A New Approach for Multipurpose Stratification in Agriculture Surveys

Marcello D'Orazio, Elena Catanese[1]

Abstract

The stratified random sampling is frequently used in sample surveys on businesses and farms because of its efficiency and practical advantages. The stratification of the target population is a crucial step; it is based on the information available in the sampling frame being related to the phenomena under investigation. The task is not straightforward in multipurpose surveys, where different phenomena are investigated at the same time. Several stratification criteria can be applied when dealing with a single auxiliary variable while few methods are available to tackle the problem in the presence of a set of many auxiliary variables.

This paper shows a relatively new simple procedure to stratify the sampling frame in presence of a set of continuous auxiliary variables. Its main advantages and disadvantages are highlighted. The procedure is compared with one of the reference methods by applying both to the design of some sample surveys in the agricultural sector.

Key Words: stratified random sampling, multivariate stratification.

1. Introduction

Traditionally sample surveys on enterprises and farms are based on one stage stratified sampling; it consists in partitioning the sampling frame into non-overlapping *subpopulations* (or *strata*) and selecting an independent sample in each subpopulation (*stratum*). If strata consist of units being homogenous with respect to the investigated phenomena then the stratified sampling allows for reduction of the sampling error and permits to derive reliable estimates for each subpopulation (Cochran, 1977). In agriculture surveys, usually homogeneous strata can be achieved by partitioning the farms according to geographical information, type of farming (specialist crops, specialist livestock, mixed) and some measures of farm's size (e.g. size of areas with crops, livestock, etc.). The variables for stratification purposes should be chosen among those available in the sampling frame (e.g. Register of active farms, Administrative register, the previous Census data); the more the auxiliary variables are correlated with the target variables, the higher will be the benefits in using them in stratification. It is worth noting that, for practical purposes it would be preferable to carry out stratification -so to derive estimation domains by simple aggregation of elementary strata.

Stratifying a population does not pose problem when it is performed through categorical variables (e.g. geographical regions), while eventual continuous auxiliary variables need to be categorized in advance. Multipurpose surveys pose additional problems; the main

[1] Elena Catanese (email: catanese@istat.it). Marcello D'Orazio (email: madorazi@istat.it), Italian National Institute of Statistics, Via C. Balbo, 16, Rome, Italy.

difficulty consists in creating strata of units being homogenous with respect to the different phenomena to study. Moreover, the sampling frame may provide several auxiliary variables correlated with the target ones but uncorrelated with respect to each other; thus increasing difficulties in choosing the stratification variables.

This paper tackles the problem of stratifying a population in presence of a set of continuous auxiliary variables, by exploring a relatively new procedure, illustrated in Section 3. 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 in Section 4. Main features and notation of stratified random sampling design are provided in Section 2.

2. Stratified Sampling

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

Stratification allows for different independent selection schemes in each subpopulation; the common practice in business and agriculture surveys consists in applying *simple random sampling without replacement* in all the strata, because of its practical and theoretical advantages. The sample size is decided according to the desired precision for the main survey estimates (expressed in relative terms: desired sampling error divided by the quantity to estimate, denoted usually as CV). For instance, in European Union (EU) Countries the desired CVs in estimating the total amount for the main variables through agriculture surveys (e.g. Farm Structure Survey) are explicitly listed in EU regulations. The allocation of sample among the strata may follow different rules: equal allocation, proportional allocation, Neyman allocation, power allocation etc. The choice is related to the desired precision characterizing the final survey estimates and to the stratification strategy.

2.1 Main characteristics of stratified random sampling

Let U be the finite population under investigation, consisting of N units. At first, U 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, $N = \sum_{h=1}^{H} N_h$. Then, a simple random sample without replacement, s_h of n_h ($n_h \leq N_h$) units is selected independently stratum by stratum; the overall sample size is $n = \sum_{h=1}^{H} n_h$. An estimate of the total amount of the target variable Y in U, $t_y = \sum_{k \in U} y_k$, is provided by:

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

And the associated sampling variance is:

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

It is usually expressed in relative terms: $CV_{t_y}=\sqrt{V\left(\hat{t}_y\right)}\Big/t_y$, known as *relative standard error*.

