Journal of Official Statistics, Vol. 23, No. 1, 2007, pp. 35–50

Efficient Stratification Based on Nonparametric Regression Methods

Enrico Fabrizi[1] *and Carlo Trivisano*[2]

Classical rules for optimal one-way stratification, such as the Dalenius and Hodges rule, are applied under the assumption that a single stratification variable is to be used. In this article, we consider an information setting in which a set of candidate stratification variables is available and a proxy of the target variable (or the target variable itself) is known for a random sample of units from the population. Under these assumptions, we propose various extensions of the Dalenius and Hodges rule based either on linear prediction or on nonparametric regression methods. The resulting stratification rules are compared by means of a Monte Carlo exercise based on a set of pseudo-populations covering a wide range of possible forms of relationship between the target and the stratification variables. The application of regression trees as stratification rules, an option that may be intuitively appealing in the considered information setting, is also discussed.

Key words: Dalenius and Hodges rule; one-way optimal stratification; regression trees; additive models; MARS; boosted regression trees.

1. Introduction

There are several reasons to divide a population U into H strata: practical or administrative constraints, the need to obtain estimates of known precision for some subpopulations or because sampling problems are markedly different in different parts of the population. Stratification may also be used to improve the efficiency of estimators of population descriptive quantities. In fact, if we assume simple random sampling within each stratum, we have that for the estimation of, say, the population total t_y of a target variable y, the Horwitz-Thompson estimator associated with stratified sampling ($\hat{t}_{y,s}$) is much more efficient than the expansion estimator, whenever variance within strata is small as compared with variance between strata.

In the simplest case, a univariate auxiliary stratification variable x (known for each unit in the population) is used, and stratification consists in the definition of an ordered

[1] DMSIA, Università di Bergamo, Via dei Caniana 2, 24127 Bergamo, Italy. Email: enrico.fabrizi@unibg.it
[2] Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, Via Belle Arti 41, 40126 Bologna, Italy. Email: trivi@stat.unibo.it
Acknowledgments: We would like to thank the Associate Editor and three anonymous referees for their useful comments and suggestions which allowed us to significantly improve the article. We also thank Daniela Cocchi and Meri Raggi of the Department of Statistics, University of Bologna, for discussing with us the subjects of this research. The work of Enrico Fabrizi was partially supported by the grants 60FABR06 and 60BIFF04, University of Bergamo. The work of Carlo Trivisano was partially supported by the grant n. 2004137478.001, PRIN 2004.

sequence of $H - 1$ boundary points,

$$x_{INF} < x_1 < x_2 < \ldots < x_{H-1} < x_{SUP}$$

which partition the domain of x and hence the N observations of population U into H groups. General stratification can be based on several auxiliary variables.

Whenever there is some latitude in the definition of strata, that is when stratification is not driven solely by practical and administrative considerations, and we have a single target variable y, it is sensible to look for an efficient stratification in terms of $V(\hat{t}_{y,s})$. This problem is known as one-way optimal stratification and has been widely debated in the literature. In passing, we note that in many situations optimal methods are applied within subpopulations identified by other stratification criteria such as geographical areas. A classical theoretical result in the field of optimal one-way stratification, under Neyman allocation and the assumption of stratification based on the target variable ($x \equiv y$), was worked out by Dalenius (1957; also in Cochran 1977, pp. 128–131). An approximated method based on the Dalenius result was then introduced by Dalenius and Hodges (1959), widely known as the $cum\sqrt{f}$ rule. An alternative to the Dalenius and Hodges approximation is given by the Ekman rule (Ekman 1959; Hedlin 2000). The merits of the two approximations are compared in Cochran (1961), Hess, Sethi, and Balakrishnan (1966) and Murthy (1967).

Later contributions introduced model-based or model-assisted stratification methods; they are discussed for instance in Singh (1971), Wright (1983), and Sweet and Sigman (1995). We note that many works consider the special topic of stratifying highly skewed populations (Lavallée and Hidiroglou 1988; Sigman and Monsour 1995). Rivest (2002) proposes stratification algorithms in which the discrepancy between the stratification and the study variable is accounted for. He proposes different models for the relationship between y and x but he does not deal with the problem of estimating model parameters and considers only the case of a single stratification variable. We note that most of these contributions rely on the assumption (up to an error term with known distribution) that either ($x \equiv y$) or the relationship linking x and y is known.

In practice, none of these assumptions hold, and the optimal stratification will be only approximated, the quality of the approximation depending on how strongly the selected stratification variable is correlated with the target variable y or on the adequacy of the assumption of the relationship between y and x.

We consider the following information setting. Suppose that a matrix $\mathbf{X}_U = \{x_{ij}, i \in U, j = 1, \ldots, p\}$ of potential stratification variables is known. Of course the realized values of the target variable $\mathbf{y}_U = (y_1, \ldots, y_N)$ are unknown before the survey is conducted, but we assume that a proxy y^* of the target variable y is observed for a random sample $S \subset U$: that is, $\mathbf{y}_s^* = \{y_i^* : i \in S\}$ is observed.

This situation may arise for instance in repeated surveys or when data from some pilot survey on a similar subject are available. In particular, in the case of pilot surveys we have that $y_i = y_i^*, i \in S$.

In this setting, it is reasonable to use sample data $d_s = \{(y_i^*, x_{1i}, \ldots, x_{pi}), i \in S\}$ to select the best stratification variable x or to estimate the relationship between y and the set of auxiliary variables $\boldsymbol{x} = \{x_1, \ldots, x_p\}$. This latter option is what we deal with in this article.

Our basic idea is to predict the values of y_i for each unit i in the population using the information contained in d_s and then use this approximated target variable to stratify the population in an efficient way. Note that "efficient" is used in its informal meaning and that the stratification methods we propose are not optimal in a strict mathematical sense. In particular, we will illustrate extensions of the $cum\sqrt{f}$ rule, but similar arguments can be put forward for the Ekman's or any other rule based on a single stratification variable.

