Journal of Data Science **9**(2011), 261-270

Using the Non-Parametric Classifier CART to Model Wood Density

Eduardo Navarrete and Miguel Espinosa
Universidad de Concepción

Abstract: To identify the stand attributes that best explain the variability in wood density, *Pinus radiata* plantations located in the Chilean coastal sector were studied and modeled. The study area corresponded to stands located in sedimentary soil between the zones of Constitución and Cobquecura. Within each sampling sector, individual tree variables were recorded and the most relevant stand parameters were estimated. Fifty trees were sampled in each sector, obtaining from each one six wood discs from different stem heights. Each disc was weighed in green and then dried to anhydrous weight, and its basic density was calculated. The profile identification to classify basic density according to stand characteristics was performed through regression trees, a technique based in the use of predictor variables to partition the database using recursive algorithms in regions with similar responses. The objective of the regression tree method is to obtain highly homogenous groups (branches), which are identified using pruning techniques that successively eliminate the branches that least contribute to the classification of the variable of interest. The results found that the stand attributes that contributed significantly to basic density classification were the basal area, the number of trees per hectare, and the mean height.

Key words: Basic density, decision trees, *Pinus radiata*, regression trees.

1. Introduction

Forestry modeling has commonly been based in attribute quantification and prediction at the stand and individual tree level. Due to the increased intensity of stand management, with the goal of increasing both wood production and quality under different management conditions, more and better information with respect to forest growth and performance is also necessary. Growth simulators are a valuable decision-making tool in the forestry sector. These are based in models that estimate future tree growth over a certain time period, and consequently make forecasts on the wood resources to be produced, evaluate silvicultural prescriptions, and economically analyze different management alternatives (Vanclay,

1994). However, these simulators are not capable of predicting wood quality under different silvicultural options (Tassisa and Burkhart, 1998). Of the wood properties that affect quality, basic density is one of the most important because it determines its utilization in saw mills, manufacturing factories, cellulose plants, and as planks. Due to this relation with the final product quality, saw mills and paper industries are interested in both density and its variability (Tian *et al.*, 1995). Haygreen and Bowyer (1996) found several factors that affect the variability in basic density, such as the site, climate, geographic location, species, age, and silviculture. Valencia and López (1999) indicate that the value of basic density and its variation depends to a great degree on the height and tree section where the sample is extracted. In general, the majority of the studies on wood properties are based in the description and evaluation of the variability and its causal factors without providing tools to estimate these properties.

Wood property estimation is commonly performed within specific zones and using traditional statistical tools such as regression models. Other non-traditional model construction methods are the so-called classification and regression trees (CART) (Breiman *et al.*, 1984), whose objective is to obtain highly homogenous groups (branches), which are achieved using pruning techniques that consist in successively eliminating the branches that contribute least to the classification of variable of interest (Larsen and Speckman, 2002; Skinner *et al.*, 2002; Moisen and Frescino, 2003; Tamminen *et al.*, 2003). The decision trees can be used to define wood use in function of stand characteristics, identifying wood production areas with defined characteristics and according to market specifications.

Due to the need of more disaggregated information with respect to the wood density of *Pinus radiata* (radiate pine) in Chile, the general objective of this study is to identify the stand attributes that best explain wood variability in this species. The specific objectives are to identify profiles that classify wood density according to its most relevant stand characteristics.

2. Materials and Methods

Study Area

The study area corresponded to radiata pine tree plantations, aged between 20 and 28 years, established in the Chilean coastal sector between Constitución and Cobquecura (34°50′ a 36°25′ S) in predominantly sedimentary-origin soils.

Description of Sampling Sector

The sampling sector was characterized at three levels: site, stand, and individual tree. At site level, the soil type was recorded; at stand level, the

site index (dominant and codominant trees height), age, density (trees/hectare), height, mean diameter, and basal area (occupation area tree) were recorded. The breast height diameter, total height, and crown initiation height (first live branch height), crown class, and height at which the wood discs were obtained were recorded for each tree sampled. The table 1 shows the main statistics of stand characteristics for sampled trees.

