Rapid Techniques for Screening Wood Properties in Forest Plantations

by

Christian R. Mora

(Under the direction of Laurence R. Schimleck)

Abstract

The use of three nondestructive techniques (NDT) for the assessment of wood properties in forest plantations: i) acoustics; ii) near infrared (NIR) spectroscopy; and iii) SilviScan, was studied. In this dissertation, a major focus was given to aspects related to the development of NIR calibration models using SilviScan data. The utilization of acoustics in Pinus taeda plantations was useful for the classification of trees based on wood stiffness. NDT values compared well to those obtained from traditional laboratory tests after accounting for differences in moisture content and the type of acoustic waves employed. For NIR modeling, using pulp yield data from Eucalyptus nitens trees, it was found that models fitted with a reduced number of selected samples showed similar performance compared to a regression fitted using all samples. Selection of samples based on NIR spectra was as successful as selection based on pulp yields. The partial least squares (PLS) algorithm was modified allowing incorporation of grouped data. It was shown that repeated measurements of wood strips could induce autocorrelation in the PLS residuals. Models fitted using the modified approach showed that the serial correlation can be completely removed; however the strength of the autocorrelation will determine whether PLS or its modified version is preferable. Net
analyte signal and figures of merit were introduced to help interpreting the effects of different pre-processing techniques on the NIR data. Sensitivity, selectivity, and signal-to-noise ratio proved to be more useful statistics than $R^2$, RMSECV, and RMSEP for this objective. Two approaches based on randomization tests were proposed for determining the dimensionality of principal component (PCA) and PLS models. Compared to these tests, it was found that cross-validation can lead, in some situations, to model underfitting. Alternatively, the development of nonlinear NIR calibrations based on kernel methods showed that some problems detected with PLS were due to nonlinearities in the properties. Kernel regressions were able to model better these properties, improving their predictive ability compared to PLS. Finally, the use of NIR values as control points for spatial interpolation was demonstrated by generating within-tree maps showing the variation of density and microfibril angle in *P. taeda* trees.

RAPID TECHNIQUES FOR SCREENING

WOOD PROPERTIES IN FOREST PLANTATIONS

by

CHRISTIAN R. MORA

Forest Engineer, Universidad de Chile, Santiago, Chile, 1996

MSc, North Carolina State University, 2003

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2009
RAPID TECHNIQUES FOR SCREENING
WOOD PROPERTIES IN FOREST PLANTATIONS

by

CHRISTIAN R. MORA

Approved:

Major Professor: Laurence R. Schimleck

Committee: Richard F. Daniels
             Michael B. Kane
             Fikret Isik

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2009
DEDICATION

To my wife Jimena, my daughter Fernanda, and my son Sebastián
This doctoral research was supported by an USDA Forest Service Agenda 2020 grant. I gratefully acknowledge support from the Wood Quality Consortium of the University of Georgia, the USDA Forest Service Southern Research Station, the North Carolina State University Cooperative Tree Improvement Program, and funding from Bioforest S.A. (Arauco) and the University of Georgia Graduate School through the Doctoral Dissertation Completion Award. However, this work would have not been possible without the support of Laurie Schimleck and Alex Clark (THANKS!). I also wish to express my sincere appreciation to Dick Daniels, Mike Kane, and Brock Stewart (University of Georgia), Jerry Mahon (Molpus Timberlands Management), Lewis Jordan (Weyerhaeuser Corporation, Southern Timberlands Group), Coby Smith, Bailian Li, Fikret Isik, and Lee Allen (North Carolina State University), Eduardo Rodríguez and Miguel Peredo (Bioforest S.A.), and of course my family.
Table of Contents

Acknowledgments ......................................................... v

Chapter

1 Nondestructive techniques for the estimation of wood properties: Introduction ............................................. 1
  1.1 Acoustics .......................................................... 1
  1.2 SilviScan .......................................................... 3
  1.3 Near infrared (NIR) spectroscopy ................................. 5
  1.4 References ....................................................... 7

2 Relationships between acoustic variables and different measures of stiffness in standing Pinus taeda trees ......... 14
  2.1 Abstract .......................................................... 15
  2.2 Introduction ....................................................... 15
  2.3 Material and methods ............................................. 18
  2.4 Results and discussion ........................................... 22
  2.5 Conclusions ....................................................... 28
  2.6 References ....................................................... 29

3 On the selection of samples for multivariate regression analysis: application to near infrared (NIR) calibration models for the prediction of pulp yield in Eucalyptus nitens ......................... 41
  3.1 Abstract .......................................................... 42
  3.2 Introduction ....................................................... 42
3.3 Material and methods ........................................... 45
3.4 Results and discussion ......................................... 51
3.5 Conclusions ....................................................... 54
3.6 References ......................................................... 55

4 Near infrared calibration models for the estimation of wood
density in *Pinus taeda* using repeated sample measurements .... 64
  4.1 Abstract ......................................................... 65
  4.2 Introduction .................................................... 65
  4.3 Material and methods ........................................... 68
  4.4 Results and discussion ......................................... 74
  4.5 Conclusions ....................................................... 77
  4.6 References ......................................................... 79

5 Determination of specific gravity on green *Pinus taeda* samples
by near infrared spectroscopy: comparison of pre-processing
methods using multivariate figures of merit ......................... 93
  5.1 Abstract ......................................................... 94
  5.2 Introduction .................................................... 94
  5.3 Material and methods ........................................... 96
  5.4 Results and discussion ......................................... 101
  5.5 Conclusions ....................................................... 107
  5.6 References ......................................................... 108

6 Significance tests for principal components and partial least
squares models: application to the study of microfibril angle
in *Pinus taeda* using near infrared (NIR) spectroscopy .......... 119
  6.1 Abstract ......................................................... 120
  6.2 Introduction .................................................... 120
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3</td>
<td><strong>Material and methods</strong></td>
<td>123</td>
</tr>
<tr>
<td>6.4</td>
<td><strong>Results and discussion</strong></td>
<td>128</td>
</tr>
<tr>
<td>6.5</td>
<td><strong>Conclusions</strong></td>
<td>132</td>
</tr>
<tr>
<td>6.6</td>
<td><strong>References</strong></td>
<td>133</td>
</tr>
<tr>
<td>7</td>
<td>Kernel regression methods for the prediction of wood properties of <em>Pinus taeda</em> using near infrared (NIR) spectroscopy</td>
<td>144</td>
</tr>
<tr>
<td>7.1</td>
<td><strong>Abstract</strong></td>
<td>145</td>
</tr>
<tr>
<td>7.2</td>
<td><strong>Introduction</strong></td>
<td>145</td>
</tr>
<tr>
<td>7.3</td>
<td><strong>Material and methods</strong></td>
<td>149</td>
</tr>
<tr>
<td>7.4</td>
<td><strong>Results and discussion</strong></td>
<td>154</td>
</tr>
<tr>
<td>7.5</td>
<td><strong>Conclusions</strong></td>
<td>159</td>
</tr>
<tr>
<td>7.6</td>
<td><strong>References</strong></td>
<td>160</td>
</tr>
<tr>
<td>8</td>
<td>Determination of within-tree variation of <em>Pinus taeda</em> wood properties by near infrared spectroscopy: Whole-tree wood property maps</td>
<td>174</td>
</tr>
<tr>
<td>8.1</td>
<td><strong>Abstract</strong></td>
<td>175</td>
</tr>
<tr>
<td>8.2</td>
<td><strong>Introduction</strong></td>
<td>175</td>
</tr>
<tr>
<td>8.3</td>
<td><strong>Material and methods</strong></td>
<td>178</td>
</tr>
<tr>
<td>8.4</td>
<td><strong>Results and discussion</strong></td>
<td>182</td>
</tr>
<tr>
<td>8.5</td>
<td><strong>Conclusions</strong></td>
<td>187</td>
</tr>
<tr>
<td>8.6</td>
<td><strong>References</strong></td>
<td>188</td>
</tr>
<tr>
<td>9</td>
<td>Final remarks and general conclusions</td>
<td>199</td>
</tr>
</tbody>
</table>
Chapter 1

Nondestructive techniques for the estimation of wood properties:

Introduction

Traditional methods for the determination of almost all wood properties are costly, time consuming, and are based on samples obtained from trees that have been destructively sampled. These restrictions limit the number of samples that can be analyzed rendering research that relies on large data sets impossible. Examples of studies that are limited in their scope, especially by the cost of the analyses, are those examining the within-tree variation of wood properties and those related to the genetics of wood properties. During the past decades, a significant effort has been made to implement nondestructive evaluation techniques in standing trees capable of predicting wood properties in a rapid and cost-efficient way and on a large-scale. Included are acoustics, near infrared (NIR) spectroscopy, and SilviScan. A description of each of these techniques as well as the advantages and disadvantages of them are briefly addressed.

1.1 Acoustics

In the recent years, the development of robust, inexpensive, field-based tools have positioned acoustics as one of the best options for assessing stiffness in a large number of trees (Chauhan et al., 2006). Stiffness or modulus of elasticity (MOE) provides a measure of the resistance of a wood member to deformation and it is often determined by a static bending test in which a wood section of fixed dimensions and moisture content is subjected to a known load.

The time-of-flight (ToF) method is the most rapid and least destructive acoustic technique available and is especially suited for assessing stiffness in standing trees (Wang et al., 2000).
Examples of ToF instruments are the TreeSonic and FAKOPP-2D tool (e.g. Kumar et al., 2002; Chauhan and Walker, 2006; Grabianowski et al., 2006; Mahon Jr. et al., 2008), the fibre-gen ST300 (e.g. Roth et al., 2007; Auty and Achim, 2008; Cherry et al., 2008) and the TreeTap tool (e.g. Lasserre et al., 2004, 2005; Toulmin and Raymond, 2007).

Depending on the device utilized, two or three probes (starting, transmitting, and receiving probes) are inserted into the stem separated by a pre-specified distance, generally 1 m centered on breast height, and a stress wave is induced in the tree by hitting the starting or transmitting probe with a hammer. The receiving probe detects the stress wave and the ToF is determined. As the distance between the probes is known the acoustic velocity can be determined using the one-dimensional relationship:

\[ E_D = \rho \times V^2 \]  

(1.1)

where \( E_D \) is dynamic stiffness (Pa), \( \rho \) is the green density of the wood (kg/m\(^3\)) and \( V \) is the acoustic velocity (m/s).

Advantages of this approach include:

- Applicability to standing trees, giving nondestructive estimates of stiffness;
- \( \rho \) is relatively constant among trees which implies that \( V \) is sufficient to evaluate standing trees and stands for their structural value; and
- A rapid technique, i.e. it allows many trees to be tested daily.

Disadvantages include:

- Provides an estimate of MOE of the outermost wood only (Auty and Achim, 2008; Raymond et al., 2008);
- ToF velocities are faster than resonance velocities obtained from logs, which correlate well with MOE determined by conventional static bending tests (benchmark); hence \( V^2 \)
calculated from ToF measured in standing trees will overestimate the “true” stiffness (Andrews, 2003). Conversion between ToF and resonance velocities is required if a realistic estimation of stiffness is sought; and

- Seasonal moisture content variation may influence ToF readings.

Several studies have noted that a skewed relationship exists between tree and log acoustic measurements (e.g. Chauhan and Walker, 2006; Wang et al., 2007b); thus acoustic velocity derived from ToF readouts must be interpreted differently when assessing wood stiffness in standing trees (Wang et al., 2007c). In addition, experimental results have shown that $E_D$ calculated from stress wave velocity and green wood density (Eq.1.1) increases as moisture content increases (Gerhards, 1975; Wang and Chuang, 2000) which disagrees with the data observed from static tests, making it necessary to account for differences in moisture content of the wood when acoustic MOE is calculated. Similarly, the assumption of a constant green wood density and its impact on predicted MOE values needs to be evaluated (e.g. Wielinga et al., 2008). The analysis of the assumptions behind this technique and the steps required for an appropriate analysis of ToF data collected in standing *Pinus taeda* trees are covered in Chapter 2.

1.2 SilviScan

SilviScan is an automated system that can be used to estimate a wide range of wood properties. Analysis is based on radial strips cut from increment cores or disks. Strip dimensions are 2 mm tangentially × 7 mm longitudinally, with the radial dimension being determined by the pith-to-bark length of the sample. The following wood properties can be measured:

- Air-dry density (kg/m$^3$) in 25 µm steps using X-ray densitometry.
- MFA (deg) over 0.1 mm intervals using scanning X-ray diffractometry (Evans, 1998, 1999).
• Wood stiffness (GPa) estimated at the same resolution as MFA by combining X-ray densitometry and X-ray diffraction data (Evans, 2006).

SilviScan determines stiffness based on the relationship:

\[ E_D = A(I_{CV} \times D)^B \]  \hspace{1cm} (1.2)

where \( A \) and \( B \) are a scaling and curvature parameters, respectively, \( D \) is the air-dry density (kg/m\(^3\)) of the sample, and \( I_{CV} \) (dimensionless) is the coefficient of variation of the amplitude of the azimuthal X-ray diffraction intensity profile (Evans, 2006). Resolution of MOE estimates is limited by the resolution of the \( I_{CV} \) measurements and can range from 0.1 (determined by the diameter of the X-ray beam) to 10 mm (limited to avoid excessive contribution of fiber orientation variation to MFA estimates).

SilviScan also utilizes an image analysis system to determine radial tracheid dimensions (\( R \)) and tangential tracheid dimensions (\( T \)). Tracheid coarseness (\( C \)), perimeter (external perimeter of rectangular tracheid cross-section, \( P \)) and tracheid wall thickness (\( w \)), are determined from relationships that have been in use in various forms for several decades (Evans, 1994):

\[ P = 2 \times (R + T) \]  \hspace{1cm} (1.3)

\[ C = R \times T \times D \]  \hspace{1cm} (1.4)

\[ w = P/8 - \sqrt{(P^2/16 - C/d)/2} \]  \hspace{1cm} (1.5)

where \( d \) is the cell wall density, approximately 1500 kg/m\(^3\) for all softwoods (Kellogg et al., 1975).

Advantages of this approach include:

• Pith-to-bark wood property profiles can be obtained;
• Wood property variation within rings and for early- and latewood can be examined at high resolution;

• Analysis time is relatively quick considering the amount of data generated (1000-3000 individual measurements/day); and

• Non-destructive evaluation of trees is possible because samples are usually cut from increment cores.

Disadvantages include:

• Analysis, particularly at high resolution, is relatively expensive on a whole-sample basis (not necessarily per measurement point);

• Sample preparation is time-consuming; and

• Dynamic MOE values are obtained; however SilviScan can be calibrated to obtain static MOE values as shown by Raymond et al. (2007).

1.3 Near infrared (NIR) spectroscopy

This technique relies on measuring diffuse reflectance NIR spectra from the surface of a wood sample. Bands in the NIR spectra of wood arise from the vibrations of chemical bonds in various components such as cellulose, lignin and extractives. Spectra that occur in the NIR region (700-2500 nm) consist of overtone and combination bands of the fundamental stretching vibrations of O-H, N-H and C-H functional groups (Osborne et al., 1993). As these bands are related to the constituents of wood, any change in them reflects changes in wood properties. Typically a characterized set of samples is used to generate a multivariate calibration model that is then used to predict the properties of interest for a new set of samples.

Several studies demonstrate that NIR spectroscopy can successfully be used to estimate a wide range of wood properties nondestructively (e.g Via et al., 2003; Kelley et al., 2004b;
So et al., 2004b; Schimleck et al., 2005b). Initially studies concentrated on wood properties directly related to some identifiable wood constituent, for example cellulose and lignin (e.g. Raymond and Schimleck, 2002; Hodge and Woodbridge, 2004; Schimleck et al., 2004b; Yeh et al., 2004), or properties directly related to such wood constituents, for example pulp yield (Michell and Schimleck, 1998a; Wright et al., 1990; Schimleck and French, 2002). Research has expanded to include physical-mechanical properties with considerable success (Hoffmeyer and Pedersen, 1995; Gindl and Teischinger, 2001; Schimleck et al., 2005b).

High resolution data provided by SilviScan (Evans, 1994) has facilitated the development of NIR calibration models for the prediction of a range of wood properties in Pinus radiata (Schimleck and Evans, 2004), Pinus taeda (Schimleck et al., 2003c; Jones et al., 2005b), and Eucalyptus nitens (Schimleck et al., 2006a). The approach of using SilviScan data for fitting the calibration models will be further explore in this dissertation.

Advantages of this approach include:

- Rapid estimates of wood properties can be obtained;
- Nondestructive technique; samples can be cut from increment cores;
- Calibrations can be based on spectra collected from a range of samples (chips, strips, disks, bolts, or short-clear samples);
- For solid wood samples surface roughness does not have a big impact on calibration performance (Schimleck et al., 2005b) and;
- A profile of radial variation of wood property variation can be obtained, resolutions of 10-, 5- and 2-mm have been reported (Schimleck and Evans, 2002b; Jones et al., 2007).

Disadvantages include:

- Calibrations are surface specific, i.e. models can only be applied to samples prepared in a similar way;
• Calibrations may be site specific; although some attempts have been made to develop global models (Jones et al., 2005b); and

• Usually calibrations based on either juvenile or mature wood only have low predictive ability because of the reduced wood property range.

NIR spectroscopy has been widely adopted in many industries for the rapid analysis of samples and quality control; however, despite its success, several important problems remain (Gómez-Carracedo et al., 2007), including:

• How to select the calibration (and validation) sample sets and how to set the proper number of samples;

• How to choose an adequate NIR data pre-processing algorithm; and

• How to assess the optimal model dimensionality.

Another issue particularly relevant to the analysis of radial wooden strips is the possible autocorrelation between consecutive sections of the strips. The problem arises partly because the cambial initials (self-perpetuating cells from which cells of the secondary xylem or wood and secondary phloem, the living tissue that carries organic nutrients, are derived) of each radial profile is influenced by its previous growth history and also because maturing wood cells experience common environmental conditions (Schimleck et al., 2005a). It is possible that autocorrelation and inflated sample size could influence the relationships that have been observed.

Each of these topics are addressed in this dissertation. Chapter 3 explores different methods for selecting calibration samples using pulp yield data measured in *Eucalyptus nitens* trees; Chapter 4 investigates if autocorrelation is important when developing NIR calibration models from radial wood strips obtained from *P. taeda* trees; Chapter 5 proposes a methodology for identifying suitable data pre-processing algorithms for the development of calibration models using green wood samples of *P. taeda*, while Chapter 6 investigates the
applicability of randomization methods that can be utilized to help identify the optimal model dimensionality.

The fitting process of NIR calibration models has been traditionally based on algorithms that assume linearity of the data (Martens and Næs, 1989). However, in some cases this assumption may not hold (Næs et al., 2002). Chapter 7 explores some alternatives for the development of calibration models based on nonlinear algorithms developed in the field of machine learning and kernel methods. Finally, Chapter 8 shows how wood properties predicted from NIR models can be combined with spatial statistics to assess the within-tree variability in *P. taeda* trees, something that has been rarely done owing to the difficulty of examining wood properties in a large number of samples.

1.4 References


Chapter 2

Relationships between acoustic variables and different measures of stiffness in standing *Pinus taeda* trees

---

2.1 Abstract

Acoustic tools are increasingly used to estimate standing tree (dynamic) stiffness; however, such measurements overestimate static stiffness, the standard technique to determine modulus of elasticity (MOE) of wood. This study aimed to identify correction methods for standing tree estimates making dynamic and static stiffness comparable. Sixty Pinus taeda trees, ranging from 14 to 19 years old, obtained from genetic tests established in the southeastern United States, were analyzed. Standing tree acoustic velocities were measured using the TreeSonic tool. Acoustic velocities were also recorded in butt logs cut from the same trees using the Director HM200. A strong but biased relationship between tree and log velocities was observed, with tree velocities 32% higher (on average) than the corresponding log velocities. Two correction methods, one for calibrating tree velocities and one for accounting for differences in wood moisture content, were used to determine an adjusted MOE. After correction, adjusted MOE estimates were in good agreement with static longitudinal MOE values measured on clearwood specimens obtained from the trees and no systematic bias was observed. The results of this study show that acoustic estimates of MOE on standing trees largely depend on how the data is processed and the reference method used.

