

---

## Forest Inventory Predictions From Individual Tree Crowns: Regression Modeling Within a Sample Framework

James W. Flewelling<sup>1</sup>

**Abstract.**—Remotely sensed data can be used to make digital maps showing individual tree crowns (ITC) for entire forests. Attributes of the ITCs may include area, shape, height, and color. The crown map is sampled in a way that provides an unbiased linkage between ITCs and identifiable trees measured on the ground. Methods of avoiding edge bias are given. In an example from a forest of young southern pine, the forest is delineated into several thousand stands. Forty stands are sampled, each with two 0.12 acre plots. The resultant estimator of a volume surrogate, tree basal area times height summed over all trees, is unbiased and has a 90-percent confidence interval of  $\pm 4.1$  percent. The root mean square errors for basal area and the volume surrogate at the stand level are estimated at 9.7 percent and 12.8 percent, respectively. That precision in basal area for individual stands is approximately the same as would have been achieved by ground sampling with ten 0.12 acre plots in each stand, making no use of the remotely sensed data.

### Introduction

Methods to obtain and interpret high spatial resolution digital imagery are evolving rapidly. Forest inventory systems are increasingly making use of such imagery with the goals of improved precision, lower field costs, and faster completion of large inventories. A logical step towards greater precision in inventory is to identify and describe the visible individual tree crowns (ITCs). Gougeon and Leckie (2003) provided an overview of the methodologies. Until recently most crown segmentation was based on digital color and infrared (CIR) photography. CIR photography is now complemented by

“fused” data from airborne Light Detection and Ranging (LIDAR), making it practical to obtain and analyze high resolution CIR photography data and LIDAR data for entire forests. Gougeon and Leckie (2003) opined that the use of ITC techniques in forest inventory would evolve slowly, starting with a few niche applications. Næsset *et al.* (2004) reported that LIDAR is starting to be used in large-area forest inventories, although applications involving the estimation of individual tree properties are still in research mode. Næsset and Nelson (2009) presented the status of an ongoing operational-scale project in Norway.

### Sampling Within a Stand

The overall sample design presented here uses stands as the primary sample units. Data from stands selected for sampling are used to develop relationships that will be applied in many stands; that strategy is the primary focus of this article. The within-stand sampling methods for multistand relationships are the same as would be proposed for sampling within individual stands, one stand at a time. That single-stand case is addressed first.

### Sample Frame

The sample frame for a stand is the map of ITCs, together with the remotely sensed characteristics assigned to each ITC. The map is assumed to have accurate scale and stand boundaries. Each ITC is required to have a specific spatial extent and a specifically identified center. ITCs may not overlap one another, but may overlap stand boundaries. Only the ITCs whose centers are within the stand boundary are included in the sample frame.

Other attributes associated with the ITCs may include any statistics derived from remote sensing. If CIR photography data are available, means or other summary statistics for the color intensities of the pixels associated with each ITC are used. If LIDAR data are available, they are summarized to provide one

---

<sup>1</sup> Consulting Biometrician, 9320 40th Avenue NE, Seattle, WA 98115. E-mail: jwflew-wmen@yahoo.com.

---

or more height statistics; for example, the height above ground of the highest LIDAR return is an obvious statistic to consider. Additionally, data obtained from a “window” in the vicinity of each subject ITC may be available; for example, a statistic similar to top height could be calculated from the LIDAR heights of nearby ITCs.

### **Sample Selection and Field Work**

A fixed number of sample points ( $n$ ) are sought for a stand. Coordinates are randomly selected from a two-dimensional region that fully encompasses the mapped boundaries of the stand. If a random coordinate falls within the stand, it is accepted as a sample point. In addition to the  $n$  desired points, several additional random points within the stand are kept in reserve. The coordinates of the randomly selected points become the preliminary estimates of the locations of the sample points.

Field work consists of traveling to the preselected coordinates and installing fixed-area field plots around each point. The plots are located on the ground with portable Global Positioning System (GPS) equipment. The field crew travels to a point near the desired location. At that point, the GPS equipment is allowed to stabilize to produce a better estimate of the current location. Using that improved estimate, a bearing and distance to the preselected coordinate is determined; the field crew makes the indicated traverse and then monuments a plot center. Subsequent analysis will determine whether or not the monumented plot center is actually in the target stand; if it is not, one of the reserve coordinate pairs is used to establish an alternative plot center.

