

Penalized Balanced Sampling

BY F. J. BREIDT

Department of Statistics, Colorado State University, Fort Collins, Colorado 80523–1877, USA
jbreidt@stat.colostate.edu

G. CHAUVET

Laboratoire de Statistique d'Enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France
chauvet@ensai.fr

SUMMARY

Linear mixed models are flexible and extensible models that cover a wide range of statistical methods. They have found many uses in estimation for complex surveys, particularly in small area estimation and in extensions of generalized regression estimation. They have also been used as a means of relaxing constraints in calibration estimation. The purpose of this work is to consider methods by which linear mixed models may be used at the design stage of a survey. This paper reviews the ideas of balanced sampling and the cube algorithm, and proposes an implementation of the cube algorithm by which penalized balanced samples can be selected. Such samples have the property that Horvitz-Thompson estimators from penalized balanced samples behave like linear mixed model-assisted survey regression estimators from unbalanced samples. The methodology is evaluated by using nonparametric and temporal linear mixed models in simulation experiments, and by using a spatial linear mixed model in an artificial but realistic sampling application, motivated by a 1991–1996 Environmental Protection Agency survey of lakes in the Northeastern states of the United States.

Some key words: Model-assisted estimation; Nonparametric regression; Ridge calibration; Small area estimation.

1. INTRODUCTION

Linear mixed models are flexible and extensible models that cover methods ranging from multiple regression and analysis of variance, to penalized splines, to low-rank kriging, among many others (Robinson, 1991; Ruppert et al., 2003). Naturally, they have found many uses in estimation for complex surveys, particularly in small area estimation (Fay & Herriot, 1979; Battese et al., 1988; Datta & Ghosh, 1991; Ghosh & Rao, 1994; Rao, 2003; Opsomer et al., 2007) and in extensions of generalized regression estimation (Zheng & Little, 2003, 2004; Breidt et al., 2005). Linear mixed models have also been used, explicitly or implicitly, as a means of stabilizing weights and relaxing constraints in calibration estimation and related methods. These uses include robust case weighting (Bardsley & Chambers, 1984; Chambers, 1996), ridge calibration (Rao & Singh, 1997; Beaumont & Bocci, 2008; Montanari & Ranalli, 2009), and other methods for satisfying range restrictions (Park, 2002; Park & Fuller, 2005) or smoothing weights (Lazzeroni & Little, 1998; Elliott & Little, 2000).

The purpose of this work is to consider methods by which linear mixed models may be used at the design stage of a survey, rather than at the estimation stage. That is, if one believes a priori

49 that a linear mixed model of the form (1), below, is a reasonable working model for key response
 50 variable(s) y , how should one draw a sample? In §2.1, we introduce our linear mixed model
 51 notation and discuss penalization, smoothing, and the degrees of freedom of a smooth. Then, to
 52 motivate our approach we begin with the simpler case of $Z = 0$ in (1). This leads to a review
 53 of generalized regression estimation and weighting in §2.2, as well as balanced sampling and
 54 the cube algorithm in §2.3. In §2.4, we discuss extension of the generalized regression estimator
 55 to linear mixed model-assisted survey regression estimation, and use this to motivate our main
 56 results on penalized balanced sampling in §2.5. As noted above, linear mixed models in survey
 57 estimation can be used to relax calibration constraints. Our use of linear mixed models in survey
 58 design allows us to relax balance constraints.

59 Penalized balanced samples are selected via a particular implementation of the cube algorithm.
 60 They have the property that Horvitz-Thompson estimators from penalized balanced samples be-
 61 have like linear mixed model-assisted survey regression estimators. We make this statement pre-
 62 cise in a pair of propositions. We demonstrate the methodology with simulation results using a
 63 nonparametric linear mixed model (penalized spline) in §3.1 and using a temporal linear mixed
 64 model (regression with random walk errors) in §3.2. We then present in §3.3 a penalized bal-
 65 anced sample driven by a spatial linear mixed model (low-rank kriging), and motivated by a real
 66 US Environmental Protection Agency survey of lakes in the Northeastern United States.

67 2. METHODOLOGY

68 2.1. Linear mixed models, smoothing and penalization

69 Consider a finite population with index set $U = \{1, 2, \dots, N\}$. Suppose that a particular re-
 70 sponse variable $y_U = [y_j]_{j \in U}$ follows a linear mixed model of the form

$$71 \quad y_U = X\beta + Zb + \epsilon_U, \quad (1)$$

72 where

$$73 \quad \mathbb{E} \begin{bmatrix} b \\ \epsilon_U \end{bmatrix} = 0, \quad \text{var} \left(\begin{bmatrix} b \\ \epsilon_U \end{bmatrix} \right) = \sigma^2 \begin{bmatrix} \lambda^{-2}Q & 0 \\ 0 & I \end{bmatrix},$$

74 where X is a full rank $N \times q$ matrix, Z is a full rank $N \times K$ matrix, and where I will denote an
 75 identity matrix of appropriate dimension. We suppose that Q is positive definite and known, and
 76 in our examples it is typically an identity matrix. The parameter σ^2 is unknown and the parameter
 77 λ^2 is to be determined.

78 The parameter λ^2 can be interpreted as a penalty for model complexity. To see this, write
 79 $C = [X, Z]$ and $\Lambda = \text{blockdiag}(0, \lambda^2 Q^{-1})$. Then

$$80 \quad (C^T C + \Lambda)^{-1} C^T y_U$$

81 gives the best linear unbiased estimators of β and best linear unbiased predictors of b , and

$$82 \quad C(C^T C + \Lambda)^{-1} C^T y_U$$

83 gives the best linear unbiased predictors of y_U . Because $C(C^T C + \Lambda)^{-1} C^T$ predicts or smooths
 84 y_U , it is common (Hastie & Tibshirani, 1990, p. 52) to interpret

$$85 \quad \text{df} = \text{tr} \left\{ C(C^T C + \Lambda)^{-1} C^T \right\} = \text{tr} \left\{ (C^T C + \Lambda)^{-1} C^T C \right\}$$

86 as the degrees of freedom of this smooth. If $\lambda^2 = 0$, there is no penalty for model complexity,
 87 and the model has $q + K$ degrees of freedom, corresponding to a regression on X and Z fixed
 88
 89
 90
 91
 92
 93
 94
 95
 96

97 effects. As $\lambda^2 \rightarrow \infty$, $df \rightarrow q$, corresponding to a regression on X only. Values of λ^2 can be
 98 chosen to trade off the goodness-of-fit of the linear mixed model and its complexity.