Table 1: Stand characteristics for sampled trees

	Age (years)	**Site Index (m)**	**Stand Density (tree/ha)**	**Basal Area (m^2/ha)**	**Mean height (m)**	**Mean diameter (cm)**
Average	26	26.7	627	39.4	30.7	26.8
Minimum	20	22.9	440	21.2	27.7	23.2
Maximum	28	30.4	867	77.1	32.9	32.6
Deviation	3	2.5	196	21.5	1.8	3.2

Sampling and Laboratory Procedures to Basic Wood Density Determination

In the tree selection process, the diameter classes that were principally represented in the distribution that characterized the stand structure (regular, with classes 22 to 32 cm diameter) were considered, choosing stems that were cylindrical, straight, and free of bifurcations and defects. Eighty two trees were sampled on five stands (standard deviation of 5 kg/m^3 and a confidence level of 95%). From each sampled tree, six discs were extracted: at the stump, at 5% (Dap), 25%, 50%, and 75% of the tree's commercial height as well as at the utilization limit diameter (ULD) 8 cm (sample total: 82 trees $\times$ 6 disc/tree = 492). Each one of the samples (wood disc) were weighed and measured green, following Chilean Standards NCh 176/1 and 176/2 (INN, 1985, 1986), and subsequently dried in a stove at $103 \pm 2^{\circ}$ until obtaining anhydrous weight. The basic density was calculated relating the sample's anhydrous weight with respect to its green state volume.

Statistical Analysis

Regression Trees Construction

With the descriptive stand variables for the study area (Table 1), regression trees were constructed using software S-PLUS 2000 (Mathsoft, 2000). The process is the following: Consider the multiple regression problem $y_i = f(x_{i1}, \cdots, x_{ip}) + \varepsilon_i$, $i = 1, \cdots, n$, where f is unknown and not easily parameterized; x_{ij} are independent known variables, and ε_i are random error terms with zero mean. A

node N is a subgroup of indexes $\{1, \cdots, n\}$. The deviance of node N is defined as:

$$D\,(N) \;=\; \sum_{i\,\in\,N} (y_i \;-\; \bar{y}\,(N))^2, \tag{1}$$

Where $\bar{y}(N)$ is the mean of the observations in node N. The root node consists in all the observations, and in each step, the parental node is divided recursively in two child nodes: a left node (N_L) and a right node (N_R) in order to minimize $(N_L)+D(N_R)$. Node partition is performed considering, in the case of continuous variables, all the divisions of the formula $N_{Lj} = \{i \in N : x_{ij} \leq t\}$, $N_{Rj} = \{i \in N : x_{ij} > t\}$ for constant t.

For each independent variable, all the possible partitions are considered, calculating the deviance for the following node to be divided $D(N_L) + D(N_R)$. The candidate partitions are calculated for each independent variable, and variables that produce the best divisions (with less deviance) are selected to partition node N. The algorithm proceeds recursively until the next partition cannot be performed according to predetermined criteria. Normally, a number of nodes or stops is specified a priori when the deviance of the node is above a certain level (Larsen and Speckman, 2002).

The selection of one tree with respect to another will generally depend on the estimation of its error rate $R(T)$. This rate can be estimated in several manners, where the most notable is cross validation. This estimation method consists in estimating $R(T)$ to the estimator by validation sample in a reiterate and analogous manner. Each time, a fraction k^{-1} of the total sample size is removed from the tree construction sample. In this way, k estimates $R^{(1)}(T), \cdots, R^{(k)}(T)$ are obtained and averaged in using the following formula:

$$R^{cv}(T) = \frac{R^{(1)}(T) + \cdots + R^{(k)}(T)}{k}, \tag{2}$$

where $R^{cv}(T)$ means R cross-validation. In the case that the tree constructed for each one of the sub-samples is different from the others, the previous expression would not be valid.

