Forest Science, Vol. 38, No. 1, pp. 21–33

Empirical Bayes Development of Honduran Pine Yield Models

EDWIN J. GREEN
WILLIAM E. STRAWDERMAN
CHARLES E. THOMAS

ABSTRACT. Occasionally it is of interest to calibrate a given growth or yield model to data from several regions. If it is expected that the parameters from different regions can somehow be regarded as similar, then a Bayesian approach suggests itself. A cursory examination of the literature reveals that most of the theoretical work on empirical Bayes estimation for the linear model has focused on simultaneously estimating the coefficients in one model. In these methods the usual least squares estimates of the model parameters are shrunk toward their mean. One set of parameter estimates results. In contrast we desire a method whereby multiple sets of parameter estimates are produced. We report the results of using such a method to calibrate yield models for unthinned Honduran pine plantations to data from 21 soil–site groups. The models compare favorably to those developed via traditional methods and allow estimation of regression coefficients even for soil–site groups for which the design matrix is not of full rank. FOR. SCI. 38(1):21–33.

FOREST BIOMETRICIANS ARE OFTEN REQUIRED TO ESTIMATE many parameters simultaneously. Burk and Ek (1982), Green and Strawderman (1986), and Green et al. (1987) have shown that if the parameters are similar in some respect, then considerable gains in, say, mean squared error may be realized by using empirical Bayes (*EB*) estimation in lieu of classical techniques (e.g., sample means). However, these authors dealt with the case in which a number of univariate means were to be estimated simultaneously. In this paper we use a method for simultaneously estimating a number of regression coefficient vectors. In particular we shall be simultaneously estimating Honduran pine yield equation coefficients for a number (21) of different soil–site groups. The method we present is useful when one needs to estimate many regression coefficient vectors simultaneously, but when one has limited data with which to estimate each vector individually.

DATA

The data for this study consist of stand level variables measured on unthinned Honduran pine plantations in five Caribbean countries: Costa Rica, Jamaica, Puerto Rico, Trinidad, and Venezuela. Each country was further subdivided into soil–site groups. The sample size (number of plots) for each country and soil group is displayed in Table 1. Note that there are 21 soil–site groups overall. The

TABLE 1.

Sample size (n) per soil group for each country.[1]

Country	code	n	Soil group
Costa Rica	c1	25	Pre-Montane Rain
Costa Rica	c2	23	Tropical Rain
Costa Rica	c3	10	Tropical Wet
Jamaica	j1	8	Cuffy Gully
Jamaica	j2	9	Halls Delight
Jamaica	j3	6	Limestone
Jamaica	j4	9	Mixed Series
Jamaica	j5	9	Valda
Puerto Rico	p1	9	Deep Clays higher than 300m above msl[2]
Puerto Rico	p2	8	Deep Clays lower than 300m above msl
Puerto Rico	p3	4	Lowland Sands
Puerto Rico	p4	2	Shallow Clays
Puerto Rico	p5	4	Upland Sands
Trinidad	t1	6	Alluvial
Trinidad	t2	4	High Upland
Trinidad	t3	12	Interm. Upland
Trinidad	t4	19	Terrace
Venezuela	v1	10	G1 Moderate Rain
Venezuela	v2	11	G2 Moderate Rain
Venezuela	v3	12	Dry Climate
Venezuela	v4	10	Wet Climate

[1] Puerto Rico is a territory of the United States, not an independent country.
[2] msl = mean sea level.

following variables were measured on each plot: volume (m^3/ha), basal area (m^2/ha), trees/ha, height of dominants and codominants (m), and age (yr). Additional details on the data and data collection procedures may be found in Liegel (1991).

OBJECTIVES

This study was initiated by the Institute for Quantitative Studies at the Southern Forest Experiment Station, USDA Forest Service. An earlier attempt to obtain parameters for growth and yield models for Honduran pine had resulted in a series of solutions within countries, but not at the soil–site level due to the small sample sizes for some soil–site groups (Table 1). Ordinary least squares solutions were reported in an internal, unpublished format consisting of analyses by country (Parresol et al. 1987, Liegel 1991). It was apparent that a yield equation of the form:

$$\ln(\hat{V}) = \hat{\beta}_0 + \hat{\beta}_1(A^{-1}) + \hat{\beta}_2\ln(H) + \hat{\beta}_3\ln(N) \quad (1)$$

where $\hat{V}$ is predicted volume (m^3/ha), A is plantation age, H is the height of dominants and codominants (m), N is the number of trees per ha, and the $\hat{\beta}_i$ are

the estimated regression coefficients, was appropriate for all of the countries and we adopted it for this study. The internal report format satisfied the country forestry cooperators; however, it seemed less than satisfactory given apparent differences among soil–site groups. For example, a standard F-test indicated differences among soil groups in Venezuela. Furthermore, the variances between soil–site groups in Trinidad were found to be significantly different using Bartlett's test. Dissatisfaction with the country-level yield models provided the impetus for this study. Our specific objective was to develop a separate yield model for each of the 21 individual soil–site groups. We explicitly allow for heterogeneity among the soil–site group variances.

