10. The representativeness of the 1999 Spanish FADN survey

Ricardo Mora [1] Carlos San Juan [2] and José Eusebio de la Torre [3], Carlos III University Spain

10.1 Introduction

The Farm Accountancy Data Network (FADN) is a European Union (EU) farm-level survey widely used to analize policy impacts and the effect of changes in the level of producers protection. The survey is frequently used by researches and policy makers for calculating production, costs, and income of commercial agricultural holdings in the EU among other purposes. It provides valuable information on areas such as productive orientation and localization of the holdings, data which cannot be known from the results of the economic accounts of the agrarian sector. Guaranteeing that the survey is representative of the reality of the sector is a major priority.

Since 1986 the Spanish program of the National Farm Accountancy Network (RECAN from the Spanish acronym) has been integrated into the community FADN, adopting its methodology so that results obtained are, in principle, comparable to those from other EU countries.

However, sample sizes vary greatly among member states so that in order to obtain meaningful comparisons between aggregated results, the computation of sampling weights becomes an essential methodological issue. Furthermore, the complexity of the information gathered as well as the need to consider the active participation of the head of the holding implies that random selection is not a realistic sampling procedure. Both the rate of no response to the survey (in reality a simplified accounting of the holding) and the costs of obtaining the collaboration of a random sampling of agricultural business men and women would be quite high. For this reason, in practice, the sampling method has traditionally been based on the stratification of the field of observation.

The purpose of this work is to develop and implement a methodology to study how representative the 1999 Spanish RECAN sample is by using the recently released 1999 Agricultural Census data. We first propose a sampling design which maintains the advantages of the traditional stratification procedure and deals with well-known reported sampling problems in the RECAN survey. Then, we analyse how representative the actual 1999 RECAN sample is with respect to the 1999 Agricultural Census.

[1] Universidad Carlos III Madrid, Department of Economics Department. ricmora@eco.uc3m.es
[2] Universidad Carlos III Madrid, Department of Economics Department. csj@eco.uc3m.es
[3] Universidad Carlos III Madrid, Department of Economics Department. ejtorre@eco.uc3m.es

10.2 The 1999 Agricultural Census for Spain

The source of population data used in the sampling design is the Spanish Agricultural Census for 1999. The target population in RECAN is the set of the commercial holdings [1], which is a sub-set of the target population in the 1999 Agricultural Census. The strata are defined by three characteristics: region, type of farming, and economic size. In order to design the sampling plan it is also necessary to focus our attention on the results on one or more characteristics or economic variables.

The design of the stratification involves unavoidable *a priori* decisions as to the definition of the cell. While the regional stratification by the Spanish NUTS-2 regions (Comunidades Autónomas) is desirable for sociological reasons, the definition of cells by farming type or economic size can be more difficult. In the following, we accurately define the cell that will be used in the design of the RECAN sample and the 1999 Agricultural Census variables used in the design of the plan for the RECAN sample.

10.2.1 Size classes

The concept of Standard Gross Margins (SGM) is used to determine the economic size of the farms, which are expressed as European Size Units (ESU). The SGM of a crop or livestock item is defined as the value of output from one hectare or from one animal less the cost of variable inputs required to produce that output. The SGM is calculated from three year averages on a regional level for more than 82 crops and livestock items. For 1999, the period of reference for the SGM corresponds to 1995, 1996 and 1997.

The economic size of the farms is calculated by multiplying the farm's hectares and Livestock Units by the corresponding regional SGM. The result, the Total Gross Margin (TGM) of the farm whether in pesetas, ECUs or Euros, can be expressed as ESUs by keeping in mind the fixed relation between the ECU and the ESU.[2]

In the typology used by the European Community, ten farm sizes are considered.[3] Given the target of defining cells as finely as possible, an attempt should be made to minimize the number of observations within each cell without jeopardizing the global

[1] The definitions of commercial holdings, and therefore of the definition of the target population will be covered in more detail below.

[2] The value of the ECU by ESU has changed over time to reflect inflation, passing from 1,000 in 1980 to 1,200 in 1984. The exchange rate for conversion Pesetas/Euros is 167,119.

[3] These are: I (fewer than 2 ESUs), II (from 2 to fewer than 4), III (from 4 to fewer than 6), IV (from 6 to fewer than 8), V (from 8 to fewer than 12), VI (from 12 to fewer than 16), VII (from 16 to fewer than 40), VIII (from 40 to fewer than 100), IX (from 100 to fewer than 250), and X (more than 250 ESUs).

Table 10.1 Main variables by economic size units (ESU) 1999 agricultural census

(ESU)	Farms			Gross Total Margin			Utilised Agricultural Area (UAA)			Labour			Livestock		
	Number	a)	b)	ESU	a)	b)	Ha	a)	b)	AWU	a)	b)	TAU	a)	b)
[0,2)	781131	47.10		595220	3.83		1281246	5.26		234299	19.93		333635	2.22	
[2,4)	249442	15.04		718749	4.63		1157533	4.75		119911	10.20		369612	2.45	
[4,8)	224405	13.53	35.73	1284410	8.27	9.03	1974202	8.11	9.01	150644	12.81	18.34	800602	5.32	5.58
[8,16)	182756	11.02	29.10	2074915	13.35	14.59	3248939	13.35	14.83	181460	15.43	22.09	1633141	10.84	11.37
[16,40)	150483	9.07	23.96	3716363	23.92	26.13	6365798	26.15	29.06	210973	17.94	25.68	3284446	21.81	22.88
[40,100)	52458	3.16	8.35	3118411	20.07	21.92	5337838	21.93	24.37	121174	10.31	14.75	3555044	23.60	24.76
[100,250)	14068	0.85	2.24	2083549	13.41	14.65	2999781	12.32	13.69	71729	6.10	8.73	2730045	18.13	19.01
[250,∞)	3831	0.23	0.61	1947617	12.53	13.69	1978650	8.13	9.03	85534	7.28	10.41	2354219	15.63	16.40
Total	1658574	100.00	100.00	15539235	100.00	100.00	24343987	100.00	100.00	1175724	100.00	100.00	15060745	100.00	100.00

a) % overall farms; b) % farms under 4 ESU.

representation of the sample. However, an ambitious stratification design is in danger of not being implementable due to the high demands of collaboration which FADN imposes on the farmers. A compromise must be reached to solve this dilemma. Based on previous studies on the Spanish RECAN, a simplified farm size typology will be adopted in this study so that only eight size classes are considered: Extremely small (Fewer than 2), Verysmall (From 2 to fewer than 4), Small (From 4 to fewer than 8), Medium low (From 8 to fewer than 16), Medium (From 16 to fewer than 40), Medium high (From 40 to fewer than 100), Large (From 100 to fewer than 250), and Very large (More than 250).

The main difference with respect to the FADN farm size variable consists in the merging of types III and IV and also types V and VI. In our opinion, this simplification does not have a negative effect on the accuracy of the sample, as these categories are very densely populated and we assume that sampling can be very cost-effective in these strata. Table 1 shows the census distribution by size of the national totals of (i) number of farms, (ii) Total Gross Margin (TGM) in ESUs, (iii) Utilised Agricultural Area (UAA) in hectares, (iv) farm labour, in Agricultural Work Units (AWU[1]), and (v) livestock, in Total Livestock Units (LU).[2]

In the 1999 Agricultural Census, the total number of farms ammounts to 1,658,574 farms which exploit some 24,343,987 ha of UAA and 15,060,745 LUs, providing employment for 1,175,724 AWUs and generate a TGM of 15,539,235 ESU (the equivalent of 18,647,082,000 euro). Close to half of the farms (47.10%) do not reach 2 ESU of TGM, which in Spain and some other member countries is still the minimum TGM requirement defining a commmercial farm. Because of this, all farms appearing in the lowest size group in table 10.1 are currently not considered commercial. This large group of 'non-commercial' farms provides less than 4% (3.8%) of total TGM and exploit less than 6% of total UAA and less than 3% of LU, yet they absorb nearly one fifth (19.93%) of AWU, the majority of which is family labour.