The field plot is a fixed-area circular plot. Every tree within the plot, whose diameter at breast height (d.b.h.) ( $D$ ) exceeds a certain threshold ( $D_{\min}$ ), is recorded for diameter, species or species group, and distance and bearing from the monumented plot center. A subset of the trees is measured for height.

### **Plot Registration**

The field plot is not assumed to have been located perfectly. The determination of the location of the field plot on the crown map is accomplished with a computer-assisted system that overlays the field-determined stem map on a representation of the remotely sensed data. Initially this procedure is performed

in the field. If it appears likely that the center of the field plot is outside the stand boundary, a replacement plot within the stand is required. The plot registration process can be refined as part of the overall analysis. After the registration is completed, the location of the field plot's center on the ITC map becomes the accepted location for the sample. The vector change in location of each sample point is due to the sum of errors in the map and in the GPS.

### **Crown and Tree Matching**

A procedure is required whereby a mapping between ITCs and field trees is developed. An ITC may be matched to no trees or to any number of trees. A tree may be matched to one ITC or may be unmatched. A requirement of the matching process is that it identifies all of the trees associated with ITCs whose centers are within the boundaries of a crown analysis plot. That plot is a circular plot, centered on coordinates determined in the plot registration process. The size of the plot is less than that of the field plot. A secondary requirement is that all the unmatched trees within a ground analysis plot be identified. The ground analysis plot is a circular plot whose center is the same as the field plot and whose size is smaller than that of the field plot. In the example inventory described later in the article, the field plot size is 0.12 acres and the crown analysis plot and the ground analysis plot both are 0.08 acres.

The mechanism of crown and tree matching is almost entirely automatic. The boundaries of each ITC are extended by up to 1 m, but the boundaries are not allowed to overlap. This process is applied to all crowns within a processing area whose extent must be greater than that of the field plot. Trees whose locations fall within the extended boundaries for an ITC are tentatively associated with that ITC. Of those trees, the one or two with the largest diameters are accepted as being matched, and the remainder is considered to be unmatched. Some subjectivity in the matching is allowed for leaning trees.

### **Sampling Theory**

Every tree within the stand is considered to belong to one of two subpopulations. The first, referred to as the associated trees, consists of all the trees that would be matched to an ITC if the ITC were sampled. The second subpopulation, referred to as the unassociated trees, consists of those trees that would be

unmatched even if all the ITCs in their vicinity were sampled. This imposes a constraint upon the matching process: the decision on matching for an ITC or for a field tree must not be affected by the location of the plot center relative to the ITC or field tree. This constraint is addressed in the discussion section. Separate estimators are developed for the two subpopulations.

Models are developed to predict numbers and sizes of trees associated with each ITC. Before dealing with that complexity, however, the requisite theory to calculate sample weights that allow the data to fairly represent the population of trees within the stand is presented. Although the weights are to be used within regression analysis of ITC data, a simpler application may offer better motivation. Consider how weights would be calculated if the stand's basal area due to associated trees were to be estimated as the product of a sample-determined ratio and the summed areas of the ITCs within a stand. A suitable estimator for that purpose is described. Separately, I address how the observed, unmatched tree data can be used to make an unbiased estimate of the stand's basal area due to the unassociated trees.

Three weight calculations all deal with avoiding biases related to sampling near stand boundaries. Relative weights are assigned to the multiple sample points within a stand. Also weights are associated with individual ITCs within the crown analysis plot, and weights are associated with individual unmatched trees within the ground analysis plot. Ideally, the sample plot weights would be the same for all plots within the stand, which would be the situation if the maps and the GPS equipment were perfect; each sample plot would have been centered at the coordinates of the originally chosen random location. Instead the established plot locations have errors. The distribution of errors is observed in the plot registration process. That distribution can be modeled, and the model can be used to determine the probability density function (*pdf*) for sample point locations within the stand. Since the complete stand geometry is available, the determination of the *pdf* could be a numeric process. Alternatively, for large stands with smooth boundaries, the relative value of the *pdf* can be approximated as a function of the shortest distance to the edge of the stand (*d*):

$$pdf(d) \propto 1 - .5 \times \exp(k \times d)$$

where *k* is a constant related to the variance of errors in plot location.

