1989

Methods of Parameter Estimation of Linear Regression Models for Yield Prediction.

Man Yong Shin
Louisiana State University and Agricultural & Mechanical College

Order Number 9025340

Methods of parameter estimation of linear regression models for yield prediction

Shin, Man Yong, Ph.D.

The Louisiana State University and Agricultural and Mechanical Col., 1989

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

METHODS OF PARAMETER ESTIMATION OF LINEAR REGRESSION MODELS FOR YIELD PREDICTION

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in
The School of Forestry, Wildlife, and Fisheries

by
Man Yong Shin
B.A., Kyung-Hee University, 1981
M.A., Kyung-Hee University, 1983
M.S., Iowa State University, 1986
December 1989

ACKNOWLEDGEMENTS

I wish to express my special thanks and sincere gratitude to my major professor, Dr. Quang V. Cao, for providing me the opportunity to pursue this program and for his understanding and guidance during the course of this research. I would also like to thank the other members of my committee, Drs. Jimmy L. Chambers, Luis A. Escóbar, James E. Hotvedt, and Ben D. Jackson, for their advice and encouragement during the course of this study.

My appreciation is also extended to the Committee on Southern Forest Tree Improvement and the USDA Forest Service, for the data used in this study. I am also grateful to both graduate and undergraduate students in the School of Forestry, Wildlife, and Fisheries for their friendship and advice during my stay at LSU.

The deepest gratitude goes to my parents in Korea, for supporting me in many ways throughout my long academic career. And finally, to my wife Tae Hee, who has understood and supported me for last six and half years in the United States of America. I would also like to share this great pleasure with my lovely kids, Gee Hae and Samuel.

TABLE OF CONTENTS

LIST OF TABLES

STUDY III

ABSTRACT

Different parameter estimation methods of yield prediction models were investigated using data from the Southwide Loblolly Pine Seed Source Study. This project consisted of three distinct studies. Each study dealt with a possible situation in which other parameter estimation methods rather than the ordinary least squares (OLS) estimator might be used. Three different evaluation statistics were computed to select the "best" estimation method for each situation.

The objective of the first study was to select the best estimator for a yield model which had multicollinearity among independent variables. Three types of biased estimators were compared with the ordinary least squares estimator in terms of the predictive ability of the yield model. Ridge estimators were better than the OLS estimator in dealing with multicollinearity problems. Among methods used for selecting the ridge parameter k, Mallows's (1973) C_k statistic provided the best ridge estimator. On the other hand, principal components and Stein-rule estimators performed poorly compared to the OLS estimator in prediction problems. However, the improvement of yield prediction by ridge estimator was not enough in terms of volume per acre. Thus, the OLS estimator might be preferable due to the simplicity.

The second study dealt with the calibration of yield prediction models to a specific locality and seed source by using Stein-rule estimators. The Stein-rule estimators provided better

yield prediction for a specific locality than OLS estimators. For seed sources, however, the Stein-rule estimators offered little gain in prediction compared with the OLS estimators.

In the third study, Kalman filter estimators were used to update yield prediction models by combining OLS estimators from the sample data with some prior information. Two different sources of prior information were applied in this study. Kalman filter estimators performed better in both cases than OLS estimators. Kalman filter estimators also predicted yield better when prior information was obtained from inside the study area than from outside of the study area.

INTRODUCTION

Regression techniques are used to predict the variable of interest based on the relation between two or more quantitative variables. Yield prediction models are regression equations that express yield per unit area as a function of age, measures of site quality, and stand density. Yield per unit area can be easily predicted by substituting the stand attributes from inventory data into the yield model.

Coefficients of yield models have primarily been estimated by using the ordinary least squares (OLS) estimation method, due to several favorable properties of the OLS estimator. The OLS estimator is unbiased and has the smallest variance among all linear unbiased estimators. However, precise yield prediction under some situations may not be accomplished with the OLS estimation method. Other parameter estimation techniques should be considered as alternative to OLS in order to improve yield prediction.

There are several problems associated with forestry data. When a yield prediction model is developed, multicollinearity might be considered because of its bad effects on the yield prediction. The adverse effects of multicollinearity on regression models have been emphasized by numerous authors (Hoerl and Kennard 1970a, 1970b; Brown and Beattie 1975; Chatterjee and Price 1977; Mitchell and Hann; 1979; Belsley et al. 1980; Bare and Hann 1981). To deal

with the multicollinearity problem, biased estimation methods often have been used as alternatives to the OLS technique. These methods result in estimators that are biased but have lower mean squared error compared to OLS estimators. However, research is needed to determine which biased estimator performs better than the OLS estimator in terms of prediction capability of yield models.

Another possible problem that forest managers may face is the application of yield models, that are based on the entire region, to a small subregion of interest. The variability in environmental factors of the small subregion is not fully explained by the regional model. So far, a few estimation techniques have been adopted to adjust the regional parameters. A Stein-rule estimator can be used to calibrate yield prediction models to different subregions. In this process, the regional parameter estimates are appropriately weighted to fit to a specific locality or seed source.

The other problem is to improve the precision of yield models without collecting more sample data. This can be done by updating OLS parameter estimates with some prior information. These types of modification methods are known as feedback procedures.

Kalman filter estimation is one of the feedback procedures that combines the OLS estimate with prior information by using Bayesian estimation methodology. The Kalman filter estimator is simple and very straightforward in application because no assumption is made in the distributional form of the sample data except for the assumption on the errors. The Kalman filter

estimator should be evaluated against OLS estimator to select a proper parameter estimation method for yield prediction models.

With the consideration of the possible situations described above, this study was divided into three distinct cases and conducted to identify the "best" estimation method for each case. The objectives of this study are as follows:

1) to improve the precision of yield models by using biased estimations,
2) to calibrate yield models for different localities and different seed sources by using a Stein-rule estimator, and
3) to update yield models by using Kalman filter estimation techniques.

STUDY I

BIASED ESTIMATION OF PARAMETERS FOR YIELD PREDICTION MODELS

ABSTRACT

Three types of biased regression estimators were compared to the ordinary least squares (OLS) estimator in order to select the "best" estimator when multicollinearity existed. The biased estimators were ridge regression, principal components regression, and Stein-rule estimators. The evaluation was conducted based on the predictive ability of a yield model developed by Matney et al. (1988). A total of 522 plots from the data of the Southwide Loblolly Pine Seed Source study was used in this study.

All three ridge estimators were better than OLS in terms of predictive ability. The ridge estimator obtained by using Mallows's (1973) C_k statistic performed the best. On the other hand, the other two biased estimators, principal components and Stein-rule estimators (James and Stein 1961), performed poorly when compared to the OLS estimator. Thus, ridge estimators can be recommended as an alternative estimator when multicollinearity exists among independent variables. However, The performances of all estimators did not show any enough difference in terms of evaluation statistics. Thus, the OLS estimator might be preferable due to its simplicity.

INTRODUCTION

Foresters are often required to make estimates of wood volume yield. Yield estimation accomplishes a key role in supporting management plans and determining the amount of cutting on the forest. Therefore, accurate yield prediction is essential to effective forest management planning.

Multiple linear regression techniques have been employed in the development of yield prediction models since Mackinney and Chaiken (1939) first applied them to loblolly pine stands. Model parameters usually have been estimated using the ordinary least squares (OLS) method, that produces estimates that have lower variance than other linear unbiased estimators. However, the OLS estimators can have large variance when multicollinearity exists among variables in the data.

Yield prediction models require stand variables such as age, density, and site index as independent variables. Since the yield models are developed by multiple linear regression techniques, the presence of multicollinearity should be considered in the estimation of parameters for the prediction models. If high correlation exists between some of the independent variables, then the regression model is said to contain multicollinearity between these variables. Problems can arise depending on the degree of multicollinearity that the regression model exhibits (Marquardt 1970; Kmenta 1971). When high multicollinearity is involved in a regression model, there are

some adverse effects on parameter estimates such as imprecise estimates and incorrect signs of regression coefficients.

To avoid most of the pitfalls of the OLS method in the presence of multicollinearity, biased estimation techniques such as ridge regression, principal components regression, and Stein-rule estimators have been used. Since the 1970's, much research has been conducted on obtaining biased estimators with better overall performance than OLS when multicollinearity is present (McDonald and Galarneau 1975; Gunst and Mason 1977; Dempster et al. 1977; Bare and Hann 1981).

The concerns of multicollinearity have been recently addressed in forestry. Mitchell and Hann (1979) discussed ridge regression methodology for dealing with multicollinearity and also presented an algorithm for obtaining the coefficients in ridge regression. Bare and Hann (1981) concluded, in the development of a basal area growth model for ponderosa pine, that the use of ridge regression produced precise and stable estimates of model parameters.

Past works on the biased estimation methods, especially in the field of forestry, mainly focused on mean squared error (MSE) of parameter estimates for the selection of good regression estimators. However, this study concentrated on the predictive ability of the models in selecting the "best" estimator because that is what the users (forest managers) are interested in.

In this study, biased estimation techniques for dealing with multicollinearity are presented and evaluated to select the "best" estimator in terms of predictive ability of yield models.

LITERATURE REVIEW

Among the many possible estimators of coefficients in a linear regression model, least squares estimator has been the most popular. It is an unbiased estimator of the regression parameters and has the smallest variance of all unbiased linear functions. However, the least squares estimator can be extremely unstable when there exists multicollinearity in the data. There are two major adverse effects of high multicollinearity. First, it results in the possibility of very imprecise estimates of the regression coefficients. Second, high multicollinearity can cause wrong signs of regression coefficients from what are expected (Hoerl and Kennard 1970a, 1970b; Brown and Beattie 1975). Chatterjee and Price (1977) pointed out that when a new independent variable is added or deleted, regression coefficients affected by multicollinearity are drastically changed. To obtain appropriate estimators under conditions of multicollinearity, therefore, considerable attention has been focused on biased estimation of the parameters of a linear regression model.

A number of alternatives to OLS may be preferable although they produce biased estimates. The objection to bias may not be strong depending upon the intended use of the regression models (Hocking 1976). The important issue would appear to be whether or not the resulting estimators perform better than the OLS estimation method.

Ridge Regression

Ridge regression sacrifices unbiasedness to obtain parameter estimates that have a smaller mean squared error (MSE). The ridge estimator proposed by Hoerl and Kennard (1970a) is

$$b_{RR} = (X'X + kI)^{-1}X'y \qquad (1)$$

where

b_{RR} = the ridge estimator,

X = standardized matrix of independent variables,

X' = transpose of X,

y = standardized dependent variable vector,

I = identity matrix, and

k = ridge parameter.

Since the 1970's, there has been much interest in ridge regression. The concept of ridge regression has been examined by many researchers (Marquardt 1970; Mayer and Willke 1973; McDonald and Schwing 1973; McDonald and Galarneau 1975). Much of the discussion centered around the choice of the constant k. It is recognized that the OLS estimator is unlikely to be a satisfactory estimator when the design matrix (X'X) is badly conditioned due to multicollinearity. Ridge regression can be used to remedy this problem. The important step in ridge regression is to choose a value for k such that the ridge estimator has smaller mean squared error than the OLS estimator. To improve the coefficients of the models, numerous methods have been proposed for determining the value of k. The ridge trace is one common technique proposed by Hoerl and Kennard (1970a, 1970b). The ridge trace is a plot of all

regression coefficients over a range of values for k. A k value is chosen when the regression coefficients first become stable in the ridge trace. Marquardt (1970) proposed another method based on the variance inflation factor (VIF). VIF is the diagonal elements of the inverse of the correlation matrix. Marquardt proposed a k value such that the maximum VIF of ridge estimator is between 1 to 10, and close to 1 if possible. Simulation studies have been conducted to determine the improvement of the mean squared error of estimates (McDonalds and Galarneau 1975; Dempster et al. 1977; Hoerl and Kennard 1976; Gibbons 1983).

Some other criteria have been proposed to select k when the prediction capability of the model is more important than the precision of coefficients of the models. Research on this topic has been sketchy so far. Myers (1986) summarized three general criteria to select the value of k for prediction performance of regression models. The criteria are Mallows's (1973) C_p-like statistic, Allens's (1974) PRESS-like statistic, and the generalized cross validation (GCV) proposed by Golub et al. (1979).

C_p was proposed by Mallows (1973) as a criterion for selecting a regression model. C_p is a measure of total squared error. Mallows's criterion in a ridge regression context, C_k, has been used by some researchers to select k. Erikson (1981) used ridge regression to directly estimate lagged effects in marketing and discussed the C_k statistic as one of the prediction criteria for ridge regression. Li (1986) discussed the asymptotic optimality of C_k in the setting of ridge regression.

Allen (1974) proposed PRESS (predicted residual sum of squares) as a cross validation technique for the selection of a suitable regression model. When prediction capability is an important criterion for a choice of k, a PRESS-like statistic can be used in ridge regression. This statistic is very similar to the PRESS statistic in OLS. The method consists of dropping one observation at a time, estimating the model, and predicting its left-out observation. The sum of squares of the predicted residuals is computed for each choice of k. Delaney and Chatterjee (1986), using Monte Carlo simulation technique, evaluated several methods of choosing ridge parameter k including the PRESS-like statistic. Erickson (1981) also reviewed the PRESS-like statistic and compared it with other prediction criteria.

The generalized cross validation (GCV) advocated by Golub et al. (1979) provides another criterion to choose k for improving the prediction capability of a model. This technique selects the k that minimizes a weighted mean squared prediction error. The weights are derived as a function of the design matrix. Golub et al. (1979) showed that the GCV does not require an estimate of variance. The GCV statistic has been used to choose ridge parameter k in several studies (Erikson 1981; Delaney and Chatterjee 1986; Li 1986; Bates et al. 1987).

Bare and Hann (1981) introduced ridge regression to the field of forestry, using it to select independent variables during the development of a basal area growth model for ponderosa pine. They concluded that the use of ridge regression produced a meaningful

predictive model with interpretable coefficients. However, no study so far has been done to improve the predictive capability of yield models based on data with multicollinearity problems.

Principal Components Regression

Principal components regression has received considerable attention as a method for dealing with ill-conditioned data (Massy 1965; Johnson et al. 1973; Lott 1973; Fomby and Hill 1978). Principal components regression is a method of inspecting the sample data or design matrix X'X for directions of variability and using this information to reduce the dimensionality of the estimation problem.

The occurrence of small eigenvalues of correlation matrix X'X is a warning of the presence of multicollinearity problem. Terms that have reasonably small eigenvalues of X'X are deleted to obtain principal components estimators. Thus, the principal components estimator is given by

$$\underline{b}_{PC} = \sum_{J=1}^{r} e_J^{-1} c_J \underline{v}_J \qquad (2)$$

where

$\underline{b}_{PC}$ = the principal components estimator,

e_J = Jth eigenvalue of X'X,

$\underline{v}_J$ = Jth eigenvector of X'X,

c_J = $\underline{v}_J'X'\underline{y}$, and

r = number of terms to be retained so that (p-r) terms are deleted from p parameters.

The most important thing in principal components regression is

how to determine the (p - r) terms to be deleted in order to reduce the dimensionality of the estimation problem. Judge et al. (1985) discussed two approaches on this topic. The first approach, which was somewhat arbitrary, involved deleting those components associated with small eigenvalues. The second approach was based upon tests of hypotheses using classical or MSE tests. Hill et al. (1977) provided a listing of such tests and their interpretations. Lott (1973) and Massy (1965) discussed alternative methods of selecting terms to eliminate. The methods utilized the observed values of the response variable and did not necessarily result in eliminating the terms with the smallest eigenvalues. The disadvantage is that R^2 may decrease as terms are deleted. Mansfield (1975) demonstrated that even the procedure using response variables to decide which terms to eliminate did not consistently identify the proper terms in cases of strong multicollinearity.