A basic technique in tree construction suggests the construction of leafy trees, arriving to the maximum possible tree $A_{máx}$ without considering error rates, and pruning can be performed after their construction by choosing the tree that provides the lowest error rate. Once the entire tree $A_{máx}$ has been constructed, and is adjusted to fit the data, a pruning algorithm is applied to obtain a sequence of sub-trees through the successive suppression of the branches that provide less information in terms of discrimination between the class of the response variable Y. Tree pruning is a procedure that is analogous to the "Backward" selection in

regression: removes some of the terminal nodes. Finally, the sub-tree A^* that provides the lowest error rate is selected (Puerta, 2002).

According to Larsen and Speckman (2002), a bonding criterion for adjustment of tree T with $\{N_k\}$ terminal nodes is defined as:

$$D(T) = \sum^* D(N_k), \tag{3}$$

where $\sum^*$ means summing over all the terminal nodes N_k. If a tree T' is a sub-tree of T, clearly $D(T) \leq D(T')$. The pruning algorithm successively removes pairs of terminal nodes corresponding to the partition with the least deviance. In other words, if T has terminal nodes $\{N_k\}$, then each pair $\{N_{2j}, N_{2j+1}\}$ is the result of the division of a larger node N_j with $D(N_j) \geq D(N_{2j}) + D(N_{2j+1})$. The totality of the terminal node pairs are examined, and the pair with the lowest $D(N_j) - D(N_{2j}) - D(N_{2j+1})$ is removed to create a new sub-tree T'. This method is analogous to removing the least significant variable in the "backward" selection process. The process is repeated to create a nestled set of trees $T_m \subset \cdots \subset T_0$, where T_0 is the entire tree and T_m corresponds to the tree with only a root node.

To select the quantity of tree sequences, Breiman *et al.* (1984) propose a measure of cost-complexity for the tree T,

$$D_\alpha(T) = D(T) + \alpha \times size(T), \tag{4}$$

Where α is a parameter chosen to adjust for cost-complexity. For a certain α, there is at least one tree that minimizes $D_\alpha(T)$. Where $\alpha = 2\sigma^2$, it corresponds to the automatic information criterion (AIC). The deviance of the nestled sequence is a decreasing function of α.

Breiman *et al.* (1984) suggests that an optimal tree should be one with the least possible quantity of terminal nodes, with a standard minimum error, and with the lowest cost from the point of view of the information that it should contribute.

3. Results and Discussion

The constructed regression tree found that the stand variables that principally contribute to the wood density classification are in their order of importance in the discriminatory process: the basal area per hectare, the mean height, stand density, and site index (Figure 1). The tree's left branch has grouped the stands with low average basic density (415.7 kg/m^3), corresponding to those stands with a basal area below 22.1 m^2/ha. The right branch has grouped the stands with higher basic density defined by the variables mean height, stand density, and site index. The stands with the highest basic density (482.3 kg/m^3) are found at a mean height lower than 32.2 m and a stand density below 455 trees/ha (Figure 1).

