EURAREA Proyect
(Enhancing Small Area Estimation Techniques to meet European Needs)

December 2005

1 Foreword

The EURAREA project (Enhancing Small Area Estimation Techniques to meet European needs) has been carried out within the Fifth Framework Programme of R&D developed by the European Union from 2000 - 2004. Spain has taken part in this programme together with 6 other European countries (United Kingdom, Italy, Sweden, Norway, Finland and Poland).

The aim of this project is to evaluate the effectiveness of standard estimation techniques for small areas (synthetic estimators, GREGs and composite estimators). The studies carried out up until now were based on sampling designs with equal selection probabilities. In order to undertake this project, it was necessary to study the existing theory as well as to develop new theories that make it easier to obtain estimation techniques and their mean squared error when other sampling plans are used that are more similar to those applied in official statistics in the real world. Finally, all the theory developed has been implemented in a SAS IT application whose use has been widely documented so that any user is able to apply the programme to their own data.

The project focuses the research mainly into four topics:

1) The use of ancillary information from the past.
2) The use of ancillary information from other geographical areas.
3) The adaptation of standard estimators to complex sampling designs, in other words, with the use of unequal probabilities and in particular, with the selection of conglomerates.
4) The obtaining of estimations for cross-classifications.

Since the beginning of 2001, the National Statistics Institute (INE) has worked together with the Miguel Hernández de Elche University (UMH) on the development of topic 3. This document therefore centres on a description of the work undertaken from this perspective and, in particular, with the aim of studying the impact of sampling weights on small area estimators.

This document has 13 sections and corresponding subsections in some cases. Section 2 describes the process followed to create an artificial population on which different simulation experiments could be carried out that allow us to evaluate the small area estimators. Section 3 provides the concepts applied to small area. Section 4 illustrates the complementary sources to the created artificial population that provide ancillary information broken down at area level. Sections 5, 6 and 7 define the population parameters under study, the sampling designs applied to estimate them and the standard estimators used. Section 8 provides the evaluation measures calculated in the simulation experiments carried out to value the estimators. Section 9 offers the new theory developed to calculate the EBLUP estimators (Empirical Best Linear Unbiased Predictor),

which are assisted by mixed linear models with a random area factor when the model parameters are estimated using individual weights and the area weights. All of this theory has been subsequently implemented in SAS/IML, as described in section 10.

In 2002, the simulation experiments started in order to test the standard estimators and later on, in 2003, the tests continued with the application of more complex designs and also new estimators based on the new theory. It is worth highlighting the following experiments:

- Calculation of the standard estimators, applying the APS and LCS type designs described in section 6 for the estimation of three population parameters.
- Calculation of standard estimators and of new estimators for the ILO unemployment estimation by applying a modified design of the APS type, which provides a non-self-weighted sample on a strata level that is beneficial for the study on the impact of the use of sampling weights.
- Calculation of standard estimators for the estimation of income by applying the LCS type design and a collection of covariables that are different from those used with the standard estimators and that include taxable income from income tax, which is very useful when analysing the effect of this covariable on the estimations obtained.
- Calculation of standard estimators for the estimation of income by applying the LCS type design, but incorporating a non- response mechanism correlated with household income. By doing this, we were able to study the impact of the use of informative weights.

All of the experiments carried out and the results obtained, both in 2002 and 2003, are described in sections 11 and 12 respectively. Finally, section 13 provides references for readers that have queries.

Once the EURAREA project had finished and following common practice in the Fifth Framework Programme of European projects, the results were widely disseminated with special attention paid to the website.

Both the final volume, which includes the results produced by the INE summarised in this document, and the associated theory and software are available on the project's official website (http://www.statistics.gov/eurarea).

It is also worth mentioning the project's last objective, which was to create a discussion forum. This was achieved in August of this year and the first international conference on Small Area Estimates (SAE 2005) was held in Jyväskylä (Finland) from the 28-31 August.

International figures from the field took part in this conference (D. Pfeffermann, J.N.K. Rao, C. Särndal, ...). Alongside these speakers there were other presentations and informative displays given by approximately 100 participants

from the official statistics field and from universities. In particular, together with the Miguel Hernández de Elche University, the INE presented a paper on the application of these estimation techniques to the APS (see the document *Small Area Estimation in the Spanish Labour Force Survey*).

The work presented includes the results obtained in the last few years in the field of Small Area Estimation and is linked to:

- research (theoretical and methodological development)
- production (application to the real world)

On the official conference website, http://www.stat.jyu.fi/sae2005/, it is possible to find more information and in particular, the Proceedings and abstracts of the papers presented. All of these presentations, or at least the vast majority of them, are currently being revised for publishing in the Polish magazine *Statistics in Transition.*

2 Creation of the artificial population (APES)

One of the EURAREA project's activities in its first stage, which forms part of the actions developed for the study of any of the four topics mentioned in the foreword, is the creation of a data file that contains unemployment, household income and composition as objective variables, together with a wide range of social and demographic variables used as ancillary variables.

This database, described below, was compiled during the project's first year and has been widely documented. It contains Spain's artificial population in the project named APES (Artificial Population EURAREA-Spain). All of the variables in the file are named APES+no.

The APES file contains 40 variables and 38,872,268 entries. An entry is able to take up to 90 characters. For each entry, the first 35 variables come from the 1991 Population and Dwellings Census, in other words, the entry unit is the person resident in a main family dwelling in Spain on the census reference data (1 March 1991). The household to which the person belongs can also be identified using a common identification number for all members of the same household.

In each entry, 5 new variables have been generated: 2 imputed using the information contained in ancillary files and 3 obtained from the transformation of previous variables, which it is not strictly necessary to include.

The imputed variables are:

- *Registration at the public unemployment office* (APES501) according to the Active Population Survey (APS). This variable was obviously not present in the original census register, but it is necessary in APES as an explanatory variable in all models in order to estimate ILO unemployment with simulated samples (objective variables, present as an 'actual' variable in APES). The person requesting work is the person who is registered at the National Unemployment Office belonging to the Ministry for Work (National Institute for Employment, INEM) in order to request work. The people interviewed in the APS are asked whether they are requesting work, meaning that this variable is entered in APES using the information collected from the APS in the second quarter of 1991.

- *Total annual net income for the household* (APES502), which is obtained by imputation using the Household Budget Survey (HBS) 1990-91. This variable, which in this case is an objective EURAREA variable, but which is not available in Spanish population censuses, is defined in the HBS as the total net income as a result of the household's annual monetary income in the year prior to the interview. Capital and property income have been excluded from EURAREA applications, as these components are not suitable for simulations. The non-monetary components (such as imputed rent of owned dwellings, self-consumption and self-supply) have also been excluded given the lack of international comparability.

The APS file contains 199,231 entries (individuals) with 23 APES variables and the HBS file contains 21,155 entries (households) with 21 APES variables. Some of these variables are discreet and others are continuous, meaning that the general regression models that allow us to predict their value have been fitted for the imputation of the APES501 and APES502 variables. In the terminology of linear models, discreet variables are called *factors* and continuous variables are called *covariables*. For factors with *a* levels, *a-1* parameters are estimated (the parameter for the last level is zero). However, only one parameter is estimated for the covariables.

The sampling designs of both surveys select independent samples in the different Autonomous Communities. As a result, two possibilities can be considered for estimating the models: fitting a single model to the national sample or fitting 18 models, one to each regional sample. In this case, an in-depth understanding of the Spanish economy and society has been decisive in the preference for the region to region estimation. The sample size in the HBS however is smaller than for the APS and therefore the solution of estimating different regressions in each region in order to predict the APES502 variable will not always be possible due to the existing disparity between the number of parameters to be estimated in the regional model and the sample size in the Autonomous Community. To do this, single province regions have not received individual treatment, rather they have been added to another similar region, with the exception of the Autonomous Community of Baleares, given its status as an archipelago.

With the aim of predicting the value of APES501 for all individuals in the APES artificial population, a study was undertaken to select the best functioning linear regression model with a binary response variable and the result was the application of logistic regression models (logit). A similar piece of research was undertaken in order to predict the APES502 values and the final decision was to use the log-normal type model.

Having estimated the models selected with the least square method, a fitting level indicator was obtained for each model. The following percentage was obtained for the logistic models used to impute APES501:

$$Q = \left(1 - \frac{deviationof\ \mathrm{mod}\,el}{zerodeviation} \right) 100$$

where the numerator and the denominator are the deviation of the chosen model and the deviation of the zero model (one that contains a single

parameter) to the saturated model (one with as many parameters to be estimated as observations).

For the log-normal models fitted to predict the APES502 variable, the determination coefficient has been used as the indicator R^2 from the model defined by the coefficient:

$$R^2 = \frac{VE}{VT}$$

where the numerator and the denominator represent the variability of the observations explained by the model and the total variability respectively.

Tables 2.1 and 2.2 presented below show the ancillary variables used in the models together with the indicator value of the fitting quality. The number of observations used in the model fitting is given in the column called ***n*** whereas the number of parameters estimated is given in the last column.

Table 2.1. Factors and covariables from the 1991 APS file that appear in the APES501 logit models

			Factors														Covariables			
Autonomous Community	**n**	**(%)**	103	104	202	206	207	208	210	211	301	303	304	306	307	403	203	405	409	**parameters**
Andalucía, Ceuta y Melilla	23.016	49.88	X	X	X		X	X	X	X		X	X	X	X		X		X	88
Aragón	5.508	57.20	X		X		X	X	X	X		X	X				X		X	50
Asturias y Cantabria	6.735	63.23	X	X	X	X	X	X		X					X	X	X		X	76
Baleares	2.466	53.35		X			X	X	X	X		X				X	X		X	57
Canarias	5.912	50.26	X		X		X	X	X				X				X			21
Castilla-León	12.595	61.99	X		X	X	X	X	X	X							X			52
Castilla-La Mancha y Murcia	12.063	58.40	X	X	X	X	X	X	X				X				X			40
Cataluña	13.306	72.16	X	X	X	X	X	X		X			X				X			53
Valencia	10.783	56.57	X		X		X	X	X	X	X		X	X			X		X	47
Extremadura	4.994	46.96	X	X	X	X	X	X	X	X							X	X		52
Galicia	8.591	60.10	X	X	X		X	X	X	X	X						X			47
Madrid	6.284	77.17				X	X	X								X	X			31
Navarra y La Rioja	4.722	53.28	X		X		X	X		X			X				X			35
País Vasco	7.349	60.43	X		X	X	X	X	X	X	X						X			47

FACTORS				**COVARIABLES**	
APES 103	Province	APES 211	Socio-economic situation	APES 203	Age
APES 104	Strata	APES 301	Sex of reference person (RP)	APES405	Number of employed people
APES 202	Sex	APES 303	Highest level of education finished by RP	APES 409	Size of household
APES 206	Relationship with the referente person	APES 304	RP's relation with activity		
APES 207	Highest level of education finished	APES 306	RP's professional sit.		
APES 208	Relation with activity	APES 307	RP's socio-economic situation		
APES 210	Professional situation	APES 403	Type of household		

Table 2.2 Factors and covariables from the 1991 HBS file that appear in the APES502 log-normal models

			Factors														Covariables			
Autonomous Community	**n**	**R^2**	103	104	301	303	304	306	403	404	405	406	407	408	410	411	413	302	409	412	**parameters**
Andalucía, Ceuta y Melilla	3.895	0.580	X	X		X	X	X	X		X	X	X	X	X	X	X		X	X	79
Aragón	1.105	0.702	X	X		X		X	X	X	X	X		X	X	X	X			X	55
Asturias y Cantabria	805	0.584		X	X	X	X	X	X		X	X				X		X	X	X	48
Baleares	429	0.641			X	X			X		X	X			X			X			34
Canarias	771	0.575	X		X	X			X	X	X					X		X		X	60
Castilla-León	3.157	0.625	X	X		X	X	X	X		X				X				X	X	51
Castilla-La Mancha y Murcia	2.220	0.608	X	X		X		X	X		X				X				X		46
Cataluña	1.642	0.645		X		X	X	X	X		X				X	X	X		X	X	52
Valencia	1.706	0.589	X			X	X	X	X	X			X	X		X		X	X		50
Extremadura	829	0.510		X		X		X	X		X						X	X	X		42
Galicia	829	0.550	X	X		X		X	X		X				X				X	X	44
Madrid	762	0.605		X		X	X	X	X		X								X	X	38
Navarra y La Rioja	724	0.612				X		X	X		X				X				X		54
País Vasco	1.694	0.594		X	X	X	X		X		X				X					X	41

FACTORS				**COVARIABLES**	
APES 103	Province	APES 405	Number of employed people	APES 302	RP's age
APES 104	Strata	APES 406	Number of unemployed people	APES 409	Size of household
APES 301	Sex of reference person (RP)	APES 407	Number of people under 16	APES 412	Dwelling's useful area (m^2)
APES 303	Highest level of education finished by RP	APES 408	Number of people over 64		
APES 304	RP's relation with activity	APES 410	Heating		
APES 306	RP's professional sit.	APES 411	Air conditioning		
APES 403	Type of	APES 413	Dwelling's tenancy regime		

3 What is understood by small area in Spain?

The surveys carried out by the INE have been designed to provide periodical information with a large number of characteristics and to find a specific balance between cost and accuracy not only on a national level, but also at other sub-population or domain levels.