So far, limited studies have been conducted in the field of forestry using principal components regression. Fries (1965) used eigenvalues and eigenvectors to find the pattern of variation in stem form for different species. Kozak and Smith (1966) also used similar approaches to estimate tree taper but concluded that simpler methods were adequate. Principal components regression (Liu and Keister 1977) was used to develop equations for defining stem tapers. Newcomer and Myers (1984) also adopted principal components analysis to separate form variance from size variance for 7 tree species and to express form variance as a set of independent linear functions of the measured variables.

James-Stein Estimation

James and Stein (1961) proposed a compromise estimator for the mean of a multivariate normal distribution having a uniformly lower mean squared error than the sample mean:

$$\theta_{JSE(i)} = (1 - C)\bar{X}_i \qquad (3)$$

where

$\theta_{JSE(i)}$ = James-Stein estimator for group i,

$\bar{X}_i$ = sample average for group i,

$C = (k - 2)V / \sum_{i=1}^{n_i} X_i^2$,

n_i = number of observations in group i,

k = number of groups, and

V = common variance of groups.

There exists a risk of the estimator (3) being smaller than that of $\bar{X}_i$ for k>2 (Stein 1955; James and Stein 1961). Efron and Morris (1972a, 1972b, 1973a, 1973b, 1975) used the empirical Bayes approach to develop the James-Stein rule. Their estimator is modified by Lindley and Smith (1972) as follows:

$$\theta_{JSE(i)} = U + (1 - D)(\bar{X}_i - U) \qquad (4)$$

where

$U = \sum_{i=1}^{k} X_i / k$,

$D = (k - 3)V/S$, and

$S = \sum_{i=1}^{k} (X_i - U)^2$.

James-Stein estimators have not performed consistently well in simulation studies. In Vinod's (1978) simulation they did well in only one of three cases. Gunst and Mason's (1977) simulation showed that there is no proof that mean squared error of James-Stein

estimators is lower than that for OLS. Their results indicated that James-Stein estimators performed better than OLS when the columns of the design matrices were not close to being dependent, but were not much of an improvement for nearly collinear data.

James-Stein estimators were first used in forestry in the early 1980's by Burk and Ek (1982) to improve estimation efficiency in forest inventory problems. After comparing the estimator with maximum likelihood estimators, they concluded that the James-Stein estimator improved the precision of inventory in terms of total mean squared error. Green (1986) discussed the James-Stein estimator as an empirical Bayes estimation to update forest inventory.

For the estimation of regression coefficients, Mayer and Willke (1973) discussed the use of Stein-rule estimators of the form:

$$\underline{b}_{SR} = d * \underline{b} \qquad (5)$$

where

$\underline{b}_{SR}$ = Stein-rule estimator,

$\underline{b}$ = ordinary least squares estimator.

d = max [0, $(1 - cv/\underline{b}'\underline{b})$] for $0 < c < 2(p-2)/(h+2)$,

v = the error sum of squares using $\underline{b}$,

p = the number of eigenvalues of $X'X$, and

h = the number of degrees of freedom on which v is based.

They chose a weight d that provided smaller mean squared error than least squares estimators. James and Stein (1961) showed that mean squared error of the Stein-rule estimator (5) was minimized if $c=(p - 2)/(h + 2)$.

MATERIALS AND METHODS

Data

Data for this study came from the Southwide Loblolly Pine Seed Source Study, which was established in 1952-1953 to determine the genetic variation associated with geographic variation for loblolly pine (Wells and Wakeley 1966). Seeds from 15 geographic areas involving 9 Southern states were obtained. The seedlings from these sources were planted at each of 12 locations in a randomized complete block design with 4 replications. Because of drought after establishment, however, only 2 replications remained in 3 locations. Each replication in this study was regarded as a plot. A total of 522 plots was available. Each seed source plot contained 121 trees on a 6 ft x 6 ft spacing. The inner 49 trees on each plot were measured at 1, 3, 5, 10, 15, 20, and 25 years after planting, although the last three measurements at some locations were made at age 16, 22, and 27 instead.

Height of the 49 measurement trees on each plot were noted at time of planting, and survival was recorded the first May and June thereafter. Diameter at breast height was recorded starting at the tenth growing season.

Total cubic-foot volume outside bark per acre was computed using Burkhart et al.'s (1972) individual tree volume equation. Also, the mean height of the tallest 50 percent of surviving trees at each age was used as average height of the dominants and

codominants for each plot. This approach was employed by Golden et al. (1981) on the same data set because crown class data were not available.

The number of plots used in this study by locality and seed source is presented in Table 1. Because only data after the tenth growing season are generally available for the development of growth and yield models, data collected before age 10 were not used to estimate parameters of yield prediction models. Furthermore, remeasurements from these permanent plots formed time series data. The autocorrelation among the error terms of the time series data was detected ($p > 0.1$) by Durbin-Watson test (Neter et. al. 1985). To remove the effect of autocorrelation problems on yield prediction models, only one age class from each plot was randomly selected. This process was adopted to simulate the temporary plot data similar to those used for developing yield models.

Yield prediction data for this study were divided randomly into a fit data set and a test data set. Regression coefficients of the model were estimated from the fit data set. The test data set was used to validate the ability of the yield models to accurately predict volume yield for an independent data set. The fit data set consisted of 261 plots randomly selected from a total of 522 plots available. The remaining 261 plots were withheld to form the test data set. This half-and-half data splitting method is popular when the collection of new data is neither practical nor possible for model validation (Snee 1977). The fit and test data sets were found to be similar in stand attributes (Table 2).

Table 1. Number of plots present in the Southwide Loblolly Pine Seed Source Study Data, by locality and seed source

Locality	Seed Source Number															Total
Number	301	303	305	307	309	311	315	317	319	321	323	325	327	329	331	
	Number of plots															
03	4	4	4		4		4		4		4	4	4			36
07	6	10	6	4	6	4		4	5	4	10	6	10	4	4	84
13		4		4		4		4		4	4		4	4	4	36
15	4	4	4		4		4		4		4	4	4			36
17		4		4		4		4		4	4			4	4	32
25		4		4		4		4		4	4		4	4	4	36
26	4	4	4		4		4		4		4	4	4			36
28	2	4	2	2	2	2		2	2	2	4	2	4	2	2	34
29		4		4		4		4		4	4		4	4	4	36
32	4	8	4	4	4	4		4	4	4	8	4	8	4	4	68
36	2	2	2		2				2		2	2	2			16
40	4	8	4	4	4	4	4	4	4	4	8	4	8	4	4	72
Total	30	60	30	30	30	30	16	30	30	30	60	30	56	30	30	522

Table 2. Data summary of stand variables for the fit and test data sets

Variable [a/]	Number of obs.	Minimum	Maximum	Mean
	- - - - - - - - Fit data set - - - - - - - - -			
Age (years)	261	10	27	18
H_d	261	18	79	48
N	261	24	1185	540
V	261	123	6779	2597
	- - - - - - - - Test data set - - - - - - - - -			
Age (years)	261	10	27	18
H_d	261	14	78	49
N	261	49	1160	480
V	261	121	6751	2620

[a/] Notations:

H_d = Average height of the dominant and codominants in feet.

N = Number of trees per acre.

V = Total volume per acre in cubic-foot outside bark.

Procedure

The process of data standardization was employed before fitting the model. Standardization is merely a transformation on variables that eliminates all units of measurements and forces the standardized variables to have the same mean and the same amount of variability. The standardized variables are computed from:

$$y_i^* = (1/\sqrt{n-1})\ (y_i - \bar{y})/s_y \qquad (6)$$

$$x_{ij}^* = (1/\sqrt{n-1})\ (x_{ij} - \bar{x}_j)/s_j \qquad (7)$$

where

y_i^* = the ith observation of the standardized dependent variable,

y_i = the ith observation of the original dependent variable,

x_{ij}^* = the ith observation of the standardized jth independent variable,

x_{ij} = the ith observation of the original jth independent variable,

$\bar{y}$ = mean of the observations for the original dependent variable,

$\bar{x}_j$ = mean of the observations for the original jth independent variable,

s_y = the standard deviation of the original dependent variable,

s_j = the standard deviation of the original jth independent variable, and

n = number of observations.

Two main advantages of standardization of data are known. One is to eliminate rounding error when precision is low for computing inverse of the X'X matrix. The other is to enable regression coefficients to be more directly comparable. Parameter estimates for the original yield prediction model are given by

$$\underline{b}_J = (s_y/s_J)\ \underline{b}_J^* \quad (9)$$

$$\underline{b}_0 = y - \sum_{j=1}^{p} \underline{b}_J X_J \quad (10)$$

where

$\underline{b}_J$ = parameter estimate of the original jth independent variable,

$\underline{b}_J^*$ = parameter estimate of the standardized jth independent variable, and

$\underline{b}_0$ = parameter estimate of the intercept for the original model.

Model form for yield prediction

The model form developed by Matney et al. (1988) for yield prediction was used for this study:

$$\ln(V) = b_0 + b_1(1/A) + b_2 \ln(H_d)/A + b_3 \ln(N)/A + b_4 \ln(H_d) \quad (11)$$

where

V = total cubic-foot volume outside bark per acre,

A = total stand age in years,

H_d = average height of the dominants and codominants in feet

N = number of surviving trees per acre, and

ln(x) = natural logarithm of x.

Multicollinearity diagnostics

Multicollinearity means that the model has redundant information because of linear dependency among independent variables. In this study, four diagnostics (simple correlations among independent variables, variance inflation factors (VIFs), system of eigenvalues of X'X, and variance decomposition proportions) were used to detect the strength of the linear

dependencies and how much the variance of each regression coefficient is inflated.

Correlation is a measure of the intensity of association. In multiple regression, however, the simple correlations do not always underscore the extent of the multicollinearity problem because multicollinearity often involves associations among multiple independent variables. Even though the simple correlations do not indicate the extent of multicollinearity, they may provide guideline values to see which one-on-one associations exist (Myers 1986). The values of simple correlations among independent variables are presented in Table 3. As a general rule if the correlation coefficient between the values of two independent variables is greater than 0.8 or 0.9, then multicollinearity is a problem (Judge et al. 1988). In this study, the absolute values of correlation coefficients among independent variables ranged from 0.8385 to 0.9863, signifying a degree of multicollinearity.

The VIFs represent the inflation that each regression coefficient experiences above the ideal level if the correlation matrix were an identity matrix. They provide more a productive approach for detection than do simple correlations. They indicate which coefficients are adversely affected and to what extent. It is generally known that if VIF exceeds 10 there should be at least some concern with multicollinearity (Myers 1986). As shown in Table 4, the VIFs of variables 1/A and $\ln(H_d)/A$ were 222.1 and 120.8, respectively, indicating that a multicollinearity problem should be suspected.

Table 3. Simple correlations among independent variables used in the yield prediction model

Variable [a/]	1/A	$\ln(H_d)/A$	ln(N)/A	$\ln(H_d)$
1/A	1.0000	0.9863	0.9646	-0.9085
$\ln(H_d)/A$		1.0000	0.9472	-0.8385
ln(N)/A			1.0000	-0.8858
$\ln(H_d)$				1.0000

[a/] Notations:

A = Stand age in years.

H_d = Average height of the dominant and codominants in feet.

N = Number of trees per acre.

ln(x) = Natural logarithm of x.

Table 4. Variance inflation factor analysis for the fit data set

Variable [a/]	Variance inflation factor
1/A	222.1
$\ln(H_d)/A$	120.8
ln(N)/A	14.5
$\ln(H_d)$	18.8

[a/] Notations:

A = Stand age in years.

H_d = Average height of dominants and codominants in feet.

N = Number of trees per acre.

ln(x) = Natural logarithm of x.

Eigenvalues of the correlation matrix can also be used to detect the multicollinearity problem. A near-zero eigenvalue indicates a strong linear dependency. Multicollinearity can be measured in terms of the condition number of correlation matrix which is given by

$$\emptyset_i = \sqrt{\frac{\lambda_{max}}{\lambda_i}} \qquad (12)$$

where

$\emptyset_i$ = the condition number of the ith eigenvalue,

λ_{max} = the largest eigenvalue of the correlation matrix, and

λ_i = the ith eigenvalue of the correlation matrix.

A large condition number is evidence that the regression coefficients are unstable. When the condition number exceeds 30, multicollinearity should be suspected (Belsley et al. 1980). Table 5 shows that the smallest eigenvalue in this study had a condition number of 36.38, signifying a multicollinearity problem.

It should be emphasized that a serious multicollinearity does not deposit its effect on only one regression coefficient. The variance decomposition proportions should be analyzed to determine what proportion of the variance of each coefficient is attributed to each dependency. According to the analysis of variance proportions in this study (Table 5), the precision of estimating regression coefficients for 1/A and $\ln(H_d)/A$ was damaged by the linear dependency with high variance proportions for the smallest eigenvalue. It seems that the variable $\ln(H_d)$ does not have a lot

Table 5. Condition numbers and variance proportions for the fit data set as multicollinearity diagnostics

		Variance proportion			
Eigenvalue	Condition number	1/A	$\ln(H_d)/A$	ln(N)/A	$\ln(H_d)$
3.767380	1.0000	0.0003	0.0006	0.0046	0.0033
0.177395	4.6084	0.0006	0.0115	0.0090	0.2114
0.052379	8.4809	0.0075	0.0270	0.9475	0.0214
0.002846	36.3821	0.9916	0.9609	0.0389	0.7640

of variation. Thus, based on the analysis of variance proportions, the variables 1/A and $\ln(H_d)/A$ basically seem to be the same.

Based on the above diagnostics, some multicollinearity was detected in the data. As a result, an alternative estimation method to OLS should be recommended for the yield prediction model.

Biased estimation of parameters for yield models

In addition to ordinary least squares estimators, biased estimators such as ridge regression, Stein-rule estimator, and principal components regression were obtained to determine which estimation technique performs best in terms of the improvement of the prediction capability of the yield model.

Ridge Regression --- The performance of the ridge regression estimator depends on how well the ridge parameter k is determined. Obviously, in yield prediction models with multicollinearity, the prediction capability should be improved by using an appropriate value for k. In this study, three criteria of choosing k were Mallows's (1973) C_p-like statistic, Allens's (1974) PRESS-like statistic, and the generalized cross validation (GCV) proposed by Golub et al. (1979).

Mallows's criterion in a ridge regression context is

$$C_k = SSE_k/\hat{\sigma}^2 - n + 2 + 2\ tr(H_k) \qquad (13)$$

where

SSE_k = the sum of squared error using ridge regression,

$\hat{\sigma}^2$ = the mean squared error from OLS estimation,

n = number of observation,

H_k = hat matrix in ridge regression, which is computed by $X(X'X + kI)^{-1}X'$, and

$tr(H_k)$ = trace of the hat matrix for ridge regression.

The PRESS-like statistic, a modification of Allens's PRESS, used in this study is given by

$$PR(Ridge) = (1/n)\sum_{i=1}^{n} [\, e^2_{i,k}/(1 - h_{ii,k})^2 \,] \qquad (14)$$

where

$e_{i,k}$ = the ith residual for specific value of k, and

$h_{ii,k}$ = the ith diagonal elements of hat matrix.