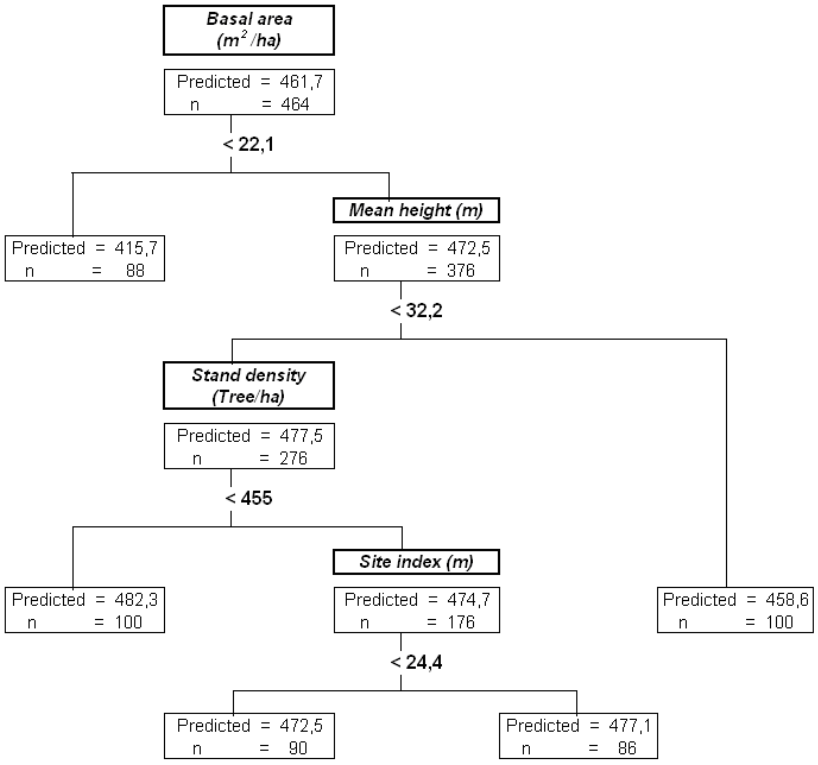

Figure 1: Complete tree with five terminal nodes using only stand characteristics. The sample sizes of each node are indicated by "n".

Since the complete regression tree is usually generated by the CART method, the model is over-adjusted. Consequently, pruning methods need to be used to eliminate the terminal nodes that least contribute to the classification of the variable of interest. For this reason, the tree was pruned using the cross-validation method (Breiman *et al.*, 1984). Figure 2 shows the deviance behavior considering several tree sizes (number of terminal nodes), indicating that with three nodes it is possible to diminish model deviance, reducing the dimensionality of the original tree from five to only three terminal nodes, avoiding in this way over-adjusting the model.

After pruning, the regression tree was constituted by only two stand variables (basal area per hectare and mean height), and with which the model could adequately discriminate the stands according to wood density. The highest average basic density (477.5 kg/m^3) is given by stands with a mean height lower than 32.2 m and a basal area larger than 22.1 m^2/ha (Figure 3).

If the stand's mean height is considered as a quality predictor variable, given its close relation with the site index, then it reasonable to think that the stands

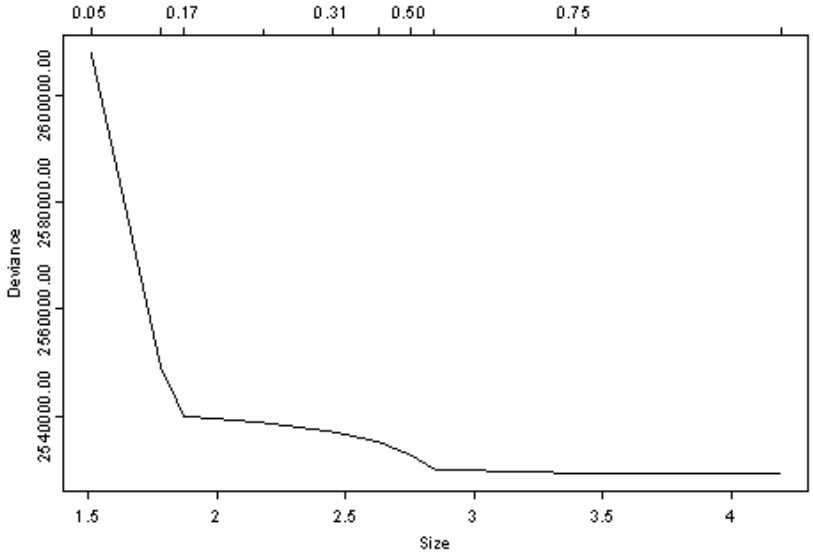

Figure 2: Fitted deviance reduction of the complete tree model, using a tree with three terminal nodes.