2.2 Introduction

Mechanical properties are the most important wood characteristics for structural applications. Particularly important is the modulus of elasticity (MOE) or stiffness which provides a measure of the resistance of a wood member to deformation. It is often determined by a static bending test in which a wood section of fixed dimensions and moisture content is subjected to a known load. The procedure involves the utilization of the load-deflection relationship of a supported beam to determine static longitudinal MOE ($E_s$) (Pellerin and Ross, 2002). Despite its simplicity, this technique is costly as it requires destructive sampling of trees to provide samples for testing (Raymond et al., 2007). As a consequence research
has been undertaken to develop nondestructive evaluation (NDE) techniques capable of predicting wood stiffness in a rapid and cost-efficient way (Wang et al., 2007b). Acoustics, near infrared spectroscopy, and SilviScan have been found to be good alternatives to traditional bending tests (e.g. Schimleck and Evans, 2002b; Schimleck et al., 2005b; Wang and Ross, 2002; Evans, 2006). Of the three, the development of robust, inexpensive, field-based acoustic tools (Chauhan et al., 2006) have made it the best option for assessing stiffness in a large number of standing trees.

Standing tree acoustic measurements are obtained by hammering two sensor probes (transmitting and receiving probes) at a pre-specified distance apart into the stem, usually around breast height (1.4 m from the ground). A stress wave is induced by striking the transmitting probe with a steel hammer, and the time-of-flight (ToF) is recorded (Wang et al., 2000). The ToF approach, pioneered by Fakopp Enterprise (Chauhan et al., 2006), provides a nondestructive measurement of dynamic MOE ($E_D$) in a column of outerwood 2-3 cm thick and approximately 1 m long (Auty and Achim, 2008), calculated on the basis of a constant green wood density (Grabianowski et al., 2006; Toulmin and Raymond, 2007). Variations to this approach exist, for example the acoustic tool used and how the probes are positioned in the stem (e.g. Lasserre et al., 2004; Mahon Jr. et al., 2008; Raymond et al., 2008).

Several studies have examined the relationships between acoustic measurements on standing trees and $E_S$ of wood with mixed results. Wang et al. (2001) found that acoustic $E_D$ determined on western hemlock ($Tsuga heterophylla$) and Sitka spruce ($Picea sitchensis$) trees ranging from 38 to 70 years old, was moderately correlated with $E_S$ of short clearwood specimens obtained from the same trees ($r = 0.66$). Lindström et al. (2004) reported a strong correlation ($r = 0.89$), at a clonal mean level, between acoustic $E_D$ and $E_S$ measured by axial compression loading on wood billets cut from 3-year-old $Pinus radiata$ trees. Recently, Eckard (2007) reported only a moderate phenotypic correlation ($r = 0.67$) between stress wave velocity (squared) and $E_S$ of short clearwood specimens in 8-year-old $Pinus taeda$
clones; while Auty and Achim (2008) found a coefficient of determination of $R^2 = 0.53$ ($r = 0.73$) between $E_S$ and stress wave velocity in Scots pine ($Pinus sylvestris$) trees ranging from 45 to 72 years old. Similarly, Raymond et al. (2008) working with $P. radiata$ trees ranging from 28 to 43 years old, reported a correlation of 0.79 between short clear $E_S$ and $E_D$ measured in standing trees.

Conversely, Matheson et al. (2002) reported relationships between stress wave velocity and lumber stiffness for $P. radiata$ (approximately 30 years old) that ranged from $r = 0.01$ (no correlation) to $r = 0.33$ (weakly correlated). Kumar et al. (2002) found a weak phenotypic correlation ($r = -0.47$) between the transit time measured in 12-year-old $P. radiata$ trees and stiffness measured on small clearwood samples cut from the trees; and Cherry et al. (2008), comparing $E_S$ with acoustic $E_D$ in 25-year-old Douglas-fir ($Pseudotsuga menziesii$) trees, also found a weak phenotypic correlation ($r = 0.45$) between both traits.

Wood mechanical properties are affected by silvicultural practices, especially stand density (Wang and Ross, 2002), and acoustics have had an important role in assessing the impact of initial stocking or thinning practices on wood stiffness. For example, Wang et al. (2001) found that trees from unthinned stands had higher acoustic velocities and stiffness compared to trees from medium and heavily thinned stands. Lasserre et al. (2005) reported that planting density (ranging from 833 to 2500 trees/ha) significantly influenced stiffness in 11-year-old $P. radiata$ trees, with values in the high-density plots exceeding those observed for low-density plots by 34%. Similarly, Roth et al. (2007) reported an increase of 31% in stress wave velocity (squared) measured in 6-year-old $P. taeda$ trees when planting density increased from 1334 to 2990 trees/ha; while Waghorn et al. (2007a) found an increase of 37% in acoustic $E_D$ when stocking increased from 275 to 2551 trees/ha in 17-year-old $P. radiata$ trees.

According to Andrews (2002), it is generally accepted that ToF tools used on standing trees are less accurate than resonance tools, now commonly used for sorting logs (Carter et al., 2006). It has been noted that a biased relationship between tree and log acoustic
measurements exists (Chauhan and Walker, 2006; Wang et al., 2007c); thus acoustic velocity derived from ToF readings must be interpreted differently when assessing wood stiffness in standing trees (Wang et al., 2007b). In addition, because of the systematic bias in the relationship between ToF and resonance, Andrews (2003) concluded that squaring the ToF speed will badly overestimate wood stiffness in standing trees. Experimental results have also shown that $E_D$ calculated from stress wave velocity and green wood density increases with an increase in moisture content (Gerhards, 1975; Wang and Chuang, 2000) which disagrees with the data observed from static tests, making it necessary to account for differences in wood moisture content when acoustic $E_D$ is calculated. Similarly, the assumption of a constant green wood density and its impact on predicted MOE values needs to be evaluated (e.g. Wielinga et al., 2008). Furthermore, in most studies involving the evaluation of standing trees the inherent differences between static and dynamic estimates of wood stiffness have been overlooked which also produces a bias in the relationship between predicted and observed MOE values, as noted by Ilic (2001b) and Raymond et al. (2007).

Despite these apparent unresolved issues, standing tree acoustic tools are of great value for tree breeding as the wood properties of potential parents for future generations can be nondestructively and rapidly measured (Kumar et al., 2002). The main objective of this study was to examine the relationships between acoustic variables and wood stiffness measured in $P. \text{taeda}$ trees obtained from 3 genetic field tests established in the southeastern United States. The specific objectives were: $i$) to analyze the relationship between ToF measured on trees and resonance measured on merchantable logs cut from the trees, and $ii$) to study their relationships with the different techniques used to measure MOE (static bending and SilviScan).
2.3 MATERIAL AND METHODS

2.3.1 *P. taeda* TREES

Trees were obtained from progeny tests established by members of the North Carolina State University Cooperative Tree Improvement Program between 1987 and 1992. The field tests were located in 3 main regions (central Georgia (GA), south-west South Carolina (SC), and North Carolina (NC) Atlantic coast). A 6-tree disconnected diallel mating design was utilized to produce progeny of 12 parents and a randomized complete block design was used on each test location. Thirty full-sib families plus one check lot were available for sampling on each test site. A total of 20 trees representing different full-sib families were destructively sampled in each test (Table 2.1). Trees that were suppressed, atypical in form, or infected by fusiform rust were excluded from sampling.

2.3.2 ACOUSTIC MEASUREMENTS ON STANDING TREES

Tree acoustic velocity ($V_{ToF}$), measured in m/s, was calculated as the average of 3 consecutive ToF readouts (µs/m) obtained on each tree with the Fakopp TreeSonic microsecond timer (Fakopp Enterprise, Ágfalva, Hungary). The probes were positioned on the same side of the stem, 1 m apart, at approximately 45° with respect to the main axis of the trunk, centered around breast height (1.4 m) and always on the same aspect to minimize environmental variation. Stress waves were induced by striking the transmitting probe with a steel hammer.

2.3.3 ACOUSTIC MEASUREMENTS ON LOGS AND WOOD SAMPLES

After recording the acoustic transit times, trees were measured (diameter and height), felled, delimbed, and destructively sampled. Each tree was cut into merchantable logs (approximately 4.9 m in length) with the number of logs dependent on the length of the tree. Only butt logs were considered for the analysis described in this study. Log acoustic velocity ($V_R$),
in m/s, was calculated as the average of 5 velocity readouts (in order to minimize experimental error) obtained on each log using the Director HM200 resonance tool (fibre-gen, Christchurch, New Zealand). The readouts were taken by holding the acoustic tool firmly against the large-end diameter of each log and hitting the end of the log with a steel hammer to induce the stress waves.

From each log, four 4-cm thick wood disks were extracted at 1.5 m intervals from the base and used for green density, basic density, and moisture content determination according to ASTM Standard D 2395 - 07a (2007). A duplicate disk was obtained at 1.4 m and a 12.5 mm by 12.5 mm radial section was cut from the center of this disk for SilviScan analysis. In addition, wood billets (0.6 m long) were cut from each log, at an average height of 1.82 m (±0.02 m). The billets were sawn through the center and used to obtain wood specimens of dimensions 25 mm by 25 mm by 410 mm for static bending testing, starting from the position closest to bark, according to ASTM Standard D 143-94 (2007).

2.3.4 Stiffness determinations

Static bending

Static MOE ($E_S$) of short clearwood specimens was obtained by a 3-point static bending test using a Tinius Olsen 5000 machine (Tinius Olsen Inc., Horsham, PA) as described by Schimleck et al. (2005b). Wood samples were tested at 12% equilibrium moisture content over a 355.6-mm span with center loading and the pith up until failure. A continuous load was applied at a head speed of 1.78 mm per minute. The formulas used to calculate $E_S$ are given in ASTM Standard D 143-94 (2007).

SilviScan

Radial strips, 2 mm tangentially by 7 mm longitudinally, were cut from the 12.5 mm by 12.5 mm radial sections using a twin-blade saw and used for SilviScan-3 analysis (Paprican, Vancouver, BC). The length of the radial strips varied depending on the pith-to-bark length.
of the wood sections. The strips were not resin extracted. SilviScan MOE estimates were obtained using a combination of X-ray densitometry and X-ray diffractometry data collected at 5-mm resolution using the expression:

\[ E_{D(SS)} = A(I_{CV} \times AD)^B \]  

(2.1)

where \( E_{D(SS)} \) is the dynamic MOE (GPa), \( A \) and \( B \) are a scaling and curvature parameters, respectively, \( AD \) is the air-dry density (kg/m\(^3\)) of the sample, and \( I_{CV} \) (dimensionless) is the coefficient of variation of the amplitude of the azimuthal X-ray diffraction intensity profile (Evans, 2006). All measurements were made in a controlled environment at 40% relative humidity and a temperature of 20°C.

ACOUSTICS

Traditionally, acoustic dynamic stiffness \((E_D)\), either in trees \((E_{D(ToF)})\) or logs \((E_{D(R)})\), has been calculated using the one-dimensional equation:

\[ E_D = V^2 \times \rho \]  

(2.2)

where \( E_D \) is dynamic MOE, \( V \) is longitudinal wave velocity (m/s), and \( \rho \) is green density (kg/m\(^3\)).

However, as noted by several authors, a biased relationship between stress wave velocities calculated from ToF and velocities measured by resonance is commonly observed, with tree velocities being, in general, higher than log velocities (e.g. Chauhan and Walker, 2006; Mahon Jr. et al., 2008; Wang et al., 2007b). In order to reduce these differences, Andrews (2003) and Wang et al. (2007c), developed a correction method for tree velocities based on dilatational wave theory that makes use of acoustic information collected on both trees and logs.

According to Andrews (2003), in an elastic material there are two propagation speeds, the dilatational speed and the shear speed. The speed in a rod can be viewed as the result
of multiple interferences and reflections of these two waves, and in equilibrium, the sum is a
plane wave moving along the rod at a speed \( V = \sqrt{\frac{E_D}{\rho}} \), which is simply a re-expression
of eq. 2.2. Wang et al. (2007c) point out that tree velocities are consistently higher than the
corresponding log velocities, which can be interpreted as a good indication that ToF in trees
is dominated by dilatational waves rather than one-dimensional plane waves.

The theoretical ratio between the dilatational speed (\( V_{\text{ToF}} \)) and the one-dimensional
speed, which according to Andrews (2003) is assumed to be the acoustic velocity measured
by resonance (\( V_R \)), is related to the Poisson ratio of the material (\( \nu \)):

\[
\frac{V_{\text{ToF}}}{V_R} = \sqrt{\frac{1 - \nu}{(1 + \nu)(1 - 2\nu)}}
\]  

(2.3)

A rod-like specimen subjected to uniaxial tension will exhibit some shrinkage in the
lateral direction for most materials. The ratio of lateral strain and axial strain is defined
as Poisson’s ratio. The Poisson ratio of green wood is not known but it can be numerically
solved from eq. 2.3 when both \( V_{\text{ToF}} \) and \( V_R \) have been measured on the same trees. Thus, eq.
2.2 can be rewritten to calculate an adjusted dynamic MOE (\( E'_{D(\text{ToF})} \)) in standing trees as:

\[
E'_{D(\text{ToF})} = \left( \frac{V_{\text{ToF}}}{\sqrt{\frac{1 - \nu}{(1 + \nu)(1 - 2\nu)}}} \right)^2 \times \rho
\]  

(2.4)

In addition, eq. 2.4 can be further modified to take into account the effects of moisture
content on \( E'_{D(\text{ToF})} \) by replacing \( \rho \) by the effective density of the wood (\( \rho^* \)), defined as (Wang
and Chuang, 2000):

\[
\rho^* = \rho \left\{ 1 - \frac{(1 - k)(MC - MC_{\text{FSP}})}{100 + MC} \right\}
\]  

(2.5)

where \( MC \) is wood moisture content (%), \( MC_{\text{FSP}} \) is the moisture content at fiber saturation
(assumed to be 30% in this study), and \( k \) is the mobility of the free water, defined as
the ratio of the weight of free water that vibrates in the same phase with wood cell walls
to the total weight of free water (Wang and Chuang, 2000). In general, \( k \) is determined
through simulation but for this study, $k = 0.6$ as suggested by Wang and Chuang (2000) for stress wave velocity was adopted. Replacing $\rho$ by $\rho^*$ in eq. 2.4 and including the factor $K = 9.84E-10$ to incorporate gravitational acceleration and conversion constants to express stiffness in GPa, yields:

$$E''_{D(ToF)} = K \times \left( \frac{V_{ToF}}{1-\nu} \right)^2 \times \rho \left\{ 1 - \frac{0.4(MC - 30)}{100 + MC} \right\}$$

(2.6)

which is simply an adjusted version of eq. 2.2 that takes into account differences between ToF and resonance stress wave velocities as well as differences in wood moisture content to calculate dynamic MOE from acoustic measurements in standing trees. R version 2.8.1 (R Development Core Team, 2008) was used for the analyses.

2.4 RESULTS AND DISCUSSION

2.4.1 TREE VS. LOG ACOUSTIC VELOCITIES

A strong relationship between tree and log acoustic velocities was observed, characterized by a coefficient of determination ($R^2$) of 0.81 and a root mean square error (RMSE) of 130.1 m/s. Tree velocities ranged from 2493.8 to 4484.3 m/s and were, on average, 32% higher than velocities measured on logs, which varied between 1909.9 and 3228.0 m/s. Deviation from the line of equivalence increased as velocity increased (Fig. 2.1). According to Chauhan and Walker (2006) and Grabianowski et al. (2006), the higher velocity measured by TreeSonic can be attributed to the fact that single pass transit-time velocities are sensitive to the high localized stiffness of the outerwood, whereas resonance methods, such as the Director HM200 tool, assess area-weighted cross-sectional average stiffness.

Using the TreeSonic and Director HM200 on *P. taeda*, Mahon Jr. et al. (2008) also found a mean difference between tree and log velocities of over 30%. However, Chauhan and Walker (2006) and Wang et al. (2007c), reported differences lower than 32%. These authors used the Director HM200 for assessing log velocities but the Fakopp 2D (Chauhan and Walker, 2006)
and a prototype acoustic tool (Wang et al., 2007c) were used to measure tree velocities. In addition, these studies were conducted on different species and over a greater age range. Thus, the magnitude of the mean difference between tree and log velocities appears to be associated to the instrument and material utilized.

To explore the relationship between tree and log acoustic velocities further, 69 data points obtained from the work of Mahon Jr. et al. (2008) were added to the set of 60 trees utilized in this study. This data set consisted of acoustic velocities measured on *P. taeda* trees (ranging from 13 to 22 years-old) and velocities measured on butt logs cut from the same trees. The relationship between tree and log acoustic velocities for the “enhanced” data set ($n = 129$) was similar to that found for the original 60 trees, with an $R^2 = 0.80$, RMSE = 153.1 m/s, and an average difference of 32% between ToF- and resonance-based stress-wave velocities.

In addition, and despite the reduced number of data points, analysis by site gave similar results for GA ($R^2 = 0.82$; RMSE = 123.7 m/s) and NC sites ($R^2 = 0.86$; RMSE = 109.6 m/s). However, a drop in the association ($R^2 = 0.57$; RMSE = 141.5 m/s) was observed for the SC sites, presumably because of the reduced velocity range. Considering the individual test sites, i.e. site-age combination, relationships ($R^2$) between tree and log velocities ranged from 0.71 to 0.97, the exception being the 15-year-old trees from SC that showed an $R^2 = 0.38$.

Some studies have reported a negative relationship between acoustic velocity (or stiffness derived from it) and breast-height diameter (DBH) (e.g. Lasserre et al., 2005; Chauhan and Walker, 2006). Furthermore, a method for adjusting ToF velocity based on a simple nonlinear model with DBH and $V_R$ as independent variables has been proposed (Wang et al., 2007b,c). However, in this study such relationships were not observed for trees ($R^2 = 0.02$) or logs ($R^2 = 0.05$) and in all cases the estimated coefficient associated with DBH using the nonlinear model were not significantly different from zero. Similar findings were reported by Mahon Jr. et al. (2008).
From these results we can conclude that the inherent differences between $V_{ToF}$ and $V_R$ will always result in standing tree estimates of stiffness ($E_{D(ToF)}$) higher than those observed in logs ($E_{D(R)}$).

### 2.4.2 Adjusted tree acoustic velocity

To obtain adjusted tree velocities ($V_{ToF}'$) using the dilatational theory approach proposed by Andrews (2003) and Wang et al. (2007c), Poisson ratios ($\nu$) derived from eq. 2.3 were calculated for each site-age combination (Table 2.2). The average $V_{ToF}/V_R$ ratio was 1.32 ($\pm 0.01$) and the resultant $\nu$ was 0.366. This value was very close to the ratio of 0.370 suggested for dry wood in softwoods and hardwoods (see Wang et al., 2007c).

As shown in Table 2.2, $\nu$ ranged from 0.358 to 0.376 and was consistently lower in younger trees. After correction, the relationship between tree and log acoustic velocity hardly changed ($R^2 = 0.81$; RMSE = 129.5 m/s); however, most of the bias initially observed was removed (Fig. 2.2). Adjusted tree velocities ranged from 1932.3 to 3313.2 m/s and the average difference of 32% dropped to almost 0 (0.02%).

These results suggest overestimation of $E_D$ when assessing standing trees can be greatly reduced or eliminated by correcting $V_{ToF}$ based on $V_R$, i.e. $E_{D(ToF)}' \approx E_{D(R)}$. Further analysis indicates that individual estimates of Poisson ratios may not be required, as the relationship between $V_{ToF}'$, using individual $\nu$ values, and $V_{ToF}$, using the overall ratio $\nu = 0.366$, was very strong with $R^2 = 0.98$ and RMSE = 50.11 m/s.