To obtain the predicted values, $\hat{y}_i$ $i \in U$, we consider general regression models and nonparametric fitting methods that provide flexible tools for obtaining good predictions for the y_is regardless of whether the relationship between y and the auxiliary variables is linear or nonlinear.

The stratification method we propose is best suited to stratifying populations where a single variable is assumed as target and a set of auxiliaries with good predictive power is available despite a possible complex nonlinear relationship between y and $\boldsymbol{x}$. Moreover, as far as it is based on the Dalenius and Hodges rule, our method may be not suitable for the stratification of highly skewed populations such as those arising in business surveys.

More in detail, the article is organized as follows.

In Section 2, we sketch the basic ideas behind the classical Dalenius and Hodges rule and introduce a simple extension of it, in the information setting just introduced, based on the linearly predicted values of y_i, $i \in U$, using a linear model fitted on d_s. This stratification method will then be kept as a benchmark against which all other methods based on nonparametric regression techniques will be tested.

Section 3 is devoted to the discussion of regression trees as a stratification method. In the discussed information scenario they may in fact be seen as an intuitive and appealing stratification method, since the output obtained by this nonparametric regression method is a partition of the predictors' space into multivariate rectangles which actually represents a stratification of the population. In Section 4 modifications of the Dalenius and Hodges rule based on different nonparametric regression methods (Additive Models, Multivariate Adaptive Regression Splines and Boosting Regression) are introduced, and compared with the one based on linear prediction.

In Section 5, a simulation exercise that we use for comparisons is outlined. In this simulation, the relationship between y and $\boldsymbol{x}$, although nonlinear, is characterized by homoskedastic residuals. The case of heteroskedastic residuals, considered for instance in Bethel (1989), may be relevant in many practical situations but is not considered here. In Section 6, some complementary topics, such as the choice of the number of strata and sensitivity analysis of boundary determination to the adopted regression methods, are discussed. A final Section 7 contains some concluding remarks and outlines some direction for future research.

2. The Dalenius-Hodges Rule and the Proposed Modifications

To solve the one-way optimal stratification problem we have to answer three basic questions: i) which variable (or set of variables) is the best for defining strata?; ii) how to define boundaries between strata?; iii) how many strata should there be?

The answer to the first question is clearly the target variable y itself. Since y is not known for all units in the population before the survey is conducted, a common practical

alternative is to use an arbitrarily chosen "highly correlated" auxiliary variable x known for all units in the population.

Many classical theoretical results assume that a single x on which to base the actual stratification is available. We note that, in most practical situations no single highly correlated variable can be easily identified, but we have a vector $\boldsymbol{x}$ of candidate stratification variables, none of which is clearly better than the others in terms of correlation with y.

As far as the determination of boundaries between strata is concerned, we focus on a classical method, still widely used: the Dalenius and Hodges rule, known also as the $cum\sqrt{f}$ rule. A description of this method may be found, for instance, in Särndal, Swensson, and Wretman (1992, pp. 462–464).

As regards the choice of the optimal number of strata, Cochran (1977) points out that in principle, if the stratification is based on y, the multiplication of strata is beneficial but, if an auxiliary variable x is used instead, the variance reduction induced by the $cum\sqrt{f}$ rule has a global maximum for a moderate number of strata. We will discuss this topic more in detail in Sections 5.2 and 6.

The Dalenius and Hodges rule is based on two basic elements: i) an auxiliary variable $\boldsymbol{x}$ highly correlated with y, that is, a linear approximation of y; ii) the $cum\sqrt{f}$ rule.

A useful extension of the rule is as follows. Use the sample information d_s to fit the linear model:

$$y_i^* = \alpha + \mathbf{x}_i^T \beta + e_i,\ i \in U$$

with $E(e_i|\mathbf{x}_i) = 0,\ V(e_i|\mathbf{x}_i) = \sigma^2$ and use the estimated parameters to predict y_i with $\hat{y}_i = \hat{\alpha} + \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}\, \forall i \in U$. If necessary apply the ordinary $cum\sqrt{f}$ rule to $\hat{\boldsymbol{y}}_U = (\hat{y}_1, \ldots, \hat{y}_N)$. We will refer to this stratification method as the Linear Prediction Dalenius-Hodges rule (LPDH). The idea is used also in Hidiroglou and Laniel (2001), who generalize the Lavallée and Hidiroglou (1988) algorithm in a similar way.

The assumption of a linear relationship between the stratification and the target variable is realistic in many cases, but may fail in others. Thus, a natural generalization of the LPDH rule consists in assuming the more general model

$$y_i^* = g(\mathbf{x}_i) + e_i \quad E(e_i|\mathbf{x}_i) = 0 \quad V(e_i|\mathbf{x}_i) = \sigma^2 \tag{2.1}$$

where g is an unknown regression function. This model is to be fitted by some nonparametric algorithm. Once the predicted values $\hat{y}_i$ are obtained by means of $\forall i \in U$, they can be used for stratification with the $cum\sqrt{f}$ rule.

3. Regression Trees as Stratification Tools

Before discussing direct generalizations of the LPDH rule along the lines of Section 2, we now illustrate the use of regression trees for stratification since they can be a natural and appealing method for stratifying a population in the informative setting considered in this article. Moreover, they do not require the application of the $cum\sqrt{f}$ rule since their output is a partition of the predictors' space that can be straightforwardly used for stratification.

In principle, the use of regression trees has several advantages over the LPDH rule: the problems of determining the number of strata and the boundaries between them are solved

by the same algorithm; no assumptions are introduced on the type of relation between y and $\boldsymbol{x}$, and the definition of boundaries is independent of allocation of the sample units to the strata. Moreover, strata are given by multidimensional rectangles and are therefore easy to interpret.