100 2.2. Generalized regression estimation and weighting

101 Let $s \subset U$ be a sample of size n drawn via a sampling design with known, positive inclusion
 102 probabilities $\{\pi_j\}_{j \in U}$. For any column vector v_j , we write the Horvitz-Thompson estimator
 103 (Horvitz & Thompson, 1952) as

$$104 \text{HT}(v) = \sum_{j \in s} \frac{v_j}{\pi_j}.$$

107 A common circumstance in survey practice is the availability of a $q \times 1$ covariate vector x_j , for
 108 which both $\{x_j\}_{j \in s}$ and t_x are observed. This information can be incorporated into the estimation
 109 procedure through the generalized regression estimator, which we write as

$$110 \text{GREG}(v) = \sum_{j \in s} \left\{ \frac{1}{\pi_j} + (t_x - \text{HT}(x))^T M_s^{-1} \frac{x_j}{N\pi_j} \right\} v_j = \sum_{j \in s} \omega_{js} v_j \quad (2)$$

113 where

$$114 M_s = \sum_{j \in s} \frac{x_j x_j^T}{N\pi_j},$$

115 see for example Särndal et al. (1992, chap. 7). The generalized regression estimation weights

$$116 \omega_{js} = \left\{ \frac{1}{\pi_j} + (t_x - \text{HT}(x))^T M_s^{-1} \frac{x_j}{N\pi_j} \right\}$$

117 in (2) can be viewed as the original design weights π_j^{-1} modified to take into account the in-
 118 formation in x_j . A well-known property of the generalized regression estimation weights is that
 119 they are calibrated to the x -totals, in the sense that

$$120 \text{GREG}(x^T) = \sum_{j \in s} \omega_{js} x_j^T = t_x^T.$$

121 Under a standard asymptotic framework for finite population sampling with $N \rightarrow \infty$,

$$122 \text{HT}(x^T) - t_x^T = O_p \left(\frac{N}{\sqrt{n}} \right).$$

123 Here and elsewhere, the remainder notation $O_p(\cdot)$, $o_p(\cdot)$, $O(\cdot)$, or $o(\cdot)$ is interpreted element-wise
 124 for a matrix of the implied dimension; $1 \times q$ in this case. Thus,

$$125 \text{HT}(x^T) - \text{GREG}(x^T) = O_p \left(\frac{N}{\sqrt{n}} \right).$$

126 In practice, these generalized regression estimation weight adjustments to the Horvitz-
 127 Thompson estimators may be highly variable, particularly if the number of x -variables is large.
 128 Extremely large or even negative weights are possible. This instability has led a number of
 129 authors to consider alternative estimation strategies, using a variety of methods (Bardsley &
 130 Chambers, 1984; Chambers, 1996; Rao & Singh, 1997; Beaumont & Bocci, 2008; Montanari &
 131 Ranalli, 2009; Park, 2002; Park & Fuller, 2005; Lazzeroni & Little, 1998; Elliott & Little, 2000).
 132 As noted in the introduction, many of these methods are explicitly built on linear mixed models,
 133 or build on ideas of ridge regression, and hence have an implicit link to penalized regression as
 134
 135
 136
 137
 138
 139
 140
 141
 142
 143
 144

145 in the linear mixed models. Broadly speaking, these estimation methods use linear mixed models
 146 to relax calibration constraints.

148 2.3. *Balanced sampling*

149 If $\{x_j\}$ is available at the design stage for all $j \in U$, then an alternative to generalized re-
 150 gression estimation is to incorporate the x -information into the design through the selection of
 151 balanced samples. The sampling design $p(\cdot)$ is said to be balanced on variables x if the equations
 152

$$153 \text{HT}(x^T) = t_x^T \quad (3)$$

154 hold exactly. The variables x_j are called the balancing variables. If the equations (3) are satis-
 155 fied, the variance of the Horvitz-Thompson estimator is zero for any linear combination of the
 156 balancing variables x .

157 Deville & Tillé (2004) introduced the cube method, which enables the selection of exact bal-
 158 anced samples if such samples may be found, or approximate balanced samples otherwise. The
 159 cube method proceeds in two phases: the flight phase, in which balancing constraints are main-
 160 tained exactly, and the landing phase, in which the balancing constraints are relaxed to complete
 161 the sample. The flight phase consists of a discrete random walk in the N -dimensional hyper-
 162 cube, beginning at $\pi^0(s) = (\pi_1, \pi_2, \dots, \pi_N)^T$ and proceeding in steps $\pi^1(s), \pi^2(s), \dots, \pi^F(s)$,
 163 where F denotes the last step of the flight phase. At each step, one or more coordinates of $\pi^t(s)$
 164 are randomly rounded to 0 or 1, and remain there forever. At the end of the flight phase, at most
 165 q coordinates of $\pi^F(s)$ are not equal to 0 or 1.

166 The landing phase then proceeds to determine the remaining non-integer coordinates, resulting
 167 in the final sample, $\pi^T(s) = (I_1, I_2, \dots, I_N)^T$. In this paper, we focus on the variant of the
 168 landing phase that successively relaxes the balance constraints until the sample is finalized, by
 169 dropping the last balance constraint, re-running the flight phase with one fewer constraint, and
 170 repeating as necessary. In this formulation, it is necessary to order the balancing variables with
 171 respect to their relative importance, since the last balancing variables are removed first.

172 Deville & Tillé (2004) show in their Proposition 4 that, for any variant of the landing phase,
 173 the cube method achieves

$$174 | \text{HT}(x) - t_x | \leq q \max_{j \in U} \frac{|x_j|}{\pi_j}$$

175 (again, interpreted element-wise), so that under reasonable hypotheses on x_j and for standard
 176 designs with $(N \min_{j \in U} \pi_j)^{-1} = O(n^{-1})$, we have

$$177 \text{HT}(x) = t_x + O\left(\frac{Nq}{n}\right). \quad (4)$$

178 Note that the remainder term is non-stochastic, so for fixed q this can be much better than the
 179 usual rate in the unbalanced case, $O_p(N/\sqrt{n})$.

180 One desirable feature of balance on x_j is that it reduces or eliminates the need for generalized
 181 regression weight adjustments, since (4) implies that

$$182 \text{HT}(x^T) - \text{GREG}(x^T) = \text{HT}(x^T) - t_x^T = O\left(\frac{Nq}{n}\right). \quad (5)$$

2.4. Linear mixed model survey regression estimators

Linear mixed model-assisted survey regression estimators analogous to generalized regression estimators can be written

$$\text{LMM}(v) = \sum_{j \in s} \left\{ \frac{1}{\pi_j} + (t_c - \text{HT}(c))^T M_s^{-1} \frac{c_j}{\pi_j} \right\} v_j$$

where

$$M_s = \sum_{j \in s} \frac{c_j c_j^T}{\pi_j} + \begin{bmatrix} 0 & 0 \\ 0 & \lambda^2 Q^{-1} \end{bmatrix} = \sum_{j \in s} \frac{c_j c_j^T}{\pi_j} + \Lambda.$$

The penalized spline survey regression estimator of Breidt et al. (2005) is a special case with $x_j^T = [1, x_j, \dots, x_j^{q-1}]$, $z_j^T = [(x_j - \kappa_1)_+^{q-1}, \dots, (x_j - \kappa_K)_+^{q-1}]$ for fixed, known knots $\{\kappa_k\}_{k=1}^K$, and $Q = I$.

One approach to the problem of using model (1) in survey design would be to use the cube algorithm to draw samples balanced on c_j , in the sense that

$$\text{HT}(c) = t_c + O\left(\frac{N(q+K)}{n}\right). \tag{6}$$

The flexibility and power of linear mixed models, however, typically come with large K , so that (6) may have unacceptably large errors. Such balance also ignores the mixed effect structure of the linear mixed model, and treats it (essentially) as an ordinary regression model with $q + K$ fixed effects.

2.5. Penalized balanced sampling

We propose an alternative approach to the problem of using model (1) in survey design by modifying the cube method of Deville & Tillé (2004) to draw penalized balanced samples, with the property that Horvitz-Thompson estimators for the totals of balancing covariates in the model are approximately equal to linear mixed model-assisted estimators, while preserving the overall df of the linear mixed model. To use the cube method, we specify a set of balance conditions and an ordering of the balance conditions. The penalty in the linear mixed model sets the overall df of the model, and the (possibly fractional) df of each balance condition determines its priority in the landing phase. These df also appear explicitly in the determination of the error rates, as we show below.