More than half of the UAA (61.43%) and of LU (56.25%) fall within the group of medium-sized farms (8-16; 16-40 and 40-100 ESU). The TGM of this type of farm ammounts to more than half of the national total (57.34%). Finally, 43.68% of workers are employed by these farms. The large or very large farms (more than 100 ESU) use more than 20% of the UAA (20.45%) and more than 30% of LU (33.76%); their TGM represents 25.94% of the national total and they employ 13.38% of the total work employed by agriculture (measured in AWU). The smaller farms (from 4 to 8 ESUs) use around one tenth of land (8.11% of the UAA) and barely exceed 5% of the livestock (5.32% of the LU). Their TGM represents little more than 8% of the total (8.27% of the TGM). Scarcely more than one tenth of total work (12.81% of AWU) employed in agriculture is employed by this group of small farms.

[1] An AWU is equivalent to the work of one person working full-time during one year.

[2] Agricultural holdings without an assigned type of Farming are excluded from the Table. In this case, the census assigns a Farming Type of 9999 and a TGM of 0. The north African enclaves are also excluded from the analysis. There are 48 farms in Ceuta and Melilla, of which 30 have a Farming Type of 9999 which implies a TGM of 0 and only 18 have a positive TGM.

10.2.2 Distribution of farms by regions

Table 10.2 shows the distribution of farms by Spanish Autonomous Communities according to the 1999 Spanish Agricultural Census. The three most extensive regions (Andalusia, Castilla-La Mancha and Castilla y León) are the ones which contribute the most to the total number of farms. This is, in part, due to differences in land size across regions. However, land size, number of farms, and other variables, do not always follow a simple relation. For example, Galicia, with less than 3% of national UAA, employs more than 16% of AWU despite only contributing 4% to the national TGM. Behind this lack of proporcionality lies the structure of agriculture in Galicia, characterized by a large number of small farms with an important participation of low-productivity underemployed labour.

Table 10.2 Main variables by region. 1999 agricultural census

Region	Farms		Gross Total Margin		Utilised Agricultural Area (UAA)		Labour		Livestock	
	Number	%	ESU	%	Ha	%	AWU	%	TAU	%
Galicia	239194	14.42	622545	4.01	674071	2.77	191189	16.26	1460863	9.70
Asturias	42117	2.54	177460	1.14	389056	1.60	40093	3.41	417028	2.77
Cantabria	16690	1.01	100606	0.65	190362	0.78	16720	1.42	311072	2.07
País Vasco	34294	2.07	148221	0.95	245268	1.01	27532	2.34	215658	1.43
Navarra	23229	1.40	324285	2.09	536931	2.21	16300	1.39	323502	2.15
La Rioja	16911	1.02	223127	1.44	151659	0.62	12846	1.09	113874	0.76
Aragón	73724	4.45	1038839	6.69	2133391	8.76	45953	3.91	1706099	11.33
Cataluña	73762	4.45	1110062	7.14	1016158	4.17	75103	6.39	2788476	18.51
Baleares	18857	1.14	124328	0.80	220335	0.91	12968	1.10	121721	0.81
Castilla-León	156982	9.46	2009867	12.93	5304079	21.79	99666	8.48	2375654	15.77
Madrid	14893	0.90	128437	0.83	350186	1.44	7993	0.68	144646	0.96
Castilla-La Mancha	182266	10.99	1674968	10.78	4433950	18.21	92095	7.83	1127163	7.48
Comunidad Valenciana	214367	12.92	1243210	8.00	705735	2.90	83992	7.14	569420	3.78
Murcia	55386	3.34	766947	4.94	449483	1.85	53427	4.54	541641	3.60
Extremadura	105796	6.38	966165	6.22	2806672	11.53	67856	5.77	1191396	7.91
Andalucía	355638	21.44	4622126	29.74	4673890	19.20	284816	24.22	1557653	10.34
Canarias	34468	2.08	258041	1.66	62760	0.26	47175	4.01	94882	0.63
Total	1658574	100.00	15539235	100.00	24343987	100.00	1175724	100.00	15060745	100.00

10.2.3 Type of Farming

Type of farming is defined following economic, not physical, principles. The EU has established a typology for types of farming which includes 17 principal classifications, subdivided into 50 special classes.

For a better stratification of the Spanish agriculture, we felt it convenient to group some of the principal EU types of farms into a single class while splitting others into different classes. As a result, 18 main types of farming are analyzed. The farming types and

their corresponding EU codes are as follows: *Specialist cereals, oilseed and protein crops* (EU codes 1310, 1320, and 1330), *General field cropping* (1410, 1420, 1430, 1441, 1442, and 1443), *Specialist horticulture (not greenhouse)* (2011, 2013, 2021, 2023, 2031, 2033, and 2034), *Specialist horticulture (greenhouse)* (2012, 2022, and 2032), *Specialist vineyards* (3110, 3120, 3130, 3141, 3142, and 3143), *Specialist fruits and citrus fruits* (3211, 2312, 3213, 3220, and 3230), *Specialist olives* (3300), *Various permanent crops combined* (3400), *Specialist dairying* (4110, and 4120), *Cattle - combined* (4210, 4220, 4310, and 4320), *Sheep and goats* (4410, and 4430), *Sheep, cattle and other grazing livestock* (4420, and 4440), *Pigs* (5011, 5012, and 5013), *Fowl* (5021, 5022, and 5023), *Specialist combined granivores* (5031, and 5032), *Mixed cropping* (6010, 6020, 6030, 6040, 6050, 6061, and

Table 10.3 Distribution by region and type of farming of gtm. 1999 agricultural census

TYPE OF FARMING	TOTAL		PERCENT OF EACH AUTONOMUS REGION OF AGGREGATE MBT BY TYPE OF FARMING																
	GTM (ESUs)	%	Galicia	Asturias	Cantabria	País Vasco	Navarra	La Rioja	Aragón	Cataluña	Baleares	C. y León	Madrid	C. La Mancha	C. Valenciana	Murcia	Extremadura	Andalucía	Canarias
Specialist cereals, oilseed and protein crops	2114953	13.61	0.18	0.00	0.01	0.67	3.42	0.55	13.01	5.23	0.21	33.51	1.47	19.38	1.09	0.57	5.05	15.64	0.01
General field cropping	1230144	7.92	1.52	0.04	0.04	1.58	2.91	1.72	4.08	2.13	0.40	19.40	0.45	9.12	0.47	2.65	14.35	38.49	0.65
Specialist horticulture (not greenhouse)	537925	3.46	1.17	0.13	0.06	0.35	1.48	1.75	1.70	5.18	1.46	2.43	1.49	9.54	10.87	29.88	5.84	23.32	3.35
Specialist horticulture (greenhouse)	564295	3.63	1.42	0.25	0.46	0.39	0.24	0.04	0.19	2.47	0.41	0.29	0.30	0.94	4.05	13.73	0.35	65.47	8.99
Specialist vineyards	625010	4.02	1.28	0.00	0.00	2.33	5.31	14.30	4.50	10.66	0.06	5.59	0.68	31.97	9.52	3.10	5.54	4.75	0.42
Specialist fruits and citrus fruits	1723851	11.09	0.37	0.30	0.01	0.17	0.37	0.90	4.99	9.81	1.15	0.48	0.01	0.49	43.23	14.31	1.00	17.34	5.08
Specialist olives	2256424	14.52	0.00	0.00	0.00	0.00	0.04	0.03	0.34	2.17	0.30	0.16	0.37	4.00	1.27	0.14	2.66	88.51	0.00
Various permanent crops combined	583791	3.76	0.51	0.03	0.03	0.12	2.00	2.75	6.70	11.10	2.29	1.75	0.95	14.63	18.17	4.69	7.80	19.50	6.99
Specialist dairying	649001	4.18	39.82	12.20	8.43	4.81	3.15	0.36	0.68	5.25	1.71	8.50	1.39	2.46	0.74	0.77	1.16	7.79	0.80
Cattle - combined	517812	3.33	15.35	9.97	5.24	3.35	1.62	1.13	4.22	6.34	0.33	24.51	2.75	4.42	0.99	0.32	12.08	7.27	0.11
Sheep and goats	620562	3.99	0.82	0.20	0.20	1.74	2.93	1.51	6.74	3.64	0.34	28.96	1.09	25.45	1.93	2.28	9.18	11.29	1.70
Sheep, cattle and other grazing livestock	325188	2.09	4.58	5.96	2.90	4.20	3.32	0.58	6.80	3.26	0.65	27.14	1.56	9.85	1.60	0.73	14.33	11.93	0.60
Pigs	1106373	7.12	5.37	0.07	0.12	0.14	3.23	0.54	25.06	19.57	0.53	14.92	0.36	6.24	6.31	7.83	2.88	6.48	0.35
Fowl	160729	1.03	10.30	0.39	0.46	1.55	1.32	1.00	10.11	24.65	0.46	11.92	1.93	13.58	7.61	1.06	1.15	10.22	2.30
Specialist combined granivores	65975	0.42	10.69	1.17	0.82	2.39	1.34	0.98	9.60	35.97	0.40	6.13	0.05	9.94	8.09	1.85	1.24	8.43	0.92
Mixed cropping	1175114	7.56	3.14	0.19	0.02	0.35	3.67	2.17	5.40	4.92	1.76	5.35	1.34	20.66	4.19	3.83	8.66	32.89	1.46
Mixed livestock	386279	2.49	13.08	1.85	0.09	0.77	0.77	0.41	3.87	13.15	1.92	13.16	0.21	8.51	1.15	2.32	20.28	17.97	0.51
Mixed crops and livestock	895812	5.76	4.47	0.71	0.11	0.73	1.35	0.49	8.22	10.44	1.40	26.65	0.56	12.33	2.82	2.35	11.58	15.26	0.51