### 2.4.3 TreeSonic velocity and outerwood stiffness

In the literature, it is frequently mentioned that the stress wave technique when applied to standing trees measures the MOE of the outermost rings only. For example, Grabianowski et al. (2006) found that standing tree acoustic measurements were strongly associated with acoustic measurements obtained on lumber cut adjacent to the bark ($R^2 = 0.89$) and moderately correlated with corewood velocities ($R^2 = 0.74$).
In this work, TreeSonic velocities measured on standing trees in combination with green wood density values, derived from X-ray densitometry data collected on each segment, and the average moisture content of the wood measured at breast height, were used to calculate stiffness at breast height ($E_{D(ToF)}$) using eq. 2.2 in consecutive 10-mm steps measured from bark to pith. The association between $E_{D(ToF)}$ and the weighted SilviScan MOE ($E_{D(SS)}$), for the different cumulative distances, was moderate ranging from $R^2 = 0.64$ (RMSE = 2.0 GPa) to $R^2 = 0.73$ (RMSE = 1.3 GPa), as shown in Figure 2.3.

Fig. 2.3 shows that, as the amount of transition wood and/or corewood increases, the relationship between $E_{D(ToF)}$ and $E_{D(SS)}$ increasingly diverges from the line of equivalence. After a distance of 3 cm, the degree of association between the two estimates remained relatively constant with a small drop in $R^2$ when the whole breast-height section was used, mainly as a consequence of the reduction in the range of weighted stiffness values calculated from SilviScan. This figure also illustrates one of the main assumptions behind the ToF method when applied to standing trees: single pass transit-time velocities are sensitive to the high localized stiffness of the outerwood.

It is important to note that $E_{D(ToF)}$ and $E_{D(SS)}$ are both measures of dynamic MOE and, for this reason, the relationship between these estimates is less biased than the one observed between $E_{D(ToF)}$ and $E_{D(R)}$. However, SilviScan uses air-dry density and the amplitude of the azimuthal X-ray diffraction intensity profile of the sample to calculate stiffness while $E_{D(ToF)}$ uses green density and acoustic velocity. So, for any given distance from the bark, the dynamic MOE is calculated using different wood properties, and care must be taken when comparing both estimates.

These results also suggest that SilviScan can be used as an alternative to traditional static bending tests. An advantage is that complete radial profiles for MOE can be obtained, while the main disadvantage is cost. For $P. \ radiata$, Raymond et al. (2007) found a very strong association ($R^2 = 0.93$) between SilviScan MOE measured on subsamples obtained from both ends of static bending specimens. In this work, the association between $E_{D(SS)}$,
measured on the radial sections obtained at breast height, and $E_S$ measured on static bending specimens was strong ($R^2 = 0.83$; RMSE = 0.93 GPa) but not as high as the one reported by Raymond et al. (2007), probably because stiffness values were measured in this study at different heights and on different samples (not cut from the end of static bending samples as per Raymond et al. (2007)).

2.4.4 Breast-height vs. whole-log wood properties

To incorporate the concept of effective density (eq. 2.5) in the calculation of $E''_{D(ToF)}$ (eq. 2.6), estimates of $\rho$ and $MC$ are required. Measurements of these wood properties are usually taken at breast height for practical reasons, although as Auty and Achim (2008) point out, this may raise questions about the applicability of the results to describe whole stem properties.

Following Downes et al. (1997), log green density ($\rho_L$) and log moisture content ($MC_L$) were calculated as volume weighted averages of the mean sectional properties derived from the disks extracted at 1.5 m intervals. For breast-height wood properties, the values of green density ($\rho_{DBH}$) and moisture content ($MC_{DBH}$) measured on additional disks obtained at 1.4 m were used. A summary of these properties is given in Table 2.2. Very strong associations between $\rho_L$ and $\rho_{DBH}$ ($R^2 = 0.82$, RMSE = 14.8 kg/m$^3$), and between $MC_L$ and $MC_{DBH}$ ($R^2 = 0.83$, RMSE = 0.6%) were found, suggesting that $E_{D(ToF)}$ can be reasonably approximated by using $\rho^*$ derived from measurements at breast height.

2.4.5 Static MOE and tree dynamic MOE

Most of the studies dealing with acoustic determinations of stiffness in standing trees assume that $E_S$ is a measure of the “true” MOE of the wood (e.g. Wang et al., 2001; Eckard, 2007; Raymond et al., 2008) and this assumption is the basis for the corrections suggested to determine $E''_{D(ToF)}$ (e.g. Andrews, 2003; Wang and Chuang, 2000; Wang et al., 2007c). In this section, a sequential analysis is given to illustrate the effects of these adjustments on the estimates of dynamic MOE obtained from TreeSonic measurements.
To obtain a representative estimate of the weighted $E_S$ in butt logs, it was defined (for this study) that a minimum of 4 static bending specimens per wood billet were required. Owing to this restriction, only 52 of the 60 trees originally sampled were used in this section (GA: $n = 17$, SC: $n = 17$, NC: $n = 18$). Trees removed ($n = 8$) were either too small to provide the minimum number of clearwood samples or the wood specimens had defects that limited their use, such as an excessive number of knots. These problems when using small trees for the determination of $E_S$ are not uncommon and as Lindström et al. (2002) point out, most of the standards for the determination of MOE in small clear specimens are applicable only to large mature trees because young or small-sized trees rarely display pronounced internodes and it is difficult to obtain more than one or two defect-free samples per billet.

A moderate and biased relationship ($R^2 = 0.65$, RMSE = 1.09 GPa) between $E_{D(ToF)}$ and $E_S$ was observed when tree dynamic MOE was calculated using eq. 2.2, $V_{ToF}$, and the average green density measured at breast height (Fig. 2.4a).

Considering the biased relationship found between $V_{ToF}$ and $V_R$, the results shown in Fig. 2.4a were not surprising. TreeSonic overestimates static MOE values in all cases and the deviation from the line of equivalence increases as stiffness increases. Wang et al. (2007c) makes the same observation and point out that tree velocities measured by the ToF method cannot be directly used for assessing the quality of wood in standing trees. However, we support the assessment of standing trees if it is carried out with the objective of providing rapid information for ranking purposes and not with the goal of finding the “true” MOE of the trees, i.e. the bias should not be a limitation for using acoustics tools such as TreeSonic. This assumes the bias is consistent from tree to tree and that no interactions exist.

After calibrating $V_{ToF}$ on $V_R$ as described in eq. 2.3, the relationship between $E_S$ and $E'_{D(ToF)}$, calculated using eq. 2.4 and the average green density measured at breast height, showed a small improvement ($R^2 = 0.70$, RMSE = 1.02 GPa). More importantly the bias of the relationship decreased (Fig. 2.4b) compared to that observed when using $E_{D(ToF)}$. 
If, in addition, we incorporate the concept of effective density instead of green density (eq. 2.5) in the calculation of stiffness, the bias in the relationship between $E_{D(ToF)}''$, given by eq. 2.6, and $E_S$ is almost totally removed ($R^2 = 0.70$, RMSE = 1.0 GPa) as shown in Fig. 2.4c.

The initial $V_{ToF}$ values were corrected by taking into account the differences observed with the corresponding velocities ($V_R$) of the logs and the differences in moisture content of the wood. Fig. 2.4c shows that the dynamic MOE estimates obtained in standing trees, after correction, were in good agreement with the weighted static MOE values measured by static bending.

Although the degree of association between $E_{D(ToF)}''$ and $E_S$ was only moderate, the fact that the corrected relationship is close to the line of equivalence, suggests that further improvements on the degree of association between the variables can be achieved through a better control of other sources of experimental error, such as the technique used for recording the ToF in the field and the preparation of the static bending samples.

2.5 Conclusions

The results of this study showed that acoustic velocities derived from ToF measurements in standing trees can be successfully used for the NDE of wood stiffness when tree velocities are adjusted for differences between dilatational waves (measured in trees) and resonance waves (measured in logs) and after accounting for differences in wood moisture content of the trees.

After correction, adjusted dynamic MOE estimates from tree acoustic measurements were in good agreement with static MOE values measured on small clearwood samples obtained from the same trees. The bias frequently reported in the literature, when comparing these two measures of stiffness, was not observed.
Acoustic estimates of MOE on standing trees largely depend on how the data is processed and the reference method used. There are obvious differences between dynamic and static MOE estimates that suggest that a direct comparison of values may not be appropriate.

When comparing dynamic estimates of MOE, it was found that stiffness estimated from acoustic velocity in trees was well correlated with stiffness measured by SilviScan for the 3 cm of wood closest to the bark. At greater distances, the effects of increasing amounts of transition and corewood in the samples was reflected as an increased bias in the relationship between the MOE estimates (although not necessarily in the degree of association given by $R^2$).

In order to implement the adjusted ToF method to obtain a measure of stiffness in standing trees, estimates of green density and moisture content are required. These wood properties can be measured on increment cores extracted from a subsample of the trees under assessment. The results presented in this work are based on the average green density and moisture content measured on each test site which suggests that individual wood properties are not required to get accurate MOE estimates from tree acoustic velocities. The adjustment methods described in this work can be easily implemented to assess a large number of individuals in tree breeding programs for selection based on wood stiffness.

2.6 References


Table 2.1. Individual tree mean characteristics on each test site. Standard errors are given in parenthesis.

<table>
<thead>
<tr>
<th>Test</th>
<th>Location</th>
<th>Physiographic region</th>
<th>Age (years)</th>
<th>n</th>
<th>DBH (cm)</th>
<th>HT (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GA</td>
<td>Piedmont</td>
<td>15</td>
<td>10</td>
<td>18.8 (0.6)</td>
<td>13.8 (0.2)</td>
</tr>
<tr>
<td>2</td>
<td>GA</td>
<td>Piedmont</td>
<td>16</td>
<td>10</td>
<td>19.0 (0.5)</td>
<td>15.7 (0.1)</td>
</tr>
<tr>
<td>3</td>
<td>SC</td>
<td>Coastal plain</td>
<td>14</td>
<td>10</td>
<td>22.7 (0.7)</td>
<td>20.1 (0.4)</td>
</tr>
<tr>
<td>4</td>
<td>SC</td>
<td>Coastal plain</td>
<td>15</td>
<td>10</td>
<td>23.9 (0.5)</td>
<td>20.2 (0.5)</td>
</tr>
<tr>
<td>5</td>
<td>NC</td>
<td>Coastal plain</td>
<td>18</td>
<td>10</td>
<td>23.9 (1.2)</td>
<td>19.0 (0.4)</td>
</tr>
<tr>
<td>6</td>
<td>NC</td>
<td>Coastal plain</td>
<td>19</td>
<td>10</td>
<td>22.2 (1.4)</td>
<td>19.7 (0.6)</td>
</tr>
</tbody>
</table>

**Note:** DBH, diameter at breast height (1.4 m); HT, total height.
Table 2.2. Estimated Poisson ratios ($\nu$) and measured wood properties for each test site. Standard errors are given in parenthesis. The subscripts DBH and L indicate properties measured at breast height and in logs, respectively.

<table>
<thead>
<tr>
<th>Test</th>
<th>$V_{ToF}/V_R$</th>
<th>$\nu$</th>
<th>$\rho_{DBH}$ (kg/m$^3$)</th>
<th>$\rho_L$ (kg/m$^3$)</th>
<th>$MC_{DBH}$ %</th>
<th>$MC_L$ %</th>
<th>$\rho^*$ (kg/m$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.29 (0.03)</td>
<td>0.358</td>
<td>965.8 (8.6)</td>
<td>952.2 (9.5)</td>
<td>124.1 (5.6)</td>
<td>130.7 (4.2)</td>
<td>803.6</td>
</tr>
<tr>
<td>2</td>
<td>1.35 (0.03)</td>
<td>0.374</td>
<td>962.0 (9.2)</td>
<td>943.5 (8.6)</td>
<td>121.2 (4.9)</td>
<td>127.0 (4.0)</td>
<td>803.3</td>
</tr>
<tr>
<td>3</td>
<td>1.30 (0.02)</td>
<td>0.360</td>
<td>966.2 (11.1)</td>
<td>963.5 (11.1)</td>
<td>123.7 (5.0)</td>
<td>128.3 (4.7)</td>
<td>804.3</td>
</tr>
<tr>
<td>4</td>
<td>1.33 (0.03)</td>
<td>0.369</td>
<td>990.0 (9.9)</td>
<td>986.1 (10.3)</td>
<td>117.7 (5.5)</td>
<td>126.3 (5.4)</td>
<td>830.5</td>
</tr>
<tr>
<td>5</td>
<td>1.29 (0.01)</td>
<td>0.358</td>
<td>1007.0 (7.8)</td>
<td>998.0 (5.1)</td>
<td>110.4 (3.2)</td>
<td>120.7 (3.7)</td>
<td>853.1</td>
</tr>
<tr>
<td>6</td>
<td>1.35 (0.02)</td>
<td>0.376</td>
<td>1018.2 (5.5)</td>
<td>1002.8 (5.8)</td>
<td>103.0 (2.0)</td>
<td>111.3 (2.7)</td>
<td>871.7</td>
</tr>
</tbody>
</table>

Note: $V_{ToF}$, tree velocity (m/s); $V_R$, log velocity (m/s); $\rho$, green density; $MC$, moisture content (dry basis); $\rho^*$, effective density at breast height.
Figure 2.1. Comparison between acoustic velocity on standing trees ($V_{ToF}$) and in butt logs cut from the same trees ($V_R$) obtained from three genetic field tests (GA, SC, NC). The segmented line represents the line of equivalence.
Figure 2.2. Relationship between acoustic velocity of butt logs ($V_R$) and adjusted acoustic velocity of standing trees ($V'_{ToF}$) corrected by using the dilatational theory approach obtained from three genetic field tests (GA, SC, NC). The segmented line represents the line of equivalence.
Figure 2.3. Relationship between dynamic MOE estimated from ToF velocities measured on standing trees \( (E_{D(ToF)}) \) and weighted dynamic MOE at breast height derived from SilviScan analysis \( (E_{D(SS)}) \) for different cumulative distances \( (d) \) from bark to pith. The segmented line represents the line of equivalence.
Figure 2.4. Relationship between static MOE ($E_s$) and dynamic MOE estimated from unadjusted ToF velocities (a), adjusted ToF velocities by Poisson ratio (b), and adjusted ToF velocities by Poisson ratio and moisture content in trees obtained from three genetic field tests (GA, SC, NC). The segmented line represents the line of equivalence.
Chapter 3

On the selection of samples for multivariate regression analysis: application to near infrared (NIR) calibration models for the prediction of pulp yield in *Eucalyptus nitens*¹

3.1 Abstract

The effects of using reduced calibration sets on the development of near-infrared (NIR) calibration models for the prediction of kraft pulp yield in *Eucalyptus nitens* (Dean & Maiden) trees were explored. Three selection techniques based on NIR spectral data (CADEX (computer-aided design of experiments), DUPLEX, and SELECT algorithms) and one selection method based on a measured property (RANKING algorithm) were used for analysis and compared against a model using all data. The effect of using calibration sets of different sizes was also evaluated. All sample selection methods resulted in models of similar performance compared to the model fitted using all samples. For calibration purposes, RANKING selection resulted in models with the lowest errors of cross-validation, followed by the DUPLEX, CADEX, and SELECT methods. In terms of validation, the SELECT and CADEX methods resulted in lower errors of prediction compared to the DUPLEX and RANKING algorithms. In general, cross-validation and prediction errors decreased as the number of calibration samples increased. These results show that it is possible to obtain adequate NIR calibration models with a reduced number of samples allowing the remaining samples to be used for model validation and that sample selection based on NIR spectral data alone is as successful as selection based on a measured property.

3.2 Introduction

In near-infrared (NIR) analysis, a multivariate calibration model represents a mathematical expression that relates component concentrations to the absorbances of a set of known reference samples at more than one wavelength or frequency (Mark and Workman Jr., 2007). To obtain adequate multivariate calibration models, we must use efficient estimation procedures and have good calibration data (Næs et al., 2002). In general, the more samples that are used for calibration, the better the model is. However, not only the number of samples is important. What the samples represent (e.g., many samples included in a calibration set
may be very similar and add little new information) and how they are selected are also of importance (Isaksson and Næs, 1990).

Ideally, because the final use of calibration models is usually for the prediction of the concentrations of new samples, they must include all possible sources of variation that can be encountered later in real applications. According to De Maesschalck et al. (1999b), a random selection of samples for calibration may approach a normal distribution; however, a procedure that selects samples equally distributed over the factor space can lead to a flat distribution, which is more favourable from a regression point of view.

A mandatory step in the process of building any model is the statistical validation of it. Validation has been defined as the comparison of a models predictions with the observed values (Mayer and Butler, 1993). The most effective method of validating a model with respect to its prediction performance is to collect new data and to compare predictions against measured values. However, in many situations, new data are not always available at the time of model development, and it is necessary to use alternative validation methods. A reasonable approach is to use an appropriate sample selection method to split the data set into two parts: a calibration set and a validation set (Montgomery et al., 2001).

When the property of interest has already been measured, it is possible to rank the sample set in ascending order and split the data set into two uniformly distributed subsets. However, in practice, this rarely occurs, and usually, calibration samples must be selected based on their spectral information under the assumption that they will adequately represent the variation of the property of interest present in the population. In NIR analysis, the selection of a subset of samples for calibration with “unique” spectral characteristics (Honigs et al., 1985) can significantly reduce the cost of laboratory analysis required to obtain the reference values that will be used later to build the regression models.

Several algorithms for the selection of representative subsets for experimental design and regression validation have been published since the early 1970s. The most popular method is the uniform mapping algorithm or computer-aided design of experiments (CADEX) method.
developed by Kennard and Stone (1969). The procedure consists of selecting for calibration, from a list of candidate objects, the sample that is most distant (using the Euclidean distance or any other distance measure) from the samples already in the calibration set. As a starting point, the two objects that are the most distant from each other are initially selected, and the algorithm then proceeds to fill up the factor space using the remaining candidate objects. Another popular method is the DUPLEX algorithm (Snee, 1977). In this procedure, the starting points for the calibration set are selected in the same way as in the CADEX algorithm; from the remaining list of candidate objects, the two points that are farthest away from each other are included in the prediction set. The procedure continues assigning the remaining points alternatively into the calibration and validation sets according to their distances to the points already assigned to the sets. This approach selects calibration and validation data sets that tend to cover the factor space in a similar fashion.

Næs (1987) suggested a procedure based on cluster analysis for selecting calibration samples. This algorithm starts with a principal component analysis (PCA) on the original data set to obtain the matrix of scores for the number of components retained in the model. A cluster analysis is carried out on the scores matrix until the number of clusters matches the number of calibration samples desired. From each cluster, one sample is selected for calibration (usually the object that is furthest away from the mean).

Puchwein (1998) developed a procedure where the samples are first sorted based on Mahalanobis distances (De Maesschalck et al., 2000) to the centre of the data set, and the most extreme point is selected for calibration. Next, a critical distance is defined so that all samples with a distance to the selected point less than the limiting value are excluded. The most extreme object among the remaining samples is selected, and the procedure is repeated until there are no data points left. A similar approach is used in the SELECT algorithm employed by WinISI software (Shenk and Westerhaus, 1991b), where the samples are chosen based on their spectral information but using a standardized form of the Mahalanobis distance between pairs of data points.
Recently, Daszykowski et al. (2002) compared several algorithms for representative subset selection, concluding that uniform designs can deal with any kind of data structure and consequently, methods such as the CADEX selection and the OptiSim algorithm (Clark, 1997), a faster version of the Kennard and Stone (1969) method, seem to be preferable to cluster-based designs (e.g., K-means, Næ's method).

Independent of the method chosen to split the data set, the decision of how many samples to include in the calibration set will always depend on the analyst and will be closely related to the final complexity of the model (De Maesschalck et al., 1999b).

To our knowledge, there are no references in the wood properties literature on the use of sample selection algorithms for model fitting; hence, the objectives of this study were (i) to explore the use of sample selection techniques in the development of NIR calibration models and (ii) to evaluate the effects of using different calibration set sizes on the fit statistics and predictive ability of the resulting models. Three spectral-based selection algorithms (the CADEX, DUPLEX, and SELECT methods) and one constituent-based selection method (the RANKING algorithm) were used for the analyses.

