Sampling estimators of total mill receipts for use in timber product output studies

John P. Brown and Richard G. Oderwald

Abstract: Data from the 2001 timber product output study for Georgia was explored to determine new methods for stratifying mills and finding suitable sampling estimators. Estimators for roundwood receipts totals comprised several types: simple random sample, ratio, stratified sample, and combined ratio. Two stratification methods were examined: the Dalenius–Hodges (DH) square root of the frequency method and a cluster analysis method. Three candidate sizes for the number of groups were selected from the cluster analysis and subsequently used in the DH stratification as well. Relative efficiency improved when the number of groups increased and when using a ratio estimator, particularly a combined ratio. The two stratification methods performed similarly. Neither the DH method nor the cluster analysis method performed better than the other. Six bound sizes (1%, 5%, 10%, 15%, 20%, and 25%) were considered for deriving samples sizes for the total volume of roundwood receipts. The minimum achievable bound size was found to be 10% of the total receipts volume for the DH method using a two-group stratification. This was true for both the stratified and combined ratio estimators. In addition, for the stratified and combined ratio estimators, only the DH method stratifications were able to reach a 15% bound on the total (six of the 12 stratified estimators). These results demonstrate that the utilized classification methods are compatible with stratified totals estimators and can provide users with the opportunity to develop viable sampling procedures as opposed to complete mill censuses.

Résumé : Les données de l'étude de 2001 sur les produits de transformation du bois en Géorgie ont été analysées pour identifier de nouvelles méthodes de stratification des usines et pour trouver des estimateurs d'échantillonnage appropriés. Quatre estimateurs du volume total des arrivages de bois rond ont été considérés : l'échantillon aléatoire simple, le rapport, l'échantillon stratifié et le rapport combiné. Deux méthodes de stratification ont été examinées : la méthode de la racine carrée cumulée de la fréquence de Dalenius–Hodges (DH) et une méthode d'analyse par grappes. Trois tailles candidates pour le nombre de grappes ont été sélectionnées à partir de l'analyse par grappes et ensuite elles ont aussi été utilisées dans la méthode de stratification de DH. L'efficacité relative s'est améliorée avec l'augmentation du nombre de grappes et l'utilisation d'un estimateur de rapport, en particulier un rapport combiné. Les deux méthodes de stratification avaient une performance similaire : la méthode de stratification de DH et la méthode d'analyse par grappes étaient équivalentes. Six tailles limites (1 %, 5 %, 10 %, 15 %, 20 %, et 25 %) ont été considérées pour calculer la taille des échantillons pour le volume total des arrivages de bois rond. La taille minimale réalisable était de 10 % du volume total pour la méthode de DH basée sur une stratification en deux groupes. Cette taille minimale était valable à la fois pour l'estimateur stratifié et pour l'estimateur par rapport combiné. En outre, pour l'estimateur stratifié et l'estimateur par rapport combiné, seule la méthode de stratification de DH a pu atteindre la limite de 15 % du volume total (six des 12 estimateurs stratifiés). Ces résultats démontrent que les méthodes de classification utilisées sont compatibles avec les estimateurs des totaux stratifiés et peuvent fournir aux utilisateurs l'occasion de développer des méthodes d'échantillonnage viables par opposition au recensement complet des usines.

[Traduit par la Rédaction]

Introduction

Periodic assessments of timber removals are an important component of the US Forest Service's national forest inventory and monitoring program. One method of assessment is the timber product output (TPO) survey, which is conducted under the direction of each US Forest Service research station's forest inventory and analysis (FIA) program. During a TPO survey, a complete canvass of all primary wood-using mills is conducted in a state. The primary measure of interest is roundwood receipts, which is the cubic foot volume of roundwood harvested in the state plus roundwood imported from other states. The data collected provide important information on the amount, county of harvest, and species composition of roundwood harvested within each state.

TPO designs vary by US Forest Service research station and sometimes even among states within a specific research station's purview. What is common to all is that a complete census is attempted by canvassing all primary mills in each state (T.G. Johnson, C.E. Keegan, R.J. Piva, and E.H. Wharton, personal communications (2005)). Complete canvasses can be problematic due to lack of response, limited monetary

Received 31 August 2011. Accepted 29 November 2011. Published at www.nrcresearchpress.com/cjfr on 21 February 2012.

J.P. Brown. USDA Forest Service, 241 Mercer Springs Road, Princeton, WV 24740, USA.
R.G. Oderwald. Department of Forest Resources and Environmental Conservation, Virginia Tech, 138 Cheatham Hall, Blacksburg, VA 24061, USA.

Corresponding author: John P. Brown (e-mail: jpbrown@fs.fed.us).

 doi:10.1139/X11-183

resources, lack of available personnel, and other reasons. Quality information can often be obtained more readily and more easily from sampling. Given past reliance on canvasses, estimation procedures require development, specifically procedures for a state's total receipts. Budgetary constraints are providing a strong impetus to save money by discontinuing censuses and moving to samples. New sampling methods are a clear need.

Previous research has not attempted sampling in TPO surveys. Some states have mill populations that are thought to be too variable for effective sampling, with population strata sizes that are not conducive to estimating sampling totals with adequate precision. The difficulty of sampling a small, heterogeneous population motivated this study. Sampling techniques that address variance reduction could solve this issue. Of secondary concern to sampling is that mill surveys need to be short in completion time per the United States Office of Management and Budget. Therefore, mill profile information that is readily available and potentially correlated with receipts (i.e., number of employees) could be useful in estimation. Methods to be considered include simple random samples, stratification, ratio estimators, and combined (stratified) ratio estimators. Stratified estimators are often chosen for use when stratification produces more homogenous strata. Ratio estimators are selected when there is a linearly related auxiliary variable. Stratification through cluster analysis was chosen due to the potential use of Ward's (1963) minimum variance method. The Dalenius-Hodges (DH) square root of the frequency method (Dalenius and Hodges 1959) also creates estimators with small variances.

The objectives of this study are to (*i*) develop stratifications for mills using the DH method and cluster analysis, (*ii*) employ stratified and ratio totals estimators to reduce variance with a concurrent reduction in confidence interval size, and (*iii*) determine if the sample calculated for the different estimators at the several levels of sampling error is feasible for the population studied.

This analysis used Georgia data to examine new methods for sampling mills. Georgia TPO reports have separate receipts data dating back to 1971 (Welch and Bellamy 1976). Older TPO reports for Georgia dating as early as 1937 (Spillers 1943) are available for roundwood production, which is the total of roundwood cut from a state and reported both within and outside that state (i.e., data collected in an outside state's TPO report.)

Methods

Survey data

TPO data were collected for the year 2001 from a 100% canvass of Georgia primary wood-using plants. These data were collected in 2002 by the Southern Research Station's FIA unit. Questionnaires were mailed with additional information sought by telephone or through personal contact as needed for completion (Johnson and Wells 2002). A total of 85 mills out of 170 responded. For this study, receipts totals by mill and the number of employees were of primary interest. Data were entered and checked by the Southern Research Stations's FIA unit and filed into a database.

Receipts totals, which were given in a variety of units by product type and species, were converted into cubic metre totals using conversion factors from Johnson and Wells (2002). Conversion factors not provided in that report were obtained from C. Steppleton in the Southern Research Station's FIA unit (personal communication (2006)). These conversion factors were then utilized to calculate the total cubic metres removed of each species at the mill level. Then the receipts for all species at a mill were summed for each mill.

General assumptions

All sample size estimates were calculated from the standpoint of a 95% confidence interval. An assumption was made that the sample sizes generated are sufficiently "large" such that the use of a z value of 2 was appropriate. The error bounds examined were 1%, 5%, 10%, 15%, 20%, and 25% of the estimated mean total (of all calculated totals using the several estimators). All formulas for estimators and associated variances come from Schaeffer et al. (2006).