6062), *Mixed livestock* (7110, 7120, 7210, 7220, and 7230), *and Mixed crops and livestock* (8110, 8120, 8130, 8140, 8210, 8220, 8231, and 8232).

Table 10.3 shows the distribution of TGM by farming type in each Autonomous Region. Well known geographical specialization features can be highlighted by close scrutiny of the table: the concentration of olive trees in Andalusia, vineyards in Castilla-La Mancha, fruit and citrus in Valencia, horticulture in Andalusia, Murcia and Valencia, dairy cattle in Galicia, Asturias and Cantabria, and pigs in Castilla y León, Cataluña and Aragón. In general, regions tend to concentrate in one or two farming types: in Andalusia, for example, the highest percentage is found in olives, in Valencia and Murcia in fruits and citrus. Farming in Cataluña however, is more diversified among the defined farming types. Similar conclusions are obtained if we restrict the analysis to farms with TGM of 4 ESUs or more.

10.3 Sample size and sampling errors

Sample size can either be fixed exogenously, and therefore be considered as a constant, or it can be determined depending on a target on the significance level and/or a relative sampling error. As will be shown later, due to logistical limitations, the size of the sample for the RECAN survey cannot exceed a certain number, n, which should therefore be considered a pre-set constant. Nonetheless, we exploit the relation between sample sizes and relative sampling errors to distribute the sample of size n into two sumbsamples. In turn, this partition will provide us with a flexible tool for harmonising all of the restrictions that must be kept in mind when carrying out the sampling plan for RECAN.

For simplicity, we will develop a formula of a general nature. Consider a population of size N, divided into H cells of size H_h each. In this population we can take data from a statistical variable which we will call Y. Y_{hi} will be the value of Y for the element of the population *i-th* in cell h. Taking a sample of the population, we can estimate a characteristic of the distribution of the variable Y in the population.

We will concentrate on the total aggregate of variable Y across the population, **Error! Objects cannot be created from editing field codes.**. Given that the population is divided into cells, if we have a sample of size n, then we have n_h elements of the sample which pertain to the *h-th* cell. Therefore **Error! Objects cannot be created from editing field codes.**. In order to estimate Y_T, we use the weighted mean: **Error! Objects cannot be created from editing field codes.** where **Error! Objects cannot be created from editing field codes.** is the probability that the element of the *i-th* population of cell h pertains to the sample of nh units obtained in the h-th cell. If we supposed that **Error! Objects cannot be created from editing field codes.**[1], meaning that the elements of the population are selected using a simple random sampling without repositioning within cell h, then **Error!**

[1] As will be seen later, this quotient receives the name of elevation factor.

Objects cannot be created from editing field codes. where **Error! Objects cannot be created from editing field codes.** is the average sample of Y in the *h-th* cell. **Error! Objects cannot be created from editing field codes.** is an unbiased estimator with variance equal to **Error! Objects cannot be created from editing field codes.**, where **Error! Objects cannot be created from editing field codes.** is the population quasi-variance of the variable Y in the cell *h-th*.

Our objective is to calculate the sample size so that **Error! Objects cannot be created from editing field codes.**, where *r* is the maximum admissible relative error and **Error! Objects cannot be created from editing field codes.** is the level of significance. If we suppose that **Error! Objects cannot be created from editing field codes.** follows a normal distribution or, as is the case, that the sample size is large, then we can calculate the optimal sampling size *n* based on the earlier formula for a given value of *r* and **Error! Objects cannot be created from editing field codes.**.

To see this, note that since **Error! Objects cannot be created from editing field codes.** follows a normal distribution, then **Error! Objects cannot be created from editing field codes.** follows a distribution N(0,1). Therefore **Error! Objects cannot be created from editing field codes.** and **Error! Objects cannot be created from editing field codes.**.

As mentioned earlier, Z is a *N(0,1)*, so that if *FN(0,1) (x)* is the function of distribution of a random variable with distribution *N(0,1)*, then **Error! Objects cannot be created from editing field codes.**, which implies that **Error! Objects cannot be created from editing field codes.**. Let us define **Error! Objects cannot be created from editing field codes.** so that **Error! Objects cannot be created from editing field codes.**. Substituting **Error! Objects cannot be created from editing field codes.** for its value:

Error! Objects cannot be created from editing field codes.

Calling **Error! Objects cannot be created from editing field codes.**, then:

Error! Objects cannot be created from editing field codes.

Solving for *n* from the earlier equation, gives the desired sample size:

Error! Objects cannot be created from editing field codes.

Now, if the design feature of the sample is of minimum variance then **Error! Objects cannot be created from editing field codes.**. Thus, substituting in the formula for *n*, the expression for the size of the sample simplifies to:

Error! Objects cannot be created from editing field codes.

Now we can customize the earlier formula for the chosen clusters:

Error! Objects cannot be created from editing field codes. (1)

Where:

Error! Objects cannot be created from editing field codes. is the quantile of order **Error! Objects cannot be created from editing field codes.** in the distribution N(0,1).

$V(Y_{jhk})$ is the population quasi-variance of the variable Total Gross Margin, within the cluster defined by the size class k (k = 1, 2, …, M_{jh}) the farming type h(h = 1, 2, …, Tj) within the region j (j = 1, 2, …, 17).
Y_T is the total population of the variable Total Gross Margin.
r is the maximum admissible relative error.
N_{jhk} is the number of farms in the field of observation of class k, within the farming type h and belonging to region j.

Equation (1) can be used to determine the approximate sample size needed to estimate the variable of reference (the national Total Gross Margin) with a fixed degree of precision. It must be born in mind that this is a valid approximation to the extent to which the sampling is random within each cell.

With a stratified random sampling and large samples, the maximum relative error in the estimation of the total of the variable studied in region j, with Farming Type h and size class k, is approximated by the formula from the normal distribution:
Error! Objects cannot be created from editing field codes. (2)
in which:
r_{jhk}: denotes the maximum relative error which can be expected, with a probability **Error! Objects cannot be created from editing field codes.**, in the estimation of the total of variable Y in the cell jhk.
N_{jhk}: denotes the number of farms in the field of observation belonging to a region j (j = 1, 2, …, 17), farming type h(h = 1, 2, …, T_j) and size class k (k = 1, 2, …, M_{jh}).
Error! Objects cannot be created from editing field codes.is the sampling rate within the region j (j = 1, 2, …, 17), farming type h(h = 1, 2, …, Tj), and size class k (k = 1, 2, …, M_{jh}).
n_{jhk}: denotes the size of the sample in cell jhk.
$V(Y_{jhk})$: denotes the population quasi-variance of the variable being studied in the cell jhk.
Y_{Tjhk}: is the total of the variable being studied in the cell jhk.
Error! Objects cannot be created from editing field codes. is the quantile of the order **Error! Objects cannot be created from editing field codes.** in the distribution $N(0,1)$.