3.3 Material and methods

Data set

The data set used in this work was utilized by Schimleck et al. (2005e) for the development of whole-tree kraft pulp yield calibration models for Eucalyptus nitens (Dean & Maiden) Maiden using NIR spectroscopy. Briefly, a total of 126 E. nitens whole-tree composite chip samples, each representing a single tree, were used to develop the models. Representative samples of the whole-tree composite were removed for analysis by NIR spectroscopy and for the determination of kraft pulp yield. The samples were dried to a nominal moisture content of 10% and milled in a Wiley mill through a 1 mm mesh screen. Diffuse reflectance NIR spectra were collected from each sample over the range of 1100-2498 nm at 2 nm intervals.
using a FOSS NIRSystems Model 5000 scanning spectrophotometer (FOSS NIRSystems, Inc., Laurel, Md.).

**Preprocessing and PCA**

The NIR data was converted to second derivative using the Savitzky-Golay filter (Savitzky and Golay, 1964) with left and right gaps of 8 nm and second order degree polynomial as in Schimleck et al. (2005e) and compressed later by PCA using singular value decomposition (Jolliffe, 2002). Leave one out (row-wise) cross-validation (Bro et al., 2008) was used to determine the number of components needed to describe the data variance. Nine components, explaining 97.1% of the total variance, were used for further calculations and outlier detection.

**Outlier detection in PCA**

For outlier detection, the $Q$ statistic and Hotelling’s $T^2$ statistic of each sample was calculated. The principal components (PC) model was expressed as

$$ X = TP^T + E $$

where $X$ is the mean-centered $n \times p$ matrix of preprocessed NIR spectra, $T$ the $n \times k$ matrix of scores, $P$ the $p \times k$ matrix of loadings, $E$ the $n \times p$ residual matrix, $n$ is the number of samples, $p$ is the total number of wavelengths, and $k$ the number of components retained in the model. The superscript $T$ denotes transpose. In this study $n = 126$, $p = 692$, and $k = 9$.

For the $i$th sample $x_i$, the $Q$-statistic was obtained as

$$ Q_i = e_i^T e_i = x_i(I - PP^T)x_i^T $$

where $e_i$ corresponds to the $i$th row of $E$, and $I$ to the identity matrix. The $Q$ statistic indicates how well each sample conforms to the PC model (lack of fit). It can be interpreted...
as a measure of the difference between the sample and its projection onto the \( k \) principal components retained in the model.

The confidence limit for \( Q \) was obtained as (Jackson and Mudholkar, 1979):

\[
Q_\alpha = \phi_1 \left( \frac{c_\alpha \sqrt{2\phi_2 h_0^2}}{\phi_1} + 1 + \frac{\phi_2 h_0 (h_0 - 1)}{\phi_1^2} \right)^{1/h_0} \tag{3.3}
\]

where

\[
h_0 = 1 - \frac{2\phi_1 \phi_3}{3\phi_2^2} \tag{3.4}
\]

and

\[
\phi_i = \sum_{j=k+1}^{K} \lambda_j^i \quad ; i = 1, 2, 3 \tag{3.5}
\]

In eq. 3.3, \( c_\alpha \) is the normal deviate corresponding to the upper \( 1-\alpha \) percentile, set to \( \alpha = 0.05 \) in this study. In eq. 3.5, \( \lambda_j \) represents the eigenvalue associated to the component \( j = k + 1 \), and \( K \) the total number of principal components not included in the PC model.

From the measurement of the \( i \)th sample, \( x_i \), the Hotelling's \( T^2 \) statistic was obtained as

\[
T^2_i = t_i \lambda^{-1} t_i^T = x_i P \lambda^{-1} P^T x_i^T \tag{3.6}
\]

where \( t_i \) corresponds to the \( i \)th row of the \( T \) matrix and \( \lambda^{-1} \) represents the inverse of the diagonal matrix containing the eigenvalues. The \( T^2 \) statistic is a measure of the distance from the multivariate mean to the projection of the sample onto the principal components.

The confidence limit for \( T^2 \) (\( T^2_{\text{LIM}} \)) was calculated using the \( F \) distribution for the chosen level of significance \( \alpha \) with \( k \) and \( n - k \) degrees of freedom as

\[
T^2_{\text{LIM}} = \frac{k(n-1)}{n-k} F_{(k),(n-k),\alpha} \tag{3.7}
\]
Sample selection algorithms

CADEX algorithm

Following Kennard and Stone (1969), let $k$ be the number of components retained in the PC model, $n$ be the number samples, and $T$ be the matrix of scores of dimension $n \times k$. The procedure starts with a list $L$ of $n$ ($k$-dimensional) candidate samples and an empty calibration set $C$ of size $m$. The calibration samples are chosen sequentially to cover uniformly the factor space based on the Euclidean distance between objects.

The squared Euclidean distance ($D^2$) between every two candidate samples $t_v = (t_{1v}, t_{2v}, \ldots, t_{kv})$, $t_\mu = (t_{1\mu}, t_{2\mu}, \ldots, t_{p\mu})$ for $v, \mu = 1, 2, \ldots, n$ is calculated as

$$D^2_{v\mu} = \| t_v - t_\mu \|^2 = \sum_{q=1}^{k} (t_{qv} - t_{q\mu})^2$$ (3.8)

The two samples that are furthest apart from each other ($P_1^*$ and $P_2^*$) are the first two objects assigned to $C$ and deleted from $L$:

$$D^2_{\text{max}} = \max_{v, \mu \neq \mu} \| t_v - t_\mu \|^2$$ (3.9)

Let $P_1^*, P_2^*, \ldots, P_{i^*}, \ldots, P_{w^*}$, $w^* < m$, be $w$ samples that have been assigned to $C$. The remaining $(m - w)$ samples are selected from $L$ using the expression:

$$\Delta^2_{w+1} = \max_{v \neq \mu} \{ \Delta^2_v(w) \} = \max_{v \neq \mu} \{ \min_{i^*} \{ D^2_{1*}, D^2_{2*}, \ldots, D^2_{q*} \} \}$$ (3.10)

Equation 3.10 shows that the remaining calibration samples are selected from the candidate objects whose distances to the points already in $C$ are most distant (maximal). The algorithm continues until $C$ reaches the size $m$. The objects not assigned to $C$ are used as validation samples ($V$).
DUPLEX ALGORITHM

The DUPLEX method is a modification of the CADEX algorithm (Snee, 1977) developed with the objective to divide the data into sets (calibration and validation) that cover approximately the same region with similar statistical properties.

First, the Euclidean distance between all possible pairs of samples is calculated as described in eq. 3.8. The two samples which are farthest apart (eq. 3.9) are assigned to $C$ and deleted from $L$. Next, the two samples in the remaining list which are farthest apart (eq. 3.9) are assigned to $V$ and deleted from $L$. Alternation between $C$ and $V$ continues until all samples in the list have been assigned to one of the two data sets using eq. 3.10.

RANKING ALGORITHM

This algorithm simply ranks the samples in ascending order based on some given response or property value. The vector of responses is then split into two uniformly distributed subsets based on a determined number of sample quantiles. The ratio of the number of samples in the two sets ($C$ and $V$) is determined by the user.

SELECT (WinISI) ALGORITHM

This algorithm developed by Shenk and Westerhaus (1991b) eliminates samples with similar spectra from the data set. The data is first preprocessed as desired, and PCA is carried out to obtain the matrix of loadings $P$. This matrix is used to calculate the distance between all pairs of spectra using a modified version of the Mahalanobis distance. A minimum distance is set such that, if the distance between two spectra is less than this minimum, one of the spectra is deleted from the set.

Next, the algorithm identifies the spectrum that has the most neighboring spectra closer than the minimum distance. This spectrum is retained and assigned to $C$, and its neighbours are discarded. The algorithm then evaluates all remaining samples to identify the sample that has the most neighbours, and the elimination process is repeated. These steps continue until
no samples remain with neighbours closer than the defined minimum distance. As with the CADEX method, the samples not assigned to $C$ are used as validation samples ($V$).

Calibration sets

Five calibration sets per sample selection algorithm were generated. The subsets were of size $m = 63, 75, 88, 100, \text{ and } 113$ corresponding to $50\%, 60\%, 70\%, 80\%$, and $90\%$ of the total number of samples, respectively. In total, 20 calibration sets were created for fitting the partial least squares (PLS) models.

Kraft pulp yield NIR calibration models

Nonlinear iterative PLS regressions (Martens and Næs, 1989) were used to fit the calibration models using leave one out cross-validation and a maximum of six factors. The numbers of factors was set to six to facilitate the comparison with the model presented by Schimleck et al. (2005e).

Calibration performance was assessed using the calibration $R^2$ and the root mean square error of cross-validation (RMSECV). Prediction performance, calculated on the validation set $V$, was evaluated using the prediction $R^2$ and the root mean square error of prediction (RMSEP). Both RMSECV and RMSEP are expressed in pulp yield units (\%).

Following Mevik and Cederkvist (2004), the general expression for RMSECV is

$$\text{RMSECV} = \left( \frac{1}{m} \sum_{s=1}^{S} \sum_{i \in C_s} (f_s(x_i) - y_i)^2 \right)^{0.5}$$  \hspace{1cm} (3.11)

where $m$ is the number of observations in $C$, which is divided into $C_s$ segments, $s = 1, 2, \ldots, S$, $f_s$ is the predictor trained on the observations not contained in $C_s$, and $y_i$ is the reference value. The inner sum is taken over the observations in the $s$th segment. For leave-one-out cross validation, as in this work, $S = m$.

Similarly, RMSEP was expressed as
\[ \text{RMSEP} = \left( \frac{1}{n_V} \sum_{i=1}^{n_V} (f_c(x_{i,V}) - y_{i,V})^2 \right)^{0.5} \] (3.12)

where \( n_V \) is the number of observations in the validation set \( V \) \( (n_V = n - m) \), \( f_c \) is the predictor trained on \( C \), and \( y_{i,V} \) is the reference value of the \( i \)th sample in \( V \).

R version 2.7.1 (R Development Core Team, 2008) was used for the analyses. The regression models were fitted using the \texttt{pls} package (Mevik and Wehrens, 2007). The SELECT algorithm is commercially available in WinISI III version 1.63 (InfraSoft International LLC, State College, Pa.).

3.4 Results and discussion

As mentioned in the preprocessing and PCA section, nine principal components were retained in the PC model after cross-validation. This number was used to obtain the \( T^2 \) and \( Q \) plots plus their 95% confidence limits (Fig. 3.1).

The influence plot in Fig. 3.1 shows that samples 1 and 106 can be considered potential outliers. In addition, sample 34 shows a \( Q \) statistic well beyond the confidence limit. Based on these results, it was decided to remove these three samples from the list of candidate objects and assign them to the validation set \( V \) (giving a “worst case” scenario for validation). As a result, the initial sizes of \( L \) and \( V \) were set to 123 and 3, respectively.

The rows of the \( T \) matrix represent the coordinates of each sample. These coordinates were used to obtain the distances between all pairs of objects for the spectral-based algorithms (the CADEX, DUPLEX, and SELECT algorithms). For the RANKING algorithm, the vector of pulp yields was used to sort the samples in \( L \) and to generate \( C \) and \( V \).

Snee (1977) suggests that a measure of the statistical properties of \( C \) and \( V \) can be obtained by comparing the \( k \)th root of the determinants of their corresponding \( \mathbf{X}^T\mathbf{X} \) matrices, where \( k \) in this case represents the number of principal components retained in the PC model. The lower the ratio, the more similar the space covered by both data sets.
The distribution of the calibration and validation samples in the PC space and the ratio of their hypervolumes (HV ratio) for a calibration set size of 70% of the original data set size are shown in Fig. 3.2.

A visual inspection of the sample scores in Fig. 3.2 in different perspectives (not shown) and the HV ratios show that the CADEX and SELECT methods tend to split the original data set in a similar fashion, where validation samples are restricted within the cloud of calibration points (HV ratio > 3). On the other hand, the DUPLEX algorithm tends to distribute the calibration and validation samples more evenly throughout the factor space as reflected by an HV ratio < 2, with several validation samples located on the boundary of the hypervolume defined by the sample scores. Sample selection by the RANKING algorithm gave an HV ratio between 2 and 3 indicating a compromise between the CADEX and DUPLEX methods. Volume ratios between $C$ and $V$ change depending on the size of $C$, with the DUPLEX method consistently having the lowest ratios. An HV ratio of 1 (i.e., $C$ and $V$ cover the same factor space) was observed when 50% of the data was selected for calibration using only the DUPLEX method.

The fit statistics for the PLS models is given in Table 3.1.

The calibration $R^2$ had large values for six factors, ranging from 0.88 (using the DUPLEX method and 90% of the data for calibration) to 0.93 (using RANKING selection and 50% – 60% of the data for calibration). The RMSECV was < 1% in all cases with a maximum of 0.92%. RANKING selection consistently gave the lowest RMSECV followed by the DUPLEX, CADEX, and SELECT algorithms.

In general, except for the RANKING algorithm, RMSECV decreased with an increase in the number of samples used for calibration from a mean for all methods of 0.814% to a mean of 0.791%. The coefficient of variation of RMSECV calculated across sampling methods also showed a decrease as the size of the calibration sets increased, ranging from a mean of 11% for $m = 63$ (i.e., when 50% of the data was used for calibration) to a mean of 1% for $m = 113$ (i.e., when 90% of the data was used for calibration).
Of the different selection methods, the DUPLEX algorithm showed the lowest variation in RMSECV in relation to the number of samples used for calibration, followed by the CADEX, SELECT, and RANKING algorithms.

On average, across-sample selection methods and calibration set sizes, the RMSECV was 0.05 percentage points higher than the cross-validation error reported by Schimleck et al. (2005e), which used all samples. The fitted PLS models for a calibration set size of 70% of the original data set size and the calibration using all samples are shown in Fig. 3.3.

The prediction statistics are given in Table 3.2.

The prediction $R^2$ ranged from a minimum of 0.80 (using the CADEX algorithm and 50% of the data for calibration) to a maximum of 0.92 (using the DUPLEX algorithm and 90% of the data for calibration). The RMSEP varied between 0.633% and 0.901%. In general, the SELECT algorithm gave the lowest error of predictions, followed by the CADEX, DUPLEX, and RANKING methods. Similar to the error of cross-validation, RMSEP decreased as the number of samples used for calibration increased from a mean of 0.832% to a mean of 0.763%. As expected, the RMSEP coefficient of variation showed an increase as a result of the reduced number of samples used for validation when the calibration set size reached 80% - 90% of the total number of samples, ranging from a mean of 8% for $m = 63$ i.e., $n_V = 63$) to a mean of 14% for $m = 113$ (i.e., $n_V = 13$). The pulp yield predictions for a calibration set size of 70% of the original data set size are shown in Fig. 3.4.

These results show that there is a clear trade-off between obtaining good fit statistics for calibration and lower errors of prediction or validation. The RANKING and DUPLEX methods, which showed the lowest errors of cross-validation (i.e., better calibration models), also showed the highest errors of prediction. On the contrary, the CADEX and SELECT algorithms, which resulted in calibration models of apparent lower quality, showed better performance in the validation step.

Montgomery et al. (2001) point out that a potential disadvantage in splitting a data set is that there is no assurance that the samples used for validation are real extrapolation points.
In other words, there is no guarantee that the validation data set “stresses” the regression model severely enough to test its real predictive ability. As explained before, the primary focus of the CADEX and SELECT algorithms is the selection of samples for calibration giving no consideration to the samples used for validation. By definition, these two algorithms will select all the “extreme” samples for calibration (samples that are farthest apart). This can explain the high RMSECV and low RMSEP errors obtained in the calibration and validation phases, respectively. On the other hand, because of the way in which the DUPLEX and RANKING algorithms split the data set, the calibration and validation sets are likely to contain some of the extreme samples. For this reason, we can generally expect better calibration models but with higher prediction errors using these two methods, as was found in this study. These high RMSEP values may be more realistic than the RMSEP obtained using the CADEX and SELECT methods.

It was mentioned in the introduction that selection based on RANKING can only be achieved if a property of the sample set has already been measured. This imposes several practical difficulties on the analysis, such as the effort associated with obtaining large sets of samples to span the range of variation of the property under consideration, the costs of analyzing each sample for the property of interest with an accepted analytical technique, and the computational requirements to handle large amounts of data. Usually, the calibration samples are selected assuming that they will adequately represent the variation of the property of interest present in the population, i.e., a range of genotypes or sites of varying site quality. Here, we have shown that calibrations based on samples selected solely because of their spectral characteristics provide similar calibration and prediction statistics to those based on samples selected to maximize the range of a known property (in this study, pulp yield). Identification of the most suitable samples for calibration based on spectral characteristics prior to laboratory analysis would provide substantial savings in analytical costs particularly for a property such as pulp yield, which is very expensive to measure.
3.5 Conclusions

The results show that it is possible to obtain adequate NIR calibration models with a reduced number of samples allowing the remaining samples to be used for model validation. Sample selection based on NIR spectra collected from samples was as successful as selection based on measured pulp yields. The identification of suitable samples for calibration based on spectral characteristics prior to laboratory analysis would provide substantial savings in analytical costs.

All sample selection methods resulted in models of similar performance compared with the model fitted using all samples (Schimleck et al., 2005e). For calibration purposes, RANKING selection resulted in models with the lowest RMSECV, followed by the DUPLEX, CADEX, and SELECT methods. In terms of validation, the SELECT and CADEX methods resulted in lower RMSEP compared with the DUPLEX and RANKING algorithms. In general, RMSECV and RMSEP decreased as the number of calibration samples increased.

Because of the way in which samples are selected, the performance of the CADEX and SELECT algorithms may be highly influenced by the presence of extreme samples resulting in over optimistic RMSEP values. Therefore, it is recommended to remove any anomalous sample(s) from the list of potential samples for calibration before running the algorithms. The DUPLEX and RANKING methods overcome this problem by assigning some of the extreme samples to both the calibration and validation sets. For this reason, although these two selection methods showed higher errors of prediction, they probably give more realistic results.

A clear compromise between calibration and validation was evident: a good fit does not necessarily imply good predictions.
3.6 References


Table 3.1. Fit statistics for the partial least squares regression models using the different sample selection algorithms and five different calibration set sizes (m).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>63</th>
<th>75</th>
<th>88</th>
<th>100</th>
<th>113</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(50)</td>
<td>(60)</td>
<td>(70)</td>
<td>(80)</td>
<td>(90)</td>
</tr>
<tr>
<td>Calibration $R^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CADEX</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>DUPLEX</td>
<td>0.90</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>SELECT</td>
<td>0.92</td>
<td>0.89</td>
<td>0.90</td>
<td>0.90</td>
<td>0.89</td>
</tr>
<tr>
<td>RANKING</td>
<td>0.93</td>
<td>0.93</td>
<td>0.90</td>
<td>0.89</td>
<td>0.90</td>
</tr>
<tr>
<td>RMSECV (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CADEX</td>
<td>0.917</td>
<td>0.892</td>
<td>0.852</td>
<td>0.832</td>
<td>0.790</td>
</tr>
<tr>
<td>DUPLEX</td>
<td>0.845</td>
<td>0.793</td>
<td>0.789</td>
<td>0.782</td>
<td>0.786</td>
</tr>
<tr>
<td>SELECT</td>
<td>0.788</td>
<td>0.922</td>
<td>0.872</td>
<td>0.826</td>
<td>0.803</td>
</tr>
<tr>
<td>RANKING</td>
<td>0.704</td>
<td>0.708</td>
<td>0.778</td>
<td>0.815</td>
<td>0.785</td>
</tr>
</tbody>
</table>

Note: RMSECV, root mean square error of cross-validation; $n$, total number of samples (126).
Table 3.2. Prediction statistics using the different sample selection algorithms and five different validation set sizes ($n_V = n - m$).