Regression trees are based on the assumption of a general regression model as in (2.1) in which the regression function g is estimated by means of a multivariate step function:

$$\hat{g}_i(d_s) = \sum_{l=1}^{H} \bar{y}_l^* \mathbf{1}(\mathbf{x}_i \in r_l) \quad i = 1, \ldots, n \tag{3.1}$$

where the H sets r_l are the multivariate regions (i.e., strata) into which the predictors' space is partitioned by the tree algorithm, $\bar{y}_l^* = |r_l|^{-1} \sum_{j \in r_l} y_j^*$ and $|r_l|$ is the number of elements in r_l.

The estimator $\hat{g}$ is obtained by a recursive partitioning algorithm introduced by Breiman et al. (1984, pp. 228–237) in which at each step a group of observations is bi-partitioned in order to maximise the "between" deviance of the newly created subgroups.

The correct size of the tree is usually determined by bi-partitioning the observations until a stopping criterion is met (e.g., a lower threshold for the number of observations in the group to be partitioned). Then the tree is pruned in order to find the optimal tree in terms of prediction error in the following way. The conditional prediction error for observation i in fitting model (2.1) can be decomposed as:

$$E(y_i^* - \hat{g}_i(\mathbf{x}_i))^2 = \{g(\mathbf{x}_i) - E[\hat{g}(\mathbf{x}_i)]\}^2 + E\{\hat{g}(\mathbf{x}_i) - E[\hat{g}(\mathbf{x}_i)]\}^2$$

(see Hastie et al. 2003, par. 7.3), where $i \in U$, and $\hat{g}(\mathbf{x}_i)$ is the prediction of y_i^* based on (3.1). The first term, the squared bias, measures the average distance between the approximating and the true regression function and is therefore a measure of accuracy. The second term is the sampling variance of $\hat{g}$. Regression trees are usually characterized by a sampling variance much larger than the bias (Breiman 1998).

The right-sized tree is identified by means of a pruning rule which can be described as a tool to balance the trade-off between bias and variance. As a consequence the number of strata determined by the regression tree is sample dependent.

4. Generalizations of the LPDH Based on Nonparametric Regression Methods

We consider three further methods of fitting (2.1) to sample data d_s. These are very popular in the applied nonparametric literature: Additive regression Models (AM), Multivariate Adaptive Regression Splines (MARS) and BOOSTed regression trees (BOOST). We selected these three methods among the many proposed in the literature because they can be interpreted as generalizations of either the linear model, on which the LPDH rule is based, or the regression trees discussed in Section 3. The AM generalize the linear additive models, while MARS and BOOST can both be viewed as methods intended to robustify and stabilize regression trees.

The predicted values of y_i, for all $i \in U$, will then be used to construct stratification rules based on the $cum\sqrt{f}$ rule. The stratification methods consisting of fitting (2.1)

followed by application of the $cum\sqrt{f}$ rule will be denoted LPDH (defined in Section 2), AMDH, MARSDH and BOOSTDH.

A detailed description of the nonparametric regression methods AM, MARS and BOOST is beyond the scope of this article; for an excellent and comprehensive introduction to them see Hastie et al. (2003, Chapters 9, 10). Let us sketch the basic ideas underlying the various methods and the chosen options for their implementation.

The AM algorithm approximates a nonlinear relationship by means of a sum of arbitrary smooth univariate functions. It estimates the regression function g by means of the following functions:

$$\hat{g}_i(d_s) = \hat{\alpha} + \sum_{j=1}^{p} \hat{g}_j^*(x_{ij})$$

where $i \in S$ and p is, as before, the number of auxiliary variables. We set the $\hat{g}_j^* s$ to be estimated cubic splines.

MARS can be viewed as a generalization of stepwise linear regression or a modification of Regression Trees to improve the performances of these methods. It estimates g by means of the following functions:

$$\hat{g}_i(d_s) = \hat{\alpha} + \sum_{m=1}^{M} \hat{\gamma}_m h_m(\mathbf{x}_i)$$

where $h_m(\mathbf{x}_i)$ are functions or products of k functions in $\mathbb{C}$, where $\mathbb{C}$ is the collection of piecewise linear basis functions. For further details see Hastie et al. (2003, p. 283). We set $k = 2$ and $M = 21$. The $\hat{\gamma}_m$ are estimated regression coefficients.

It is known that MARS automatically accommodates interactions between variables and variable selection and is well suited to high-dimensional problems (Friedman 1991). The technique of boosting is one of the most powerful tools introduced in the literature on nonparametric regression in recent years and it is based on combining results of many "weak" regression methods (in most cases trees) to create a more powerful predictor. In particular, boosting M regression trees reduces their potentially large individual variances. Among the many available boosting algorithms, we consider the gradient boosting (Friedman 2001). It estimates g by means of the following set of functions:

$$\hat{g}_i(d_s) = \sum_{m=1}^{M} \hat{\beta}_m \hat{g}_{im}(d_s) \quad i = 1, \ldots, n$$

where $\hat{g}_{i1}$ is a regression tree based prediction as in (3.1) while $\hat{g}_{im}$ $(m \geq 2)$ are regression trees recursively defined in order to minimize the squared sum of the residuals obtained from regression already calculated. The $\hat{\beta}_m$ are estimated weights designed to optimize the combination of the M trees in terms of squared prediction error. We set $M = 500$ and the number of terminal nodes for each tree to 6.

The settings we selected represent in most cases standard options of popular softwares packages for the implementation of these methods.

5. The Simulation Exercise

We compare the LPDH rule with both its generalizations based on regression models (2.1) and the regression trees, by means of a simulation that can be described by means of the following steps:

- we generate synthetic populations characterized by different assumptions on the form of the relationship between y and $\boldsymbol{x}$;
- we generate also the population values of $y*$, $\mathbf{y}^*_U = (y^*_1, \ldots, y^*_N)$ according to a predetermined value of $y*$ $Corr_U(\mathbf{y}_U, \mathbf{y}^*_U)$ $(Corr_U(\mathbf{y}_U, \mathbf{y}^*_U) = \dfrac{N^{-1}\sum_{i\in U} y_i y^*_i - \bar{\mathbf{y}}_U \bar{\mathbf{y}}^*_U}{\sqrt{V_U(\mathbf{y}_U) V_U(\mathbf{y}^*_U)}}$ and $V_U(\bullet)$ denotes the descriptive variance in population U);
- R simple random samples of fixed size n are drawn and d_s is assumed to be observed $(s = 1, \ldots, R)$. This emulates a situation where a sample of values of a proxy variable y^* is available to the survey analyst at the design stage.