Equation (2) is quite useful because it provides an estimator for the relative errors by stratum for a given sampling design. Since the representativeness of the sample by regions has a special relevance for sociological reasons, it is interesting to adapt Equation (2) for the case of the estimation of regional TGM. In this case, the sampling error in the region j is **Error! Objects cannot be created from editing field codes.**and the total for the country is **Error! Objects cannot be created from editing field codes.**where Y_{Tj} represents the total of the variable of the study in the region j and Y_T is the total of the variable of the study for the country as a whole.

The above formula assumes that the estimator of interest is unbiased. This is unwarranted if some cells remain without surveyed observations. In this case, the above formula would underestimate the real relative sampling error, even in those cells for which there are data. This is relevant in our study case since, as it will be seen below, various cells will be considered irrelevant in our sampling design, imposing an additional restriction on the field of observation in the RECAN. If we are merely interested in the new field of observation, then the above formulas are still valid.

However, if we are interested in the field of the RECAN before considering those cells to be irrelevant and we would like an evaluation of the discrepancy between the estimations of the RECAN and the parameter to be estimated since we considered those cells irrelevant, then it is necessary to correct for the bias in the estimator. This formula will also be valid for evaluating the real RECAN sample in case there are cells for which there are no observations.

For simplicity, we will develop a formula of a general nature. Its application to each case of interest is straightforward. Suppose that there are *M* cells for which there are no observations in the cell and the elevation factor is therefore not defined. This implies that the estimator **Error! Objects cannot be created from editing field codes.** is biased. Let Y_{T1} be the average of the estimator of Y_T in the presence of the bias and let Y_{T2} be the aggregate of the variable in the cells for which there are no observations, then: $Y_T = Y_{T1} + Y_{T2}$. We are interested in knowing the relative sampling error as a function of the various groups defined by size, region, and farming type. In the case, since the average of **Error! Objects cannot be created from editing field codes.** is not Y_T, from the definition of the relative sampling error it can be shown that

Error! Objects cannot be created from editing field codes. (3)

The second term of the right-hand part of the above equation indicates the percentage of the variable of interest which lies out of reach ($Y_{T2} = Y_T - Y_{T1}$). This equation tells us that if there are cells with no observations in the sample, then the relative sampling error is equal to the relative error that is committed in the cells with sample plus the proportion of the variable of interest found in the cells with no sample.

10.4 The sampling design methodology

In this section, the methodological choices needed to implement the sampling design and the evaluation of the actual 1999 RECAN survey are described. First, we define the field of observation of the RECAN. In particular, we propose a minimum size for a farm to be considered commercial and we also identify strata to be considered irrelevant in the sampling. Finally, we look at restrictions on the sampling plan imposed for practical purposes and set up the algorithm that aims at guaranteeing that all restrictions are fulfilled.

10.4.1 Minimum Farm Size

The target population of the FADN is the grouping of commercial farms in the EU of at least one hectare and/or those with less than one hectare which commercialise a specific quantity (which differs among Member States) of their production.

In order to define commercial farms, the Commission follows directives specified in Regulation 79/65/EEC and the subsequent modifications and adopts a pragmatic approach based on the economic significance that the farm has for its owner. In particular, a farm is considered commercial when it is large enough to provide a sufficient income level to maintain the farmer and his or her family. Consequently, to be considered commercial, the farm must exceed a minimum economic size expressed in ESU. In Spain, the established

limit is 2 ESU, which implies that all farms with a TGM of more than 2 ESU have been included in the target universe of the RECAN.[1] For 1999, this limit is questionably low.

The 1989 Spanish Agricultural Census shows that farms with fewer than 2 ESU made up 63.40% of the total, their TGM was 9.5% and the UAA was 11.4%. As far as labour and livestock items are concerned, the figures were 26.40% and 4.9% respectively. This means that in the ten years between 1989 and 1999, there has been a significant reduction in the importance of farms with fewer than 2 ESU in Spain. This structural change in the behaviour of Spanish agriculture suggests the need to revise the lower limit for economic size of a commercial farm for two reasons. First, an increase in prices in the Spanish economy has meant that the purchasing power of income from farms with fewer than 2 ESU has dropped dramatically, raising serious doubts as to whether it is realistic to suppose that farms of 2 ESU can guarantee a sufficient economic level to sustain an average Spanish family, even in rural areas. Second, general price increases in the products have affected the intertemporal comparisons of TGM. Farms which were formerly considered non-commercial in 1989 will now fall into the category of commercial farms using the new SGMs calculated from the 1995, 1996 and 1997 RECAN surveys.

The census information reveals that the proportion of farms with fewer than 4 ESU in 1999 (62.14%) is very similar to the proportion of farms with less than 2 ESU in 1989 (63.4%). As far as the TGM is concerned, the corresponding figure for farms with fewer than 4 ESU in 1999 is 8.46% whilst the figure for farms with less than 2 ESU in 1989 is 9.5%. For the UAA, the comparison would be 10.01% compared with 11.4%. In the case of AWU and LU, the figures are 29.43% compared to 26.2 and 4.67% compared to 4.9% respectively.

In the 1999 Agricultural Census, accountancy data was collected from 1,658,574 farms.[2] A lower limit of economic size of 2 ESU of TGM would exclude 781,131 farms and the RECAN universe would consist of 877,443 farms with a TGM of fewer than 2 ESU. The sample size of the 1999 RECAN was 8,233 farms[3] of which 25 had a TGM strictly inferior to 2 ESU and, hence, only 8,208 farms would be considered commercial. This figure implies an average weight in the 1999 RECAN sampling of 106.9.

If the lower limit of the TGM for a commercial farm is 4 ESU, the 1999 RECAN universe would include only 628,001 farms. The 1999 RECAN survey has 8,080 farms with 4 or more ESU, which would imply an average sampling weight of 77.72.[4] This new

[1] This threshold varies greately across countries: Holland has established its limit at 16 whilst in Belgium is 12 ESUs. Austria, Denmark, Finland, France, Germany, Luxembourg, Sweden, and the UK (except N. Ireland) have all set the limit at 8 ESUs. Northern Ireland has the limit at 4 and Greece, Ireland, Italy and Spain at 2. Finally, Portugal has the limit at 1.

[2] Farms without a Farming Type (code 9999 in the Farm Return) and farms in Ceuta and Melilla were excluded.

[3] This quantity compares favourably to sample sizes from many other Member States. The average number of observations in the FADN sample over the last few years in Belgium has been 1,196 observations, in Denmark 2,117, in Germany 5,827, in Greece 4,834, in France 7,568, in Ireland 1,202, in Italy 16,235, in Luxembourg 278, in Holland 1,516, in Austria 2,085, in Portugal 2,932, in Finland 1,007, in Sweden 827 and in the United Kingdom 3,648.

[4] The average weight of a farm in the FADN sample in Belgium over the last few years has been 37, in Denmark 25, Germany 50, Greece 100, France 54, Ireland 106, Finland 48, Sweden 49 and in the United Kingdom 37. Note how, in general, the more restricted the definition of a business is, the lower its average weight tends to be in each observation.

lower limit leads us to a new definition of a commercial farm. As shown in table 10.1, the 628,001 farms with more than 4 ESU of TGM contribute more than 89% to the national TGM, to national UAA and to LU, and nearly employ 70% of AWU.[1]

For these reasons, and with the previously mentioned goal to reduce the number of cells under consideration and, thus, handling the management of the sample without harming its representativeness, we restrict the field of observation by subtracting from the 1,658,574 farms in the Census those without a farming type and farms whose TGM is less than 4 ESU.

