istat
working
papers

N.10
2016

A simple approach for stratification of units in multipurpose in business and agriculture surveys

Marcello D’Orazio, Elena Catanese

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N.10
2016

A simple approach for stratification of units in multipurpose in business and agriculture surveys

Marcello D'Orazio, Elena Catanese

Istat

Istat Working Papers

A simple approach for stratification of units in multipurpose in business and agriculture surveys

N. 10/2016

ISBN 978-88-458-1897-4

A simple approach for stratification of units in multipurpose in business and agriculture surveys[1]

Marcello D'Orazio, Elena Catanese

Sommario

Il campionamento casuale stratificato è frequentemente utilizzato nelle indagini campionarie sulle imprese e sulle aziende agricole. Un passo cruciale del disegno di campionamento è rappresentato dalla stratificazione della popolazione di interesse. In genere, tale operazione viene condotta sfruttando le informazioni ausiliarie, disponibili nella lista di campionamento e legate ai fenomeni oggetto di indagine. La stratificazione pone problemi nelle indagini che si prefiggono di osservare congiuntamente più fenomeni. In letteratura esistono diversi metodi di stratificazione univariata, applicabili cioè utilizzando una sola variabile ausiliaria; al contrario pochi sono i metodi che gestiscono più variabili ausiliarie. Questo lavoro propone un nuovo approccio per risolvere tale problema. La nuova procedura viene confrontata con uno dei metodi di riferimento, nel contesto della progettazione di alcune indagini campionarie nel settore agricolo.

Parole chiave: campionamento casuale stratificato, stratificazione multivariata.

Abstract

The stratified random sampling is often used in sample surveys on businesses and farms. The stratification of the target population is a key step in the sampling design. It is usually performed by using the auxiliary information available from the sampling frame and related to the phenomena under investigation. The task becomes complex in multipurpose surveys, where different phenomena have to be investigated at the same time. Different stratification criteria have been proposed in presence of a single auxiliary variable while few methods are available to tackle the problem in the presence of various different auxiliary variables. In this paper is introduced a new procedure to solve this problem. The procedure is compared with one of the reference methods, in the framework of the design of some sample surveys in the agricultural sector.

Keywords: stratified random sampling, multivariate stratification.

[1] The views expressed in this paper are those of the authors and do not necessarily represent the institutions with which they are affiliated.

Contents

1. Introduction

Traditionally sample surveys on enterprises and farms are based on one stage stratified sampling. In practice the sampling frame is divided in non-overlapping subpopulations (or *strata*) and sampling is performed independently in each subpopulation. Stratification, as noted by Cochran (1977), allows for reduction of the sampling error and permits to derive reliable estimates for each subpopulation. In agriculture surveys, usually the strata are formed by considering geographical information, type of farming (specialist crops, specialist livestock, mixed) and some measures of farms' size (e.g. size of areas with crops, livestock, etc.). The variables used for stratification are chosen among the ones available in the sampling frame (e.g. Register of active farms, administrative sources, the previous Census data), the more the auxiliary variables are correlated with the target variables the higher will be the benefits in using them for stratification purposes.

The stratification of a population does not pose problem when performed through categorical variables such as geographical regions, while eventual continuous auxiliary variables need to be categorized. Multipurpose surveys pose additional problems: it is not simple to divide units in strata being homogenous with respect to the different phenomena to investigate; in particular, the sampling frame may provide several auxiliary variables positively correlated with the target ones but uncorrelated or negatively with respect to each other; in this case the choice of the auxiliary variable for stratification purposes becomes more complex. This paper tackles this problem i.e. stratification in presence of continuous auxiliary variables, by suggesting a relatively new procedure, introduced in Section 3. In Section 4 this new procedure is compared with a multivariate method proposed by Ballin and Barcaroli (2013) by applying both to design samples for three agriculture surveys; the main findings are summarized. Main features and notation of stratified random sampling design are provided in Section 2.

2. Stratified sampling

The application of stratified sampling requires a number of decisions strictly related each other's: (i) how to stratify the population and how many strata to create; (ii) which selection scheme employ in each subpopulation (simple random sampling, systematic, probability proportional to size, etc.) and, finally, (iii) the size of the whole sample and corresponding partitioning among the strata (so called *allocation*).