In the context of sample surveys, a domain estimator is called *direct* if only the data collected with the sample for this domain are used to compile it. It is often necessary to obtain estimates for certain domains that have not been taken into account when designing the surveys, either because the necessary resources were not available, or because this need arose after the design was undertaken.

In domains such as those described above, it is probable that the survey statistic has few observations or even none. Under these circumstances, a *small area* is a domain for which it is not possible to obtain direct estimations with sufficient precision and this usually refers to cases where the domain is defined geographically.

Within the framework of the EURAREA project, each of the participating countries had to obtain estimations for two geographic levels: the province (NUT3) and any other level below this one.

In the case of Spain, the geographic levels chosen were *the province* (NUT3, 52 in the whole of Spain) and the *EURAREA-areas*, which is the "ad hoc" territorial unit at NUT4 level for the EURAREA project research. This territorial division consists of an intermediate geographic area between the province and municipality levels used by the INE as quality control areas for the monitoring of field work in the censuses. These areas are the responsibility of a field work inspector and their average population size is around 60,000 inhabitants.

This division is considered to be valid, for the purposes of analysis, in substitution of a possible NUT4 type territorial division (region?), which was not available at the time when the EURAREA project was started. Currently, definition work on the NUT4 level is very advanced in Spain, meaning that it will be possible to apply the EURAREA results to this *actual* territorial division in the short-term.

4 Complementary sources for the generation of area covariables

Using the microdata level information contained in the APES file, it is possible to obtain ancillary breakdown information on small areas. In addition, ancillary information on a small area level can be provided from complementary sources, such as the *Register of those requesting work at the INEM* and *Taxable Income Tax*, both broken down on a EURAREA-region level.

4.1 REGISTER OF THOSE REQUESTING WORK AT THE INEM

The National Institute for Employment provided the total numbers of people requesting employment by municipality which, combined adequately with data from the 1991 and 1998 APS, enabled us to obtain data on those people requesting employment for the geographic levels required in the project and for 1991.

4.2 ADMINISTRATIVE REGISTER ON TAXABLE INCOME TAX

Each year, the State Tax Administration Agency (AEAT) collects the annual income declared by contributors and for the first time, it has provided the broken down total on a post code level according to the source of origin (pensions, unemployment, agricultural activities, etc.). As with the total number of people requesting employment, totals have been obtained on a census section level relating to 1991 by applying the deflation indices for the census and postal code section crosses. Subsequently, the AEAT can now directly supply this data on a census section level to the INE, which means that this last step will no longer be necessary in current applications.

5 Definition of the population parameters

The APES artificial population parameters, which are estimated in the EURAREA project, are:

- *The proportion of the population that is ILO unemployed.* This parameter corresponds to the proportion of unemployed people in the population of people aged 16 years old and over. In terms of the APES file, this proportion can be expressed as:

$$\frac{\text{Total}\,(\text{APES503} = 1)}{\text{Total}\,(\text{APES203} \geq 16)}$$

- *The average annual income per consumption unit in households*. This parameter corresponds to:

$$\frac{1}{\text{N}} \sum_{i=1}^{\text{N}} \frac{\text{APES502}}{\text{APES505}}$$

where N is the total number of households and the APES505 variable represents the number of Consumption Units in the household according to the modified OECD scale. This scale allocates the following coefficients:

  - 1 for the main breadwinner
  - 0.5 for the remaining adults (14 years old or above)
  - 0.3 for children (under 14 years old)

The sum of household members weighted by these coefficients is what is called the number of consumption units in a household.

- *Proportion of single-person households.* This parameter corresponds to the proportion of the population of households with one single member. In particular, and in terms of the APES file, this proportion is expressed in the following way:

$$\frac{\text{Total}\,(\text{APES504} = 1)}{\text{Total}\,(\text{APES206} = 1)}$$

where the identification of the household is made using the household's reference person.

6 Sampling designs applied in the project

Within the framework for the EURAREA project, all participating countries must evaluate the *standard estimators* using their own database and obtain estimations of the 3 population parameters of interest defined in the section above.

The evaluation of the results mainly consists of obtaining estimations of the bias and mean squared error estimators using simulations of the high number of sample repetitions with similar designs to those used in official surveys in the real world.

Thus, two complex sample designs have been chosen that are similar to those commonly used in European surveys: the Active Population Survey (APS) to estimate ILO unemployment and the Living Conditions Survey (LCS) to estimate household income and composition.

6.1 SIMILAR DESIGN TO THE ACTIVE POPULATION SURVEY (APS)

The type of sample used is two-stage with stratification in the first stage units.

The first stage units are the census sections and they are grouped by strata according to the type of municipality to which they belong (demographic importance) and by applying the following classification:

- Stratum 1: Province capital municipalities
- Stratum 2: Self-represented municipalities, important areas in comparison with the capital.
- Stratum 3: Other self-represented municipalities, important areas in comparison with the capital or municipalities with more than 100,000 inhabitants.
- Stratum 4: Municipalities with between 50,000 and 100,000 inhabitants
- Stratum 5: Municipalities with between 20,000 and 50,000 inhabitants
- Stratum 6: Municipalities with between 10,000 and 20,000 inhabitants
- Stratum 7: Municipalities with between 5,000 and 10,000 inhabitants
- Stratum 8: Municipalities with between 2,000 and 5,000 inhabitants
- Stratum 9: Municipalities with under 2,000 inhabitants

The second stage units are made up of the households and within these units no sub-sample is carried out. Information is collected from all people whose usual residence is the households.

The sample selection has been carried out independently in each province and in such a way that within each strata, any individual has the same probability of being chosen, in other words, self-weighted samples are obtained within each strata.

To do this, the first stage units have been selected without replacements and with probability that is proportional to size according to the number of households. Within each section selected in the first stage, 20 households have been chosen with equal probabilities and without replacement.

The artificial population to be used in the EURAREA experiments has been reduced in accordance with all the participating countries in the interests of achieving a better balance between the sizes to be worked on in the different countries and due to resource costs in the data process.

In this way, the population under study in order to estimate unemployment is defined as all people aged 16 years or above in the Spanish EURAREA universe, in other words, people who belong to the Autonomous Communities of Andalucía, Canarias, Galicia, Valenciana and Madrid (approximately more than half of the APES artificial population).

Table 6.1.1 includes the first stage sample sizes:

Table 6.1.1 Sample sizes used in the first stage of the APS type design by strata and province

Provinces	1	2	3	4	5	6	7	8	9	Total
Alava	27				3		6			36
Albacete	15				6		3	6	6	36
Alicante	18	9		12	12	6	9	3	3	72
Almería	15				3	6	3	9		36
Avila	12						9		15	36
Badajoz	24				12	6	9	12	9	72
Baleares	30				12	12	9	9		72
Barcelona	60		30	12	18	9	6	6	3	144
Burgos	18				6		3		9	36
Cáceres	18				6	3	12	15	18	72
Cádiz	15	12	6	12	12	9	6			72
Castellón	24				15	12	3	9	9	72
Ciudad Real	12	9			12	9	15	9	6	72
Córdoba	30				12	9	12	9		72
Coruña (La)	21			12	6	12	15	6		72
Cuenca	12						6	6	12	36
Girona	15				12	12	9	12	12	72
Granada	24				12	6	12	18		72
Guadalajara	15						6		15	36
Guipúzcoa	24			6	15	15	6	6		72
Huelva	12					9	6	9		36
Huesca	12					9	6		9	36
Jaén	15	6			12	12	12	15		72
León	24	9				6	18		15	72
Lleida	12					3	3	6	12	36
Logroño	21					6	6	6	9	48
Lugo	12					6	9	9		36
Madrid	99		21	9	9		6			144
Málaga	36			6	12	6		12		72
Murcia	24	12		6	12	9	9			72
Navarra	30				3	6	6	15	12	72
Ourense	12					3	9	12		36
Oviedo	21	24		18	15	12	9	9		108
Palencia	15						9		12	36
Palmas (Las)	36			6	12	9	6	3		72
Pontevedra	12	24			9	15	9	3		72
Salamanca	18					3	3		12	36
S.Cruz Tenerife	24	12			12	9	9	6		72
Santander	24	9				12	6	12	9	72
Segovia	15						6		15	36
Sevilla	48			6	18	12	12	12		108
Soria	12						9		15	36
Tarragona	18	12			6	12	6	9	9	72
Teruel	12					3	9		12	36
Toledo	15	15					15	15	12	72
Valencia	48			6	24	9	9	6	6	108
Valladolid	24					3	3		6	36
Vizcaya	30	6		12	6	6	6	6		72
Zamora	12					6			18	36
Zaragoza	48					6	9		9	72
Ceuta	12									12
Melilla	12									12
Total	1,170	159	57	123	321	321	384	300	309	3,168

6.2 DESIGN SIMILAR TO THE LIVING CONDITIONS SURVEY (LCS)

A two-stage stratified sample has been used to select the sample in the first stage units, which are the census sections. The second stage units are households.

With this criteria, an independent sample has been selected in each Autonomous Community.

The stratification variable in the census sections is the size of the municipality to which it belongs, but with slight differences in relation to the stratification applied in the previous design, as described below:

- Stratum 0: Municipality of Barcelona.
- Stratum 1: Other province capital municipalities.
- Stratum 2: Municipalities with more than 100.000 inhabitants.
- Stratum 3: Municipalities with between 50,000 and 100,000 inhabitants.
- Stratum 4: Municipalities with between 20,000 and 49,999 inhabitants.
- Stratum 5: Municipalities with between 10,000 and 19,999 inhabitants.
- Stratum 6: Municipalities with under 10,000 inhabitants.

In the first stage, the census sections are selected without replacement and with probabilities in proportion to the size according to the number of households. In the second stage, 8 households were selected with equal probabilities and without replacement from each census section selected in the previous stage.

In order to estimate the household's income and composition, the population being researched is made up of all households belonging to the Spanish EURAREA universe, in other words, those that belong to the Autonomous Communities of Andalucía, Canarias, Galicia, Valenciana and Madrid (more than half of the APES population approximately).

Table 6.2.1 includes the sample sizes used in the first stage:

Table 6.2.1 Sample sizes used in the first stage of the LCS type design by strata and Autonomous Community

Regions	0	1	2	3	4	5	6	Total
Andalucía		60	5	18	30	24	43	180
Aragón		34				6	21	61
Asturias (Principado)		11	14	10	5	11	9	60
Baleares (Illes)		21			10	8	12	51
Canarias		26	5	4	13	8	13	69
Cantabria		17		5		7	17	46
Castilla-La Mancha		45		3	3	7	49	107
Castilla y León		15		5	6	5	41	72
Cataluña	47	8	30	15	22	16	34	172
C. Valenciana		36	6	9	29	14	28	122
Extremadura		11			9	5	33	58
Galicia		18	10	6	8	21	34	97
C. de Madrid		85	27	10	11		8	141
Murcia (Región de)		18	9	4	12	9	6	58
C. F. de Navarra		17			3	6	24	50
País Vasco		29	5	11	10	12	15	82
Rioja (La)		19				5	16	40
Ceuta y Melilla		34						34
TOTAL	47	504	111	100	171	164	403	1500

7 The standard estimators

In order to estimate each of the population parameters researched, more than 20 small area estimators have been tested using simulation experiments. One of the EURAREA project's main objectives however is to evaluate the standard methodology. For this reason, below you will find a complete definition of the small area estimators named standard in the EURAREA context.

Prior to this, we will make some comments on the notes that we will use throughout the document:

- Sub-indices: s is used to designate samples.
  h=1, ,H for the strata
  d=1,.,.D for the small areas
  i for the units researched
- Sizes: N for the population researched and n for the sample selected. When N or n have a sub-index, it indicates the size of the sub-group defined by the sub-index. For example, n_d is the size of the sample selected in the small area d.
- Totals: $Y \acute{o} X$. When $Y \acute{o} X$ has a sub-index, it indicates the total for the sub-group corresponding to the sub-index. For example, Y_d denotes the total Y in the small area d.
- Averages: $\overline{Y} \acute{o} \overline{X}$. When $\overline{Y} \acute{o} \overline{X}$ has a sub-index, it indicates the average of the sub-group corresponding to the sub-index. For example $\overline{Y}_d$ denotes the average $\overline{Y}$ of the small area d.
- Weights: w_i is used for the sample weight of unit i. When it has a sub-index, it also indicates the sum of the weights corresponding to the sample units belonging to the sub-population defined by the sub-index.

