A New Sampling Design for Swiss Farm Accountancy Data Network (FADN) Data

Andreas Roesch
Agroscope Reckenholz-Tänikon Research Station ART
Tänikon 1
8356 Ettenhausen
E-mail: andreas.roesch@art.admin.ch
Tel: 0041 52 368 3470

Paper prepared for presentation at the EAAE 2011 Congress
Change and Uncertainty
Challenges for Agriculture,
Food and Natural Resources

August 30 to September 2, 2011
ETH Zurich, Zurich, Switzerland

1. Introduction

In order to provide an economic monitoring system, accountancy data from farms is gathered, analysed and published in Switzerland. In addition to financial-economic information, a broad set of data on the technical-economic and socioeconomic situation of the farm is collected. The Farm Accountancy Data Network (FADN) is administered by Agroscope Reckenholz-Tänikon Research Station ART, which is part of the Federal Office for Agriculture (FOAG). Similar systems which are linked in the Pacioli network (www.pacioli.org) exist in all European countries.

For various reasons, the existing system for sampling and analysing data must be adapted to future requirements. The two major drawbacks of the current system are (i) the non-random sampling of Swiss farms, and (ii) the marked over- and under-representation of groups (strata) in the sample. In the latter case, a significant lack of data for a number of groups must be addressed. This has led to the formulation of a strategy for overcoming all identified drawbacks (Lips et al. 2010).

In the past, many different sampling designs were investigated. The use of effective sampling techniques is an ideal way to acquire knowledge about important aspects of a population. Stratified sampling is one of the most frequently used techniques. The optimal allocation approach is often applied to determine sample size within each stratum of a stratified sample scheme. This is done by solving a cost-constrained non-linear optimisation problem in which the objective function is the variance. Traditionally, this problem has been solved by using the Cauchy-Schwarz inequality (Stuart, 1954) or Lagrange's multiplier method (see Sukhatme et al., 1984). These methods may be applied to both univariate and multivariate variables. In its univariate form, the size of the sample is computed by taking into account a single key variable. Many of the optimisation approaches developed are based on multiple variables, and allow for the allocation of different costs (Cochran, 1977). Khan et al. (1983) suggested an optimal multivariate stratified sampling design using dynamic programming. They developed a criterion which they called 'compromise allocation', and which gives the optimum sample size in some sense for all characteristics. Nevertheless, taking multiple response variables into account may cause serious problems, since the variables are largely weighted on an arbitrary level. We have therefore decided to base the minimisation of the sample's mean variance on a single variable, furthermore we assume constant costs for all units in the population. In addition, it will be shown that the (per-stratum) variances of many key economic variables such as farm income, cash flow and standard gross margin are often highly correlated with one another, rendering the application of the multivariate variables somewhat awkward.

This study describes the new sample design, based on disproportional stratified random sampling, which avoids some of the major shortcomings in the sampling design currently used for the Swiss FADN. In addition, emphasis is placed on a detailed discussion of the accuracy of both technical and economic key variables. Note that the recruitment of farms and related problems such as refusals will not be covered by this paper.

The two main aims of this paper are (i) the presentation of a new sampling design for the Swiss FADN, and (ii) a detailed analysis of expected accuracies.

The paper is organized as follows: The data used are specified in Section 2. Section 3 describes the stratification scheme. Section 4 deals with the structure of the FADN population, while Section 5 provides a detailed description of the selection plan. The analyses of expected results and accuracies of aggregated values on both the (Swiss) national and stratum level are presented in Section 6, while the findings are summarised in Section 7.

2. Data

The Swiss agricultural census data are collected at the Federal Office for Agriculture (FOAG). These data provide a detailed insight into the structural, technical and sociodemographic situation of almost all Swiss farms. The census does not cover economic data, which are of high importance for agricultural policy reasons. The current investigation is based on the FOAG census from 2007 covering a total population of 61,763 farms. Based on the census data, the FADN population (the farms that have a non-zero probability of joining the FADN sample) will be defined by excluding farms below a certain threshold of standard output, with the aim of eliminating small farms run strictly as a hobby or sideline.

In addition to the FOAG census data, frequent use is made of economic information from the FADN sample data. This comprehensive database includes detailed information on cost accounting from 3,328 farms in the year 2007.

3. Design of the stratification scheme

Stratification is a statistical technique used to increase sampling efficiency. The target population is broken down into similarly structured subgroups or strata which should be as homogeneous as possible, forming mutually exclusive groups. Minimising the number of farms therefore helps to achieve a certain degree of accuracy in the estimated mean values. The current stratification scheme for the Swiss FADN is based on three groupings: farm size, type of farming and region. The region is in turn subdivided into three levels: 'mountain', 'hill' and 'plain'. Since fundamental differences in terms of aspects such as climate, orography, production and farming system exist between these regions, it is wise to retain the three regions in the future design as well. The existing 11 farm types will not be redefined, as this would lead to serious problems regarding long-term time series and public understanding. In addition. size of farm will also be retained as an important dimension, given that small farms often react differently to policy measures and/or market changes from larger farms (Andersen et al., 2006). The number of size classes will be reduced from the current five down to two, however, in order to avoid empty or nearly empty strata. A detailed analysis has shown that it is ideal to distinguish farms with a utilised agricultural area (UAA) of below and above 20 hectares. This means that the future selection plan will be based on $H = 2\ (land\ classes) \cdot 11\ (farm\ types) \cdot 3\ (regions) = 66$ strata. The reduction in the number of strata from the current 165 (5·11·3) to 66 will not lead to much loss of information, since too many strata increases the chance that separate strata will not be represented by a sufficiently high number of sample farms to provide accurate estimates of strata means.

4. Definition of the FADN population

Given that there is not enough time and/or money to gather information from all farms in the census, the aim is to find a representative sample (or subset) of that population. Therefore, in sampling, the definition of a target population from which the sample is drawn is of utmost importance. This reduced population will be referred to as the field of survey (FS) or FADN population (Figure 1). The FS can be defined as including all farms with the characteristic that one wishes to understand. The FADN population need not represent the entire census population, however, as we are deliberately excluding farms below a certain (low) economic threshold (cf. Section 4.1).

4.1 Threshold based on standard output

The currently used criterion is based on 11 thresholds relating to the size of agricultural land or livestock numbers, one of which at least must be exceeded (Meier, 2005). As these thresholds rely on assumptions with no underlying statistical considerations, a new, simple approach will henceforth be applied for the future sampling design.

The new approach is based on the standard output (SO) of the farms, defined as the monetary value of the gross agricultural output (Schürch et al., 2010). SO was selected for two reasons: (i) it is the key monetary measure equal to the gross agricultural output at the farm-gate price, and (ii) the EU determines the economic size of agricultural holdings by using SO. SO can thus be interpreted as the "economic size of the farm", which in itself is an indicator of farm size. In the effective productive stage, ART will use SO values compiled per farm by the Federal Statistical Office. This implementation will be fulfilled by mid 2011, together with the 2010 Agricultural Census. The threshold value for 2007 may therefore differ slightly from the actual experimental pilot values.