On the other hand, the generalized cross validation (GCV) advocated by Golub et al. (1979) is to select a value for k that minimizes a weighted mean squared error prediction. The criterion is given by

$$GCV = SSE_k/(n - [1 + tr(H_k)])^2 \qquad (15)$$

Most ridge regression is applied to the standardized form of the model. The ridge estimator (1) in standardized form is given by

$$\underline{b}^*_{RR} = (R_{xx} + kI)^{-1} r_{xy} \qquad (16)$$

where

R_{xx} = the correlation matrix of independent variables, and

r_{xy} = the vector of simple correlation of the independent variables and the dependent variable.

For different values of k from 0 to 1, the three criteria C_k statistic, PR(Ridge), and GCV were computed using the standardized form of the data. A value of k which minimized the statistic was chosen for each criterion. The parameters of the yield prediction model were then estimated from equation (16), resulting in three yield equations.

Principal components regression --- The reduction of the dimension of the estimation problem implies a trade-off that balances bias against reduced sampling variances. These considerations are particularly important in the case of principal components regression. The matrix form of a linear regression model can be transformed as follows:

$$\begin{aligned} \underline{y} &= X\underline{\beta} + \underline{e} \\ \underline{y} &= XPP'\underline{\beta} + \underline{e} \\ \underline{y} &= Z\underline{\theta} + \underline{e} \end{aligned} \tag{17}$$

where

$\underline{y}$ = vector of dependent variable from standardized data,

X = matrix of independent variables from standardized data,

P = the orthogonal eigenvectors of $X'X$,

$\underline{\beta}$ = the unknown parameters to be estimated,

$\underline{e}$ = vector of errors distributed as $N(0, \sigma^2 I)$,

$\underline{\theta} = P'$, and

$Z = XP$ which is the matrix of principal components.

The principal components estimator of $\underline{\beta}$ is obtained by deleting one or more of the principal components, applying OLS to the resulting model and making a transformation back to the original parameter space. The matrix Z can be partitioned into two parts, Z_1 to be retained and Z_2 to be deleted. Thus the model (17) can be rewritten as

$$\underline{y} = Z_1\underline{\theta}_1 + Z_2\underline{\theta}_2 + \underline{e} \tag{18}$$

When $\underline{\theta}_2$ is set equal to zero, the least squares estimator of $\underline{\theta}_1$ is easily computed from $\hat{\underline{\theta}}_1 = (Z_1'Z_1)^{-1}Z_1'\underline{y}$. The principal components

estimator is obtained from an inverse linear transformation:

$$\underline{b}_{PC} = P_1\hat{\underline{\theta}}_1 \qquad (19)$$

The major question in the principal components regression is how to select components for deletion. The simplest way is to select Z_2 associated with small eigenvalues. Based on the eigenvalue analysis, in this study, one eigenvalue had the condition number of 36.38 (Table 5). Thus, one principal component corresponding to the small eigenvalue was deleted and the principal components estimator (19) was computed.

Stein-rule Estimator --- Vinod and Ullah (1981) discussed a Stein-rule estimator in the regression context. The estimator is derived by the Bayesian interpretation with a prior distribution of $\underline{\beta} \sim N(0, \sigma_\beta^2(X'X)^{-1})$, where σ_β^2 is the variance of $\underline{\beta}$. In other words, the Stein-rule estimator is a compromise between the sample information and prior information of the estimates. Since the prior information $\underline{\beta}$ has mean zero, the estimator shrinks the OLS estimator toward the origin. As a result, the Stein-rule estimator is given by

$$\underline{b}_{SR} = [1 - (ps^2/\underline{b}'X'X\underline{b})]\,\underline{b} \qquad (20)$$

where

$\underline{b}_{SR}$ = Stein-rule estimator,

$\underline{b}$ = ordinary least squares estimator,

p = the number of eigenvalues of $X'X$,

s^2 = the mean squared error from the OLS, and

X = matrix of independent variables from the standardized data.

The Stein-rule estimator weights the OLS estimator by the

factor $0 \leq 1 - (ps^2/\underline{b}'X'X\underline{b}) \leq 1$. The yield prediction model was fitted to estimate coefficients using the estimator (20). The resulting equation should have a lower MSE over the OLS equation.

Evaluation criteria

Parameter estimates of the yield prediction model were obtained from the fit data set using each of the biased estimation methods. In addition, the OLS technique was employed to estimate the parameters of the model. Thus, six final equations were evaluated to determine which method provided the "best" results in terms of prediction performance of the model under the multicollinearity situation.

To evaluate the estimation methods, candidate estimators were compared based on the following three evaluation criteria.

1. Mean difference, which is a measure of bias of a model.

$$\overline{Diff} = (1/n)\sum_{i=1}^{n} Diff_i$$

where

$Diff_i = y_i - \hat{y}_i$ = difference between the ith observed and predicted volume per acre, and

n = the number of observations.

2. Mean absolute difference, which is a measure of precision of a model.

$$\overline{|Diff|} = (1/n)\sum_{i=1}^{n} |Diff_i|$$

3. Mean squared difference, which is similar to the mean absolute difference, but is more more sensitive to outliers.

$$\overline{Diff^2} = (1/n)\sum_{i=1}^{n} (Diff_i)^2$$

These statistics were computed separately for the test data and the pooled data (both fit and test data sets). The test data represented an independent data set, whereas the pooled data were regarded as the representative of the population.

The evaluation criteria were computed based on volume per acre rather than the logarithm of volume which was the dependent variable in the yield model. This was because volume per acre was really the variable of interest.

The final six equations were ranked relative to one another based on each criterion, with rank 1 corresponding to the smallest value. Then the overall rank was calculated as the sum of the ranks over three criteria. The "best" system of yield prediction equation was the one with the smallest overall rank.

RESULTS AND DISCUSSION

The k values from prediction-oriented selection criteria ranged from 0.00012 to 0.00065. The minimum values for C_k, the PRESS-like statistic, and generalized cross validation were obtained when k was 0.00013, 0.00065, and 0.00012, respectively. These k values were conservative (close to zero). Hocking (1976) reported that for his data, C_k statistic was more conservative in producing a smaller k value than the ridge trace and VIF criteria. In this study, the PRESS-like criterion produced the least conservative (largest k) biased estimation of the coefficients, whereas C_k and GCV resulted in similar values for k.

Six sets of coefficients of the yield prediction model (11) were obtained from the fit data set (Table 6). The six estimation methods were OLS, three ridge estimators based on different criteria of choosing k, principal components regression, and Stein-rule estimator. The results of evaluation on the test data set and the pooled data set are presented in Table 7.

For both validation data sets, three ridge regression methods performed slightly better than the OLS. Especially, ridge estimator based on the C_k statistic performed better than the OLS for all evaluation statistics. The principal components estimator had the smallest mean difference and the largest mean absolute and squared difference for both data sets. Thus, this estimator was not only the least biased but also the least precise for yield

Table 6. Parameter estimates of the yield prediction model from six different estimation methods

Estimator [a/]	Parameter Estimates				
	b_0	b_1	b_2	b_3	b_4
OLS	0.2038	-61.7273	3.0816	8.7655	1.8884
J-S	0.2186	-61.6054	3.0755	8.7482	1.8847
C_k	0.0998	-59.9155	2.6386	8.7320	1.9162
PR(Ridge)	-0.2487	-53.9804	1.2856	8.6080	2.0048
GCV	0.1035	-60.0494	2.6873	8.7345	1.9142
PC	-2.5309	-21.3997	-7.1846	8.4873	2.5945

[a/] Notation:

OLS = Ordinary least squares estimator,

J-S = James-Stein estimator,

C_k = Ridge estimator based on Mallows's (1973) statistic (k = 0.00013),

PR(Ridge) = Ridge estimator based on Allens's (1972) PRESS-like statistic (k = 0.00065),

GCV = Ridge estimator based on the generalized cross validation (k = 0.00012), and

PC = Principal component estimator.

Table 7. Evaluation statistics from six estimation methods for the test data set and the pooled data set

Estimator	Test data set			Pooled data set		
	$\overline{Diff}$ [a/]	$\overline{\lvert Diff \rvert}$ [b/]	$\overline{Diff^2}$ [c/]	$\overline{Diff}$	$\overline{\lvert Diff \rvert}$	$\overline{Diff^2}$
OLS	118.92	395.58	330725	182.40	422.36	365786
J-S	120.11	396.32	332054	183.53	423.07	367181
C_k	117.97	394.40	329112	180.76	421.61	363798
PR(Ridge)	118.80	395.16	330482	180.24	423.68	365003
GCV	118.94	395.33	330436	181.98	422.62	365417
PC	101.66	405.09	335177	157.51	435.73	369511

a/ $\overline{Diff} = (1/n)\sum_{i=1}^{n} Diff_i$, where $Diff_i = y_i - \hat{y}_i$ = difference between the ith observed and predicted volume per acre.

b/ $\overline{|Diff|} = (1/n)\sum_{i=1}^{n} |Diff_i|$.

c/ $\overline{Diff^2} = (1/n)\sum_{i=1}^{n} (Diff_i)^2$.

prediction. On the other hand, the Stein-rule estimator consistently performed worse than the OLS estimator for all evaluation statistics in both validation data sets.

The ranks based on the three criteria are presented in Table 8. The overall ranks were similar for both the test data set and the pooled data set, indicating that each estimator performed consistently for an independent data set as well as for the population.

Ridge estimators performed slightly better than OLS estimators. The C_k criterion produced the best improvement in terms of prediction capability of the model. The PRESS-like and GCV criteria also provided some improvement of prediction over the OLS and ranked second and third, respectively, in both validation data sets (Table 8). However, the ridge estimators gained 1 to 2 cubic feet per acre in mean difference and mean absolute difference for both validation data sets. This amount of improvement by ridge estimators over OLS estimators may not be meaningful in practical applications. These results were similar to those obtained by Delaney and Chatterjee (1986), who compared ridge estimators to OLS estimator through Monte Carlo simulations. They concluded that, for the predictive ability, the OLS estimator performed as well as the ridge estimator from PRESS-like statistic and even better than the ridge estimator from GCV.

This study showed that the use of OLS estimators might be preferable for the predictive ability of the model when the data have a multicollinearity problem. Judge et al. (1988) discussed a

Table 8. Ranks[a/] of evaluation statistics from six estimation methos for the test data set and the pooled data set

Estimator	Test data set: $\overline{Diff}$	Test data set: $\overline{\lvert Diff \rvert}$	Test data set: $\overline{Diff^2}$	Test data set: Total	Polled data set: $\overline{Diff}$	Polled data set: $\overline{\lvert Diff \rvert}$	Polled data set: $\overline{Diff^2}$	Polled data set: Total	Rank sum	Overall rank
OLS	4	4	4	12	5	2	4	11	23	4
J-S	6	5	5	16	6	4	5	15	31	6
C_k	2	1	1	4	3	1	1	5	9	1
PR(Ridge)	3	2	3	8	2	5	2	9	17	2
GCV	5	3	2	10	4	3	3	10	20	3
PC	1	6	6	13	1	6	6	13	26	5

[a/] Numbers to represent relative performances of six estimation methods (1 being the best and 6 being the worst). The overall ranks were determined by the sum of the ranks over three evaluation statistics.

near-exact multicollinearity situation in which the ill effects of small eigenvalues were cancelled out, resulting in good predictions from the OLS estimator.

The principal components estimator ranked fifth overall, below the OLS estimator (Table 8). It ranked first in terms of mean difference for both validation data sets, but ranked last in the other two criteria (mean absolute difference, and mean squared difference). Residual plots revealed that principal components estimator produced residuals which were more symmetrical about the zero line as compared with OLS estimators. In other words, the principal components estimator provided more systematical overprediction and underprediction than the OLS estimator. As a result, bias based on mean difference was lower for principal components estimators.

The better overall performance of the OLS estimator over the principal components estimator was not expected and might be due to the data structure. Kozak and Smith (1966) suggested that OLS estimators were adequate for estimating tree taper rather than principal components regression methods.

The Stein-rule estimator consistently performed poorly for both data sets in this study (Table 7). The Stein-rule estimator ranked last overall for both data sets (Table 8). This indicates that James-Stein estimator did not improve prediction in this study when multicollinearity was involved. It is known that Stein-rule estimator is better than OLS in terms of lower mean squared error (MSE) of estimates, provided there are at least three parameters to

be estimated. However, this estimator was not frequently used in the 1960's and early 1970's, despite its theoretical superiority (Vinod and Ullah 1981). The lack of faith on this estimator was the main reason this estimator was not frequently used. Researchers were not sure whether or not their data from practical problems could meet the assumptions such as normal distribution and independence of errors in order to use this estimator. Moreover, some simulation studies failed to prove the superiority of this estimator over the OLS in terms of mean squared error (Gunst and Mason 1977; Vinod 1978). Draper and Van Nostrard (1979) suggested that Stein-rule estimator did not produce much of an improvement for nearly collinear data.

SUMMARY AND CONCLUSION

This study was conducted to select the "best" estimation method of linear regression yield models with multicollinearity. Attention has been focused on biased estimation techniques for dealing with multicollinearity. Several biased estimators were compared to select the best estimator in terms of predictive ability of yield models with the OLS estimator.

Based on three evaluation statistics, ridge estimators were slightly better than the OLS in their performances. However, care should be focused on the method of choosing ridge parameter k. In this case, the choice of k in ridge regression should be restricted to prediction-oriented selection
criteria such as C_k, PRESS-like, and GCV statistics.

Ridge estimator with k based on the C_k statistic was the "best" in terms of the predictive ability. The Stein-rule and principal component estimators did not perform as well as OLS estimators in prediction problems for the data used in this study.

Even though ridge estimators performed well in this study, the gain in yield prediction was small. OLS might be used safely in estimating parameters of yield equations even though multicollinearity problems exist.

STUDY II
CALIBRATION OF YIELD PREDICTION MODELS FOR A SPECIFIC LOCALITY AND A SPECIFIC SEED SOURCE

ABSTRACT

A Stein-rule estimator was employed to calibrate yield prediction models to a specific locality and a specific seed source. Data from 12 localities and 15 seed sources from the Southwide Loblolly Pine Seed Source Study were used in this study. The yield model form developed by Burkhart et al. (1972) was used. Three approaches for parameter estimates of the yield prediction model were evaluated: the ordinary least squares (OLS) for the entire region, the OLS for a specific subregion, and a Stein-rule estimator which is a compromise of the previous two approaches.

The Stein-rule estimator provided more precise yield prediction than the two OLS estimators for calibrating the model to a specific locality and to a specific seed source. The gain in terms of the predictive ability by the Stein-rule estimator was not as pronounced for seed sources as for localities.

INTRODUCTION

Mathematical models for yield prediction have been fitted to data from wide geographical areas and broad ranges of site using the ordinary least squares (OLS) estimation technique. In many cases, however, forest managers are interested in yield prediction for specific subregions such as counties and stands. The yield models do not necessarily provide precise prediction for specific applications of small subregions (Smith 1983). The main reason for poor performance is that regional yield models do not fully account for the variation in site quality, climatic conditions, drainage pattern, and genotypic characteristics of a specific forest area (Gertner 1984). For regional estimates these unexplained factors are usually averaged out, but for subregional estimates, this may not be the case. Thus, the resulting estimates may have large variances.