In a general way, the population parameters researched in EURAREA can be considered population means constructed using values $y_1,\ldots,y_N$. In other words, they can be expressed in the following way:

$$\overline{Y}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} y_i$$

The standard estimators are therefore the following:

Estimator 1*:* direct estimator

$$\hat{\bar{Y}}_d^{DIRECT} = \frac{1}{\hat{N}_d} \sum_{i \in s_d} w_i \, y_i \qquad \text{where} \quad \hat{N}_d = \sum_{i \in s_d} w_i$$

Estimator 2*:* generalised regression estimator (GREG)

$$\hat{\bar{Y}}_d^{GREG} = \hat{\bar{Y}}_d^{DIRECT} + \left(\bar{X}_d - \hat{\bar{X}}_d^{DIRECT} \right)^T \hat{\beta}$$

where $\bar{X}_d = (\bar{X}_{d1}, K, \bar{X}_{dp})^T$ is the vector column for population means of the ancillary p variables included in the regression model adopted:

$$y_i = \alpha + x_i^T \beta + e_i$$

where $x_i = (x_{i1}, K, x_{ip})^T$ is the vector column for the values of the ancillary variables associated with unit i, assuming that $E(e_i) = 0$ y $V(e_i) = \sigma_e^2$. $\hat{\beta}$ is the parameter estimator β obtained using the least square method.

Estimator 3: synthetic estimator under model A (regression model with individual data and random area effects):

$$y_i = x_i^T \beta + u_d + e_i$$

where $u_d \sim N(o, \sigma_u^2)$ y $e_i \sim N(o, \sigma_e^2)$ are independent.

The synthetic estimator is therefore expressed in the following way:

$$\hat{\bar{Y}}_d^{SYNTHA} = \bar{X}_d^T \hat{\beta}$$

Estimator 4: synthetic estimator under model B (regression model with small area data):

$$\bar{Y}_d = \bar{X}_d^T \beta + u_d \qquad \text{y} \qquad \hat{\bar{Y}}_d^{DIRECT} = \bar{Y}_d + e_d$$

where. $u_d \sim N(o, \sigma_u^2)$ y $e_d \sim N(o, \sigma_e^2)$ are independent.

The estimator is therefore expressed in the following way:

$$\hat{\bar{Y}}_d^{SYNTHB} = \bar{X}_d^T \hat{\beta}$$

Estimator 5: synthetic estimator under model C (logistic regression model with small area data):

$$\text{logit}\,(p_d) = \overline{X}_d^T \beta + e_d$$

where $e_d \sim N(o, \sigma_e^2)$ y p_d represents the probability of value one of the binary variables under study in the small area.

Estimator 6: EBLUP estimator (Empirical Best Linear Unbiased Predictor) under model A:

$$\hat{\overline{Y}}_d^{EBLUPA} = \hat{\gamma}_d \hat{\overline{Y}}_d^{GREG} + (1-\hat{\gamma}_d)\overline{X}_d^T \hat{\beta}$$

where $\hat{\gamma}_d = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \dfrac{\hat{\sigma}_e^2}{n_d}}$

Estimator 7: EBLUP estimator under model B:

$$\hat{\overline{Y}}_d^{EBLUPB} = \hat{\gamma}_d \hat{\overline{Y}}_d^{DIRECT} + (1-\hat{\gamma}_d)\overline{X}_d^T \hat{\beta}$$

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}$$

8 Evaluation measures calculated in the simulations

With the aim of evaluating the suitability of the proposed estimators K, independent samples from each sampling design have been extracted from the Spanish universe in EURAREA and the corresponding estimates have been calculated using each of them.

Where $\hat{\bar{Y}}_d(k)$ is the mean population estimate $\bar{Y}_d$ obtained with sample K, the measures for evaluation the estimation's performance are the following:

1. *Mean relative bias associated with small area d:*

$$ARB_d = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\hat{\bar{Y}}_d(k)}{\bar{Y}_d} - 1\right)100$$

2. *Relative bias average:*

$$\overline{ARB} = \frac{1}{D}\sum_{d=1}^{D} ARB_d$$

where D is the total of small areas.

3. *Square root of relative mean square error associated with the small area d:*

$$RMSE_d = \frac{100}{\bar{Y}_d}\sqrt{\frac{1}{k}\sum_{k=1}^{K}\left(\hat{\bar{Y}}_d(k) - \bar{Y}_d\right)^2}$$

4. *Relative error square average:*

$$\overline{RMSE} = \frac{1}{D}\sum_{d=1}^{D} RMSE_d$$

During 2002 and after finishing the construction of the APES artificial population, a number of approximation exercises were carried out in order to evaluate the use of time and resources when generating a high number of samples, applying estimates and calculating evaluation measures. Thus, 10,000 samples were selected at the beginning using a non-stratified random sample and subsequently with stratification. More than 20 different estimators were applied to them, which were evaluated. During this time, it was noticed that some relative errors were excessively large in some areas given that the population parameter value to be estimated was very close to zero (particularly when estimating proportions). As a result of this, the relative error concept was replaced with the absolute error concept and the following measures began to be calculated:

5. *Square root of mean square error associated with the small area d:*

$$EMSE_d = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\overline{Y}}_d(k) - \overline{Y}_d\right)^2}$$

6. Error root average:

$$\overline{EMSE} = \frac{1}{D}\sum_{D=1}^{D} EMSE_d$$

9. Theory

In small area estimation theory, combined linear models with a random factor are used as tools to obtain the EBLUP estimators (empirical best linear unbiased predictor) of population means or totals. The parameter estimators of such models (regression coefficients and variance components) only contain efficiency properties when the analysed data come from sampling designs with equal inclusion probabilities. When a design sample is used with unequal inclusion probabilities, the model parameter estimators lose their optimum properties. In this way, the following two questions can be asked: is it necessary to adopt model fitting algorithms in the case of unequal sampling weights (or more generally, complex sampling designs)?, how should this be done?, what differences would there be in the small area estimators?

In this section, we provide a number of modifications to the Fisher scoring algorithm in order to fit combined linear models with a random factor when the samples are obtained from complex sampling designs (samples that are different from a simple random sample). In particular, we study the problem of how to introduce sampling weights (inverse of inclusion probabilities) into fitting algorithms.

9.1. EBLUP ESTIMATE IN COMPLEX SAMPLING DESIGNS

This section describes the standard theory on empirical best linear unbiased predictors under combined linear models with a random factor and with samples obtained from complex sampling designs.

9.1.1 The Fisher scoring census model and algorithm

If we consider a population with *N* units and *D* small areas. N_d is the number of units in the small area *d*. *Y* is the variable of interest taking the values $\boldsymbol{y} = (y_1,...,y_N)$. We suppose that this population vector is a carrying out of the variables $Y_1,...,Y_N$ distributed according to the model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}, \tag{9.1}$$

where $\boldsymbol{y} = \boldsymbol{y}_{N\times 1}$, $\boldsymbol{X} = \boldsymbol{X}_{N\times p}$ is a matrix of constants with the values of the ancillary variables in columns, $r(\boldsymbol{X}) = p$, $\boldsymbol{\beta} = \boldsymbol{\beta}_{p\times 1}$ is the coefficient regression vector of the fixed covariables or effects, $\boldsymbol{Z} = \boldsymbol{Z}_{N\times D} = diag\left(\mathbf{1}_{N_1},\mathrm{K},\mathbf{1}_{N_D}\right)$ where $\mathbf{1}_{N_d}$ is a column vector of some of a size N_d, $\boldsymbol{u} = \boldsymbol{u}_{D\times 1} \sim N\left(\boldsymbol{0},\sigma_u^2\boldsymbol{I}_D\right)$ is independent of $\boldsymbol{e} = \boldsymbol{e}_{N\times 1} \sim N\left(\boldsymbol{0},\sigma_e^2 I_N\right)$ and $\boldsymbol{I}_a = diag\left(1,\mathrm{K},1\right)_{a\times a}$. Also note that the model (9.1) can be written alternatively as in Prasad and Rao (1990) in other words,

$$y_{dj} = \boldsymbol{x}_{dj}\beta + u_d + e_{dj}, \quad d = 1,\mathrm{K}\ D,\ j = 1,\mathrm{K}\ , N_d\ , \tag{9.2}$$

where y_{dj} is the characteristic of interest for the unit *j* of area *d* and $\boldsymbol{x}_{dj}$ is the row (*d*, *j*) of matrix $\boldsymbol{X}$ which contains the corresponding ancillary variables. The model (9.2) can be interpreted as a model with ordinates in its random origin.

Let $\theta = \left(\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2\right)^t$ be the parameter vector. The vector's density function $\boldsymbol{y}$ under (9.1) is

$$f_\theta(\boldsymbol{y}) = c\left|\boldsymbol{V}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^t \boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)\right\},$$

where $\boldsymbol{V} = \mathrm{var}(\boldsymbol{y}) = \sigma_e^2 \boldsymbol{I}_N + \sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}^t = diag(\boldsymbol{V}_1, ..., \boldsymbol{V}_D)$, $\boldsymbol{V}_d = \sigma_e^2 \boldsymbol{I}_{N_d} + \sigma_u^2 \mathbf{1}_{N_d} \mathbf{1}_{N_d}^t$, *d=1,.D*, and *c* is a constant. The log-density is

$$l(\theta) = \ln f_\theta(\boldsymbol{y}) = \ln c - \frac{1}{2}\ln\left|\boldsymbol{V}\right| - \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^t \boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)$$

The scoring vector, evaluated at point θ, is

$$\boldsymbol{S}(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \left(\frac{\partial l(\theta)}{\partial \beta}, \frac{\partial l(\theta)}{\partial \sigma_u^2}, \frac{\partial l(\theta)}{\partial \sigma_e^2}\right) = \left(\boldsymbol{S}_\beta^t, \boldsymbol{S}_{\sigma_u^2}, \boldsymbol{S}_{\sigma_e^2}\right)^t,$$

and the maximum realistic estimators are obtained by solving the equation $\mathbf{0} = \boldsymbol{S}(\theta)$. The Fisher scoring algorithm is frequently used to numerically calculate these estimators. The algorithm starts with some initial estimator values (seeds), $\theta^0 = \left(\boldsymbol{\beta}_0^t, \sigma_{u,0}^2, \sigma_{e,0}^2\right)$, and they are updated in each iteration using the equation

$$\theta^{i+1} = \theta^i + \boldsymbol{F}\left(\theta^i\right)^{-1} \boldsymbol{S}\left(\theta^i\right), \qquad (9.3)$$

where

$$\boldsymbol{F}\left(\theta\right) = -E\left[\frac{\partial \boldsymbol{S}(\theta)}{\partial \theta}\right] = \begin{pmatrix} \boldsymbol{F}_{\beta\beta} & \boldsymbol{F}_{\beta\sigma_u^2} & \boldsymbol{F}_{\beta\sigma_e^2} \\ \boldsymbol{F}_{\beta\sigma_u^2} & \boldsymbol{F}_{\sigma_u^2\sigma_u^2} & \boldsymbol{F}_{\sigma_u^2\sigma_e^2} \\ \boldsymbol{F}_{\beta\sigma_e^2} & \boldsymbol{F}_{\sigma_u^2\sigma_e^2} & \boldsymbol{F}_{\sigma_e^2\sigma_e^2} \end{pmatrix}$$

is the Fisher information matrix in point θ. Defined as

$$\boldsymbol{y}_d = \begin{pmatrix} y_{d1} \\ \mathrm{M} \\ y_{dN_d} \end{pmatrix}, \quad \boldsymbol{X}_d = \begin{pmatrix} x_{d11} & \mathrm{L} & x_{d1p} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ x_{dN_d 1} & \mathrm{L} & x_{dN_d p} \end{pmatrix}, \quad d=1,.,D. \qquad (9.4)$$