<table>
<thead>
<tr>
<th>$n_V$ ($m$)</th>
<th>63</th>
<th>75</th>
<th>88</th>
<th>100</th>
<th>113</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>(50)</td>
<td>(60)</td>
<td>(70)</td>
<td>(80)</td>
<td>(90)</td>
</tr>
<tr>
<td>Prediction $R^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CADEX</td>
<td>0.80</td>
<td>0.82</td>
<td>0.84</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>DUPLEX</td>
<td>0.89</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>SELECT</td>
<td>0.82</td>
<td>0.86</td>
<td>0.82</td>
<td>0.86</td>
<td>0.91</td>
</tr>
<tr>
<td>RANKING</td>
<td>0.81</td>
<td>0.83</td>
<td>0.84</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>RMSEP (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CADEX</td>
<td>0.799</td>
<td>0.772</td>
<td>0.735</td>
<td>0.670</td>
<td>0.764</td>
</tr>
<tr>
<td>DUPLEX</td>
<td>0.764</td>
<td>0.828</td>
<td>0.863</td>
<td>0.901</td>
<td>0.831</td>
</tr>
<tr>
<td>SELECT</td>
<td>0.865</td>
<td>0.663</td>
<td>0.675</td>
<td>0.633</td>
<td>0.614</td>
</tr>
<tr>
<td>RANKING</td>
<td>0.901</td>
<td>0.887</td>
<td>0.856</td>
<td>0.730</td>
<td>0.842</td>
</tr>
</tbody>
</table>

Note: RMSEP, root mean square error of prediction; $n$ the total number of samples (126); $m$, number of samples used for calibration.
Figure 3.1. Results of the PCA using nine principal components. The first principal component, PC1, explained 36.88% of the variance and the second principal component, PC2, explained 31.69% of the variance. Samples with Q statistic and Hotelling’s $T^2$ statistic beyond the corresponding 95% confidence limits (broken line) are identified with their sample numbers.
Figure 3.2. Distribution of samples in the hyperspace defined by components 7, 8, and 9. Open circles represent the samples selected for calibration (when 70% of the data is used for calibration) and solid circles the samples used for validation. HV ratio, ratio between the hypervolumes defined by the calibration and validation sets after compressing the original data to nine principal components.
Figure 3.3. NIR kraft pulp yield calibration models obtained with 70% of the original data \((n = 88)\) and the different sample selection algorithms. The original fit is displayed in the top left part of the figure. SEC, standard error of calibration (%); RMSECV, root mean squared error of cross-validation (%).
Figure 3.4. NIR kraft pulp yield predictions for the 38 samples selected for validation by the different sample selection algorithms. SEP, standard error of prediction (%); RMSEP, root mean squared error of prediction (%).
Chapter 4

Near infrared calibration models for the estimation of wood density in Pinus taeda using repeated sample measurements

---

4.1 Abstract

Near infrared (NIR) diffuse reflectance was used for the estimation of air-dry density and basic density in wood radial strip samples obtained at breast height (1.4 m) from 60 *Pinus taeda* trees established in three progeny tests in the south-eastern United States. NIR calibration models were fitted using raw spectra and pre-processed spectra with second derivative, multiplicative scatter correction and orthogonal signal correction. Successful calibrations were obtained for both wood properties using data collected in consecutive 10 mm sections from the samples. Data pre-processing did not result in model improvements compared to the models fitted using raw data. The effects of using repeated measures were evaluated by incorporating serial correlation into the partial least squares regression algorithm. The empirical autocorrelation of the normalised residuals showed that serial dependence among residuals was successfully removed by using an autoregressive correlation structure of second order. However, because the initial dependence among observations was not strong, the predictions were similar using the modified algorithm to those obtained with the traditional approach. These results indicate that the use of repeated measurements does not represent a serious problem for the development of NIR calibration models for the prediction of wood properties using radial samples measured in 10 mm sections and that the specification of the correlation structure may not be required when the models are used only for predictive purposes.

4.2 Introduction

Globally, wood is a very important raw material, being used in a vast array of products. Population growth and increasing affluence continue to raise the demand for wood and wood products. In response to this pressure, forest product companies must optimise their wood utilisation practices and identify and grow the best possible trees in plantations. A rapid provision of basic wood property information is then required. Several different measurement
techniques have been utilised for this purpose of which one of the most promising is near infrared (NIR) spectroscopy (Schimleck and Workman Jr., 2004; Tsuchikawa, 2007).

In tree breeding programmes the non-destructive estimation of the wood properties of standing trees is frequently required and this is achieved by extracting an increment core from the stem and measuring its wood properties. The most common property evaluated is wood density, owing to its ease of measurement and strong relationship with a number of other properties. Wood density is often expressed in the scientific literature as basic density (oven-dry mass per green volume), while in the forest products industry, air-dry density (weight and volume are determined at the moisture content of the wood) is commonly used (Saranpää, 2003).

When NIR spectroscopy has been used for estimating the density of increment cores, the spectra have usually been collected in consecutive 10 mm sections from the radial longitudinal face of strips cut from the cores, and strong relationships between the spectral data and measured wood properties have been reported (Schimleck et al., 2003a,c; Jones et al., 2005b). However, as the density is repeatedly measured within the strips and these measurements are from adjacent sections of the samples, it is possible that autocorrelation and the inflated sample size could influence the relationships that have been observed.

In general, the presence of replicates or repeated measures in the calibration set is often ignored and it has been suggested to average them prior to analysis (Esbensen, 2002). An alternative approach was presented by Schimleck et al. (2005a), where NIR calibration models developed for the estimation of different wood properties of *Pinus taeda* using consecutive 10 mm sections were compared against models developed using unrelated 10 mm sections obtained from independent samples. The results showed that the calibration statistics were similar in both cases and the authors concluded that the success of the calibrations was not related to the presence of replicates in the data.

It is known that partial least squares regression (PLSR) has several advantages compared to other techniques for the analysis of NIR data because it can analyse strongly collinear
and noisy data with numerous predictor variables while simultaneously modelling several response variables (Wold et al., 2001a). However, the analysis of data showing some sort of specific structure (for example, grouped data, repeated measures) has not been explicitly addressed in the chemometrics literature. Some authors (Mark, 1991; Burns, 2008) have proposed the alternative of including dummy variables as part of the independent set of predictor variables or the use of functional data techniques (Saeys et al., 2008) in order to account for grouping structures in the data but there are few reported examples in which these approaches have been successfully applied.

When dependence among observations is observed or suspected, specification of the correlation structure of the variance-covariance matrix is the common approach used in the analysis of longitudinal data (one-dimensional case) (Mora et al., 2007) and in the analysis of spatial data (two-dimensional situation) (Jordan et al., 2008). Consecutive measurements within wood samples can be regarded as a longitudinal data problem. Since data collected from an individual sample tend to be more alike than different they tend to be correlated and any two measurements that are closer in time or position are likely to be more closely correlated than two measurements that are more distant. Serial correlation in wood samples exists in part because the cambial initials (self-perpetuating cells from which cells of the secondary xylem or wood and secondary phloem, the living tissue that carries organic nutrients, are derived) of each radial profile is influenced by its previous growth history and also because maturing wood cells experience common environmental conditions (Schimleck et al., 2005a).

Data of this structure accommodate analysis using linear or non-linear generalised modelling techniques which allow for the inclusion of multiple sources of variation and can account for heteroscedasticity and serial correlation. However, these types of model are not designed to handle data having numerous predictor variables such as absorbance measured at different wavelengths. It follows that a possible solution would be to combine the traditional PLSR technique with the generalised models theory. In this regard, the work developed by
Luo et al. (2007) represents an interesting alternative for the development of NIR calibrations for the estimation of wood properties based on radial strips. By using simulated data with a known correlation structure, these authors demonstrated that by combining generalised linear regression models with traditional PLSR models, a lower prediction error can be obtained in a separate validation set of similar correlation structure.

To our knowledge, there are no published references on the use of PLSR models that explicitly account for serial correlation of the data. For this reason, the primary objective of this work was to implement the approach proposed by Luo et al. (2007) for the development of NIR calibration models for the estimation of wood density in *P. taeda* radial strip samples. The specific objectives were: (1) To compare calibration models for basic density and air-dry density obtained from measurements on the same samples; (2) to compare calibration models based on raw spectra and spectra treated using three common pre-processing algorithms: derivatives, multiplicative scatter correction, and orthogonal signal correction; and (3) to evaluate if repeated measures on the same samples represents a problem in the development of calibration models based on wood radial strips by incorporating the concept of serial correlation into the model fitting process.

4.3 Material and methods

4.3.1 Data set

Wood samples were obtained from *P. taeda* progeny tests established by members of the North Carolina State University Cooperative Tree Improvement Program between 1987 and 1992 in the south-eastern United States. The tests were located in three regions (North Carolina Atlantic coast, south-west South Carolina, and central Georgia) and consisted of two disconnected diallels arranged in a randomised complete block design. Thirty full-sib families plus one check lot were available for sampling on each test site.

A total of 20 trees representing different full-sib families were destructively sampled in each region. Trees that were suppressed, atypical in form, or infected by fusiform rust were
excluded from sampling. From each selected tree, a 4 cm thick wood disc was extracted at breast height (1.4 m from the base) and a 12.5 by 12.5 mm radial strip was obtained from the centre of the disc and used for sample preparation.

4.3.2 Sample preparation and determination of wood density

Each radial strip was divided at the pith and one half was prepared for data collection. The selected half was dried at 50°C for 24 hours, glued into wood holders and sectioned along its longitudinal axis using a twin-blade saw to produce a radial longitudinal sample of approximately 2 mm thick from the centre of strip. The samples were not resin-extracted.

Wood basic density (BD) was determined by X-ray densitometry using a Quintek Measurement Systems density profiler (QMS, Knoxville TN, USA) starting from the bark end of the samples using a resolution step of 0.06 mm. Air-dry density (AD) was determined by X-ray densitometry using the SilviScan (Evans, 1994) system (SS) starting from the position closest to the pith using a resolution of 0.025 mm. The AD data was then inverted to match the BD profiles. All measurements were made at 40% relative humidity (RH) and 20°C. Both properties are expressed in kg/m³. The average density profiles for the 10 mm sections of individual samples are shown in Figure 4.1.

4.3.3 Near infrared spectroscopy

NIR diffuse reflectance spectra were obtained in consecutive 10 mm sections from the radial longitudinal face of each strip using a FOSS NIRSystems (Laurel MD, USA) Model 5000 spectrophotometer. The first spectrum was collected from the bark end of the samples to match the density profiles. The strips were held in a custom-made holder and a 5 by 10 mm mask was used to ensure that a constant area was tested (Schimleck et al., 2001b). NIR spectra were collected over the range of 1100-2498 nm with a resolution of 2 nm and the instrument reference used was a ceramic standard. The estimated instrumental noise obtained by replicate measurements of instrument reference was 1.2E-5 AU (absorbance
units). The measurements were made in a conditioned atmosphere at 40% RH and 20°C. The results of the radial density measurements and a radial strip showing a 10 mm section used to collect the NIR spectra for one sample is shown in Figure 4.2.

4.3.4 **Pre-processing algorithms and sample selection for calibration**

Raw NIR data (RAW) and pre-processed spectra using second derivative (2D), multiplicative scatter correction (MSC), and orthogonal signal correction (OSC) were used for the analyses. Each pre-treatment was applied individually to the raw data and they were not combined. Second derivative spectra were obtained using the Savitzky-Golay filter (Savitzky and Golay, 1964) with a nine-point window and second order degree polynomial as described by Wentzell and Brown (2000). The MSC correction was obtained by using the approach described by Geladi et al. (1985). In the case of OSC pre-treatment, the raw spectra were corrected by removing two orthogonal components using 50 iterations and a tolerance level of 99.9%. We refer the reader to Wold et al. (1998) for additional details.

The raw and transformed spectra were compressed by non-linear iterative partial least squares principal component analysis (NIPALS-PCA) (Martens and Næs, 1989) using and a maximum of 15 components. Sixty segments representing individual samples (of length ranging between 6 and 11 depending on the number of spectra collected per sample) were used for cross-validation to determine the optimal number of components.

For outlier detection (in the calibration phase), the $Q$-statistic and Hotelling’s $T^2$-statistic of each sample was calculated (Mora and Schimleck, 2008a). Samples identified as potential outliers were assigned to the validation set. The remaining samples were ranked by density and, based on the rank, 45 trees (15 per region) covering the entire range of air-dry density and basic density were selected for calibration. Table 4.1 gives a summary of the property analysed for each data set.
According to Pinheiro and Bates (2000), correlation structures are used to model dependence among observations but in the context of mixed-effects models and extended linear and nonlinear models, they are used to model dependence among the within-group errors. This is the main idea behind the approach proposed by Luo et al. (2007), i.e. the use of extended linear models within the NIPALS algorithm to incorporate the concept of dependence among observations into the PLSR models.

Let \( i = 1, \ldots, m \) be the \( i \)th sample in the calibration set and \( j = 1, \ldots, n_i \) the \( j \)th measurement for the \( i \)th sample. Assuming isotropy, the correlation between two within-sample errors \( \varepsilon_{ij} \) and \( \varepsilon_{ij^*} \) depends only on their one-dimensional positions \( p_{ij} \) and \( p_{ij^*} \) and not on the direction and particular values they assume. Then the general within-sample correlation structure among observations can be expressed as:

\[
\text{cor}(\varepsilon_{ij}, \varepsilon_{ij^*}) = h(|p_{ij} - p_{ij^*}|, \rho) \tag{4.1}
\]

where \( \rho \) is a vector of correlation parameters and \( h(\cdot) \) is a correlation function taking values between \(-1\) and \(1\), and such that \( h(0, \rho) = 1 \).

In the analysis of time-series data, the correlation function \( h(\cdot) \) is referred to as the autocorrelation function. The empirical autocorrelation function (ACF), the non-parametric estimation of the true autocorrelation function, is commonly used for investigating serial correlation and is defined as:

\[
\text{ACF} = \hat{\rho}(l) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i-1} r_{ij}r_{i(j+l)} / N(l)}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} r_{ij}^2 / N(0)} \tag{4.2}
\]

where \( r_{ij} = (y_{ij} - \hat{y}_{ij}) / \hat{\sigma}_{ij} \) are the standardised residuals associated to the \( m \) samples used to calibrate the model and \( N(l) \) is the number of residuals pairs used at lag \( l \).

Serial correlation can be incorporated into classical linear models through the specification of the variance-covariance matrix as (Pinheiro and Bates, 2000):
\[ y_i = X_i \beta + \varepsilon_i, \quad \text{var}(\varepsilon_i) = \Sigma_i, \quad i = 1, \ldots, m \]  

(4.3)

The model in Equation 4.3 is known as the generalised least squares problem, where \( \Sigma_i \) is a known positive-definite matrix. For such a matrix, it is always possible to find a square-root matrix \( \Sigma_i^{1/2} \) that has the property \( (\Sigma_i^{1/2})^T \Sigma_i^{1/2} = \Sigma_i \). By multiplying both sides of Equation 4.3 by \( \{ (\Sigma_i^{1/2})^T \}^{-1} \), a modified classical regression model is obtained:

\[ y_i^* = X_i^* \beta + \varepsilon_i^*, \quad \text{var}(\varepsilon_i^*) = \sigma_i^2 I, \quad i = 1, \ldots, m \]  

(4.4)

Let \( X^* \) be the matrix obtained by stacking up the \( X_i^* \) matrices, \( y^* \) the stacked vector \( y_i^* \), and \( \tilde{\Sigma} \) the block diagonal matrix with \( m \) covariance matrices of dimensions \( n_i \times n_i \) along the diagonal, then the maximum likelihood (ML) estimator for \( \hat{\beta} \) may be written as:

\[ \hat{\beta} = \{ (X^*)^T X^* \}^{-1} (X^*)^T y^* = (X^T \tilde{\Sigma}^{-1} X)^{-1} X^T \tilde{\Sigma}^{-1} y \]  

(4.5)

This estimator is called the generalised least squares estimator of \( \beta \) where \( \Sigma_i \) is not known and its parameters need to be estimated.

Following Luo et al. (2007), in the context of PLS modelling we assume that the original explanatory variables can be summarised by a smaller group of \( k \) uncorrelated factors in the form:

\[ \tilde{y} = T_{m^*,k} q + \xi \]  

(4.6)

where \( \tilde{y} \) is the centred response, \( m^* \) is the number of samples \( (m \times n_i) \), \( T_{m^*,k} \) is the scores matrix, \( q \) is the y-loadings vector, and \( \xi \) is the residual matrix under \( k \) factors.

Under this formulation, the score vectors are orthogonal to each other in the calibration set; therefore, the estimation of \( q \) can be done by a series of simple univariate regressions on the consecutive score vectors (Martens et al., 2001). By replacing Equation 4.5 with the appropriate terms, the generalised predictor of \( q \) is given by the expression:
\[ \hat{q}_{\text{GLS}} = \left( T_{m^*,k}^T \tilde{\Sigma}^{-1} T_{m^*,k} \right)^{-1} T_{m^*,k}^T \tilde{\Sigma}^{-1} \tilde{y} \quad (4.7) \]

The generalised predictor of \( q \) is also the minimum variance linear unbiased estimator (BLUE) in the generalised regression model and the transformation has no effect on the derivation of the scores because it is only applied in the second step of the bilinear model. Essentially, the matrix of scores and the vector of loadings are first obtained using the traditional PLS algorithm and then \( T_{m^*,k} \) is used to recompute \( q \) in order to obtain \( \hat{q}_{\text{GLS}} \).

In the rest of the text, the modified PLSR model will be noted as GPLSR to indicate that \( q \) is estimated using generalised linear models.

4.3.6 Wood density NIR calibration models

Non-linear iterative partial least squares regressions (NIPALS-PLSR) and generalised partial least squares regressions (GPLSR) were used to fit the calibration models with a maximum of 15 factors. Forty-five segments representing the samples in the calibration sets were used for cross-validation. The optimal number of components was determined as (Esbensen, 2002):

\[ NF = \text{Min}\{\text{RMSECV}(k) + 0.01 \times k \times \text{RMSECV}(0)\} \quad (4.8) \]

where \( NF \) represented the optimal number of factors, \( k \) the current dimensionality of the model, \( \text{RMSECV}(k) \) the root mean square error of cross-validation at dimensionality \( k \), and \( \text{RMSECV}(0) \) was the initial error. Model performance was assessed using the coefficient of determination for calibration \( (R_c^2) \), the standard error of calibration (SEC), the root mean square error of cross-validation (RMSECV), the coefficient of determination for prediction \( (R_p^2) \) and the root mean square error of prediction (RMSEP).
4.3.7 Software

R version 2.8.0 (R Development Core Team, 2008) was used for the analysis. Calibration models were fitted using the package pls (Mevik and Wehrens, 2007) and routines programmed to allow the computation of $\hat{q}_{\text{GLS}}$ at each step of the cross-validation process.

4.4 Results and discussion

4.4.1 Air-dry density, basic density, and preliminary PLSR models

The relationship between AD and BD is shown in Figure 4.3. Empirical results have suggested that BD is proportional to AD (Reyes et al., 1992) but this was not observed in this study (intercept statistically different from zero). The relationship shows that the higher the AD the higher the deviation from the one-to-one correspondence with BD.

This deviation arises because BD and AD are derived from green volume and from volume at any given moisture content, respectively. Therefore, the volume used to determine AD varies as a function of the density of wood being measured. Despite the bias, it is still possible to transform one property into the other without losing interpretability of the data. This facilitates the comparison of models fitted using either measure of wood density.

Considering that AD and BD are closely related to each other and because it is known that NIR spectroscopy is a powerful tool for measuring the physical properties of wood, it is natural to expect that calibration models based on these two measures of density will have similar performance. To confirm these assumptions, two preliminary calibration models were fitted to the density data using raw NIR spectra with PLSR. The models are shown in Figure 4.4.