The performance measure of each stratification method is the design effect,

$$Deff(\hat{\boldsymbol{t}}_{y,s\_M=m}) = \frac{V_D(\hat{\boldsymbol{t}}_{y,s\_M=m})}{V_D(\hat{\boldsymbol{t}}_{y,srs})} \tag{5.1}$$

where $\hat{\boldsymbol{t}}_{y,s\_M=m}$ is the stratification estimator of the $\boldsymbol{t}_y$ based on stratification method m and V_D denotes the variance with respect to the randomization distribution. Neyman allocation of the samples to strata (assuming known variances within strata) is considered.

5.1. Setup

We consider the five synthetic populations discussed in Banks et al. (2003). All populations are characterized by the general structure $y_i = g(\mathbf{x}_i) + u_i$, $i \in U$ where $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is a deterministic function and u_i is a zero-mean disturbance term. The population size is set to be moderately large: $N = 20,000$. In fact, in many situations efficient methods are to be applied to the stratification of subpopulations.

More in detail the deterministic components are given by:

- LIN: $g(\mathbf{x}_i) = p^{-1}\sum_{j=1}^{p} x_{ij}$
- INDGAU: $g(\mathbf{x}_i) = (2\pi)^{-p/2}(|.25\boldsymbol{I}|)^{-1/2}\exp\{-\frac{1}{2}\mathbf{x}_i^T(.25\mathbf{I})^{-1}\mathbf{x}_i\}$

  where $\mathbf{I}$ is the $p \times p$ identity matrix;
- CORGAU: $g(\mathbf{x}_i) = (2\pi)^{-p/2}(|.25\boldsymbol{A}|)^{-1/2}\exp\{-\frac{1}{2}\mathbf{x}_i^T(.25\mathbf{A})^{-1}\mathbf{x}_i\}$

  where $\mathbf{A}$ is a $p \times p$ matrix and such that $\mathbf{A}_{ii} = 1$ and $\mathbf{A}_{ij} = \rho_1 = 0.8$;
- MIXT: $g(\mathbf{x}_i) = (2\pi)^{-p/2}(|.16\mathbf{I}|)^{-1/2}\exp\{-\frac{1}{2}\mathbf{x}_i^T(.16\mathbf{I})^{-1}\mathbf{x}_i\} + (2\pi)^{-p/2}(|.16\mathbf{I}|)^{-1/2}$ $\exp\{-\frac{1}{2}(\mathbf{x}_i - \mathbf{1})^T(.16\mathbf{I})^{-1}(\mathbf{x}_i - \mathbf{1})\}$ where $\mathbf{1}$ is a p-dimensional vector of ones;
- PROD: $g(\mathbf{x}_i) = \left\{\prod_{j=1}^{p} x_{ij}\right\}^{\frac{1}{p}}$

The number of auxiliary variables is set to $p = 6$ and they are all assumed to be uniformly distributed over the unit interval $\mathbf{x}_{Uj} \sim Uni(0, 1)$, with $\mathbf{x}_{Uj}$ being the $N \times 1$ column vector with the values of the j-th auxiliary variable. These auxiliary variables are set to be equally and mildly correlated: $Corr_U(\mathbf{x}_{Uj}, \mathbf{x}_{Uj'}) \cong 0.3 \ \forall j \neq j'$.

They are generated using an algorithm proposed by Fackler (1999). As regards the error term we set:

$$V_U(\mathbf{u}) = \frac{1 - \rho_2^2}{\rho_2^2} V_U\{g(\mathbf{X}_U)\}$$

where $\mathbf{u} = (u_1, \ldots, u_N)$ and the vector $g(\mathbf{X}_U) = \{g(\mathbf{x}_i), i \in U\}$

As a consequence, $Corr_U\{\mathbf{y}_U, g(\mathbf{X}_U)\} = \rho_2$. We set this parameter to 0.9.

A plot of $(\mathbf{X}_U, g(\mathbf{X}_U))$ for a population generated in that manner is given in Figure 1 for $p = 2$. This plot provides an idea of the relationship between $\mathbf{y}_U$ and $\mathbf{X}_U$ in the five populations.

The relationship between y and $\boldsymbol{x}$ in the five populations ranges from exact linearity to nonlinearity. Marginally, the population densities of y range from the exact symmetry of population LIN to moderately skewed situations.

To summarize, it is assumed that in each sample, a "proxy" variable y^* such that $y^* = y + w$ is observed, where w is a zero-mean disturbance. The values of w are generated in order to satisfy the following conditions as almost exactly:

- $\overline{\mathbf{w}}_U = 0$
- $V_U(\mathbf{w}) = \phi^{-2}(1 - \phi^2)V_U(\mathbf{y})$ with $\phi = 0.9$ so that $Corr_U(\mathbf{y}^*_U, \mathbf{y}_U) = .9$

We note that in the case of pilot surveys $y \equiv y^*$ and $Corr_U(\mathbf{y}^*_U, \mathbf{y}_U) = 1$.

$R = 2,000$ independent simple random samples were taken from each synthetic population.