Simple random sample estimators and sample size estimates

Simple random sampling (SRS) is one of the most basic methods for estimating population parameters such as the mean, variance, and total. This analysis uses SRS estimators as the basis for comparison with other more sophisticated estimators.

Estimate of SRS total

$$[1] \qquad \hat{\tau} = N\frac{\sum_{i=1}^{n} y_i}{n} = N\bar{y}$$

where N is the total number of mills in the state, y_i is the total receipts from mill i, n is the number of mills sampled, and

$$[2] \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Estimated variance of SRS total

$$[3] \qquad \hat{V}(\hat{\tau}) = \hat{V}(N\bar{y}) = N^2\hat{V}(\bar{y})$$

where

$$[4] \qquad \hat{V}(\bar{y}) = \frac{s^2}{n} \times \frac{N-n}{N}$$

$$[5] \qquad s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

where N, n, i, and y_i are as in eq. 1 and $\bar{y}$ is as in eq. 2.

Sample size to estimate SRS total

$$[6] \qquad n_{\text{samp}} = \frac{N\sigma^2}{(N-1)\frac{B^2}{4N^2} + \sigma^2}$$

where N is as for eq. 1, B is the error bound desired, and σ^2 is the population variance. This method uses s^2 (eq. 5) as an estimate of σ^2.

Strata selection by cluster analysis

Cluster analysis is a statistical technique that attempts to gather similar objects into a meaningful collection of groups. Data are collected on p variables for m objects. This analysis utilizes hierarchical clustering, one of many clustering techniques. The similarity matrix (or resemblance matrix) uses Euclidean distance. Agglomeration of clusters is based on the Ward's (1963) minimum variance linkage, with objects starting out singly and eventually being placed into groups based upon minimizing the sum of the squared distances weighted by cluster size (McGarigal et al. 2000). Here, two variables are included in the clustering: total receipts and number of employees. A total of 85 mills have data for both variables.

A cluster analysis creates cluster sizes from one cluster — all cases in one group — up to the n cluster — all groups single cases. Strict criteria for determining the number of clusters do not exist. However, simulation studies performed by Milligan and Cooper (1985) and Cooper and Milligan (1988) indicate that the pseudo F statistic, the pseudo t^2 statistic, and the cubic clustering criterion (CCC) all perform well in determining the best number of clusters. Cluster sizes can be chosen by creating a scree plot of the CCC versus the number of clusters and examining the figure for peaks (Sarle 1983). McGarigal et al. (2000) suggested a combined view of the CCC, pseudo F statistic, and pseudo t^2 statistic where large pseudo F statistics and increasing pseudo t^2 statistics from one cluster to the next are examined at local peaks on the CCC scree plot. This combination then suggests a stopping size for the number of clusters.

Strata selection by the DH cumulative square root of the frequency method

The second method chosen to generate a class variable for stratification was the DH cumulative square root of the frequency method (Dalenius and Hodges 1959).

The DH method is another method used to generate estimators with small variance. An example is provided in Table 1 using a population ranging from 0 to 1000, bin size $b = 100$, and three strata ($L = 3$). The DH method first bins the observations on the response variable and a frequency table is constructed. The interval between bins is chosen by the experimenter much like a standard histogram bin size is chosen. Once the frequencies for the bins are found, the square root of each frequency is taken. Then, a cumulative total is taken for the square roots of the frequencies. The experimenter will decide how many strata are desired. The cumulative sum is then divided by the number of strata, arriving at a rough category size c. Stratum boundaries are then found by picking the bins closest in cumulative frequency to $1 \cdot c$, $2 \cdot c$, ..., $(L - 1) \cdot c$. For consistency, L will be set to the number of strata suggested from the clustering analysis.

Stratified random sample estimators and sample size estimates

Strata were developed using results of both the cluster analysis and DH methods. Each method will have one or

Table 1. Example for the DH method.

Observations	y	(1) Bins Lower	(1) Bins Upper	(2) Frequency	(3) $\sqrt{\text{Frequency}}$		$\sqrt{\text{Frequency}}$	Subtotal	(6) Strata Lower	(6) Strata Upper
1	610	0	100	4	2.0		2.0		0	400
2	705	100	200	6	2.4		2.4		400	600
3	959	200	300	8	2.8		2.8		600	1000
4	127	300	400	15	3.9	(5) $c = 30.1/3 \sim 10$	3.9	11.1		
5	35	400	500	21	4.6		4.6			
6	581	500	600	18	4.2		4.2	8.8		
7	810	600	700	11	3.3		3.3			
8	225	700	800	9	3.0		3.0			
		800	900	5	2.2		2.2			
100	486	900	1000	3	1.7		1.7	10.2		
					(4) Total = 30.1					

Note: (1) Bin size 100; (2) frequencies calculated for bins; (3) calculate square feet; (4) calculate sum of square feet; (5) determine category size, $c \sim 10$; (6) find strata boundaries.

more group size solutions. Once initial groupings are determined from each method, summary statistics will be generated for all groups in a solution: minimum, maximum, mean, and standard deviation. These values will help determine a modified solution that finalizes the employee ranges on which groups will be stratified.

The modified solution ranges will be determined from the following two procedures for each group size solution. If there is no overlap in the range of the number of employees for a group with that of another within a solution, then the minimum and maximum numbers of employees for the considered group are used as draft bounds for that group. These draft bounds will be rounded to tens of employees. If there is overlap between two groups within a solution, the boundary division will be determined as follows. First, take the difference between the group 1 and group 2 means:

$$[7] \qquad d = \bar{y}_2 - \bar{y}_1$$

Second, calculate a weight for the group one mean that is the group 1 standard deviation (SD) divided by the sum of the SD of group 1 and the SD of group 2:

$$[8] \qquad w_1 = \frac{s_1}{s_1 + s_2}$$

Third, multiply the difference of the means by the weight for group 1 and add that value to the group 1 mean:

$$[9] \qquad \text{Upper}_1 = w_1 \times d + \bar{y}_1$$

This becomes the upper limit for group 1 and the lower limit for group 2. Fourth, adjust this limit to reasonable tens of employees values. Repeat for any other overlapping groups. This method weights the distance between the strata means and considers the overlap.

Calculating sample sizes for stratified totals is more involved than for a SRS. To determine the total sample size, an allocation strategy is required. The costs of obtaining a mill's information are expected to be nearly equal, yet differences in strata variances are evident, so Neyman allocation is employed to calculate the strata weights. These weights are then incorporated into the formula for the total sample size. Strata sample sizes are obtained from multiplying out the weights and the total size.

Estimate of the stratified population total

$$[10] \qquad \hat{\tau} = N\bar{y}_{\text{st}}$$

where N is as for eq. 1 and

$$[11] \qquad \bar{y}_{\text{st}} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i$$

where L is the number of strata, N_i is the number of sampling units in stratum i, and $\bar{y}_i$ is the estimated mean using eq. 2 for mills from stratum i.

Estimated variance of the stratified total

$$[12] \qquad \hat{V}(\hat{\tau}) = N^2 \hat{V}(\bar{y}_{\text{st}})$$

where N is as for eq. 1,

$$[13] \qquad \hat{V}(\bar{y}_{\text{st}}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(\frac{N_i - n_i}{N_i}\right)\frac{s_i^2}{n_i}$$

where N, i, L, and N_i are as in eq. 11 and n_i is the sample size for stratum i, and

$$[14] \qquad s_i^2 = \frac{\sum_{j=1}^{n_i} (y_j - \bar{y}_i)^2}{n_i - 1}$$

where y_j is the jth observation in stratum i and $\bar{y}_i$ is as in eq. 11.

Neyman allocation for determining stratum allocation weight

$$[15] \qquad w_i = \frac{N_i \sigma_i}{\sum_{k=i}^{L} N_k \sigma_k}$$

where w_i is the sample weight for stratum i, N_i (N_k) is the number of sampling units in stratum i (k), and σ_i (σ_k) is the population variance for stratum i (k). This method uses s_i (eq. 14) as an estimate of σ_i.