Even if the stratification allows for different independent selection schemes in each subpopulation, the common practice in business and agriculture survey is to apply simple random sampling without replacement in all the strata because of its practical and theoretical advantages. In most of the cases the sample size is decided according to the desired precision for the main survey estimates (expressed in terms or *relative standard error*: desired sampling error divided by the quantity to estimate, denoted usually as CV). In some agriculture survey the desired CVs in estimating the total amount for the main variables are decided at European Union (EU) level and explicitly mentioned in EU regulations. A crucial role is played also by the sample allocation between the strata. Different allocation rules may be employed: equal allocation, proportional allocation, Neyman allocation, power allocation etc. The choice depends on the desired precision characterizing the final survey estimates and the stratification strategy.

2.1 Basic notation of stratified random sampling

Let U be the finite population under investigation, consisting of N units. This population is divided into H non-overlapping subpopulations or strata ($U = U_1 \cup U_2 \cup \ldots \cup U_H$) whereas N_h denotes the number of units in the stratum h and, consequently, $N = \sum_{h=1}^{H} N_h$. Stratified random sampling consists in selecting a simple random sample without "replacement", s_h of n_h $(n_h \leq N_h)$ "units", independently stratum by stratum; the overall sample size is $n = \sum_{h=1}^{H} n_h$.

If *Y* is the target variable, an estimate of its total amount $t_y = \sum_{k \in U} y_k$ in *U* is provided by:

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_h} y_k \qquad (1)$$

whose sampling variance is:

$$V\left(\hat{t}_y\right) = \sum_{h=1}^{H} V\left(\hat{t}_{yh}\right) = \sum_{h=1}^{H} N_h \left(N_h - n_h\right) \frac{S_{yh}^2}{n_h} \qquad (2)$$

The *relative standard error* is $CV_{t_y} = \sqrt{V\left(\hat{t}_y\right)} \Big/ t_y$.

By fixing in advance the desired relative error *d* in estimating the total amount of *Y* it is possible to determine the required sample size:

$$n_{opt} = \frac{\sum_{h=1}^{H} \frac{N_h^2 S_{yh}^2}{n_h / n}}{t_y^2 d^2 + \sum_{h=1}^{H} N_h S_{yh}^2} \qquad (3)$$

The partitioning of n_{opt} among the strata can be done according to different criteria; in the *proportional allocation* the stratum sampling fraction is set equal to the relative size of the stratum $n_h = n_{opt} N_h / N$, thus ensuring equal including probabilities to all the units; with *optimal allocation* (often called *Neyman allocation*) the sampling fraction is higher in more heterogeneous strata having $n_h \propto N_h S_h$; *power allocation* (cf. Särndal et al., 1992, pp. 470-471) avoids underrepresentation of small strata and is a compromise solution between the Neyman and an allocation ensuring constant precision for each of the strata estimates. The derivation of the optimal sample size and the corresponding allocation between the strata requires information concerning the *Y* variable, usually not known in advance. For this reason it is taken into account an auxiliary variable *X*, known for all the units in the population and being highly correlated with *Y*; for instance, the optimal sample size is estimated by using the expression (3) whereas all characteristics related to *Y* are substituted by the corresponding ones but computed on *X*.

In multipurpose surveys a unique sample should satisfy precision requirements concerning several target variables; in this case the decisions concerning the overall sample size and the corresponding allocation can be approached in a multivariate setting by expressing it as a convex mathematical programming problem; Bethel (1989) provides a solution to this problem, an alternative solution is given by the Chromy (1987) algorithm.

In this setting, the choice of the sample size and its allocation among the strata is a step that can be performed only after the stratification of *U*, i.e. given *H* and the corresponding partitioning of *U* into non-overlapping strata ($U = U_1 \cup U_2 \cup \ldots \cup U_H$). Therefore the first decision in stratified sampling usually concerns how to partition the target population.

2.2 Univariate stratification of the population

The expression (2) shows that given a non-null sample size n_h $(1 \leq n_h \leq N_h)$ the sampling variance will reduce if the variance of *Y* in U_h is small; therefore an efficient stratification should try to derive strata as homogeneous as possible for the target variable. Unfortunately the variable *Y* is not known in advance and consequently the stratification is carried out on one or more auxiliary variables strictly related with *Y*. As said before, the stratification does not pose any particular problem when it is based on a categorical *X* variable (typically Regions, NACE in case of business surveys or Farm Typology as far as farms are concerned). Difficulties arise when *X* is a continuous variable and therefore its categorization is necessary.