In this way, the target universe is made up of 628,001 farms with more than 4 ESU of TGM according to the 1999 Agricultural Census. Although the proportion of farms with fewer than 4 ESU is very high in all regions (with values ranging from 44.62% in Cataluña to an astonishing 84.38% in Galicia), this class of farm contributes little to regional TGM and employs a small proportion of the available resources, with the exception perhaps, of Galicia, where it accounts for 25.7% of TGM.

10.4.2 Irrelevant cells

In order to define the field of observation, it is also necessary to identify cells considered to be irrelevant for the purpose of the survey. We proceed following previous studies and practices in the RECAN design and assume a farming type to be irrelevant in a region if its TGM represents less than 1% of the regional TGM. All of the cells which fulfill this condition are outside the field of observation of the RECAN. Following this criterion, the farming types considered irrelevant by region are presented in figure 10.1.

It is important to point out that the exclusion of irrelevant farms from the sample does not substantially alter the coverage rate of this group's commmercial farms.

Finally, table 10.4 shows the number of cells by Autonomous Region and the number of observations in each cluster that fulfil the three following conditions: (a) each observation represents one commercial farm, i.e. a farm with TGM in excess of 4 ESU; (b) all cells are relevant, i.e. they represent farming types which contribute to regional TGM in excess of 1 percent; and (c) cells for which the 1999 Agricultural Census does not record any farm are also excluded. To sum up, the proposed 1999 RECAN universe is formed by 612,921 farms distributed across 1,195 strata or cells.

[1] Of course, the proportion is lower if we consider salaried labour since it is concentrated in farms with greater economic size.

Region	Farming Types exclude as irrelevant
Galicia	Cereals, oilseed and protein crops (except rice); Horticulture (except greenhouse); Horticulture in greenhouse; Vineyards; Fruit and Citrus; Olives; Various permanent crops combined; Sheep and Goats; Mixed cropping
Asturias	Cereals, oilseed and protein crops; Various crops combined; Horticulture (except greenhouse); Horticulture in greenhouse; Vineyards; Fruit and Citrus; Olives; Various permanent crops combined; Sheep and Goats; Pigs; Fowl; Various granivores combined; Mixed cropping
Cantabria	Cereals, oilseed and protein crops; Various crops combined; Horticulture (except greenhouse); Horticulture in greenhouse; Vineyards; Fruit and Citrus; Olives; Various permanent crops combined; Sheep and Goats; Fowl; Various granivores combined; Mixed livestock; Mixed crops and livestock
Basque Country	Horticulture (except greenhouse); Various permanent crops combined; Mixed livestock
Navarra	Horticulture in greenhouse; Olives; Fowl; Various granivores combined; Mixed livestock
La Rioja	Horticulture in greenhouse; Olives; Sheep and Cattle + various grazing livestock; Fowl; Various granivores combined; Mixed livestock
Aragon	Horticulture (except greenhouse); Horticulture in greenhouse; Olives; Dairy cattle; Various granivores combined
Cataluña	Sheep and Cattle + various grazing livestock
Baleares	Vineyards; Fowl; Various granivores combined
Castilla y León	Various granivores combined; Horticulture in greenhouse; Fruit and Citrus; Olives; Various permanent crops combined; Fowl; Various granivores combined
Madrid	Fruits and Citrus; Various granivores combined; Mixed livestock
Castilla - La Mancha	Horticulture in greenhouse; Fruit and Citrus; Various granivores combined
Valencia	Various crops combined; Dairy cattle; Beef and mixed cattle; Sheep and Cattle + various grazing livestock; Various granivores combined; Mixed livestock
Murcia	Olives; Dairy cattle; Beef and mixed cattle; Sheep and Cattle + various grazing livestock; Fowl; Various granivores combined
Extremadura	Horticulture in greenhouse; Dairy cattle; Fowl; Various granivores combined
Andalusia	Vineyards; Beef and mixed cattle; Sheep and Cattle + various grazing livestock; Various granivores combined
Canarias	Cereals, oilseed and protein crops; Vineyards; Olives; Beef and mixed cattle; Sheep and Cattle + various grazing livestock; Various granivores combined; Mixed livestock

Figure 10.1 Irrelevant farming types by autonomous region

Table 10.4 Number of cells in the proposed 1999 sampling universe by autonomous region

Region	Number of Cells	Number of Farms in 1999 Census
Galicia	53	35,546
Asturias	31	11,987
Cantabria	28	6,328
País Vasco	73	7,161
Navarra	75	11,650
Rioja	69	8,981
Aragón	78	37,403
Cataluña	101	40,475
Baleares	82	5,428
Castilla-León	65	76,115
Madrid	85	5,074
Castilla-La Mancha	90	65,801
Valencia	70	64,262
Murcia	71	19,727
Extremadura	84	33,252
Andalucía	78	174,152
Canarias	62	9,579
Total	1,195	612,921

10.4.3 Restrictions in the sampling design

Given the sample size and one method for assigning a quota to each cell, it is straightforward to compute an *unrestricted* sampling design. In practice, this design will not be fully implementable for a variety of reasons which we discuss below. However, an implementable sampling design can be obtained after imposing several restrictions to the design process. We propose an iterative algorithm which has two stages at any given step. In the first stage, we obtain *unrestricted* quotas for each cell. In a second stage, we obtain *restricted* quotas by sequentially imposing a pre-set number of restrictions/conditions that our design should follow.

Looking for a compromise between the best sampling design and its implementability, a number of restrictions have been considered. These restrictions are based on experience obtained through data collection in earlier editions of the RECAN, impositions from the EU to ensure quality of the data, and budgetary constraints. The first restriction to consider is that the sample size is bounded due to budget considerations.

Restriction 1

The total number of farms in the sample must be less than or equal to 9,500.

The experience of earlier editions of the RECAN indicates that farms of a large economic size are difficult to sample. Accountancy agencies have, over the years, been unable

to obtain collaboration for many large size farms. The year under study, 1999, was no exception. This situation leads us to propose the following.

Restriction 2

The number of farms in the RECAN sampling plan with more than 250 ESU must be fewer than or equal to 350.

In order to ensure the quality of estimations, the FADN requires that the elevation factor of the farms should not be greater than 500.

Restriction 3

The ratio between the number of existing farms and the number of farms in the RECAN sample must be less than or equal to 500 in each stratum.

In previous RECAN sampling designs, two additional restrictions had been included ex-post: (i) the minimum number of farms in a cell is set to 5, and (ii) the maximum number of farms in a cell has to be 50. The minimum of 5 was chosen in order to ensure a representative minimum for all cells being considered whilst the choice of 50 as upper limit was taken on the assumption that a greater sample size would not improve significantly the precision of the estimations but could increase the cost of the survey. These restrictions are not without serious problems. First, in a significant number of cases there are less than five farms in a cell. Then, the design would imply to sample farms in excess of all existing farms in the cell, so that the sampling quota would be over 100%. Second, the maximum bound does not prevent this problem to appear also in large cells so that it is perfectly possible to have less existing farms than those assigned in the design and still satisfy the upper boundary. An obvious solution to these problems is to impose the following restriction in the design.

Restriction 4

The number of farms in the RECAN sample must always be smaller or equal to the number of existing farms.

However, the limits of 5 and 50 still imply a practical problem as they are incompatible with Restriction 2. To see this, consider that there is a total of 77 relevant cells with strictly fewer than 5 existing farms and with more than 250 ESU, which add up to a total of 169 existing and, thus, assigned farms. On the other hand, there are 89 cells with 5 or more existing farms with more than 250 ESU. If there must be a minimum of 5 farms in the sampling plan in each cell, then we have, at least, a total of 5 x 89 = 445 farms for this group. This means a total of 614 farms to be surveyed, a number which violates Restriction 2.

A direct solution consists of relaxing the lower limit, at least for farms of more than 250 ESU. According to this, it is straightforward to compute the number of farms in the sample with more than 250 ESU if the minimum number of farms by cell is lowered successively to: 4, 3, 2, or 1. If we set 4 as a lower bound: 105 + 4 105 = 525. If 3 is the lower bound: 72 + 3 116 = 420. If it is 2: 28 + 2 138 = 304. Finally, with 1, we have 28.