The FS is generated as follows: Based on the farms ranked according to their SO, we now include all farms, starting with the farm with the highest SO, until the summed SO reaches 95% of the total SO in the census, hence ignoring the farms on the lower tail of the SO density curve, which are smallholdings run more for pin money or as a hobby than as a career or occupation. This means that the FS covers 95% of the summed SO generated by all census farms. For the 2007 census data, this leads to a threshold value of approximately CHF 63,080. (Economically) small farms are thus excluded from the sampling frame, since their SO lies below the given threshold separating the FS from the census data. This threshold value will be adjusted annually on the basis of the census data, and will therefore vary slightly from year to year. The authors suggest using multi-year averages in order to limit the impact of extreme events such as the rapid decline in the (Swiss) milk price in 2009.

As in the European Union (EU), an upper threshold will not be applied to exclude very large farms from the field of survey. Nevertheless, experience from previous years has clearly shown that it is difficult to obtain data from (economically) large farms.

As some of the statistics estimated from the FS and from the census data differ quite substantially, it is essential to analyse the coverage of different key variables as outlined in Section 4.2.

4.2 Coverage

Coverage compares the FS with the agricultural census data, i.e. the extent to which the census is represented in the FS. Given that it is only farms in the FS that have a non-zero probability of joining the FADN sample and that it is therefore the properties of the FS which directly apply to the FADN sample, this is clearly an important issue. The coverage of some activities and technical data is presented in Table 1. It indicates the extent to which the FS covers the entire census population. It shows that little more than 70% of all census farms (corresponding to 43,964 farms) are included in the FS. This means that a substantial percentage of the census farms (28.8%) is excluded from the sampling framework. Nevertheless, Table 1 demonstrates that the FS largely covers the census population with respect to most key variables. As given from the method, 95% of production (measured as SO) is covered by the FS. The FADN population covers almost 88% of the UAA, with a distinctly higher coverage for arable land (94.4%) than for grassland (85.4%), which predominates in the hill and mountain regions of Switzerland. Livestock is also well represented, with an almost full coverage of pigs (99.8%) and poultry (98.5%). Since sheep and goat farms typically have a low SO, sheep are poorly covered in FS (65%). Arable farms are well covered. Almost 95% of the 0.275 million hectares of Swiss arable farmland (26% of total UAA) is covered by the FADN population. The FS covers almost all potato (98.7%), sugar beet (97.1%) and vegetable cultivation (98.6%), while horticulture is very poorly covered with less than 40%. In summary, it can be stated that although a substantial proportion of the farms have been excluded from the field of survey, the percentage of production value not covered is quite low. Only farms of small (economic) size are excluded from the sample. Such farms are predominantly found in the hill and mountain areas. This assumption is confirmed by the data: whereas 79% of all census farms in the plain region are members of the FS, this figure decreases to 74% and 56% for the hill and mountain regions, respectively. This means that the FS includes little more than every second mountain farm from the census data.

In addition to analysing coverage, it is essential to analyse the percentage coverage for the Swiss FADN stratification type, distinguishing between 11 farm types and (the currently used) five size classes (Table 2). The third classification variable, Region, has been ignored here for reasons of simplicity.

Table 2 clearly shows that coverage for small-sized farms below 10 ha is low, while farms with UAAs above 30ha are largely included for most farm types. This matches our specification, since small-sized farms tend – with some important exceptions – to be run as a hobby or sideline. Significant differences in coverage are found among the 11 Swiss FADN farm types. For Special crops, Pigs/poultry, Combined dairy/arable and Combined pigs/poultry, the FS consists almost entirely of farms with a UAA of between 10 and 20 ha. This is in line with the findings from Table 1 depicting a high coverage for the production branches Pigs, Poultry, Vegetables and – to a lesser extent –Dairy. Conversely, the low coverage for Suckling cows and Other cattle is in line with the low values in Table 1 for Suckling cows, Sheep and Goats.

4.3 Sensitivity

In order to better assess further differences between the FS and the census population, it is beneficial to analyse the impact of the value of the SO threshold on some key variables. Figure 2 displays the sensitivity curves for coverage for a number of important selected variables. The figure clearly reveals the extremely high SOs generated by Swiss pig farms: even with a threshold of CHF 400,000, 80% would be included in the FS. By contrast, less than 10% of the sheep population are covered, assuming the same threshold for separating farms between the FADN and the census population. The figure reveals that a significant proportion of the 6,200 organic farms currently operating in Switzerland fail to achieve a reasonable SO, with little more than one-third of all organic farms reaching more than CHF 100,000 (standard) output from their products.

5. Selection Plan

Due to a fixed budget, it is not possible to collect accountancy data from all farms in the FADN population. Estimates for the entire population will therefore be based on a limited sample of farms. The selection plan must ensure that the farms included in the FADN sample are representative of the population as a whole. The selection-plan specification is described in the following sections.

5.1 Proportional and optimal allocation

In stratified sample surveys, the sample sizes for all strata must be known before drawing the sample.

The notation used to describe the sampling procedure is as follows: The population FS of size N is first divided into L subpopulations (strata) of N_1, $N_2, \ldots N_L$ elements, respectively. These strata are non-overlapping, and $N=N_1+N_2+\ldots+N_L$. Strata sample sizes are denoted by n_1, $n_2, \ldots,$ n_L, respectively. The total sample size n equals the sum $\sum_{h=1}^{L} n_h$. The mean (the first moment) of a variable y in strata h is denoted as $\overline{y_h}$, the standard deviation (positive square root of second moment) as σ_h. When applying proportional allocation, the total sample mean $\bar{y}$ of a stratified sample is computed as a weighted mean of $\overline{y_h}$, $h=1,\ldots,L$, with weights W_h equal to the fraction N_h/N. In other words, the same percentage of farms is chosen in every stratum. In order to minimise $\text{var}(\bar{y})$ for a fixed total cost c, the method of optimal (rather than proportional) allocation will be applied (Cochran, 1972). This method is discussed in the following section.

The number of farms per stratum will be determined such that the (total) sampling variance of the estimator standard gross margin (SGM) will be minimised for a specified cost. This can be achieved by applying the optimal allocation according to Neyman-Tschuprov (Cochran, 1977), assuming the per-unit cost $c_h=c$ to be the same for all strata. By restricting the optimisation to a single variable (SGM), we avoid allocating arbitrary weights for the computation of the weighted sum of the sampling variances for multiple variables. The variance of the estimated mean $\bar{y}$ of a sample with L strata is given by Equation (1):

$$\text{var}(\bar{y}) = \sum_{h=1}^{L} \frac{N_h^2}{N^2} \sigma_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) \qquad (1)$$

where

σ_h = standard deviation of a given target variable in stratum h

N_h = number of elements in stratum h in the FADN population

n_h = sample size in stratum h

n = total sample size

N = total size of census (number of farms)

L = total number of strata.