One might criticize this study for not having a balanced design and claim that better planning should have been done. However, it is often the case that unplanned elements resulting from a study assume importance after completion of the study or inventory. Purcell and Kish (1979), in their study of small domain estimation, designate these as *unplanned domains*. When little chance exists to supplement the data that has been collected, *EB* estimation can often improve the utility of the existing data. Such was the case with the Honduran pine data. The original expectation was to have additional data collected, but the reality was that there was no funding for the followup. We employed *EB* methodology in order to provide the best estimates for the small unplanned sampling domains. *EB* estimators are generally considered useful in situations where many parameters are to be estimated, but data available for the estimation of any one particular class or stratum parameter is limited.

EMPIRICAL BAYES ESTIMATOR

We begin by making the Bayesian assumption that the β_i are realizations from some underlying distribution with a mean and covariance matrix. For convenience, assume the β_i are normal and make the usual regression assumption that the $\ln(V_i)$ are independent and normally distributed about the regression surface. We then have:

$$\mathbf{y}_i|\mathbf{X}_i,\ \beta_i \sim \mathbf{N}(\mathbf{X}_i\beta_i,\ \sigma_i^2\mathbf{I}) \tag{2}$$

$$\beta_i \sim \mathbf{N}(\boldsymbol{\theta},\ \boldsymbol{\Sigma}) \tag{3}$$

where $\mathbf{y}_i$ is an $(n_i \times 1)$ vector of the logarithms of plot volumes for the ith soil group; $\mathbf{X}_i$ is the $(n_i \times 4)$ design matrix for the ith soil group (i.e., the jth row of $\mathbf{X}_i$ is: $1, A_{ij}^{-1}, \ln(H_{ij}), \ln(N_{ij})$); n_i is the number of observations or plots per soil group; $\sigma_i^2\mathbf{I}$ is the covariance matrix for $\mathbf{y}_i$; and $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ are the mean and covariance for β_i, respectively. This is what is meant by similar, i.e., data from regions other than region i have some relevant information on the parameter vector for region i, in the form given by (3).

In the terminology of Bayesian statistics, (2) specifies the likelihood, while (3) represents the prior distribution of β_i. As a check on the normality assumption in (3), Q—Q plots of the empirical distribution function of the 20 OLS estimates for each β_i in (1) against a standard normal cumulative distribution function were

examined. The plots indicated that the normality assumption was acceptable for β_2 and β_3, while there was some indication of non-normality for β_1 and β_4. However, inasmuch as the departures from normality for these two coefficients were not extreme, and given that empirical Bayes procedures are generally considered to be robust to the distributional assumptions on the prior, we concluded that the normality assumption was acceptable.

In some cases, model (3) may be insufficient. One might possess information on some auxiliary variables, or covariates, and one might hypothesize that the β_i are correlated with these covariates. In this case, rather than assuming $\beta_i \sim \mathbf{N}(\boldsymbol{\theta},\boldsymbol{\Sigma})$, it might be more sensible to assume $\beta_i \sim \mathbf{N}(\boldsymbol{\Delta\Gamma},\boldsymbol{\Sigma})$, where $\boldsymbol{\Delta}$ is a matrix of covariates, and $\boldsymbol{\Gamma}$ is matrix of hyperparameters. Such a procedure is discussed in Braun and Jones 1985, Braun 1987, and Green et al. 1990. Alternatively, one might prefer to group soil–site groups into classes which were felt to be similar, and then model these classes separately, e.g.,

$$\mathbf{y}_{ij}|\mathbf{X}_{ij},\ \beta_{ij} \sim \mathbf{N}(\mathbf{X}_{ij}\beta_{ij},\ \sigma_{ij}^2\mathbf{I})$$
$$\beta_{ij} \sim \mathbf{N}(\boldsymbol{\theta}_i,\ \boldsymbol{\Sigma}_i)$$

where $\mathbf{y}_{ij}$ represents the log-volume for the jth soil–site group in the ith class, and the definitions of $\mathbf{X}_{ij}$, β_{ij}, σ_{ij}^2, $\boldsymbol{\theta}_i$, and $\boldsymbol{\Delta}_i$ follow in an obvious way. Green et al. (1990) examined the former method, and we believe the added complexity over the model proposed here was unwarranted. Regarding the latter method, it was not obvious how many classes would be necessary or which soil–site groups belonged in separate classes. Thus we adopted the model shown in (2) and (3). As will be seen later, this implies that the solution for soil–site groups with exceedingly sparse data will tend to be weighted heavily towards an overall mean solution for all the soil–site groups.

Lindley and Smith (1971) presented a fully Bayesian solution for models (2) and (3), where homogeneous variance was assumed, i.e., $\sigma_i^2 = \sigma^2$ for all i. Their technique requires specification of vague prior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. An alternative fully Bayesian solution was reported by Liu (1981). Liu's technique was also based on the assumption of homogeneous variance, but differed from that of Lindley and Smith in the choice of the prior distribution for $\boldsymbol{\Sigma}$. Liu presented some heuristic arguments supporting his method, but there is no universal agreement on the proper choice of a vague or noninformative prior for a covariance matrix such as $\boldsymbol{\Sigma}$. Furthermore, neither of the above studies allowed for heterogeneous variance. For these reasons, we adopted an empirical Bayes approach. The *EB* method is similar to fully Bayes procedures, except that instead of specifying prior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ these quantities are estimated from the data. Note that the term *empirical Bayes* is here used to refer to a procedure in which the parameters of the prior distribution ($\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$) are estimated from the data. This is in keeping with Robbins' (1951) original use of the term. We do not mean to imply that we have observed historical data and derived the prior distribution by looking at the empirical distribution function of the data [indeed Green (1990) has argued that the latter procedure is true Bayes]. For a more complete discussion of empirical Bayes methods, see Morris (1983) or Casella (1985), among others.