Therefore, under the population model (9.1) the scores are:

$$S_{\beta} = \frac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\varsigma_d - \frac{\gamma_d}{N_d}\boldsymbol{X}_d^t\mathbf{1}_{N_d}\mathbf{1}_{N_d}^t\varsigma_d\right),\quad S_{\sigma_u^2} = -\frac{1}{2}\sum_{d=1}^{D}\frac{1-\gamma_d}{\sigma_e^2}N_d + \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d}{\sigma_e^2}\right)^2\varsigma_d^t\mathbf{1}_{N_d}\mathbf{1}_{N_d}^t\varsigma_d$$ and

$$S_{\sigma_e^2} = -\frac{1}{2}\sum_{d=1}^{D}\frac{N_d-\gamma_d}{\sigma_e^2} + \frac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[\varsigma_d^t\varsigma_d + \frac{\gamma_d\left(\gamma_d-2\right)}{N_d}\varsigma_d^t\mathbf{1}_{N_d}\mathbf{1}_{N_d}^t\varsigma_d\right],$$ where

$\varsigma_d = \boldsymbol{y}_d - \boldsymbol{X}_d\boldsymbol{\beta}$ and $\gamma_d = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/N_d}$, *d=1,..,D.*

The Fisher information matrix elements are

$$F_{\beta\beta} = \frac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\boldsymbol{X}_d - \frac{\gamma_d}{N_d}\boldsymbol{X}_d^t\mathbf{1}_{N_d}\mathbf{1}_{N_d}^t\boldsymbol{X}_d\right),\quad F_{\sigma_u^2\sigma_u^2} = \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d}{\sigma_e^2}\right)^2 N_d^2,$$

$$F_{\sigma_u^2\sigma_e^2} = \frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d}{\sigma_e^2}\right)^2 N_d,\quad F_{\sigma_e^2\sigma_e^2} = \frac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[N_d + \gamma_d\left(\gamma_d-2\right)\right]$$ and $$F_{\beta\sigma_u^2} = F_{\beta\sigma_e^2} = 0.$$

9.1.2 Empirical best linear unbiased predictor

Let n_d be the number of population units in the sample of area d and we define $f_d = n_d/N_d$. The empirical best linear unbiased predictor (EBLUP) of $\overline{Y}_d$ is

$$\hat{\bar{Y}}_d^{eblup} = (1-f_d)\hat{\bar{Y}}_d^{eblupa} + f_d\left[\hat{\bar{Y}}_d + \hat{\beta}\left(\overline{\boldsymbol{X}}_d - \hat{\bar{X}}_d\right)\right],$$

where

$$\hat{\bar{Y}}_d^{eblupa} = \overline{\boldsymbol{X}}_d\hat{\beta} + \hat{\gamma}_d^w(\hat{\bar{Y}}_d^{direct} - \hat{\bar{X}}_d^{direct}\hat{\beta}),\qquad \hat{\gamma}_d^w = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + (\hat{\sigma}_e^2/w_{d.})},\qquad \overline{\boldsymbol{X}}_d = \frac{1}{N_d}\sum_{j=1}^{N_d}\boldsymbol{x}_{dj},$$

$$\hat{\bar{X}}_d = \frac{1}{n_d}\sum_{j=1}^{n_d}\boldsymbol{x}_{dj},\quad \hat{\bar{X}}_d^{direct} = \frac{1}{\hat{N}_d}\sum_{j=1}^{n_d}w_{dj}\boldsymbol{x}_{dj},\quad \hat{\bar{Y}}_d = \frac{1}{n_d}\sum_{j=1}^{n_d}y_{dj},\quad \hat{\bar{Y}}_d^{direct} = \frac{1}{\hat{N}_d}\sum_{j=1}^{n_d}w_{dj}y_{dj},\quad \hat{N}_d = \sum_{j=1}^{n_d}w_{dj}.$$

The estimators $\hat{\beta}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ can be calculated using a sample version of the Fisher scoring algorithm presented in (9.3).

Let $\Omega=\{1,...,N\}$ be a finite population. Let $s\subset\Omega$ and $r=\Omega-s$ be the groups of sampled and unsampled units respectively. It is interesting to observe that

$$\hat{\bar{Y}}_d^{eblup} = \frac{1}{N_d}\sum_{j=1}^{N_d}\hat{\bar{Y}}_{dj}^{eblup}$$

where

$$\hat{\bar{Y}}_{dj}^{eblup} = \begin{cases} \boldsymbol{x}_{dj}\hat{\beta} + \hat{\gamma}_d^w\left(\hat{\bar{Y}}_d^{wdirect} - \hat{\bar{X}}_d^{wdirect}\hat{\beta}\right) & if \quad y_{dj}\in r \\ y_{dj} & if \quad y_{dj}\in s \end{cases}$$

so that $\hat{\bar{Y}}_d^{eblup}$ is also called a *predictive estimator*. On the other hand, in the EURAREA project the following *projective estimator* is used (EBLUPA).

$$\begin{aligned}\hat{\bar{Y}}_d^{eblupa} &= \frac{1}{N_d}\sum_{j=1}^{N_d}\left\{\boldsymbol{x}_{dj}\hat{\beta} + \hat{\gamma}_d^w\left(\hat{\bar{Y}}_d^{direct} - \hat{\bar{X}}_d^{direct}\hat{\beta}\right)\right\} \\ &= (1-\hat{\gamma}_d^w)\overline{\boldsymbol{X}}_d\hat{\beta} + \hat{\gamma}_d^w\left(\hat{\bar{Y}}_d^{direct} + \left(\overline{\boldsymbol{X}}_d - \hat{\bar{X}}_d^{direct}\right)\hat{\beta}\right) = (1-\hat{\gamma}_d^w)\hat{\bar{Y}}_d^{syntha} + \hat{\gamma}_d^w\hat{\bar{Y}}_d^{greg}.\end{aligned}$$

9.1.3 Inclusion probabilities and weights

In probability sampling, a sampling plan (or sampling design) is a process by which a sample is chosen so that each sub-population (sample) *s* of units has a probability *p (s)* of being selected. Let's suppose that a sample of size *n* is extracted in accordance with a sampling design with inclusion probabilities

$$\pi_{dj} = \sum_{s\in S(d,j)} p(s),$$

where *S*(*d*, *j*) is the total of all samples of size *n* that contain the individual *j* from small area *d*. The weights $w_{dj} = 1/\pi_{dj}$, can be interpreted as the number of population units represented by the sampling unit *j* of small area *d*. Let us also consider the inclusion probabilities $\pi_d = P(s\cap d\neq\varnothing),$ the conditional probabilities $\pi_{j|d} = \pi_{dj}/\pi_d$ and their corresponding weights $w_d = 1/\pi_d$ and $w_{j|d} = 1/\pi_{j|d}$.

Complex sampling designs are frequently used in national surveys in order to reduce costs and to take into account the geographical and socio-economical characteristics of the population under study. When two-stage sampling designs are used, it is normal that the first stage units do not coincide with the small areas of interest. For this reason, this section illustrates the calculation of inclusion probabilities in such small areas in a two-stage sampling design with stratification in the first stage. Let us suppose that the first stage units are territories that are completely contained within a small area and with the aim of

using a common name, we'll call them census sections. Census sections are selected without replacement and equal probabilities. The second stage units are the dwellings and 20 of these are selected without replacement using simple random sampling. Included in the sample are all individuals (final units) belonging to a selected dwelling. This is a modified version of the sampling design in the Active Population Survey.

Let H be the number of strata, m_h the number of census sections in strata h selected in the sample, M_h the number of census sections in strata h in the population, M_{hd} the number of census sections in small area d in strata h in the population, N_h the number of dwellings in strata h in the population, N_{hi} the number of dwellings in census section i in strata h in the population, N_{hd} the number of dwellings in small area d in strata h in the population, π_{hij} the inclusion probability of dwelling j in census section i in strata h, π_{hi} the inclusion probability of census section i of strata h, $\pi_{j|hi}$ the inclusion probability in the second stage of dwelling j in census section i when in the first stage census section i of strata h has been selected and finally, let π_d be the inclusion probability of small area d. Thus:

$$\pi_{j|hi} = 1 - \frac{\binom{N_{hi}-1}{20}}{\binom{N_{hi}}{20}} = \frac{20}{N_{hi}}, \qquad \pi_{hi} = 1 - \frac{\binom{M_h-1}{m_h}}{\binom{M_h}{m_h}} = \frac{m_h}{M_h} \quad \text{and} \quad \pi_{hij} = \pi_{j|hi}\pi_{hi} = \frac{20m_h}{M_h N_{hi}},$$

so that the sampling weights of the individuals in dwelling j in census section i in strata h are

$$w_{hij} = \frac{1}{\pi_{hij}} = \frac{M_h N_{hi}}{20 m_h}.$$

On the other hand, if we suppose that the population size is large enough, in the following calculation we can accept that the selection of first stage units has been with replacement. In this case, $w_d = 1/\pi_d$ is obtained with

$$\pi_d = 1 - \prod_{h=1}^{H} P(d \cap h \cap s = \varnothing) \approx \prod_{h=1}^{H} \left(1 - \frac{M_{hd}}{M_h}\right)^{m_h} = 1 - \exp\left\{\sum_{h=1}^{H} m_h \ln\left(1 - \frac{M_{hd}}{M_h}\right)\right\}.$$

9.2 SAMPLING VERSIONS OF THE FISHER SCORING CENSUS ALGORITHM

This section considers sampling versions of the Fisher scoring algorithm (9.3) supposing that the model (9.2) , or a modification of the model, is also valid for the sample.

9.2.1 Fisher scoring census algorithm

In order to obtain EBLUP estimates of small areas, statisticians usually suppose that the model (9.2) is also valid for the sample; in other words, they suppose that

$$y_{dj} = \boldsymbol{x}_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad d = 1,\text{K } D,\ j = 1,\text{K },n_d \text{ ,} \tag{9.5}$$

where $u_d \sim iid\ N\left(\boldsymbol{0},\sigma_u^2\right)$ are independent and $e_{dj} \sim iid\ N\left(\boldsymbol{0},\sigma_e^2\right)$. The EBLUP estimators are therefore obtained by fitting the model (9.5).

The Fisher scoring algorithm without weights uses the updating equation (9.3) with scores

$$S_\beta = \tfrac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\varsigma_d - \tfrac{\gamma_d}{n_d}\boldsymbol{X}_d^t\boldsymbol{1}_{n_d}\boldsymbol{1}_{n_d}^t\varsigma_d\right), \quad S_{\sigma_u^2} = -\tfrac{1}{2}\sum_{d=1}^{D}\tfrac{1-\gamma_d}{\sigma_e^2}n_d + \tfrac{1}{2}\sum_{d=1}^{D}\left(\tfrac{1-\gamma_d}{\sigma_e^2}\right)^2\varsigma_d^t\boldsymbol{1}_{n_d}\boldsymbol{1}_{n_d}^t\varsigma_d \text{ ,}$$

$$S_{\sigma_e^2} = -\tfrac{1}{2}\sum_{d=1}^{D}\tfrac{n_d-\gamma_d}{\sigma_e^2} + \tfrac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[\varsigma_d^t\varsigma_d + \tfrac{\gamma_d\left(\gamma_d-2\right)}{n_d}\varsigma_d^t\boldsymbol{1}_{n_d}\boldsymbol{1}_{n_d}^t\varsigma_d\right],$$

and elements of the Fisher information matrix

$$F_{\beta\beta} = \tfrac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\boldsymbol{X}_d - \tfrac{\gamma_d}{n_d}\boldsymbol{X}_d^t\boldsymbol{1}_{n_d}\boldsymbol{1}_{n_d}^t\boldsymbol{X}_d\right), \quad F_{\sigma_u^2\sigma_u^2} = \tfrac{1}{2}\sum_{d=1}^{D}\left(\tfrac{1-\gamma_d}{\sigma_e^2}\right)^2 n_d^2, \quad F_{\sigma_u^2\sigma_e^2} = \tfrac{1}{2}\sum_{d=1}^{D}\left(\tfrac{1-\gamma_d}{\sigma_e^2}\right)^2 n_d \text{ ,}$$

$$F_{\sigma_e^2\sigma_e^2} = \tfrac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[n_d + \gamma_d\left(\gamma_d - 2\right)\right] \quad \text{and} \quad F_{\beta\sigma_u^2} = F_{\beta\sigma_e^2} = 0 \text{ ,}$$

where $\boldsymbol{y}_d$ y $\boldsymbol{X}_d$ are taken from (9.4) with n_d in place of N_d, $\varsigma_d = \boldsymbol{y}_d - \boldsymbol{X}_d\beta$ and $\gamma_d = \tfrac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/n_d}$

By proposing model (9.5) for the sample, you are implicitly assuming that the distribution of the sampling vector (y_1,...,y_n) is obtained directly from (9.2) and

that as a result, the sample's random selection mechanism is ignored. Unfortunately, this fact can not be justified with probability arguments (see, for example, section 2.6.2 of Vaillant et al. (2000)). As a result, it is necessary to research more deeply in order to clarify when the model (9.5) can be used to obtain the EBLUP estimator for small areas associated with the model (9.2).