Stratification ensures that variation between strata does not enter into the variance by taking account of this source when drawing the sample. In actual surveys the sampling

error is usually expressed in terms of the coefficient of variation (COV), a normalised measure defined as the ratio of the standard deviation to the mean.
The task of minimising Equation (1) subject to the condition of the cost function $c = \sum_h n_h c_h = c_0$ can be accomplished by using the Lagrange multiplier. For a given total sample size n, and assuming per-unit cost to be the same in all strata, this leads to the following optimal sample sizes n_h in stratum h:

$$n_h = n \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h}, \; h = 1, 2, \dots, L. \tag{2}$$

This implies that the sample fractions n_h/N_h differ between the strata, justifying the alternative term 'disproportional stratified sampling'. Equation (2) indicates that the sample size n_h is high when (i) the stratum h accounts for a large part of the population, and (ii) the variance within stratum h (σ_h) is large, i.e., the sampling fraction is higher to compensate for heterogeneity.
Since the population studied by farm samples is often markedly skewed, the non-robust measure σ_h may be strongly affected by a few large observations or 'outliers'. Hence, in any stratum there are likely to be a few very large SGMs which account for a large proportion of the stratum variance. These sampling units, referred to as "surprise units" in Dalenius (1962), may be (i) kept in the sample, resulting in a large variance; (ii) rejected, leading to a bias; or (iii) downweighted, e.g. as per occurrences in previous years. In order to yield a reasonable estimate of σ_h, we have decided to follow approach (ii), using only the values between the 0.25% and 99.75% quantile. In addition, a simple outlier detection has been applied, with an outlier being defined as data outside the band $\overline{y_h}$ +/- $10 \cdot \sigma_h$. Bootstrapping (Efron, 1993) and sensitivity analyses on the 2007 census data have shown that this procedure is appropriate for providing reasonable (and robust) estimates of the variability measure σ_h.

5.2 Confidence intervals

If either (i) the sample sizes within each stratum are large, or (ii) the sampling design has a large number of strata, we can approximate the 95% interval for the sample mean $\bar{y}$ by using the variance from Equation (1) as follows:

$$\bar{y} \pm t_{0.975,f} \cdot SE(\bar{y}) \tag{3}$$

with $t_{0.975,f}$ being the 97.5 percentile of the t-distribution with f degrees of freedom, and SE the standard error of $\bar{y}$, i.e. the square root of $var(\bar{y})$. The degrees of freedom are computed as the difference between the total sample size and the number of strata. In order to characterise the accuracy of the entire sample, we define the parameter q as the ratio of half the length of the confidence interval divided by $\bar{y}$:

$$q = \frac{t_{0.975,f} \cdot SE(\bar{y})}{\bar{y}} \tag{4}$$

It is evident that the value of the standard error, and thus the parameter q, strongly depends on the selected variable. This implies that the selection of the variables used for variance minimisation is crucial for the success of the sampling plan. For economic studies, the work income or some closely related measure would be most appropriate for estimating the variations within strata. Since this variable is unfortunately available only for the roughly 3,300 non-randomly sampled FADN farms (covering approximately 5% of all Swiss farms), it is, for statistical reasons, beneficial to derive variability estimates of variables that are available in the FSO census data. In order to avoid any arbitrary weighting using a multivariate technique, variabilities will be estimated on the basis of a single key variable only. Extensive evaluations have shown that the standard gross margin (SGM) is ideal for measuring the within-strata variability. SGMs are calculated per unit area of crops and per head of livestock on the basis of standardised SGM coefficients for each type of crop and livestock. The SGM of a crop or livestock type 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. SGM is an important tool for providing information on the economic size of a farm as well as its type, and has therefore been extensively used to classify farms in the EU and in economic studies up to the year 2009 (Boone, 2002; De Bont et al., 2003).

5.3 Random sampling

The advantages of random sampling versus non-probability methods such as quota sampling have already been addressed in the early literature (Bowley, 1926; Hubback, 1927). Due to the requirements of Swiss federal statistics (Kilchmann 2007), following the guidelines of random sampling was therefore strongly recommended.

Based on the number of farms to be recruited (Section 6.1), a random draw (without replacement) per stratum will be performed. Assuming expected non-response rates, the number of farms drawn per stratum must be at least three times higher than the required number of farms in order to ensure sufficient responses. Experience from several farm surveys shows a return rate of between 15 and 50 per cent. It is important to analyse carefully the farms that refused to participate. These refusals will cause problems if they differ (with respect to the mean and variability of key variables) from the farms that participate in their place. If high non-response rates occur, it will be necessary to investigate the reasons carefully (e.g. via questionnaire) in order to reduce the bias in the estimates. Sample A and B farms from the same stratum will be analysed to identify possible differences in response rate and characteristics between the two samples (cf. Section 5.4).c As the new sampling plan is not yet in used in practice, however, the consequences of unexpectedly high non-response rates and differences between sample A and B can only be analysed at a later stage.

5.4 Sampling plan

The sampling plan is based on the framework of two samples, A and B. Whereas sample A is taken from each stratum, sample B is only taken from ’economically‘ important strata. This distinction was motivated by both practical and statistical considerations which will be outlined below.

Sample A will include only a few goal variables such as agricultural income and cash flow, and will be drawn from strata with either (i) a limited number of farms in the census, or (ii) strata with farm types that are difficult to recruit, such as Special crops or Horticulture. By contrast, farms belonging to sample B must provide a detailed set of bookkeeping data with variable direct costing in order to allow for analyses of specific production branches. This two-sample design effectively reduces the bias, since sample A guarantees the collection of key bookkeeping data for farm types or regions that are (almost) absent from the current FADN sample. The improved geographic coverage is likely to lead to a marked reduction in (probable) systematic bias. Sample B, by contrast, covers all strata of economic importance and/or whose farm managers were generally seen to be highly motivated to contribute their accountancy data over the last decade (M. Lips, D. Schmid; pers. comm., 2009).

Table 3 displays the strata sampling farms for samples A and B. Perusal of the table reveals that sample B strata include more than 10% of the farms in the respective region. It is planned to compare farms from the same strata in samples A and B in order to investigate possible biases due to different response rates.

The number of farms to be sampled per stratum is computed according to the following multi-step procedure. The aim is to specify the sample sizes n_h in stratum h for both sample A and B.

(i) Definition of the FADN population (cf. Chapter 4), i.e. specifying the set of farms with a non-zero probability of joining the population.

(ii) The accuracy parameter q (Sect. 5.2, Equation 4) is set, based on available monetary resources and accuracy considerations. This parameter is now computed for different total sample sizes n on the basis of the equation set (1)-(4), as well as the FADN and census data for 2007. Figure 2 displays the derived relation between q and the total number n of farms included in the sample. The figure clearly shows that for low normalised confidence intervals q below 0.05, the total sample number n must be strongly enhanced for higher accuracy in the total sample mean. The number of sample farms required rapidly decreases from 8,300 (q=1%) to 2,230 (q=3%) and 1,020 (q=5%). The sensitivities or first discrete derivatives ($\Delta n / \Delta q$) rapidly become smaller for increasing q-values, and level off for very low accuracies (high q values). Slightly different values for samples A (q_A) and B (q_B) have been suggested (see summary in Table 4) in order to account for different accuracy requirements.

