTE	4.64	3.94	3.83	4.27	4.08	4.05	4.10	3.83	3.84
DAO-LAFOES	2.64	2.27	2.20	2.25	2.22	2.27	2.21	2.19	2.19
PINHAL INTERIOR SUL	5.83	5.39	4.86	10.04	5.09	6.03	5.07	6.66	4.84
SERRA DA ESTRELA	13.03	12.02	11.16	12.86	11.44	12.20	11.02	11.74	11.32
BEIRA INTERIOR NORTE	2.70	2.74	2.52	2.66	2.52	2.57	2.52	2.48	2.46
BEIRA INTERIOR SUL	3.80	4.11	3.53	3.91	3.55	3.10	2.88	3.25	3.23
COVA DA BEIRA	6.16	6.30	5.92	6.10	5.94	5.85	5.84	5.83	5.82
OESTE	4.49	4.48	4.29	4.71	4.32	4.28	4.34	4.14	4.10
GRANDE LISBOA	10.94	10.90	10.18	10.76	10.28	9.24	10.18	9.24	9.19
PENINSULA DE SETUBAL	12.71	12.74	11.67	12.82	11.83	12.12	10.34	11.80	11.60
MEDIO TEJO	6.30	6.39	6.07	7.62	6.33	6.24	6.43	5.90	5.83
LEZIRIA DO TEJO	5.56	5.73	5.40	5.53	5.43	5.30	5.54	5.40	5.38
ALENTEJO LITORAL	5.20	5.98	5.07	8.60	5.29	5.53	5.43	5.11	4.98
ALTO ALENTEJO	2.42	3.96	1.98	5.41	2.07	2.17	2.88	2.34	2.01
ALENTEJO CENTRAL	2.90	4.23	2.61	4.49	2.62	2.67	2.70	2.68	2.57
BAIXO ALENTEJO	2.27	3.34	2.01	2.55	2.05	1.89	2.04	1.90	1.88
ALGARVE	2.88	2.88	2.88	2.88	2.88	2.80	2.88	2.71	2.69
R. AUTONOMA ACORES	2.40	2.40	2.40	2.40	2.40	2.11	2.40	2.03	2.04
R. AUTONOMA MADEIRA	7.37	7.37	7.37	7.37	7.37	7.28	7.36	7.24	7.25
COEFICIENTE DE VARIA-ÇÃO MÉDIO	**5.00**	**4.91**	**4.45**	**5.31**	**4.55**	**4.50**	**4.49**	**4.41**	**4.29**