We are specifically interested in mixed models in which b_1, \dots, b_K are random coefficients of K basis functions, which occur in penalized splines and their semiparametric extensions, and in smoothing with radial basis functions; see Ruppert et al. (2003) for these and many other examples. It is convenient to first orthogonalize the fixed and random effects in model (1) via

$$C = [X, (I - P_X)Z]$$

where $P_X = X(X^T X)^{-1} X^T$. Then define

$$M = C^T C + \Lambda = \begin{bmatrix} X^T X & 0 \\ 0 & Z^T (I - P_X) Z + \lambda^2 Q^{-1} \end{bmatrix}$$

and compute

$$M^{-1} C^T C = \begin{bmatrix} I & 0 \\ 0 & A_1 D A_2^T \end{bmatrix},$$

where $D = \text{diag}\{d_1, \dots, d_K\}$, $A_1 D A_2^T$ is the singular value decomposition of

$$\left\{ Z^T (I - P_X) Z + \lambda^2 Q^{-1} \right\}^{-1} Z^T (I - P_X) Z,$$

and $1 \geq d_1 \geq d_2 \geq \dots \geq d_K \geq 0$. Note that the q singular values corresponding to the fixed effects are identically equal to one, and that $q + \sum_{i=1}^K d_i = \text{tr}(C M^{-1} C^T)$ is the df of the linear mixed model.

Finally, we define the balancing conditions as the columns of the $N \times (q + K)$ matrix

$$B = [b_j^T] = C \begin{bmatrix} I & 0 \\ 0 & A_1 D \end{bmatrix}. \quad (7)$$

Note that the first q columns of B are exactly X in this formulation.

An alternative to the balancing conditions (7) that is useful in practice is to keep only the first r columns of B , where $\sum_{i=r+1}^K d_i \leq \alpha \ll 1$; that is, dropping columns that all together account for much less than one degree of freedom. In any case, penalized balanced sampling can be implemented using existing software for the cube method, such as the R function `samplecube` in the `sampling` library (Tillé & Matei, 2008; R Development Core Team, 2008), or the SAS Macro `fastcube` (Chauvet & Tillé, 2005). The user simply specifies the balance conditions and their ordering, selecting the landing phase variant that drops columns sequentially until the sample is finalized.

Figure 1 shows the balance conditions (columns of B) and corresponding singular values for the linear mixed model corresponding to a 7 df penalized spline, with $q = 2$ and $K = 35$. The first two degrees of freedom correspond to fixed effects for the intercept and a linear term; subsequent terms (approximately quadratic, cubic, etc.) have fractional degrees of freedom, which sum to 7 across all 37 singular values. These additional columns of B are all orthogonal to X by construction. See Section 3.1 for details.

Whether or not the sample is balanced, we have that

$$\begin{aligned} & \text{LMM}(c^T) - \text{HT}(c^T) \\ &= \sum_{j \in s} \left\{ \frac{c_j^T}{\pi_j} + (t_c - \text{HT}(c))^T M_s^{-1} \frac{c_j c_j^T}{\pi_j} \right\} - \text{HT}(c^T) \\ &= (t_c - \text{HT}(c))^T \left\{ M^{-1} C^T C + o_p(1) \right\} \\ &= (t_c - \text{HT}(c))^T \left\{ \begin{bmatrix} I & 0 \\ 0 & A_1 D \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & A_2^T \end{bmatrix} + o_p(1) \right\} \\ &= (t_b - \text{HT}(b))^T \begin{bmatrix} I & 0 \\ 0 & A_2^T \end{bmatrix} + o_p\left(\frac{N}{\sqrt{n}}\right), \end{aligned} \quad (8)$$

where we have assumed that $t_c - \text{HT}(c) = O_p(N/\sqrt{n})$; that is, its error rate is no worse than the standard rate. The dominant term in the error is then $t_b - \text{HT}(b)$, which we now consider in detail for balanced and unbalanced sampling.

PROPOSITION 1. *Under penalized balanced sampling with B as defined in (7),*

$$|t_{b_k} - \text{HT}(b_k)| \leq q \max_{j \in U} \frac{|x_{jk}|}{\pi_j}, \quad (9)$$

for the fixed effects $k = 1, \dots, q$, and

$$|t_{b_{q+k}} - HT(b_{q+k})| \leq (q+k) d_k \max_{j \in U} \frac{\left(\sum_{k=1}^K \tilde{z}_{jk}^2\right)^{1/2}}{\pi_j} \quad (10)$$

for the random effects $k = 1, \dots, K$, where $\tilde{z}_j^T = (\tilde{z}_{j1}, \dots, \tilde{z}_{jK})$ denotes the j th row of $(I - P_X)Z$.

Proof: For the q fixed effects, which are the last to be dropped in the landing phase, the result follows directly from Proposition 4 of Deville & Tillé (2004); see equation (4). For the K random effects, the result follows from Proposition 4 of Deville & Tillé (2004) and inequality (A1) in the Appendix.

Proposition 1 shows explicitly the dependence of the estimation error on the number of parameters and the corresponding df. We now obtain a similar expression for the expected estimation error under a general sampling design, not necessarily balanced.

PROPOSITION 2. *Under a probability sampling design with second-order inclusion probabilities π_{ij} , define $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$. Then, with B as defined in (7),*

$$\begin{aligned} E|t_{b_k} - HT(b_k)| \\ \leq \left(N \max_{j \in U} \pi_j\right)^{1/2} \left(1 + \frac{N \max_{i,j \in U: i \neq j} |\Delta_{ij}|}{\max_{j \in U} \pi_j}\right)^{1/2} \max_{j \in U} \frac{|x_{jk}|}{\pi_j}. \end{aligned} \quad (11)$$

for $k = 1, \dots, q$, and

$$\begin{aligned} E|t_{b_{q+k}} - HT(b_{q+k})| \\ \leq \left(N \max_{j \in U} \pi_j\right)^{1/2} \left(1 + \frac{N \max_{i,j \in U: i \neq j} |\Delta_{ij}|}{\max_{j \in U} \pi_j}\right)^{1/2} d_k \max_{j \in U} \frac{\left(\sum_{k=1}^K \tilde{z}_{jk}^2\right)^{1/2}}{\pi_j} \end{aligned} \quad (12)$$

for $k = 1, \dots, K$.

The proof of this proposition is given in the appendix.

To facilitate comparisons, it is useful to simplify the design features in Propositions 1 and 2 by introducing some assumptions on the asymptotic behavior of the design. We assume a sequence of designs $p_N(\cdot)$ drawing samples s_N of size n_N from the finite populations U_N , with $n_N \rightarrow \infty$ as $N \rightarrow \infty$, and

$$\left(N \min_{j \in U_N} \pi_{jN}\right)^{-1} = O(n^{-1}); \quad N \max_{j \in U_N} \pi_{jN} = O(n); \quad N \max_{i,j \in U_N: i \neq j} |\Delta_{ijN}| = O(nN^{-1});$$

see, for example, A6 in Breidt & Opsomer (2000) for analogous hypotheses.

It is immediate from these assumptions and equations (9) and (11) that the relevant orders for direct comparison of the errors of the fixed effects are $qO(Nn^{-1})$ for penalized balanced sampling and $O_p(Nn^{-1/2})$ for unbalanced sampling.