A well-known approach consists in using the *cumulative* $\sqrt{f}$ *rule* proposed by Dalenious and Hodges (1959). Unfortunately such procedure is not suitable for *X* variables showing highly skewed distributions; a typical situation in most business and agriculture surveys where most of the auxiliary and target variables present high positive skewness. A common adopted strategy to manage such situation is to separate few large units in a specific stratum and to include all of them in the sample with certainty (so called *take-all* stratum; cf. Hidiroglou and Lavallée, 2009); in practice, given that these units contribute at large extent to the total amount of the target population, separating them in a stratum that is censused allows to lower the whole sample size.

Hidiroglou (1986) proposed an iterative algorithm to identify the threshold b_c such that all the units with *X* values exceeding it ($x_k > b_c$) are put in the take-all stratum; the procedure requires the specification of the desired CV. Once identified the take-all stratum, the remaining units can be further stratified according to one of the available methods, e.g. the cum $\sqrt{f}$ *rule*. Lavalleé and Hidiroglou (1988) introduced a unified procedure for identifying the take-all stratum and then stratify the remaining units. It starts by specifying the desired level of precision (CV) and ends up with a stratification that minimizes the overall sample size. The partitioning of the sample between the take-some strata follows the power allocation criterion. The Lavalleé and Hidiroglou method is based on an iterative procedure which unfortunately may not converge to a global minimum. Convergence problems can partly be solved by using the procedure suggested by Kozak (2004).

To overcome the problem of an allocation performed on a variable, *X*, that does not correspond to the target one (*Y*), Rivest (2002) suggested using anticipated moments of *Y* given *X* in the Lavalleé and Hidiroglou procedure. Recently, Baillargeon and Rivest (2009) suggested a modification of the procedure by introducing the possibility of separating small units in a stratum that is not sampled (*take-none* stratum). In fact small units could have a negligible contribution to the total amount of the interest variable, which usually holds true in presence of highly positive skewed distributions. The same authors, provided an important contribution for applying the various above mentioned methods by developing the software package “stratification” (Baillargeon and Rivest 2011, 2014) freely available for the R environment (R Core Team, 2015).

The stratification problem can also be approached within a model based framework where the strata are formed by considering similar values of $V_\xi(y_k)$, the model variance of *Y* (cf. Särndal et al., 1992, Section 12.4). In particular when dealing with a linear super-population model with $E_\xi(y_k) = \beta x_k$ and $V_\xi(y_k) = \sigma_0^2 x_k^\gamma$ ($\gamma > 0$, large γ denotes more pronounced heteroscedasticity) the stratification can be performed by grouping units with similar values of $V_\xi(y_k) = \sigma_0^2 x_k^\gamma$. In this context the optimal sampling design (i.e. that minimizes the anticipated variance) is the one which ensures inclusion probabilities proportional to the model standard deviation i.e.:

$$\pi_k = n \frac{\sqrt{V_\xi(y_k)}}{\sum_{k \in U} \sqrt{V_\xi(y_k)}} = n \frac{x_k^{\gamma/2}}{\sum_{k \in U} x_k^{\gamma/2}} \qquad (4)$$

where *n* is the expected sample size. A simple fixed-size design which ensures $\pi_k \propto x_k^{\gamma/2}$ is a stratified random sampling design where: (i) the *H* strata are formed by means to the *equal aggregate* σ *-rule* (cf. Särndal et al., 1992, Section 12.4), i.e. the strata are formed by grouping units homogeneous with respect to the $x_k^{\gamma/2}$; (ii) the sample is allocated equally among the strata, $n_h = n/H$; and, (iii) the combined ratio estimator is used for estimating the total amount of *Y* in the population.

Usually γ lies in the interval $\left(0,2\right]$; in most establishment surveys $1 \leq \gamma \leq 2$ (cf. Särndal et al., 1992, Section 12.5); when $\gamma = 2$ the optimal model based design provides the same inclusion of probability proportional to size (PPS) "sampling", $\pi_k = n \; x_k / t_x$.