Table A3 **–*Total of cereals*: coefficients of variation by NUTIII**

NUT III	t_{d1}	t_{d2}	t_{d3}	$t_{d2\ mod}$	$t_{d3\ mod}$	t_{q1}	t_{q2}	t_{R1}	t_{R2}
MINHO-LIMA	36.83	36.95	39.42	37.34	37.58	30.92	37.10	33.52	33.67
CAVADO	21.36	20.87	20.53	20.94	20.65	15.95	20.91	16.34	16.35
AVE	15.62	15.61	15.24	15.62	15.29	10.62	14.96	11.30	11.21
GRANDE PORTO	15.79	15.44	15.40	15.44	15.65	18.27	14.97	14.07	14.09
TAMEGA	11.34	11.37	11.40	11.33	11.35	10.45	11.40	9.95	9.94
ENTRE DOURO E VOUGA	38.79	39.07	39.49	40.40	41.48	27.36	42.21	29.42	29.57
DOURO	7.73	7.56	7.43	7.59	7.56	5.55	7.54	5.33	5.33
ALTO TRAS-OS-MONTES	5.85	5.88	5.84	5.94	5.90	6.23	5.93	5.42	5.41
BAIXO VOUGA	15.57	15.49	15.46	17.14	17.30	17.80	17.06	17.10	17.00
BAIXO MONDEGO	12.06	11.85	11.75	11.95	11.88	9.62	11.96	9.12	9.14
PINHAL LITORAL	10.98	10.88	10.53	10.87	10.80	7.24	10.78	7.59	7.57
PINHAL INTERIOR NORTE	17.94	18.14	17.86	18.30	18.22	14.85	17.92	15.04	15.12
DAO-LAFOES	11.70	11.72	11.33	11.71	11.51	8.78	11.48	8.47	8.47
PINHAL INTERIOR SUL	21.24	21.71	20.88	30.78	22.74	24.37	23.32	23.71	18.53
SERRA DA ESTRELA	32.57	31.24	29.74	32.16	27.59	13.52	26.18	26.26	23.30
BEIRA INTERIOR NORTE	16.47	16.41	16.51	16.42	16.55	16.74	16.75	16.24	16.35
BEIRA INTERIOR SUL	17.28	17.08	16.98	18.15	17.19	23.36	13.77	17.46	16.88
COVA DA BEIRA	7.35	7.46	7.34	7.23	7.35	5.54	7.21	6.49	6.60
OESTE	4.08	4.05	3.83	4.00	3.89	2.97	3.86	3.00	3.00
GRANDE LISBOA	12.49	12.09	12.01	17.85	18.36	10.47	17.89	10.56	10.63
PENINSULA DE SETUBAL	11.12	10.83	10.47	11.39	10.81	10.71	10.82	9.79	9.75
MEDIO TEJO	7.37	7.36	7.19	7.89	7.35	5.99	7.22	6.04	6.03
LEZIRIA DO TEJO	6.79	6.69	6.54	7.34	7.10	6.20	7.36	5.79	5.78
ALENTEJO LITORAL	21.03	20.98	21.00	29.27	40.37	31.72	40.53	44.04	46.13
ALTO ALENTEJO	29.74	29.24	29.59	29.42	29.26	19.33	29.97	19.68	19.52
ALENTEJO CENTRAL	14.28	14.98	14.52	15.35	14.35	10.72	14.55	10.44	10.40
BAIXO ALENTEJO	21.19	20.55	19.99	20.55	19.81	16.41	19.86	16.61	16.44
ALGARVE	3.62	3.62	3.62	3.62	3.62	3.44	3.38	3.16	3.16
R. AUTONOMA ACORES	8.14	8.14	8.14	8.14	8.14	6.07	8.18	6.44	6.45
R. AUTONOMA MADEIRA	9.31	9.31	9.31	9.31	9.31	9.56	9.19	8.96	8.95
COEFICIENTE DE VARIAÇÃO MÉDIO	**15.52**	**15.42**	**15.31**	**16.45**	**16.30**	**13.36**	**16.14**	**13.91**	**13.69**

Table A4 ***–Total of fresh fruits*: coefficients of variation by NUTIII**

REFERENCES

BATTESE, G.E., e FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.

BATTESE, G.E., HARTER, R.M., e FULLER, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

COELHO, P. (1996). Estimação em pequenos domínios. *Working Paper*, Lisboa, ISEGI/UNL.

COELHO, P. (1996). Estimadores combinados para pequenos domínios. *Revista de Estatística*, 2, 23-43.

COELHO, P. (2000). Estimação em domínios no inquérito à estrutura das explorações agrícolas. *Methodologica*. To appear.

COELHO, P. e GUERRA, HELENA (1999). Estimação em domínios no inquérito à estrutura da exploração agrícola: Relatório final.

COELHO, P. e GUERRA, HELENA (1998). Estimação em domínios no inquérito à estrutura da exploração agrícola: Segundo relatório intercalar.

CRONKHITE, F. (1987). Use of regression techniques for developing state and area employment and unemplyment estimates. In *Small Area statistics*. (Eds. R. Platek et al.). New York: Wiley.

DURBIN, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling* (Eds. N. Johnson e H. Smith), Wiley-Interscience.

FAY, R.E. (1987). Application of multivariate regression to small domains estimation. In *Small Area statistics*. (Eds. R. Platek et al.). New York: Wiley.

LITTLE, R. J. A. (1983). Estimating a finite population meanfrom unequal prabability samples. *Journal of the American Statistical Association*, 78, 596-604.

PFEFFERMANN, D. e BARNARD, C. (1991). Some new estimators for small-area means with aplication to the assesment of farmland values. *Journal of Business and Economics Statistics*, 9, 73-84.

PFEFFERMANN, D. e BURCK, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.

RAO, J.N.K. (1985). Conditional Inference in Survey Sampling. *Survey Methodology*, 11, 15-31.

SÄRNDAL, C.E. (1980). On p inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SÄRNDAL, C.E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.

SÄRNDAL, C.E. (1984). Design-consistent versus model-dependent estimators for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SÄRNDAL, C.E., e HIDIROGLOU, M.A. (1989). Small Domain Estimation: a conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

SÄRNDAL, C.E., SWENSSON, B., e WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, M.P., GAMBINO, J., e MANTEL, H.J. (1994). Issues and Strategies for Small Area Data. *Survey Methodology*, 20, 3-22.

SINGH, M.P., e TESSIER, R. (1976). Some estimators for domain totals. *Journal of the American Statistical Association*, 71, 322-325.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.