Similarly, for the random effects, we have from (10) that with balance on B ,

$$|t_{b_{q+k}} - HT(b_{q+k})| \leq (q+k) d_k \max_{j \in U_N} \frac{\left(\sum_{k=1}^K \tilde{z}_{jk}^2\right)^{1/2}}{\pi_{jN}}$$

$$\begin{aligned}
&\leq (q+k) d_k \max_{j \in U_N} \left(\sum_{k=1}^K \tilde{z}_{jk}^2 \right)^{1/2} \frac{N}{N \min_{j \in U_N} \pi_{jN}} \\
&= (q+k) d_k \max_{j \in U_N} \left(\sum_{k=1}^K \tilde{z}_{jk}^2 \right)^{1/2} O\left(\frac{N}{n}\right).
\end{aligned}$$

Without balance on B , we have from (12) that

$$\begin{aligned}
&E|t_{b_{q+k}} - \text{HT}(b_{q+k})| \\
&\leq O(\sqrt{n_N}) \left(1 + \frac{NO(n_N N^{-1})}{N \min_{j \in U_N} \pi_{jN}} \right)^{1/2} d_k \max_{j \in U_N} \left(\sum_{k=1}^K \tilde{z}_{jk}^2 \right)^{1/2} \frac{N}{N \min_{j \in U_N} \pi_{jN}} \\
&= d_k \max_{j \in U_N} \left(\sum_{k=1}^K \tilde{z}_{jk}^2 \right)^{1/2} O(\sqrt{n_N}) \left(1 + O(n_N) O(n_N^{-1}) \right)^{1/2} O(Nn_N^{-1}) \\
&= d_k \max_{j \in U_N} \left(\sum_{k=1}^K \tilde{z}_{jk}^2 \right)^{1/2} O\left(\frac{N}{\sqrt{n_N}}\right).
\end{aligned}$$

The relevant orders for direct comparison of the errors of the random effects are then $(q+k)d_k O(Nn^{-1})$ for penalized balanced sampling, and $d_k O_p(Nn^{-1/2})$ for unbalanced sampling. The first is deterministic, and of smaller order than the second, provided $q+K$ can be regarded as fixed relative to the sample size. Unlike (6), the order for penalized balanced sampling includes the factor d_k , which decays rapidly to zero for many linear mixed models of interest. The d_k 's are directly tied to the variance component structure of the linear mixed model and are interpretable as fractional degrees of freedom.

Complete treatment of asymptotics with $K \rightarrow \infty$ is beyond the scope of this paper, but the above arguments suggest that penalized balanced sampling will dominate unbalanced sampling for the linear mixed model, provided $K = o(\sqrt{n_N})$. A reasonable way to ensure this for a given design is to fix the overall df and truncate B by dropping columns that together account for less than one degree of freedom, as described above.

3. EMPIRICAL RESULTS

3.1. Simulation with penalized spline balance

In this section we discuss a Monte Carlo study comparing a number of design/estimation strategies, including balanced sampling guided by a penalized spline expressed as a linear mixed model. Following Breidt et al. (2005), we consider eight mean functions:

- constant: $m_1(x) = 8$,
- linear: $m_2(x) = 1 + 2(x - 0.5)$,
- quadratic: $m_3(x) = 1 + 2(x - 0.5)^2$,
- bump: $m_4(x) = 1 + 2(x - 0.5)^2 + \exp(-200(x - 0.5)^2)$,
- jump: $m_5(x) = \{1 + 2(x - 0.5)I_{\{x \leq 0.65\}}\} + 0.65I_{\{x > 0.65\}}$,
- exponential: $m_6(x) = \exp(-8x)$,
- cycle1: $m_7(x) = 2 + \sin(2\pi x)$,
- cycle4: $m_8(x) = 2 + \sin(8\pi x)$.

385 The population is of size $N = 1000$. The population x_j are generated as independent and
 386 identically distributed $\text{Uniform}[0, 1]$. For all variables other than `constant`, the mean function
 387 is shifted and scaled so that the minimum equals 0 and the maximum equals 2. The population
 388 values $y_{ik}; i = 1, \dots, 8$ are created from the mean functions by generating a set of independent
 389 $N(0, \sigma^2)$ errors and adding this same random sequence to each of the shifted and scaled mean
 390 functions. We use two possible values for the model standard deviation of the errors, namely
 391 $\sigma = 0.1$ and $\sigma = 0.4$.

392 The linear mixed model corresponds to the penalized-spline model-assisted survey regression
 393 estimator of Breidt et al. (2005), with $Q = I_K$,

$$394 \quad X = \begin{bmatrix} 1 & x_j & \dots & x_j^{q-1} \end{bmatrix},$$

395 and

$$396 \quad Z = \begin{bmatrix} (x_j - \kappa_1)_+^{q-1} & \dots & (x_j - \kappa_K)_+^{q-1} \end{bmatrix}$$

397 where $(a)_+ = \max\{0, a\}$. The knots $\{\kappa_j\}$ are fixed and given by the $1/(K+1), \dots, K/(K+1)$
 398 quantiles of the variable x_j . In this application, we take $q = 2$ and $K = 35$, so that $\lambda^2 = 0$
 399 corresponds to a continuous, piecewise linear fit between the knots (37 degrees of freedom), and
 400 $\lambda^2 \rightarrow \infty$ corresponds to a global line (2 degrees of freedom). Following Breidt et al. (2005),
 401 we consider 7 degrees of freedom and compare to a variety of other sampling and estimation
 402 strategies.
 403
 404
 405

406 The penalized balanced sampling with Horvitz-Thompson estimation strategy dominates sim-
 407 ple random sampling followed by regression estimation or penalized spline estimation, as well
 408 as the four stratified strategies, since it is often much better (root mean square error ratio > 1.2
 409 in 34 out of 96 cases) and rarely much worse (root mean square error ratio < 0.95 in 10 out
 410 of 96 cases). Many of the biggest wins occur for the cases that are poorly approximated by
 411 a low-order polynomial: `bump`, `jump`, and `cycle4`. The penalized balanced sampling with
 412 Horvitz-Thompson estimation loses in low noise to simple random sampling followed by regres-
 413 sion estimation for `linear`, `quadratic`, `exponential`, and `cycle1`, which are the four
 414 non-constant cases best approximated by the polynomial used in regression estimation. Also, it
 415 loses to simple random sampling followed by penalized spline estimation for the same cases, and
 416 same reason. The remaining two losses are to stratification with 20 strata. Fine stratification is a
 417 natural competitor for the proposed method, but penalized balanced sampling does well against
 418 it in this simple example. See Section 3.3 for a case in which constructing fine strata would be
 419 complicated, but penalized balanced sampling is straightforward.

420 The strategy of penalized balanced sampling followed by linear mixed model estimation is
 421 better than penalized balanced sampling followed by Horvitz-Thompson estimation in all of the
 422 low-noise cases (root mean square error ratios ranging from 0.78–0.99), and essentially equiva-
 423 lent in the high-noise cases (root mean square error ratios from 0.97–0.99). The penalized bal-
 424 anced sampling design does not achieve exact balance, and so even on the fixed effects there are
 425 typically calibration adjustments, meaning that the linear mixed model-assisted estimator differs
 426 from the Horvitz-Thompson estimator. The simulation results here are consistent with those of
 427 Deville & Tillé (2004), who also found calibration after balanced sampling to be the best strategy.