At this point, we note that the structure of models (2) and (3) is identical to a structure often used in random coefficient problems in econometrics (see, e.g.,

Judge et al. 1985, p. 806–809). The usual econometrics approach is to estimate the β_i with a generalized least squares procedure. Unfortunately, such procedures generally do not guarantee a positive definite estimate for $\boldsymbol{\Sigma}$. The *EB* procedure does provide this guarantee.

It is well known that the combination of a normal likelihood and a normal prior distribution results in a normal posterior distribution (e.g., see Box and Tiao 1973, p. 74–75). Thus, models (2) and (3) imply $(\beta_i|\boldsymbol{\theta},\boldsymbol{\Sigma},\sigma_i^2) \sim \mathbf{N}(\boldsymbol{\delta}_i, \boldsymbol{\Psi}_i)$ where

$$\boldsymbol{\delta}_i = (\boldsymbol{\Sigma}^{-1} + \sigma_i^{-2}\,\mathbf{X}_i'\mathbf{X}_i)^{-1}\,[\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\theta} + \sigma_i^{-2}\,\mathbf{X}_i'\mathbf{y}_i] \tag{4}$$

and

$$\boldsymbol{\Psi}_i = (\boldsymbol{\Sigma}^{-1} + \sigma_i^{-2}\,\mathbf{X}_i'\mathbf{X}_i)^{-1} \tag{5}$$

In keeping with traditional Bayesian analysis, we will estimate β_i with the posterior mean (or mode; the two are identical because the posterior is normal), $\boldsymbol{\delta}_i$. In the *EB* approach we find maximum likelihood estimates of $\boldsymbol{\Sigma}$, $\boldsymbol{\theta}$ and σ_i^2 and insert these for the appropriate quantities in (4) to yield the *EB* estimate of $\boldsymbol{\delta}_i$.

Solutions to this problem *seem* equivocal; we can't solve for the $\boldsymbol{\delta}_i$ without knowing $\boldsymbol{\Delta}$, $\boldsymbol{\theta}$, and σ_i^2, and we can't obtain estimates for $\boldsymbol{\Sigma}$, $\boldsymbol{\theta}$, and σ_i^2 without $\boldsymbol{\delta}_i$. Conventional methods cannot be used. An iterative solution suggests itself, but starting values for some of the parameters must be assumed and could be critical to convergence. Fortunately an algorithm has been developed that *guarantees* convergence to the maximum likelihood estimates of $\boldsymbol{\Sigma}$, $\boldsymbol{\theta}$ and σ_i^2. As reported by Green et al. (1990), this problem has been considered by researchers at the Educational Testing Service (see, e.g., Rubin 1980, Braun and Jones 1981, 1985, Braun et al. 1983, Braun 1987). Their solution, which we adopt here, was to use the *EM* algorithm to estimate $\boldsymbol{\delta}_i$.

EM Algorithm

The *EM* algorithm was formalized by Dempster et al. (1977). *EM* is iterative and consists of an estimation (*E*) step in which conditional values of the sufficient statistics are determined, followed by a maximization (*M*) step during which the likelihood function is maximized based on the conditional values of the sufficient statistics. For the *EB* model specified in (2) and (3), the *EM* algorithm reduces to the following: Let $\hat{\sigma}_i^2$, $\hat{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{\theta}}$ be initial guesses for σ_i^2, $\boldsymbol{\Sigma}$, and $\boldsymbol{\theta}$. The *E*-step consists of computing conditional values for the sufficient statistics for $\boldsymbol{\delta}_i$ and $\boldsymbol{\Psi}_i$, the parameters of the posterior distribution for β_i:

$$\hat{\boldsymbol{\delta}}_i = E(\beta_i|\mathbf{y}_i,\, \mathbf{X}_i,\, \hat{\sigma}_i^2,\, \hat{\boldsymbol{\Sigma}},\, \hat{\boldsymbol{\theta}}) \tag{6}$$

$$= (\hat{\boldsymbol{\Sigma}}^{-1} + \hat{\sigma}_i^{-2}\,\mathbf{X}_i'\mathbf{X}_i)^{-1}\,[\hat{\boldsymbol{\Sigma}}^{-1}\,\hat{\boldsymbol{\theta}} + \hat{\sigma}_i^{-2}\,\mathbf{X}_i'\mathbf{y}_i] \tag{7}$$

and

$$\hat{\boldsymbol{\Lambda}}_i = E(\beta_i\beta_i'|\mathbf{y}_i,\, \mathbf{X}_i,\, \hat{\sigma}_i^2,\, \hat{\boldsymbol{\Sigma}},\, \hat{\boldsymbol{\theta}}) \tag{8}$$

$$= \hat{\boldsymbol{\delta}}_i\hat{\boldsymbol{\delta}}_i' + (\hat{\boldsymbol{\Sigma}}^{-1} + \hat{\sigma}_i^{-2}\,\mathbf{X}_i'\mathbf{X}_i)^{-1}. \tag{9}$$