(iii) In addition to Swiss sample means, customers also request stratum means. The number of farms per stratum obtained in (ii) is therefore reviewed critically so as to guarantee a specified accuracy in each stratum. This is accomplished by restricting the maximum inaccuracy using the two parameters $q_{h,A}$ and $q_{h,B}$, denoting the minimum accuracy level required for all strata h for samples A ($q_{h,A}$) and B ($q_{h,B}$). Again, different values for $q_{h,A}$ and $q_{h,B}$ have been proposed (cf. Table 4).

(iv) For feasibility reasons, a maximum response rate of 30% is assumed. This means that $n_{h,A}$ (the number of sampled farms for stratum h in sample A) and $n_{h,B}$ both follow the simple constraint:

$$n_{h,S} \leq 0.3 \cdot N_h \text{ , with } S = \{A, B\} \text{ and } h = 1, \ldots, 66 \qquad (5).$$

Since restrictions (iii) and (iv) automatically exclude any stratum with a (too-) low number of sampled farms, no further constraint describing a minimum number of farms per stratum is required.

The technical properties of samples A and B are summarised in Table 4. For the test year 2007, the sample populations for A and B are approx. 43,960 and 33,420 farms, respectively. The sample A (B) size is equivalent to 2,535 (2,361) farms, corresponding to sample percentages of 5.8% (7.1%). The value of the accuracy parameter q at the Swiss level (q_A, q_B) and at the stratum level ($q_{h,A}$ and $q_{h,A}$) clearly shows that the maximum normalised confidence interval is approx. one order of magnitude greater at the stratum level than at the Swiss level.

5.5 Weighting

The application of a weighting system at single-farm level is a crucial prerequisite for sample mean estimates that represent the FS mean with the maximum possible accuracy. The sampling weight represents the number of farms in the population represented by the sample member. For each stratum, the weights are computed as the ratio of the number of farms from the FADN population to the sample. The sampling is therefore simply the reciprocal of the probability of selection. A weighted average of the sample observations gives a good estimate for the entire FADN population. Note that the FS average may differ from the census mean, as substantial differences exist between the census and the FADN population for certain strata (cf. Table 2).

6. Results

6.1 Number of farms per stratum

The number of sampled farms per stratum is tabulated in Table 5 according to the sampling plan described in Section 5 above. Table 5 provides the results based on the 2007 census data. The total number of farms in samples A and B are given in Table 4, leading to a mean sampling percentage of 5.8% and 7.1% for sample A and B, respectively. The results clearly show that the farm-type stratum ‘Special crops‘ requires relatively high sampling percentages to meet accuracy constraints owing to its inhomogeneous structure, and hence its high variability. Furthermore, the sample proportion for the larger size class (UAA > 20ha) is often considerably larger than for the small-scale farms. This is hardly surprising given that – for any farm type – strata with smaller farms are typically more homogeneous than the corresponding farm stratum with UAA > 20ha.

6.2 Simulated sampling

This section provides a detailed assessment of the expected accuracy of sample means at both the Swiss and stratum level. Since economic variables for agricultural data are often skewed, we use simulated sampling (SS), which – in combination with the exhaustive data pool from the census data – allows for detailed analysis of the expected variabilities of estimated sample means. Farm weights are calculated as the ratio of the number of farms in FS to the number of farms in the sample (cf. Section 5.5). This means that the random sample should ideally reproduce the statistics derived from the FADN population (which differs from the results computed from all agricultural census data, cf. Section 4). Note that in order to obtain consistent results, we have assumed throughout this section that the random sample was constructed according to the constraints given in Table 4 for sample B, but without omitting certain strata. A single SS simulation consists of randomly selected farms according to the sampling plan (Section 6.1), from which the weighted (Swiss) mean is computed for the designated target variable.

Figure 3 displays the probability density function (PDF) for the Swiss SGM mean based on 2,000 simulations using SS. Assuming Gaussian-distributed simulated data, the 95% confidence interval for the SGM mean values is equal to 1.96 · CHF 831 = CHF 1,629, or 1.5% of the overall simulated mean (CHF 104,430). The true mean (sample population) is CHF 104,346. The estimated value is well within the range of the accuracy constraint required for sample B (q_B=2.1%, cf. Tab. 4). Confidence intervals may also be estimated by computing confidence intervals from the simulations. Since the confidence intervals assuming normal distributed means reproduce the simulated confidence interval very well, however, we omit additional specification of the simulated confidence intervals.

Accuracy clearly depends on the target variable. Knowing the 'true' FS mean for the (technical) parameters available in the agricultural census allows direct comparison of the results from SS and the true mean. Table 6 gives an overview of accuracy based on the previously presented sampling plan and SS.

It is clear from Table 6 that the random sample is large enough to accurately estimate the Swiss mean of selected key target variables. The confidence interval e.g. for the utilised agricultural area (UAA) can be interpreted as follows: When drawing a random sample 100 times according to the selection plan described above, 95 of the 100 estimated Swiss UAA means will be within 21.24 +/- 0.28 ha. The coefficient of variation (COV), computed as the quotient of the mean and the standard deviation, is approximately equal to one-quarter of the normalised confidence interval, assuming a normal distribution. Since the COV is generally below 1%, we conclude that the sampling plan described above allows accurate predictions at the Swiss level.

From simple statistical considerations, it is evident that reliability decreases for subsets of Switzerland, such as a single stratum. This is evident in Table 7, which shows the results from SS for Arable crop farms (farm type 11) in the plain region with UAA < 20ha, hereinafter referred to as CROPs. This sample contains 177 farms, corresponding to a sampling percentage of 10.2%. Tables 6 and 7 show that the ratio of the COV for the Swiss mean to CROP varies between 2.7 and 10.8 for OAA and LS, respectively. This demonstrates that for any variable, it is difficult to compute confidence intervals for specific strata from the accuracy achieved at the Swiss level. The ranking of the variables according to their COVs, and hence the reliability of the estimates, is clearly different for CROP and the whole of Switzerland. Further analyses have shown that these findings also apply to most other variables and strata.

6.3 The method "PROCPDF"

In addition to accuracy estimates of (technical) census variables gathered for all Swiss farms, it is of utmost importance also to provide expected confidence intervals for goal variables that are known in the FADN sample but unknown for the FADN population.