428 One general concern about balanced sampling is the size of the support of the design. That
 429 is, one possible drawback of balanced sampling could be a lack of entropy associated with only
 430 a small number of samples likely to be selected, as is the case with systematic sampling. In
 431 the case of penalized balanced sampling, it might be expected that the support of the design is
 432 particularly small, since many balancing constraints are used. However, in this study each of the

Table 1. *Root mean square error of seven strategies divided by root mean square error of Horvitz-Thompson estimator for penalized balanced sample with a 7 degree of freedom spline, based on 1,000 samples of size $n = 100$ from a fixed finite population of size $N = 1,000$.*

	σ	SI-GREG	SI-LMM	ST4	ST5	ST10	ST20	PBS-LMM
constant	0.1	1.02	1.01	1.00	0.95	1.03	1.00	0.99
	0.4	1.02	1.01	1.00	0.95	1.03	1.00	0.99
linear	0.1	0.84	0.83	1.48	1.24	0.95	0.86	0.82
	0.4	1.01	1.00	1.07	0.98	1.02	0.99	0.99
quadratic	0.1	0.83	0.83	2.75	2.33	1.35	1.00	0.81
	0.4	1.00	0.99	1.26	1.14	1.06	1.00	0.97
bump	0.1	1.47	1.25	2.28	1.60	1.41	0.98	0.81
	0.4	1.07	1.03	1.20	1.02	1.07	1.00	0.99
jump	0.1	1.54	1.37	2.29	1.90	1.62	1.12	0.93
	0.4	1.09	1.06	1.19	1.07	1.11	1.02	0.99
exponential	0.1	0.80	0.86	2.15	1.80	1.16	0.91	0.78
	0.4	0.99	0.99	1.15	1.07	1.04	0.99	0.97
cycle1	0.1	0.87	0.89	2.85	2.34	1.40	0.97	0.85
	0.4	1.01	1.00	1.28	1.13	1.07	0.98	0.98
cycle4	0.1	4.74	3.80	4.72	4.61	3.11	1.76	0.92
	0.4	2.01	1.69	1.97	1.94	1.49	1.13	0.98

SI-GREG, simple random sampling without replacement with 6th degree polynomial regression estimation; SI-LMM, simple random sampling without replacement with a 7 degree of freedom penalized spline linear mixed model estimation; ST4, ST5, ST10, ST20, stratified simple random sampling without replacement with 4, 5, 10, or 20 equally-sized strata and Horvitz-Thompson estimation; PBS-LMM, penalized balanced sampling selection followed by 7 degree of freedom penalized spline linear mixed model estimation

1000 simulations performed led to a different estimate, suggesting that the support of the design is reasonably large.

We also considered the issue of variance estimation for penalized balanced sampling. Since second-order inclusion probabilities are difficult to compute exactly, both the Yates-Grundy variance estimator and the Horvitz-Thompson variance estimator may not be used directly. Alternatively, one may consider the variance approximation proposed by Deville & Tillé (2005) or the martingale-difference variance approximation proposed by Breidt & Chauvet (2009). The martingale-difference variance approximation is based on a representation of the cube method that enables the construction of an efficient simulation-based approximation of the design variance-covariance matrix. This approximation is obtained through an independent run of 10,000 simulations. Both the Deville and Tillé variance approximation and the martingale-difference variance approximation are compared to the variance of the Horvitz-Thompson estimator in Table 2. Clearly, the approximation of Deville and Tillé tends to underestimate the true variance. The negative bias is particularly high with a small σ coefficient, that is, if the balancing variables have good explanatory power. In fact, only the flight phase is taken into account in the variance approximation of Deville & Tillé (2005), whereas the landing phase may prove to contribute non-negligible variance when the variable of interest is highly related to the balancing variables. The martingale-difference variance approximation clearly dominates the Deville and Tillé variance approximation, since it is always better in all the low-noise cases and in all the high-noise cases but two, indicating that the method succeeds in accounting for the whole sampling process in the variance approximation.

3.2. Simulation 2: Random walk balance

In this section, we discuss a second Monte Carlo study, with penalized balance on a temporal linear mixed model. The response variables in this example follow Lazzeroni & Little (1998),

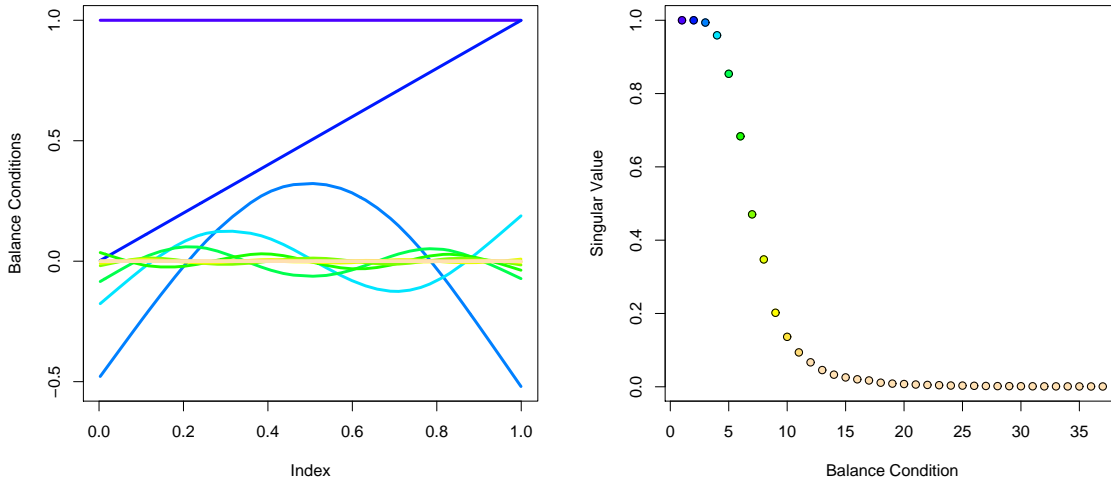


Fig. 1. Left: Balance conditions (b_{jk} versus index = j/N for columns $k = 1, \dots, 37$ of B). Right: Singular values (d_k versus k) for penalized balanced sampling with 7-df penalized spline. In each plot, k is indexed by color, with lighter color corresponding to lower df.

Table 2. Variance of the Horvitz-Thompson estimator for penalized balanced sample based on 1,000 samples of size $n = 100$.

σ		constant	linear	quadratic	bump	jump	exponential	cycle1	cycle4
0.1	SIM	82	133	124	150	139	127	132	197
	DT	88	88	88	88	91	88	88	91
	MD	88	134	130	144	145	132	117	208
0.4	SIM	1510	1461	1566	1497	1475	1544	1296	1473
	DT	1412	1412	1412	1412	1413	1412	1412	1414
	MD	1413	1461	1456	1473	1470	1457	1441	1534

SIM, direct simulation-based variance approximation; DT, Deville and Tillé variance approximation; MD, martingale-difference variance approximation

where models for smoothing post-stratification weights were compared with other procedures on an artificial data set. Our population was constructed with $H = 20$ domains, representing ordered age group categories. The domains are denoted as U_1, \dots, U_H , with randomly-generated domain counts N_h , $h = 1, \dots, H$ ranging from 29 to 112. These counts were simulated with mean 62.9 and standard deviation 28.3, one-tenth of the corresponding values for Non-Hispanic Females in East Los Angeles used by Lazzeroni & Little (1998). Our simulated population size, $N = 1,317$, is thus approximately one-tenth the size of the original study (12,575).