Weight calculations are required for individual ITCs within a crown analysis plot and for individual unmatched trees within a ground analysis plot. The weight computation depends on the location of the ITCs or the location of the unmatched trees; the same computation procedure applies to both. The most efficient, unbiased weighting scheme is the tree concentric method (Schreuder *et al.* 1993). This scheme considers an object-centered circular plot. Each object's weight is the inverse of the proportion of the plot's area within the stand. The plot size is that of the crown analysis plot or the ground analysis plot, depending on whether ITCs or trees are being addressed. The determination of plot area within the stand is a routine computation for GIS software.

A sample-based estimate of the unassociated trees in the stand is simply a weighed list of the unmatched trees in ground analysis plots. For a single plot, each observed tree represents the number of trees per acre calculated as the product of the tree's weight and the reciprocal of the size of the ground analysis plot. The overall estimate of the distribution of unmatched trees is the weighted average of the estimates derived from the individual plots, using the plot weights specified above. The resultant estimated distribution is an unbiased estimate of the true distribution.

For a simple ratio-of-means estimate of the basal area for ITC-associated trees, the computations are as follows. The sample ratio of basal area to crown area is the weighted sum of the ITC-associated basal areas divided by the weighted sum of the ITCs' areas. The weights here are the product of the plot weights and the weights associated with the individual ITCs. The estimator of stand basal area is the product of the sample ratio and the sum of the areas of all the ITCs in the stand. The estimator is asymptotically unbiased. This estimator is presented for the sole purpose of demonstrating that an ITC sampling approach can be used within the context of a familiar sample estimator, the ratio of means.

---

## Sampling Within Multiple Stands

The situation to be addressed is for a large forest holding with many stands in one or more strata. The intent is that a single set of regression relationships between ITC characteristics and the associated trees should apply to all the stands in a stratum. No expectation exists that the yield statistics of the stands in a stratum are similar. All strata are to be analyzed separately from one another; hence, the discussion need only address a single stratum. That single stratum is assumed to have a large number of stands. All stands are to have remotely sensed imagery, and ITC's are to be delineated for the entire areas of all the stands. Only a small fraction of the stands are selected for sampling. The within-stand sampling protocol is the same as was presented for single-stand sampling.

Two sets of sample stands are used: a development set and a calibration set. The development set of sample stands may be chosen by any means. The calibration set is a random selection from all the stands in the stratum. For the latter, sampling with replacement with probability proportional to size is assumed.

Regression equations are developed to predict the number, species, and sizes of trees associated with each ITC. The form of the equations is selected by viewing the data from the crown analysis plots in the development set of sample stands. Two types of regressions are needed. Logistic regressions predict the number and species of the matched trees for each ITC, and conditional regression equations predict the diameter and height of each tree. After the equation forms are chosen, the calibration and development data sets are combined to fit the coefficients.

The prediction equations receive a final modification based on the calibration data only. The modification to tree counts is a simple multiplier. For the d.b.h. equation, the modification is similar to a multiplier on basal area per tree, with an offset such that trees with diameter  $D_{\min}$  are unmodified. Another modifier equation predicts altered values of tree height. The coefficients of the modifier equations are set in this order: trees, diameters, and heights. The values of the coefficients are set such that the weighted sum of predictions for the sample ITCs exactly equals the weighted sums of statistics from the trees matched to the same set of ITCs. The statistics being matched are trees per

acre, basal area per acre, and product of basal area and Lorey height. That product is computed as the sum over trees of basal area times tree height and expressed on a per-acre basis.

The tree predictions sum to stand predictions, which in turn sum to the prediction of a stratum total. The estimator of a stratum total is similar to a ratio of means estimator where the independent variable is the unmodified regression prediction and the dependent variable is the true value. In the present case, however, the calibration sample influences the regression predictions; hence, the usual properties of the ratio estimator cannot be assured. The generalized regression (GREG) estimator (Särndal *et al.* 1992) is similar to that proposed here in that both adjust the estimator of the total on the basis of the weighted errors in the sample. The GREG estimator does so with an additive term while the approach here uses a multiplicative term. The ratio estimator is asymptotically unbiased, and the GREG estimator is approximately asymptotically design unbiased. The estimator proposed here is also expected to be approximately asymptotically design-unbiased. Asymptotic unbiasedness could be assured by limiting the use of the calibration data to the final modification step; doing so is not recommended because of the diminution of the sample size available for fitting the regression models.