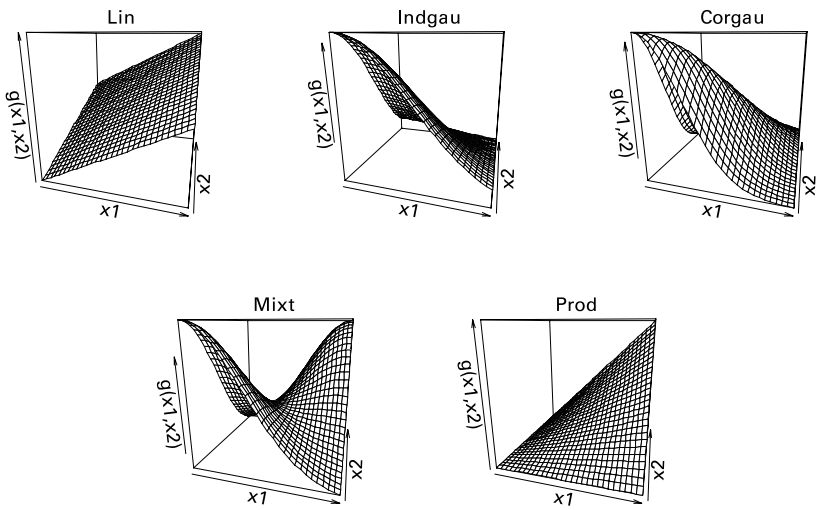

Fig. 1. Diagrams of $(\mathbf{X}_U, g(\mathbf{X}_U))$ *for the simulated population in the case of* $p = 2$

The simulation is intended to represent a situation where there is good information for stratification even if no single auxiliary variable has a strong correlation with the target variable.

5.2. Results

We first compare the LPDH rule and regression trees according to the simulation exercise described in Section 5.1. In particular we consider two sample sizes: $n = 100$ and $n = 1,000$.

As regards the number of strata in which the population is to be partitioned, we consider for both the LPDH rule and the regression tree the number of strata identified by the latter as a result of the pruning rule.

Tree regressions are computed using the software package R and the library RPART (Therneau and Atkinson 1997). For pruning, we consider a cost-complexity rule (Breiman et al. 1984, pp. 66–86) in which bias and variance are estimated by means of a 10-fold cross-validation algorithm. In particular we adopt the usual "1 *standard error*" rule (Breiman et al. 1984, pp. 78–81).

The simulations results are shown in Table 1.

We note that with $n = 100$, regression trees perform very poorly and are not reasonable competitors to the LPDH rule. This is a consequence of the fact that in general most nonparametric methods need large sample sizes to work properly and that regression trees, in particular, are often characterized by large sampling variance, that is, a small change in the data can result in a very different set of splits.

With $n = 1,000$, trees perform well when the relationship between y and $\boldsymbol{x}$ is highly nonlinear (CORGAU and MIXT populations); but when this relationship is linear or nearly so, the LPDH rule is dramatically better. This second fact highlights one of the drawbacks of regression trees, that is, their difficulty in identifying additive structures.

Table 1. Stratification based on Regression Trees vs LPDH rule: simulation results averaged over the $R = 2,000$ MC replications (standard deviations within parenthesis). $\overline{H}$ is the average number of strata over the replication space

			Average *Deff*	
Population	Sample size	$\overline{H}$	Regression trees	LPDH
LIN	100	4.812 (2.666)	0.645 (0.106)	0.324 (0.141)
INDGAU	100	2.092 (1.547)	0.867 (0.152)	0.764 (0.262)
CORGAU	100	1.172 (0.715)	0.982 (0.053)	0.974 (0.072)
MIXT	100	1.256 (1.030)	0.986 (0.032)	0.987 (0.042)
PROD	100	1.798 (1.313)	0.892 (0.147)	0.831 (0.231)
LIN	1,000	18.572 (4.975)	0.395 (0.023)	0.185 (0.003)
INDGAU	1,000	16.526 (4.192)	0.404 (0.034)	0.227 (0.009)
CORGAU	1,000	10.110 (3.420)	0.486 (0.054)	0.515 (0.018)
MIXT	1,000	20.716 (7.694)	0.605 (0.081)	0.745 (0.167)
PROD	1,000	15.238 (4.585)	0.427 (0.043)	0.232 (0.029)

Moreover, with few exceptions, the trees tend to identify "too many" strata and the number of strata they identify is quite unstable. We say "too many" meaning that they are far more than the number of strata widely recognized in the literature as optimal (see also Section 6).

We also ran the simulation with sample sizes larger than 1,000 (up to $n = 5,000$). The results in terms of (5.1) are not very different from the case of $n = 1,000$; moreover the number of strata identified by regression trees grows to unacceptably large values. To save space, the results are not reported here.

Our general aim is to find a stratification method that, at least for large samples, is as effective as the LPDH rule in the linear case and better when the relationship between y and $\boldsymbol{x}$ is nonlinear. In this sense the regression trees fail, even though they can still be used and be effective in particular circumstances.

In the generalization of the LPDH rule we do not have any automatic method for determining the optimal number of strata. In comparing the performances of the three nonparametric regression methods we present results for different values of H ($H = 3, 6, 12$).

The reason for this choice can be indicated by considering that when Model (2.1) holds and g is a known linear function, little reduction in variance is to be expected beyond $H = 6$ (Cochran 1977 p. 132). We also consider the case $H = 12$ in order to verify whether this holds when in Model (2.1) g is nonlinear (see also Section 6).

For the estimation we used the packages mgcv, mda and gbm which have been implemented in R. They are freely available on http://cran.r-project.org

The results of the simulation exercise are shown in Tables 2 and 3.