Next the M-step is performed, during which updated estimates of σ_i^2, and the prior parameters, $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$, are computed, conditional upon $\hat{\boldsymbol{\delta}}_i$ and $\hat{\boldsymbol{\Lambda}}_i$:

$$\hat{\boldsymbol{\theta}} = m^{-1} \sum_{i=1}^{m} \hat{\boldsymbol{\delta}}_i \tag{10}$$

$$\hat{\boldsymbol{\Sigma}} = m^{-1} \left(\sum_{i=1}^{m} \hat{\boldsymbol{\Lambda}}_i - \hat{\boldsymbol{\theta}}\, \hat{\boldsymbol{\theta}}' \right) \tag{11}$$

and

$$\hat{\sigma}_i^2 = n_i^{-1} \left(\mathbf{y}_i'\mathbf{y}_i - 2\, \hat{\boldsymbol{\delta}}_i \mathbf{X}_i \mathbf{y}_i + \sum_{j,k} (s_{jk}^i)(w_{jk}^i) \right), \tag{12}$$

where s_{jk}^i is the jk^{th} element of $\hat{\boldsymbol{\Lambda}}_i$, w_{jk}^i is the jkth element of $(\mathbf{X}_i'\mathbf{X}_i)$, and m is the number of soil–site groups, i.e., the number of regression coefficient vectors desired. Following the M-step, the E-step is then repeated with the updated guesses for σ_i^2, $\boldsymbol{\Sigma}$, and $\boldsymbol{\theta}$. Iterations continue until $\hat{\boldsymbol{\delta}}_i$, $\hat{\boldsymbol{\Lambda}}_i$, $\hat{\sigma}_i^2$, $\hat{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{\theta}}$ converge. As reported by Dempster et al. (1977), successive iterations *always* increase the likelihood function and convergence implies a stationary point on the likelihood, hence the *MLE*. As with all iterative procedures, use of *EM* requires specification of a stopping rule. Two stopping criteria readily suggest themselves: stopping when the change in the value of likelihood function between successive iterations is sufficiently small, or stopping when the change in the values of the estimated parameters between successive iterations is sufficiently small. In this study, we chose the latter method, although we could just as easily have selected the former. For more details on the application of *EM* to models (2) and (3), see Rubin (1980), Braun et al. (1983), or Braun and Jones (1985). Note that following Braun and Jones (1985), we do not obtain simultaneous empirical Bayes estimates of $\boldsymbol{\delta}_i$ and σ_i^2 (i.e., no prior is specified for σ_i^2; thus although the *MLE* for σ_i^2 is included in the M-step, there are no sufficient statistics for the prior parameters of σ_i^2 in the E-step). This is a much more complicated problem and cannot be solved directly.

One interesting aspect of the *EB* method is that it permits estimation of $\boldsymbol{\delta}_i$ *even if* $(\mathbf{X}_i'\mathbf{X}_i)$ *is not of full rank.* Inspection of (6) through (12) reveals that we are never required to compute $(\mathbf{X}_i'\mathbf{X}_i)^{-1}$. Thus it is possible to compute an *EB* estimate of $\boldsymbol{\delta}_i$ even in situations where there is too little data to permit computation of the usual *OLS* estimator.[1] When $(\mathbf{X}_i'\mathbf{X}_i)$ is nonsingular, the estimate of $\boldsymbol{\delta}_i$ at the $(k+1)$th iteration can be written as the precision (inverse variance) weighted combination of the kth estimate of $\hat{\boldsymbol{\theta}}$ and the OLS estimate $\hat{\boldsymbol{\beta}}_i$:

$$\begin{aligned} \hat{\boldsymbol{\delta}}_{i,k} = {} & (\hat{\boldsymbol{\Sigma}}_k^{-1} + \hat{\sigma}_{i,k}^{-2} \mathbf{X}_i'\mathbf{X}_i)^{-1} \hat{\boldsymbol{\Sigma}}_k^{-1}\, \hat{\boldsymbol{\theta}}_k \\ & + (\hat{\boldsymbol{\Sigma}}_k^{-1} + \hat{\sigma}_{i,k}^{-2} \mathbf{X}_i'\mathbf{X}_i)^{-1}\, \hat{\sigma}_{i,k}^{-2}\, \mathbf{X}_i'\mathbf{X}_i\, \hat{\boldsymbol{\beta}}_i \end{aligned}$$

[1] However, if there are groups with p or fewer observations, where P = length of $\hat{\boldsymbol{\beta}}_i$, one may encounter difficulties in maximizing the likelihood function (see section on Final Estimates). This seems to have been overlooked by previous investigators.

Thus the *EB* estimate can be thought of as a weighted average between the estimate of the mean vector $\hat{\boldsymbol{\theta}}$ and the OLS estimate, where the weight is proportional to the inverse of each component's variance. Hence *EB* may be regarded as a compromise between using $\hat{\boldsymbol{\theta}}$ and the *OLS* estimate, where ***the data are permitted to dictate the weight applied to each.*** As mentioned previously, the *EB* estimate for soil–site groups with sparse data will thus be weighted heavily toward $\hat{\boldsymbol{\theta}}$.