Here, we face the problem that SS cannot be applied directly, since economic variables are generally available only for the FADN farms.
Accountancy data is estimated for the entire FADN population by separately carrying out the following three-step procedure, hereinafter referred to as PROCPDF, i.e. *PRO*Cedure using the *P*robability *D*ensity *F*unction (PDF): For a given stratum and a given economic variable v: (i) (randomly) select a farm i from the FADN sample between 2003-2008, and extract its target-variable value v_i ; (ii) assign this value v_i to a randomly selected farm from the FADN population of the year 2007; (iii) repeat steps (i) and (ii) until the allocation has met every single farm within the stratum. In order to approximate a real distribution at farm level, a (small) normal random error has been added to each variable value v_i. This algorithm guarantees that, for any given variable, the probability density function (PDF) of the 2007 FADN population equals the respective PDF of the 2003-2008 FADN sample data. Consideration of the six-year period 2003-2008 instead of the one-year period 2007 aims at a better approximation of the FADN PDF through enhanced sample sizes, mainly in strata with (too) few elements. Note, however, that this procedure may provide biased density functions, as (i) FADN farms are not randomly sampled, and (ii) data from several years are used to approximate the distribution for one year, ignoring e.g. effects such as inflation.
Applying the procedure described above allows the computation of approximated confidence intervals for all FADN variables on the basis of SS. Table 8 sheds some light on the expected accuracies of several key variables for both Switzerland and the CROP stratum (‘Arable crop’ farms with UAA >20ha, Plain Region). At the Swiss level, we expect agricultural income to be CHF 60,465+/- CHF 1,021 with COV=0.9%, while the respective numbers for CROP are CHF 68,185+/- CHF 4,506 with COV=3.4%, respectively. As shown in Table 7, for structural data available only in the census data, Table 8 reveals that the estimated COVs for CROP farms are higher by a factor of 3 to 5 than those of the Swiss level. The reliability of the estimated means plainly increases when the sample size rises from 177 (CROP stratum) to 3,121 (total sample size). The tabulated figures for labour and major economic indicators reveal that the ’true‘ mean (mean sample population) and the simulated overall mean do not differ, even at very high (>99.99%) confidence levels.

Summary and conclusions

The new Swiss sampling design is described and analysed in terms of its expected accuracy. The core of the new sampling plan consists of the following three aspects: (i) limiting the FADN population from the census data using SO; (ii) random sampling; and (iii) constraining the length of confidence intervals for mean values at both the stratum level and the Swiss level.
The present study demonstrates that SO is an appropriate parameter for defining the FADN population from the census data. The advantage of defining the threshold by requiring 95% of the census SO to be above this value is threefold: (i) the threshold is intuitively clear from both a statistical and economic point of view; (ii) the measure is easily customisable on an annual basis; and (iii) SO is used as a new measure for the EU farm-typology classification from 2010 onwards.
An in-depth analysis of expected sample means was performed for the year 2007. This paper shows that the selection plan presented is well adapted for the reasonably accurate

estimate of aggregated values for key economic variables at both the country and stratum level. SS was identified as an ideal tool for analysing expected accuracies of aggregated variables for variables that are often distinctly skewed. The simulations revealed that the optimum number of farms per stratum (based on SGM) is also well suited for estimating the mean of other key economic variables.

The problem of estimating aggregated values of variables that are *not* recorded in the census surveys is tackled via the methods PROCPDF.

The implementation of the new sampling design has only been (theoretically) tested on the basis of past census and FADN sample data. The performance of the new sampling plan has not yet been tested under actual conditions. The accuracy of means from randomly sampled farms is, for example, heavily dependent on the response rate. It is of the utmost importance to compare the main characteristics of the farms prepared to participate with those of the farms refusing to join the sample. In addition, possible differences in response rates between samples A and B must be analysed thoroughly. This problem will be investigated by sampling dairy farms for both sample A and B. Future efforts must also be dedicated to an in-depth investigation of the quality of the current farm-type classification. This will be accomplished by applying multivariate statistics such as clustering or discriminant analysis.

List of abbreviations

AI	Agricultural income
ART	Agroscope Reckenholz-Tänikon Research Station
AWU	Annual work unit
CF	Cash flow
COV	Coefficient of variation
EQ	Equity capital
FA	Farm assets
FADN	Farm Accountancy Data Network
FAWU	Family annual work unit
FOAG	Federal Office for Agriculture
FS	Field of survey
FSO	Federal Statistical Office
GL	Grassland
LC	Loan capital
LS	Livestock
LU	Livestock Unit
MD	Mahalanobis distance
OAA	Open arable land
PDF	Probability density function
SGM	Standard gross margin
SLU	Standard labor unit
SM	Statistical matching

SO	Standard output
SS	Simulated sampling
UAA	Utilised agricultural area
WIFM	Work Income per family worker

References

Andersen, E., D. Verhoog, B. Elbersen, F. Godeschalk, and B. Koole. *A Multidimensional Farming System Typology*, SEAMLESS, report no. 12, 2006. Available at http://www.seamless-ip.org/Reports/Report_12_PD4.4.2.pdf.

Boone, J.A., C.J.A.M. de Bont and K.J. Poppe (eds.), *Inputs for AgrIS- Checking the consistency of the agricultural sector database*, Report 8.02.4, LEI, The Hague, 2002.

Bowley, A.L., 1926: *Measurements of the precision attained in sampling*. Proc. Internat. Statist. Inst., **12**, 6-62.

Causey, B.D., 1983: *Computational Aspects of Optimal Allocation in Multivariate Stratified Sampling*. Siam Journal of Science Statistics, **4**, 2, 322-329.

Chatterjee, S., 1967. *A Note on Optimum Allocation*, Scandinavian Actuarial Journal, **50**, 40-44.

Cochran, W. G. C., 1977: *Sampling Techniques*, Wiley, New York.

Dalenius, T., 1962. *Recent Advances in Sample Survey Theory and Methods*. The annals of mathematical statistics, **33**, 325-344.

De Bont, C.J.A.M., Van Everdingen, W.H., Koole, B., 2003. *Standard gross margins in the Netherlands*. LEI Report 1.03.04, Den Haag.

Efron, B., and Tibshirani, R. J., 1993. *An introduction to the bootstrap*. New York: Chapman & Hall.

Garica, J.A.D. and Cortez, L.U., 2006: *Optimal Allocation in Multivariate Stratified Sampling: Multi Objective Programming*. Comunicacion Technica. No I-06-06, 1-22.

Hubback, J.A., 1927: *Sampling for rice yield in Bihar and Orissa*. Imperical Agricultural Research Institute, Pusa, Bulletin no. 166.

Kilchmann, D., 2007: *Studie bemepro: Auswahlplan 2012 für die Zentrale Auswertung von Buchhaltungszahlen*, Aktennotiz, Bundesamt für Statistik, Neuenburg.

Kish, L., 1976: *Optima and Proxima in Linear Sample Design*. Journal of the Royal Statistical Society. Series A, **139**, 1, 80-95.

Lips, M., Mühlethaler, K., Hausheer Schnider, J., Roesch, A. and Schmid, D., 2010. *Stichprobenkonzept für das Schweizer Buchhaltungsnetz landwirtschaftlicher Betriebe* (in

English: Sampling System for the Swiss Farm Accountancy Data Network). Jahrbuch der Österreichischen Gesellschaft für Agrarökonomie.

Mahalanobis P.C., 1936: *On the generalised distance in statistics*. Proceedings National Institute of Science, India, **2**, 1, pp. 49-55.

Meier, B., 2005*: Analyse der Repräsentativität im schweizerischen landwirtschaftlichen Buchhaltungsnetz. Messung und Verbesserung der Schätzqualität ökonomischer Kennzahlen in der Zentralen Auswertung von Buchhaltungsdaten*. Dissertation ETH no. 15868, Zürich, Switzerland.

Schürch, D. and Schmid, D: Standard Output: *XXXX*. Agrarforschung, submitted.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., and Asok, C., 1984: *Sampling Theory of Surveys with Applications*. Iowa State University Press. Ames. IA.