2.3 Multivariate stratification

Stratification has been deeply explored when dealing with a single auxiliary *X* variable supposed to be highly associated/correlated with the target variable *Y*. The problem becomes more complex in multipurpose surveys with many target variables not necessarily related one each other. In such cases there may be a number of auxiliary variables *X* related at different levels with the various target variables; a stratification on a single *X* variable may not be efficient for all the target variables. Kish and Anderson (1978) highlighted that the advantages of using several stratification variables are greater in multipurpose surveys. The gains however depend on the two factors: (i) the relationship between the stratification variables and the target ones, and (ii) intercorrelations among the stratification variables.

The simplest strategy consists in: 1) selecting a minimal subset of the *X* variables which are connected with most of the target ones; 2) performing univariate stratification on each of the single *X* variables, and then 3) derive the final stratification by cross-classifying units according to the chosen categorized variables. For the sake of parsimony the selected *X* variables should not be related to each other's (at least not more than weakly), and among highly associated variables one should take into account for stratification purposes only the one with higher relative variability to avoid any lack of information. This overall described strategy can determine too many strata, with some of them too small in terms of size.

In literature there are some proposals to perform stratification in the bivariate case. Kish and Anderson (1978) suggested to apply the cum $\sqrt{f}$ rule independently on each of the variables and then to obtain the final stratification as a combination of the two results. Roshwalb and Wright (1991) extended the model based stratification approach to the bivariate case.

When dealing with more than two stratification variables, Hagood and Bernert (1945) suggested performing stratification on the first principal components computed starting from the set of the predictors. The same approach is followed by Pla (1991) but attention is limited just to the first component. Kish and Anderson (1978) warned about principal components approach because it may provide final strata that are not readily interpretable moreover principal components analysis (PCA) deals with intercorrelations among the stratification variables rather than their relationship with the target variables. Barrios *at al* (2013) noted that PCA is not able to deal with high skewed variables with few units exhibiting very high values, they try to alleviate the problem by performing PCA on the log-transformed variables but the results obtained in their simulation studies on agriculture surveys require further investigations.

Benedetti *et al* (2008) suggested using a unique procedure to perform both stratification and sample allocation in a multivariate framework, once defined a set of desired CVs for the target variables. Their proposal consisted in a tree-based approach that, starting from a set of strata, identifies finer and finer partitions of the units by minimizing at each step the sample allocation.

A similar framework is considered in Ballin and Barcaroli (2013) paper. Their sequential procedure starts with a very fine stratification performed on each of the available stratification variables, then it performs iterative strata collapsing. The procedure aims at minimizing the final sample size given the target precision (CVs) required for a set of target variables (or available proxies which can be the same auxiliary variables used to create the initial fine partition) under the optimal Bethel allocation. The proposed procedure makes use of the genetic algorithm and is implemented in the package "SamplingStrata" (Barcaroli, 2014) available for the R environment. The procedure is very effective in achieving a small sample given the target CVs, however the identified final stratification, obtained by various collapsing steps of intermediate strata, is not readily interpretable. Moreover the procedure requires an initial subjective choice concerning the stratification variables to use and the corresponding categorization: the initial 'atomic' stratification is obtained by cross-classifying the so obtained categorical variables. The procedure involves setting a high number of initial parameters many tuning steps are required; finally a number of iterations is required to achieve the final result, thus implying a non-negligible computational effort.

3. A new procedure for multivariate stratification

Method proposed here addresses the problem of stratification in the presence of a series of auxiliary variables, supposed to be related to the target variables, in a context similar to the one based on PCA, in the sense that stratification is carried out on a finale score variable, but not the principal component. The proposed approach makes use of some findings of model-based univariate stratification and, in particular, by considering the multivariate version of the Brewer's optimal selection, i.e. the *Maximal Brewer Selection* (MBS) also known as *Multivariate Probability Proportional to Size* (MPPS) (Kott and Bailey, 2000). As shown in Section 2.2, in the univariate case, the optimal design in a model based framework is the one that guarantees $\pi_k \propto \sqrt{V_\xi(y_k)}$, coupled with a ratio estimator of the total. When dealing with a set of J ($J \geq 2$) available auxiliary variables, the MBS (1997) consists in setting:

$$\pi_k^* = \min\left\{1,\ \max_j\left[\pi_{1,k},,\ldots\pi_{j,k},\ldots,\pi_{J,k}\right]\right\},\quad k = 1,2,\ldots,N \qquad (5)$$

where

$$\pi_{j,k} = n_j \frac{x_{j,k}^{\gamma/2}}{\sum_{k\in U} x_{j,k}^{\gamma/2}},\quad j = 1,2,\ldots,J\ ;\quad k = 1,2,\ldots,N \qquad (6)$$