Sample size to estimate the total using stratified sampling

$$[16] \qquad n_{\text{samp}} = \frac{\sum_{i=1}^{L} N_i^2 \sigma_i^2 / w_i}{\frac{B^2}{4} + \sum_{i=1}^{L} N_i \sigma_i^2}$$

where N_i and w_i are as in eq. 15 and B is as in eq. 6. This method uses s_i^2 (eq. 14) as an estimate of σ_i^2.

Ratio estimators and sample size estimates

Ratio estimators use a well-correlated auxiliary variable as an aid to estimate either means or totals. Additionally, an estimate of the population ratio may be of interest. While potentially biased, this bias is negligible (of the order $1/n$) and disappears if the regression of y on x (variable of interest, auxiliary variable) is a straight line through the origin (Schaeffer et al. 2006).

Estimate of the population total

$$[17] \qquad \hat{\tau}_y = r\tau_x$$

where τ_x is the employee population total and

$$[18] \qquad r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

where n and y_i are as in eq. 1 and x_i is the number of employees at mill i.

Estimate of the population variance

$$[19] \quad \widehat{V}(\widehat{\tau}_y) = \tau_x^2 \widehat{V}(r)$$

where τ_x is as in eq. 17,

$$[20] \quad \widehat{V}(r) = \left(\frac{N-n}{N}\right)\left(\frac{1}{\mu_x^2}\right)\frac{s_r^2}{n}$$

where N and n are as in eq. 1 and μ_x is the employee population mean, and

$$[21] \quad s_r^2 = \frac{\sum_{i=1}^{n}(y_i - rx_i)^2}{n-1}$$

where y_i is as in eq. 1 and r and x_i are as in eq. 18.

Sample size to estimate the total using a ratio estimator

$$[22] \quad n_{\mathrm{samp}} = \frac{N\sigma^2}{\frac{B^2}{4N} + \sigma^2}$$

where N is as in eq. 1 and σ^2 and B are as in eq. 6. This method uses s_r^2 (eq. 21) as an estimate of σ^2.

Stratified ratio estimators (combined)

The two varieties of stratified ratio estimators are separate ratio estimators and combined ratio estimators (Schaeffer et al. 2006). Separate ratio estimators estimate r (sample ratio) within the strata and these strata estimates for r are then combined with weighted strata means and totals to arrive at population estimates for the mean or total. The combined ratio estimator estimates r from the stratified means of both the variable of interest and its auxillary variable. The combined estimate for r is then used with the weighted stratum means or totals to arrive at the desired population estimates.

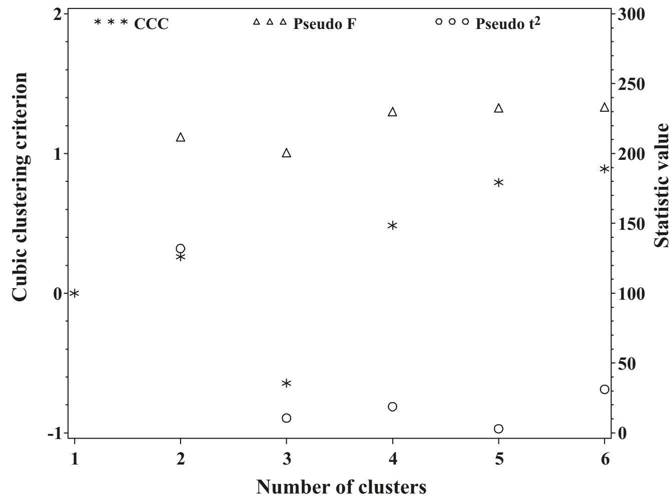
For small sample sizes within strata, it is recommended that the combined ratio estimator be used (Schaeffer et al. 2006). For larger sample sizes, the separate estimator generally provides narrower confidence intervals, especially when strata ratios differ markedly. This analysis will use combined ratio estimators due to smaller sample sizes. Sample sizes for the combined ratio estimator are calculated similarly to those of the stratified estimator (eq. 6) with the exception that s_{ri}^2 is used to estimate σ_i^2 (as opposed to using s_i^2 (eq. 5)).

Estimate of population total

$$[23] \quad \widehat{\tau}_{y\mathrm{C}} = \frac{\overline{y}_{\mathrm{st}}}{\overline{x}_{\mathrm{st}}}\tau_x$$

where $\overline{x}_{\mathrm{st}}$ is the stratified mean for mill employees using eq. 11, $\overline{y}_{\mathrm{st}}$ is as in eq. 11, and τ_x is as in eq. 17.

Fig. 1. CCC, pseudo F, and pseudo t^2 values for the cluster analysis.



Estimated variance of population total

$$[24] \quad \widehat{V}(\widehat{\tau}_{y\mathrm{C}}) = N^2\widehat{V}(\widehat{\mu}_{y\mathrm{C}})$$

where N is as in eq. 1,

$$[25] \quad \widehat{V}(\widehat{\mu}_{y\mathrm{C}}) = \sum_{i=1}^{L}\left(\frac{N_i}{N}\right)^2\left(\frac{N_i - n_i}{N_i}\right)\frac{s_{ri}^2}{n_i}$$

where L, N_i, and n_i are as in eq. 11,

$$[26] \quad s_{ri}^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - r_{\mathrm{C}}x_{ij})^2}{n_i - 1}$$

where x_{ij} is the number of employees for mill j from stratum i and y_{ij} is the cubic metres total for mill j from stratum i, and

$$[27] \quad r_{\mathrm{C}} = \frac{\overline{y}_{\mathrm{st}}}{\overline{x}_{\mathrm{st}}}$$

where $\overline{x}_{\mathrm{st}}$ and $\overline{y}_{\mathrm{st}}$ are as in eq. 23.

Sample size to estimate the total using a ratio estimator

Calculations utilize eq. 16 with the exception that s_{ri}^2 is used to estimate σ_i^2 (as opposed to using s_i^2 (eq. 14)).

Comparison of estimators

Totals for each of the estimators were calculated using the 85 responding mills. Variances for each of the estimators were compared using the estimated relative efficiency

$$[28] \quad \widehat{\mathrm{RE}}\left(\frac{E_1}{E_2}\right) = \frac{\widehat{V}(E_1)}{\widehat{V}(E_2)}$$

where E_1 and E_2 are estimators and $\widehat{V}(E_1)$ and $\widehat{V}(E_2)$ are the respective estimated variances.

Results

Cluster analysis stratification and estimators

Values for the CCC indicated a local peak at two clusters with a plateau approached at six clusters (Fig. 1). Cluster

Table 2. Cluster analysis group statistics and solution ranges based on number of employees.

No. of groups	Group	Cluster solution Minimum	Maximum	Mean	SD	Modified cluster solution Minimum	Maximum
2	1	1	381	73.22	91.48	1	400
	2	400	907	691.00	159.71	401	907
4	1	1	66	20.90	17.15	1	40
	2	37	381	168.15	95.21	41	330
	3	400	907	657.83	188.50	331	730
	4	692	800	757.33	57.46	731	907
6/5	1	1	66	20.90	17.15	1	50
	2	37	169	113.05	37.45	51	200
	3	240	381	299.12	47.13	201	360
	4	400	826	597.50	183.44	361	720
	5	692	800	757.33	57.46	720	907
	6	650	907	778.50	181.73		

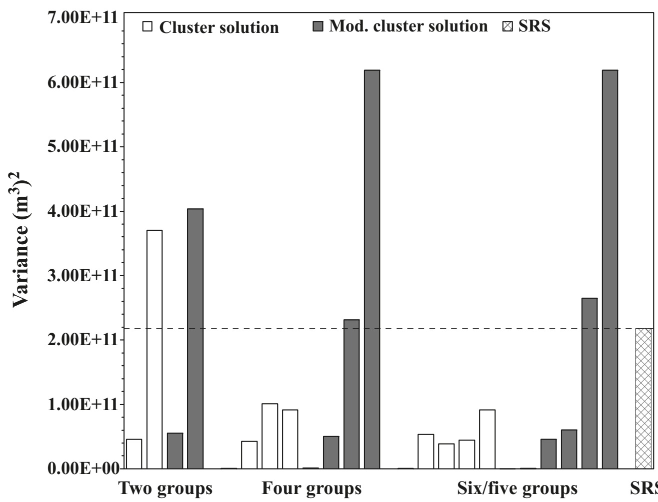
numbers of larger than six were rejected due to insufficient potential sample sizes within clusters. The pseudo t^2 and pseudo F statistics suggested two, four, and six groups; the pseudo t^2 statistic showed local peaks at four and six, while the pseudo F had higher values at two, four, five, and six. Therefore, the data were divided into groups of two, four, and six by liberally considering all criteria.