9.2.2 Fisher scoring algorithm with unit weights

An alternative process consists of using the sample to construct an artificial population repeating unit *(d,j)* w_{dj} times. For this artificial population we propose, similarly to (9.2), the model

$$y_{djk} = \boldsymbol{x}_{djk}\boldsymbol{\beta} + u_d + e_{djk}, \quad d = 1,\text{K } D,\ j = 1,\text{K },n_d, k = 1,...,w_{dj}, \tag{9.6}$$

where $u_d \sim iid\ N\left(\boldsymbol{0},\sigma_u^2\right)$ are independent and $e_{djk} \sim iid\ N\left(\boldsymbol{0},\sigma_e^2\right)$. The EBLUP estimators are obtained by fitting the model (9.6). In (9.6), instead of taking w_{dj}, we take the closest whole to $1/\pi_{dj}$. Note that in the surveys carried out by national statistics institutes, $1/\pi_{dj}$ is usually greater than 100, meaning that this last approximate is admissible.

It is also possible to consider the following model for the sample.

$$y_{dj} = \boldsymbol{x}_{dj}\boldsymbol{\beta} + u_d + e_{dj}, \quad d = 1,\text{K } D,\ j = 1,\text{K },n_d, \tag{9.7}$$

where $u_d \sim iid\ N\left(\boldsymbol{0},\sigma_u^2\right)$ and $e_{dj} \sim iid\ N\left(\boldsymbol{0},w_{dj}^{-1}\sigma_e^2\right)$ are independent. The EBLUP estimators can be obtained by fitting the model (9.7).

In section 4.5 of Morales and Molina (2002), it is shown that if σ_u^2 and σ_e^2 are known, then the most reliable estimators/predictors $\hat{\beta}$ and $\hat{u}_d$, *d=1,...,D*, in the models (9.6) and (9.7) coincide. Under the same hypothesis, the corresponding BLUP estimators also coincide. Remember that Henderson (1975) demonstrated that the BLUP estimator of $\boldsymbol{l}^t\boldsymbol{\beta} + u_d$ is $\boldsymbol{l}^t\hat{\beta} + \hat{u}_d$ in combined linear models of the (9.6) or (9.7) type with σ_u^2 and σ_e^2 known.

Using sound arguments, if the artificial population with the model (9.6) is considered a reasonably good approximate of the actual population with the model (9.1), then we propose fitting the model (9.7) in order to estimate/predict β and u_d, *d=1,...,D*, of model (9.1).

The Fisher scoring algorithm with weights in units uses the updating equation (1.3) with scores

$$S_{\beta}=\frac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\boldsymbol{W}_d\varsigma_d-\frac{\gamma_d^w}{w_{d\cdot}}\boldsymbol{X}_d^t\boldsymbol{w}_{n_d}\boldsymbol{w}_{n_d}^t\varsigma_d\right),\quad S_{\sigma_u^2}=-\frac{1}{2}\sum_{d=1}^{D}\frac{1-\gamma_d^w}{\sigma_e^2}w_{d\cdot}+\frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d^w}{\sigma_e^2}\right)^2\varsigma_d^t\boldsymbol{w}_{n_d}\boldsymbol{w}_{n_d}^t\varsigma_d,$$

$$S_{\sigma_e^2}=-\frac{1}{2}\sum_{d=1}^{D}\frac{n_d-\gamma_d^w}{\sigma_e^2}+\frac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[\varsigma_d^t\boldsymbol{W}_d\varsigma_d+\frac{\gamma_d^w\left(\gamma_d^w-2\right)}{\omega_{d\cdot}}\varsigma_d^t\boldsymbol{w}_{n_d}\boldsymbol{w}_{n_d}^t\varsigma_d\right]$$

and elements of the Fisher information matrix

$$F_{\beta\beta}=\frac{1}{\sigma_e^2}\sum_{d=1}^{D}\left(\boldsymbol{X}_d^t\boldsymbol{W}_d\boldsymbol{X}_d-\frac{\gamma_d^w}{w_{d\cdot}}\boldsymbol{X}_d^t\boldsymbol{w}_{n_d}\boldsymbol{w}_{n_d}^t\boldsymbol{X}_d\right),\quad F_{\sigma_u^2\sigma_u^2}=\frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d^w}{\sigma_e^2}\right)^2w_{d\cdot}^2,\quad F_{\sigma_u^2\sigma_e^2}=\frac{1}{2}\sum_{d=1}^{D}\left(\frac{1-\gamma_d^w}{\sigma_e^2}\right)^2w_{d\cdot},$$

$$F_{\sigma_e^2\sigma_e^2}=\frac{1}{2\left(\sigma_e^2\right)^2}\sum_{d=1}^{D}\left[n_d+\gamma_d^w\left(\gamma_d^w-2\right)\right],\quad F_{\beta\sigma_u^2}=F_{\beta\sigma_e^2}=0,$$

where $\boldsymbol{y}_d$ y $\boldsymbol{X}_d$ are taken from (9.4) with n_d instead of N_d, $\varsigma_d=\boldsymbol{y}_d-\boldsymbol{X}_d\boldsymbol{\beta}$,

$$\boldsymbol{W}_d=diag\left(w_{d1},\mathrm{K},w_{dn_d}\right)_{n_d\times n_d},\quad \boldsymbol{w}_{n_d}=\left(w_{d1},\mathrm{K},w_{dn_d}\right)_{n_d\times 1}^t,\quad w_{d\cdot}=\mathbf{1}_{n_d}^t\boldsymbol{w}_{n_d}=\sum_{j=1}^{n_d}w_{dj}$$ and

$\gamma_d^w=\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2/w_{d\cdot}}$, *d=1,.,D.*

Morales, Molina and Santamaría (2002) have obtained some computational results that demonstrate that the EBLUP estimators obtained by fitting the model (9.7) are more effective (in bias and relative mean squared error) than the EBLUP estimators obtained from the sampling model (9.1). If the individual weights are all the same as one, this algorithm coincides with the Fisher scoring algorithm without weights.

9.2.3 Fisher scoring algorithm with weights in the units and in areas

Section 9.2.2 implicitly proposes a calculation of the realistic population (census) using the sampling reality and the obtaining of consistent estimators using this last reality. This procedure is widely accepted by statisticians when standard linear regression models are fitted (with random disruptions at just one level) to the sampling data. In such a case, the values of the variable under study in the elemental units of the finite population are considered independent, meaning that the census reality is a sum that can be consistently estimated by weighting the observations. Pfeffermann and others (1998) give the following reasons for why the multi-level models are different to the single-level models with regards the weighting of observations.

1. The values observed in the units of a finite population are not independent in such models and therefore the census log-reality is not a simple sum across the population. This implies that it can not be estimated using the method of weighting the sampling observations.

2. The inclusion probabilities of the last sampling units do not provide enough information to carry out a correction of the relevant biases, which is the opposite of what happened in the case of single-level regression models.

With the aim of reducing the bias of estimators from the model (9.5), Pfefferman and others (1998) suggest reproducing the method of estimating the population with sampling observations, using the introduction of two-level weights. They suggest replacing each population sum of units *j* within the area *d* with the sampling sums with values weighted by $w_{j|d}$ and each population sum of the small areas *d* with the corresponding sum with weighted elements with $w_{d.}$ In order to apply Pfefferman's suggestion, in the census scores expressions and elements of the Fisher matrix, we have replaced the sums $\sum_{j=1}^{N_d} b_{dj}$ and $\sum_{d=1}^{D} a_d$ with $\sum_{j=1}^{n_d} w_{j|d} b_{dj}$ i $\sum_{d=1}^{D} w_d a_d$ respectively. The sizes N_d are also replaced with the corresponding estimators $\hat{N}_d = \sum_{j=1}^{n_d} w_{j|d}$. Pfeffermann and others (1998) justify their suggestion by arguing that the population sums are estimated in a biased and consistent way (with regards the sample distribution) by the corresponding weighted sampling sums.

Let $\boldsymbol{y}_d$ y $\boldsymbol{X}_d$ be the vector and the matrix defined in (9.4), but for the sampling units; in other words, of the sizes n_d y $n_d \times p$ respectively. In addition, we define $\varsigma_d = \boldsymbol{y}_d - \boldsymbol{X}_d \beta$, $\gamma_{|d} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2 / w_{d.}}$, $\boldsymbol{W}_{|d} = diag\left(w_{1|d}, w_{2|d}, \mathrm{K}, w_{n_d|d}\right)$ and $\boldsymbol{w}_{|d} = \left(w_{1|d}, w_{2|d}, \mathrm{K}, w_{n_d|d}\right)^t$, *d=1,.,D*. The sampling estimators of the scores and Fisher matrix elements from the model (9.2) with two-level sampling weights are

$$S_\beta = \frac{1}{\sigma_e^2} \sum_{d=1}^{D} w_d \left(\boldsymbol{X}_d^t \boldsymbol{W}_{|d} \varsigma_d - \frac{\gamma_{|d}}{w_{d.}} \boldsymbol{X}_d^t \boldsymbol{w}_{|d} \boldsymbol{w}_{|d}^t \varsigma_d \right), \quad S_{\sigma_u^2} = -\frac{1}{2} \sum_{d=1}^{D} w_d \frac{1-\gamma_{|d}}{\sigma_e^2} w_{d.} + \frac{1}{2} \sum_{d=1}^{D} w_d \left(\frac{1-\gamma_{|d}}{\sigma_e^2} \right)^2 \varsigma_d^t \boldsymbol{w}_{|d} \boldsymbol{w}_{|d}^t \varsigma_d,$$

$$S_{\sigma_e^2} = -\frac{1}{2} \sum_{d=1}^{D} w_d \frac{w_{d.} - \gamma_{|d}}{\sigma_e^2} + \frac{1}{2\left(\sigma_e^2\right)^2} \sum_{d=1}^{D} w_d \left[\varsigma_d^t \boldsymbol{W}_{|d} \varsigma_d + \frac{\gamma_{|d}\left(\gamma_{|d} - 2\right)}{w_{d.}} \varsigma_d^t \boldsymbol{w}_{|d} \boldsymbol{w}_{|d}^t \varsigma_d \right],$$

$$F_{\beta\beta} = \frac{1}{\sigma_e^2} \sum_{d=1}^{D} w_d \left(\boldsymbol{X}_d^t \boldsymbol{W}_{|d} \boldsymbol{X}_d - \frac{\gamma_{|d}}{w_{d.}} \boldsymbol{X}_d^t \boldsymbol{w}_{|d} \boldsymbol{w}_{|d}^t \boldsymbol{X}_d \right), \quad F_{\sigma_u^2 \sigma_u^2} = \frac{1}{2} \sum_{d=1}^{D} w_d \left(\frac{1-\gamma_{|d}}{\sigma_e^2} \right)^2 w_{d.}^2,$$

$$F_{\sigma_u^2 \sigma_e^2} = \frac{1}{2} \sum_{d=1}^{D} w_d \left(\frac{1-\gamma_{|d}}{\sigma_e^2} \right)^2 w_{d.}, \quad F_{\sigma_e^2 \sigma_e^2} = \frac{1}{2\left(\sigma_e^2\right)^2} \sum_{d=1}^{D} w_d \left[w_{d.} + \gamma_{|d} \left(\gamma_{|d} - 2 \right) \right] \text{ and } F_{\beta \sigma_u^2} = F_{\beta \sigma_e^2} = 0.$$

In the previous expressions, the sums $\sum_{d=1}^{D}$ are not evaluated in fact in relation to all areas *d=1,...,D*, rather only in relation to those represented in the sample. *The Fisher scoring algorithm with weights in units and in areas* is obtained by replacing the scores and Fisher matrix elements in (9.3) with their corresponding estimators. If the individual weights and areas are all equal to one, this algorithm coincides with the Fisher scoring without weights. However, if the weights in small areas are only equal to one, this algorithm differs from the Fisher scoring algorithm with weights in the units.