On the other hand, the OLS estimation fitting to sample data for the subregion may not result in good yield prediction due to the small size of the data set. Therefore, it is frequently desirable to adjust the regional parameters to different subregions using sample data from the subregions. An alternative to OLS should be adopted which uses all available information on the subregion and therefore improves the predictive ability of yield models.

Stein-rule estimators can be used in calibrating models for this purpose. These estimators incorporate prior information (which

is previous knowledge about the parameters) from the entire region with sample information from the subregion to provide precise yield prediction for the subregion of interest. When forest managers have many localities of interest and want to precisely estimate regression coefficients of yield models for each of the localities, a Stein-rule estimator can be computed by combining information from all localities and from that specific locality.

Similarly, forest geneticists may be interested in yield prediction equations for different seed sources in order to reveal the genetic effects on volume yield. In this case, the Stein-rule estimator can also be employed to improve the predictive ability of yield models for a specific seed source, which is considered a subregion.

The objective of this study is to calibrate yield prediction models for a specific locality and a specific seed source by using a Stein-rule estimation method in order to improve the predictive ability of yield models.

LITERATURE REVIEW

The following literature review is focused on the calibration efforts of regression models in forestry and on the applications of Stein-rule estimators.

Calibrating Regression Models in Forestry

Calibration methods are used to adjust the parameters of a regional model to a subregion of interest. Calibration techniques use all available information on the subregion and possibly provide more precise parameter estimates for the subregional model. Thus, in the field of forestry, some calibration techniques for regression models have been used. Stage (1981), in his forest growth projection system (PROGNOSIS), employed a regression revision procedure for localizing an individual tree diameter increment model. The model was localized by revising only the intercept term, while the other parameters were kept constant.

Smith (1983) used an annual adjustment factor to localize estimates of annual diameter growth for individual trees provided by STEMS, a regional growth projection system. The annual adjustment factor is simply the ratio of the mean observed diameter growth from the subregion to the mean regional predicted diameter growth.

A sequential Bayesian procedure was adopted by Gertner (1984) to localize a nonlinear diameter increment model. The regional parameters of the model were sequentially adjusted for each time

period, using information from previous periods. Unlike linear models, the nonlinear model required an iterative procedure for parameter adjustment.

Stein-Rule Estimation

James and Stein (1961) proposed a biased estimation technique. The estimator is

$$\hat{\theta}_{iJS} = (1 - (k-2)\sigma^2 / \sum_{i=1}^{k} \bar{y}_i^2)\ \bar{y}_i \qquad (1)$$

where

$\hat{\theta}_{iJS}$ = James-Stein estimator for group i,

y_i = sample average for group i,

σ^2 = common variance for groups, and

k = number of groups.

Since that time, efforts for the application of this estimator have been made, mainly by Efron and Morris (1972a, 1972b, 1973a, 1973b, 1975). They used the empirical Bayes approach to develop a Stein-rule estimator. The development provides more useful information for both identifying appropriate applications and for generalizing and extending the James-Stein results.

Morris (1977) used formal Bayesian ideas given by Baranchik (1970) to derive an estimator very similar to that of James and Stein (1961). The estimator is minimax and admissible for the equal variance case. Morris (1977) derived estimators for both the equal and unequal variance among groups.

Literature dealing with applications of the Stein-rule estimators are limited. Carter and Rolph (1974) applied a procedure

very similar to the James-Stein estimator to estimation of fire alarm probabilities. Stein-rule estimator was used by Fay and Herriot (1979) with census data to improve income estimates for small communities. Looney and Brock (1979) used Stein-rule estimator to improve small area estimates based on the data from National Center for health statistics.

In forestry literatures, Stein-rule estimators have mainly been employed to estimate forest inventory. A Stein-rule estimator was first used in forestry in the early 1980's by Burk and Ek (1982). They applied Stein procedures to simultaneous estimation problems in forest inventory. Green (1986) discussed the James-Stein estimator as an empirical Bayes estimation to update forest inventory. More recently, Green et al. (1987) estimated volume harvested per acre in softwoods and hardwoods by county in Louisiana, using a Stein-rule estimator.

Stein-rule estimators have been also used in the context of regression models (Mayer and Willke 1973; Gunst and Mason 1977; Vinod and Ullah 1981). Vinod and Ullah (1981) discussed Stein-rule estimators and derived a Stein-rule estimator using Bayesian interpretation for regression models. Efron and Morris (1972b) showed that an empirical Bayes approach can be used to derive a Stein-rule estimator of linear regression models. Lindley and Smith (1972) modified the James-Stein estimator to obtain a Stein-rule estimator:

$$\underline{b}_L = \underline{b} + (1 - A/(A + V))\ (\underline{b}_i - \underline{b}) \qquad (2)$$

where

$\underline{b}_L$ = Lindley's Stein-rule estimator,

$\underline{b}$ = parameter estimates obtained by OLS for the entire region,

$\underline{b}_i$ = parameter estimates obtained by OLS for the ith subregion,

A = common variance of sample data for that subregion, and

V = variance of parameter estimates for the entire region.

In the field of forestry, the application of Stein-rule estimators for regression models is limited. Green and Strawderman (1986) used a Stein-rule estimator to simultaneously estimate coefficients in 18 eastern hardwood volume equations. They concluded through simulation that the Stein-rule estimator was biased, but produced better predictions than the least squares estimates.

MATERIALS AND METHODS

Data

Data used in this study also came from the Southwide Loblolly Pine Seed Source study (described in study I). The data set consists of 12 localities and 15 seed sources.

Similar to the biased estimation study, only data for age 10 and thereafter were used to estimate parameters of yield models for different localities and seed sources. Also, one age class from each plot was randomly selected in order to remove the effect of autocorrelation due to remeasurements of each plot.

For each of 12 localities, data were randomly divided into a fit data set and a test data set using a half-and -half data splitting method. The fit data set (248 plots) was used to estimate coefficients of the yield model. The remaining 249 plots which formed a test data set were withheld to validate the prediction capability of the model. The pooled data (fit and test data sets) were also used to validate the model's prediction capability for the population. The stand attributes for the fit and test data sets are presented by locality in Table 1 and by seed source in Table 2.

Total cubic-foot volume outside bark per acre was computed by using Burkhart et al.'s (1972) individual tree volume equation. Average height of dominants and codominants for each plot was computed by the mean height of the tallest 50 percent of surviving trees at each age.

Table 1. Stand attributes for the fit and test data sets, by locality

Locality Number	Location	Volume [a/]	Age [b/]	H_d [c/]	TPA [d/]	Number of Plots	Volume	Age	H_d	TPA	Number of Plots
		- - - Fit Data Set - - -					- - - Test Data Set - - -				
03	Worcester County, MD	2659	18.4	46	621	18	2467	17.5	43	672	18
07	Craven County, NC	2349	18.2	52	440	41	2489	18.2	53	457	41
13	Newberry County, SC	3076	19.7	48	739	18	2745	15.3	40	928	18
15	Dooly County, NC	2300	16.4	38	815	18	2160	16.1	38	834	18
17	Spalding County, GA	3403	17.8	47	561	16	3498	19.3	51	493	15
25	Coosa County, AL	1777	17.9	45	280	18	1833	17.5	48	297	18
26	Talladega County, AL	3051	16.1	48	784	9	3567	16.4	52	779	11
28	Pearl River County, MS	2357	17.4	39	986	17	1813	15.2	33	1105	17
29	Winston County, MS	1638	17.8	49	200	18	1948	18.4	51	216	17
32	Washington Parish, LA	4354	19.6	58	611	34	4657	17.7	56	743	34

Table 1. (Continued).

Locality Number	Location	Volume	Age	H_d	TPA	Number of Plots	Volume	Age	H_d	TPA	Number of Plots
		- - - Fit Data Set - - -					- - - Test Data Set - - -				
36	Cherokee County, TX	1845	14.2	45	275	6	1150	13.3	40	337	6
40	Clark County, AR	4955	16.9	52	566	35	4458	18.8	56	480	36

a/ Total volume per acre outside bark in cubic-foot.

b/ Stand age in years.

c/ Average height of dominants and codominants in feet.

d/ Number of trees per acre survived.

Table 2. Stand attributes for the fit and test data sets, by seed source

Seed Source Number	Reigon	Volume	Age	H_d	TPA	Number of plots	Volume	Age	H_d	TPA	Number of plots
		- - - Fit Data Set - - -					- - - Test Data Set - - -				
301	Eastern MD	2409	18.1	49	591	14	3082	18.8	54	568	13
303	Southeastern NC	3085	19.4	56	437	30	2863	17.6	50	538	28
305	Eastern NC	2122	15.9	46	515	15	2899	17.3	50	569	14
307	Western SC	2372	16.8	48	493	15	2799	19.9	54	378	15
309	Southwestern GA	2271	16.4	46	513	14	2775	16.6	49	583	14
311	Northwestern GA	2416	17.6	46	436	15	2396	18.1	49	449	15
315	Northern AL (Cullman)	3133	16.4	47	624	7	3329	19.6	48	726	7
317	Northeastern AL	1552	15.9	43	424	15	2261	15.5	43	497	15
319	Northern AL (Jefferson)	2682	18.1	51	498	12	2724	19.2	50	531	15
321	Northeastern MS	2491	19.3	52	388	15	2754	17.6	49	536	14
323	Southeastern LA	3034	19.5	56	485	28	2473	16.7	48	533	29

Table 2. (Continued).

Seed Source Number	Reigon	Volume	Age	H_d	TPA	Number of plots	Volume	Age	H_d	TPA	Number of plots
		- - - Fit Data Set - - -					- - - Test Data Set - - -				
325	Eastern TX	4390	19.0	45	638	13	2570	14.5	44	640	15
327	Southwestern AR	2507	17.0	45	596	28	2497	16.6	42	710	25
329	Western TN	1862	15.3	42	502	15	2518	19.0	49	454	15
331	Northwestern GA	3256	20.5	55	459	15	2602	18.5	50	406	15

Procedure

Let us assume that data are available for the entire region, but our interest is focused on a specific subregion. There exist three possible approaches for parameter estimates of yield prediction models.

The first approach involved the use of OLS technique to estimate regression coefficients of yield models using data from the entire region. This approach had an advantage of utilizing all data. However, the wide range of data from the entire region resulted in predictions not specific enough for the subregion. In the second approach, the OLS technique was employed to estimate regression coefficients of yield models using only data from the subregion of interest. This approach concentrated on that specific subregion at the expense of losing information from other subregions. A Stein-rule estimator was used in the last approach to combine sample information from the subregion and prior information from the entire region. This approach provides a compromise of the previous two approaches.

Furthermore, two scenarios were considered in this study. In the first scenario, each of the 12 localities was considered as a subregion. The second scenario assumed that each of the 15 seed sources was a subregion.

Model forms for yield prediction

In this study, a yield prediction model developed by Burkhart et al. (1972) was used. The model form is given by

$$\log(V) = b_0 + b_1(1/A) + b_2(H_d/A) + b_3(N/100) + b_4(A)[\log(N)] \quad (3)$$

where

V = total cubic-foot volume outside bark per acre,

A = stand age in years,

H_d = average height of the dominants and codominants in feet.

N = number of surviving trees per acre, and

log(x) = logarithm (base 10) of x.

Stein-rule estimator for calibrating yield prediction models

Stein-rule estimators can be considered as the weighted average of least squares estimators from the subregion and from the entire region. Vinod and Ullah (1981) discussed a Stein-rule estimator in the regression context based on Lindley and Smith's (1972) approach. The estimator is given by

$$\underline{b}_s = \underline{b} + [\,1 - ps^2/\{(\underline{b}_i - \underline{b})'X'X(\underline{b}_i - \underline{b})\}](\underline{b}_i - \underline{b}) \quad (4)$$

where

$\underline{b}_s$ = Stein-rule estimator,

$\underline{b}$ = ordinary least squares estimates obtained from fitting yield model over the entire region,

$\underline{b}_i$ = ordinary least squares estimates obtained from fitting yield models over the ith subregion,

p = the number of independent variables in yield model,

s^2 = the estimated mean squared error obtained from the data of the ith sample data, and

X = matrix of independent variables for sample data.

Since this estimator was derived using Bayesian interpretation,

several underlying assumptions related to Bayes theory needed to be made. The yield prediction model can be expressed in general with the following mathematical notation:

$$\underline{y} = X\underline{\beta} + \underline{e} \qquad (5)$$

where

$\underline{y}$ = vector of dependent variable in the yield prediction model,

$\underline{\beta}$ = parameters to be estimated in the yield prediction model, and

$\underline{e}$ = vector of errors in the yield model.

One assumption for this model is that the dependent variable $\underline{y}$ is normally distributed with mean $X\underline{\beta}$ and variance $\sigma^2 I$. Another assumption is that the parameter $\underline{\beta}$ of the yield prediction model for the entire region is also normally distributed with mean $\underline{b}$ and covariance $\sigma^2_{\underline{\beta}}(X'X)^{-1}$, where $\sigma^2_{\underline{\beta}}$ is the variance of $\underline{\beta}$.

The Stein-rule estimator (4) was employed to calibrate the yield prediction model (3) for a specific subregion.

Evaluation criteria

The three approaches in each of the two scenarios mentioned above were evaluated based on three evaluation criteria, which included mean difference ($\overline{Diff}$), mean absolute difference ($\overline{|Diff|}$), and mean squared difference ($\overline{Diff^2}$). These criteria defined in study I.

These three statistics were computed separately for the test data and the pooled data (both fit and test data sets). The test data represented an independent data set, whereas the pooled data represented the population.

A ranking method was then adopted to evaluate the performances of the three approaches. A rank of one to three (one being best) was given to each criterion. The overall rank was computed as the sum of the ranks for all subregions (localities or seed sources) separately for the test data and pooled data sets. Finally, the sum of overall ranks for test data and pooled data sets was used to decide the "best" estimation method for calibration of yield prediction models.

RESULTS AND DISCUSSION

For each of the 12 localities and 15 seed sources, the performances of the Stein-rule and OLS estimators in terms of predictive ability were evaluated. Results and discussion for each scenario are give separately as follows.

Calibration of yield models to a locality

Ordinary least squares estimates of regression coefficients of the yield model (3) were obtained for all localities, using the fit data set. Stein-rule estimator was then computed for each locality (Table 3). The resulting yield prediction equations of 12 localities were evaluated. The three statistics and their corresponding ranks were found for both the test and pooled data sets (Table 4). Overall ranks for the estimators were then determined based on the ranks of all localities (Table 5).

As expected, Stein-rule estimators performed consistently well on both data sets and ranked first overall. Out of 12 localities, the Stein-rule estimators performed better than the two OLS estimators in 8 and 9 localities for the test data set and pooled data set, respectively. For the rest of localities, they ranked second on both data sets.