It turns out that Restriction 2 is only compatible with a minimum number of 2 or 1. For the sake of improving accuracy at the lowest cost, we set the minimum number to 2 and extend it to those cells in which there are farms with fewer than 250 ESU to avoid a bias in the representativeness of the cells with assigned numbers between 2 and 5. Then, the last restriction becomes.

Restriction 5

The number of farms in a cell must belong to the interval:
[min{2, existing farms}, min{50, existing farms}].

Therefore, restrictions 1 to 5 are applied to the *unrestricted* design in the following section to develop the RECAN sampling plan.

10.4.4 Choice of quota assignment method

Several methods can be used to assign a quota to each cell or stratum. Amongst them, we will consider the following two: the minimum variance method and the proportional method. The first one identifies the assignment which minimizes the variance of the target estimation, in our case, national TGM. In general, if the objective of the RECAN were to estimate the aggregate TGM only, the method of minimum variance would be the 'ideal' method. Although the estimation of the aggregate TGM is an important goal, the RECAN survey also pursues other goals. For example, RECAN gives detailed information on labour requirements at farm level. As productivity differs widely between farms of different size, it seems that a minimum variance method based on national TGM will underemphasize the need to sample strata with low production levels and large labour requirements.

It is theoretically possible to propose an assignment method which minimizes the variance of, for example, a linear combination between TGM and AWU. However, in practice, this strategy will always reduce to an *ad-hoc* proposal of a trade off between the variance of TGM and the variance of AWU. These *ad-hoc* choices can become even more controversial if we are willing to consider more than two variables, as it is the case for the RECAN survey.

The proportional method assigns quotas such that the ratio between the number of farms in the sample and the number of farms in the population is approximately constant. This is, therefore, an extremely straightforward method that does not require ad-hoc assumptions on the objectives of the survey. In fact, the proportional method may prove very effective if the variables of interest, no matter which ones and how many, show little variance across cells. However, it becomes less and less reliable if the dispersion of variance across cells is very important. In particular, if variance changes according to size, so that large farms have less variance in, say farm TGM, the proportional method will tend to overemphasize sampling in precisely those strata for which RECAN has traditionally been less successful: the very large farms.

For these reasons, we propose an intermediate simple solution which consists of assigning, for each cell, the average between the minimal variance quota and the proportional quota.

10.4.5 The algorithm for quota calculations

It is possible to compute *unrestricted* assigned quotas using both the minimum variance method and the proportional method. An *unrestricted* average method would simply consist of the average between the previously computed two assignments in each cell using the 1999 Agricultural Census data. This procedure, however, does not satisfy restrictions 1-5 and the result is, therefore, not satisfactory.

An ideal way to ensure that restrictions 1-5 are met would be to find the assignment that minimizes the variance of the estimator of the aggregated TGM, subject to restrictions 1-5. This strategy is not implementable as it implies a complex nonlinear problem with an enormous number of variables (one for each defined cell). Furthermore, it does not solve the above-mentioned problem of compromising amongst multiple objectives. For this reason, a different strategic line will be followed so that the sampling plan closely verifies restrictions 1 to 5 and the average method is implemented.

First, the RECAN target population is divided into two subpopulations: farms with strictly fewer than 250 ESU and farms greater than or equal to 250 ESU. These populations shall be referred to as SP1 and SP2, respectively. At this stage, the total sample size is 9,500.

We start by setting a significance level and a relative sampling error for each subpopulation. The level of significance and the sampling error depend on the sample size of the two subpopulations as shown in Equation (1). Thus, we can minimize the admissible relative error subject to restrictions 1 and 2 by choosing a partition of the total sample into the two subpopulations.

Once we have obtained the sample size for SP1 and SP2, we calculate the quotas following the minimum variance and proportional methodology and, using those results we obtain the initial quota using the averaging method.

Of course, restrictions 1 and 2 are already met but restrictions 3, 4 and 5 are not. Given a set of assigned quotas, we are going to implement at any step of the algorithm a sequential testing procedure which gives predominance to restriction 5. We start by checking whether the quota is lower than two. If this is case, we set the quota to 2. If it is not, then we keep the quota unchanged. Then we see whether the quota is smaller than the number of existing farms. In the positive case, we leave the quota unchanged. If the answer is negative, then we set the quota equal to ¼ of the number of existing farms. Finally, we check whether the quota is smaller or equal to 50. If this is the case, we leave it unchanged. In the opposite case, we fix it to 50. The final quota assigned is the resulting quota rounded off to cero decimal digits.

We check whether restrictions 1, 2, 4, and 5 are met. If that is case, the algorithm is finished. Otherwise, we must first ensure that restrictions 1 and 2 are met. When they are not met, we recalculate the sample size of the two populations by using a search grid and marginally changing the different relative sampling errors. In general, the relative sampling errors vary inversely with the size of the sample. Thus we choose the smallest increase or decrease in the relative sampling error so that restrictions 1 and 2 are met in the new partition.

Once we have this new partition, we can procede with the algorithm. Although it is not possible to establish convergence, a fundamental advantage of the proposed algorithm

is that it is very simple to implement. In our application to the design of the 1999 RECAN sample, a solution was reached in the second step. However, note that restriction 3 is never imposed thoughout the algorithm. As will be seen below, after the algorithm was implemented, restriction 3 was satisfied in all but two cells (that is, in 1,193 out of 1,195 cells). We feel that this does not affect the quality of the design as we will argue in the next subsection.

10.5 The sampling design results

10.5.1 Summary of results

The relative sampling errors of the sampling design satisfying the restrictions are 0.43% for SP1 and 15.00% for SP2, resulting in a sample size for the whole population of 9,485 farms and for the subpopulation of 250 ESU or more a size of 349 farms. The relative sampling errors were set to be as near as possible to the maximum of farms permitted for the whole population and the subpopulation SP2. It can be seen that for SP1, the relative sampling error is quite low, while for SP2 it is high. This is due to the fact that 349 farms is a very low number to be able to get a small relative sampling error.

By class size, the Medium class strata receives the highest quota (2,387 farms, 25.17% of the total sample size), while the Very Large class is assigned the smallest quota (349 farms, 3.68% of the total). By region, the plan assigns the highest quota to Andalusia; 1,816 farms (19.15% of the total), followed by Castilla-La Mancha (1,138 farms) and Castilla y Leon (1,095). Inversely, Cantabria is the Region with the lowest quota (125 farms). By farming type, the highest is assigned to *Cereals, oilseed and protein crops* (1,244 farms or 13.12% of the sample), a coherent number given that its contribution to national TGM for commercial farms as a whole in Spain is 14.14%. The smallest quota goes to *Various granivores combined*, with only 41 farms. This is mainly due to the fact that this farming type is only relevant in three Regions: Galicia, The Basque Country and Cataluña.

The average elevation factor is 66.21%. By size classes, the elevation factor averages decrease from 118.98 for the Small class down to 10.98 for the Very Large class. By farming types, the largest average is obtained for *Olives* (218.18). This is due to the fact that the number of farms by cell is restricted to be equal or smaller than 50, and farms in *Olives* tend to vary hugely in size, with the presence of some very large farms with volatile TGM. Although the sampling plan assigns the largest quota available (50), elevation factors still remain high. By region, the elevation factors are on average smaller than 100 except in Valencia (104.79). The lowest average is obtained for Madrid (26.34).

In general, the elevation factors are below the limit of 500 established by the EU. Table 10.5 offers a brief summary of the distribution of the elevation factors by cells. There are 5% of cells with an elevation factor of 1, which means that all farms belonging to these cells must be sampled. Only 1% of the cells show an elevation factor greater than 127.34.

As already mentioned, restriction 3 is not fulfilled in two cells. These cells are: (I) Andalusia, *Olives*, Small (with an elevation factor of 758.16); and (II) Andalusia, *Olives*, Medium low (with an elevation factor of 515.98).

The problem arises simply because the required maximum number of farms is simply too restrictive. There are 37,980 and 25,799 existing farms in the two cells respectively. Thus, restrictions 3 and 5 are incompatible in these two cells. We give priority to restriction 5, but note that given the current sample size, it would always be possible at least to fix quotas of 63 (37,980/500≅76) and 52 (25,799/500≅52) for the two cells. Of course, the effect on the relative errors for national TGM is negligible.