Here n_j are the "target" sample sizes (cf. Kott and Bailey, 2000) for each of the J ($J \geq 2$) available auxiliary variables, while $0 < \gamma \leq 2$ (Authors suggest setting $\gamma = 3/2$ in agriculture surveys). In practice the final inclusion probability of a unit is obtained as the maximum of the various inclusion probabilities derived by considering each of the available auxiliary variables.

The idea in this paper is to perform the stratification on the π_k^*. In particular the proposed strategy consists in:

a) Derivation of the π_k^* by means of expressions (5) and (6);

b) Stratification of the units in H strata by means of the *equal aggregate* σ*-rule* applied on the π_k^* :

b.1) the population units are ordered according to increasing magnitude of π_k :

$$\pi_1^* \leq \pi_2^* \leq \ldots \leq \pi_k^* \leq \ldots \leq \pi_N^*\ ;$$

b.2) Once decided *H*, the number of final strata, include in the first stratum the N_1 elements such that:

$$\sum_{k=1}^{N_1} \pi_k^* = \frac{1}{H}\sum_{k=1}^{N} \pi_k^*$$

and so on for the remaining strata.

c) Multivariate allocation of the sample by means of the method proposed by Chromy (1987), once set the precision criteria in estimating the total amount of the interest *Y* variables.

In other words the stratification procedure is designed to create strata of units characterized by similar inclusion probabilities accounting for the various auxiliary variables being considered. From a practical viewpoint a procedure is very straightforward but a number of choices should be made. At first, in order to determine the π_k^*, it is necessary to set in advance the target sample size n_j for each of the *J* auxiliary variables being considered; Kott and Bailey (2000) suggest taking the ratio between σ_0 and the desired anticipated coefficient of variance C_j^*; in such a case however it is necessary to estimate σ_0 (from previous survey data) or make a guess for it. A simplifying choice would be that of considering a constant value, i.e. $n_j = n_0$ ($n_0 > 0$) for $j = 1, 2, \ldots, J$. This choice, in the proposed procedure corresponds to perform the stratification directly on the values of a new variable *Z*:

$$z_k = \max_j \left[\frac{x_{1k}^{\gamma/2}}{\sum_{k=1}^{N} x_{1k}^{\gamma/2}}, \ldots, \frac{x_{jk}^{\gamma/2}}{\sum_{k=1}^{N} x_{jk}^{\gamma/2}}, \ldots, \frac{x_{Jk}^{\gamma/2}}{\sum_{k=1}^{N} x_{Jk}^{\gamma/2}} \right], \quad k = 1, 2, \ldots, N \qquad (7)$$

In practice the values z_k substitute π_k^* in steps (b.1) and (b.2).

The value of γ depends on the heteroscedasticity. Usually $0 < \gamma \leq 2$ but in most establishment surveys a narrower interval $1 \leq \gamma \leq 2$ can be considered. Särndal *et al.* (1992) claim that $\gamma = 1$ is a good compromise choice, another suggestion favors $\gamma = 3/2$ (cf. Kott and Bailey, 2000). In theory, it is possible to set different values of γ for each of the *X* variables being considered. But this would introduce a further element of complexity which may be unnecessary. Some Authors (Godfrey et al., 1984; Särndal & Wright, 1984) suggest methods to estimate γ. In the present setting we will consider it fixed and equal to 2.

As far as step c) is concerned, in traditional model based stratification it is suggested to allocate *n* equally into the *H* strata, i.e. $n_h = n/H$ ($h = 1, 2, \ldots, H$). Unfortunately in multipurpose survey the same sample should ensure accurate estimates for a number of target variables, and the equal aggregate rule may not serve for this purpose. For this reason, in the proposed procedure the equal allocation is replaced with the optimum allocation in case of multi-character surveys.

The decision concerning *H*, the final number of strata, is as usual a subjective choice. In the univariate case, Cochran (1997, pp. 132-134) shows that when stratification is performed on *X*, under the hypothesis of linear relationship between the target variable *Y* and *X*, the reduction of sampling variance is negligible beyond $H = 6$ unless there is a very strong correlation between *Y* and *X* (greater than 0.95).