Based on the mean difference (Diff), which represented a measure of bias of the model, Stein-rule estimator provided less bias than OLS estimators in both validation data sets. This result

Table 3. Parameter estimates of Burkhart et al. (1972) 's model for twelve localities in the fit data set, using OLS and Stein-rule estimators

Locality Number	Estimators	b_0	b_1	b_2	b_3	b_4
All	OLS[a/]All	1.0421	2.1140	0.3418	0.0273	0.0232
03	OLS[b/] [c/]	2.1857	-5.0778	0.2595	0.0444	0.0121
	Stein	2.0843	-4.4401	0.2668	0.0249	0.0131
07	OLS	1.5950	-5.0524	0.3881	0.0532	0.0148
	Stein	1.5187	-4.0627	0.3817	0.0496	0.0160
13	OLS	1.3101	-5.3441	0.4991	0.0490	0.0150
	Stein	1.3053	-5.2109	0.4963	0.0486	0.0152
15	OLS	2.6563	-6.9597	0.2745	0.0222	0.0065
	Stein	2.6069	-6.6823	0.2766	0.0224	0.0070
17	OLS	3.9555	-15.2941	0.1058	0.0541	-0.0024
	Stein	3.8441	-14.6283	0.1148	0.0531	-0.0014
25	OLS	1.6598	4.9101	-0.0092	0.0281	0.0260
	Stein	1.5469	4.3992	0.0549	0.0280	0.0255
26	OLS	1.5741	0.2206	0.2780	0.0233	0.0178
	Stein	1.4611	0.6230	0.2916	0.0242	0.0190
28	OLS	0.7869	4.5444	0.3480	0.0184	0.0250
	Stein	0.8111	4.3138	0.3474	0.0192	0.0248
29	OLS	1.2517	-1.0563	0.3178	0.0649	0.0236
	Stein	1.2071	-0.3805	0.3229	0.0569	0.0235
32	OLS	1.0666	-0.8278	0.4588	0.0265	0.0192
	Stein	1.0623	-0.3188	0.4386	0.0266	0.0199

Table 3. (Continued).

Locality Number	Estimators	b_0	b_1	b_2	b_3	b_4
36	OLS	2.7394	-19.6690	0.6254	0.1058	-0.0127
	Stein	1.8580	-8.3564	0.4781	0.0651	0.0060
40	OLS	1.7632	-0.7301	0.2054	0.0341	0.0216
	Stein	1.6058	-0.1092	0.2352	0.0326	0.0220

a/ Ordinary least squares estimates for all twelve localities.

b/ Ordinary least squares estimates for that locality.

c/ Stein-rule estimates for that locality.

Table 4. Evaluation statistics for three estimation methods, by criterion and locality.

Locality Number	Estimator	Test Data Set: $\overline{Diff}$ [a/]	Test Data Set: $\overline{\lvert Diff \rvert}$ [b/]	Test Data Set: $\overline{Diff^2}$ [c/]	Pooled data set: $\overline{Diff}$	Pooled data set: $\overline{\lvert Diff \rvert}$	Pooled data set: $\overline{Diff^2}$	Rank Sum
03	OLSAll	138.33 (3)	408.95 (3)	273400 (3)	131.82 (3)	388.91 (3)	238053 (3)	19
	OLS	-7.57 (2)	190.79 (1)	65868 (1)	-2.15 (1)	188.79 (1)	60649 (1)	7
	Stein	6.94 (1)	209.08 (2)	77141 (2)	11.57 (2)	199.02 (2)	67213 (2)	11
07	OLSAll	21.56 (1)	404.38 (3)	246598 (3)	-61.03 (2)	342.82 (3)	180894 (3)	15
	OLS	127.09 (3)	386.33 (2)	231703 (2)	73.03 (3)	298.44 (2)	151382 (2)	14
	Stein	113.75 (2)	383.41 (1)	227475 (1)	55.80 (1)	296.66 (1)	148615 (1)	7
13	OLSAll	-263.73 (3)	393.49 (3)	340570 (3)	-315.32 (3)	472.11 (3)	420857 (3)	18
	OLS	73.16 (2)	179.41 (2)	45812 (2)	36.49 (2)	158.04 (2)	36824 (2)	12
	Stein	67.66 (1)	179.23 (1)	45777 (1)	31.05 (1)	157.84 (1)	36536 (1)	6
15	OLSAll	-129.18 (3)	178.54 (1)	112007 (3)	-151.23 (3)	260.13 (3)	181894 (3)	16
	OLS	-55.03 (1)	184.88 (3)	45653 (2)	-25.26 (1)	130.59 (2)	33167 (2)	11
	Stein	-56.72 (2)	181.84 (2)	43528 (1)	-28.36 (2)	129.23 (1)	31920 (1)	9
17	OLSAll	370.82 (3)	593.84 (1)	575946 (2)	531.13 (3)	747.15 (3)	896183 (3)	15
	OLS	-227.82 (2)	638.44 (3)	614763 (3)	-103.85 (2)	460.51 (2)	407907 (2)	14
	Stein	-198.41 (1)	618.74 (2)	571623 (1)	-73.35 (1)	450.44 (1)	389206 (1)	7
25	OLSAll	75.08 (1)	227.17 (1)	100802 (1)	125.41 (1)	322.54 (1)	198869 (1)	6
	OLS	406.41 (3)	626.45 (3)	681389 (3)	245.80 (3)	638.42 (3)	752711 (3)	18
	Stein	358.22 (2)	562.28 (2)	541566 (2)	236.33 (2)	585.02 (2)	607689 (2)	12

[d/]

Table 4. (Continued).

Locality Number	Estimator	Test Data Set $\overline{Diff}$	Test Data Set $\overline{\lvert Diff \rvert}$	Test Data Set $\overline{Diff^2}$	Pooled data set $\overline{Diff}$	Pooled data set $\overline{\lvert Diff \rvert}$	Pooled data set $\overline{Diff^2}$	Rank Sum
26	OLSAll	-592.66 (3)	592.66 (3)	699595 (3)	-647.30 (3)	647.30 (3)	798822 (3)	18
	OLS	228.60 (2)	313.68 (2)	248019 (2)	132.83 (2)	289.01 (1)	197424 (2)	11
	Stein	69.99 (1)	298.03 (1)	174291 (1)	-18.10 (1)	304.53 (2)	167860 (1)	7
28	OLSAll	-454.90 (3)	454.90 (3)	416210 (3)	-522.24 (3)	539.86 (3)	695191 (3)	18
	OLS	39.05 (2)	155.34 (2)	48594 (2)	17.72 (1)	168.62 (2)	62277 (1)	10
	Stein	-1.98 (1)	132.95 (1)	39254 (1)	-27.44 (2)	163.90 (1)	66107 (2)	8
29	OLSAll	272.95 (3)	366.55 (3)	227317 (3)	252.25 (3)	337.96 (3)	201415 (3)	18
	OLS	-143.56 (2)	295.14 (2)	219932 (2)	-70.56 (2)	253.73 (2)	153660 (2)	12
	Stein	-44.49 (1)	250.45 (1)	150602 (1)	6.00 (1)	226.16 (1)	112592 (1)	6
32	OLSAll	-54.45 (1)	412.32 (1)	329939 (1)	-59.75 (3)	515.81 (3)	523989 (3)	12
	OLS	-90.08 (3)	513.22 (3)	399306 (3)	-32.36 (1)	437.66 (2)	305544 (2)	14
	Stein	-81.21 (2)	486.82 (2)	357941 (2)	-33.43 (2)	431.05 (1)	293531 (1)	10
36	OLSAll	182.87 (1)	311.82 (1)	186567 (1)	131.79 (1)	334.70 (2)	175253 (2)	8
	OLS	337.49 (3)	425.26 (3)	225389 (3)	1696.63 (3)	1696.63 (3)	4044788 (3)	18
	Stein	272.50 (2)	374.90 (2)	199914 (2)	161.91 (2)	298.38 (1)	136382 (1)	10

Table 4. (Continued).

Locality Number	Estimator	Test Data Set $\overline{Diff}$	Test Data Set $\overline{\lvert Diff \rvert}$	Test Data Set $\overline{Diff^2}$	Pooled data set $\overline{Diff}$	Pooled data set $\overline{\lvert Diff \rvert}$	Pooled data set $\overline{Diff^2}$	Rank Sum
40	OLSAll	654.51 (3)	807.18 (1)	1302473 (1)	939.08 (3)	1084.20 (2)	7409968 (2)	12
	OLS	-607.15 (2)	1008.64 (3)	2584182 (3)	-248.75 (2)	1197.18 (3)	7512683 (3)	16
	Stein	-292.29 (1)	842.80 (2)	1589345 (2)	48.14 (1)	1056.87 (1)	6769766 (1)	8

a/ $\overline{Diff} = (1/n)\sum_{i=1}^{n} Diff_i$, where $Diff_i = y_i - \hat{y}_i$ = difference between the ith observed and predicted volume per acre.

b/ $\overline{|Diff|} = (1/n)\sum_{i=1}^{n} |Diff|$.

c/ $\overline{Diff^2} = (1/n)\sum_{i=1}^{n} (Diff)^2$.

d/ Values in parentheses denote the ranks of the estimators relative to one another for that statistic.

Table 5. Sum of ranks over twelve localities for three estimation methods

Estimator	Test Data Set a/ $\overline{Diff}$	$\overline{\lvert Diff \rvert}$	$\overline{Diff^2}$	Total	Number of 1st	Pooled data set $\overline{Diff}$	$\overline{\lvert Diff \rvert}$	$\overline{Diff^2}$	Total	Number of 1st	Rank sum	Overall rank
	Sum of ranks					Sum of ranks						
OLSAll	28	24	27	79	4	31	32	32	95	1	174	3
OLS	27	29	28	84	1	23	25	25	73	2	157	2
Stein	17	19	17	53	8	18	15	15	48	9	101	1

a/ The number of localities where that estimator was ranked first (out of 12 localities).

was not expected because OLS estimators are unbiased, whereas Stein-rule estimators are biased. There are two possible explanations. First, the evaluation was conducted using validation data sets different from the fit data set. Second, the evaluation statistics were based on volume per acre, not logarithm of volume which is the dependent variable of the yield model. OLS estimators therefore did not provide unbiased prediction for volume yield.

The mean absolute difference ($\overline{|Diff|}$) and mean squared difference ($\overline{Diff^2}$) of volume per acre were measures of precision of the model. The ranks based on these two statistics show that Stein-rule estimators provided more precise yield predictions than the other two estimators.

The OLS estimator derived from a specific locality ranked second, better than the overall OLS estimator in predicting volume yield. It was expected that OLS estimates for the entire region provided poorer yield prediction for a specific locality because the yield model for the entire region did not fully explain the variation among localities in site quality, local climatic changes, interaction between trees, etc. (Turnbull 1977).

In order to reveal the amount of improvement from the Stein-rule estimator over the OLS from a specific locality, the average mean difference and the average mean absolute difference from 12 localities were computed for these two estimators in both validation data sets. In the test data set, the Stein-rule estimator was better than the OLS by 64.90 cubic feet per acre in mean difference and by 33.09 cubic feet per acre in mean absolute

difference. In the case of pooled data set, the Stein-rule estimator also gained by 162.83 and 134.88 cubic feet per acre for the mean difference and the mean absolute difference, respectively.

Therefore, the calibration of yield models to a specific locality using the Stein-rule estimator should provide large improvement in terms of bias and precision when this technique is applied to large areas.

Stein-rule estimators appeared to be promising for calibrating yield prediction models to a specific locality. Thus, the prediction capability of yield models could be improved by incorporating information from the entire region with sample information from that locality. For this purpose, Stein-rule estimators was useful under the usual normality assumptions.

Calibration of yield models to a seed source

Stein-rule estimation technique was adopted to calibrate the yield model (3) to each of the 15 seed sources. The resulting parameter estimates from three estimators are presented for 15 seed sources (Table 6).

Using three statistics, the performances of the three estimators were evaluated for each seed source. Ranks by each seed source (Table 7) and overall ranks for the three estimators (Table 8) were then determined based on the evaluation statistics.

Stein-rule estimator performed well on both validation data sets and ranked first overall. Although the Stein-rule estimator was superior to the other two estimators on both validation data

Table 6. Parameter estimates of Burkhart et al. (1972) 's model for fifteen seed sources in the fit data set, using OLS and Stein-rule estimators

Seed source Number	Estimator	b_0	b_1	b_2	b_3	b_4
All	OLSAll [a/]	0.9482	4.1577	0.3174	0.0177	0.0251
301	OLS	0.4403	4.3980	0.3883	0.0380	0.0266
	Stein	0.4840	4.3773	0.3822	0.0362	0.0264
303	OLS	1.3633	-5.7465	0.4893	0.0663	0.0140
	Stein	1.3248	-4.8286	0.4733	0.0618	0.0150
305	OLS	2.2068	-2.8618	0.2371	0.0182	0.0129
	Stein	1.9287	-1.3105	0.2548	0.0181	0.0156
307	OLS	1.2585	-6.3059	0.5402	0.0708	0.0152
	Stein	1.1662	-3.1917	0.4739	0.0550	0.0181
309	OLS	1.8227	-4.4894	0.3304	0.0380	0.0147
	Stein	1.6678	-2.9571	0.3281	0.0344	0.0165
311	OLS	2.3403	0.0732	0.0108	0.0340	0.0173
	Stein	2.2478	0.3445	0.0311	0.0329	0.0178
315	OLS	1.2630	-2.2461	0.5234	0.0339	0.0152
	Stein	1.2628	-2.2440	0.5233	0.0339	0.0152
317	OLS	0.5441	2.8020	0.4174	0.0375	0.0282
	Stein	0.6518	3.1632	0.3908	0.0322	0.0274
319	OLS	1.4156	-5.4168	0.4587	0.0497	0.0155
	Stein	1.2403	-1.8256	0.4057	0.0377	0.0191
321	OLS	1.6868	-5.6826	0.4532	0.0528	0.0115
	Stein	1.3449	-1.1272	0.3903	0.0365	0.0178
323	OLS	2.0418	-4.6460	0.3001	0.0300	0.0133
	Stein	1.8670	-3.2391	0.3029	0.0280	0.0152

Table 6. (Continued).

Seed source Number	Estimator	b_0	b_1	b_2	b_3	b_4
325	OLS	1.9774	-6.1293	0.3591	0.0280	0.0150
	Stein	-0.3653	17.2871	0.2641	0.0046	0.0380
327	OLS	0.7727	8.5346	0.2755	-0.0005	0.0274
	Stein	0.7869	8.1809	0.2788	0.0010	0.0272
329	OLS	1.2565	1.8834	0.2774	0.0294	0.0237
	Stein	1.0512	3.3980	0.3040	0.0216	0.0246
331	OLS	1.9363	-7.5163	0.3803	0.0593	0.0113
	Stein	1.7327	-5.1112	0.3674	0.0507	0.0142

[a/] Ordinary least squares estimates for all fifteen seed sources.