By fixing in advance the desired relative error in estimating the total amount of a continuous variable *Y*, it is possible to determine the required sample size n_{opt} (see e.g. formulas in Section 5.9 in Cochran, 1977). The partitioning of n_{opt} among the strata can follow different criteria; in - *proportional allocation* the stratum sampling fraction is set equal to the stratum relative size, $n_h=n_{opt}N_h/N$ so to ensure equal inclusion probabilities to all the units in *U*; in *optimal allocation* (or *Neyman allocation*) the sampling fraction is higher in more heterogeneous strata, being $n_h \propto N_h S_h$; *power allocation* (cf. Särndal et al., 1992, pp. 470-471) is a compromise between the Neyman and an allocation ensuring constant precision for each of the strata estimates. The derivation of n_{opt} and its allocation between the strata requires information concerning the *Y* variable, usually not known in advance. For this reason it is considered an auxiliary variable *X*, known for all the units in the population and being highly correlated with *Y*.

In multipurpose surveys the same sample should provide estimates for several target variables fulfilling the precision requirements (CVs); for this reason, the decisions concerning the overall sample size and the corresponding allocation should be approached in a multivariate framework by setting up a convex mathematical programming problem; Chromy (1987) and Bethel (1989) provide solutions to this problem.

In the traditional approach to stratified sampling, the sample size and its allocation among the strata are derived -given –a certain stratification of *U*, i.e. once decided *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 step should be necessarily the stratification of the target population.

2.2 Stratification of the Population in the Univariate Case

Consider a single continuous target variable *Y*; an efficient stratification should try to derive strata as homogeneous as possible in terms of *Y* values. Unfortunately, the values of *Y* are not known in advance and consequently the stratification is carried out on an auxiliary variable *X* whose values are known for all the units in the population and - strictly related with *Y*. When *X* is a categorical variable (e.g. geographical regions, NACE in case of business surveys or Farm Typology as far as farms are concerned) the stratification is straightforward: strata are formed by units with the same (or similar) *X* category. On the contrary, categorization is needed when *X* is continuous.

Different criteria are available for categorizing a continuous *X* variable for stratification purposes. A widespread criterion is the *cumulative* $\sqrt{f}$ *rule* (Dalenious and Hodges, 1959). Unfortunately it performs poorly when *X* shows a highly skewed distribution; a frequent scenario in business and agriculture surveys, where most of the continuous variables presents high positive skewness. This problem can be solved by separating the

few large units in a specific stratum; all these units are included with certainty in the sample (so called *take-all* stratum; cf. Hidiroglou and Lavallée, 2009); in practice, given that these units contribute at large extent to the total amount in 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 for identifying the take-all stratum (all the units with $x_k > b_c$); the procedure requires setting the desired CV. Once identified the take-all stratum, the remaining units can be further stratified according the *cum* $\sqrt{f}$ *rule* (or other criteria).

Lavalleé and Hidiroglou (1988) introduced a unified procedure for both identifying the take-all stratum and stratifying the remaining units. Once decided the desired level of precision (CV), the procedure provides a stratification that minimizes the overall sample size. The partitioning of the sample between the take-some strata follows the power allocation criterion. Unfortunately, the Lavalleé and Hidiroglou procedure is based on an iterative algorithm which may not converge to a global minimum; this problem can partly be solved by applying the Kozak (2004) algorithm.

To overcome the problem of an allocation performed on a variable, *X*, assumed to be correlated with the unknown target one (*Y*), Rivest (2002) suggested using anticipated moments of *Y* given *X* in the Lavalleé and Hidiroglou procedure. Recently, Baillargeon and Rivest (2009) introduced the possibility of separating very small units in a stratum that is not sampled (*take-none* stratum); these units 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, 2016).

The stratification problem can also be approached in a model based framework (cf. Särndal et al., 1992, Section 12.4). In particular, if it is assumed 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), then the stratification can be performed by grouping units with similar values of the model variance $V_\xi(y_k)$. 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:

$$\pi_k = n \frac{\sqrt{V_\xi(y_k)}}{\sum_{k \in U} \sqrt{V_\xi(y_k)}} = n \frac{x_k^{\gamma/2}}{\sum_{k \in U} x_k^{\gamma/2}} \qquad [3]$$

where *n* is the expected sample size. A simple fixed-size design which maintains $\pi_k \propto x_k^{\gamma/2}$ is a stratified random sampling design where: (i) the *H* strata are formed by applying 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

$(0,2]$; 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 probabilities of *probability proportional to size* (PPS) sampling, $\pi_k = n\ x_k / t_x$.

2.3 Stratification of the Population based on Many Variables

Stratification becomes more complex in a multipurpose survey with many target variables not necessarily related to one each other. In this case there may be a high number of auxiliary variables *X* related - differently - with the various target ones; stratification based just on a *X* variable may not be efficient for all the target variables. According to Kish and Anderson (1978) the advantages of using several stratification variables are greater in multipurpose surveys, but potential gains depend on the (i) the relationship between the stratification variables and the target ones, and (ii) intercorrelations among the stratification variables.