The efficiency of the proposed approach is explored in the next Section through a series of simulations carried out with real data related to surveys in the agriculture context. Moreover the proposed procedure has been compared with the multivariate procedure suggested by Ballin and Barcaroli (2013).

4. The new stratification procedure in some agriculture surveys

To explore the efficiency and the issues related to the implementation of the proposed stratification strategy some simulations were performed by using data and problems encountered in agriculture surveys carried out on a regular basis. In particular three surveys are considered: (i) the Early Estimates for Crop Products Survey (EECPS) (Regulation EC No. 543/2009); (ii) the livestock survey (LS) carried out twice a year (Regulation EC No. 1165/2008); and (iii) the Farm Structure Survey (FSS) carried out every three years (Regulation EC No. 1166/2008).

4.1 The survey characteristics

The Early Estimates for Crop Products Survey (EECPS) is an annual survey aimed at providing forecasts for the main crops (cereals for the production of grain, dried pulses and protein crops for the production of grain, root crops, industrial crops and plants harvested green) at national level. In practice the survey is designed to provide also estimates of the total amounts at NUTS1 level (5 macro areas in Italy) characterized by a relative standard error (CV) not exceeding the 3%. As usual stratified random sampling design is employed, and strata are obtained by cross-classifying NUTS1 regions with the categorizations of few relevant auxiliary variables observed in the latest 2010 Agriculture Census. For the sake of simplicity, the stratification procedure introduced in Section 3 is tested and applied to one NUTS1 domain South and Islands of Italy, which is the richest in terms of crops. The Census Frame consists of 282 017 farms with more than 1 hectare devoted to crops. The Table 1 (2nd column) provides the list of the target Y variables which are of interest in this domain.

Table 1 – Auxiliary and target variables used for stratification and allocation purposes.

	EECPS			LS			FSS	
X variables (areas in ha)	Y variables (areas in ha)	CVs	X variables (No. animals)	Y variables (No. animals)	CVs	X variables	Y variables	CVs
Cereals	Durum Wheat	0.03	Bovines	Bovines	0.010	Cereals	Cereals	0.05
	Barley	0.03		Cows	0.015	Industrial crops	Oil seed crops	0.05
	Oats	0.03	Pigs	Pigs	0.020	Harvested green	Harvested green	0.05
Legumes	Legumes	0.03	Sheep	Sheep	0.020	Permanent grassland	Permanent grassland	0.05
Harvested green	Harvested green	0.03	Goats	Goats	0.050	Vineyards	Vineyards	0.05
Vegetables	Tomatoes	0.03				Bovines	Dairy cows	0.05
Potatoes	potatoes	0.03					Other bovines	0.05
						Pigs	Pigs	0.05
						Poultry	Poultry	0.05

The livestock survey (LS) aims at estimating the total number of bovine animals, pigs, sheep and goats kept for farming purposes; the bovine and pigs livestock estimates shall be produced twice a year (a given day in May/June and a given day in November/December) while the sheep and goat livestock estimates should be provided just in the November/December survey edition. The survey should provide estimates of the total amounts at national level characterized by a CV not exceeding a threshold ranging from 1 to 5% according to the different livestock categories (see Table 1). The target population consists of 173 617 farms that in occasion of the 2010 Census had at least 1 head for the above mentioned livestock categories. The sampling design in use is again stratified random, the stratification and the joint sample allocation are determined through the procedure suggested by Ballin and Barcaroli (2013) starting from a very fine cross-classification of the farms by considering as auxiliary variables 2010 Census's values for important livestock subcategories.

The third survey considered is the Farm Structure Survey (FSS), a very important survey carried out every three years (in the Census occasion it is substituted by the Census itself) which should provide estimates of a wide set of characteristics of the farms at national and regional level (NUTS 2). The precision requirements concern the regional estimates and should not exceed the 5% for the most important crops or livestock characteristics identified for each region (with the exception of small regions) according to a series of rules explicitly defined at EU level. It is worth noting that, the target population of the FSS excludes the smallest agricultural holdings which together contribute 2 % or less to the total *utilized agricultural area* (UAA) and 2 % or less to the total number of farm livestock units. The sample design must be a stratified sampling design, as mentioned in EU

Regulation; the main stratification variable is the NUTS 2 domain, then in each region a finer stratification of the farms is defined by cross-classifying the farms according to opportune categories of UAA (change region by region) and of Livestock Size units (LSU) characteristics. The largest farms at regional level in terms of UAA or LSU are included in the sample with certainty. The whole sample size of the 2013 FSS consisted of about 42 000 farms out of the 1 138 214 farms in the target population. For our purposes it is considered the Veneto Region where, according the FSS Regulation, many farm characteristics should be investigated (for details see Table 1). The target population in Veneto consists of 119 384 farms.