9.2.4 Seeds for the Fisher scoring algorithms

It is advisable to use the following initial estimator values.

$$\beta_0 = \left(\sum_{d=1}^{D} \boldsymbol{X}_d^t \boldsymbol{W}_d \boldsymbol{X}_d - \sum_{d=1}^{D} \tfrac{1}{w_{d.}} \boldsymbol{X}_d^t \boldsymbol{w}_{n_d} \boldsymbol{w}_{n_d}^t \boldsymbol{X}_d \right)^{-1} \left(\sum_{d=1}^{D} \boldsymbol{X}_d^t \boldsymbol{W}_d \boldsymbol{y}_d - \sum_{d=1}^{D} \tfrac{1}{w_{d.}} \boldsymbol{X}_d^t \boldsymbol{w}_{n_d} \boldsymbol{w}_{n_d}^t \boldsymbol{y}_d \right),$$

$$\sigma_{u,0}^2 = \frac{1}{D} \left[\sum_{d=1}^{D} \frac{1}{w_{d.}^2} \boldsymbol{y}_d^t \boldsymbol{w}_{n_d} \boldsymbol{w}_{n_d}^t \boldsymbol{y}_d - 2\boldsymbol{\beta}_0^t \sum_{d=1}^{D} \frac{1}{w_{d.}^2} \boldsymbol{X}_d^t \boldsymbol{w}_{n_d} \boldsymbol{w}_{n_d}^t \boldsymbol{y}_d + \boldsymbol{\beta}_0^t \left(\sum_{d=1}^{D} \frac{1}{w_{d.}^2} \boldsymbol{X}_d^t \boldsymbol{w}_{n_d} \boldsymbol{w}_{n_d}^t \boldsymbol{X}_d \right) \boldsymbol{\beta}_0 \right],$$

$$\sigma_{e,0}^2 = \frac{1}{n - r(\boldsymbol{X})} \left[\sum_{d=1}^{D} \boldsymbol{y}_d^t \boldsymbol{W}_d \boldsymbol{y}_d - 2\boldsymbol{\beta}_0^t \sum_{d=1}^{D} \boldsymbol{X}_d^t \boldsymbol{W}_d \boldsymbol{y}_d + \boldsymbol{\beta}_0^t \left(\sum_{d=1}^{D} \boldsymbol{X}_d^t \boldsymbol{W}_d \boldsymbol{X}_d \right) \boldsymbol{\beta}_0 \right].$$

10 Software

In the EURAREA project, the Spanish research team has developed C++ and SAS/IML software. The 2002 simulations were carried out with C++ and the 2003 simulations with SAS/IML. The sample selection has always been undertaken with C++, as it is a "low level" programming language that is more flexible, allows the memory to be better managed and produces significant gains in the speed of algorithms. However, SAS/IML is a "high-level" programming language, meaning that it has a range of incorporated statistical procedures that can be used directly and simply in the estimation stage and in addition, the national statistics institutes generally use SAS/IML. For this reason, in the EURAREA project the estimation software in small areas should be developed using SAS/IML and not C++.

The main reasons for having produced C++ software and for having carried out simulations in the aforementioned programming language are: the file size contained in the artificial universe for the simulations (2.4 Gb in the case of Spain), the difficulty in extracting samples with complex sampling designs and the need to carry out simulation experiments with a high number of repetitions. In this sense, the 2002 simulations were carried out with 10,000 repetitions extracted from a complete sample in each repetition. Except in the case of the Fisher scoring algorithms described in section 9, the software was not developed in a closed and ready to use way. The programming strategy was not aimed at the production of software, rather at obtaining maximum efficiency gains in the simulation experiments. This means, for example, that within each repetition each time an amount is calculated, the partial calculations made are not repeated. The 2003 simulations were done with 500 repetitions and the process was structured in two parts. In the first part, 500 samples were randomly extracted using C++. In this case, sub-routines were produced for different sampling designs. In the second part, SAS/IML was used to sequentially treat the 500 samples.

Additional information on the implementation of the Fisher scoring algorithms, both in the case of using unit weights and unit and area weights, can be found in the documents mentioned in the previous section.

11 Simulations carried out in 2002

The simulation experiments carried out in order to value the estimators rest on the basis of selecting independent samples, calculating the parameter estimations in the small areas and comparing them with the known population values.

Given the quality of operations to be carried out in the simulations undertaken to estimate each of the three parameters researched, it was decided to start with the most simple sampling design but respecting the sampling sizes described in section 6. In this way, the first independent samples extracted from the Spanish universe in EURAREA corresponded to a sampling scheme with equal probabilities and later on, stratified samples with equal probabilities in each strata were obtained.

Firstly, a test with the registers in the Comunidad Valenciana was carried out and 10,000 unstratified samples were selected and another 10,000 stratified samples. All of these were subsequently processed in order to evaluate the estimators on a provincial and regional level. When the estimators were based on or assisted by models, the model was fitted with sampling data obtained in the Autonomous Community.

Later on, this work was extended to the whole Spanish universe in EURAREA and the number of processed samples was 2,000, both in terms of unstratified and stratified samples. Estimates were only calculated on a provincial level and, as a consequence, the models were fitted with sampling data from the whole universe.

We also wanted to analyse the most convenient way of fitting the models, in other words, if it was more beneficial to estimate a model for each region or to fit a model to the whole universe being researched. The estimation of household income appears to be suitable for carrying out a comparison of the results obtained after applying both types of fitting. In terms of the standard estimators, the best results were obtained when the estimate of the model parameters were based solely on the sampling data from the Autonomous Community.

The following table lists the estimators evaluated in this case.

Table 11.1. Estimators used in the estimation of income

Estimators	Observations
1 Direct (with N_d known +Horvitz-Thompson)	Standard estimator 1 (DIRECT).
2 Direct (with N_d estimated)	
3 Post-stratified	Qualitative variables *A*, *B* or *C*.
4 Basic synthetic	Qualitative variables *A*, *B* or *C*.
5 Regression synthetic	*APES*409 as covariable.
6 GREG synthetic	*APES*409.
7 GREG1	*APES*409. Standard estimator 2 (GREG).
8 BLUP version (Best Linear Unbiased Predictor) of GREG1 estimator	APES409.
9 EBLUE1	APES409. Standard estimator 6 (EBLUPA).
9s Synthetic from EBLUE1	APES409. Standard estimator 3 (SYNTHA).
10 EBLUP version (Empirical BLUP) of EBLUE1	APES409.
11 Compound dependent on sample size (SSD1)	Qualitative variables *A, B* or *C*. Result of combining estimators 3 and 4.
12 Compound dependent on sample size (SSD2)	APES409. Result of combining estimators 7 and 6.
13 Compound dependent on sample size (SSD3)	Qualitative variables *A, B* or *C*. Result of combining estimators 2 and 4.
14 GREG2	APES409 and APES412. Standard estimator 2 (GREG).
15 BLUP version of GREG2 estimator	APES409 and APES412.
16 GREG3	Qualitative variables *A, B* or *C,* together with APES409 and APES412. Standard estimator 2 (GREG).
17 BLUP version of GREG3 estimator	Qualitative variables *A, B* or *C,* together with APES409 and APES412.
18 Fay-Herriot	*Taxable income tax as* covariable. Result of combining estimators 2 and 18s. Standard estimator 7 (EBLUPB).
18s Synthetic Fay-Herriot	*Taxable income tax* as covariable. Standard estimator 4 (SYNTHB).

Where the ancillary variables mentioned in the table are:

- *Qualitative variable A*. A variable derived from the *APES403* variable (Type of household), which groups together households in 6 types according to whether the household is single person, is made up of only two adults, of one adult and one or more children, of two adults and one or more children, of three adults and one or more children and other types of household.
- *Qualitative variable B*. A variable derived from the *APES208* individual variables (Relation with activity) and the *APES211* (Socio-economic situation) variables in the household components. Thus, the households are grouped together in 4 types according to whether none of the members are employed, some members of the household are employed but none works in a specific activity (agriculture, military,...), idem but only one member works in the aforementioned activities or finally idem but only 2 or more members work in the aforementioned activities.

- *Qualitative variable C*. A variable derived from the *APES211* variable (Studies at more complete levels), which groups together households in 4 types according to whether all adults in the household have finished secondary education, only 50% or more, less than 50% but at least one, or none.
- *APES409*. Size of household.
- *APES412*. Useful area of dwelling in square meters.

Tables 11.2 and 11.3 present the results obtained for the standard estimators ($ARE_d = ARB_d + 100$).

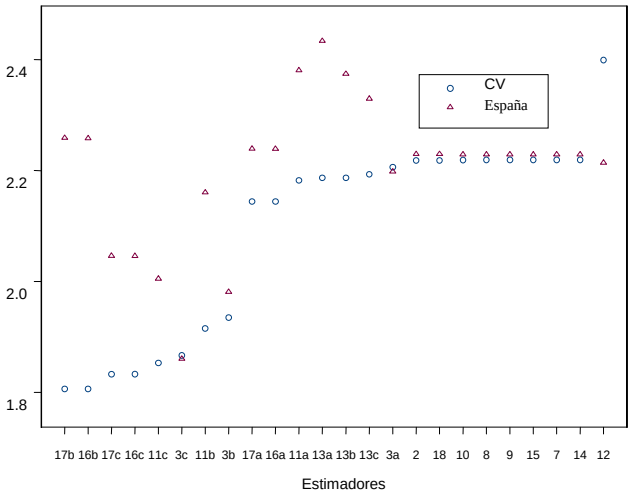
Table 11.2. Provincial estimators in the Comunidad Valenciana when only sa -mpling data from the region is used to fit the models.

	Alicante		Castellón		Valencia		Mean	
Estimator	ARE_1	$RMSE_1$	ARE_2	$RMSE_2$	ARE_3	$RMSE_3$	$\overline{ARE}$	$\overline{RMSE}$
Direct	99,986	1987	99,976	3,105	100,034	1,563	99,999	2,218
GREG1	99,989	1982	99,981	3,113	100,039	1,563	100,003	2,219
GREG2	99,989	1982	99,981	3,113	100,039	1,563	100,003	2,219
EBLUPA	99,988	1982	99,980	3,113	100,039	1,563	100,002	2,219
SYNTHB	94,001	6,147	99,182	1,635	105,931	6,121	99,705	4,634
EBLUPB	99,986	1987	99,976	3,105	100,034	1,563	99,999	2,218

Table 11.3. Provincial estimators in the Comunidad Valenciana when sampling data from the whole universe are used for fitting the models.

	Alicante		Castellón		Valencia		Mean	
Estimator	ARE_1	$RMSE_1$	ARE_2	$RMSE_2$	ARE_3	$RMSE_3$		
Direct	99,923	1,929	100,060	3,173	100,060	1,588	100,014	2,230
GREG1	99,933	1,901	100,070	3,188	100,057	1,598	100,020	2,229
GREG2	99,933	1,901	100,070	3,188	100,057	1,598	100,020	2,229
EBLUPA	99,932	1,901	100,070	3,188	100,057	1,598	1,598	2,229
SYNTHB	98,223	1,906	103,636	3,708	110,688	10,716	104,182	5,443
EBLUPB	99,923	1,929	100,060	3,173	100,060	1,588	100,014	2,230

The next graph represents the numerical values of $\overline{\text{RMSE}}$:



Graph 11.1 $\overline{\text{RMSE}}$ of the estimators for the estimation of income using only data from the Comunidad Valenciana (CV) and those from the whole universe (Spain).

Finally, it is important to note that in 2002, the definition of the estimated parameter relating to ILO unemployment was the proportion of the economically active population that was unemployed. However, this definition was modified in order to carry out the 2003 simulations given that the economically active population is, in general, an unknown quantity and the definition given in section 5 was adopted.

12 Simulations carried out in 2003

In November 2002, the participants of the EURAREA project had a meeting at which professor Tim Holt was present as an expert on the topic of small area estimators. He suggested the need for all participants to carry out simulations to try the standard estimators under similar work conditions to order to validate the comparison of results between the different countries.

Unfortunately, the homogenization of simulations process also implies limitations in terms of the number of questions to be researched, as not all countries have the same resources available. Inevitably therefore, the standard estimator simulations avoided the analysis of some issues which, outside the project context, are undoubtedly of general interest.