The 'traditional' strategy to carry out stratification in presence of a large set of continuous *X* variables consists in: 1) selecting the *X* variables highly correlated with most of the target ones; 2) performing univariate stratification on each of the selected *X*s, and, then 3) deriving the final stratification by cross-classifying units according to the chosen categorized *X* variables. Parsimony should be the guiding principle in step (1), the chosen variables should not be related to each other's (at least not more than weakly); moreover, in presence of a set of highly correlated *X* variables, it would be preferable to select just the one with higher relative variability, to avoid any lack of information. This strategy can determine too many strata, with some of them too small in terms of size.

In literature there are other proposals to perform stratification in the multivariate framework. For instance, in the bivariate case Kish and Anderson (1978) suggested to apply the *cum* $\sqrt{f}$ *rule* independently on each of the *X* variables; then the final stratification as a combination of the two results. An extension of the model based stratification to the bivariate case can be found in Roshwalb and Wright (1991).

When dealing with more than two stratification variables Hagood and Bernert (1945) suggested performing stratification on a subset of the first principal components computed starting from the set of the *X*s. Pla (1991) suggested considering just the first component. Kish and Anderson (1978) warned that principal components approach may provide final strata that are not readily interpretable; moreover principal components analysis (PCA) consider just intercorrelations among the stratification variables and not their relationship with the target variables. Barrios *at al* (2013) noted that PCA is not suitable for high skewed variables with few units exhibiting very high values. Performing PCA on the log-transformed variables may not solve the problem.

Benedetti *et al* (2008) suggested a unique procedure tackling both stratification and sample allocation in a multivariate framework. This procedure requires setting the desired CVs for proxies of the target variables, then a tree-based approach identifies finer and finer partitions of the units by minimizing at each step the overall sample size.

A similar approach is suggested by Ballin and Barcaroli (2013). Their sequential procedure starts with a very fine stratification and then iteratively collapses the strata with the objective of minimizing the overall sample size given the target precision (CVs) required for a set of proxy variables (can be the same auxiliary variables used to create

the initial fine partition) under the optimal Bethel allocation. The proposed procedure makes use of a 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 size given the target CVs, however the identified final stratification, obtained by subsequent collapsing steps of intermediate strata, is not readily interpretable. Moreover the procedure requires a subjective choice concerning the initial stratification; a possible starting point can be the stratification obtained by cross-classifying the chosen *X* variables conveniently categorized. It requires setting a high number of input parameters and a high number of iterations is necessary to achieve valuable final results, thus implying a non-negligible computational effort.

It is worth noting that recently Ballin *et al* (2016) explored the problem of stratification of a sampling frame in the multivariate setting by using the functionalities of the R environment. This paper illustrates an interesting comparison between the 'traditional' approach to multivariate stratification and the Barcaroli and Ballin (2013) one.

3. A New Procedure for Stratification in a Multivariate Setting

Recently d D'Orazio and Catanese (2016) suggested a new procedure to tackle the problem of stratification in presence of a series of auxiliary variables, supposed to be related to the target ones. The procedure follows the same reasoning of model based stratification in the univariate case, but the stratification is performed on the inclusion probabilities obtained by applying the *Maximal Brewer Selection* (MBS; also known as *Multivariate Probability Proportional to Size*, MPPS) (Kott and Bailey, 2000). In particular, stratification is obtained by applying the *equal aggregate* σ *-rule* to the probabilities

$$\pi_k^* = \min\left\{1,\ \max_j\left[\pi_{1,k},,\ldots\pi_{j,k},\ldots,\pi_{J,k}\right]\right\},\quad k=1,2,\ldots,N \qquad [4]$$

Where

$$\pi_{j,k} = n_j \frac{x_{j,k}^{\gamma/2}}{\sum_{k\in U} x_{j,k}^{\gamma/2}},\qquad j=1,2,\ldots,J\ ;\quad k=1,2,\ldots,N \qquad [5]$$

In practice, to derive the π_k^*, it is necessary to set in advance the "target" sample size n_j for each of the J ($J \geq 2$) auxiliary variables being considered (cf. Kott and Bailey, 2000). A simplifying choice consists in setting $n_j = n_0$ ($0 < n_0 < N$) for $j = 1,2,\ldots,J$ (i.e. a constant value). As a consequence the stratification is performed directly on the values:

$$z_k = \max_j\left[\frac{x_{1k}^{\gamma/2}}{\sum_{k=1}^N x_{1k}^{\gamma/2}},\ldots,\frac{x_{jk}^{\gamma/2}}{\sum_{k=1}^N x_{jk}^{\gamma/2}},\ldots,\frac{x_{Jk}^{\gamma/2}}{\sum_{k=1}^N x_{Jk}^{\gamma/2}}\right],\quad k=1,2,\ldots,N \qquad [6]$$

Where the constant 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 $\gamma = 3/2$.