Table 7. Evaluation statistics for three estimation methods, by criterion and seed source

Seed Source Number	Estimators	Test Data Set $\overline{Diff}$	Test Data Set $\overline{\lvert Diff \rvert}$	Test Data Set $\overline{Diff^2}$	Pooled data set $\overline{Diff}$	Pooled data set $\overline{\lvert Diff \rvert}$	Pooled data set $\overline{Diff^2}$	Rank Sum
301	OLSAll	-516.37 (3)	547.60 (3)	508638 (3)	-490.95 (3)	552.12 (3)	480053 (3)	18
	OLS	247.83 (2)	350.12 (2)	204761 (2)	127.38 (2)	333.25 (2)	184555 (2)	12
	Stein	109.28 (1)	288.90 (1)	137447 (1)	13.45 (1)	308.62 (1)	149647 (1)	6
303	OLSAll	67.15 (3)	382.37 (1)	277070 (1)	67.26 (3)	420.98 (1)	359642 (2)	11
	OLS	7.41 (1)	427.49 (3)	346606 (3)	32.60 (1)	448.27 (3)	381291 (3)	14
	Stein	26.31 (2)	392.95 (2)	280078 (2)	33.60 (2)	425.83 (2)	346113 (1)	11
305	OLSAll	2.70 (1)	256.37 (2)	110087 (1)	-39.22 (2)	326.75 (2)	193169 (1)	9
	OLS	217.13 (3)	312.09 (3)	201196 (3)	82.28 (3)	389.52 (3)	382509 (3)	18
	Stein	69.00 (2)	253.72 (1)	112424 (2)	-6.93 (1)	309.38 (1)	240236 (2)	9
307	OLSAll	191.18 (3)	549.21 (1)	491328 (1)	56.41 (1)	470.82 (1)	381020 (1)	8
	OLS	25.85 (1)	754.00 (3)	1032742 (3)	-138.65 (3)	656.15 (3)	1172340 (3)	16
	Stein	91.19 (2)	671.86 (2)	751746 (2)	-69.58 (2)	569.75 (2)	831497 (2)	12
309	OLSAll	132.09 (3)	367.98 (1)	313867 (2)	50.79 (2)	316.34 (3)	209099 (3)	14
	OLS	122.35 (2)	376.18 (3)	318312 (3)	63.52 (3)	273.00 (1)	185709 (2)	14
	Stein	103.36 (1)	375.48 (2)	300758 (1)	36.57 (1)	286.20 (2)	184024 (1)	8
311	OLSAll	-61.41 (2)	403.97 (1)	291178 (2)	47.97 (1)	392.28 (1)	330587 (2)	9
	OLS	-431.72 (3)	593.89 (3)	759760 (3)	-257.31 (3)	525.01 (3)	642930 (3)	18
	Stein	-10.81 (1)	414.79 (2)	282619 (1)	83.48 (2)	414.87 (2)	329281 (1)	9

Table 7. (Continued).

Seed Source Number	Estimators	Test Data Set $\overline{Diff}$	Test Data Set $\overline{\lvert Diff \rvert}$	Test Data Set $\overline{Diff^2}$	Pooled data set $\overline{Diff}$	Pooled data set $\overline{\lvert Diff \rvert}$	Pooled data set $\overline{Diff^2}$	Rank Sum
315	OLSAll	-51.65 (1)	500.70 (1)	331473 (1)	177.04 (3)	513.68 (3)	487152 (3)	12
	OLS	-234.58 (3)	704.82 (3)	724669 (3)	-105.50 (2)	497.43 (1)	439744 (2)	14
	Stein	-120.56 (2)	542.07 (2)	385224 (2)	83.44 (1)	500.17 (2)	410936 (1)	10
317	OLSAll	286.87 (3)	340.02 (1)	277843 (3)	911.38 (3)	1137.00 (1)	1783145 (3)	14
	OLS	233.49 (1)	378.08 (3)	242690 (1)	788.54 (1)	1179.68 (3)	1018005 (1)	10
	Stein	257.66 (2)	349.17 (2)	249748 (2)	851.43 (2)	1148.72 (2)	1419312 (2)	12
319	OLSAll	-173.88 (2)	513.91 (2)	548003 (2)	-129.32 (2)	455.28 (2)	422300 (2)	12
	OLS	-375.38 (3)	631.26 (3)	664273 (3)	-203.11 (3)	628.66 (3)	533578 (3)	18
	Stein	-145.56 (1)	491.04 (1)	467813 (1)	-85.49 (1)	384.53 (1)	317255 (1)	6
321	OLSAll	252.85 (2)	295.64 (2)	205883 (3)	307.02 (3)	388.52 (3)	299121 (3)	16
	OLS	-285.73 (3)	341.46 (3)	201497 (2)	-149.27 (2)	359.03 (2)	241702 (2)	14
	Stein	-19.97 (1)	248.75 (1)	93590 (1)	76.34 (1)	282.50 (1)	159338 (1)	6
323	OLSAll	57.77 (2)	364.62 (1)	246338 (1)	5.39 (1)	460.50 (2)	343355 (2)	9
	OLS	176.82 (3)	376.61 (2)	295527 (3)	113.67 (3)	438.03 (1)	331025 (1)	13
	Stein	-43.84 (1)	398.62 (3)	251226 (2)	-94.20 (2)	510.93 (3)	411279 (3)	14
325	OLSAll	-72.81 (1)	390.32 (1)	390485 (1)	718.70 (1)	1163.95 (1)	1843489 (2)	7
	OLS	2087.98 (3)	2087.98 (3)	5005210 (3)	2876.93 (3)	2876.93 (3)	2295348 (3)	18
	Stein	815.51 (2)	856.27 (2)	1190010 (2)	1409.94 (2)	1573.35 (2)	1600384 (1)	11

Table 7. (Continued).

Seed Source Number	Estimators	Test Data Set $\overline{Diff}$	Test Data Set $\overline{\lvert Diff \rvert}$	Test Data Set $\overline{Diff^2}$	Pooled data set $\overline{Diff}$	Pooled data set $\overline{\lvert Diff \rvert}$	Pooled data set $\overline{Diff^2}$	Rank Sum
327	OLSAll	94.66 (1)	350.18 (1)	219000 (1)	72.21 (1)	364.11 (1)	252013 (1)	6
	OLS	122.90 (3)	412.02 (3)	249514 (3)	86.51 (3)	420.37 (3)	269830 (3)	18
	Stein	108.01 (2)	410.27 (2)	243811 (2)	73.01 (2)	418.73 (2)	265718 (2)	12
329	OLSAll	152.20 (2)	429.14 (3)	410001 (2)	183.20 (3)	325.32 (3)	262264 (3)	16
	OLS	-167.28 (3)	416.34 (2)	432248 (1)	-78.95 (2)	301.72 (2)	246575 (2)	14
	Stein	-26.76 (1)	386.81 (1)	386812 (3)	37.02 (1)	277.28 (1)	229176 (1)	6
331	OLSAll	34.95 (2)	371.08 (3)	250146 (3)	32.07 (3)	420.20 (3)	360742 (3)	17
	OLS	-55.45 (3)	273.73 (1)	146170 (1)	-16.47 (2)	326.14 (1)	249939 (1)	9
	Stein	-3.31 (1)	296.92 (2)	173347 (2)	7.40 (1)	351.66 (2)	289299 (2)	10

Table 8. Sum of ranks over fifteen seed sources for three estimation methods

Estimator	Test Data Set a/ $\overline{Diff}$	$\overline{\lvert Diff \rvert}$	$\overline{Diff^2}$	Total	Number of 1st	Pooled data set $\overline{Diff}$	$\overline{\lvert Diff \rvert}$	$\overline{Diff^2}$	Total	Number of 1st	Rank sum	Overall rank
	Sum of ranks					Sum of ranks						
OLSAll	31	24	27	82	7	32	30	33	95	5	177	2
OLS	37	40	39	116	2	36	34	34	104	3	220	3
Stein	22	26	24	72	7	22	26	23	71	8	143	1

a/ The number of seed sources where that estimator was ranked first (out of 15 seed sources).

sets, the performance of the Stein-rule estimator was not as good as expected. Out of 15 seed sources, the Stein-rule estimator performed better than the two OLS estimators in only 7 and 8 seed sources for the test data set and the pooled data set, respectively (Table 8). This result might be due to the fact that the difference in yield prediction was not as pronounced in seed sources as in localities.

On the other hand, the OLS estimator for all seed sources ranked second overall and performed almost as well as the Stein-rule estimators in the test data set. In the pooled data set, Stein-rule estimators were somewhat better than the OLS for all seed sources in terms of predictive ability of volume yield.

Unlike the results from different localities, the OLS estimator for a specific seed source performed worst overall among the three estimators (Table 8). Since yields from different seed sources were similar, OLS estimators derived from the entire data set should give better yield predictions than those from a specific seed source (with fewer observations).

The average mean difference for 15 seed sources was 143.24 cubic feet for the OLS from the entire region and 130.08 cubic feet for the Stein-rule estimator in the test data set. Thus, the Stein-rule estimator resulted in a reduction of 13.16 cubic feet per acre in mean difference for test data set. The Stein-rule estimator was also 21.80 cubic feet per acre lower in mean difference for the pooled data set.

However, the OLS from the entire region was better than the

Stein-rule estimator in terms of mean absolute difference for both validation data sets. By using the Stein-rule estimator, the precision of the yield model was lost by 20.96 and 3.64 cubic feet per acre for the test data and pooled data sets, respectively.

Even though Stein-rule estimator ranked first overall, the gain obtained from the Stein-rule estimator may not justify the complex calibration procedures. OLS technique might be appropriate in this case and a single regression equation might be adequate for all 15 seed sources in this study.

SUMMARY AND CONCLUSIONS

The main objective of this study was to calibrate yield prediction models to a specific locality or seed source by using a Stein-rule estimator. Twelve localities and fifteen seed sources were used for this study. OLS technique was employed to obtain the parameter estimates for the entire region and also for each subregion (locality or seed source). By combining these two types of estimators, the Stein-rule estimator was employed to provide more precise yield prediction for a specific locality and a specific seed source of interest.

As expected, Stein-rule estimators performed well for calibrating a yield prediction model to a specific locality, ranking first. For seed sources, Stein-rule estimators were just slightly better than OLS estimators, but might not be worth the extra efforts.

STUDY III
USE OF THE KALMAN FILTER ESTIMATION TECHNIQUE TO UPDATE YIELD PREDICTION MODELS

ABSTRACT

The Kalman filter estimation technique was employed to update yield prediction models. Two different sources of prior information were used to modify the estimates from the sample data using the Kalman filter. The Kalman filter and two OLS estimators were evaluated based on the predictive ability of the resulting yield models.

The Kalman filter estimator performed better than the other estimators for both validation data sets. Also, plot data collected inside of the study area formed better prior information than those from outside of the sample data range. This indicated that the quality of prior information was important in using feedback procedures such as the Kalman filter approach.

INTRODUCTION

The ordinary least squares (OLS) estimation methods have been adopted to estimate parameters of yield prediction models using sample data collected from the area of interest. Researchers always face the dilemma of choosing between lower cost of data acquisition and better model performance. Obviously the more data collected, the better such models perform. It is thus desirable to develop a system that efficiently uses all available information rather than collecting additional data to improve yield estimates.

This system can be developed by feedback procedures that modify parameter estimates of models by combining prior information with existing sample data. The feedback procedures have been mainly conducted by using Bayesian estimation methodology for updating forest inventory (Ek and Issos 1978a, 1978b; Green and Strawderman 1985; Green 1986). Kalman filter estimation technique is another feedback procedure. Unlike Bayesian estimators, the Kalman filter is simple and intuitive because no assumption is made of the distributional form of the prior and sample data. The only assumption is that the errors are independent and identically distributed. The Kalman filter has been used in forest inventory systems (Dixon and Howitt 1979) and in localizing site index equations (Walters and Burkhart 1987). A similar method can be applied to the improvement of yield estimates from regression methods.

In this study, the Kalman filter estimator was used to update yield prediction models. Its performance was then evaluated against those of traditional OLS estimators.

LITERATURE REVIEW

The following literature review is focused on past work on updating forestry inventory and on updating regression coefficients in forestry. Literature related to Kalman filter estimator is also reviewed.

Updating Forestry Inventory

Updating parameters from a model means improving the precision of the model using all possible information. The updating efforts in forestry fields have mainly centered on forest inventory. Much research for updating forest inventory has been done by Ek and associates (Ek and Issos 1978a, 1978b; Burk and Ek 1982). They applied James-Stein and empirical Bayes procedures to increase the precision and efficiency of estimates for stand basal area and stand volume. Prior information from nearby stands were merged with current information based on a forest survey from the area of interest. Through simulation studies and analytical methods they found that significant gains in efficiencies of the estimates could be realized, particularly when current information is limited due to small survey data.

Dixon and Howitt (1979) used the Kalman filter in a forest inventory system. They provided the conditional mean and conditional covariance of the inventories using the Kalman filter approach. They also compared the Kalman filter to a recursive

estimator proposed by Ware and Cunia (1962) and concluded that the variance of the Kalman filter estimator was almost always less than the variance of the Ware and Cunia estimator.

Green (1986) reviewed some updating procedures such as empirical Bayes and composite estimator for forestry inventory. The composite estimator, a kind of Bayesian estimator, is basically a weighted average of two or more other estimators.

Updating Regression Parameter Estimates

Green and Strawderman (1985), in the development of individual tree volume equations for both pine and hardwood, examined the feasibility of using empirical Bayes estimators (Zellner 1971; Box and Tiao 1973) to construct volume equations with greater predictive ability. They compared empirical Bayes estimators to weighted least squares estimators and concluded that the empirical Bayes estimators should be used to improve the predictive ability of volume equations only when good prior information was available. They also found that the estimators could be used to reduce the amount of field data necessary to produce an estimate with a stated allowable error.

More recently, Walters and Burkhart (1987) presented a procedure for the prediction of height-age relationship through the use of updated equations. A site index equation was updated to a particular stand by applying the Kalman filter estimator.

Kalman Filter Estimation

Filtering is the estimation of the current state of a system based on the current sample and all prior samples and information. It is similar to empirical Bayes estimation. However, filtering does not need any assumptions except that the errors are independent and identically distributed, whereas the empirical Bayes estimation requires the standard normality assumptions. Kalman filter theory was introduced as an alternative approach to the classical estimation problem by Kalman (1960). The theory, which is commonly used in engineering fields, is a sequential implementation of the Goldberger-Theil mixed estimator (Theil 1963) that combines prior information in linear models. Diderrich (1985) derived the updating step of the Kalman filter estimator that is equivalent to the Goldberger-Theil mixed estimator.

Many researchers (Bierman 1976; Mehra 1979; Sorenson 1980; Sallas and Harville 1981; Diderrich 1985) indicated the connection between least squares estimation and the Kalman filter theory. Sallas and Harville (1981) used the Kalman filter to obtain recursive estimators, which were extended to mixed models. Diderrich (1985) concluded that the Kalman filter is just least squares estimation made into a recursive process by combining prior information with sample information. However, such oversimplifications result in loss of important insight as an estimation of a dynamic process (Welch 1987). The Kalman filter estimator can successfully be used in time series data.

Duncan and Horn (1972) introduced parameter update equations

based on a random coefficients regression theory as a natural extension of conventional regression theory. By expressing prior expectation as part of the observation vector, they derived a linear unbiased estimator with the minimum MSE for the coefficients of a simple regression equation and then extended the results to the Kalman filter model.

A Bayesian approach to regression theory presents another way to view the Kalman filter derivation. Meinhold and Singpurwalla (1983) derived the basic equations of Kalman filter theory from a Bayesian point of view. They established the joint density of the parameter and the predicted residuals, conditional on previous observations. Broemeling (1985) viewed Kalman filtering as part of a Bayesian treatment of general linear models.

In the field of forestry, Kalman filter estimators were used in a forestry inventory system (Dixon and Howitt 1979) and localizing site index equations (Walters and Burkhart 1987).

MATERIALS AND METHODS

Data

A portion of the data set from the Southwide Loblolly Pine Seed Source study was used in this study. A total of 226 plots from the West Gulf region (Louisiana, Mississippi, Arkansas, and Texas) was chosen to develop a yield prediction model. The detailed information about the data used in this study is presented in Table 1. Similar to the previous studies, only one age class from each of 226 plots was randomly selected to form a data set to simulate temporary plot data often used for developing yield models.