The stratification of mills into groups when sampling will ultimately be based solely on the number of employees. While the receipts data are a useful auxiliary variable, they are unknown before sampling (particularly for new mills), whereas the employee data are easily obtainable. This constraint necessitated examining the cluster solution and developing a stratification that modified the original cluster analysis group characteristics. Prior mill receipts data are readily obtainable, as most states have previous TPO data. Future mill stratifications can therefore be conducted on a range of employee values that potentially have incorporated some information from the previous receipts relationship. In the event that prior receipt data are not available, a pilot study could be conducted.

The employee ranges obtained from modifying the original cluster solution are presented in Table 2. Note that for the six-group cluster solution, a sixth group was not obtainable based on the distribution of groups from the initial clustering. There is separation in two dimensions, but not reasonable separation along the employee dimension. Therefore, the modified cluster solution has five groups as the largest group size examined and is referred to as the six/five-group solution.

The sample variances for the various groups formed from clustering into two, four, and six/five groups are presented in Fig. 2. Several groups had variances smaller than 1.00×10^{10}. These are the groups shown as gaps in the figure. The majority of the groups had variances less than that for the SRS variance, with only five out of the 23 groups exceeding this value.

Fig. 2. Sample variances for the cluster, modified cluster, and SRS groupings. Gaps indicate strata with variances too small to plot.



DH method stratification and estimators

The divisions that led to the groupings by the DH method are presented in Table 3. Bin sizes of 1 million ft^3 were used to develop groupings. Once the cumulative square root of the frequencies (CSQRTF) was calculated, group divisions were made.

As an example, consider the four-group solution. The CSQRTF is 37.72, which is divided by 4. This results in an interval of around 9.4. Divisions are then made at 9.38, 18.77, and 27.65, creating four categories of mill sizes. There are 47, 13, 12, and 13 mills in each category, respectively. This grouping provides a reasonable number of mills per stratum (and it does so for the other two group sizes as well).

The employee ranges obtained from modifying the original cluster solution are presented in Table 4. There were no issues with overlapping groups, as this method used one variable to separate the groups.

The sample variances for the two-, four-, and six-group stratifications for both the DH solution and the modified DH solution along with the SRS variance are presented in Fig. 3. Again, several of the variances near or below the value 1.0×10^{10} do not appear in the figure. Similar to the cluster solutions, the majority of the variances are below the value for the SRS solution, with only six out of 24 groups having variances greater than that of the SRS solution.

Table 3. DH method cutoff values for two-, four-, and six-group sizes.

Range					Subgroup total		
Lower (1000s m^3)	Upper (1000s m^3)	Frequency	√Frequency	Cumulative √Frequency	2	4	6
0	28	35	5.92	5.92			35
28	57	12	3.46	9.38		47	
57	85	3	1.73	11.11			
85	113	1	1.00	12.11			16
142	170	2	1.41	13.53			
170	198	2	1.41	14.94			
198	227	2	1.41	16.35			
227	255	2	1.41	17.77			
283	311	1	1.00	18.77	60	13	9
340	368	3	1.73	20.50			
396	425	1	1.00	21.50			
425	453	1	1.00	22.50			
453	481	2	1.41	23.92			
481	510	1	1.00	24.92			8
538	566	3	1.73	26.65			
566	595	1	1.00	27.65		12	
595	623	2	1.41	29.06			
680	708	2	1.41	30.48			
765	793	2	1.41	31.89			10
991	1019	2	1.41	33.30			
1359	1387	2	1.41	34.72			
1812	1841	1	1.00	35.72			
1982	2010	1	1.00	36.72			
2407	2435	1	1.00	37.72	25	13	7

Note: Serrated line indicates two-group cutoff, dashed lines four-group cutoffs, and dotted lines six-group cutoffs.

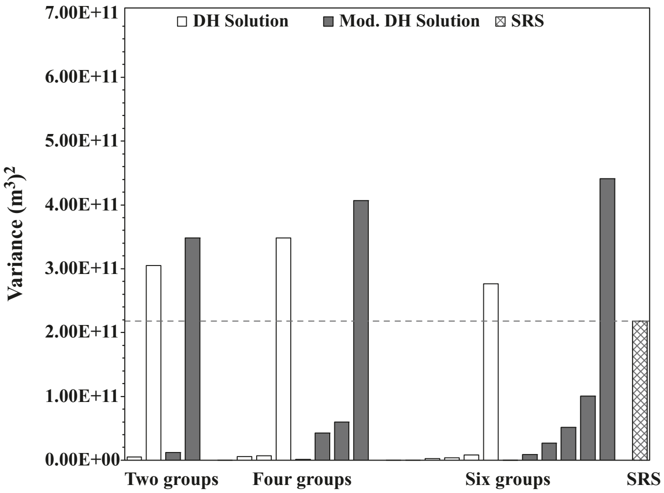
Table 4. DH method group statistics and solution ranges based on number of employees.

		DH solution				Modified DH solution	
No. of groups	Group	Minimum	Maximum	Mean	SD	Minimum	Maximum
2	1	1	285	40.30	54.12	1	100
	2	70	907	374.64	270.58	101	907
4	1	1	103	21.04	19.59	1	40
	2	17	285	109.92	78.92	41	150
	3	70	381	193.33	103.43	151	290
	4	150	907	542.00	270.94	291	907
6	1	1	40	13.51	10.95	1	20
	2	14	285	58.69	64.17	21	90
	3	37	240	111.78	61.71	91	130
	4	70	381	167.00	108.66	131	220
	5	134	826	328.70	231.72	221	530
	6	400	907	677.58	174.59	531	907

Comparison of the modified cluster and DH classifications

The employee ranges obtained for the groupings in both the DH and cluster classifications differed when considering the same number of groups. The two-group solution for the cluster analysis was divided nearly in the middle of the em-

Fig. 3. Sample variances for the DH, modified DH, and SRS groupings. Gaps indicate strata with variances too small to plot.



ployee range, while the DH method had a much lower break at about 100 employees (Fig. 4). For the larger number of groups, four and six/five, the cluster solution exhibited more even employee ranges in its group sizes. The DH solution had variable employee ranges among the strata for both the four- and six-strata solutions.

The modified cluster solution generally has group variances higher than those of the cluster solution, although it should be noted that the groups do not always comprise the same mills. The groups having the greatest number of employees tended to have the highest variances, and those of the modified cluster solution were larger than those of the cluster solution. The two-group solution had the greatest similarity between the cluster solution and the modified cluster solution.

Relative efficiencies of estimators

The estimated relative efficiencies of the estimators are included in Table 5. Efficiencies greater than 1 suggest that the estimator in the row position of the table is more efficient that the estimator in the column. However, it should be noted that these are calculated variances and not theoretical variances, and any differences may be due to random chance.

Many comparisons are presented in Table 5. The number of groups is the foremost factor to consider, and these comparisons are possible at many levels: within a stratification method and type of means analysis (e.g., DH method, stratified mean analysis), across stratification methods (e.g., cluster stratification versus DH stratification), and across means methods (cluster stratified mean versus combined ratio (cluster)).