The unassociated trees are dealt with separately. One approach would be to simply calculate an average stand table of unassociated trees using the weighted data from the calibration set. Alternatively, characteristics of the distribution of unassociated trees can be related to the predictions for associated trees. A complex model is warranted and possible only if a fair number of unassociated trees are in the sample data, and if the unassociated component is of silvicultural significance.

Error statistics are sought at the per acre level not the tree level. The variances of the estimators of strata-level means for trees per acre, basal area per acre, and basal area times Lorey height can be estimated from the sample results by stand; the individual plot data are needed only for computing stand-level statistics. As a first approximation, the predicted statistics for the stand can be treated as having come from a ratio estimator where the data are the stand-level means for the plot data in the calibration sample, and the means of the corresponding predic-

---

tions before the final calibration. The variance of the ratio may be approximated with standard formula for a ratio estimator, making no finite population correction. Alternatively, the bootstrap variance estimator may be used.

## Example Inventory

### Population, Sample Design, and Data

The sampling and regression methodologies were applied in a commercial forest with two strata. The single stratum addressed here consists of young plantations of a southern pine species, with minor hardwood components in some stands. A lower limit on stand age was chosen so that most of the dominant trees would be expected to have diameters greater than 3 in. Some 2,456 stands met the age criterion and were included in the stratum. Color infrared photography was obtained with a Digital Mapping Camera with 8-bit pixel radiometric resolution, multiple charge-coupled device camera heads, and forward motion compensation. Pixel size was 0.6 m on a side. Airborne LIDAR data was from an Optech ALTM 3100 Sensor System (100,000 laser pulses per second) that had an average of five first-returns per square meter. The LIDAR data were pixelized, 0.5 m on a side, and smoothed.

ITCs were delineated with algorithms similar to those described by Hyypä *et al.* (2001). Manual training processes were used to set parameters for the algorithms. Each delineated ITC consisted of a specific set of LIDAR-derived pixels. A center point was defined as the geometric mean of the pixels associated with an ITC. A color was obtained for each ITC as the average color over the ITC by overlaying the ITC shape pattern on the CIR data. The height of the ITC was calculated as the height of the highest pixel.

Management data were available for all the stands. The data that were used included stand age, year of most recent thinning, and stand area. The development sample consisted of data from three sample stands which were thinned and ten which were unthinned. All of the unthinned stands in the strata had been placed into a  $3 \times 3 \times 3$  orthogonal array based on age, height, and crown closure. Statistics for stand height and crown closure had been derived from the LIDAR data. One sample

stand was drawn from each of the corner cells of the array, and two were drawn from the center cell. Sampling was proportional to stand area. The resultant sample broadly covered the array of stand conditions but was not a representative sample. The calibration sample consists of 25 stands selected from the stratum; the selections were independent, with replacement, and with probability proportional to stand area.

Each sample stand was sampled with two 0.12 acre circular plots. The plots were located using the registration protocol stated earlier, with the exception that plots, whose centers were determined to be outside the targeted stand, were excluded but not replaced. The lower diameter limit ( $D_{\min}$ ) was 2.95 in. All trees above that limit had their diameters, species, and coordinates recorded. Every fourth tree was measured for height; additional height measurements were taken in plots where the number of height trees per species group would otherwise have been low. The species groups were pine and hardwood. Height diameter curves were fit, usually by species group within a plot, to impute height for unmeasured trees. For stands with few hardwoods, height diameter data were sometimes shared between the plots in a stand.