Table 2. Stratification based on nonparametric regression methods vs LPDH rule: simulation results for $n = 100$ averaged over the $R = 2,000$ replications (standard deviations within parenthesis; results for the best stratification method are bold)

		Average *Deff*			
H	Population	LPDH	AMDH	MARSDH	BOOSTDH
3	LIN	**0.334** (0.012)	0.375 (0.037)	0.469 (0.051)	0.401 (0.028)
3	INDGAU	**0.467** (0.014)	0.513 (0.054)	0.501 (0.062)	0.508 (0.033)
3	CORGAU	**0.719** (0.025)	0.819 (0.077)	0.721 (0.077)	0.731 (0.046)
3	MIXT	0.882 (0.085)	0.785 (0.067)	**0.638** (0.071)	0.867 (0.054)
3	PROD	**0.515** (0.013)	0.532 (0.049)	0.516 (0.063)	0.552 (0.039)
6	LIN	**0.247** (0.014)	0.296 (0.045)	0.401 (0.052)	0.326 (0.031)
6	INDGAU	**0.333** (0.028)	0.433 (0.061)	0.438 (0.063)	0.398 (0.046)
6	CORGAU	0.639 (0.052)	0.784 (0.095)	0.682 (0.083)	**0.637** (0.065)
6	MIXT	0.802 (0.157)	0.747 (0.072)	**0.587** (0.083)	0.754 (0.080)
6	PROD	**0.363** (0.030)	0.449 (0.072)	0.454 (0.068)	0.436 (0.049)
12	LIN	**0.221** (0.013)	0.271 (0.044)	0.382 (0.055)	0.298 (0.027)
12	INDGAU	**0.290** (0.030)	0.407 (0.068)	0.414 (0.062)	0.354 (0.048)
12	CORGAU	0.606 (0.059)	0.771 (0.099)	0.657 (0.083)	**0.586** (0.067)
12	MIXT	0.785 (0.166)	0.734 (0.083)	**0.572** (0.085)	0.693 (0.078)
12	PROD	**0.306** (0.041)	0.431 (0.072)	0.444 (0.066)	0.381 (0.045)

Table 3. Stratification based on nonparametric regression methods vs LPDH rule: simulation results for $n = 1,000$ *averaged over the* $R = 2,000$ *replications (standard deviations within parenthesis; results for the best stratification method are bold)*

		Average *Deff*			
H	Population	LPDH	AMDH	MARSDH	BOOSTDH
3	LIN	**0.304** (0.003)	0.306 (0.004)	0.315 (0.021)	0.327 (0.005)
3	INDGAU	0.422 (0.003)	0.402 (0.006)	0.366 (0.013)	**0.339** (0.009)
3	CORGAU	0.655 (0.006)	0.654 (0.016)	0.557 (0.024)	**0.461** (0.022)
3	MIXT	0.831 (0.084)	0.594 (0.009)	0.448 (0.020)	**0.411** (0.015)
3	PROD	0.469 (0.003)	0.421 (0.006)	0.385 (0.013)	**0.365** (0.011)
6	LIN	**0.216** (0.002)	0.219 (0.003)	0.229 (0.007)	0.244 (0.006)
6	INDGAU	0.281 (0.003)	0.322 (0.011)	0.298 (0.014)	**0.271** (0.011)
6	CORGAU	0.543 (0.008)	0.576 (0.032)	0.492 (0.030)	**0.389** (0.026)
6	MIXT	0.767 (0.152)	0.529 (0.014)	0.382 (0.023)	**0.347** (0.019)
6	PROD	0.305 (0.004)	0.337 (0.011)	0.319 (0.014)	**0.294** (0.012)
12	LIN	**0.191** (0.002)	0.194 (0.003)	0.204 (0.007)	0.221 (0.006)
12	INDGAU	0.247 (0.004)	0.297 (0.011)	0.278 (0.014)	**0.234** (0.011)
12	CORGAU	0.496 (0.012)	0.551 (0.034)	0.475 (0.031)	**0.370** (0.028)
12	MIXT	0.739 (0.167)	0.507 (0.014)	0.366 (0.025)	**0.326** (0.020)
12	PROD	**0.241** (0.004)	0.309 (0.014)	0.298 (0.015)	0.269 (0.013)

With a "training" sample d_s with $n = 100$, a situation that is likely to occur when stratifying primary sampling units in multistage surveys, the linear model is the more efficient kernel for the Dalenius-Hodges method in most of the cases (LIN, INDGAU, PROD populations), the rest being represented by situations in which the relationship between y and $\boldsymbol{x}$ is highly nonlinear. For the MIXT population MARSDH is clearly better than the other methods. For the CORGAU population LPDH is best for $H = 3$ but for larger values of H, BOOSTDH shows similar or better performances. AMDH never emerges as the best method.

With $n = 1,000$ or larger (we conducted the simulation also for $n = 5,000$), the methods based on nonparametric regression are almost as efficient as the one based on parametric linear prediction when the relationship between y and $\boldsymbol{x}$ is linear or approximately so, while they are much more efficient in the rest of the cases. The LPDH rule remains the best performer for the LIN population. In almost all other cases, BOOSTDH turns out to be the best method. Nonetheless for the PROD population as the number of strata grows large, its advantage over LPDH dwindles (for $H = 12$ we have that LPDH is better). The performance of MARSDH is close to that of BOOSTDH for all values of H and for all populations except CORGAU. As regards AMDH, as for the case $n = 100$, it never emerges as the best method.

As a general comment, we note that to work properly nonparametric regression methods require large sample sizes. When this is the case, BOOSTDH seems to emerge as the best stratification rule, since it is comparable to LPDH in linear or nearly linear cases and far better in the nonlinear ones.

6. Sensitivity and Consistency Checks

In this section we deal with two distinct issues that complement our discussion of the generalizations of the LPDH rule based on nonparametric regression methods. The first issue is the sensitivity of the methods described in Section 4 to the choice of the number of strata, while the second regards the consistency of the boundaries identified by the various methods, in cases where they show close performances.

As far as the sensitivity of the results to the selected number of strata is concerned, we may note from Tables 2 and 3 that we have large gains in precision in passing from simple random sampling to stratification with three strata; smaller, though sometimes still relevant gains in passing from three to six strata; but then only minor gains in efficiency in passing to twelve strata. This is so for all populations and all methods. To assess this result in a clearer way, we plot the performances of the various stratification rules against the number of strata, for $R = 2,000$ and $n = 100$ and 1,000.

From Figures 2 and 3 it is apparent that, for all stratification rules, we have huge gains in efficiency up to 5 or 6 strata regardless of the population and the kind of relationship between y and $\boldsymbol{x}$. After this threshold, the design effects stabilize quickly and increasing the number of strata becomes immaterial for efficiency. On the other hand, the multiplication of strata does not seem to worsen the efficiency of the stratified mean estimator.