First, consider the comparisons for group size within a stratification method and type of means analysis. These are found in the 3 × 3 blocks (dark gray blocks) along the main diagonal of Table 5. For each of the four combinations produced from crossing stratification method and means method (e.g., combined ratio (cluster)), relative efficiency improves for increasing numbers of groups. That is, six groups are more efficient than four groups, which are more efficient than two groups, and these relationships are transitive.

Next, group sizes can be compared within the stratified mean method. When comparing the cluster method with the DH method within stratified means, the relative efficiencies range from 0.8 to 1.3 (Table 5, main diagonals within diagonal striped blocks). A similar range in relative efficiencies, 0.7–1.4, results for the cluster versus DH comparison within the combined ratio method (Table 5, main diagonals within light gray blocks).

The final group size comparison considered is across means methods. Comparing the cluster analysis method across means methods (Table 5, main diagonals, horizontal striped blocks), relative efficiency ranges from 0.4 to 2.4. For the DH method, the range across means methods is 0.4–2.3 (Table 5, main diagonals within vertical striped blocks).

The last set of comparisons is each individual estimator versus the SRS (Table 5, first column). All estimators have greater than 1 relative efficiency versus the SRS. Additionally, all specific stratification and means method types show increasing efficiency versus the SRS with increasing group size.

Confidence intervals for totals

The 95% confidence intervals for the totals of the various methods range from a lower limit of 1.07 billion ft^3 to a high of 2.10 billion ft^3 (Fig. 5). Confidence interval ranges are narrowest for the combined ratio estimators, while the cluster and DH stratified totals are slightly broader. The widest interval (1.26, 2.10) billion ft^3 is that of the SRS.

Estimated sample sizes for SRS and ratio totals

SRS sample sizes ranged downwards from 170 to 82 to achieve bounds of 1%–25% of the total state receipts (volume of logs) (Fig. 6). (Strictly speaking, the 170 sample size would force a total census.) Ratio sample sizes ranged from 169 down to 44 for the same bounds. Ratio sample sizes were less than SRS sample sizes for all considered bounds.

Estimated sample sizes for stratified totals

Stratum weights were calculated for all strata in each of the stratification methods using the class sample standard deviation estimates and the number of sampling units available for each stratum (Table 6). One important factor here is the breakdown of stratum sampling units into the classes generated from the two stratification methods. The cluster analysis solutions have, in all cases, one stratum with 10 or fewer mills. Further, within both the four and six/five classifications, there is a stratum with only four sampling units available.

The estimated strata standard deviations obtained from the two stratification methods differ among classes almost exclusively by no more than 1 order of magnitude. The DH 6 solution is the only one where this is not true and sample standard deviations differ by 2 orders of magnitude.

Weights for both stratification methods range from 0.05 to 0.86, with most weights ranging from 0.10 to 0.25. The Cluster 2 solution Stratum 1 weight is the greatest and the DH 6 Stratum 1 weight is the smallest. The Cluster 4 Stratum 2 and DH 4 Stratum 4 weights, at 0.55 and 0.43, are the most dissimilar within all of the solutions (Table 6).

These weights are then used to calculate the stratum sample sizes needed for each method from the calculated total sample size necessary for each bound (Fig. 7). Due to a finite

Fig. 4. Comparison of the modified cluster and modified DH classifications.

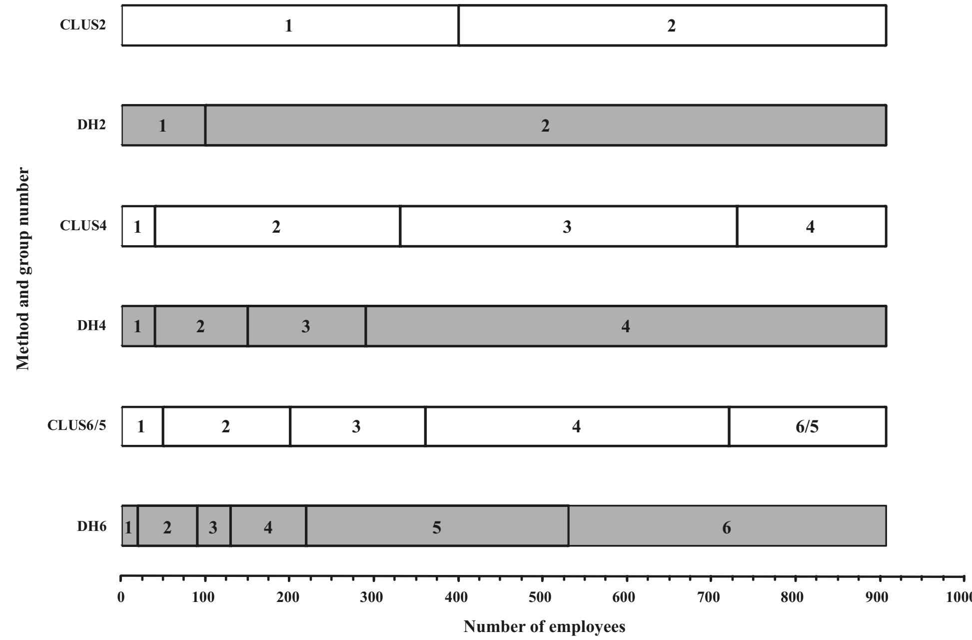


Table 5. Estimated relative efficiencies (RE) for all estimators.

RE (row/cols)	SRS	Ratio	Clus. Stratification 2	Clus. Stratification 4	Clus. Stratification 6	DH Stratification 2	DH Stratification 4	DH Stratification 6	Comb. Ratio 2 (Clus.)	Comb. Ratio 4 (Clus.)	Comb. Ratio 6 (Clus.)	Comb. Ratio 2 (DH)	Comb. Ratio 4 (DH)	Comb. Ratio 6 (DH)
SRS	1.0	0.2	0.3	0.1	0.1	0.4	0.2	0.1	0.1	0.1	0.1	0.2	0.1	0.1
Ratio	**4.4**	1.0	**1.2**	0.6	0.4	**1.6**	0.8	0.4	0.5	0.4	0.3	0.7	0.5	0.3
Clus. Stratification 2	**3.5**	0.8	1.0	0.5	0.4	**1.3**	0.6	0.3	0.4	0.4	0.2	0.6	0.4	0.2
Clus. Stratification 4	**7.0**	**1.6**	**2.0**	1.0	0.7	**2.6**	**1.2**	0.6	0.8	0.7	0.5	**1.1**	0.8	0.4
Clus. Stratification 6	**10.0**	**2.3**	**2.8**	**1.4**	1.0	**3.7**	**1.8**	0.8	**1.2**	**1.0**	0.6	**1.6**	**1.2**	0.6
DH Stratification 2	**2.7**	0.6	0.8	0.4	0.3	1.0	0.5	0.2	0.3	0.3	0.2	0.4	0.3	0.2
DH Stratification 4	**5.6**	**1.3**	**1.6**	0.8	0.6	**2.1**	1.0	0.5	0.7	0.6	0.4	0.9	0.7	0.4
DH Stratification 6	**12.2**	**2.8**	**3.5**	**1.8**	**1.2**	**4.5**	**2.2**	1.0	**1.4**	**1.2**	0.8	**1.9**	**1.4**	0.8
Comb. Ratio 2 (Clus.)	**8.5**	**2.0**	**2.4**	**1.2**	0.9	**3.2**	**1.5**	0.7	1.0	0.9	0.6	**1.4**	1.0	0.5
Comb. Ratio 4 (Clus.)	**9.9**	**2.3**	**2.8**	**1.4**	1.0	**3.7**	**1.7**	0.8	**1.2**	1.0	0.6	**1.6**	**1.1**	**0.6**
Comb. Ratio 6 (Clus.)	**15.4**	**3.5**	**4.4**	**2.2**	**1.5**	**5.7**	**2.7**	**1.3**	**1.8**	**1.6**	1.0	**2.5**	**1.8**	1.0
Comb. Ratio 2 (DH)	**6.3**	**1.4**	**1.8**	0.9	0.6	**2.3**	**1.1**	0.5	0.7	0.6	0.4	1.0	0.7	0.4
Comb. Ratio 4 (DH)	**8.6**	**2.0**	**2.4**	**1.2**	0.9	**3.2**	**1.5**	0.7	**1.0**	0.9	0.6	**1.4**	1.0	0.5
Comb. Ratio 6 (DH)	**16.0**	**3.7**	**4.5**	**2.3**	**1.6**	**5.9**	**2.8**	**1.3**	**1.9**	**1.6**	**1.0**	**2.6**	**1.9**	1.0

Note: Bolded RE indicates that the row estimator is more efficient than the column estimator.