COMPARISON OF *EB* AND *OLS*

To assess the performance of yield equations developed via the *EB* procedure, we decided to compare them with equations fitted using *OLS*. The comparison was accomplished by repeated random divisions of the data set into calibration and validation sets, fitting the models with the two candidate procedures on the calibration set, and comparing the performance of the fitted equations on the validation data. Approximately 75% of the observations for each soil group were included in the calibration set, and the remaining observations were put into the validation set. In order to have enough data in the calibration and validation sets to make meaningful comparisons it was necessary to eliminate some soil–site groups from consideration. We decided to exclude any soil group with fewer than 10 observations. This reduced the number of soil–site groups from 21 to 9.

As mentioned earlier, when fitting the yield equations via the *EB* method, we had to specify a stopping rule for the iterative *EM* algorithm. We chose to terminate the iterations when the maximum change over all the individual elements of $\hat{\boldsymbol{\delta}}_i$, $\hat{\boldsymbol{\Lambda}}_i$, $\hat{\sigma}_i^2$, $\hat{\boldsymbol{\Sigma}}$, and $\hat{\boldsymbol{\theta}}$ on successive iterations was less than 0.1%.

The data-splitting process was performed 1000 times. The average number of iterations required by the *EM* algorithm to achieve the stopping criterion was

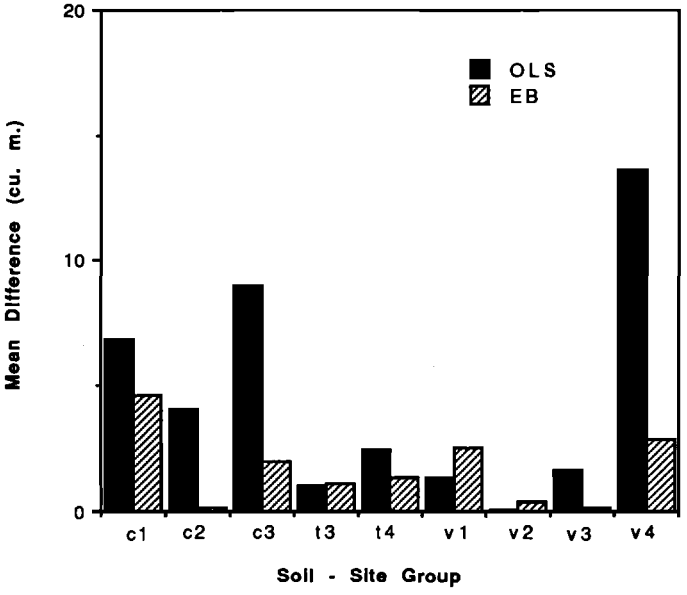

FIGURE 1. Absolute value of mean arithmetic difference by soil–site group.

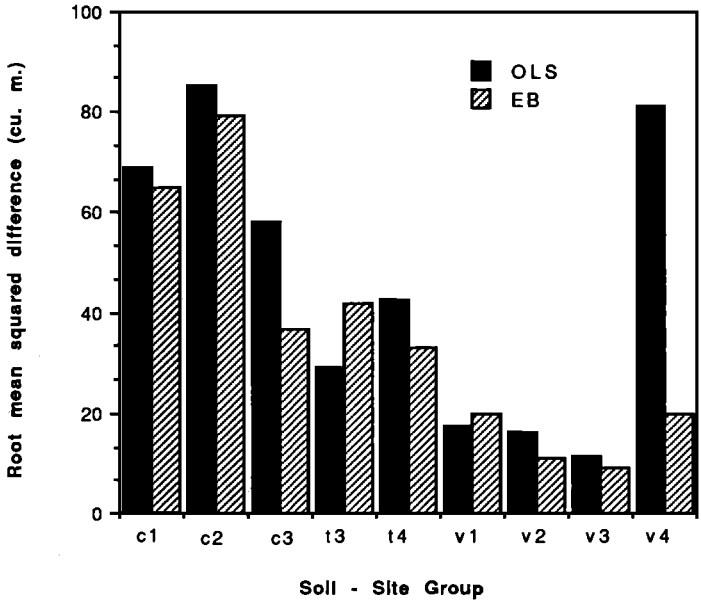

FIGURE 2. Root mean squared difference by soil–site group.

110.8. The mean arithmetic, squared, and absolute differences between observed and predicted values on the validation set, averaged over the 1000 simulations, are shown for each estimation procedure by soil group in Figures 1, 2, and 3 respectively. Note that model [1] predicts $\ln(V)$, but we are really interested in V. Thus the differences in Figures 1, 2, and 3 are in terms of V. For both estimation procedures, we simply exponentiated the model predictions to obtain predictions