12.1 EVALUATION OF THE STANDARD ESTIMATORS

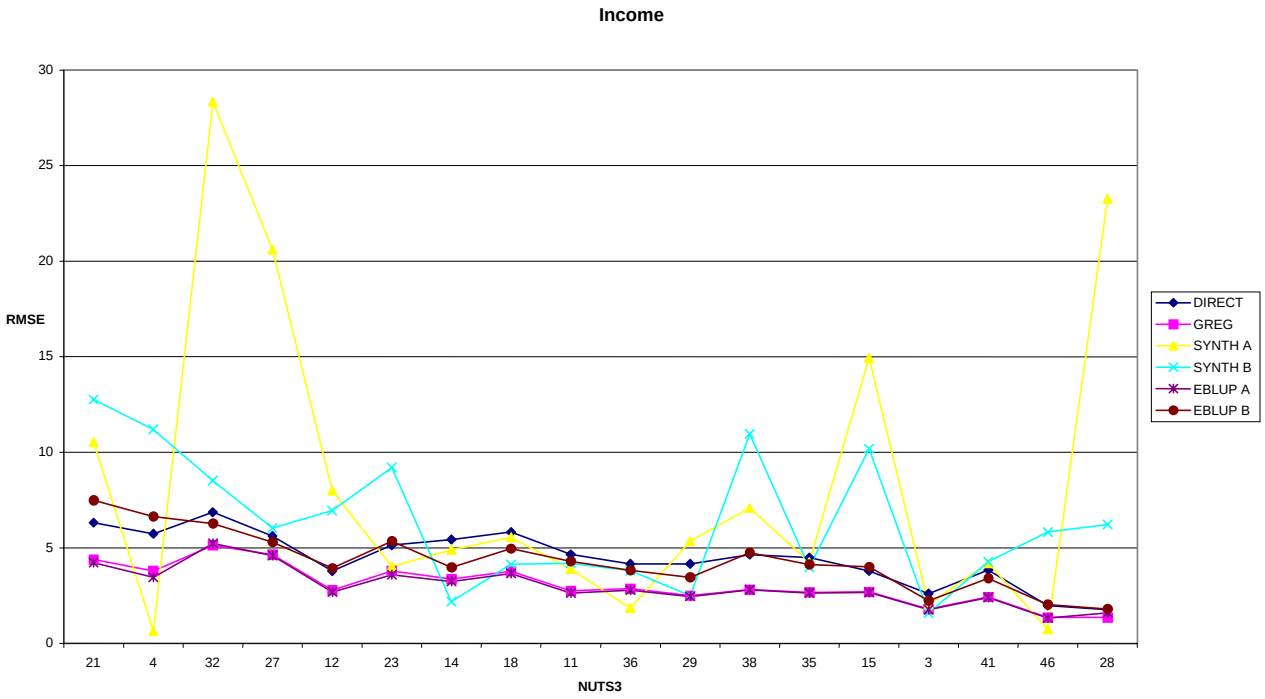
Having discussed at length the best way of implementing "standard simulation" in each participating country in order to obtain maximum comparability in the results, the following decisions were taken:

- The definitions of the population parameters given in section 5 were adopted.
- All countries used the same set of ancillary variables in the models, with small variations in a few cases.
- In each country, the selection of samples was as similar as possible to the selection method currently applied for the estimation of the parameters being researched.
- The models were fitted with the data from the complete sample.
- In order to obtain comparable results, the Office for National Statistics in the UK (ONS), as project coordinator, urged all countries to use the software developed by its team, which uses SAS language and which allows the standard estimators to be calculated and its mean squared errors to be estimated.
- The number of repetitions was 500 in each simulation experiment and in each country.

As a result, 500 samples similar to the APS and another 500 samples similar to the LCS (see section 6) were extracted independently from the Spanish universe in EURAREA. In terms of the first of these samples, estimates of ILO unemployment were obtained and with regards the second set of samples, household income and composition were estimated. The models were fitted using the complete sample and the standard estimators were applied in order to obtain provincial estimates (NUT3) and for the EURAREA-regions (NUT 4 provisional). In addition, the ONS software was used, ignoring the sampling weights when fitting the models.

The following graphs show the results relating to RMSE_d for provincial and EURAREA-region estimates.

Graph 12.1. Mean RMSE values for EURAREA-regions taken 10 in 10 and ordered in ascending sampling size.



Graph 12.2. RMSE of provinces ordered in ascending sampling size.

It can be clearly seen in both graphs that the sampling errors decrease as the sampling sizes increase, especially in the case of the direct estimator and GREG. For the EURAREA-regions, the smallest areas considered in these experiments, the performance of the standard estimators is similar to the mean by which the sample size in the area grows, but if the area has few observations in the sample, the synthetic estimators behave erratically compared with the other estimators. In general, the best performance can be seen in the GREG and EBLUPA estimators.

12.2 ESTIMATION OF ILO UNEMPLOYMENT UNDER THE MODIFIED APS SAMPLING DESIGN

In all the experiments mentioned up until now, the calculation of the estimators assisted by or based on models has been undertaken without taking into consideration the sampling weights for the model fitting.

However, within the issue of complex designs, there are two points to research in relation to the sampling weights that require the carrying out of new simulation experiments. The two issues to be researched are:

- The performance of the estimators when the models are fitted with sampling weights.
- The development, use and evaluation of a two-level sampling weight system, in other words, a system in which both the sampling units and the small areas are allocated sampling weights.

With the aim of meeting these objectives, the research is focussed on the estimation of ILO unemployment. On the other hand, it is well-known that if the sampling weights are equal, their use or not in the fitting of models is irrelevant. Thus, we also decided to increase the variability of the sampling weights in the APS type design used up until now in order to obtain the unemployment estimate.

To do all this, the APS type design described in section 6.1 was modified. Therefore, from this new perspective, the sampling units in the first stage (census sections) were selected with probabilities that were equal instead of proportional to size. In this way, the selection probability of an individual depends on the census section to which it belongs and not only on the stratum, meaning that the sampling weights of individuals are much more heterogeneous that before.

For the calculation of provincial and regional estimators, 500 independent samples were selected with this new design and all the sample data were used to estimate the models. In the model fitting process, the sampling weights appeared in different ways, as described below:

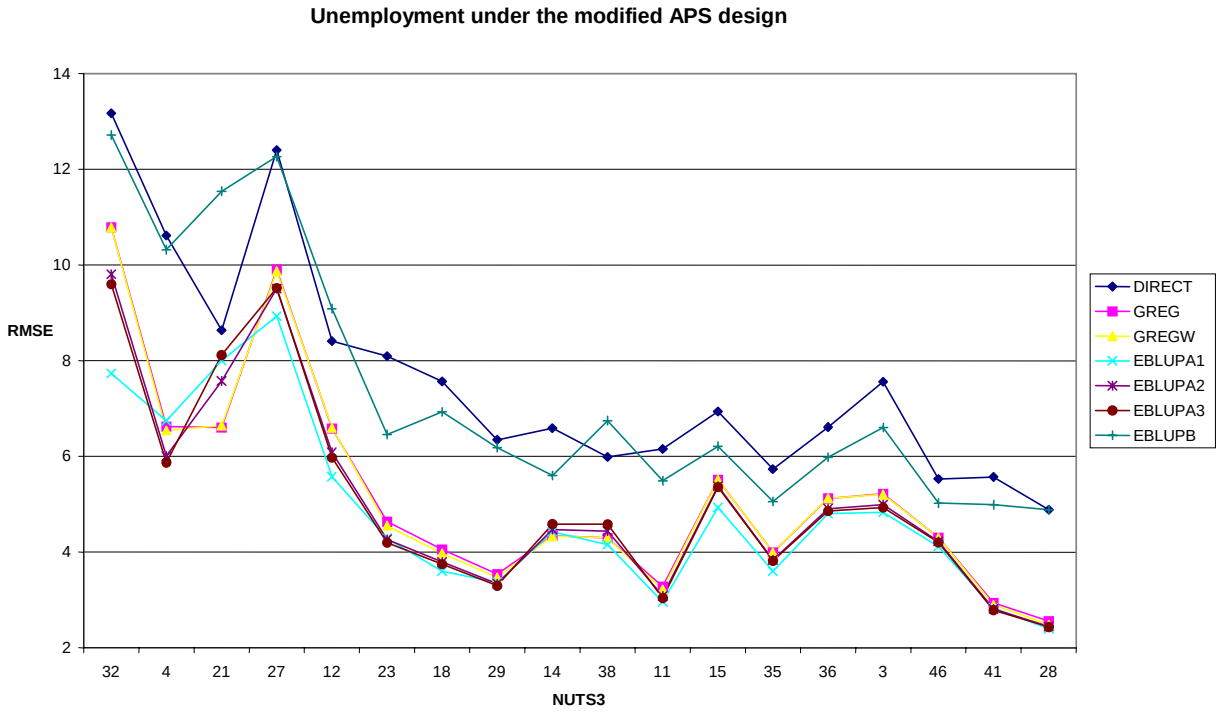
- The DIRECT, EBLUPB, SYNTHB and SYNTHC estimators were calculated as in the case of the standard estimators, in other words, without using the individual sampling weights to estimate the model parameters.
- The GREG estimator was similar in two ways: fitting the model with weights (GREGW) and without weights (GREG). This last method is the same as that used in the standard simulations.
- The EBLUPA and SYNTHA estimators are calculated for 3 different estimation methods relating to model A:
    - Method 1: as in the standard simulations (EBLUPA1 and SYNTHA1)
    - Method 2: using individual sampling weights (EBLUPA2 and SYNTHA2)
    - Method 3:using the two-level weights system developed by the UMH (EBLUPA3 and SYNTHA3)

For the provinces, in general, the best performance corresponds to the GREG and EBLUPA estimators, whereas the synthetic estimators are erratic and the RMSE values very high. In relation to the weights system used, if we focus our

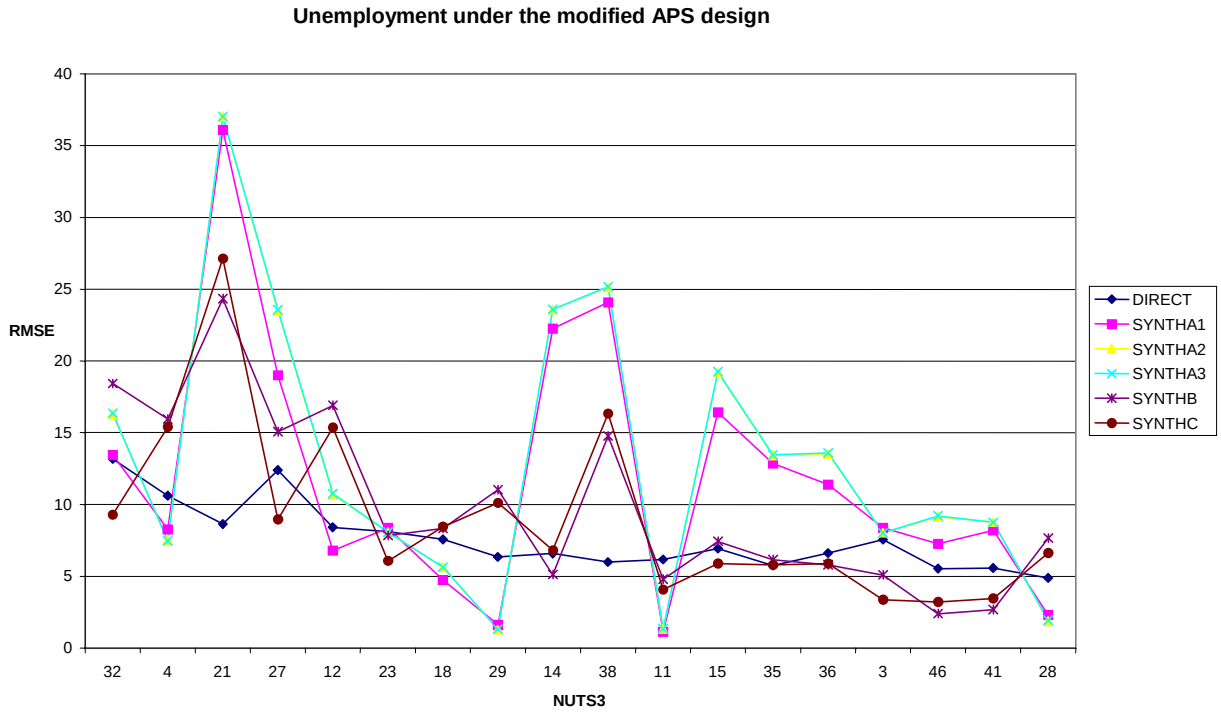
attention on the different EBLUPA estimators calculated, method 2 provides slightly better results.

For the EURAREA-regions, the performance of the synthetic estimators is significantly better than the other estimators, as the sampling sizes in the others are generally very small. As with the standard simulations, the sampling errors decrease as the sampling sizes grow, both in terms of the provinces and the EURAREA-regions. On the other hand, there are no significant differences between the performance of the GREG and GREGW estimator, except a slight difference at the EURAREA-reion level.