Study variables were generated as described in Lazzeroni & Little (1998), §5.2, with normal and log-normal versions of 9 response variables. In the 9 normal cases, the variables are generated as $y_{hi} | \mu_h$ independent and identically distributed $N(0, \sigma^2)$, with:

- NULL: $\mu_h = 0, \sigma^2 = 1$
- XRE: μ_h iid $N(0, 1/3)$ (exchangeable), $\sigma^2 = 1$ (HIGH) or $\sigma^2 = 1/3$ (LOW)
- AR1: $\mu_h \sim N(0, 1/3)$, $\text{Corr}(\mu_h, \mu_{h+k}) = 0.9^{|k|}$, $\sigma^2 = 1$ or $\sigma^2 = 1/3$
- REG: μ_h independent $N(0.2h, 1/3)$, $\sigma^2 = 1$ or $\sigma^2 = 1/3$

Table 3. *Root mean square error of six strategies divided by root mean square error of Horvitz-Thompson estimator for penalized balanced sample with 4 degree of freedom random walk, based on 1,000 samples of size $n = 100$ from a fixed finite population of size $N = 1,317$.*

	SI-GREG	SI-RW	ST4	ST5	ST10	ST20
Normal Cases						
NULL	0.98	0.97	0.99	1.00	0.96	0.98
XRE-LOW	1.40	1.35	1.33	1.29	1.25	0.93
XRE-HIGH	1.14	1.12	1.13	1.11	1.08	0.96
AR1-LOW	1.42	1.17	1.12	1.08	1.01	0.97
AR1-HIGH	1.14	1.04	1.03	1.03	0.98	0.97
REG-LOW	2.59	1.48	1.54	1.29	1.23	0.93
REG-HIGH	1.71	1.18	1.22	1.12	1.07	0.96
QUAD-LOW	2.25	1.88	1.73	1.49	1.26	0.92
QUAD-HIGH	1.56	1.37	1.30	1.21	1.08	0.95
Log-Normal Cases						
NULL	0.94	0.93	1.01	0.98	0.97	1.00
XRE-LOW	1.21	1.18	1.21	1.15	1.15	0.98
XRE-HIGH	1.04	1.02	1.08	1.04	1.04	0.99
AR1-LOW	1.23	1.06	1.09	1.02	1.00	1.00
AR1-HIGH	1.04	0.98	1.04	0.99	0.98	1.00
REG-LOW	2.09	1.27	1.36	1.14	1.14	0.98
REG-HIGH	1.43	1.06	1.15	1.03	1.03	0.99
QUAD-LOW	1.87	1.58	1.50	1.30	1.16	0.97
QUAD-HIGH	1.34	1.20	1.20	1.10	1.05	0.99

SI-GREG, simple random sampling without replacement with 3rd degree polynomial regression estimation; SI-RW, simple random sampling without replacement with a 4 degree of freedom random walk estimation; ST4, ST5, ST10, ST20, stratified simple random sampling without replacement with 4, 5, 10, or 20 equally-sized strata and Horvitz-Thompson estimation

- QUAD: μ_h independent $N(4.41 - 0.84h + 0.04h^2, 1/3)$, $\sigma^2 = 1$ or $\sigma^2 = 1/3$

In the 9 log-normal cases,

$$y_{hi} = \mu_h + \sigma\sqrt{64/151} (\exp(z_{hi}) - 13/8),$$

with $\{z_{hi}\}$ iid $N(0,1)$, and μ_h and σ as defined above.

In this application, the linear mixed model is linear ($q = 2$) with random walk errors among domains. That is,

$$X = \left[[1 \ (h/H)]_{N_h \times 2} \right]_{h=1}^H,$$

and

$$Z = [I_{\{j \in U_1\}} \ I_{\{j \in U_2\}} \ \dots \ I_{\{j \in U_H\}}] \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}_{H \times H}.$$

We consider 4 degrees of freedom and compare to six other design/estimation strategies in Table 3. Penalized balanced sampling followed by either Horvitz-Thompson estimation or by random walk estimation gives virtually identical results, so results for the latter strategy are omitted. For the five strategies excluding the fine stratification, penalized balanced sampling followed by Horvitz-Thompson estimation dominates since it is often much better (root mean square error ratio > 1.2 in 34 out of 90 cases) and rarely much worse (root mean square error ratio < 0.95 in only 2 out of 90 cases). Penalized balanced sampling followed by Horvitz-Thompson

577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624

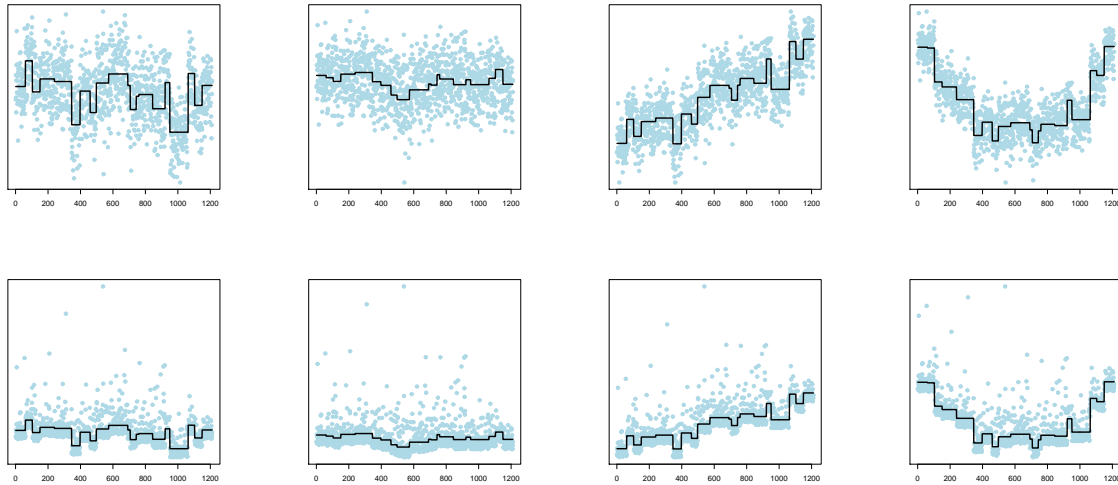


Fig. 2. Response variables with exchangeable, first-order autoregressive, linear and quadratic domain effects, from left to right. The upper plots correspond to the low-noise ($\sigma^2 = 1/3$) normal versions of the response variables, and the lower plots to the low-noise log-normal versions. The piecewise constant curves correspond to μ_h as defined in the text.

estimation is essentially identical to fine stratification for the log-normal variables, while fine stratification is slightly better for the normal variables. Fine stratification is a natural competitor to penalized balanced sampling in this example, which is ordered in one dimension and hence easy to stratify. In the example of the next section, fine stratification would be much harder to implement.

3.3. Application: spatial sampling with low-rank kriging

This example was used to illustrate a nonparametric small area estimation methodology in Opsomer et al. (2007). Between 1991 and 1996, the Environmental Monitoring and Assessment Program (EMAP) of the US Environmental Protection Agency conducted a survey of lakes in the Northeastern states of the United States. The survey was based on a population of over 21,000 lakes from which 334 lakes were surveyed. A major focus of the Opsomer et al. (2007) application was estimation of the mean acid neutralizing capacity (ANC) for each of 113 small areas defined by 8-digit Hydrologic Unit Codes (HUC) within the region of interest. The linear mixed model fitted in that application included fixed effects for the intercept and elevation, random effects for the HUCs, and random effects for a low-rank version of a thin-plate spline, with $K = 80$ knots selected by the space-filling algorithm implemented in the `cover.design()` function in the `FUNFITS` package for S-plus (Nychka et al., 1998).

For purposes of illustration, we treat this as an artificial but realistic design problem and consider the sampling of $n = 334$ lakes from the $N = 21,363$ Northeastern lakes with non-missing locations and elevations. These lakes are plotted in Figure 3, with color scale indicating lake elevation. We seek a sample balanced on elevation and with some degree of spatial balance.