Vrolijk, H., 2004: *STARS: statistics for regional studies. Proceedings of Pacioli: New roads for farm accounting and FADN*. Report 8.04.01. LEI, The Hague.

Vrolijk, H., W. Dol, and T. Kuhlman, 2005: *Integration of small area estimation and mapping techniques. Tool for Regional Studies*. LEI, the Hague, 60pp.

Xiang, S., F. Nie, C. Zhang, 2008: *Learning a Mahalanobis distance metric for data clustering and classification*, Pattern Recognition, **41**, 12, pp. 3600-3612.

Appendix A: Figures

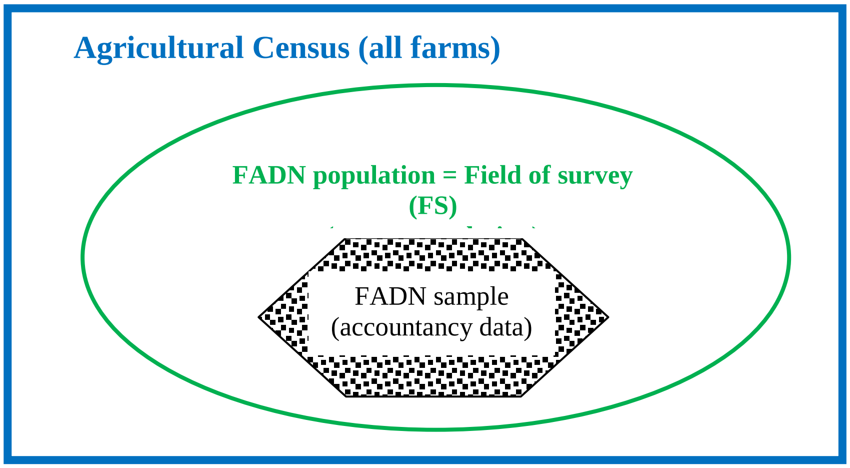

Fig.1: Illustration of the relationship between the Agricultural Census, the FADN population and the FADN sample. The field of survey (FS), also referred to as the FADN population or target population, includes the set of farms with non-zero probability of joining the FADN sample. Accountancy data are only available for the farms of the FADN sample.

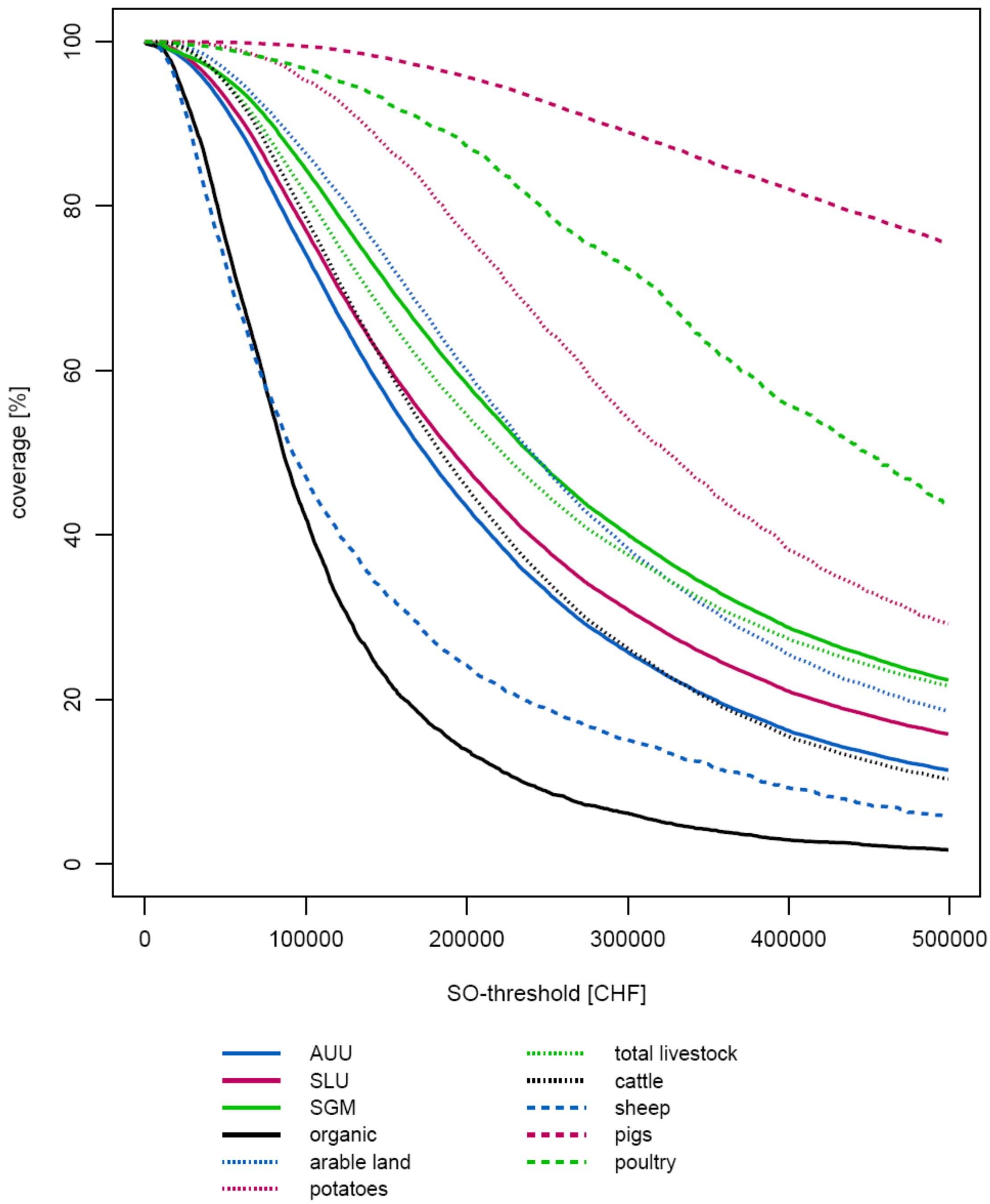

Fig.2: Sensitivity curves. Relationship between the threshold for SO and coverage for some selected key variables. Abbreviations are as follows. UAA: utilised agricultural area; SLU: standard labor unit; SGM: standard gross margin.

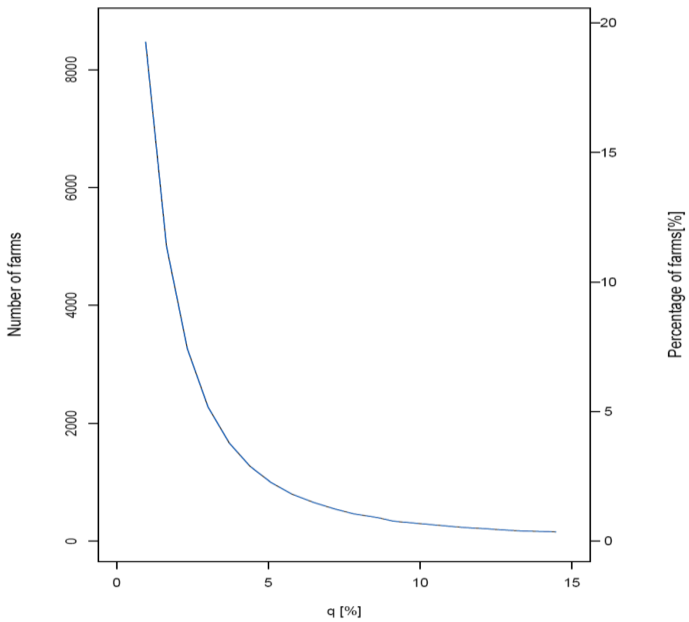

Figure 3: Percentage of farms (number of farms in the sample divided by the number of farms in the FADN population (in %)) vs. accuracy parameter q for the Swiss sample mean. The parameter q (Eq.4) indicates the half-length of the confidence interval, normalised with the sample mean. The y-axis is labelled separately in order to relate the number of farms to its respective percentage. Note that this estimate is based on the assumption that the sample comprises all strata.