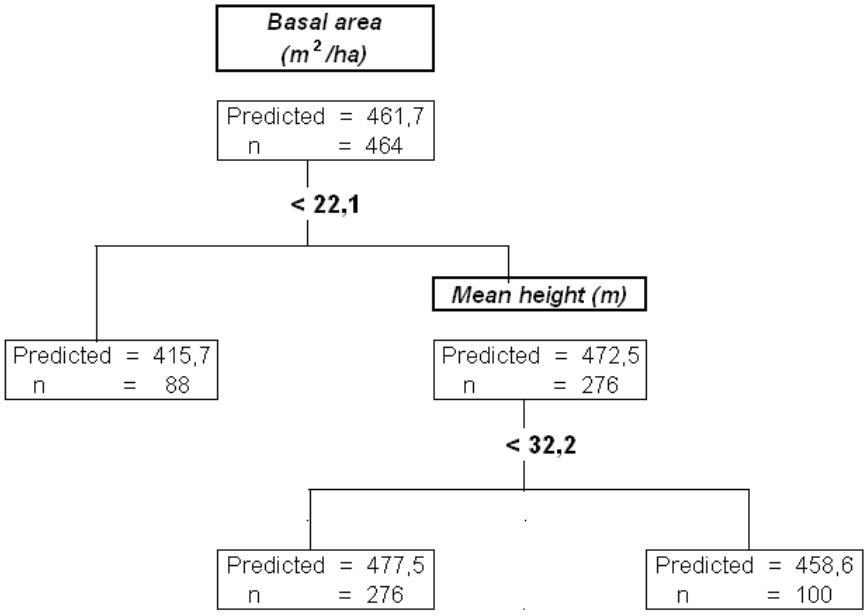

Figure 3: Final tree after pruning to three terminal nodes. The sampling sizes of each node are indicated by "n"

that presented a higher basic density corresponded to those with lower mean height. With respect to this point, numerous authors have demonstrated that

wood density increases as site quality diminishes. Schutz *et al.* (1991), in a study of *Pinus patula* in different sites, demonstrated that the site influenced 61% of the total variation in basic density. Morales (1968) found for radiata pine in Chile that the wood's specific weight increased as the site's quality diminished, increasing 21% between a good and a bad site. This result can be explained considering that the specific weight is a function of the ratio that exists between the volume occupied by the cellular walls and the volume of empty spaces. Logically, if the length of the tracheids is lower in poor sites, then the volume occupied by the cellular walls will grow, translating into an increase of wood density because less space would be occupied by cellular cavities (Haygreen and Bowyer, 1996). Several authors have demonstrated as well that ring width is relatively important in the specific weight. Consequently, if growth velocity decreases in a lower quality site, this will translate into a diminishment of ring width, increasing wood density (DeBell *et al.*, 1994).

With respect to the basal area, as the first discriminatory variable of wood density and given that it corresponds to a measure of stand density, it is logical to conclude that at higher basal area values, the basic density will be higher. According to González and Molina (1989), a forest's growth rate will be affected by the quantity of trees per surface unit since the growth potential is distributed on these. With less individual per surface unit, and consequently lower basal area, growth will be faster, generating wider rings, and consequently less density. According to Larocque and Marshall (1995), stand density closely affects wood density in *Pinus resinosa*, generally presenting a decreasing tendency in wood density as tree spacing increases. With respect to this point, Cown and McConchie (1982) signal that the principal factor affecting wood's intrinsic properties is tree age, which closely controls wood density and the development of later wood development. The growth rate per se has been demonstrated to have a minimal effect on wood density, although several studies have demonstrated that stand density levels and the wood density are negatively correlated (Cown and McConchie, 1982).

4. Conclusions

Considering the complete regression tree, the stand attributes that significantly contribute to the classification of basic density are: basal area, number of trees per hectare, mean height, and site index.

The regression tree reduced by cross validation diminishes the dimensionality of the final model, incorporating the basal area and mean height as the only stand variables in the classification of basic wood density.