Once stratified the units in the desired *H* strata, the overall sample size and the corresponding allocation is determined by applying the Bethel algorithm.

3.1 A Simulation Study

The efficiency of the D'Orazio and Catanese (2016) procedure (DC hereafter) is investigated through a series of simulations carried out with real data of agricultural holdings. Moreover, comparisons with the procedure suggested by Ballin and Barcaroli (2013) (BB hereafter) are performed. The simulation study considers three sample surveys carried out on Italian Agriculture holdings: (i) the annual Early Estimates for Crop Products Survey (EECPS); (ii) the livestock survey (LS), carried out twice a year; and (iii) the Farm Structure Survey (FSS), carried out every three years, where both livestock and crops are investigated. In practice DC and BB stratification 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 data used for stratification and allocation purposes are those collected in the 2010 Census occasion.

Table 1 provides the desired CVs which should be achieved, as explicitly mentioned in the corresponding EU regulations at either national, or regional (NUTS1 or NUTS2) level. For the sake of the present work, for EECPS the sampling frame consists of 282017 agricultural holdings having at least one hectare of crops (target population) in the south and islands of Italy (one NUTS1 region). As far as LS is concerned, the considered sampling frame includes all the Italian farms having at least one head of bovine animals, pigs, sheep and goats, thus consisting of 173617 farms. Finally, in FSS case all the farms of Veneto (one NUTS2 region) are considered (119384 farms); here, according to EU regulations, there are 5 target variables in terms of crop aggregates and 3 for livestock. The Table 1 provides the list of the variables observed in the 2010 Agriculture Census that in the simulations were used for stratification purposes (*X*) or as survey target variables (*Y*) (a variable can be used both for stratification and as proxy of the target one); for each *Y* variables it is also reported the associated desired CV.

Table 1: Auxiliary and target variables used for stratification and design purposes.

EECPS			*LS*			*FSS*		
X (areas in ha)	*Y (areas in ha)*	*CVs*	*X (No. animals)*	*Y (No. animals)*	*CVs*	*X*	*Y*	*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	Harvest. green	Harvest. green	0.05
Legumes	Legumes	0.03	Sheep	Sheep	0.020	Perm. grassland	Perm. grassland	0.05
Harvest. green	Harvest. green	0.03	Goats	Goats	0.050	Vineyards	Vineyards	0.05
Vegetab.	Tomatoes	0.03				Bovines	Dairy cows	0.05
Potatoes	Potatoes	0.03					Other bovines	0.05
						Pigs	Pigs	0.05

		Poultry	Poultry	0.05

Table 2 summarizes the main results of the simulation study in terms of the achieved overall optimal sample size, given *H*: a stratification and allocation procedure that achieves the same target CVs with a smaller sample size is obviously the preferred one.

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

EECPS			*LS*			*FSS*		
H	*DC*	*BB*	*H*	*DC*	*BB*	*H*	*DC*	*BB*
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

Results are not -homogeneous with respect to the surveys: in designing the EECPS the DC procedure is very efficient and performs better than BB in almost all cases with the exception of $H = 150$. In the FSS case, the BB procedure performs always better than the DC and the distance in terms of final sample size increases as the total number of strata grows. Finally, the BB procedure outperforms DC in LS, in this case a finer stratification (i.e. increasing *H*) does not imply a reduction of the final sample size.

In general, it seems that summarizing a high number of variables with a unique score (*Z*) subsequently used for stratification purposes may not be a good solution when the variables have different nature as in FSS (areas and animals), but this is not the only reason, given that in LS all the *X* variables refer to animals. A possible explanation to this situation has to be searched by exploring at the inter-correlations between the *X*s. In particular, in the LS case (DCworstperformance) it can be seen that 'Bovines' variable is negatively correlated with all the remaining ones (Table 3). This situation suggests –to test the DC stratification strategy by applying it separately at two score variables: Z_1 derived starting just from 'Bovines' *X* variable, and Z_2 derived summarizing the remaining variables ('Pigs', 'Sheep', 'Goats') through the expression [6].