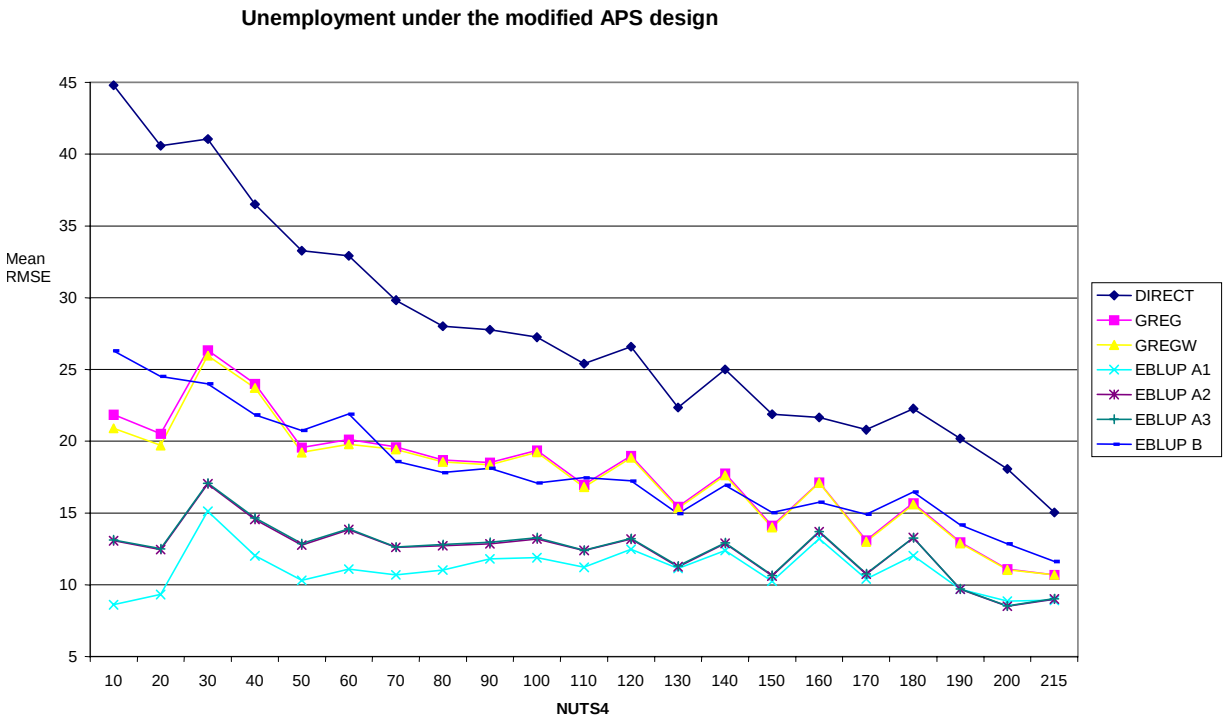
The following graphs show the results:



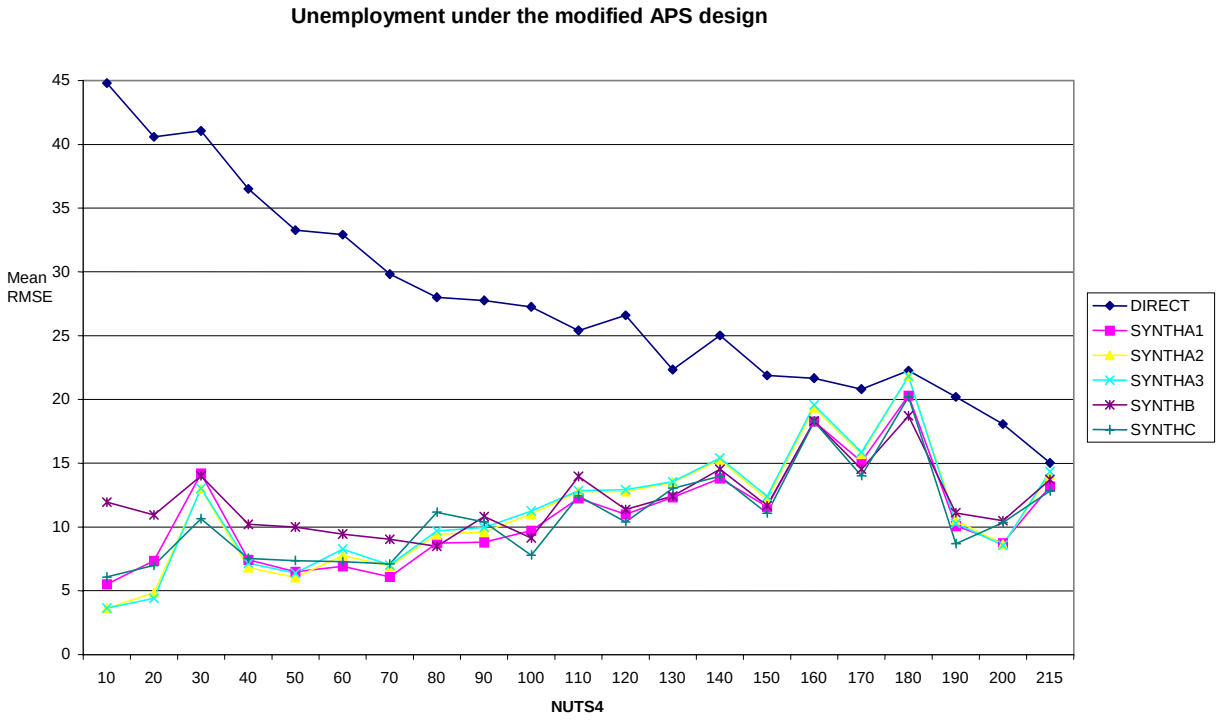
Graph 12.3. Province RMSE ordered in ascending sampling size.



Graph 12.4 Province RMSE ordered in ascending sampling size



Graph 12.5. Mean values for EURAREA-region groups taken 10 in 10 and for RMSE ordered in ascending sampling size.



Graph 12.6. Mean RMSE values for EURAREA-region groups taken 10 in 10 and ordered in ascending sampling size.

12.3 ESTIMATION OF INCOME USING TAXABLE INCOME TAX AS THE EXPLANATORY VARIABLE IN THE MODELS.

In the last few years, the statistical system has been strengthened by it collaboration with the AEAT for the use of tax sources for statistical purposes. In particular and in the EURAREA project context, the AEAT has provided broken down data at small geographic area level, maintaining the confidentiality of the tax data.

In the standard simulations carried out previously, this ancillary information has not been used as it is not available in any of the other countries participating in the project. However, once the comparative analysis of the standard results obtained by all participants has ended, it was agreed that each country could carry out any extra work that it considered necessary to meet its own objectives.

In our case, interest was aimed at covering this gap and we began to work again with the income estimate and the 500 LCS type samples selected during the standard simulations.

In order to derive the income estimations, taxable income tax broken down at small area level (province or EURAREA-regions) was used as a covariable in the models considered, which in turn, were fitted in relation to the whole sample with the following procedures.

- Method 1: as in the standard simulations
- Method 2: using the sampling weights.

In order to analyse the impact of ancillary information on the small area estimators, we're going to compare the results from Method 1 with those from the standard simulations. Note that both experiments are based on the same set of samples (the 500 LCS type sample) and the estimators have been calculated using the same methods (without using weights to fit the models), however the covariables used are different. The ancillary variables used in the non-standard simulations (Method 1) are more realistic in the sense that the INE has access to this information in the real world. We are going to call this set of variables $A_2 = \{X_1,..,.X_6\}$ where:

X_1 = *APES409* (Size of household)

X_2 = *Qualitative variable B* (Socio-economic situation of the household derived from direct variables)

X_3 - X_6 = *taxable income tax* according to different income sources (total, pensions, unemployment and agrarian)

On the other hand, the set of ancillary variables used in the standard simulations we shall call $A_1 = /\xi_1$,..,.ξ_6 } *where:*

= *APES405* (Number of employed people in household)

ξ_2 = *APES409* (Size of household)

ξ_3 = *APES412* (Useful area of dwelling m^2)

ξ_4 = *Sum of ages of male members*

ξ_5 = *Sum of ages of female members*

ξ_6 = *Variable derived from the qualitative variables C* (1 if no adults in the household have finished secondary education and 0 if the opposite)

they are less realistic in the sense set out previously.

The tables below summarise the mean RMSE and EMSE values:

Table 12.1. Mean values of ARB relative bias in relation to provinces as small areas

Experiment	DIRECT	GREG	EBLUPA3	EBLUPB
A_2 / with income tax	-0.097	-0.070	-0.244	-0.121
A_1 / without income tax	-0.097	-0.013	0.084	-0.226

Table 12.2. Mean values of RMSE relative error in relation to provinces as small areas

Experiment	DIRECT	GREG	EBLUPA3	EBLUPB
A_2 / with income tax	4.489	3.205	3.036	4.147
A_1 / without income tax	4.489	3.049	2.982	4.323

Table 12.3. Mean values of ARB relative bias in relation to EURAREA-regions as small areas

Experiment	DIRECT	GREG	EBLUPA3	EBLUPB
A_2 / with income tax	-0.082	0.068	0.472	0.492
A_1 / without income tax	-0.082	0.046	0.762	-0.008

Table 12.4. Mean values of RMSE relative error in relation to EURAREA-regions as small areas

Experiment	DIRECT	GREG	EBLUPA3	EBLUPB
A_2 / with income tax	12.347	9.711	6.610	8.571
A_1 / without income tax	12.347	8.941	7.692	7.806

On a province level, our conclusion is that the use of taxable income tax is recommendable, particularly with the EBLUPA estimators assisted by area models, as a reduction is observed in both bias and relative error. On a EURAREA-region level and surely due to the increase in the number of small areas, the use of taxable income tax is recommendable, particularly with the EBLUPA estimators assisted by household models with a random area factor, as a reduction is observed in both bias and relative error.

In any case, it is difficult to set out definitive conclusions and we shouldn't forget that the variable being researched, household income, is an imputed variable, as it is not collected in the 1991 Population Census and as a result, there is a potential source of error that is not under complete control in the experiments analysed.

12.4 ESTIMATION OF INCOME WITH NON-RESPONSE

With the aim of expanding the research from the effect of sampling weights, we include in the work of the Spanish EURAREA team a study on the impact of the use of *informative sampling designs* (the weights depend on the objective variable) with estimators assisted by or based on models.

To do this, in each of the 500 LCS type samples a non-response mechanism was introduced that was correlated with household income. Thus, the probability of households responding decreased as their income increased. Having generated the non-response in each sample according to this mechanism, the final sampling weights obviously depended on the household's income.

In order to derive the income estimations, taxable income tax broken down at small area level (province or EURAREA-regions) was used as a covariable with models, which in turn were fitted in relation to the whole sample with the following procedures:

- Method 1: as in the standard simulations
- Method 2: using sampling weights.

In other words, the same covariables were used as before, but with different weights.

In order to study the effect of informative weights, we're going to compare the results of these two experiments. The tables below summarise the mean ARB and RMSE values:

Table 12.5. Mean values for relative ARB bias in relation to provinces as small areas

Experiment	DIRECT	GREG	EBLUPA3	EBLUPB
A_2 / without sampling weights	-0.083	-0.063	-4.131	-1.382
A_2/ with sampling weights	-0.083	-0.061	0.054	0.806

Table 12.6 Mean values for relative RMSE error in relation to provinces as small areas

Experiment	DIRECT	GREG	EBLUPA3	EBLUPB
A_2 / without sampling weights	5.031	3.708	4.983	4.707
A_2/ with sampling weights	5.031	3.692	3.397	5.394

Table 12.7. Mean values for relative ARB bias in relation to EURAREA-regions as small areas

Experiment	DIRECT	GREG	EBLUPA3	EBLUPB
A_2 / without sampling weights	-0.254	-0.127	-2.754	-1.586
A_2/ with sampling weights	-0.254	0.086	0.868	2.116

Table 12.8. Mean values for relative RMSE error in relation to EURAREA-regions as small areas

Experiment	DIRECT	GREG	EBLUPA3	EBLUPB
A_2 / without sampling weights	14.077	11.106	7.023	9.401
A_2/ with sampling weights	14.077	11.183	7.375	7.350

In general, our conclusion is that the use of informative sampling weights reduces the bias of estimators, both on a provincial and EURAREA-regions level. On the other hand, in relation to relative error, this reduction is not as significant except in the case of the EBLUPA estimator based on an area model and applied to EURAREA-region type small areas.

13 References

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under selection model. *Biometrics*, **31**, 423-427.

Morales D. y Molina I. (2002). Small area mixed linear models for normal variables. Not published.

Morales D., Molina I. y Santamaría L. (2002). A comparative study of small area estimators with applications to surveys on income and living conditions. Not published.

Pfefferman, D., Skinner, C.J., Holmes, D.J., Goldstein, H. y Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, **60**, 23-40.

Prasad, N.G.N. y Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163-171.

Saralegui, J. y Herrador, M. (2003). El problema de la estimación en áreas pequeñas para la estadística oficial. Recientes progresos en España. *27 National Statistics and Operational Research Congress.*

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference. A Prediction Approach*. John Wiley. New York.