A half-and-half data splitting method was adopted to divide the West Gulf region data into a fit and test data set. The fit data set, representing sample data, was used to estimate parameters of yield prediction models using the OLS estimator. The test data set, representing an independent data set, was withheld to evaluate the performances of yield models from different regression estimators. The pooled data set, which was the combined fit and test data sets, was used to represent the population. The summary of the stand attributes for both fit and test data sets is shown in Table 2.

Procedure

In this study, regression coefficients of a yield model was

Table 1. Number of plots for the West Gulf region states from the Southwide Loblolly Pine Seed Source Study

Locality number	State	Number of plots
28	Mississippi	34
29	Mississippi	36
32	Louisiana	68
36	Texas	16
40	Arkansas	72
Total		226

Table 2. Stand attributes of the fit and test data sets used in this study

Variable	Number of observations [a/]	Mean	Minimum	Maximum
	- - - - - - - - - - Fit Data Set - - - - - - - - - -			
Age	113	19	10	27
H_d	113	53	18	80
N	113	505	24	2099
V	113	3824	168	7133
	- - - - - - - - - - Test Data Set - - - - - - - - - -			
Age	113	17	10	27
H_d	113	47	18	77
N	113	510	49	2198
V	113	2915	106	6307

[a/] Notation:

Age = Plantation age in years,

H_d = Average height of the dominant and codominants in feet,

N = Number of trees per acre, and

V = Total outside-bark volume per acre in cubic-foot.

updated to improve its predictive ability using feedback procedures. The Kalman filter estimator resulted from combining the sample data and prior information.

Model form for yield prediction

The model form for yield prediction developed by Burkhart et al. (1972) was used in this study. The model form is given by

$$\log(V) = b_0 + b_1(1/A) + b_2(H_d/A) + b_3(N/100) + b_4(A)[\log(N)] \quad (1)$$

where

V = total cubic-foot volume outside bark per acre,

A = stand age in years,

H_d = average height of dominants and codominants in feet,

N = number of surviving trees per acre, and

log(x) = logarithm (base 10) of x.

Total cubic-foot volume outside bark per acre was computed using Smalley and Bower's (1968) individual tree volume equation. The mean height of the tallest 50 percent of surviving trees at each age was considered as average height of the dominants and codominants for each plot. Also, as in the previous studies, the dependent and independent variables were standardized such that they have the same mean and variance. This process enhanced the precision in computing the inverse matrices.

Kalman filter estimator for updating yield prediction models

If there exists some prior information, it can be combined with sample data to update yield models. The prior information is

defined as

$$\underline{b}_p = \underline{\beta} + \underline{e}_p \qquad (2)$$

where

$\underline{b}_p$ = a prior estimate of the parameter $\underline{\beta}$,

$\underline{\beta}$ = parameter to be estimated, and

$\underline{e}_p$ = error vector of mean 0 and covariance matrix P.

Also, the sample information can be defined as follows:

$$\underline{y} = X\underline{\beta} + \underline{e} \qquad (3)$$

where

$\underline{y}$ = vector of dependent variable,

X = Matrix of independent variables, and

$\underline{e}$ = error vector of mean 0 and covariance matrix W.

In addition, the error vectors $\underline{e}_p$ and $\underline{e}$ are assumed to be uncorrelated. With this assumption, the Kalman filter estimator is given by:

$$\underline{b}_{KF} = \underline{b}_p + K(\underline{y} - X\underline{b}_p) \qquad (4)$$

where

$\underline{b}_{KF}$ = the Kalman filter estimator, and

$K = PX'[W + XPX']^{-1}$ which is the gain calculation.

In this study, the Kalman filter estimator (4) was employed to update the yield prediction model (1) combining information from sample sample data with prior information.

Prior information

In order to access the importance of the quality of prior information, two scenarios were considered in this study. The

first scenario involved using an available data set as prior information. Data from 55 plots in a different study conducted at the Hill Farm Research Station were selected for this purpose. The yield model (1) was fitted to the Hill Farm data set using OLS to obtain parameter estimates and the covariance matrix. This information was combined with the sample data from the West Gulf region to update the yield model using the Kalman filter estimator (4).

In the second scenario, it was assumed that no data was available. Prior information came from different sets of coefficients for the same yield model. For this purpose, regression coefficients were obtained from each of the seven remaining localities of the Southwide Seed Source Study. Data summary for these localities is presented in Table 3. Prior information of the parameters for the yield model was assumed to have mean and covariance equal to the sample mean and sample covariance of the seven sets of parameter estimates. Green and Strawderman (1985) used the published 6 coefficients of individual tree volume equations as prior information.

The two different types of prior information used in this study might reveal how prior information affected the results of updating yield models. The first source of prior information was obtained from inside of the sample data range. On the other hand, the second source of prior information came from outside of the West Gulf region. Thus, this study may provide insights on the

Table 3. Stand attributes of two different sources of prior information used in this study

Variable	Number of observations	Mean	Minimum	Maximum
	- - - - - - - Hill Farm data set - - - - - - - - -			
Age	55	17	10	29
H_d	55	51	19	78
N	55	507	92	1200
V	55	3133	182	6211
	- - - Southwide Seed Source localities outside of the West Gulf region - - - - - - - -			
Age	296	18	10	27
H_d	296	46	22	72
N	296	735	123	2642
V	296	3493	709	12484

importance of the quality of prior information in feedback procedures.

Evaluation criteria

Three yield models from different parameter estimation methods were evaluated in this study. One included the OLS estimates fitted to the sample data. Another had its estimates based on prior information. Sometimes these estimates can be directly applied to the stand of interest without collecting other data. The third included Kalman filter estimates obtained by combining OLS estimates with the prior information.

The three candidate yield models were evaluated based on mean difference ($\overline{Diff}$), mean absolute difference ($\overline{|Diff|}$), and mean squared difference ($\overline{Diff^2}$). These evaluation criteria were described in detail in the previous studies. These criteria were computed separately for the test data and the pooled data (combined fit and test data sets). The test data set represented an independent data set, whereas the pooled data set represented the population.

The yield models were ranked based on each criteria, with rank number one being best. For each estimator, the overall rank was calculated as the sum of ranks for all three criteria. Finally, the "best" estimator was determined by the one with the smallest overall rank.

RESULTS AND DISCUSSION

It is apparent that the success of updating models depends upon the choice of the prior information available. In this study, two different sources of prior information were used to update yield prediction models with the Kalman filter estimator.

Prior information from the Hill Farm data set

The parameter estimates of the yield equation and their covariance matrix were obtained from the Hill Farm data set (Table 4). These values used as prior information. The three estimates - - OLS estimates from the sample data and from the prior information, and the Kalman filter estimates - - are presented in Table 5. The resulting three yield prediction equations were evaluated based on three statistics for both the test data set and the pooled data set (Table 6). The Kalman filter estimator provided gains of 16.94 and 11.81 cubic feet per acre over the OLS in mean difference for test and pooled data sets, respectively. In addition, this estimator also reduced mean absolute difference by 22.63 and 23.67 cubic feet per acre for the test and pooled data sets.

Based on these evaluation statistics, overall ranks of the estimators were determined (Table 7). As expected, the Kalman filter estimator ranked first in both of the validation data sets, whereas the OLS estimator based on the sample data ranked second,

Table 4. Prior information used in this study based on Hill Farm data set

Parameter Estimates

b_0	b_1	b_2	b_3	b_4
2.3375	-7.9635	0.2896	-0.0062	0.0119

Covariance matrix of parameter estimates

	b_0	b_1	b_2	b_3	b_4
b_0	0.253414	-0.684965	-0.045962	0.000644	-0.001712
b_1	-0.684965	3.142385	0.076970	-0.008212	0.006555
b_2	-0.045962	0.076970	0.010387	0.000069	0.000224
b_3	0.000644	-0.008212	0.000069	0.000071	-0.000015
b_4	-0.001712	0.006555	0.000224	-0.000015	0.000016

Table 5. Parameter estimates of the yield model when prior information was based on the Hill Farm data set, by estimation method

Estimator [a/]	b_0	b_1	b_2	b_3	b_4
OLS	1.3319	-3.5991	0.4567	0.0319	0.0180
Prior	2.3357	-7.9635	0.2896	-0.0062	0.0119
Kalman filter	1.6925	-7.8390	0.4682	0.0344	0.0142

[a/] Notation:

OLS = Ordinary least squares estimates from the fit data set,

Prior = Parameter estimates from the prior information based on the Hill Farm data set, and

Kalman filter = Kalman filter estimates obtained by combining the sample data with the prior information.

Table 6. Evaluation statistics for three estimation methods when the prior information was based on the Hill Farm data set

Estimator	Diff [a]	|Diff| [b]	$Diff^2$ [c]
	- - - - - - - - - - Test Data Set - - - - - - - - - -		
OLS	-44.26	400.99	416481
Prior	16.33	495.41	623339
Kalman filter	-27.32	378.37	354731
	- - - - - - - - - - Pooled Data Set - - - - - - - - - -		
OLS	-33.47	435.33	468967
Prior	51.39	567.68	677481
Kalman filter	-21.66	411.66	403114

[a] Mean difference.

[b] Mean absolute difference.

[c] Mean squared difference.

Table 7. Ranks of evaluation statistics for three estimation methos when the Hill Farm data set was used as prior information

	Test data set				Polled data set					
Estimator	$\overline{Diff}$	$\overline{\lvert Diff \rvert}$	$\overline{Diff^2}$	Total	$\overline{Diff}$	$\overline{\lvert Diff \rvert}$	$\overline{Diff^2}$	Total	Rank sum	Overall rank
OLS	3	2	2	7	2	2	2	6	13	2
Prior	1	3	3	7	3	3	3	9	16	3
Kalman filter	2	1	1	4	1	1	1	3	7	1

and the OLS estimator from prior information ranked last overall.

Kalman filter estimators can be expected to perform better than OLS estimators only when good prior information was available. In this study, the prior information came from data collected in Northern Louisiana, which is located within the study area of the West Gulf region. Thus, this type of prior information should be valuable in improving parameter estimates of yield models. Green and Strawderman (1985) traced different results in tree volume prediction to the quality of prior information. Previous height measurements was used by Walters and Burkhart (1987) as excellent prior information for refining parameter estimates of site index equations.

Prior information from other localities of the Southwide Loblolly Pine Seed Source

A different set of prior information was adopted in updating yield models. Parameter estimates from the other seven localities of the Southwide Loblolly Pine Seed Source Study are shown in Table 8. The mean and covariance matrix computed from these estimates constituted prior information in this scenario. The OLS estimator was used to estimate parameters of the yield model (1) using the same sample data from the West Gulf region. The Kalman filter estimator (4) was then employed to modify the OLS estimates with the prior information. Parameter estimates for the three estimation methods are shown in Table 9. For the resulting yield prediction equations, the same statistics were used to evaluate

Table 8. Prior information based on parameter estimates of seven localities

Locality numbers	Location	Parameter Estimates b_0	b_1	b_2	b_3	b_4
03	Maryland	1.8445	-4.0227	0.2781	0.0522	0.0159
07	North Calorina	1.7265	-2.3223	0.2899	0.0266	0.0190
13	South Calorina	1.7558	-6.8030	0.4093	0.0541	0.0127
15	North Calorina	2.0505	-3.1658	0.3026	0.0239	0.0120
17	Gorgia	2.4767	-10.2309	0.3590	0.0658	0.0065
25	Alabama	1.4499	-5.0658	0.5096	0.0568	0.0181
26	Alabama	2.0876	-3.9396	0.2853	0.0278	0.0125

Covariance matrix of parameter estimates

	b_0	b_1	b_2	b_3	b_4
b_0	0.107621	-0.460573	-0.013310	0.000310	-0.001247
b_1	-0.460573	7.197977	-0.092083	-0.037601	0.008482
b_2	-0.013310	-0.092083	0.007375	0.000880	0.000043
b_3	0.000310	-0.037601	0.000880	0.000296	-0.000022
b_4	-0.001247	0.008482	0.000043	-0.000022	0.000018

Table 9. Parameter estimates of the yield model when prior information was based on seven localities from the Southwide Loblolly Pine Seed Source Study

Estimator	b_0	b_1	b_2	b_3	b_4
OLS	1.3319	-3.5991	0.4567	0.0319	0.0180
Prior	1.9131	-5.0786	0.3477	0.0439	0.0138
Kalman filter	1.4365	-3.8187	0.4346	0.0336	0.0173

their performances for both validation data sets (Table 10). The resulting ranks from these evaluation criteria are shown in Table 11.

The Kalman filter estimator also performed better than the OLS as in the previous situation. However, in this case, the amount of improvement over the OLS was not as large. The Kalman filter reduced the mean difference only by 3.39 and 5.74 cubic feet per acre for the test and pooled data sets, respectively. A reduction of mean absolute difference by 4.99 and 9.14 cubic feet per acre was obtained by the Kalman filter estimator over the OLS.

The results were consistent for all statistics for both data sets. Again, the Kalman filter estimator ranked first, with the OLS estimator second. The estimator from prior information ranked last again and provided worse results than in the previous case where the prior information was from the Hill Farm data set. This might be due to the difference in the qualities of two sources prior information.

The prior information used in this case was from plots outside of the West Gulf region, but the Hill Farm data were collected in the same region as the sample data. The estimator based solely on prior information did provide poorer prediction of volume yield when the prior information was outside of the sample data range. The same logic probably explained why the Kalman filter estimator consistently performed better when the prior information was from the Hill Farm data set, based on the evaluation statistics. The results indicated how important the quality of prior information

Table 10. Evaluation statistics for three estimation methods when prior information was based on seven localities from the Southwide Loblolly Pine Seed Source Study

Estimator	$\overline{Diff}$	$\overline{\lvert Diff \rvert}$	$\overline{Diff^2}$
	- - - - - - - Test Data Set - - - - - - - - -		
OLS	-44.26	400.99	416481
Prior	184.68	860.12	744735
Kalman filter	-40.85	396.00	397205
	- - - - - - - Pooled Data Set - - - - - - - -		
OLS	-33.47	435.33	468967
Prior	55.67	931.11	933674
Kalman filter	-27.73	426.19	453329

Table 11. Ranks of evaluation statistics for three estimation methos when the seven localities were used as prior information

Estimator	Test data set				Polled data set				Rank sum	Overall rank
	$\overline{Diff}$	$\overline{\lvert Diff \rvert}$	$\overline{Diff^2}$	Total	$\overline{Diff}$	$\overline{\lvert Diff \rvert}$	$\overline{Diff^2}$	Total		
OLS	2	2	2	6	2	2	2	6	12	2
Prior	3	3	3	9	3	3	3	9	18	3
Kalman filter	1	1	1	3	1	1	1	3	6	1

was in applying the Kalman filter estimation technique. Green and Strawderman (1985) found little difference between empirical Bayes and least squares methods for loblolly pine in terms of predictive ability of individual tree volume equations. Moreover, they concluded that for red maple, least squares was superior to empirical Bayes estimation method. This might be because the prior information used in that study was based on equations from data collected at areas outside of the range of the sample data.

SUMMARY AND CONCLUSIONS

The objective of this study was to update yield prediction models using the Kalman filter estimation method. A total of 226 plots from the West Gulf region (Louisiana, Mississippi, Arkansas, and Texas) comprised the sample data. After the data were randomly divided into the fit and test data sets, the yield prediction model was fitted to the fit data set using the OLS technique. The OLS estimates were then modified by additional information (prior information) using the Kalman filter estimator. In this study, two different sets of prior information were used. One was the Hill Farm data set collected within the study area. The other came from parameter estimates of the yield model fitted to data from seven localities located outside of the study area.