### Predictions and Error Analysis

The prediction process is summarized here and is not given in detail. CIR data from the matched ITCs from plots in the development data set, supplemented with other subjectively chosen trees, were used in a discriminant analysis to obtain a preliminary prediction of the probability that an ITC would have a pine as its largest matched tree. That preliminary probability estimate, together with other ITC data and stand variables, were used as inputs in logistic regressions to predict the probabilities that zero, one, or two trees were associated with the ITC and to predict the species of the associated trees. Hence, the discriminant analysis had, as its only purpose, the collapsing of the CIR data to a single probability value for each ITC. A series of least-squares regression equations predicted the d.b.h. and heights of trees, conditioned on species and whether the tree was the largest or second largest associated with the ITC. For each ITC, the predictions were a series of outcomes with associated probabilities; the probabilities sum to one. The final step in developing the equations was to bring in adjustment factors which forced the weighted mean predictions for the calibration

---

data to equal the weighted mean of the observations. This step was done separately by species. The first adjustment fixed trees per acre, the second adjustment fixed basal area per acre, and the third adjustment fixed the basal area height product. In application, the predicted probabilities of various outcomes for every ITC in the population were converted to expected values and summed to generate stand tables.

Error analyses, as briefly described earlier, were used to estimate the variance of strata means estimators. One approach was to make predictions for each sample stand as the average of the predictions for the two plots, use the corresponding observed data to construct the ratio of means estimator, and then calculate the variance of the ratio estimator. That approach used the simplifying assumption of a common ratio for both species. Another approach was the bootstrap; its implementation has separate factors by species. Both approaches yielded similar results. The 90 percent confidence interval for basal area per acre and basal area times height is  $\pm 4.1$  percent and  $\pm 5.4$  percent, respectively. Actual errors were not known for any stands. The magnitude of the stand level errors, however, could be inferred through the use of a mixed model with terms for plot error and for stand error. The mixed model implied that stand level results had a root mean square error of 9.7 percent for basal area and 12.8 percent for basal area times height. To put the basal area result into perspective, the same level of uncertainty at the stand level would have been achieved on average by a system that did not use crown data but instead estimated each stand's basal area as the mean basal area observed on ten 0.12 acre plots, randomly located within the stand.

## Discussion

Much of this manuscript is focused on how maps of ITCs can be used for sampling in a way that allows for unbiased estimation of basal area and a few other statistics. Subject to a few caveats, unbiased estimation is possible in a stand which is being sampled. If the unbiasedness property were not to hold at the stand level, it would not hold at the strata level. Furthermore, without the unbiasedness property, there would be little hope of obtaining accurate error assessments. What

has been demonstrated is that unbiased estimators are possible in spite of errors in ITC delineation, errors in matching, and errors in modeling. To the extent that the mentioned sources of errors can be reduced, the precision of the estimators can be improved.

The within-stand sampling techniques presented here are potentially better than most cruising techniques in terms of randomly locating plots and successfully avoiding edge bias. The avoidance of edge bias will become increasingly important as stands become more fragmented and as management systems seek to delineate more special features such as riparian zones. One substantive difference between sampling from a map and sampling on the ground is the determination of stand boundaries. The use of crown maps for this purpose may be ideal in situations where differences in the remotely sensed imagery prompted the decision to draw a stand boundary. An extreme example of this type of stand boundary is embodied in the Gougeon and Leckie (2003) system of drawing stand boundaries so as to separate regions of differing ITCs.

The method of weighting plots within a stand based on a numerically determined *pdf* requires some effort, but it is feasible. Implicit in that method is an assumption of no differential bias between map coordinates and GPS coordinates. If all the map coordinates were off by several meters in one direction, however, this would cause one edge of the stand to be under-represented in the actual field sampling. A costly solution would be to expand the region in which random map points are located to include the stand plus a buffer with a width equal to the maximum anticipated coordinate bias. A less expensive solution would be to estimate the location bias before entering the stand and then compensate for the bias. Any accessible feature in or close to the stand that can be located on the ground and in the imagery could be used to estimate the location bias.