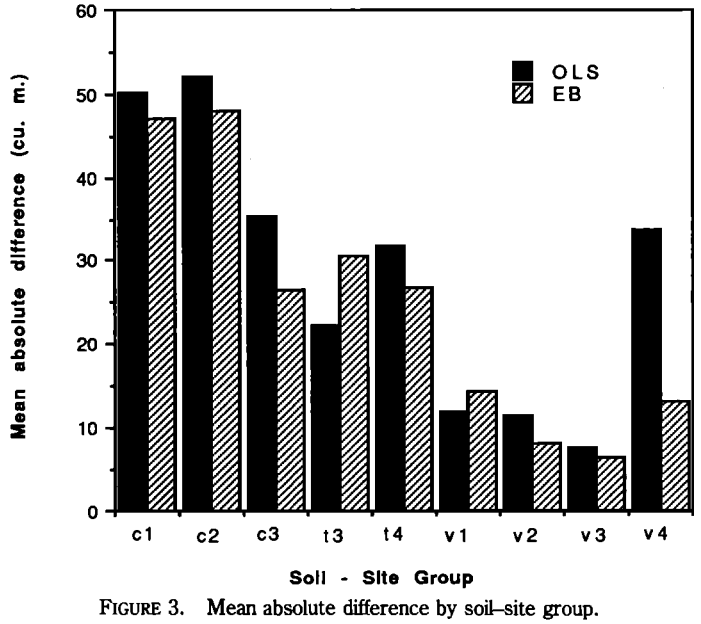

FIGURE 3. Mean absolute difference by soil–site group.

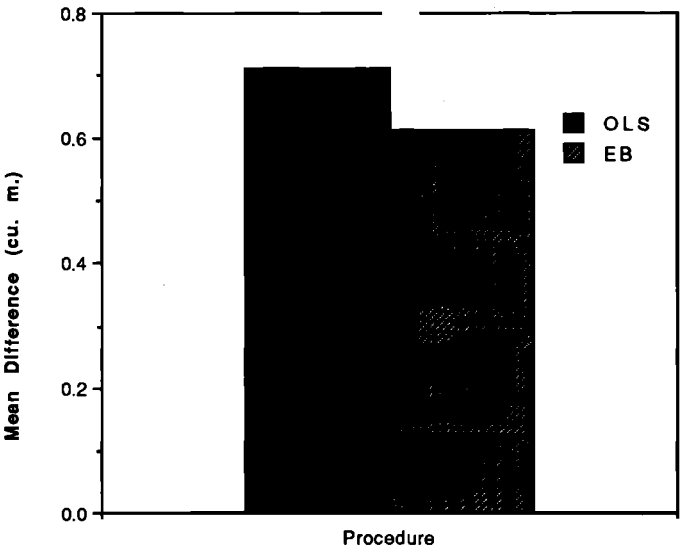

FIGURE 4. Absolute value of mean arithmetic differences over all soil groups.

of V. For ease of interpretation, we display the absolute value of the mean differences in Figure 1. Actually, the mean difference was negative for soil–site groups c3, t4, v1, v3, and v4 for the *OLS* procedure, and for soil–site groups c2, t3, v1, v3, and v4 for the *EB* procedure.

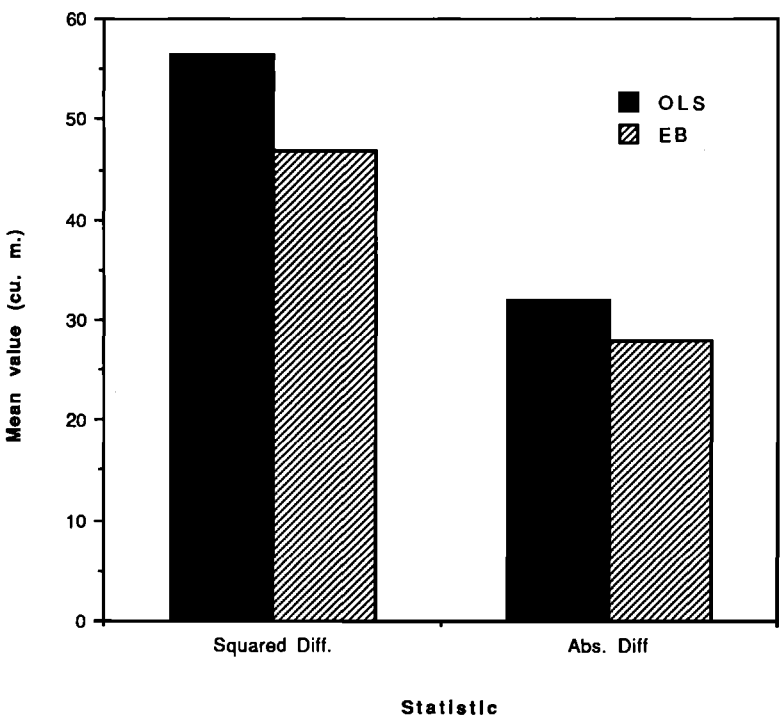

FIGURE 5. Mean squared difference and mean absolute difference over all soil groups.

Figures 1, 2, and 3 indicate that *EB* was superior in predictive ability to *OLS* on a per soil group basis. *EB* yielded lower absolute mean arithmetic differences for six of nine soil–site groups, and lower root mean squared differences and mean absolute differences for seven of the nine groups.

In Figure 4, we display the absolute value of the mean arithmetic differences over the nine soil–site groups. Note that these are the mean differences over the entire validation set, averaged over the 1000 simulations. *EB* yielded a 14% reduction in mean difference over all nine soil–site groups. The root mean squared differences and mean absolute differences over all nine soil–site groups are displayed in Figure 5. *EB* yielded a 17% reduction in overall root mean squared error and a 12% reduction in overall absolute error.

Given the results in Figures 1–5, it appears that the *EB* models were both more precise (as judged by squared and absolute error) and more accurate (as judged by arithmetic error) than the *OLS* models, both on a per soil group basis, and on an overall basis.