This result depends on the underlying assumption regarding the strength of the relationship between y and $\boldsymbol{x}$. We also ran a simulation exercise identical to the one described, except for the fact that we set parameter $\rho_2 = Corr_U\{\mathbf{y}_U, g(\mathbf{X}_U)\} = 0.7$. Results from this second experiment, which are not reported here, are consistent with those

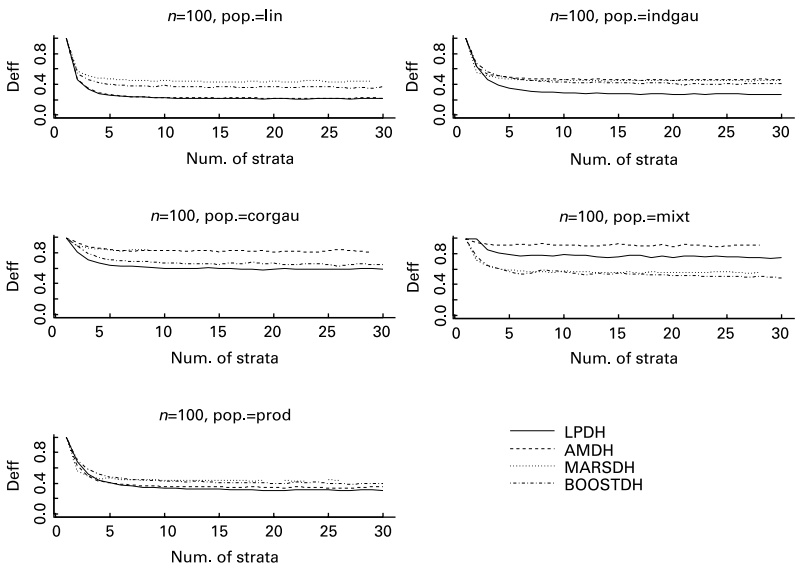

Fig. 2. Performances of different stratification rules against the number of strata ($n = 100$), averaged over $R = 2,000$ MC replicates

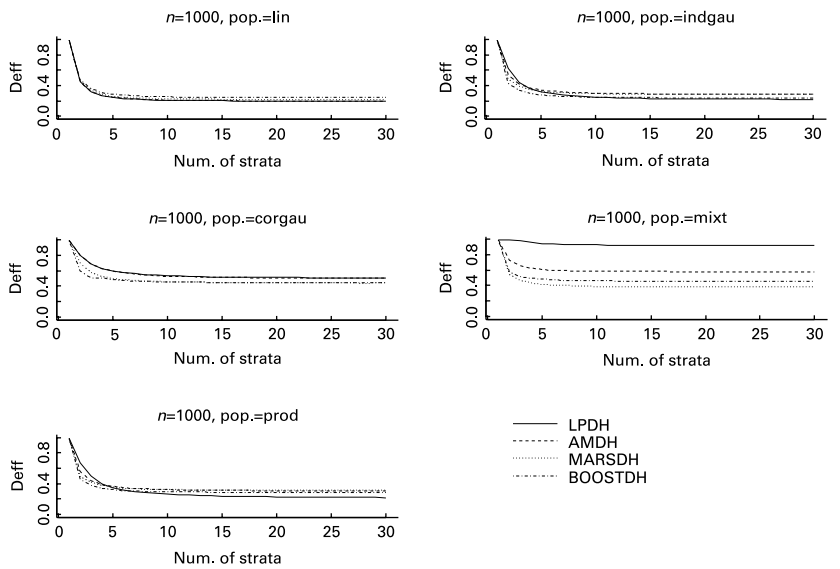

Fig. 3. Performances of different stratification rules against the number of strata ($n = 1,000$), averaged over $R = 2,000$ MC replicates

we described as far as the ranking of regression methods is concerned. By the way, the weaker correlation yields minor gains in efficiency after three strata.

Consistently with theory for the linear case, we have that gains in terms of (5.1) stabilize rapidly as the number of strata grows. In the context of our simulation, for all populations, the best choice for the number of strata seems to depend on the strength of the relationship between y and $\boldsymbol{x}$: the weaker it is, the lower is the recommendable number of strata.

We may expect that, when the performances of different methods are close in terms of *Deff*, they are also consistent in the sense of producing close boundaries between the strata. To check this we consider the distribution of boundaries in the $R = 2,000$ Monte Carlo replicates. The results, not reported here, show that, for the LIN, PROD, and INDGAU populations, all the methods are consistent with each other in terms of stratum boundaries. In contrast, for the CORGAU and MIXT populations, the poor performances of the LPDH rule displayed in Tables 2 and 3 translate into the shrinkage of strata boundaries towards the sample mean of y. As a consequence the strata formed in this way are not consistent with those based on the other methods.

7. Concluding Remarks

In this article, we investigate some extensions of the popular Dalenius and Hodges rule for optimal one-way stratification in an information setting in which a proxy of the target variable is known for a sample from the population and a possibly large set of potential stratification variables is available. The discussed methods are capable of handling both linear and a nonlinear relationship between y and $\boldsymbol{x}$. Model-assisted methods encompassing nonlinear relationship between y and $\boldsymbol{x}$ in the analysis of survey samples have been studied by many authors in recent years (Breidt and Opsomer 2000; Wu and

Sitter 2001). Examples of nonlinear relationship may arise, for instance, in environmental and agricultural surveys (see e.g., Opsomer et al. 2006). The basic idea behind practical implementation of optimal stratification rules is that of having an auxiliary variable known for each unit in the population which approximates as closely as possible the (single) target variable y. Here, we propose to build this "ideal" auxiliary variable using predicted values of y obtained by means of nonparametric regression methods.