Table 10.5 Distribution percentiles of the elevation factors in the proposed sampling plan by cells

Quantiles (%)	Elevation Factors
1	1
5	1
10	2
25	9
50	25.5
75	64
90	84
95	88.36
99	127.34

10.5.2 Relative sampling errors

One way of evaluating a sampling plan is through the calculation of the relative sampling error, meaning the largest percentage deviation between the estimation of national TGM from its real value given a significance level.

Equation (2) is an approximation which does not take into account the missing information in the irrelevant (*a priori* not surveyed) cells. If we wish to consider the discrepancy due to the irrelevant cells, then Equation (3) is a better approximation. Both approximations will be more accurate when sampling is random within cells and large samples justify the assumption of normality of the TGM estimator. Our opinion is that these equations, in particular, Equation (3) are well justified in this context. The RECAN sample is a random sample within the universe of non-irrelevant farms which are willing to collaborate. The controversy then is simply whether the strata are sufficiently detailed to avoid biases due to lack of collaboration within the cell. As already discussed, this is unlikely to be a serious problem since farm size is a major predictor of lack of collaboration, and size has been partitioned into a significant number of categories. On the other hand, the large overall sample - over 9,000 observations - implies that the normality assumption is perfectly reasonable.

Note that relative sampling errors were set in order to obtain the sampling plan. These errors are established before filtering the quotas by the algorithm. Thus, the new reassignment obtained through the algorithm will result in relative errors different (generally greater) than the pre-assigned ones. We compute theoretical sampling errors for the design

Table 10.6 Relative errors by autonomous region

Region	Relative errors at relevant cells (%)	Relative errors at all cells (%)
Galicia	4.82	8.78
Asturias	5.20	9.21
Cantabria	4.46	8.74
País Vasco	4.02	6.30
Navarra	4.52	6.87
La Rioja	4.35	7.32
Aragón	5.54	7.82
Cataluña	4.62	5.57
Baleares	4.15	5.30
Castilla-León	3.17	5.61
Madrid	4.63	5.39
Castilla-La Mancha	3.89	5.00
Comunidad Valenciana	11.56	14.34
Murcia	17.90	19.73
Extremadura	5.06	6.40
Andalucía	4.60	7.28
Canarias	22.10	24.20
Total	2.09	4.31

Table 10.7 Relative errors by type of farming

Type of farming	Relative errors at relevant cells (%)	Relative errors at all cells (%)
General field cropping	9.34	9.74
Specialist horticulture (not greenhouse)	22.68	27.12
Specialist horticulture (greenhouse)	13.06	15.31
Specialist vineyards	2.45	6.91
Specialist fruits and citrus fruits	9.02	9.92
Specialist olives	6.12	6.48
Various permanent crops combined	8.68	10.41
Specialist dairying	2.21	5.61
Cattle - combined	2.30	11.80
Sheep and goats	2.39	3.14
Sheep, cattle and other grazing livestock	5.82	25.30
Pigs	9.50	9.56
Fowl	15.44	43.36
Specialist combined granivores	13.01	64.32
Mixed cropping	6.03	6.55
Mixed livestock	7.36	10.87
Mixed crops and livestock	2.94	3.01
Total	2.09	4.31

both for only relevant cells and for the entire population of commercial farms. Results are shown in tables 10.6, 10.7, and 10.8.

The precision of the estimate of national TGM is 2.09% for the relevant cells and 4.31% for all commercial farms. By region, Castilla Leon (3.17%) and Castilla-La Mancha (3.89%) present the lowest relative error for relevant cells, whilst Castilla-La Mancha (5.00%) and Baleares (5.30%) for all farms. Canarias shows the greatest relative errors in the two categories (22.10% and 24.20% respectively).

By farming type, the largest errors appear in *Horticulture (except greenhouse)* (22.68% and 27.12%), but it is worth pointing out that errors in *Fowl and various granivores* are very high especially for all cells due to the large number of farms which are considered irrelevant. This fact is related to the low level of importance of these farming types across regions.

Table 10.8 Relative errors by economic size units

Economic Size Units	Relative errors at relevant cells (%)	Relative errors at all cells (%)
[4,8)	1.31	4.05
[8,16)	1.18	3.48
[16,40)	1.33	3.40
[40,100)	1.41	3.48
[100,250)	1.74	3.86
[250, ∞)	14.68	17.13
Total	2.09	4.31

By size class, the errors are moderate except in the case of farms with more than 250 ESU (14.68%), which are obviously paying the price of restriction 2.

10.6 The 1999 RECAN sample: an evaluation

The purpose of this final section is to analyse how representative the actual 1999 RECAN sample is with respect to the 1999 Agricultural Census. To do this, we decompose the errors in national TGM estimation based on the 1999 RECAN sample into either errors due to difficulties in covering sampling quotas (cells without observations) or sampling error.

In order to do so, we must determine how much of the deviation in the TGM estimation carried out by RECAN is due to the lack of farms in some cells in the sample and how much is due to the sampling. We assume that the 1999 RECAN sample was obtained by stratified simple random sampling method without reposition. Therefore, a linear aggregate TGM estimator in Spain is equal to the total of the TGM (calculated from the 1999 RECAN data) in each defined cell times the inverse of the probability that the farm will be selected in the cell (i.e., the elevation factor).

The percentage deviation of the actual 1999 RECAN sample estimate of national TGM from the 1999 Agricultural Census estimate is not negligible: -20.01%. This relative error has its origins in two different sources. The first source is the fact that there are various cells for which there are observations in the 1999 RECAN sample but the elevation factor is undefined, causing a bias in the estimator and increasing the relative error of the estimation. The second source is that the design of the sample necessarily implies a theoretical sampling error, which is reflected in the estimated relative error.

It can be seen that, with the exception of *Olives*, there is an underestimation of the TGM in all categories considered, be they regions, size or type of farming. This is due, as we shall see below, to the lack of observations in numerous cells causing a downward slant in the estimation of national TGM. By Autonomous Region, the estimated TGM of The Balearic islands is 85.34% lower in the RECAN than the value assigned by the 1999 Census, while the estimation which comes closest to the value given in the Census is Andalusia (9.12% lower). By farming types, *Olives* are overestimated by 5.78%, and the rest is underestimated. By economic size, all estimations were below the real value, especially the estimation of TGM of farms of more than 250 ESU (-65.10%).

To quantify which part of the total relative error is due to the selection type which RECAN 1999 followed in choosing the farms, and which part is due to the existence of cells with no observations, we divide the population of surveyed farms into two subpopulations: (i) farms which pertain to cells with a number of farms sampled by the 1999 RECAN strictly greater than zero and, (ii) farms in cells without farms in the 1999 RECAN.

By studying the 1999 Agricultural Census, it is possible to compute national TGM that was beyond the scope of the 1999 RECAN population target and thus, given the stratification and the chosen estimators, the size of the error in the national TGM estimation using the actual 1999 RECAN sample. The percentage of national TGM in this situation is 20.38%. Therefore, if we estimate national TGM for the subpopulation of farms for which there are data in the 1999 RECAN without error, we would still continue underestimating the aggregate TGM for Spain by 20.38%. For the Autonomous Regions, the greatest percentage of TGM which is beyond the scope of the 1999 RECAN is in Baleares (98.83%), while that of Castilla Leon and Andalusia is only 10.26% and 11.66% respectively. By farming type, in *Mixed livestock* and *Fowl*, 61.25% and 60.67% of the TGM were beyond the 1999 RECAN scope, whereas in *Olives*, only 2.13% is left out. As far as the size classes, the 1999 RECAN tends to concentrate in Medium farms (between 8 and 40 ESU), while it leaves out considerable TGM mainly from large farms (more than 100 ESU), and to a lesser degree in small farms (between 4 and 8 ESU).