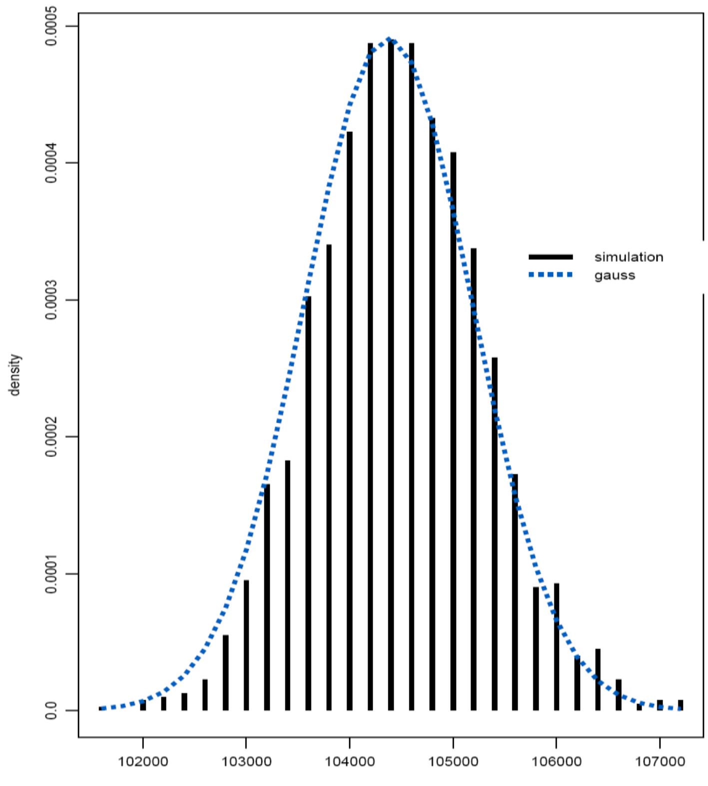

Figure 4: Probability density function for the (weighted) Swiss SGM mean. Method: SS, 2,000 replications. For the number of randomly sampled farms per stratum, see Table 5. Dotted line: Theoretical normal distribution, given the mean from the 2,000 Swiss mean values (CHF 104,430) and the standard deviation (CHF 831) computed from SS.

Appendix B: Tables

*Table 1: Coverage of the field of survey (FS) compared to the agricultural census (2007). Areas are given in hectares, livestock numbers in livestock units. Monetary quantities are presented in millions of Swiss Francs (CHF). *The percentage for SO is 95%, as this is required from the procedure which forms the FS.*

Agricultural Census Variable	*Unit*	*Number according to Census*	*Share of FS [%]*
Standard output (SO)	*Millions of CHF*	*11887.0*	*95.0**
Farms	*-*	*61,763*	*71.2*
Farm managers	*-*	*61,763*	*71.2*
Family labour (full time)	*-*	*64,582*	*93.0*
Organic farms	*-*	*6,199*	*66.7*
Standard gross margin (SGM)	*Millions of CHF*	*491.1*	*93.4*
Utilised agricultural area (UAA)	*ha*	*106,0256*	*87.9*
Arable land	*ha*	*275,1091*	*94.4*
Grassland	*ha*	*742,647*	*85.4*
Vegetables	*ha*	*9,517*	*98.6*
Horticulture	*ha*	*1,418*	*38.9*
Fruit and berries	*ha*	*7,845*	*97.7*
Vineyards	*ha*	*12,894*	*86.5*
Winter wheat	*ha*	*76,275*	*94.1*
Maize	*ha*	*17,461*	*89.5*
Potatoes	*ha*	*11,745*	*98.7*
Sugar beet	*ha*	*20,656*	*97.1*
Rye	*ha*	*1,780*	*87.9*
Number of animals			
Total	*LU*	*129,3290*	*92.2*
Cattle	*LU*	*948,226*	*91.6*
Suckling cows	*LU*	*93,545*	*73.9*
Dairy cows	*LU*	*614,795*	*95.1*
Pigs	*LU*	*198,759*	*99.8*
Poultry	*LU*	*47,474*	*98.5*
Sheep	*LU*	*43,500*	*65.0*
Goats	*LU*	*10,368*	*81.4*

Table 2: Percentage coverage of the field of survey (FS) compared to the Agricultural Census (2007). For currently used farm type and size classification, see Meier (2005).

Farm type/ UAA	**<10ha**	**10-20ha**	**20-30ha**	**30-50ha**	**>50ha**
11 Arable crops	12.6	68.4	98.2	99.9	100.0
12 Fruit/ vegetable/ vines	42.9	99.9	100	100.0	100.0
21 Dairy farms	47.1	91.4	98.5	99.9	100.0
22 Suckling cows	6.8	36.9	59.4	77.3	88.2
23 Other cattle	12.6	54.5	78.9	87.9	91.1
31 Horses/sheep/goats	29.7	71.9	94.7	99.0	100.0
41 Pigs/poultry	89.1	96.6	96.4	100.0	100.0
51 Combined dairy/arable	77.9	99.5	100	100.0	100.0

52 Combined suckling cows	21.1	81.0	99.3	100.0	100.0
53 Combined pigs/poultry	87.7	98.9	99.7	100.0	100.0
54 Combined others	23.0	81.1	96.9	99.1	99.6

Table 3: Allocation of strata to sample A and sample B. A: Sampling for sample A only. ***T****: Sampling for both sample A and sample B. All strata are represented in sample A. Figures in brackets give the percentage of farms per stratum based on the 2007 census (figures in each row add up to 100%). Figures greater than or equal to 10% are given in red.*

Type/ Region	***11***	***12***	***21***	***22***	***23***	***31***	***41***	***51***	***52***	***53***	***54***
Mountain	*A (0)*	*A (2)*	***T*** *(48)*	***T*** *(10)*	***T*** *(20)*	***T*** *(10)*	*A (1)*	*A (0)*	*A (0)*	*A (3)*	*A (5)*
Hill	*A (1)*	*A (2)*	***T*** *(45)*	*A (7)*	*A (6)*	*A (5)*	*A (4)*	*A (3)*	*A(2)*	***T*** *(12)*	***T*** *(14)*
Plain	***T*** *(15)*	*A (13)*	***T*** *(13)*	*A (3)*	*A (2)*	*A (3)*	*A (3)*	***T*** *(14)*	*A (3)*	***T*** *(12)*	***T*** *(18)*