The models show similar ranks and fit statistics. The absolute errors obtained from AD are higher than those for BD but expressed in relative terms (RMSECV/mean) the differences are minimised indicating that both models indeed perform similarly. The relative error for AD and BD was 8.3% and 7.8%, respectively. The plots of raw residuals against
fitted values (not shown) did not reveal any sign of heteroscedasticity. The results confirmed that, independent of the system used to measure the wood property (SS or QMS), NIR is an effective tool for the prediction of wood density from radial longitudinal samples.

4.4.2 PLSR models based on pre-processed NIR spectra

Additional calibration models were fitted using pre-processed data to evaluate if any systematic variation, unrelated to wood density, was present in the raw spectra. A summary of the fit statistics of the models is given in Table 4.2.

The coefficients of determination for calibration ($R^2_c$) varied from moderate to good. In the case of the AD, they were relatively constant across pre-treatments, ranging from 0.82 to 0.84. For the BD models the variation was higher, ranging from 0.78 to 0.87. The rank of the models was also affected by the pre-treatments, varying from one to nine, with OSC and MSC models requiring the lowest and highest number of components, respectively. The coefficients of determination for prediction ($R^2_p$) were variable for both properties, ranging from 0.73 to 0.84. Lower $R^2_p$ values were consistently observed for the AD models compared to the BD models, especially for the pre-processed data. The plot of RMSEP against the number of PLS components is shown in Figure 4.5.

The prediction error, for PLS components 1 to 15, was consistently lowest for the raw data closely followed by 2D. Figure 5 shows that OSC and MSC transformations did not perform well in this study, particularly for BD. The good performance of models fitted with raw spectra was also reported by Mora and Schimleck (2008b). In their work; however, 2D performed the worst in terms of predictive ability. These results show that pre-processing is a data-dependent problem and its application will not necessarily result in better models. The lack of significant improvement by transforming the spectra might indicate that the particle size, as a function of the surface roughness of the samples, is a potentially useful characteristic for predicting wood density.
4.4.3 Generalised partial least squares models (GPLSR)

Although pre-processing was not required in this study, the models fitted using pre-treated spectra were further utilised to show the practical effects of the approach proposed by Luo et al. (2007). The autocorrelation function given in Equation 4.2 was applied to the standardised PLSR residuals and empirical ACF plots for each model were generated (Figure 4.6).

Figure 4.6 shows some interesting results. First, some degree of autocorrelation in the PLSR residuals is observed which, at some lags, exceeds the 95% confidence limits. The upper and lower confidence bands, denoted by the segmented lines, were obtained based on the standard normal quantiles and the number of residuals pairs at the specific lags according to Pinheiro and Bates (2000). Because the number of pairs of data decreases with lag number, the correlation values calculated at greater distances tend to be less reliable than those obtained at shorter distances. For this reason some plots (for example, calibration model for BD using raw data) show high correlations at higher lags.

Second, the ACF values are influenced by the pre-processing technique. The link between ACF and pre-treatment used is given by the rank of the PLSR models which affects the values of the residuals used to calculate the autocorrelation. Figure 4.6 shows that ACF is more important for the OSC model (one factor) than for the models fitted using RAW (four factors) or 2D (three to five factors) spectra. The exception is given by the MSC models. Although the rank of the MSC models was the highest (seven for AD and nine for BD), the ACF plot still shows significant correlations at lags 2 and 4. The reason is not clear but it is possible that by transforming the spectra using MSC, artificial autocorrelation is introduced in the data. This observation has not been confirmed and additional analyses are required.

To properly account for serial correlation, the structure of the correlation matrix must be known. When prior information is not available, the approach commonly used is to fit a series of models with different correlation structures. Autoregressive and moving-average models of first and second order (AR(1), AR(2), MA(1) and MA(2)) as well as different autoregressive-
moving average models (ARMA(1,1), ARMA(2,1), etc.) were applied to the data in analysis. To assess the adequacy of these correlation structures, the normalised version of the residuals (residuals weighted by the estimated variance-covariance matrix for the within-sample errors) was used (Pinheiro and Bates, 2000). The ACF plots of the normalised GPLSR residuals are shown in Figure 4.7 indicating that, in this study, the AR(2) structure adequately represented the within-sample dependence of the density measurements.

4.4.4 Wood density predictions using PLSR and GPLSR models

PLSR and GPLSR models were used to predict AD and BD of the samples in the validation set. Predicted values obtained using the 2 models are plotted against each other in Figure 4.8.

It is clear from Figure 4.8 that serial correlation did not represent an issue in this work. The mean difference between predictions ranged from a minimum of 0.12 kg/m$^3$ to a maximum of 2.45 kg/m$^3$. Theoretically, the specification of the variance and covariance structures is required for modelling grouped data with repeated measurements. The reason for this is that $n$ dependent observations “contain less information” about the marginal mean than $n$ independent observations. However, the performance of the GPLSR approach largely depends on how strong is the within-sample serial correlation. Because the autocorrelation observed in this study was fairly weak, probably as a consequence of working with 10 mm averages instead of individual ring values, the vector of regression parameters and hence the predicted values of the test set, were almost identical for the GPLSR and PLSR models.

4.5 Conclusions

Near infrared calibration models for the prediction of the air-dry density and basic density of $P. taeda$ radial samples were successfully fitted using raw spectra and density data collected in consecutive 10 mm sections from samples in the calibration sets. The relative errors of
calibration were 8.3% and 7.8% for the air-dry density and basic density models, respectively; confirming that independent of the system used to measure these two properties, NIR represents an effective tool for their prediction.

The use of spectral correction techniques did not result in model and prediction improvements and the plot of the prediction errors obtained from a separate test set versus model dimensionality indicates that the selection of the pre-processing method is a data-dependent problem. It is possible that surface roughness plays an important role in the prediction of wood density which is lost by transforming the spectra prior to modelling.

Traditionally, NIR models used for the prediction of wood properties based on sequential measurements of solid wood samples have been fitted without considering the fact that samples in the calibration set are not independent. The grouping structure of the data is given at the level of the samples which contain a certain number, not necessarily equal, of nested observations. Because data of these characteristics tend to be correlated, it is natural to think that within-sample correlation must be included in the development of the calibrations.

Through a small modification of the traditional NIPALS algorithm for PLS regression, it was possible to account for the grouped structure of the data analysed. The difference between the modified and traditional PLS methods was given in the estimation of the loadings for the response variables which were calculated using a generalised linear model approach allowing the specification of the individual variance-covariance matrices for each sample. Because each sample had its own variance-covariance matrix, differences in the number of observations per sample were automatically considered in the analysis. By using this approach it was possible to account for within-sample autocorrelation and the analysis of the normalised residuals showed that the serial dependence initially observed was completely removed.

Although the inclusion of serial dependence is a sensible thing to do in modelling grouped data, for practical purposes if the dependence among observation is weak, and especially when
no inferences are going to be made about the vector of regression parameters, the assumption of having a set of independent samples does not appear to be a problem. In this work, a bigger autocorrelation was expected because of the nature of wood formation. However, probably as a consequence of working with 10 mm averages instead of individual rings (which is a more natural scale on which to model the correlation), the expected dependence was diminished. One factor limiting the use of NIR in studies of radial variation of wood properties on an annual-ring basis is that in order to capture the within-ring properties, a reduction in the size of the window used to collect the spectra is required. As Jones et al. (2007) point out, decreasing the size of the window decreases the area of the sample exposed to NIR energy and can affect the quality of the NIR spectrum and possibly the calibration obtained using the spectra.

In this analysis it was assumed that all the samples used for calibration were from the same group. In other words, the modification of the NIPALS algorithm does not consider the situation where samples, besides having repeated measurements, can belong to different groups (for example different physiographic regions). In some cases the presence of a nominal-level variable can also induce correlation (sometimes referred to as intraclass correlation). This results in non-independent samples (among-samples variation) and to the situation where some of the variability in the dependent variable can be accounted for simply by virtue of the fact that samples are grouped in some particular way. While autocorrelation and positive intraclass correlation have comparably adverse consequences in statistical modelling, they require different remedies. One method for dealing with intraclass correlations in traditional linear and non-linear models is the use of random coefficients and/or multi-level regression, however the applicability of these approaches to PLS modelling and chemometrics, to our knowledge, is an area that has not been explored yet.

In summary, the results indicate that the use of repeated measurements in the calibration set does not represent a practical problem for the development of near infrared calibration
models for the prediction of wood properties of *P. taeda* based on radial samples measured at intervals of 10 mm.

4.6 References


URL http://dx.doi.org/10.1007/s00226-008-0235-0


Table 4.1. Descriptive statistics of wood density for the calibration and prediction sets. All values are expressed in kg/m$^3$.

<table>
<thead>
<tr>
<th></th>
<th>Calibration Set</th>
<th></th>
<th>Prediction Set</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
<td>Mean</td>
<td>Std</td>
</tr>
<tr>
<td>Basic density</td>
<td>263</td>
<td>623</td>
<td>399</td>
<td>73</td>
</tr>
<tr>
<td>Air-dry density</td>
<td>327</td>
<td>811</td>
<td>509</td>
<td>103</td>
</tr>
</tbody>
</table>
Table 4.2. Fit statistics for the partial least squares regression models fitted using the different pre-processing algorithms. RMSECV and RMSEP are expressed in kg/m$^3$. See text for other definitions.

<table>
<thead>
<tr>
<th>Method</th>
<th>Basic density</th>
<th>Air-dry density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NF</td>
<td>$R^2_c$</td>
</tr>
<tr>
<td>RAW</td>
<td>4</td>
<td>0.83</td>
</tr>
<tr>
<td>2D</td>
<td>5</td>
<td>0.87</td>
</tr>
<tr>
<td>MSC</td>
<td>9</td>
<td>0.83</td>
</tr>
<tr>
<td>OSC</td>
<td>1</td>
<td>0.78</td>
</tr>
</tbody>
</table>
Figure 4.1. Longitudinal basic density (BD) and air-dry density (AD) profiles for the samples in the calibration and validation sets. Each line represents a sample and the values correspond to the average wood property measured in the corresponding 10 mm section.
Figure 4.2. Radial density profiles from one selected sample. Top: AD (air-dry density). Bottom: BD (basic density). The inner window shows one of the 10 mm sections used to collect the NIR data.
Figure 4.3. Relationship between air-dry density (AD) and basic density (BD) for the samples analysed. The segmented line represents the 1:1 correspondence.
Figure 4.4. PLS calibration models for AD (air-dry density) and BD (basic density) using NIR raw data.
Figure 4.5. RMSEP versus number of PLS components for AD (air-dry density) and BD (basic density) data sets and pre-processing method
Figure 4.6. Plot of the empirical autocorrelation function of the standardised residuals versus lag number obtained after fitting the PLSR models for basic density (BD) and air-dry density (AD). The segmented lines correspond to the approximate 95% two-sided critical bounds for the autocorrelation.
Figure 4.7. Plot of the empirical autocorrelation function of the normalised residuals versus lag number obtained after fitting the GPLSR models for basic density (BD) and air-dry density (AD).
Figure 4.8. Plot of the predicted basic density (BD) and air-dry density (AD) values obtained by PLSR and GPLSR models. The segmented line represents the 1:1 correspondence.
Chapter 5

Determination of specific gravity on green *Pinus taeda* samples by near infrared spectroscopy: comparison of pre-processing methods using multivariate figures of merit

---

5.1 Abstract

Near infrared diffuse reflectance was used for the determination of specific gravity in green *Pinus taeda* L. wood samples representing simulated increment cores obtained at breast height and merchantable green logs. The effects of using three pre-processing methods (second derivative, multiplicative scatter correction, and orthogonal signal correction) to reduce the scatter observed in the original spectra were evaluated. The effectiveness of each method was assessed in terms of the average predictive ability of the models and in terms of multivariate figures of merit derived from net analyte signal theory. Specific gravity was successfully modeled using green wood samples. No differences in predictive ability among models were found, although more parsimonious regressions were obtained using transformed spectra. The incorporation of figures of merit for the characterization of calibration models proved to be valuable tool for understanding the effects of the pre-processing alternatives on the final results.

5.2 Introduction

Basic information regarding wood properties is becoming increasingly important with forest products companies aiming to optimize utilization of raw materials and tree breeders seeking to improve wood properties. It is essential that such information be provided rapidly, at low cost and preferably non-destructively, however methods typically employed to measure important wood properties are time consuming, expensive and often destructive, i.e. trees have to be cut down.

Near infrared (NIR) spectroscopy provides an alternative for estimating many wood properties (Schimleck and Workman Jr., 2004; Tsuchikawa, 2007), while these studies have been successful they have largely been applied to dried samples. Ideally NIR spectroscopy would be employed to estimate the properties of green wood in the field and in real-time, however
the high and variable moisture content of wood reduces the strength of calibration statistics (Schimleck et al., 2003c, 2004a).

Few studies utilizing green wood have been published in the literature but often it has been found that calibrations based on NIR spectra collected from green wood are weaker when compared to those based on dry wood. For example, Thygesen (1994) developed calibrations for wood density using green Norway spruce disks and shavings that varied in moisture content (below fiber saturation or ‘dry’, green and fully saturated or ‘wet’). The calibrations developed using green samples provided calibration statistics that were inferior to those obtained using dry samples. In a recent study, comparing calibrations based on green wood and the same data set when dried, Schimleck et al. (2003c) found that the predictive error for air-dry density was relatively large for the green samples with an $R^2$ of 0.74 and a root-mean square prediction error (RMSEP) of 57.6 kg/m$^3$, while the dry-wood calibration demonstrated a strong predictive relationship with an $R^2$ of 0.90 and a RMSEP of 32.8 kg/m$^3$.

Calibrations developed for the analysis of green wood by NIR spectroscopy may be improved by using spectral preprocessing techniques. However, data pre-treatment methods can also remove significative information from the spectra and, therefore, it may be difficult to evaluate which method is better based only on fit statistics such as $R^2$ and RMSEP. A way to obtain better descriptions of the calibration models is to incorporate the concept of net analyte signal (NAS) and multivariate analytical figures of merit into the analysis (Lorber et al., 1997; Faber, 1999b).

Figures of merit for calibration models are well known in chemometrics (Olivieri et al., 2006). In the case of NIR, with severely overlapping spectra, it has historically been difficult to quantify these figures because of the inability to distinguish between interferences and the analyte of interest. The first attempt to apply these concepts to first-order calibration models was the work by Lorber (1986), but the results were limited to direct calibration models which require knowledge of the pure spectra of all contributing constituents in the calibration
samples. Lorber et al. (1997) later extended this work for the case of inverse first-order calibration models which are less demanding than direct calibration models because they only require the knowledge of the concentrations or properties of interest in the calibration set.

For multivariate data, NAS is defined as the part of a measured signal that is orthogonal to all interferents present in the data set. Therefore, when building a calibration model, it can be helpful to visualize the total instrument response \((r)\) as the sum of two orthogonal components: the interferences \((r^\perp)\) and the net analyte signal \((r^*)\). A linear combination of the interferents produces \(r^\perp\); therefore, the signal orthogonal to \(r^\perp\) belongs exclusively to the analyte or property of interest. Multiple algorithms for calculating NAS have been published (e.g. Lorber et al., 1997; Faber, 1998a; Bro and Andersen, 2003) and, to our knowledge, there are no references in the wood properties literature on the use of NAS for characterizing NIR calibration models.

The aim of this study was to use the concept of NAS and multivariate figures of merit to compare estimates of green *Pinus taeda* specific gravity obtained from NIR calibration models based on raw spectra and spectra treated using three common pre-processing algorithms: derivatives, multiplicative scatter correction, and orthogonal signal correction.

5.3 Material and methods

5.3.1 Data sets

**Data set 1: *Pinus taeda* green wood radial strips**

This data set was utilized by Schimleck et al. (2003c, 2004a) for the prediction of wood properties by NIR spectroscopy of twenty *P. taeda* radial strips obtained at breast height (1.4 m) from 22-year-old trees. Briefly, NIR spectra were obtained in 10-mm steps from the radial-longitudinal face of each sample and used to develop calibrations for each property. NIR spectra were collected over the range of 1100-2498 nm at 2 nm intervals using a FOSS
NIRSystems Model 5000 scanning spectrophotometer when the wood was green (dry-basis moisture content ranging from 100% to 154%). Samples were held in a custom made holder and a 5 by 10 mm mask was used to ensure that a constant area was tested. Wood properties of each sample were determined by SilviScan (Evans, 1994). From the measured properties, only specific gravity (a dimensionless quantity) was selected for the current analysis. Air-dry density was first converted to basic density using the relationship given by Evans et al. (2000) and then to specific gravity.

**Data set 2: Pinus taeda green logs**

This set corresponds to a study where the application of NIR spectroscopy for the estimation of the specific gravity of green *P. taeda* logs was investigated (dry-basis moisture content ranging from 70% to 200%). A total of 162 merchantable logs (4.9 m long) representing 81 trees ranging in age from 13 to 19 years were analyzed. On each end-diameter of the logs, eight NIR diffuse reflectance spectra (1100-2498 nm) were collected in four quadrants (two per quadrant) when the wood was green using a FOSS XDS NIR System coupled with a SmartProbe Analyzer. One spectrum representing the juvenile wood (near the pith) and one representing mature wood (near the bark) was collected per quadrant. For analysis, spectra were truncated to 1100-2280 nm owing to excessive spectral noise at higher wavelengths. The analysis presented here is based on the average spectra of each log (16 per log, 8 from each end). Specific gravity was determined by water displacement of 2.5 cm thick wood disks taken every 0.15 m along the stem.

### 5.3.2 Pre-processing methods

#### Second derivative (2D)

The NIR data was converted to second derivative using the Savitzky-Golay filter with a 9-point window (filter width) and second order degree polynomial as described by Wentzell
and Brown (2000). This is a common and very effective method used to remove both the baseline offset and the slope from a spectrum.

**Multiplicative scatter correction (MSC)**

This is a method used to correct differences in baseline offsets and path length due to differences in particle size of the samples (Geladi et al., 1985; Martens and Næs, 1989). The correction is achieved by regressing a measured spectrum against a reference spectrum and then correcting the measured spectrum using the intercept and slope of the linear fit. MSC works primarily for cases where the scatter effect is the dominant source of variability.

**Orthogonal signal correction (OSC)**

OSC is a method of spectral filtering developed by Wold et al. (1998). The aim is to construct a filter that removes from the NIR spectra only the part that is unrelated to the response vector. This is an iterative algorithm that requires three settings: (1) The number of OSC components, defining how many times the entire process will be performed, (2) the number of iterations, i.e. how many cycles will be used to rotate the initial principal component loading to be as orthogonal to the response vector as possible, and (3) the tolerance level, defining the percent variance that must be captured by the partial least squares (PLS) model of the orthogonalized scores. The authors refer the reader to Wold et al. (1998) for further details. Preliminary PLS models were used to detect the combination of values that resulted in the lowest prediction errors. The optimal values for this study were two OSC components, 50 iterations, and a tolerance of 99.9%.

5.3.3 **Multivariate figures of merit**

**Net analyte signal**

The NAS ($r^*$) vector for the $i$th sample and $k$th component was obtained using the method proposed by Bro and Andersen (2003) as:
\( r_{k,i}^* = b (b^T b)^{-1} b^T r_i \) \hfill (5.1)

where \( r_i \) is the sample spectrum and \( b \) the vector of regression parameters from the PLS calibration model. NAS vector was then transformed into a scalar form using the expression:

\[ r_{k,i}^* = \frac{r_{k,i}^T b}{\| b \|_E} \] \hfill (5.2)

where the operator \( \| \cdot \|_E \) is used to denote the Euclidean norm of any enclosed vector.

**Multivariate sensitivity**

Sensitivity \((s)\) characterizes the extent of signal variation as a function of analyte concentration. For first-order inverse calibration models, sensitivity is defined as the NAS generated by an analyte concentration equal to unity and represents the inverse of the slope of the (pseudo)univariate calibration model and was calculated as (Lorber et al., 1997):

\[ s = \frac{1}{\| b \|_E} \] \hfill (5.3)

Sensitivity is reported in units of instrument intensity per concentration. Recall that specific gravity is a dimensionless quantity though.