For both types of prior information, the Kalman filter estimator ranked better that the OLS estimators. Also, the Kalman filter estimates from the first source of prior information (Hill Farm data set) provided better prediction of volume yield than those from the second source of prior information (outside of the sample data range). The OLS estimators ranked second and the estimators based solely on the prior information ranked last overall as expected.

The Kalman filter technique is a promising approach to update yield prediction models. However, this estimator should be used with caution because the improvement in prediction requires good prior information.

LITERATURE CITED

Allen, D. M. 1974. The relationship between variable selection, data augmentation, and a method of prediction. Technometrics 16:125-127.

Baranchik, A. J. 1970. A family of minimax estimators of the mean of a multivariate normal distribution. Ann. Math. Stat. 41:642-645.

Bare, B. B. and D. W. Hann. 1981. Applications of ridge regression in forestry. Forest Sci. 27:339-348.

Bates, D. M., M. J. Lindstorm, G. Wahba, and B. S. Yandell. 1987. GCVPACK - routines for generalized cross validation. Commun. in Stat. 16(1):263-297.

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. Regression diagnostics. John Wiley & Sons, New York. 292 p.

Bierman, G. J. 1976. Factorization methods for discrete sequential estimation. New York, Academic Press. 241 p.

Box, G. E. and G. C. Tiao. 1973. Bayesian inference in statistical analysis. Addison-Wesley Publ. Co, Reading. MA. 588p.

Broemeling, L. 1985. Bayesian analysis of linear models. New York, Marcel Dekker.

Brown, W. G. and B. R. Beattie. 1975. Improving estimates of economic parameters by use of ridge regression with production function applications. Am. J. Agric. Econ. 57:21-32.

Burk, T. E. and A. R. Ek. 1982. Application of empirical Bayes/ James-Stein procedure to simultaneous estimation problems in forest inventory. Forest Sci. 28:753-771.

Burkhart, H. E., R. C. Parker., M. R. Strub, and R. G. Oderwald. 1972. Yield of old-field loblolly pine plantations. Va. Poly. Inst. and State Univ. Pub. FWS-3-72, 51 p.

Carter, G. M. and J. E. Rolph. 1974. Empirical Bayes methods applied to estimating fire alarm probabilities. J. Am. Stat. Assoc. 69:880-885.

Chatterjee, S. and B. Price. 1977. Regression analysis by example. John Wiley & Sons, New York. 228 p.

Delaney, N. J. and S. Chatterjee. 1986. Use of the bootstrap and cross-validation in ridge regression. J. Busi. & Econ. Assoc. 72:77-93.

Dempster, A. P., M. Schatzoff and N. Wermuth. 1977. A simulation study of alternatives to ordinary least squares. J. Am. Stat. Assoc. 72:77-93.

Diderrich, G. T. 1985. The Kalman filter from the perspective of Goldberger-Theil estimators. Am. Statistician 39:193-198.

Dixon, B. L. and R. E. Howitt. 1979. Continuous forest inventory using a linear filter. Forest Sci. 25:675-689.

Draper, N. R. and R. C. Van Nostrand. 1979. Ridge regression and James-Stein estimation; Review and comments. Technometrics 21(4):451-466.

Duncan, D. B. and S. D. Horn. 1972. Linear dynamic recursive estimation from the viewpoint of regression analysis. J. Am. Stat. Assoc. 67:815-821.

Efron, B. and C. Morris. 1972a. Limiting the risk of Bayes and empirical Bayes estimators - part II: The empirical Bayes case. J. Am. Stat. Assoc. 67:130-139.

Efron, B. and C. Morris. 1972b. Empirical Bayes on vector observations -- An extension of Stein's method. Biometrika 59:335-347.

Efron, B. and C. Morris. 1973a. Stein's estimation rule and its competitors - an empirical Bayes approach. J. Am. Stat. Assoc. 68:117-130.

Efron, B. and C. Morris. 1973b. Combining possibly related estimation problems. J. Royal Stat. Soc. Ser. B. 35:379-421.

Efron, B. and C. Morris. 1975. Data analysis using Stein's estimator and its generalizations. J. Am. Stat. Assoc. 70:311-319.

Ek, A. R. and J. N. Issos. 1978a. Bayesian theory and multi-resource inventory. In integrated inventories of renewable natural resources: Proceedings of the workshop, p 291-298. USDA For. Serv. Gen. Tech. Rep. RM-55.

Ek, A. R. and J. N. Issos. 1978b. Bayesian estimation methodology for forest inventory analysis. In Forestry inventory proceedings IUFRO Subject Groups S4.02 and S$.04, June 18-26, 1978, p 34-45.

Erickson, G. M. 1983. Using ridge regression to directly estimate lagged effects in marketing. J. Am. Stat. Assoc. 76:766-773.

Farrar, D. E. and R. R. Glauber. 1967. Multicollinearity in regression analysis. Rev. Econ. and Stat. 51:486-489.

Fay, R. E. and R. A. Herriot. 1979. Estimates of income for small places: an application of James-Stein procedures to census data. J. Am. Stat. Assoc. 74:269-277.

Fomby, T. B. and R. C. Hill. 1978. Multicollinearity and minimax conditions for Bock Stein-like estimator. Econometrica 47:211-212.

Fries, J. 1965. Eigenvector analysis show that birch and pine have similar form in Sweden and British Columbia. For. Chron. 41:135-139.

Gertner, G. Z. 1984. Localizing a diameter increment model with a sequential Bayesian procedure. Forest Sci. 30:851-864.

Gibbons, D. G. 1983. A simulation study of some ridge estimators. J. Am. Stat. Assoc. 76:131-139.

Golden, M. S., R. Meldahl, S. A. Knowe, and D. B. Boyer. 1981. Predicting site index for old-field loblolly pine plantations. South. J. Appl. For. 5(3):109-114.

Golub, G. H., M. Heath, and G. Wahba. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21:215-223.

Green, E. J. 1986. Empirical Bayes procedures for updating forest inventories. In Proceedings of the 1986 Soc. Am. For. Natl. Conv. p 67-69.

Green, E. J. and W. E. Strawderman. 1985. The use of Bayes/Empirical Bayes estimation in individual tree volume equation development. Forest Sci. 31:975-990.

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. Forest Sci. 33:1054-1061.

Gunst, R. F. and R. L. Mason. 1977. Biased estimation in regression: An evaluation using mean squared error. J. Am. Stat. Assoc. 72:616-628.

Hill, R. C., T. B. Fomby, and S. R. Johnson. 1977. Component selection norms for principal component regression. Commun. in Stat. 6:309-333.

Hocking, R. R. 1976. The analysis and selection of variables in linear regression. Biometrics 32:1-50.

Hoerl, A. E. and R. W. Kennard. 1970a. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12:55-67.

Hoerl, A. E. and R. W. Kennard. 1970b. Ridge regression: applications to nonorthogonal problems. Technometrics 12:69-82.

Hoerl, A. E. and R. W. Kennard. 1976. Ridge regression: iterative estimation of the biasing parameter. Commun. in Stat. A5:77-88.

Hoerl, A. E., R. W. Kennard, and K. F. Baldwin. 1975. Ridge regression: some simulations. Commun. in Stat. 4:105-123.

James, W. and C. Stein. 1961. Estimation with quadratic loss. Proceedings of the fourth Berkely symposium on Math. Stat. and Prob. 1:361-379.

Johnson, S. R., S. C. Reimer, and T. P. Rothrock. 1973. Principal components and the problem of multicollinearity. Metroeconomica 25:306-317.

Judge, G. G., W. E. Griffiths, R. C. Hill, and T. Lee. 1985. The theory and practice of econometrics. John Wiley & Sons, New York. 793 p.

Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lutkepohl, and T. Lee. 1988. Introduction to the theory and practice of econometrics. John Wiley & Sons, New York. 1024 p.

Kalman, R. E. 1960. A new approach to linear filtering and prediction problems. J. Basic Engin. 82:35-45.

Kmenta, J. 1971. Elements of econometrics. MacMillan, New York. 665 p.

Kozak, A. and J. H. G. Smith. 1966. Critical analysis of multivariate techniques for estimating tree taper suggests that simpler methods are best. For. Chron. 42:458-463.

Li, K. C. 1986. Asymptotic optimality of C_L, and generalized cross-validation in ridge regression with application to spline smoothing. The annals of Stat. 14:1101-1112.

Lindley, D. V. and A. F. Smith. 1972. Bayes estimators for the linear model. J. Royal Stat. Soc. Ser. B. 34:1-42.

Liu, C. J. and T. D. Keister. 1977. Southern pine stem form defined through principal component analysis. Can. J. For. Res. 8:188-197.

Looney, S. W. and D. B. Brock. 1979. An application of James-Stein and related estimators to small-are data from the National Center for Health Statistics. Paper presented at Joint Mtg Biometric Soc and Inst Math Stat, New Orleans, La. 21 p.

Lott, W. F. 1973. Optimal set of principal component restrictions on a least squares regression. Commun. in Stat. 2:449-464.

Mackinney, A. L. and L. E. Chaiken. 1939. Volume, yield, and growth of loblolly pine in the Mid-Atlantic Coastal Region. U. S. For. Serv. Tech. Note No. 33, 30 p.

Mallows, C. L. 1973. Some comments on C_p. Technometrics 15:661-675.

Mansfield, E. R. 1975. Principal component approach to handling multicollinearity in regression analysis. Ph.D. Dissert. Dep. of Stat., South. Methodist Univ. Dallas, Texas. 89 p.

Marquardt, D. W. 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics 12:591-612.

Massy, W. F. 1965. Principal component regression in exploratory statistical research. J. Am. Stat. Assoc. 4:277-292.

Matney, T. G., A. D. Sullivan, J. R. Ledbetter, and R. M. Farrar. 1988. Stand-level cubic-foot volume ratio equations for planted loblolly pine on site-prepared land in the mid-South Gulf Coastal plain. South. J. Appl. For. 12:7-11.

Mayer, L. S. and T. A. Willke. 1973. On biased estimation in linear models. Technometrics 15:497-508.

McDonald, G. C. and D. I. Galarneau. 1975. A monte carlo evaluation of some ridge-type estimators. J. Am. Stat. Assoc. 78:407-416.

McDonald, G. C. and R. C. Sehwing. 1973. Instabilities of regression estimates relating air pollution to mortality. Technometrics 12:591-612.

Mehra, R. K. 1979. Kalman filters and their applications to forecasting: In TIMS studies in management sciences, ed. M. K. Starr, Amsterdam, North-Holland. 378 p.

Meinhold, R. J. and N. D. Singpurwalla. 1983. Understanding the Kalman filter. Am. Stat. 37:123-127.

Mitchell, B. R. and D. W. Hann. 1979. A computer program for applying ridge regression techniques to multiple linear regression. USDA For. Serv. Gen. Tech. Rep. INT-51, 25 p.

Morris, C. 1977. Interval estimation for empirical Bayes generalizations of Stein's estimator. In Proc. 22nd Conf. on the design of Exps. in Army Res. Development and Testing. ARO Rep. 77-2. 219-249 p.

Munro, D. D. 1966. The distribution of log size and volume within trees: A preliminary investigation. Univ. of B.C., Fac. of For. Directed Study. 27 p. (Original not seen. For. Chron. 42:458-463).

Myers, R. H. 1986. Classical and modern regression with applications. Duxbury Press, 359p.

Neter, J., W. Wasserman, M. H. Kutner. 1985. Applied linear statistical models. Homewood, Illinois. Irwin. 1127p.

Newcomer, J. A. and W. L. Myers. 1984. Principal components analysis of total tree form in seven central Pennsylvania hardwood species. Forest Sci. 30:64-70.

Sallas, W. M. and D. A. Harville. 1981. Best estimation for mixed linear models. J. Am. Stat. Assoc. 76:860-869.

Smalley, G. W. and D. R. Bower. 1968. Volume tables and point factors for loblolly pines in plantations on abandoned fields in Tennessee, Alabama, and Georgia highlands. USDA For. Serv. South. For. Exp. Stn. Res. Pap. SO-32, 13 p.

Smith, W. B. 1983. Adjusting the STEMS regional forest growth model to improve local predictions. USDA For. Serv. Res. Note NC-297, 5 p.

Snee, R. D. 1977. Validation of regression models: Methods and examples. Technometrics 19:415-428.

Sorenson, H. W. 1980. Parameter estimation: Principles and Problems, New York, Marcel Dekker.

Stage, A. 1981. Use of self calibration procedures to adjust general regional yield models to local conditions. In Proceedings of forest resource inventory, growth models, management planning, and remote sensing, p365-375. IUFRO World Congress, Kyoto, Japan.

Stein, C. 1955. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proc. 3rd Berkeley Symp. Math. Stat. and Prob. 1:197-202. Univ. Calif. Press.

Theil, H. 1963. On the use of incomplete prior information in regression analysis. J. Am. Stat. Assoc. 58:401-414.

Turnbull, K. J. 1977. Long-term yield forecasting models; validation and iterative estimation. In growth models for long-term forecasting of timber yield. Va. Polytech & State Univ., For. & Wildl. Resour., FWS 1-78, 249 p.

Vinod, H. D. 1978. A survey of ridge regression and related techniques for improvements over ordinary least squares. Rev. Econ. and Stat. 60:121-131.

Vinod, H. D. and J. E. Ullah. 1981. Recent advance in regression methods. John Willey & Sons, New York. 267 p.

Ware, K. D. and T. Cunia. 1962. Continuous forest inventory with partial replacement. Forest Sci. Monogr. 3. 40 p.

Walters, D. K. and H. E. Burkhart. 1987. A method for localizing site index equations. Proc. of IUFRO Forest Growth Modelling and Prediction Conference. Minneapolis, MN, August 24-28.

Welch, M. E. 1987. A Kalman filtering perspective. Am. Stat. 41:90-91.

Wells, O. O. and P. C. Wakeley. 1966. Geographic variation in survival, growth, and fusiform-rust infection of planted loblolly pine. Forest Sci. 11:1-40.

Zellner, A. 1971. An introduction to Bayesian inference in econometrics. John Wiley & Sons, New York. 431 p.

VITA

Man Yong Shin was born in Seoul, Korea on December 31, 1953. He graduated from Chung-Dong High School, Seoul, Korea in 1972 and received his B.A. in Forestry at Kyung-Hee University of Korea in 1981. During that time he served in the Korean National Army for three years. He entered Kyung-Hee University in March, 1981 for graduate study in Forest Genetics. After receiving his M.A. degree, he came to the United States in August of 1983 to study at Iowa State University. He received his master of science degree in Forest Biometrics at Iowa State University in 1986. He then started his Ph.D program in Forest Biometrics at Louisiana State University and was inducted into Xi Sigma Pi National Forestry Honor Society. He is currently a candidate for Doctor of Philosophy in Forestry.

DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate: Man Yong Shin

Major Field: Forestry

Title of Dissertation: Methods of Parameter Estimation of Linear Regression Models for Yield Prediction.

Approved:

Quang V Cao
Major Professor and Chairman

F. Glen Hembry
Dean of the Graduate School

EXAMINING COMMITTEE:

Ben D Jackson

Luis G. Escobar.

James E. Hotvedt

Wayne H. Hendrick

Jim L. Chambers

Date of Examination:

November 20, 1989