Fig. 5. Estimates for total receipts by all methods.

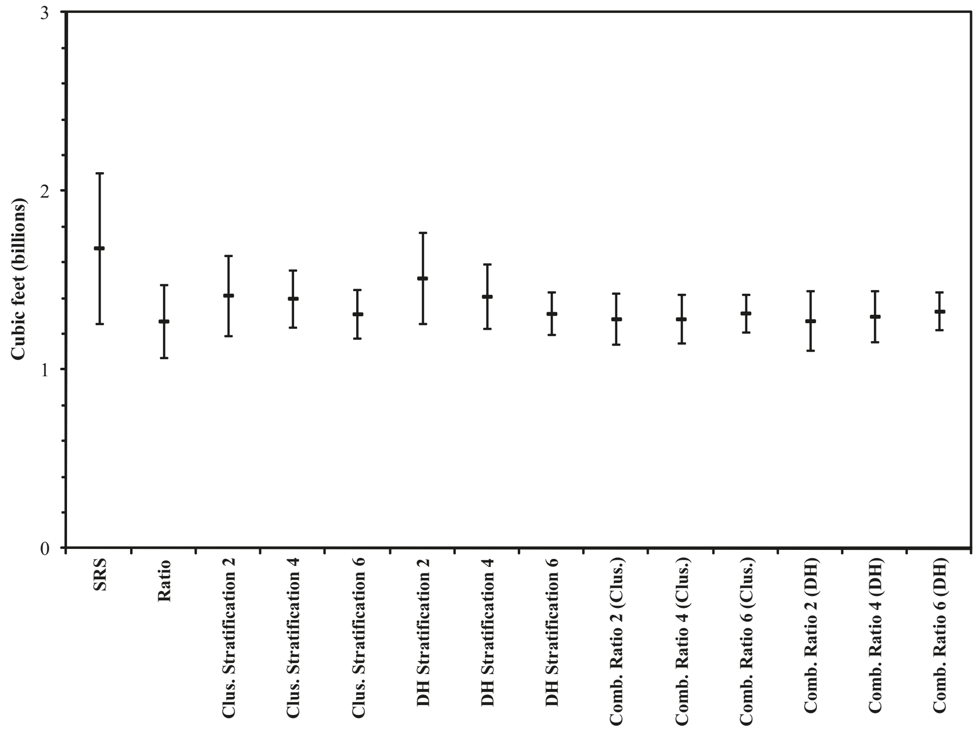
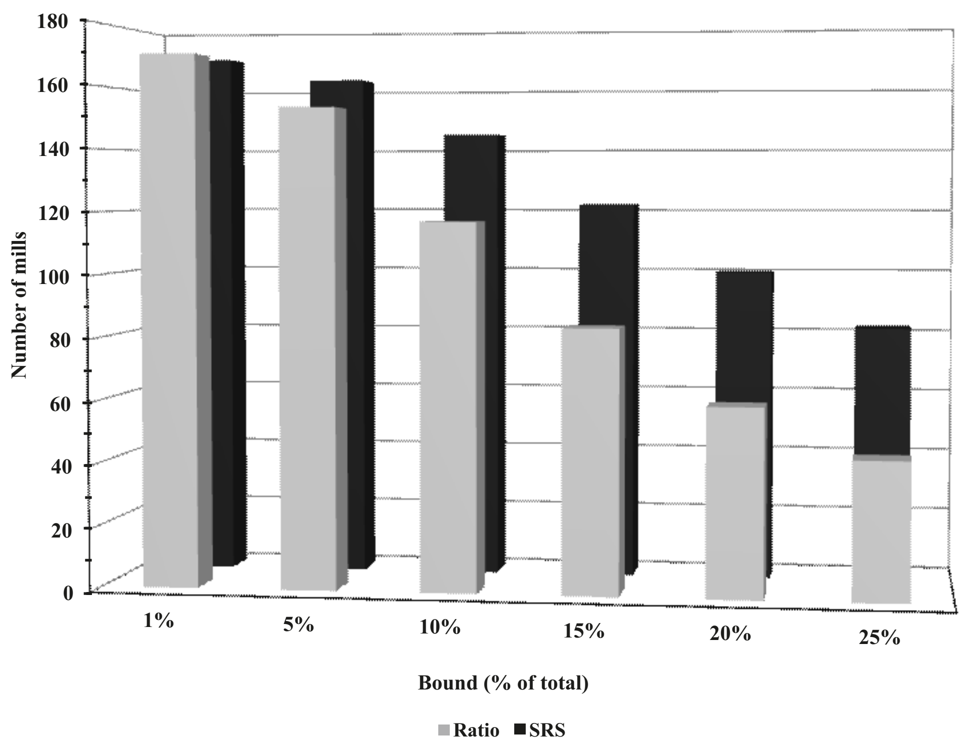


Fig. 6. Ratio and SRS estimated sample sizes for state totals from a population of 170 mills.



population, several stratum sample size estimates exceeded the total number of sampling units in the stratum and are noted in bold in Table 6 and as stippled columns in Fig. 7.

Sample size estimates are highest at 150 for the 1% bound of the cluster stratification two-group solution. The lowest sample size needed is 14 for the 25% bound on the DH stratification six-group solution. Sample sizes decrease, as expected, as bound size increases and also decrease as the number of groups used to estimate the total increases (Fig. 7).

Table 6. Stratum sample sizes for state stratified totals.

	Stratum	N_i	SD	Weight	Sample size for several bounds					
					1%	5%	10%	15%	20%	25%
Cluster 2	1	160	8.28E+06	0.86	128	120	98	76	58	44
	2	10	2.24E+07	0.14	**22**	**20**	**17**	**13**	10	8
Cluster 4	1	91	1.08E+06	0.11	8	8	6	4	3	2
	2	63	7.90E+06	0.55	43	39	30	21	15	11
	3	12	1.70E+07	0.22	**17**	**16**	**12**	9	6	5
	4	4	2.78E+07	0.12	**10**	**9**	**7**	**5**	3	3
Cluster 6/5	1	101	1.07E+06	0.13	9	8	6	4	3	2
	2	46	7.56E+06	0.43	30	27	20	14	10	7
	3	11	8.69E+06	0.12	8	7	5	4	3	2
	4	8	1.82E+07	0.18	**13**	**11**	**8**	6	4	3
	5	4	2.78E+07	0.14	**10**	**9**	**6**	**5**	3	2
DH 2	1	123	3.84E+06	0.33	31	29	25	21	17	13
	2	47	2.08E+07	0.67	**64**	**61**	**53**	43	34	27
DH 4	1	91	1.08E+06	0.10	7	7	5	4	3	2
	2	49	7.26E+06	0.36	27	24	19	15	11	8
	3	11	8.64E+06	0.10	7	7	5	4	3	2
	4	19	2.25E+07	0.44	**32**	**29**	**23**	18	13	10
DH 6	1	66	5.53E+05	0.05	3	3	2	1	1	1
	2	50	3.28E+06	0.22	16	14	10	7	5	3
	3	16	5.76E+06	0.12	9	8	5	4	3	2
	4	15	8.00E+06	0.16	11	10	7	5	3	2
	5	16	1.12E+07	0.24	**17**	15	11	7	5	4
	6	7	2.34E+07	0.22	**16**	**14**	**10**	7	5	3

Note: Values in bold indicate that the stratum sample size exceeds the stratum size.