**Selectivity**

Selectivity \((SEL)\) gives the ability of a method for the determination of a component in a complex sample without the interference of other components. It is expressed by a scalar number between 0 and 1. A value of 0 indicates that the analysis is not possible because the analyte spectrum is a linear combination of the interference spectra while a value of 1 indicates that other components do not interfere at all (Booksh and Kowalski, 1994). For the \(i\)th sample \((x_i)\) and \(k\)th component, SEL was calculated as:

\[ SEL_{k,i} = \frac{r_{k,i}^*}{\| x_i \|_E} \] \hfill (5.4)
The signal-to-noise ratio (SNR) is defined for first-order data as the ratio between NAS and the background noise (Lorber et al., 1997) and was calculated as:

\[ \text{SNR}_{k,i} = \frac{r^*_{k,i}}{\Delta r} \]  

(5.5)

where \( \Delta r \) is a measure of the instrumental noise obtained by replicate measurements of a blank sample (instrument reference). The estimated errors were \( \Delta r = 1.2 \times 10^{-5} \text{ AU} \) (absorbance units) and \( \Delta r = 1.7 \times 10^{-5} \text{ AU} \) for the FOSS NIRSystems Model 5000 scanning spectrophotometer and the FOSS XDS NIR System, respectively.

### 5.3.4 Calibration sets

Samples from data set 1 were randomly split into a calibration set of 15 samples (98 spectra) and a prediction set of 5 samples (32 spectra). The segregation of samples was the same as that used by Schimleck et al. (2003c).

Samples from the second data set were first ranked by specific gravity (Mora and Schimleck, 2008b) and based on the rank, 64 trees (128 logs) covering the entire range of specific gravity were selected for calibration and the remaining 17 trees (34 logs) were used for prediction. Table 5.1 gives the summary of the property analyzed for each data set and their corresponding calibration and prediction subsets.

### 5.3.5 Specific gravity NIR calibration models

NIPALS partial least squares (PLS) regressions (Martens and Næs, 1989) were used to fit the calibration models with leave-one-out cross-validation and a maximum of 15 factors. The optimal number of components was determined using the expression:

\[ \text{NF} = \text{Min}\{\text{RMSECV}(A) + 0.01 \times A \times \text{RMSECV}(0)\} \]  

(5.6)
where NF represented the optimal number of factors, $A$ the current dimensionality of the model, $\text{RMSECV}(A)$ the root mean square error of cross-validation at dimensionality $A$, and $\text{RMSECV}(0)$ was the initial error (Esbensen, 2002).

Calibration performance was assessed using the coefficient of determination for calibration ($R^2_c$) and the root mean square error of cross-validation (RMSECV). Prediction performance was evaluated using the coefficient of determination for prediction ($R^2_p$) and the root-mean square error of prediction (RMSEP) according to Mevik and Cederkvist (2004).

### 5.3.6 Comparison of model predictive ability

Error estimates from the calibration models were compared using the approach proposed by Reeves et al. (2006). The model used was:

$$\left(\hat{y}_{ij} - y_{ij}\right)^2 = \mu + \alpha_i + \beta_j + e_{ij}$$

(5.7)

where $\hat{y}_{ij}$ is the predicted specific gravity of the $i$th sample using the $j$th calibration model from the prediction data set, $y_{ij}$ is the measured specific gravity of the $i$th sample in the prediction set, $\mu$ is the overall mean, $\alpha_i$ is the random effect associated to the $i$th sample, $\beta_j$ is the fixed effect associated to the $j$th calibration model (pre-processing technique), and $e_{ij}$ is the random error component $\sim N(0, \sigma^2)$. Least squares means (after Box-Cox transformation) were compared using the Bonferroni correction method.

### 5.3.7 Software

**R version 2.8.0** (R Development Core Team, 2008) was used for the analysis. The regression models were fitted using the **pls** package (Mevik and Wehrens, 2007). Pre-processing algorithms and figures of merit were obtained using routines programmed in R.
5.4 Results and discussion

5.4.1 NIR spectra

The effects of the pre-processing methods on the spectra collected from the green logs are shown on Figure 5.1. Absorbance values are displayed in AU ($\times$E-5) to give an idea of the signal-to-noise ratios at the individual wavelengths as suggested by Faber (1999b). Similar results were observed for the wood strips data set (not shown).

Figure 5.1 shows high variation in absorbances among samples. This variation is likely to be caused by differences in moisture content and surface roughness. It is emphasized that spectra were collected on solid wood samples with little or no preparation. A decrease in the intensity of the signal was observed in green wood strips (not shown) compared to the green logs. Apart from differences in sample quality this may reflect differences related to the spectrometers used to evaluate each data set.

A closer inspection of the plots reveals that 2DER resulted in noisier spectra compared to RAW, MSC and OSC. Because derivatives de-emphasize lower frequencies and emphasize higher frequencies, they tend to accentuate noise. This effect can be attenuated by smoothing the spectra before calculating the derivatives. However, plots of spectra as presented in Figure 5.1 must be carefully analyzed. Brown et al. (2000) note that due to the band pass properties of the Savitzky-Golay filter, what really occurs is a reduction in both, the noise and the chemical signal, which in most cases leads to a reduction in the signal-to-noise ratio.

Note for example that the signal in Figure 5.1 is significantly diminished by 2DER (-150E-5 to +150E-5 AU) compared to OSC (-6000E-5 to +4000E-5 AU), and the RAW and MSC treatments (+60000E-5 to +160000E-5 AU). As a consequence of this reduction, the spectra tend to look much noisier using 2DER. These changes in signal intensity will have an important effect on the results of the PLS model, especially in the estimated vector of regression parameters ($\mathbf{b}$).
Chapter 5.4.2 NIR specific gravity calibration models

The fit statistics for the fitted PLS models are given in Table 5.2.

The coefficients of determination for the calibration models ($R^2_c$) ranged from moderate to good and were similar in both data sets (0.82-0.89). However, the errors of cross-validation (RMSECV) were higher for green wood strips (0.0239-0.0284) than for green logs (0.0161-0.0207). This can be explained in part by the differences in the number of latent variables required to obtain the optimum predictive model, with fewer required for wood strips (5-7) compared to green logs (9-12).

The coefficients of determination for the prediction sets ($R^2_p$) were moderate in both cases, ranging from a minimum of 0.66 to a maximum of 0.84. Lower $R^2_p$ and higher prediction errors (RMSEP) were observed in data set 1 compared to data set 2 (Figure 5.2. The fact that cross-validation errors and prediction errors in data set 2 were fairly similar in magnitude (ranging between 0.02 and 0.05), and contrary to what was observed in data set 1, indicates that the algorithm used to select the samples for calibration and prediction sets (random sampling in the case of green wood strips and rank-based sampling in the case of green logs) and the number of samples used for prediction have a critical effect on the fit statistics as discussed by Mora and Schimleck (2008b).

Figure 5.2 shows minimal differences among curves, especially for green logs. The minimum prediction error was 0.04 for green wood strips and 0.02 for green logs. Even though the error curves look similar, it is still possible to extract some specific features associated with the pre-processing methods. For green wood strips, OSC resulted in lower initial RMSECV and RMSEP values compared RAW, 2DER and MSC, and hence fewer factors were required to achieve minimum error (Table 5.2). However, this effect was not observed in the RMSECV of the green logs data set indicating that the information initially removed by the OSC components was not interfering with the signal of interest.

Another aspect to note is that for 2DER, the addition of extra factors after the optimal number of components resulted in an increased RMSECV and RMSEP values. This partic-
ular behavior of the second derivative in addition to the results given in Table 5.2 indicates that derivatives tend to do a good job in modeling the spectra of the calibration set but that does not necessarily translate into improved predictions. A similar conclusion was reported by Faber (1999b), although the derivatives reported in his work were calculated by taking differences without smoothing.

According to Shenk et al. (2008), the application of correction techniques cannot be expected to improve calibration and prediction data in all cases. The lack of significant improvement observed in this work might indicate that the particle size, as a function of the surface roughness, is a potentially useful variable in predicting green wood specific gravity. The PLS regressions constructed using raw data were able to model the data using additional components without resulting in unstable predictions. However, having a larger model (in terms of the number of latent variables) could result in a loss of predictive ability since model robustness degrades as dimensionality increases (Seasholtz and Kowalski, 1993; Wiklund et al., 2007).

5.4.3 PLS regression vectors

The vectors of estimated regression coefficients for the green logs data set are shown in Figure 5.3. The values of the vertical axis are displayed in inverse absorbance units ($\times 10^2$) to facilitate comparison among pre-treatments.

RAW, MSC, and OSC resulted in smoother regression vectors compared to 2DER which was noisier (especially in the region above 1900 nm) with no discernible patterns. As mentioned before, 2DER decreased the NIR signal compared to RAW, MSC, and OSC and hence larger coefficients were needed to make the conversion to specific gravity. This situation is clearly depicted in Figure 5.3 where the regression coefficients for RAW, MSC, and OSC vary between -0.015 and 0.020 (1/AU) and the coefficients for 2DER vary from approximately -1.5 to 1.5 (1/AU). A similar situation was observed for the green radial strips data set (not shown).
Figure 5.3 may lead to the conclusion that the increased noise in the regression vector using 2DER coupled with the decrease observed in the NIR signal (Figure 5.1) are the reasons for the high RMSEP obtained based on spectra pre-processed using this method. However, Faber (1999b) points out that it does not matter whether the resulting regression vector is noisy or smooth. The increased noise in the coefficients and the increased prediction error are due to the same cause: spectral error propagation, which is ultimately determined by the size of the vector and not by its shape.

5.4.4 Multivariate figures of merit

Table 5.3 gives the figures of merit based on the optimal number of PLS factors.

Larger scalar NAS values mean larger sensitivities. Multivariate sensitivity is the slope of the univariate regression model between scalar NAS and the property of interest; hence the values given in Table 5.3 represent the change in specific gravity by unit change in the scalar NAS for each method. As shown in the same table, OSC outperforms the other pre-treatments for both the green wood strips and the green logs data sets. The better performance in terms of sensitivity is translated to a lower contribution of the spectral measurement uncertainty to the prediction error variance for this method as discussed by Faber (1999b). RAW and MSC showed similar sensitivity values while 2DER resulted in an important decrease probably due to the introduction of new peaks in the spectra but more importantly due to the increase in the size of the regression vector.

Sensitivity was used to obtain the analytical sensitivity of each method (Olivieri et al., 2006). The inverse of the analytical sensitivity gives an indication of the minimum concentration difference which, in the absence of prediction errors, is detectable by the analytical method. Using the data sets combined, OSC resulted in the lowest inverse analytical sensitivity (average value = 2.02E-4) followed by RAW (2.25E-4), MSC (2.50E-4) and 2DER (2.38E-2). In other words, OSC, RAW, and MSC (in that order) are able to capture subtle
differences in specific gravity of the samples while 2DER can only detect much larger differences.

As defined in this work, selectivity represents the part of the signal that is lost because of the overlap of the signal of interest and other components present in the samples. Selectivity was increased by 2DER compared to the rest of the treatments. According to Faber (1999b), this is expected because derivative spectra usually exhibit more fine structure, but sometimes at the expense of increased noise. Table 5.3 shows that all the pre-processing methods (2DER, MSC, and OSC) were successful in increasing selectivity compared to raw NIR spectra.

The selectivity values obtained in this work are low compared to those reported in the literature (recall that this figure varies between 0 and 1). These differences are probably related to the characteristics of material analyzed. In general, figures of merit have been reported for the analysis of pharmaceutical and gasoline samples where some degree of sample preparation is commonly done before analysis, which was not the case of this study. However, low selectivity values are not a limitation for developing good calibration models. One of the characteristics of first-order multivariate calibration models is that they do not require total selectivity (Booksh and Kowalski, 1994).

The ability to extract the net analyte signal from the NIR data allows the representation of the fitted models as (pseudo)univariate regression models. Figure 5.4 shows the linear regressions for the green logs with calibration and prediction samples displayed together. The horizontal axis is labeled in NAS units (×E-5) for comparison with the magnitude of the instrumental error.

From Figure 5.4 it is clear that multivariate signal-to-noise ratios were lower for 2DER. According to Blanco et al. (1997), 2DER has the disadvantage that it always diminishes the SNR and is highly sensitive to the presence of noise in the original spectrum. The best SNR was achieved using OSC for both data sets, with the green logs giving the highest values.

According to Faber (1999b), in contrast to sensitivity, SNR is sample-dependent. Sensitivity is therefore preferred for formulating a general statement about the effect of data
pre-treatments. Figure 5.4 confirms that the relationship between scalar NAS ($r^*$) and specific gravity was fairly similar across pre-processing methods as discussed previously in the analysis of the error curves versus PLS components.

5.4.5 Comparison of model predictive ability

To complement the results obtained using multivariate figures of merit, ANOVA tests were performed on the square residuals obtained from the prediction samples for each data set. The results from the analysis on the green wood strips identified two groups of treatments that were statistically different at a 0.05 level of significance. The first group was composed of RAW, 2DER, and OSC, and the second group was composed by RAW, MSC, and OSC. The only significant difference was found between MSC and 2DER. The lowest square prediction errors were obtained using MSC followed, in increasing order, by RAW, OSC, and 2DER. For the green logs data set no significant differences were observed among treatments evaluated at the same level of significance.

5.5 Conclusions

NIR calibration models used for the prediction of wood properties have been traditionally characterized by means of the RMSECV and RMSEP obtained from the cross-validation and prediction phases, respectively. These parameters provide a measure of the expected average predictive ability of the calibrations. The results from this study demonstrates that calibration models can be further characterized by including multivariate figures of merit derived from the net analyte signal theory developed in the field of chemometrics.

The inclusion of figures of merit can help in getting a better understanding of the results obtained from the multivariate fitting process, especially for comparison of different pre-processing methods. For example, the relative ranking of the pre-treatment algorithms studied in terms of the figures of merit was consistent for the two data sets analyzed, while the ranking based on RMSEP and $R^2$ values was not.
The interpretation of the figures of merit is not straightforward though. The values are dependent on the number of latent variables used to fit the PLS models, which are the base for the calculation of the net analyte signal vector, and in the case of OSC the values are also dependent on the number of components removed. The alternative of assessing the quality of models for a fixed number of components was not considered in order to avoid problems of over- or under-fitting of the PLS models.

In general, specific gravity predictions were similar among pre-processing methods. Few statistical differences were observed for the green wood strips data set, where MSC resulted in lower prediction errors. Models fitted using raw data provided similar errors to those obtained using pre-treatments by increasing the number of latent variables. Some improvements were observed using OSC, especially for the green wood strips data set, which resulted in a more parsimonious model. OSC showed an increased multivariate sensitivity, inverse analytical sensitivity, and signal-to-noise ratio in both data sets. Despite of this, the RMSEP was not significantly reduced compared to, for example, MSC or RAW.

The results of applying multiplicative scatter correction to the data, a common method used in spectroscopy, were similar to those found using raw spectra. However, raw spectra resulted in a lower selectivity compared to the rest of the treatments. From the three pre-processing alternatives evaluated, 2DER was the least satisfactory in terms of RMSEP and most of the figures of merit.

5.6 References


Table 5.1. Range and standard deviation of specific gravity for calibration and prediction sets.

<table>
<thead>
<tr>
<th>Specific gravity</th>
<th>Calibration set</th>
<th>Prediction set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Data set 1</td>
<td>0.316</td>
<td>0.627</td>
</tr>
<tr>
<td>Data set 2</td>
<td>0.303</td>
<td>0.590</td>
</tr>
</tbody>
</table>
### Table 5.2. Fit statistics for the PLS models fitted using the different preprocessing algorithms

<table>
<thead>
<tr>
<th>Method</th>
<th>Data set 1</th>
<th></th>
<th></th>
<th></th>
<th>Data set 2</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NF</td>
<td>$R^2_c$</td>
<td>RMSECV</td>
<td>$R^2_p$</td>
<td>RMSEP</td>
<td>NF</td>
<td>$R^2_c$</td>
<td>RMSECV</td>
</tr>
<tr>
<td>RAW</td>
<td>7</td>
<td>0.85</td>
<td>0.0284</td>
<td>0.76</td>
<td>0.0447</td>
<td>12</td>
<td>0.84</td>
<td>0.0196</td>
</tr>
<tr>
<td>2DER</td>
<td>6</td>
<td>0.89</td>
<td>0.0239</td>
<td>0.68</td>
<td>0.0501</td>
<td>9</td>
<td>0.89</td>
<td>0.0161</td>
</tr>
<tr>
<td>MSC</td>
<td>7</td>
<td>0.86</td>
<td>0.0271</td>
<td>0.78</td>
<td>0.0416</td>
<td>10</td>
<td>0.82</td>
<td>0.0207</td>
</tr>
<tr>
<td>OSC</td>
<td>5</td>
<td>0.86</td>
<td>0.0278</td>
<td>0.77</td>
<td>0.0456</td>
<td>11</td>
<td>0.83</td>
<td>0.0205</td>
</tr>
</tbody>
</table>

NF, number of factors; RMSECV, root-mean square error of cross-validation; RMSEP, root-mean square error of prediction; $R^2_c$, coefficient of determination for calibration model; $R^2_p$, coefficient of determination for prediction
Table 5.3. Multivariate figures of merit for the PLS models fitted using the different pre-processing algorithms

<table>
<thead>
<tr>
<th>Method</th>
<th>Data set 1</th>
<th></th>
<th></th>
<th>Data set 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r^*$</td>
<td>$s$</td>
<td>SEL</td>
<td>SNR</td>
<td>$r^*$</td>
<td>$s$</td>
</tr>
<tr>
<td>RAW</td>
<td>0.0225</td>
<td>0.0476</td>
<td>0.0137</td>
<td>1728.1</td>
<td>0.0411</td>
<td>0.0962</td>
</tr>
<tr>
<td>2DER</td>
<td>0.0002</td>
<td>0.0004</td>
<td>0.0850</td>
<td>14.3</td>
<td>0.0005</td>
<td>0.0012</td>
</tr>
<tr>
<td>MSC</td>
<td>0.0191</td>
<td>0.0406</td>
<td>0.0755</td>
<td>1471.9</td>
<td>0.0407</td>
<td>0.0950</td>
</tr>
<tr>
<td>OSC</td>
<td>0.0255</td>
<td>0.0541</td>
<td>0.0729</td>
<td>1962.8</td>
<td>0.0444</td>
<td>0.1038</td>
</tr>
</tbody>
</table>

$r^*$, net analyte signal; $s$, sensitivity; SEL, selectivity; SNR, signal-to-noise ratio
Figure 5.1. Plots of NIR spectra for the green logs data set. RAW raw spectra, 2DER second derivative, MSC multiplicative scatter correction, OSC orthogonal signal correction.
Figure 5.2. Root-mean square error of cross-validation (RMSECV) and root-mean square error of prediction (RMSEP) versus number of latent variables for each data set and pre-processing method. RAW raw spectra, 2DER second derivative, MSC multiplicative scatter correction, OSC orthogonal signal correction.
Figure 5.3. Regression vector estimates for the calibration models fitted to the green logs data set. RAW raw spectra, 2DER second derivative, MSC multiplicative scatter correction, OSC orthogonal signal correction
Figure 5.4. Univariate inverse calibration model for the green logs data set. Calibration and prediction samples are represented by empty and filled circles, respectively. The model center is represented by the cross. RAW raw spectra, 2DER second derivative, MSC multiplicative scatter correction, OSC orthogonal signal correction, $s$ sensitivity, $b$ inverse sensitivity.
Chapter 6

Significance tests for principal components and partial least squares models: application to the study of microfibril angle in *Pinus taeda* using near infrared (NIR) spectroscopy

6.1 Abstract

Principal component (PCA) and partial least squares (PLS) models were used to develop near infrared (NIR) calibration models for the prediction of microfibril angle (MFA) in *Pinus taeda* wood samples. Final model dimensionality, i.e. the number of components or factors retained, was determined using cross-validation (CV) and permutation tests developed in the field of chemometrics. The permutation test aims to destroy the internal structure of the NIR matrix for PCA. For PLS, the randomization destroys the relationship between the NIR matrix and the MFA vector. The results showed that for PCA, CV performed well, resulting in model dimensionality similar to those obtained by permutation. For PLS, the permutation tests showed that CV led to severe problems of underfitting, especially for the pre-treated data. By increasing the number of factors as suggested by permutation, fit statistics for calibration were improved and errors were reduced. More importantly, the predictive ability of the models, when tested on a separate validation data set, were also improved. Compared to the raw data, the use of data pre-treatments helped to reduce some nonlinearities observed in the data.