As stated earlier, the *EB* estimates can be viewed as precision-weighted averages of the *OLS* estimates and $\hat{\boldsymbol{\theta}}$. Thus we see that the *EB* estimator is a ***shrinkage*** estimator. The *OLS* estimates are ***shrunk*** toward $\hat{\boldsymbol{\theta}}$. An occasional criticism of shrinkage estimators is that while they may do well in terms of overall performance, they can miss badly on individual components (soil–site groups). To examine this, we computed the maximum difference between observed and predicted plot volume over all plots and soil–site groups for each division of the data.

TABLE 2.

Sample variances over 1000 simulations of $\hat{\delta}_i$. Each row contains variances for $\hat{\delta}_{ij}$, $j = 1, \ldots, 4$.

	Coefficient			
Soil Group Code[1]	1	2	3	4
OLS				
c1	1.797	0.786	0.018	0.025
c2	0.724	0.568	0.028	0.012
c3	5.711	18.602	0.607	0.268
t3	0.706	3.302	0.063	0.005
t4	1.873	4.997	0.071	0.014
v1	2.680	2.129	0.057	0.034
v2	9.067	6.236	0.262	0.115
v3	0.596	3.628	0.046	0.033
v4	26.978	27.237	0.460	0.259
EB				
c1	0.344	0.553	0.010	0.006
c2	0.321	0.339	0.009	0.007
c3	0.438	2.408	0.046	0.008
t3	1.396	10.669	0.180	0.010
t4	0.432	1.393	0.021	0.004
v1	4.200	5.570	0.126	0.036
v2	0.935	1.019	0.031	0.014
v3	0.263	1.140	0.019	0.008
v4	0.640	2.833	0.048	0.008

[1] Soil group codes identified in Table 1.

The means (over the 1000 divisions of the data) of these maximum deviations were 206.19 for *OLS* and 170.01 for *EB*. Thus *EB* performed better in this regard also. Finally, to examine the stability of the estimates, we computed the sample variance over the 1000 simulations of each of the 36 coefficients (9 groups × 4 coefficients per group). These are displayed in Table 2. As might have been expected, *OLS* yielded more stable estimates for two soil–site groups; the same two for which it provided better predictions (as determined by squared or absolute error). *EB* yielded more stable estimates for the other seven groups.

FINAL ESTIMATES

Our objective was to obtain parameters for all soil–site groups of plots. Based on the results of our simulation experiment, we used the empirical Bayes approach. Model [1] was fitted to the data from all 21 soil–site types in the five countries. For this final fitting of the model, we thus included three regions with four observations, and one with two. We decided to employ maximization of the likelihood as the stopping criterion for the final estimates. We regarded the likelihood to be maximized when the change in the log-likelihood was less than

TABLE 3.

Final empirical Bayes estimates for all 21 soil groups.

Soil Group Code[1]	n	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$	$\hat{\sigma}^2$
c1	25	−4.864	−1.082	1.955	0.708	0.035
c2	23	−4.912	−0.963	1.977	0.707	0.061
c3	10	−4.784	−0.951	1.961	0.681	0.048
j1	8	−4.727	−0.920	1.941	0.675	0.047
j2	9	−4.885	−1.018	1.966	0.705	0.081
j3	6	−4.796	−0.953	1.982	0.686	0.048
j4	9	−4.798	−0.971	1.918	0.692	0.037
j5	9	−4.800	−0.933	1.968	0.686	0.018
p1	9	−5.050	−1.045	2.029	0.728	0.052
p2	8	−4.792	−0.954	1.978	0.685	0.038
p3	4	−5.120	−1.092	2.042	0.739	0.019
p4	2	−4.508	−0.637	1.862	0.630	0.019
p5	4	−4.609	−0.925	1.887	0.661	0.019
t1	6	−4.721	−0.949	1.932	0.679	0.020
t2	4	−4.750	−0.983	1.923	0.684	0.019
t3	12	−4.750	−1.047	1.922	0.687	0.015
t4	19	−4.698	−0.909	1.943	0.667	0.025
v1	10	−4.736	−1.020	1.983	0.675	0.045
v2	11	−4.867	−0.965	1.978	0.699	0.012
v3	12	−4.841	−1.069	1.957	0.700	0.010
v4	10	−4.674	−0.858	1.931	0.666	0.026
$\hat{\boldsymbol{\theta}}$		−4.800	−0.971	1.953	0.689	
$\hat{\boldsymbol{\Sigma}}$		22.045	4.432	−8.937	−3.163	
		4.432	1.012	−1.798	−0.643	
		−8.937	−1.798	3.636	1.282	
		−3.163	−0.643	1.282	0.454	

[1] Soil group codes identified in Table 1.

0.0001% between iterations of the *EM* algorithm. Not surprisingly, the *EM* algorithm tended to converge to a solution close to the perfect least squares fit for one of the soil–site groups with four observations. This caused the likelihood to increase towards infinity, with no maximization possible. To counter this we imposed the constraint that $\hat{\delta}_{11}^2 = \hat{\sigma}_{12}^2 = \hat{\sigma}_{13}^2 = \hat{\sigma}_{15}^2$ (soil–site groups 11, 12, 13, and 15 had 4, 2, 4, and 4 observations, respectively).