We propose generalizations of the Dalenius and Hodges rule, which is designed to approximately minimize the variance of the estimators of $\boldsymbol{t}_y$. All the methods (including regression trees that are not based on generalizing the Dalenius and Hodges rule) are compared in terms of $V(\hat{\boldsymbol{t}}_{y,s})$, assuming optimal allocation of the sample to the strata. For this reason we cannot say that the proposed methods are optimal in a strict sense.

A primary finding is that linear prediction is a good basis for the definition of a stratification rule whenever the available sample is small (such as $n = 100$) or the relationship holding between y and $\boldsymbol{x}$ is not far from being linear (LIN, INDGAU, and PROD populations).

When a large sample is available for "training" the prediction algorithm, nonparametric regression methods perform in much the same way as the LPDH rule when the relationship between y and $\boldsymbol{x}$ is linear or approximately linear and better in the other cases.

However, not all these methods perform equally well: boosted trees seem to offer the best basis for a modified Dalenius and Hodges stratification rule.

We also considered regression trees that may be intuitively appealing for stratification in the information setting assumed in the article. However, they perform poorly and regression trees based stratification does not seem adequate in most cases.

It should be emphasized that our methods can be applied to the generalization of rules other than the Dalenius and Hodges, such as the Ekman rule. When stratifying highly skewed populations an analogous generalization of the Lavallée and Hidiroglou (1988) rule can, in principle, be proposed.

Developments of this research in many directions are conceivable: the simulation experiment we considered is fairly general but other and more general situations may be considered, as for instance the inclusion of categorical stratification variables or the stratification of extremely skewed populations. The problem of stratifying populations characterized by a relationship between y and $\boldsymbol{x}$ affected by heteroskedastic errors may also be of interest. Moreover, we did not deal with the problem of multiple-way optimal stratification, which is a topic of relevance in many applied situations.

8. References

Banks, D.L., Olszewski, R.T., and Maxon, R.A. (2003). Comparing Methods for Multivariate Nonparametric Regression. Communication in Statistics, Simulation and Computation, 32, 541–571.

Bethel, J. (1989). Minimum Variance Estimation in Stratified Sampling. Journal of the American Statistical Association, 84, 260–265.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and Regression Trees. Chapman and Hall/CRC.

Breiman, L. (1998). Arching Classifiers (with Discussion). The Annals of Statistics, 26, 801–849.

Breidt, F.J. and Opsomer, J.D. (2000). Local Polynomial Regression Estimators in Survey Sampling. The Annals of Statistics, 28, 1026–1053.

Cochran, W.G. (1961). Comparison of the Methods for Determining Stratum Boundaries. Bulletin of the International Statistical Institute, 38, 345–358.

Cochran, W.G. (1977). Sampling Techniques, 3rd ed, New York: Wiley.

Dalenius, T. (1957). Sampling in Sweden. Contributions to the Methods and Theories of Sample Survey Practice. Stockholm: Almqvist and Wicksell.

Dalenius, T. and Hodges, J.L. (1959). Minimum Variance Stratification. Journal of the American Statistical Association, 54, 88–101.

Ekman, G. (1959). An Approximation Useful in Univariate Stratification. The Annals of Mathematical Statistics, 30, 219–229.

Fackler, P. (1999). Generating Correlated Multidimensional Variates. Working Paper retrieved from www4.ncsu.edu/~pfackler/randcorr.ps (latest check: July 21st 2006).

Friedman, J. (1991). Multivariate Adaptive Regression Splines (with Discussion). The Annals of Statistics, 19, 1–141.

Friedman, J. (2001). Greedy Functions Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29, 1189–1232.

Hedlin, D. (2000). A Procedure for Stratification by an Extended Ekman Rule. Journal of Official Statistics, 16, 15–29.

Hastie, T., Tibshirani, R., and Friedman, J. (2003). The Elements of Statistical Learning. New York: Springer.

Hess, I., Sethi, V.K., and Balakrishnan, T.R. (1966). Stratification: A Practical Investigation. Journal of the American Statistical Association, 61, 74–90.

Hidiroglou, M.A. and Laniel, N. (2001). Sampling and Estimation Issues for Annual and Sub-annual Canadian Business Surveys. International Statistical Review, 69, 487–504.

Lavallée, P. and Hidiroglou, M.A. (1988). On the Stratification of Skewed Populations. Survey Methodology, 14, 33–43.

Murthy, M.N. (1967). Sampling Theory and Methods. Calcutta: Statistical Publishing Company.

Opsomer, J.D., Breidt, J.D., Moisen, G.G., and Kauermann, G. (2006). Model-assisted Estimation of Forest Resources with Generalized Additive Models. Journal of the American Statistical Association (to appear).

Rivest, L.P. (2002). A Generalization of the Lavallée and Hidiroglou Algorithm for Stratification in Business Surveys. Survey Methodology, 28, 191–198.

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

Sigman, R.S. and Monsour, N. (1995). Selecting Samples from List Frames of Surveys. In Business Survey Methods, B. Cox, D. Binder, B.N. Chinnappa, A. Christianson, M. Colledge, and P. Kott (eds), New York: Wiley, 133–152.

Sweet, E.M. and Sigman R.S. (1995). Evaluation of Model-Assisted Procedures for Stratifying Skewed Populations Using Auxiliary Data. Proceedings of the American Statistical Association, Section on Survey Research Methods, 491–396.

Singh, R. (1971). Approximately Optimum Stratification on the Auxiliary Variable. Journal of the American Statistical Association, 66, 829–833.

Therneau, T.M. and Atkinson E.J. (1997). An Introduction to Recursive Partitioning Using the RPART Routines. Technical Report, Mayo Foundation.

Wright, R.L. (1983). Finite Population Sampling with Multivariate Auxiliary Information. Journal of the American Statistical Association, 78, 879–884.

Wu, C. and Sitter, R.R. (2001). A Model-calibration to Using Complete Auxiliary Information from Survey Data. Journal of the American Statistical Association, 96, 185–193.

Received September 2004
Revised November 2006