Estimated sample sizes for combined ratio totals

The division of sampling units is the same for each stratum here, as was noted previously (e.g., the cluster analysis classification into four groups produces the same sample frame). What differs between Table 6 and Table 7 are the estimates for each group's variance.

The estimated stratum standard deviations vary from 1 to 2 orders of magnitude within stratification solutions. Here, both the Cluster 4 and DH 6 solutions have the widest ranges in estimated standard deviations (Table 7).

Weights range from a low of 0.05 to 0.78 for the DH 6 Stratum 1 and Cluster 2 Stratum 1 strata, respectively. Weights are again in the 0.10–0.25 range, with some minor deviations from those weights as happened in the stratifications shown in Table 6.

Sample sizes were again highest at 117 for the cluster stratification into two groups (Fig. 8). The lowest sample size of 12 was for the DH stratification into six groups. The weights and stratum sample sizes calculated to estimate the combined ratio totals also result in several strata sample sizes that are unattainable (Table 7, bold values).

Discussion

Classification methods group ranges

There were clear differences between the employee ranges for each group obtained by the cluster method and those obtained by the DH method. There were many groups with narrow ranges for the DH method, whereas there is more uniformity of employee ranges for the cluster solutions within specific group size solutions, e.g., CLUS2 (Fig. 4). This is unexpected, as it was thought that the DH method might lead to more uniform groupings due to the initial binning of observations.

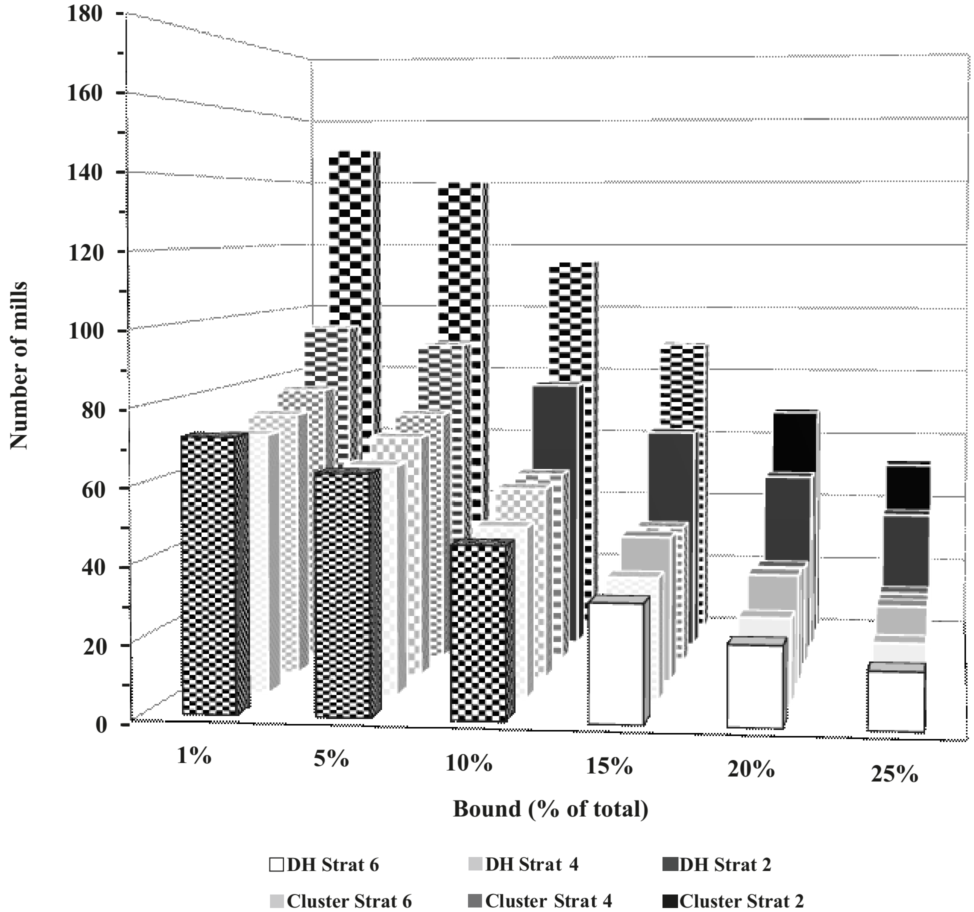
This difference in employee range sizes within the DH solution also leads to more groups for mills with fewer employees. That is not unexpected because the method naturally concentrates high frequencies into narrow range categories. Stated differently, the CSQRTF is divided into a set number of intervals, with each interval a fixed fraction of the total CSQRTF. Increasing the number of similar observations for the variable of interest will force the determined range to be narrower. That stands to be a useful technique if the variable of interest is strongly correlated with the auxiliary variable used to classify into strata.

Group sizes

There is a clear increase in estimator efficiency with increasing group size within specific stratification and means methods (Table 5, dark gray cells). The stratifications by both the cluster and DH methods allow for a decrease in variability within the groups formed by the methods. The $\widehat{RE}$ (six groups/two groups) was greatest in all instances, with a high of 4.5 calculated for the DH stratification. For the State of Georgia, stratification into four or six groups appears to be a reasonable approach to reducing estimator variance.

Group size is an important factor in reducing the number of samples needed to achieve the several bounds examined (Figs. 2 and 3). All of the cluster method by means method combinations show a decrease in the necessary sample sizes as the number of groups employed in the stratification increases.

Fig. 7. Stratified estimated sample sizes for state totals from a population of 170 mills. Stippled columns indicate that one or more stratum sample sizes exceed the stratum for that particular method.



Stratification methods

Comparison of the stratification methods for similar group sizes can be considered within a means method, i.e., stratified or ratio estimator. For these data, neither the cluster nor the DH method appears to be more efficient (Table 5, diagonals within diagonal striped and light gray cells). The relative efficiencies for these comparisons are close to 1. Therefore, neither method is recommended over the other based on this measure. However, the stratification itself is an improvement in efficiency over the SRS estimator and in most cases over the simple ratio estimator.

There were few differences in sample size estimates for the two stratification methods (cluster and DH). The differences were nearly immeasurable for the four- and six-group solutions within any means method. However, there were somewhat larger differences when comparing sample size estimates for the DH two-group solutions with the cluster solutions for the stratified total (Fig. 7). Differences for these comparisons occurred again for the combined ratio sample size estimates (Fig. 8). These differences in the two-group solution may be a result of only having two groups and that the cluster solution had an unbalanced distribution of sampling units within its two strata: $N_1 = 170$ and $N_2 = 10$.

Totals methods

By holding the stratification method and group size constant, we can examine the effect of the totals method. Here, there appears to be a clear advantage to using a combined ratio estimator versus a regular stratified mean (Table 5, diagonals within horizontal striped cells and vertical striped cells). Relative efficiencies range upward from 1.3, thereby indicating that the additional information contained within the employee numbers is providing smaller variances on the estimate of the total. In addition, the much simpler comparison for the $\widehat{\mathrm{RE}}$ (ratio/SRS) yields an efficiency of 4.4 (Table 5), further evidence that the number of employees is a useful auxiliary variable.

There is a marked reduction in the number of samples needed for a fixed group size and stratification method when switching from a stratified total to a combined ratio total in nearly all cases (Figs. 7 and 8). Giving specific consideration to the fact that some bound sizes produce unattainable stratum sample sizes, at least 20% more samples are needed by the ordinary stratification total to achieve the same bound as the combined ratio total.