The final estimates are reported in Table 3. *EM* required 11 iterations to achieve the stopping criterion. The procedure provided solutions even for the smallest soil group, which had only two plots. In addition, the parameter estimates are similar to those obtained by the USDA Forest Service (Parresol et al. 1987) using coarser data groupings.

CONCLUSIONS

We believe that this procedure will provide foresters with a valuable method for fitting models to data from multiple groups, regions, or species when data from the individual groups are scarce. Of course, the assumptions in models [2] and [3] must be tenable. If there is no reason to expect that the coefficients from two groups are similar, then it might not be prudent to use the *EB* procedure.

LITERATURE CITED

BOX, G.E.P., and G.C. TIAO. 1973. Bayesian inference in statistical analysis. Addison-Wesley, Reading, MA. 588 p.

BRAUN, H.I. 1987. Empirical Bayes methods: A tool for exploratory analysis. Educational Testing Service, Princeton, NJ. Unpublished report.

BRAUN, H.I., and D.H. JONES. 1981. The Graduate Management Admission test: Prediction bias study. ETS Res. Rep. #RR-81-25, Educational Testing Service, Princeton, NJ.

BRAUN, H.I., and D.H. JONES. 1985. Use of empirical Bayes methods in the study of the validity of academic predictors of graduate school performance. ETS Res. Rep. #84-34, Educational Testing Service, Princeton, NJ.

BRAUN, H.I., D.H. JONES, D.B. RUBIN, and D.T. THAYER. 1983. Empirical Bayes estimation of coefficients in the general linear model from data of deficient rank. Psychometrica 48:171–181.

BURK, T.E., and A.R. EK. 1982. Application of empirical Bayes/James Stein procedures to simultaneous estimation problems in forestry. For. Sci. 28:753–771.

CASELLA, G. 1985. An introduction to empirical Bayes analysis. Am. Stat. 39:83–87.

DEMPSTER, A.P., N.M. LAIRD, and D.B. RUBIN. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). JRSS, Series B 39:1–38.

GREEN, E.J. 1990. Statistics in forestry: The next ten years (a case of extrapolation beyond the range of the data). P. 105–115 *in* Proc. Division 6, IUFRO World Congress, Montreal.

GREEN, E.J., and W.E. STRAWDERMAN. 1986. Stein-rule estimation of coefficients for 18 eastern hardwood cubic volume equations. Can. J. For. Res. 16:249–255.

GREEN, E.J., C.E. THOMAS, and W.E. STRAWDERMAN. 1987. Stein-rule estimation of timber removals by county. For. Sci. 33:1054–1061.

GREEN, E.J., W.E. STRAWDERMAN, and C.E. THOMAS. 1990. Empirical Bayes methods for calibrating yield models to multiple regions. P. 89–96 *in* Proc. IUFRO Conference on Forest Simulation Systems, Wensel, L.C., and G.S. Biging (eds.).

JUDGE, G.G., ET AL. 1985. The theory and practice of econometrics. Ed. 2. Wiley, New York. 1019 p.

LIEGEL, L.H. 1991. Growth and site relationships of *Pinus caribaea* across the Caribbean. USDA For. Serv. South. For. Exp. Stn. Gen. Tech. Rep. In press.

LINDLEY, D.V., and A.F.M. SMITH. 1972. Bayes estimates for the linear model (with discussion). JRSS, B, 34:1–41.

LIU, L.M. 1981. Estimation of random coefficient regression models. J. Stat. Comp. Sim. 13:27–39.

MORRIS, C.N. 1983. Parametric empirical Bayes inference: Theory and applications (with discussion). JASA 78:47–65.

PARRESOL, B.R., K.R. DOBELBOWER, and T.R. DELL. 1987. Honduran pine yield systems. USDA For. Serv. South. For. Exp. Stn. A series of five internal reports by country. 6 p. and appendices. On file with Inst. for Quantitative Studies.

PURCELL, N.J., and L. KISH. 1979. Estimation for small domains. Biometrics 35:365–384.

ROBBINS, H. 1951. Asymptotically subminimax solutions of compound statistical decision problems. P. 157–163 *in* Proc. 2nd Berkeley Symp. Math. Stat. Prob. 1, University of California Press, Berkeley, CA.

RUBIN, D.B. 1980. Using empirical Bayes techniques in the law school validity studies (with discussion). JASA 75:801–827.

Manuscript received March 4, 1991

AUTHORS AND ACKNOWLEDGMENTS

Edwin J. Green, William E. Strawderman, and Charles E. Thomas are Associate Professor of Forest Biometrics, Cook College, Rutgers University, P.O. Box 231, New Brunswick, NJ 08903; Professor and Chairman, Department of Statistics, Rutgers University, New Brunswick, NJ 08903; and Research Forester, Institute for Quantitative Studies, South. For. Exp. Stn., 701 Loyola Avenue, New Orleans, LA 70113. Paper of the Journal Series, NJ Agricultural Experiment Station, Cook College, Rutgers University. This work was partially funded by NJAES Project Number D-17386-01-92, supported by NJAES and U.S. McIntire-Stennis Act funds, by the USDA Forest Service Southern Forest Experiment Station, and by the Program in Science and Technology Cooperation (PSTC), Project AID/SCI/E2/06, administered by the Office of the Science Advisor, US AID.