References

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont.

Cown, D. and McConchie, D. (1982). Rotation age and silvicultural effects on wood properties of four stands of *Pinus radiata*. *New Zealand Journal of Science* **12(1)**, 71-85.

DeBell, J., Tappeiner, J. and Krahmer, R. (1994). Wood density of western hemlock: effect of ring width. *Canadian Journal of Forest Research* **24**, 638-641.

Goicoechea, A. P. (2002). Imputación basada en árboles de clasificación. Eustat. Available in: http://www.eustat.es/documentos/datos/ct_04_c.pdf

González, J. and Molina, J. (1989). Consideraciones sobre los tratamientos silviculturales y los rendimientos cuantitativos y cualitativos en madera pulpable de pino radiata. Documentos Técnicos N°38-39. Revista Chile Forestal.

Haygreen, J. and Bowyer, J. (1996). *Forest Products and Wood Science*: *An Introduction*, 3rd ed. Iowa State University Press, Ames.

Instituto Nacional de Normalización (INN). (1985). Norma Chilena Oficial, NCh 176/1. Of 84, "Madera-Determinación de Humedad". Primera Edición. Santiago.

Instituto Nacional de Normalización (INN). (1986). Norma Chilena Oficial, NCh 176/2. Of 86, "Madera-Determinación de la Densidad". Primera Edición. Santiago.

Larocque, G. and Marshall, P. (1995). Wood relative density development in red pine (*Pinus resinosa* Ait.) stands as affected by different initial spacing. *Forest Science* **41(4)**, 709-728.

Larsen, D. R. and Speckman, P. L. (2004). Multivariate regression trees for analysis of abundance data. *Biometrics* **60**, 543-549.

Mathsoft. (2000). *S-Plus 2000 User's Guide*. Mathsoft Inc., Seattle.

Moisen, G. G. and Frescino, T. S. (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling* **157**, 209-225.

Morales, R. (1968). *Variación del peso específico y largo de traqueidas según edad y sitio en plantaciones de Pinus radiata D. Don*. Undergraduate Thesis, Facultad de Agronomía, Escuela de Ingeniería Forestal, Departamento Tecnología de la Madera, Universidad de Chile. Santiago.

Schutz, C., Christie, S. and Herman, B. (1991). Site relationship for some wood properties of pine species in plantation forests of southern Africa. *South Africa Forestry Journal* **156**, 1-6.

Skinner K., Montgomery, D., Runger, G., Fowler, J., McCarville, D., Reed, T. and Stanley, J. (2002). Multivariate statistical methods for modeling and analysis of wafer probe test data. *IEEE Transactions on Semiconductor Manufacturing* **15(4)**, 523-530.

Tamminen, S., Laurinen, P. and Röning, J. (1999). Comparing regression trees with neural networks in aerobic fitness approximation. *Proceedings of the International Computing Sciences Conference Symposium on Advances in Intelligent Data Analysis*, Rochester, N.Y., June 22-25, 1999, pp.414-419.

Tasissa, G. and Burkhart, H. (1998). Modelling thinning effects on ring specific gravity of loblolly pine (*Pinus taeda* L.). *Forest Science* **44(2)**, 212-223.

Tian, X., Cown, D. and McConchie, D. (1995). Modelling of *Pinus radiata* wood properties. Part 2: Basic density. *New Zealand Journal of Science* **25(2)**, 214-230.

Vanclay, J. K.(1994). *Modelling Forest Growth and Yield*: *Applications to Mixed Tropical Forests*. Cab International, Wallingford.

Receied August 24, 2009; accepted January 18, 2010.

Eduardo Navarrete
Departamento Forestal
Universidad de Concepción
Juan Antonio Coloma 0201, Los Ángeles, Chile
ednavarr@udec.cl

Miguel Espinosa
Departamento de Silvicultura
Universidad de Concepción
Víctor Lamas 1290, Concepción, Chile
mespinos@udec.cl