The principal weakness in the sampling system presented here relates to tree matching. A problem is that subjective alterations in matching are allowed under the presumption that leaning trees or minor errors in ITC positioning cause the automatic matching process to go astray. These subjective alterations, while infrequent, have the potential to introduce bias. Persson *et al.* (2002) discuss the matching problem in

some detail from the perspective of trying to identify the physically correct matches. In addition to considering relative size of potentially matched trees, they favor matches that have short distances between tree position and ITC position. Apart from any bias consideration, having trees lean out of position and become associated with the “wrong” ITCs leads to loss of precision. Gatzliolis (2009) minimizes this problem by locating in the field the three dimensional location of the tree tops, thereby improving the physical accuracy of the tree matching, particularly for conifers. His methods also improve the plot registration process. Gatzliolis suggests that submeter accuracies are attainable; results in the example inventory of southern pine support that conclusion. With sufficiently high LIDAR scanning density in a coniferous forest, spatial accuracies of 0.2 m or less would seem to be achievable. If lean were to be fully accounted for, however, the retention of unbiasedness might require that larger field plots be measured. In the example inventory, the distance between the outer edge of the analysis plot and the outer edge of the field plot was 2.28 m. With this short distance, it is likely that some crowns, whose centers were within the crown analysis plot, extended beyond the measurement plot; hence, the location of the sample point relative to an ITC could affect the matching of trees to an ITC. An analysis of distances between ITC centers and locations of the largest associated matched trees indicated that matches missed due to not having measured beyond the 0.12-acre field-plot boundary would occur for 1 out of 200 ITC’s on the outer edge of the crown analysis plot and far less frequently for ITCs located closer to the center of the crown analysis plot. If matching were to be based on the coordinates for the tops of the field trees rather than the bole locations at breast height, either the field plot size would have to increase or bias due to missed matches would increase.

The overall sampling design of having separate development and calibration samples is unusual. The deliberate assignment of the development sample within a design matrix ensures that the full range of stand conditions is being sampled in a manner that is likely to be efficient for regression modeling. These data are used to select the forms of the regression models. Model-assisted survey sampling, and possibly all sample survey estimators which claim unbiasedness, assume that the form of the estimation equation is fixed before sampling. Since ITC

regression models are in an early development stage, and would be expected to vary with forest type and sensing technology, they should not be presumed to be known before sampling. The procedures outlined here ensure that the model forms are known and fixed before the critical step of calibration. A motivation for using sampling with replacement in the calibration set is that this ensures that the stand-level observations are from identical and independent distributions. In turn, variance computations are greatly simplified, and the assumptions in standard bootstrap methodology are more easily met.

## Acknowledgments

The example inventory described here was carried out by ImageTree Corporation of Morgantown, WV. Bob Pliszka coordinated the project and supervised the field work. Olavi Kelle developed and applied the image processing algorithms. The author served as a consultant to ImageTree Corporation in the design of the inventory and the analysis of the data.

## Literature Cited

- Gatzliolis, D. 2009. Precise plot registration using field and dense LIDAR data. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C.; McWilliams, W.H., Proceedings, eighth annual forest inventory and analysis symposium: 243-250.
- Gougeon, F.A.; Leckie, D.G. 2003. Forest information extraction from high spatial resolution images using an individual tree crown approach. Information Report BC-X-396. Victoria, British Columbia, Canada: Canadian Forest Service, Pacific Forestry Center. 29 p.
- Hyypä, J.; Kelle, O.; Lehtikainen, M.; Inkinen, K. 2001. A segmentation-based method to retrieve stem volume estimates from 3-D tree height models produced by laser scanners. *IEEE Transactions on Geoscience and Remote Sensing*. 39: 969-975.
- Næsset, E.; Gobakken, T.; Holmgren, J. *et al.* 2004. Laser scanning of forest resources: the Nordic experience. *Scandinavian Journal of Forest Research*. 19: 482-499.

---

Næsset, E.; Nelson, R.F. 2009. Sampling and mapping forest volume using airborne LIDARs. In: McRoberts, R.E.; Reams, G.A.; Van Deusen, P.C.; McWilliams, W.H., eds. Proceedings, eighth annual forest inventory and analysis symposium. Gen. Tech. Rep: 297-302.

Persson, Å.; Holmgren, J.; Söderman, U. 2002. Detecting and measuring individual trees using an airborne laser scanner. *Photogrammetric Engineering and Remote Sensing*. 68: 925-932.

Särndal, C.E.; Swensson, B.; Wretman, J. 1992. Model assisted survey sampling. New York: Springer. 694 p.

Schreuder, H.T.; Gregoire, T.G.; Wood, G.B. 1993. Sampling methods for multi-resource inventories. New York: John Wiley. 446 p.