We use penalized balanced sampling driven by a linear mixed model with fixed effects for intercept and elevation. For random effects, we start with the same $K = 80$ knots as used in

Opsomer et al. (2007), but with random effects corresponding to a low-rank kriging specification using an exponential covariance function: $\gamma_\rho(r) = \exp(-|r|/\rho)$.

The linear mixed model is then

$$y_U = [1, \text{elevation}]\beta + [\gamma_\rho(\|s_j - k_i\|)]_{N \times K} \left([\gamma_\rho(\|k_h - k_i\|)]_{h,i=1}^K \right)^{-1/2} b + \epsilon_U,$$

with $\{k_i\}_{i=1}^K$ the knot locations, $\{x_j\}_{j=1}^N$ the lake locations, and

$$\begin{bmatrix} b \\ \epsilon_U \end{bmatrix} \sim \left(0, \sigma^2 \begin{bmatrix} (1/\lambda^2)I_K & 0 \\ 0 & I_N \end{bmatrix} \right).$$

We chose $\lambda^2 = (\sigma_{\text{error}}^2 + \sigma_{\text{HUC}}^2)/\sigma_{\text{spline}}^2 = ((179.5)^2 + (365.7)^2)/(71.2)^2$ to match the earlier paper, and to reflect the case in which a sample designer has prior information about sources of variation. With λ^2 fixed, the choice of $\rho = 0.34$ gives approximately 8 degrees of freedom.

We used the R function `samplecube` in the `sampling` library (Tillé & Matei, 2008) to select the penalized balanced sample shown in Figure 3, which took approximately 4 minutes on a Lenovo Thinkpad dual core laptop computer, including constructing the design matrices and computing the singular value decomposition. The sample appears to have good spatial balance, considering the irregular spatial domain and irregular spatial density of the population of lakes. Achieving such balance appears to be non-trivial using ordinary methods like spatially systematic or stratified sampling.

4. CONCLUSION

The use of linear mixed models and related methods in survey estimation is pervasive, so it seems natural to consider their use in survey design. Penalized balanced sampling produces samples for which Horvitz-Thompson estimators behave like linear mixed model-assisted survey regression estimators. The cube method and existing software make implementation of penalized balanced sampling straightforward. The linear mixed model formulation gives the survey designer considerable flexibility in specifying a design in terms of fixed effects with highest priority for achieving balance, and random effects with decreasing priority for achieving balance. Priority is parameterized by degrees of freedom of the balance conditions, ranging from 1 degree of freedom for each fixed effect to near 0 degrees of freedom for the least-important random effects. This prioritization of balance constraints is analogous to relaxation of calibration constraints in estimation, a common application of linear mixed models. Simulation results and empirical examples illustrate the methodology in nonparametric, temporal, and spatial settings and show that penalized balanced sampling leads to efficient design-estimation strategies relative to a number of alternatives.

ACKNOWLEDGEMENT

The research of the first author was partly supported by the US National Science Foundation (SES-0922142).

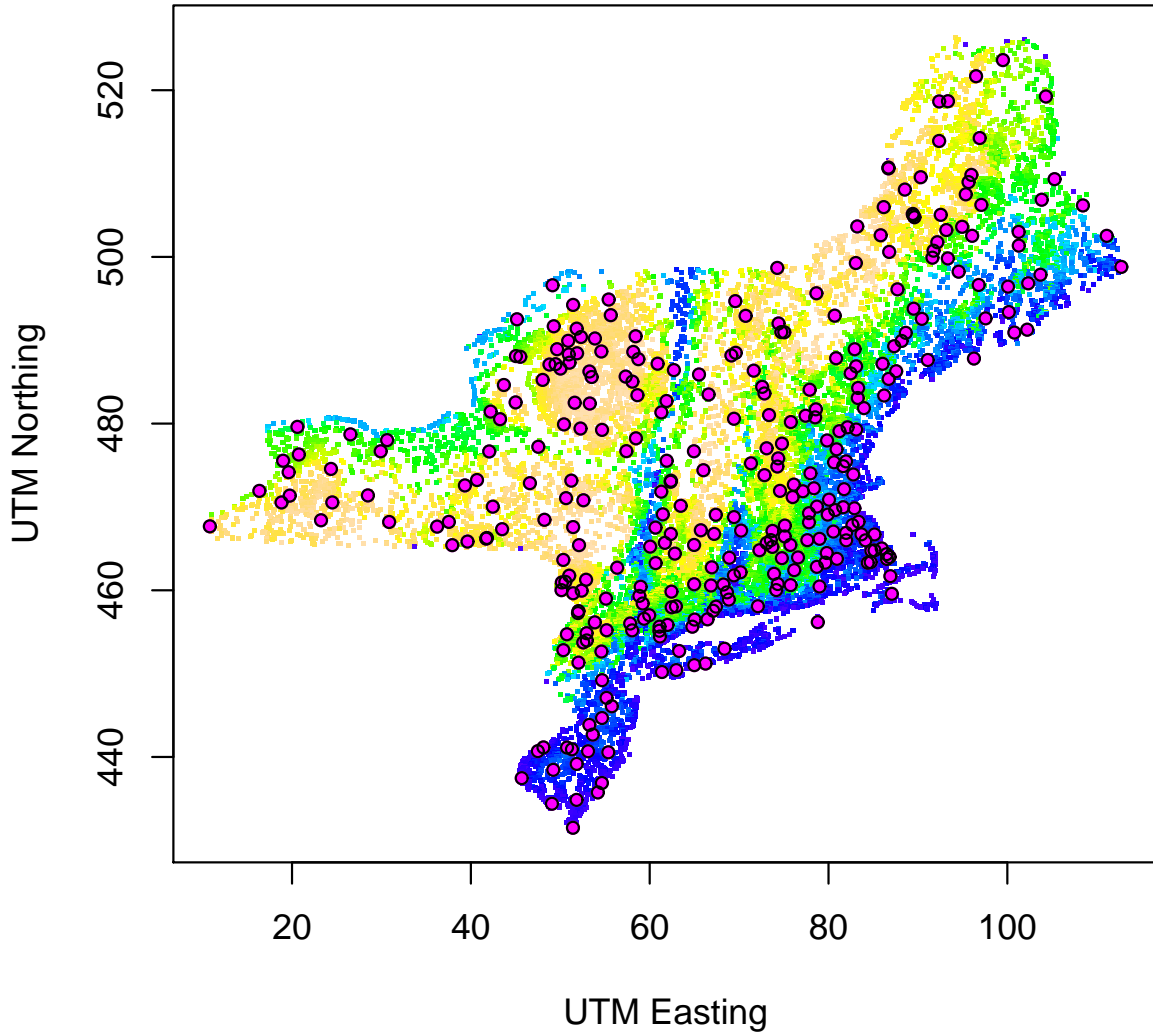


Fig. 3. Penalized balanced sample of $n = 334$ from $N = 21,363$ Northeastern lakes. Sample is balanced on elevation and on low-rank exponential autocovariance ($\rho = 0.34$, 8 df) with $K = 80$ knots, chosen according to a space-filling criterion. Color scale indicates lake elevation.