Table 4: Technical parameters for samples A and B. Number of farms in the sample and the field of survey are based on the 2007 census data. PR: plain region; HR: hill region; MR: mountain region. Farm types: 11: Arable crops; 12: Special crops; 21: Dairy; 22: Suckling cows; 23: Other cattle; 31: Horses/sheep/goats; 41: Pigs/poultry; 51: Comb. dairy/arable; 52: Combined suckling cows; 53: Combined pigs/poultry; 54: Combined others. For the definition of q, see Eq. 4. Note that for the sake of comparison with sample A, q_B *was estimated by including all strata in sample B, thereby contradicting the sampling plan.*

	Sample A	Sample B
Level of detail	low	very high
Farm types considered	all strata	PR: 11, 21, 51, 53, 54 HR: 21, 53, 54 MR: 21, 22, 23, 31
Total number of farms (2007)	2,535	2,361
Number of farms in the field of survey (2007)	43,964	33,417
Accuracy for SGM at the Swiss level	q_A=3.0%	q_B=2.1%
Accuracy constraint at the stratum level (SGM)	$q_{h,A}$<25%	$q_{h,B}$<20%

Table 5: Number of farms and associated sampling percentages (%, in brackets) for sample A, based on the 2007 agricultural census. Criteria used for stratification are (i) region, (ii) type of farm, and (iii) two size classes (UAA > 20ha, UAA <= 20ha). For the Swiss farm-type classification, see legend of Table 4. NA: empty stratum; no farms in the sample population and thus random sample.

Farm Type	Plain Region		Hill Region		Mountain Region	
11	21 (2.1%)	119 (6.8%)	28 (26.4%)		*NA*	
12	190 (9.0%)	187 (30%)	28 (11.2%)		53 (23.1%)	
21	48 (2.8%)	67 (5.6%)	75 (2.1%)	106 (4.8%)	39 (1.6%)	121 (4.1%)
22	13 (3.6%)		17 (3.1%)		14 (3.3%)	

23	15 (8.6%)	17 (25.4%)	27 (5.1%)		17 (1.1%)	
31	25 (3.0%)	15 (23.1%)	22 (3.6%)	19 (21.8%)	12 (1.6%)	14 (6.6%)
41	81 (10.7%)	10 (12.2%)	41 (6.0)		33 (15.3%)	
51	34 (3.4%)	178 (7.8%)	18 (4.7%)		*NA*	
52	11 (4.0%)	25 (5.7%)	10 (14.7%)	14 (11.2%)	*NA*	
53	119 (7.5%)	164 (12.3%)	52 (4.8%)	52 (8.9%)	11 (4.0%)	14 (7.7%)
54	94 (4.7%)	165 (8.1%)	29 (3.2%)	54 (6.1%)	15 (4.6%)	16 (7.5%)

Table 6: Estimated accuracies of some key variables (weighted Swiss averages), based on SS (2,000 simulations). The true mean is computed from the FADN population using the agricultural census data. The overall mean, standard deviation and confidence interval are computed from 2,000 simulated (Swiss) mean values. The coefficient of variation is defined as the ratio of the standard deviation to the mean. All figures refer to the year 2007. The number of (randomly) selected farms per stratum is given in Tab.5. For abbreviations, see List of Abbreviations.

Parameter	True Mean (sample population)	SS Mean	Standard Deviation	Confidence Interval (95% level)	Coefficient of Variation (COV)
UAA	21.18 ha	21.24 ha	0.14 ha	+/- 0.28 ha	0.6%
OAA	6.01 ha	6.00 ha	0.06 ha	+/- 0.12 ha	1.0%
GL	14.43 ha	14.52 ha	0.11 ha	+/- 0.22 ha	0.7%
LS	27.1 LU	27.3 LU	0.2 LU	+/- 0.4 LU	0.7%
SO	CHF 256,860	CHF 257,205	CHF 2,411	+/- CHF 4,227	0.9%
SLU	1.90	1.89	0.01	+/- 0.02	0.6%
SGM	CHF 104,346	CHF 104,430	CHF 831	+/- CHF 1,628	0.8%

Table 7: As Table 6, but for the stratum 'Arable crop' with UAA >20ha, Plain Region. This stratum contains 177 randomly sampled farms (cf. Table 5).

Parameter	True Mean (sample population)	SS Mean	Standard Deviation	95% Confidence Interval	Coefficient of Variation (COV)
UAA	35.39 ha	35.41 ha	0.95 ha	+/-0.19 ha	2.7%
OAA	29.39 ha	29.40 ha	0.80 ha	+/-1.60 ha	2.7%
GL	5.39 ha	5.39 ha	0.21 ha	+/-0.42 ha	3.8%
LS	9.01 LU	9.00 LU	0.71 LU	+/-1.39 LU	7.8%
SO	CHF 236,126	CHF 236,045	CHF 9,196	+/- CHF 18,023	3.9%
SLU	1.45	1.45	0.05	+/-0.1	3.5%

SGM	CHF 127,521	CHF 127,559	CHF 4,370	+/- CHF 8,566	3.4%

Table 8: As Table 6, but for variables surveyed in the FADN sample only. Means and accuracies for both Switzerland (bold, lines labelled 1) and CROP (italic, lines labelled 2). This stratum contains 177 randomly sampled farms (cf. Table 5). For abbreviations, see Appendix. Units: Swiss francs (CHF).

Parameter	Unit		Mean (sample population)	SS Mean	Standard Deviation	95% Confidence Interval	Coefficient of Variation (COV)
AI	**CHF**	**1**	**60,465**	**60,476**	**521**	**+/-1,021**	**0.9%**
	CHF	*2*	*68,185*	*68,161*	*2,299*	*+/-4,506*	*3.4%*
WIFW	**CHF**	**1**	**41,501**	**41,537**	**572**	**+/-1,121**	**1.4%**
	CHF	*2*	*62,000*	*61,991*	*3,643*	*+/-7,140*	*5.9%*
FAWU		**1**	**1.25**	**1.25**	**0.01**	**+/-0.02**	**0.5%**
		2	*1.03*	*1.03*	*0.03*	*+/-0.06*	*2.4%*
CF	**CHF**	**1**	**101,948**	**102,049**	**654**	**+/-1,282**	**0.6%**
	CHF	2	*107792*	*107,853*	*2,582*	*+/-5,061F*	*2.4%*
LC	**CHF**	**1**	**366,495**	**367,680**	**4,287**	**+/-8,403**	**1.2%**
	CHF	*2*	*292,238*	*292,104*	*14,586*	*+/-28,589*	*5.0%*
FA	**CHF**	**1**	**778,490**	**777,777**	**5,962**	**+/-11,686**	**0.8%**
	CHF	*2*	*718,857*	*718,735*	*20,410*	*+/-40,003*	*2.8%*
EQ	**CHF**	**1**	**454,604**	**453,565**	**4,556**	**+/-8,930**	**1.0%**
	CHF	2	*472,363*	*471,663*	*18,265*	*+/-35,799*	*3.9%*