4.2 Main results

The simulations carried out with the data consist in designing the samples needed to achieve the target CVs, as reported in Table 1, by applying the procedure proposed in Section 3 (denoted as S1) and the one suggested by Ballin and Barcaroli (2013) (denoted as S2). The auxiliary data used for stratification and allocation purposes are provided by the data collected in the 2010 Census occasion. As can be seen in Table 1, sometimes the same *X* variable is used for both stratification and allocation purposes. The procedures are compared in terms of the overall final sample size needed to achieve the desired CVs, once fixed the total number of strata *H*. Different values of *H* are considered. The computations are performed in the R environment. The Table 2 provides the main findings.

Table 2 – Overall sample size achieved with the alternative stratification strategies.

EECPS			LS			FSS		
H	S1	S2	*H*	S1	*S2*	*H*	S1	S2
20	4 107	5 601				20	2 706	2 265
30	4 020	4 996				30	2 664	2 272
40	3 926	4 706				40	2 634	2 044
50	3 821	4 465	50	11 885	3 163	50	2 619	2 103
75	3 682	3 986	75	11 517	3 130	75	2 592	1 977
100	3 498	3 626	85	11 277	3 127	100	2 554	1 851
150	3 381	3 275	110	11 160	3 109	150	2 521	1 837

In the case of EECPS our procedure (S1) is very efficient and performs better than S2 in almost all cases with the exception of $H = 150$ where the sample size achieved by S2 is smaller than the S1 one. In the FSS case, the S2 procedure performs always better than the S1 and the relative gap in terms of final sample size increases as the total number of strata grows. In both S1 and S2 the final sample size decreases by increasing number of strata, even if the decrease is faster with S2.

As far as LS is concerned, the S2 procedure outperforms S1 by providing a final sample size which is about 1/4 of the one provided by S1 roughly for each *H*. It is worth to notice that differently from the previous settings, both procedures S1 and S2 keep constant sample size as *H* increases, thus suggesting that there is no gain in efficiency by the use of a finer stratification. This results is likely due to the particular situation: a kind of separation between farms having bovines and the remaining ones (somehow shown by the negative correlation between the number of bovines and the number of heads of the others species, as reported in table A.2).

In general, the method proposed in this work (S1) seems more effective when the desired number of strata is relatively small and tend to work well when the auxiliary variables do not show negative correlation. S1 compared to S2 has a negligible computational effort (6-8 seconds vs. 3-4 hours of S2). In any case, both the procedures end up with strata that are not readily interpretable, an undesirable feature for survey practitioners that in some occasions may need to understand how a stratum has been created in order to perform particular analysis, typically to collapse strata (e.g. to compensate for unit nonresponse).

5. Conclusions

The procedure presented in the paper permits to tackle the problem of stratification in presence of many continuous stratification variables in a simple manner with a negligible computational effort. The first results obtained for the design of samples of three agriculture surveys seem promising in two of the three cases being investigated. The procedure proposed by Ballin and Barcaroli (2013) performs better in terms of final sample size but the price to pay is a higher computational effort. It is worth noting that both the procedures provide final strata which are not are not readily interpretable, an issue which may create problems in the processing stage if strata collapsing should be performed due to empty strata caused by unit nonresponse.

Further investigations are needed to overcome the problems identified for the proposed method in some settings (see LS). At this stage the procedure represent a valid fast and simple alternative to achieve an efficient stratification with a relatively small number of strata when having a small sample size is not a stringent goal (e.g. when oversampling should be performed to prevent reduction of sample size due to nonresponse).

References

Baillargeon S and Rivest L.-P. 2009 A general Algorithm for Univariate Stratification. *International Statistical Review*, 77: 331-344.