5. APPENDIX

5.1. Bounds on elements of b .

We establish a bound for $|b_{j,q+k}|$ where $j \in U = \{1, \dots, N\}$ and $k = 1, \dots, K$. These are the terms in the last K columns in B , where

$$B = C \begin{bmatrix} I & 0 \\ 0 & A_1 D \end{bmatrix} = [X, (I - P_X)Z A_1 D].$$

Let \tilde{z}_j^T denote the j th row of $(I - P_X)Z$ and let a_k denote the k th column of A_1 , recalling that this matrix is orthonormal by the singular value decomposition. Then the last K columns of B are

$$\begin{aligned} & [b_{j,q+1} \ b_{j,q+2} \ \cdots \ b_{j,q+K}]_{j \in U} \\ &= [\tilde{z}_j^T a_1 \ \tilde{z}_j^T a_2 \ \cdots \ \tilde{z}_j^T a_K]_{j \in U} \text{diag}\{d_1, d_2, \dots, d_K\} \\ &= [\tilde{z}_j^T a_1 d_1 \ \tilde{z}_j^T a_2 d_2 \ \cdots \ \tilde{z}_j^T a_K d_K]_{j \in U}. \end{aligned}$$

Hence

$$\begin{aligned} |b_{j,q+k}| &= |\tilde{z}_j^T a_k d_k| \\ &\leq d_k (\tilde{z}_j^T \tilde{z}_j)^{1/2} (a_k^T a_k)^{1/2} = d_k \left(\sum_{k=1}^K \tilde{z}_{jk}^2 \right)^{1/2}. \end{aligned} \quad (\text{A1})$$

5.2. Proof of Proposition 2

Note that

$$\begin{aligned} (\mathbb{E}|t_{b_{q+k}} - \text{HT}(b_{q+k})|)^2 &\leq \mathbb{E}|t_{b_{q+k}} - \text{HT}(b_{q+k})|^2 \\ &= \sum_{i \in U} \pi_i (1 - \pi_i) \left(\frac{b_{q+k,i}}{\pi_i} \right)^2 + \sum_{i,j \in U: i \neq j} \Delta_{ij} \frac{b_{q+k,i}}{\pi_i} \frac{b_{q+k,j}}{\pi_j} \\ &\leq \left(\sum_{i \in U} \pi_i (1 - \pi_i) + \sum_{i,j \in U: i \neq j} |\Delta_{ij}| \right) \left(\max_{j \in U} \frac{|b_{q+k,j}|}{\pi_j} \right)^2. \end{aligned} \quad (\text{A2})$$

For an asymptotic analysis, it is convenient to bound (A2) further, as

$$\begin{aligned} &\leq \left(N \max_{j \in U} \pi_j + N^2 \max_{i,j \in U: i \neq j} |\Delta_{ij}| \right) \left(\max_{j \in U} \frac{|b_{q+k,j}|}{\pi_j} \right)^2 \\ &\leq \left(N \max_{j \in U} \pi_j \right) \left(1 + \frac{N \max_{i,j \in U: i \neq j} |\Delta_{ij}|}{\max_{j \in U} \pi_j} \right) \left(\max_{j \in U} \frac{|b_{q+k,j}|}{\pi_j} \right)^2, \end{aligned}$$

from which Proposition 2 follows by application of inequality (A1).

REFERENCES

- BARDSLEY, P. & CHAMBERS, R. L. (1984). Multipurpose estimation from unbalanced sampled. *Applied Statistics* **33**, 290–299.
- BATTESE, G. E., HARTER, R. M. & FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28–36.
- BEAUMONT, J. F. & BOCCI, C. (2008). Another look at ridge calibration. *Metron: International Journal of Statistics* **66**, 5–20.
- BREIDT, F. J. & CHAUVET, G. (2009). Improved variance estimation for balanced samples drawn via the cube method. Submitted.
- BREIDT, F. J., CLAESKENS, G. & OPSOMER, J. D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika* **92**, 831–846.
- BREIDT, F. J. & OPSOMER, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics* **28**, 1026–1053.
- CHAMBERS, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics* **12**, 3–32.
- CHAUVET, G. & TILLÉ, Y. (2005). Fast SAS macros for balancing samples: User's guide. Technical report, University of Neuchâtel.
- DATTA, G. & GHOSH, M. (1991). Bayesian prediction in linear models: Applications to small area estimation. *Annals of Statistics* **19**, 1748–1770.
- DEVILLE, J. C. & TILLÉ, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika* **91**, 893–912.

- 769 DEVILLE, J.-C. & TILLÉ, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical*
770 *Planning and Inference* **128**, 569–591.
- 771 ELLIOTT, M. R. & LITTLE, R. J. A. (2000). Model-based alternatives to trimming survey weights. *Journal of*
772 *Official Statistics* **16**, 191–209.
- 773 FAY, R. E. & HERRIOT, R. A. (1979). Estimation of income from small places: An application of James-Stein
774 procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.
- 775 GHOSH, M. & RAO, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science* **9**, 55–93.
- 776 HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Washington, D. C.: Chapman and Hall.
- 777 HORVITZ, D. G. & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite
778 universe. *Journal of the American Statistical Association* **47**, 663–685.
- 779 LAZZERONI, L. C. & LITTLE, R. J. A. (1998). Random-effects models for smoothing poststratification weights.
780 *Journal of Official Statistics* **14**, 61–78.
- 781 MONTANARI, G. E. & RANALLI, M. G. (2009). Multiple and ridge model calibration. In *Proceedings of Workshop*
782 *on Calibration and Estimation in Surveys*. Statistics Canada.
- 783 NYCHKA, D., HAALAND, P., O'CONNELL, M. & ELLNER, S. (1998). FUNFITS, data analysis and statistical tools
784 for estimating functions. In *Case studies in environmental statistics*, D. e. Nychka, W. W. e. Piegorsch & L. H. e.
785 Cox, eds. New York: Springer-Verlag Inc, pp. 159–179.
- 786 OPSOMER, J. D., CLAESKENS, G., RANALLI, M. G., KAUERMANN, G. & BREIDT, F. J. (2007). Nonparametric
787 small area estimation using penalised spline regression. *Journal of the Royal Statistical Society, Series B* **70**,
788 265–286.
- 789 PARK, M. (2002). *Regression estimation of the mean in survey sampling*. Ph.D. thesis, Iowa State University, Ames,
790 Iowa.
- 791 PARK, M. & FULLER, W. A. (2005). Towards nonnegative regression weights for survey samples. *Survey Method-*
792 *ology* **31**, 85–93.
- 793 R DEVELOPMENT CORE TEAM (2008). *R: A Language and Environment for Statistical Computing*. R Foundation
794 for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- 795 RAO, J. N. K. (2003). *Small Area Estimation*. Wiley-Interscience.
- 796 RAO, J. N. K. & SINGH, A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey
797 sampling (Pkg: P57-85). In *ASA Proceedings of the Section on Survey Research Methods*. American Statistical
798 Association.
- 799 ROBINSON, G. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical*
800 *Science* **6**, 15–51.
- 801 RUPPERT, D., WAND, M. P. & CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge
802 University Press.
- 803 SÄRNDAL, C.-E., SWENSSON, B. & WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-
804 Verlag.
- 805 TILLÉ, Y. & MATEI, A. (2008). *sampling: Survey Sampling*. R package version 2.0.
- 806 ZHENG, H. & LITTLE, R. J. A. (2003). Penalized spline model-based estimation of finite population total from
807 probability-proportional-to-size samples. *Journal of Official Statistics* **19**, 99–117.
- 808 ZHENG, H. & LITTLE, R. J. A. (2004). Penalized spline nonparametric mixed models for inference about a finite
809 population mean from two-stage samples. *Survey Methodology* **30**, 209–218.
- 810
- 811
- 812
- 813
- 814
- 815
- 816