However, the percentage deviation error of the aggregate TGM for Spain for the farms with sample in the actual 1999 RECAN sample is much lower. At the maximum level of aggregation, the TGM estimator using data from the 1999 RECAN estimates the TGM with relative reliability (0.47%) if it has a sample in the cells. The relative errors for the subpopulation are low in absolute value for all regions. In Aragon, Castilla Leon, Valencia, Murcia and Extremadura the TGM is underestimated, which suggests that there is no bias in the estimator. In absolute value, the greatest relative errors show up in Baleares (31.22%), probably due to the low number of observations collected, Murcia (-18.27%) and Canarias (16.92%) surely due to the great variability in TGM in these regions. In relation to farming type, the greatest relative error is in *Horticulture (except greenhouse)*,

which is underestimated with a 16.67%, and the lowest in *Various granivores combined* (-0.59%). By class size, types below 40 ESU have overestimation, whilst types above that level are underestimated with a tendency to a lesser relative error (in absolute value) as we move towards the centre of farm size distribution.

These total relative errors can be broken down into two parts: (i) the relative error due to the sampling method and, (ii) the relative error due to the non-existence of observations. Let r be the total relative error, r1, the relative error for the subpopulation in which the observations from the 1999 exist and **Error! Objects cannot be created from editing field codes.** the proportion of the TGM in the census for those which have information in the 1999 RECAN, then **Error! Objects cannot be created from editing field codes.**.

Table 10.9 Relative error by region. the descomposition by origin: Relative error due to the type of sampling and the relative error due to lost observations

Region	Relative error due to the type of sampling (%)	Relative error due to lost observations (%)
Galicia	2.72	-19.79
Asturias	1.63	-22.62
Cantabria	4.16	-20.13
País Vasco	5.35	-26.81
Navarra	2.10	-16.10
La Rioja	1.42	-34.30
Aragón	-3.85	-13.21
Cataluña	0.99	-16.85
Baleares	3.49	-88.83
Castilla-León	-0.18	-10.26
Madrid	2.75	-29.41
Castilla-La Mancha	0.19	-26.17
Comunidad Valenciana	-0.24	-23.17
Murcia	-9.58	-47.57
Extremadura	-0.90	-34.76
Andalucía	2.54	-11.66
Canarias	3.90	-76.96
Total	0.37	-20.38

The above Equation means that the total relative error can be broken down into two parts: (a) The relative error committed in the estimation of the TGM for cells with observations in the 1999 RECAN weighted by the percentage in the TGM that represents those cells (which we will call relative error due to sampling type) minus (b) the percentage of TGM of those cells where no observations exist in the 1999 RECAN (which we will call relative error due to lost observations). It is easy to see that when there are no observations, the relative error committed is always 100%. For this reason, the weight of the total relative error is equal to **Error! Objects cannot be created from editing field codes.**.

Tables 10.9, 10.10 and 10.11 show which part of the relative error is due to the selection method followed by the 1999 RECAN, and which is due to a lack of information.

Table 10.10 Relative error by type of farming. the descomposition by origin: Relative error due to the type of sampling and the relative error due to the empty cluster in the sample

Type of farming	Relative error due to the type of sampling (%)	Relative error due to lost observations (%)
Specialist cereals, oilseed and protein crops	1.77	-8.95
General field cropping	-0.76	-11.02
Specialist horticulture (not greenhouse)	-10.18	-38.97
Specialist horticulture (greenhouse)	-2.45	-29.17
Specialist vineyards	3.09	-33.67
Specialist fruits and citrus fruits	-1.03	-24.76
Specialist olives	7.90	-2.13
Various permanent crops combined	-0.62	-37.05
Specialist dairying	1.47	-13.73
Cattle - combined	2.01	-21.31
Sheep and goats	-2.25	-13.63
Sheep, cattle and other grazing livestock	1.00	-46.55
Pigs	-6.85	-25.88
Fowl	4.04	-60.67
Specialist combined granivores	-0.32	-45.22
Mixed cropping	0.74	-26.44
Mixed livestock	-1.17	-61.25
Mixed crops and livestock	-0.54	-20.07
Total	0.37	-20.38

Table 10.11 Relative error by size. the descomposition by origin: Relative error due to the type of sampling and the relative error due to the empty cluster in the sample

Economic size units	Relative error due to the type of sampling (%)	Relative error due to lost observations (%)
4-8	10.36	-13.87
8-16	6.21	-8.15
16-40	1.21	-5.23
40-100	-4.46	-10.40
100-250	-2.83	-39.40
>250	-2.90	-62.21
TOTAL	0.37	-20.38

In general, the relative error due to lost observations is much larger than the relative error due to sampling and of the opposite side.

Acknowledgements

We acknowledge the excellent statistical work and helpful comments of Carlos García Peñas, Spanish FADN director for their assistance with the RECAN database, and also to Raimundo Fombellida, the Associate General Director of Agricultural Statistics at the Spanish Ministry of Agriculture, Fisheries and Food for his support. This study is part of a research project granted by the Ministry of Agriculture, Fisheries and Food in agreement with the Universidad Carlos III de Madrid Jean Monnet Chair in Economics of European Integration. Ricardo Mora also acknowledges financial support from DGI, Grant BEC2000-0170.

References

AGRI.A.3, *FADN: An A to Z of Methodology, European Commission*. http://europa.eu.int/comm/agriculture/rica/pdf/site_en.pdf, 2001.

Azorín, F., *Métodos y aplicaciones del muestreo*. Alianza Editorial. Madrid, 1986.

Cochran, W.H., *Técnicas de Muestreo*. C.E.C.S.A. México, 1981.

Groves, R.M., *Survey errors and Survey costs*. Wiley. New York, 1989.

Hansen, M.H.; Hurwitz, W.N. and W.H. Madow, *Sample Survey Methods and Theory*. 2 vol. Wiley, 1953.

Hansen, M.H., Madow, W.G. and B.J. Tepping, *An evaluation of model-dependent and probability sampling inferences in Sample Surveys*. Journal of the American Statistical Association. Vol 78, nº 384, 1983, pp. 776-793.

Heckman, J.J., *Sample selection bias as a specification error*. Econometrica. Vol. 47, nº 1., 1979, pp. 153-161.

INE, *Metodología del Censo Agrario 1999*, Instituto Nacional de Estadística. http://www.ine.es/daco/daco42/agricultura/meto_CENSUSag99.doc, 2002.

Ingram, S. y G. Davidson, *Methods used in designing the National Farm Survey. Proceedings of the Survey Research Methods*. American Statistical Association, 1983, pp. 220-225.

King, R.F., *Quota Sampling*. En: Madow W. G. et al. (ed) Incomplete Data in Sample Surveys. Vol. 2. Theory and Bibliographies, 1983.

Kish, L., *Muestreo de Encuestas*. Trillas. México, 1976.

Kish, L., Statistical Design for Research. Wiley. N. Y., 1986.

Little, R.J.A., *Models for nonresponse in sample surveys*. Journal of the American Statistical Association. 77, 1982, 237-250.

Little, R.J.A., *Super population models for nonresponse*. En: Madow W.G. et al. (ed) Incomplete Data in Sample Surveys. Part VI, pp. 337-413. Academic Press, 1983.

Royall, R.M., *The linear least-squares prediction approach to 2-stage sampling*. Journal of the American Statistical Association, vol. 71, 1970, p. 697.

Royall, R.M. and Herson, *Robust estimation in finite populations I*. Journal of the American Statistical Association, vol. 68, 1973a, p. 880.

Rubin, D.B., *Inference and missing data*. Biometrica, 63: 581-592, 1976.

Rubin, D.B., *Multiple imputation for nonresponse in surveys*. Willey, 1976.

Skinner, C.J., Holt and T.M.F. Smith (ed), *Analysis of complex surveys*. Wiley, 1989.

Smith, J.W., *Sex and the GSS*. General Social Survey Technical Report nº 17. Chicago National Option Research Center, 1979.

Stephenson, C.B., *Probability sampling with quotas: An experiment*. Public Opinion Quarterly. vol. 43, nº 4, 1979, pp. 477-496.

Vogel, F.A., *125 Years of Agricultural Estimates in the USDA/NASS. Proceedings of the Survey Research Methods*. American Statistical Association, 1988, pp. 53-62.