Baillargeon S and Rivest L.-P. 2011 The Construction of Stratified designs in R with the package stratification. *Survey Methodology*, 37: 53-65.

Baillargeon S and Rivest L.-P. 2014 stratification: Univariate Stratification of Survey Populations. R package version 2.2-5. http://CRAN.R-project.org/package=stratification

Ballin M and Barcaroli G. 2013. Joint Determination of optimal Stratification and Sample Allocation Using Genetic Algorithm, *Survey Methodology*, 39: 369-393

Barcaroli G. 2014. SamplingStrata: An R Package for the Optimization of Stratified Sampling. Journal of Statistical Software, 61(4), 1-24. URL http://www.jstatsoft.org/v61/i04/

Barrios E.B, Santos K.C.P. and Gauran I.I.M. (2013) Use of principal component score in sampling with multiple frames. *Proceedings 12th National Convention on Statistics*, Mandaluyong City, October 1-2, 2013.

Benedetti R, Espa G. and Lafratta G. 2008 A tree-based approach to forming strata in multipurpose business surveys. *Survey Methodology*, 34: 195-203.

Bethel J. 1989, Sample Allocation in Multivariate Surveys, *Survey Methodology*, 15: 47-57

Chromy J. 1987 Design Optimisation with Multiple Objectives, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 194-199

Cochran, W.G. 1977 *Sampling Techniques, 3rd Edition*, John Wiley & Sons, New York.

Dalenious T. and Hodges J.L. 1959. Minimum variance Stratification. *Journal of the American Statistical Association*, 54: 88-101.

Hagood M.J. and Bernert E.H. 1945 Component indexes as a basis for stratification in sampling. *Journal of the American Statistical Association*, 40: 330-341.

Hidiroglou M.A. 1986. The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40: 27-31.

Hidiroglou M.A. and Lavallée P. 2009 Sampling and Estimation in Business Surveys, in *Sample Surveys: Design, Methods and Applications, Vol. 29A*, Elsevier

Kish L. and Anderson D. W. 1978 Multivariate and Multipurpose Stratification. *Journal of the American Statistical Association*, 73: 24-34.

Kott P.S., Bailey J.T. 2000. The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling, *Proceedings of the International Conference on Establishment Surveys* (ICES-II), June 2000, Buffalo, New York.

Kozak M. 2004 Optimal Stratification Using Random Search Method in Agricultural Surveys, *Statistics in Transition*, 6: 797-806.

Lavalleé P. and Hidiroglou M.A. 1988 On the Stratification of Skewed Populations. *Survey Methodology*, 14: 33-43.

Pla L. 1991 Determining Stratum Boundaries with Multivariate Real Data. *Biometrics*, 47: 1409-1422.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rivest L.-P. 2002 A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28: 191-198.

Roshwalb A. and Wright R.L. 1991 Using information in addition to book value in sample designs for inventory cost estimation. *The Accounting Review*, 66: 348-360.

Särndal C.-E., Swensson B. and Wretman J. 1992 *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Appendix A
Correlation matrices

Table A.1 – Spearman's correlation coefficients between the *X* variables used for stratification purposes in EECPS.

	Legumes	Harv. green	Vegetables	Potatoes
Cereals	0.0975	-0.0904	-0.1439	-0.0557
Legumes		-0.0043	0.0519	0.1320
Harv. green			-0.1441	-0.0118
Vegetables				0.2045

Table A.2 – Spearman's correlation coefficients between the *X* variables used for stratification purposes in LS.

	Pigs	Sheep	Goats
Bovines	-0.1690	-0.4270	-0.2162
Pigs		0.0348	0.0284
Sheep			0.2282

Table A.3 – Spearman's correlation coefficients between the *X* variables used for stratification purposes in FSS.

	Industrial crops	Harvested green	Permanent grassland	Vineyards	Bovines	pigs	Poultry
Cereals	0.0285	-0.0112	-0.1744	-0.1911	0.0581	0.0551	0.0279
Industrial crops		-0.0389	-0.1326	-0.0726	-0.0485	-0.0038	-0.0023
Harvested green			0.0533	-0.0178	0.3365	0.0700	0.0326
Permanent grassland				-0.0018	0.3305	0.0667	0.0536
Vineyards					0.0169	0.0244	0.0009
Bovines						0.1701	0.0668
Pigs							0.2410