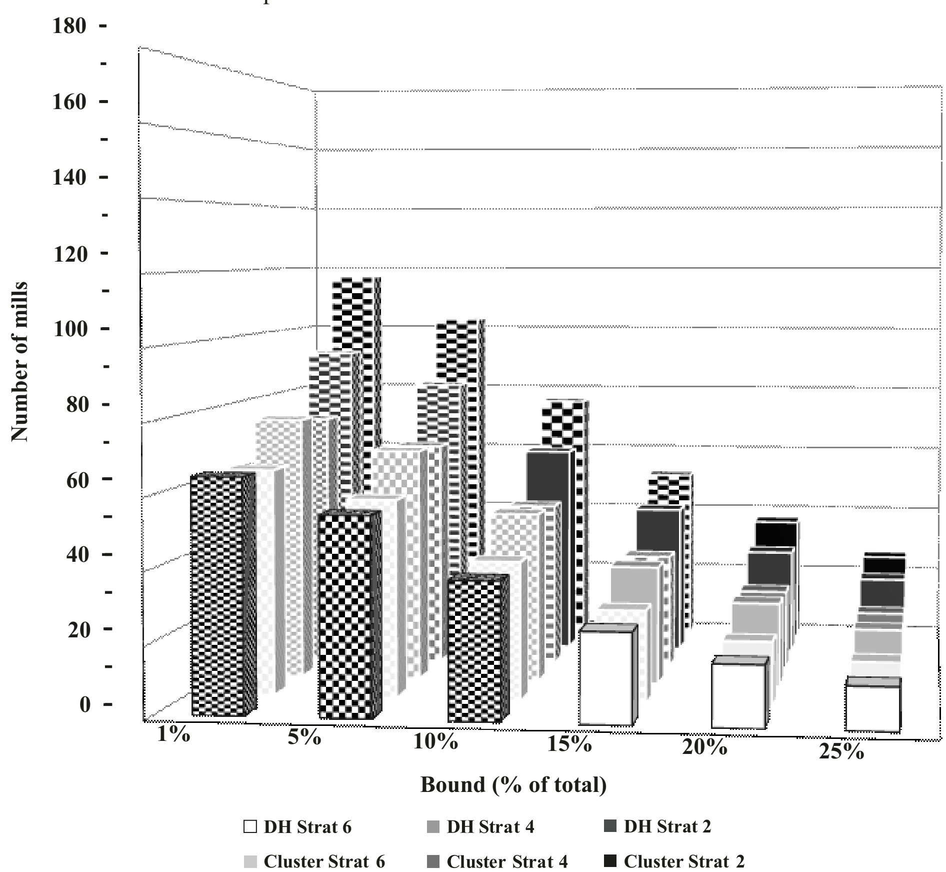
The DH 2 method using combined ratio estimators produces the only sampling strategy where a 10% bound on the total is attainable. Further, the DH method for the combined ratio estimator is the only method by which 15% bounds on the total are attainable (all group size solutions). All methods can be used if a 20% bound is an acceptable range for the total.

Table 7. Stratum sample sizes for state combined ratio totals.

					Sample size for several bounds					
	Stratum	N_i	SD	Weight	1%	5%	10%	15%	20%	25%
Cluster 2	1	160	4.95E+06	0.78	92	81	60	41	29	20
	2	10	2.19E+07	0.22	**25**	**22**	**16**	**11**	8	6
Cluster 4	1	91	9.78E+05	0.11	8	7	5	4	3	2
	2	63	7.03E+06	0.55	41	36	26	18	13	9
	3	12	1.27E+07	0.19	**14**	**13**	9	6	4	3
	4	4	2.98E+07	0.15	**11**	**10**	**7**	**5**	3	3
Cluster 6/5	1	101	1.02E+06	0.15	10	8	6	4	3	2
	2	46	6.09E+06	0.40	26	23	16	11	7	5
	3	11	7.31E+06	0.12	7	7	5	3	2	2
	4	8	1.37E+07	0.16	**10**	**9**	6	4	3	2
	5	4	2.98E+07	0.17	**11**	**10**	**7**	**4**	3	2
DH 2	1	123	2.45E+06	0.32	30	26	20	14	10	7
	2	47	1.37E+07	0.68	**63**	**57**	42	30	21	16
DH 4	1	91	9.81E+05	0.11	8	7	5	4	3	2
	2	49	5.56E+06	0.34	25	22	16	11	8	6
	3	11	1.09E+07	0.15	11	10	7	5	4	3
	4	19	1.74E+07	0.41	**31**	**27**	**20**	14	10	7
DH 6	1	66	4.88E+05	0.05	3	3	2	1	1	1
	2	50	2.51E+06	0.19	12	10	7	5	3	2
	3	16	5.59E+06	0.14	9	7	5	3	2	2
	4	15	7.97E+06	0.18	12	10	7	5	3	2
	5	16	8.02E+06	0.19	**13**	11	7	5	3	2
	6	7	2.37E+07	0.25	**16**	**14**	**10**	6	4	3

Note: Values in bold indicate that the stratum sample size exceeds the stratum size.

Fig. 8. Combined ratio estimated sample sizes for state totals from a population of 170 mills. Stippled columns indicate that one or more stratum sample sizes exceed the stratum for that particular method.



Conclusion

New approaches to conducting TPO studies demonstrate that statistical estimation procedures can be of benefit to generating estimates for a state's mill receipts. Results from this analysis indicate that stratification, coupled with ratio estimators based on employee numbers at mills, can greatly reduce the variance of estimated totals. These reductions will allow for an expected reduction in the number of samples needed to achieve desired bounds on a state's receipts total.

Stratification and the use of auxiliary information are demonstrably useful techniques in achieving reasonable bounds on a state's total receipts data. Slightly better results are produced using the DH stratification method, allowing for bounds of 15% of the estimated total for both means methods, as opposed to the cluster method attaining only the 20% level. The use of a combined ratio total estimator allows for a significant reduction in the necessary sample size. For these data, the easily obtained value for the total number of employees provides a much greater benefit through stratification versus using a SRS estimator. Given that most states have prior TPO data recorded, future sampling plans now have workable options should bounds be within acceptable tolerances. Even in cases where there has not been a prior TPO study, a pilot study could be conducted to perform necessary stratification. In times of tighter budgets, there is considerable value in sampling, as costs are traditionally lower than costs for a census.

Acknowledgements

We would like to thank Tony Johnson and Carolynn Steppleton, two members of the Southern Research Station FIA unit, for providing the data used in this work. Tony was involved in early discussions on how to improve TPO studies and he suggested the data used herein. Carolynn was extremely helpful, gracious, and patient in helping transfer and explain the data. Her assistance was invaluable.

References

Cooper, M.C., and Milligan, G.W. 1988. The effect of measurement error on determining the number of clusters in cluster analysis. *In* Data, expert knowledge and decisions. *Edited by* W. Gaul and M. Schader. Springer-Verlag, Heidelberg. pp. 319–328.

Dalenius, T., and Hodges, J.L. 1959. Minimum variance stratification. J. Am. Stat. Assoc. **54**(285): 88–101. doi:10.2307/2282141.

Johnson, T.G., and Wells, J.L. 2002. Georgia's timber industry — an assessment of timber product output and use, 1999. U.S. For. Serv. Resour. Bull. SRS68.

McGarigal, K., Cushman, S., and Stafford, S. 2000. Multivariate statistics for wildlife and ecology research. Springer-Verlag, New York.

Milligan, G.W., and Cooper, M.C. 1985. An examination of procedures for determining the number of clusters in a data set. Psychometrika, **50**(2): 159–179. doi:10.1007/BF02294245.

Sarle, W. 1983. Cubic clustering criterion. SAS® Tech. Rep. A-108. SAS® Institute, Inc., Cary, N.C.

Schaeffer, R.L., Mendenhall, W., and Ott, R.L. 2006. Elementary survey sampling. 6th ed. Thomson Brooks/Cole, Belmont, Calif.

Spillers, A.R. 1943. Georgia forest resources and industries. U.S. For. Serv. Misc. Publ. 501.

Ward, J.H., Jr. 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301): 236–244. doi:10.2307/2282967.

Welch, R. L.; Bellamy, T. R. 1976. Changes in output of industrial timber products in Georgia, 1971–1974. U.S. For. Serv. Resour. Bull. SE-RB-36.