Table 3: Spearman's correlation coefficients between stratification variables in the LS

	Pigs	*Sheep*	*Goats*
Bovines	-0.17	-0.43	-0.22
Pigs		0.03	0.03
Sheep			0.23

The procedure remains the same as in DC but final strata are derived by crossing the results of univariate stratification performed independently on Z_1 and Z_2. This new

strategy improves markedly the performances of the DC procedure in LS case, as shown in the Table 4; however the BB procedure still remains the best in terms of final overall sample size necessary to fulfill the CVs constraint, for any *H*.

Table 4: Performances of the revised DC procedure in LS

H	*DC* *One Z variable*	*DC* *Two Z variables*	*BB*
50	11 885	4 528	3 163
75	11 517	4 258	3 130
85	11 277	4 173	3 127
110	11 160	4 085	3 109

4. Conclusions

The paper deals with 'multivariate' stratification and allocation procedures in the presence of a set of continuous stratification variables. The new procedure presented in D'Orazio and Catanese (2016) is further investigated for the cases where some special attention should be kept in creating a *Z* variable needed for stratification purposes. In particular, while this method is very efficient when the variables are positively correlated or uncorrelated, if one of the initial auxiliary variables is negatively correlated to all the others, results can be very poor. A very simple solution to tackle this issue is proposed here. The whole procedure is effective because permits to overcome the problem of choosing a small subset of auxiliary variables (2 or 3 in practice in most cases) to perform separately univariate stratification, by allowing to create only one composite variable from an abundant set of variables (7 for instance in the EECPS case). Moreover the procedure is very simple, with a negligible computational effort. The results obtained for the design of samples of three agriculture surveys seem promising, and as shown, a marked improvement in some cases can be achieved with a relative additional effort. In any case, the procedure proposed by Ballin and Barcaroli (2013) remains the best if the focus is the reduction of the overall sample size. The price to pay is in general a higher number of final strata and a non-negligible computational effort. It is worth noting that both the DC and BB procedures provide final strata which are not are not readily interpretable, this is an unpleased feature for subject matter experts and may create problems when, after data collection, strata collapsing should be performed to compensate for empty strata caused by unit nonresponse.

At this stage the stratification of the transformed variables is performed by using the equal aggregate σ-rule, improvements are likely to be achieved by using more advanced univariate stratification procedures like the Lavalle-Hidiroglou (2009) one. Generally speaking, the new stratification procedure represents 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) and in presence of many target variables where no evident negative correlation among auxiliary variables is present.

References

Baillargeon, S. and Rivest. L.-P. (2009) "A general Algorithm for Univariate Stratification". *Inter-national Statistical Review*, 77, pp. 331-344.

Baillargeon, S. and Rivest, L.-P. (2011) "The Construction of Stratified designs in R with the package stratification". *Survey Methodology*, 37, pp. 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., Catanese, E. and D'Orazio, M. (2016) "Stratification in Business and Agriculture Surveys with R". *Romanian Statistical Review*, 2/2016, pp. 43-58.

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

Barcaroli, G. (2014) "SamplingStrata: An R Package for the Optimization of Stratified Sampling". *Journal of Statistical Software*, 61, pp. 1-24. 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, Man-daluyong City, October 1-2, 2013.

Benedetti, R, Espa, G. and Lafratta, G. (2008) "A tree-based approach to forming strata in multi-purpose business surveys". *Survey Methodology*, 34, pp. 195-203.

Bethel, J. (1989) "Sample Allocation in Multivariate Surveys", *Survey Methodology*, 15, pp. 47-57.

Chromy, J. (1987) "Design Optimisation with Multiple Objectives", Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 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, pp. 88-101.

D'Orazio, M. and Catanese, E. (2016) "A simple approach for stratification of units in multipurpose in business and agriculture surveys". Istat *Working Papers*, 10/2016.

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

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

Hidiroglou, M.A. and Lavallée, P. (2009) "Sampling and Estimation in Business Surveys", in D. Pfeffermann and Rao C.R. (eds.) *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, pp. 24-34.

Kott, P.S. and Bailey, J.T. (2000). "The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling". Proceedings of the Second 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, pp. 797-806.

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

Pla, L. (1991) "Determining Stratum Boundaries with Multivariate Real Data". *Biometrics*, 47, pp. 1409-1422.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 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, pp. 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, pp. 348-360.

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