6.2 Introduction

According to Booksh and Kowalski (1994), every analytical instrument or method can be classified based on the type of data it provides. For near infrared (NIR) spectroscopy, the spectrophotometers or first-order instruments are capable of generating multiple measurements for a single sample. The data obtained from these instruments, known as first-order tensor, are characterized by a high degree of multicollinearity. This makes standard regression techniques unstable and unreliable for the development of calibration models. One alternative to eliminate the problem of multicollinearity is to use data compression methods such as principal components analysis (PCA) and partial least squares (PLS) regression.
PCA is concerned with explaining the variance-covariance structure of a set of variables by obtaining a few orthogonal linear combinations of these variables which are referred to as principal components (PC) (Jolliffe, 2002). Geometrically, the PC’s represent a new coordinate system obtained by rotating the original coordinates in a way that the new axes represent the directions of maximum variability, providing a simpler and more parsimonious description of the covariance structure of the data (Jackson, 1991).

Similarly, the objective of the PLS method is to find a few linear combinations of the original variables, known as latent variables or factors, for regressing the response vector on the spectral data. By working with factors instead of the original data, irrelevant and unstable information is discarded and only the most relevant NIR variation is used for calibration (Næs et al., 2002).

Although these two compression methods have been successfully used for the analysis of NIR data for years, according to Gómez-Carracedo et al. (2007) several important problems remain without definitive answers: i) how to select the calibration (and validation) sample sets and how to set the proper number of samples, ii) how to choose an adequate NIR data pre-processing algorithm, and iii) how to assess the optimal model dimensionality. For any NIR analysis, the determination of a models rank (or dimensionality) is one of the most difficult to solve because there is no distributional theory that can provide, at least in a closed form, significance tests for PCA or PLS models.

The selection of the optimal number of components (PCA) or factors (PLS) to use is an important problem for all data compression methods. If too many components or factors are used, excessive redundancy from the spectral data is included in the analysis and the solution becomes overfitted. For PLS models, an overfitted model implies a calibration equation that is extremely data dependent and with moderate to poor predictive ability. On the contrary, using too few components or factors, known as underfitting, will result in a model that is not large enough to capture all the important variability observed in the data (Næs et al., 2002).
The issue of determining whether or not a given component or factor summarizes meaningful variation is a non-trivial problem. Denham (2000) points out, that the statistical nature of PLS regression makes it difficult to perform traditional inferential tasks such as choosing model dimensionality, assessing uncertainty in estimates, and producing prediction intervals for future responses and, according to Peres-Neto et al. (2005), when the incorrect number of components is used for a PCA model, either relevant information is lost or noise is included for subsequent analysis. An example of the importance of determining the PCA model dimensionality is given in the work by Mora and Schimleck (2008b), where the matrix of scores \( \mathbf{T} \) obtained from PCA was used as input for the selection of calibration samples using the CADEX (Kennard and Stone, 1969) and DUPLEX (Snee, 1977) algorithms. The dimensions of \( \mathbf{T} \) were defined by the total number of samples and the optimal number of PC determined by cross-validation (CV).

Several stopping rules to estimate the optimal number of PC’s and factors have been proposed in the literature. For PCA, examples of the traditional rules are the scree test (Jackson, 1991; Johnson and Wichern, 2002; Jolliffe, 2002), the test of sphericity (Jackson, 1993; Peres-Neto et al., 2005), the proportion of the total variance (Jackson, 1993; Jolliffe, 2002), the broken-stick method (Jackson, 1991), and cross-validation for studying the residual variance (Wold, 1978; Esbensen, 2002). For PLS, cross-validation (Wold, 1978; Esbensen, 2002), leverage correction (Marbach and Heise, 1990), size of the eigenvalues (see Wiklund et al., 2007), and Monte Carlo methods (Wakeling and Morris, 1993; Gourvénec et al., 2003; Xu et al., 2004b) are the most common methods.

However, the traditional methods don’t take into account one problem that is commonly observed when modeling large data sets, especially in cases where the number of samples is smaller than the number of variables: the possibility of chance effects (Wakeling and Morris, 1993). Each model has a possibility of being developed by pure chance and every regression data set contains structures which cause background correlation (Lindgren et al., 1996), i.e.
it is possible to find a certain level of correlation between the dependent variable and the set of predictors even if there is no real association between them.

A way to check how far a model is from being a result of chance correlations is by using permutation tests (Lindgren et al., 1996). A permutation or randomization test is a statistical test in which a reference distribution is obtained by calculating all possible values of a test statistic under rearrangements of the labels on the observed data (Good, 2005a). The requirement is that the observations must be exchangeable under the null hypothesis (Dijksterhuis and Heiser, 1995). Permutation tests are not new but were not popular until recently because of their computing-intensive nature (van der Voet, 1994).

Ledauphin et al. (2004) and Wiklund et al. (2007) have developed significance tests for PCA and PLS models, respectively, both based on permutation tests. These two data-driven tests represent an interesting alternative for assessing the statistical significance of PCA and PLS models applied to the study of wood properties using NIR because they don’t require unrealistic assumptions about the statistical distribution of the data under analysis. For this reason, the primary objective of this work was to implement the approaches proposed by these authors for testing the statistical significance of PCA and PLS models for the estimation of microfibril angle (MFA) in *Pinus taeda* radial strip samples using NIR spectroscopy. The specific objectives were: (i) to compare calibration models based on raw spectra and spectra treated using two common pre-treatment techniques: derivatives and standard normal variate, and (ii) to compare the dimensionality of calibration models based on the traditional CV approach with those based on permutation tests.

### 6.3 Material and methods

#### 6.3.1 Data set

Wood samples were obtained from *P. taeda* progeny tests established by members of the North Carolina State University Cooperative Tree Improvement Program between 1987 and 1992 in the south-eastern United States. The field tests were located in three regions (central
Georgia (GA), south-west South Carolina (SC), and North Carolina (NC) Atlantic coast). A 6-tree disconnected diallel mating design was utilized to produce progeny of 12 parents and a randomized complete block design was used on each test location. Thirty full-sib families plus one check lot were available for sampling on each test site.

A total of 20 trees representing different full-sib families were destructively sampled in each region (Table 6.1). Trees that were suppressed, atypical in form, or infected by fusiform rust were excluded from sampling. From each selected tree, a 4 cm thick wood disk was extracted at breast height (1.4 m from the base) and a 12.5 $\times$ 12.5 mm bark-to-bark sample was obtained from the center of the disk.

Each bark-to-bark sample was divided at the pith and one half was prepared for data collection. The selected half was dried at 50$^\circ$C for 24 hours, glued into wood holders and sectioned along its longitudinal axis using a twin-blade saw to produce a radial longitudinal strip approximately 2 mm thick from the center of each bark-to-bark sample.

### 6.3.2 Near infrared spectroscopy

Near infrared (NIR) diffuse reflectance spectra were obtained in consecutive 10-mm sections from the radial longitudinal face of each strip using a FOSS NIRSystems (Laurel MD, USA) Model 5000 spectrophotometer. The radial strips were held in a custom-made holder and a 5 $\times$ 10 mm mask was used to ensure that a constant area was tested as described by Schimleck et al. (2001b). NIR spectra were collected over the range of 1100-2498 nm with a resolution of 2 nm and the instrument reference used was a ceramic standard. The estimated instrumental noise obtained by replicate measurements of the instrument’s reference was 1.2E-5 AU (absorbance units). The measurements were made in a conditioned atmosphere at 40% relative humidity (RH) and 20$^\circ$C.
SilviScan data

Smaller radial strips, 2 mm tangentially \times 7 mm longitudinally, were cut from the strips scanned for NIR analysis and used for microfibril angle (MFA) determination. The length of the SilviScan strips varied depending on the pith-to-bark length of the samples used for NIR. MFA of the S\textsubscript{2} layer of the cell wall was measured by X-ray diffractometry at 5-mm resolution using the SilviScan-3 device (Paprican, Vancouver BC) and the variance method developed by Evans (1999, 2006). The SilviScan strips were not resin extracted and, as with NIR analysis, the measurements were made in a controlled environment of 40\% RH and a temperature of 20°C.

Spectral data pre-treatments, PCA and PLS models

Raw NIR data (RAW) and pre-treated spectra using second derivative (2D) and standard normal variate (SNV) were used for the analyses. Each pre-treatment was applied individually to the raw data and they were not combined. Second derivative spectra were obtained using the Savitzky-Golay filter (Savitzky and Golay, 1964) with a nine-point window and second order degree polynomial as described by Wentzell and Brown (2000). The SNV pre-treatment was applied using the methodology described in Barnes et al. (1989).

The mean-centered raw and pre-processed NIR matrices were compressed by principal component analysis (PCA) using the eigenvalue decomposition of the variance-covariance matrix (Jolliffe, 2002). A maximum of 15 PC's were used for the subsequent analyses. Non-linear iterative partial least squares (NIPALS-PLS) was used to fit the NIR calibration models using the mean-centered raw and pre-treated data with a maximum of 15 factors. The reader is referred to Martens and Næs (1989) for additional details about the PLS algorithm. The PCA and PLS models generated at this step of the analysis will be referred to in the text as the “initial” models.
**Model dimensionality: cross-validation**

Cross-validation (CV) is the statistical practice of partitioning the data into subsets such that the analysis is initially performed on a single subset, while the other subsets are retained for internal validation. It works by temporarily removing samples or groups of them, known as segments, from the calibration set, developing a model with those that remain and predicting the excluded samples. The data is then restored to its initial condition and the process is repeated a number of times to obtain the aggregate squared prediction error (Wold, 1978). The optimal number of PC’s or factors is the one that minimizes the overall prediction error. This methodology is sometimes referred to as pseudo-validation because there are no independently drawn validation data sets (Esbensen, 2002).

Sixty radial strips (where a single radial strip represented an individual CV segment in the analysis) were used for determining the rank of the initial PCA models. The initial PLS regressions were fitted using 45 radial strips and the remaining 15 were used to test the predictive ability of the calibration models. The optimal number of PC’s and factors was determined as (Esbensen, 2002):

\[
NF = \text{Min}\{\text{MSECV}(k) + 0.01 \times k \times \text{MSECV}(0)\} \tag{6.1}
\]

where NF represented the optimal number of PC’s or factors, \(k\) the current dimensionality of the model, \(\text{MSECV}(k)\) the mean square error of CV at dimensionality \(k\), and \(\text{MSECV}(0)\) was the initial error. In addition, the NIR calibration performance was assessed using the coefficient of determination for calibration \(R^2_c\), root mean square error of cross-validation (RMSECV), the coefficient of determination for prediction \(R^2_p\) and the root mean square error of prediction (RMSEP).
Model dimensionality: Permutation tests

Permutation test for PCA

Let $X$ be a NIR matrix of dimensions $n \times p$, where $n$ represents the total number of 10-mm sections and $p$ the number of variables or wavelengths. For $k = 1, 2, \ldots, p$, of the $k$th PC is given by $z_k = X\alpha_k$, where $\alpha_k$ is an eigenvector of the variance-covariance matrix corresponding to its $k$th largest eigenvalue $\lambda_k$. Because $\alpha_k$ is normally forced to have unit length ($\alpha_k^T\alpha_k = 1$), the variance of $k$th PC is given by $\text{var}(z_k) = \lambda_k$ (Zuur and Smith, 2007). If $z^*_k$ denotes the standardized PC associated with $z_k$ (i.e. $z^*_k = z_k / \lambda$), then:

$$\lambda = \frac{1}{n^2} z^*_k^T X X^T z^*_k \quad (6.2)$$

To test the statistical significance of the $k$th PC of $X$, the procedure proposed by Ledauphin et al. (2004) randomly permutes the rows of the $X$ matrix and for each permutation calculates $\tilde{\lambda}$ using Eq. 6.2. $\tilde{\lambda}$ is obtained by replacing the original $X$ matrix with the permuted matrix $\tilde{X}$ while keeping $z^*_k$ fixed. The procedure is repeated a large number of times (1000 in this work) and the $k$th PC is considered significant if the proportion of simulated values $\tilde{\lambda}$ that exceed the observed value $\lambda$ obtained from the initial PCA model is less than a fixed value previously chosen (5% in this work, i.e. $\text{Prob} < 0.05$).

Permutation test for PLS

The test proposed by Wiklund et al. (2007) assesses the statistical significance of each individual factor that enters sequentially in the regression model by randomly permuting the elements of the $y$ vector while keeping the corresponding values of the $X$ matrix fixed. The permutation destroys any relationship between $X$ and $y$ generating random PLS regression models, i.e. models that reflect the absence of a real association between the matrix of NIR data and the vector of the property of interest.
As with the PCA test, this procedure is repeated a large number of times (set to 1000) and for each of the random PLS models, a test statistic ($t^*_k$) is calculated as:

$$t^*_k = T^*_k y^*$$  (6.3)

Equation 6.3 represents the inner product of the scores vector after permutation for the $k$th dimension ($T^*_k$) and the permuted $y$ vector ($y^*$). The test statistic obtained from Eq. 6.3 generates the null-hypothesis distribution ($H_0$) for the $k$th factor under a scenario of chance correlations and the factor is considered statistically significant if the proportion of simulated values that exceed the observed test statistic value ($t_k$) obtained from the initial PLS model is less than a value previously chosen (set to 5%). Note that the calculation of $t^*_k$ is conducted conditional on previous components having been fully extracted from the data.

CALIBRATION SAMPLES

Once defined the number of PC’s to use for each data set (raw and pre-treated), the $Q$- and Hotelling’s $T^2$-statistic of each 10-mm section was calculated to identify potential outliers in the calibration phase as described by Mora and Schimleck (2008b). Radial strips whose 10-mm sections were identified as outliers were assigned to the prediction set. The remaining radial strips were ranked by average MFA and, based on the rank, 45 strips (the CV segments) covering the entire range of MFA were selected for calibration.

R version 2.8.1 (R Development Core Team, 2008) and the pls package (Mevik and Wehrens, 2007) were used for the analyses. All tests were programmed in R.

6.4 RESULTS AND DISCUSSION

PCA

Small differences were observed between PCA model dimensionality suggested by permutation and dimensionality recommended by the criterion defined in Eq. 6.1. For the RAW
data, both approaches (CV and permutation) recommended 2 PC’s in the final model that together explained 99.4% of the NIR variation. For the 2D data, 11 PC’s were suggested by CV while 12 PC’s were recommended by permutation, accounting for 90.0 and 90.7% of the NIR variation, respectively. For the SNV data, 6 PC’s were recommended by CV (97.2% of NIR variation) and 7 PC’s were found to be statistically significant (Prob < 0.05) by permutation (98.0% of NIR variation).

The distributions of the simulated $\tilde{\lambda}$ values, for PC’s 11 to 14, obtained by permutations of the 2D data set are shown in Fig. 6.1. PC 11 is the dimensionality suggested by CV. Based on these results, it was concluded that PC 12 was needed in the model to adequately represent the NIR variance-covariance matrix. Similar plots were obtained for the RAW and SNV data sets (not shown). When the observed $\lambda$ values, obtained from the initial PCA models, exceeded all the simulated values, the probabilities were calculated using the inverse Gaussian distribution (suited for modeling positively skewed data) as suggested by Wiklund et al. (2007) adjusted to avoid numerical overflow problems as described by Dennis et al. (1991).

Based on the recommended model dimensionality, influence plots showing the relationship between $Q$- and $T^2$-statistics for the 10-mm sections were obtained for each PCA (Fig. 6.2). The $Q$ statistic indicates how well each sample conforms to the PCA model. It can be interpreted as a measure of the difference between the sample and its projection on the $k$th PC’s retained in the model. The $T^2$ statistic is a measure of the distance from the multivariate mean to the projection of the sample on the PC’s (Mora and Schimleck, 2008b).

Figure 6.2 shows that some of the 10-mm sections, e.g. 51-10 and 22-06, were consistently located beyond the confidence limits of the $Q$- and $T^2$-statistics, independent of the pre-treatment or number of PC’s used, suggesting that they can be regarded as outliers. A similar plot was obtained for the RAW data set (not shown). Based on these plots, 6 radial strips from the RAW data set and 4 radial strips from the pre-treated data sets were removed from the list of potential radial strips for calibration and incorporated to the prediction data
sets. Due to the reduced size of the data set, no strips were deleted for the analysis, even if they appear to be clear outliers such as tree 22. Table 6.2 gives a summary of the calibration and prediction data sets for the RAW and pre-treated data sets.

Since CV and the permutation test yielded similar results, the range of MFA values for the calibration and prediction radial strips was similar among data sets (RAW and pre-treated). This result must not be generalized because it represents a particular data-dependent situation. For this work, the main advantage of the permutation test compared to the CV approach was given by the ability to assess the statistical significance of the individual PC’s, thus decreasing the uncertainty in defining the correct PCA model dimensionality.

Compared to other randomization tests available for PCA (see Peres-Neto et al., 2005), the methodology proposed by Ledauphin et al. (2004) has an advantage: the permutations are not strictly needed to assess the significance of the PC’s. The authors derived the theoretical properties of the test that can be used to avoid undergoing the permutations, pointing out that the mean of all the simulated values $\tilde{\lambda}$ can be obtained by using the expression $m(\tilde{\lambda}) = \text{tr}(X^TX)/n(n-1)$, where $\text{tr}()$ indicates matrix trace. Then, a test statistic of the form $t(\tilde{\lambda}) = (\lambda - m(\tilde{\lambda}))/\sqrt{\text{var}(\tilde{\lambda})}$ can be calculated, where $\text{var}(\tilde{\lambda})$ is the variance of $\tilde{\lambda}$, and used to assess the statistical significance of the individual PC’s. If the test statistic exceeds a threshold value $\eta$, usually between 2 and 3, then the PC is considered significant. The reader is referred to the original work by Ledauphin et al. (2004) for additional details.

In this study, the theoretical approach was also tested and in all cases (not shown) the results agreed with those obtained by actually running the permutations. For example, for the RAW data set, $\lambda_{PC1} = 0.5522$, $\lambda_{PC2} = 0.0091$, and $\lambda_{PC3} = 0.0015$. The theoretical mean of the permuted $\lambda$ was $m(\tilde{\lambda}) = 1102E-06$ with $\text{var}(\tilde{\lambda})_{PC1} = 2.314E-06$, $\text{var}(\tilde{\lambda})_{PC2} = 2.310E-06$, and $\text{var}(\tilde{\lambda})_{PC3} = 2.307E-06$. Based on these values, the calculated test statics were $t(\tilde{\lambda})_{PC1} = 362.27$, $t(\tilde{\lambda})_{PC2} = 5.25$, and $t(\tilde{\lambda})_{PC3} = 0.24$, indicating that only PC1 and PC2 (both with $t(\tilde{\lambda}) > \eta = 3$) were statistically significant.
After defining the 45 radial strips (CV segments) for calibration, PLS regression models were fitted to the NIR data to estimate MFA in the 10-mm sections. Important differences between the number of factors to retain suggested by permutation and by CV were found. For the RAW data, 10 factors (accounting for 89.1% of the MFA variation) were recommended by CV while 12 factors (90.0% of the MFA variation) were suggested by permutation. For 2D data, 6 factors were suggested by CV and 11 factors by the permutation test, explaining 86.8 and 89.5% of the MFA variance, respectively. For the SNV data, 9 factors were recommended by CV (90.2% of the MFA variation) while 14 factors were suggested by permutation (91.6% of the MFA variation).

The distribution of the simulated $t^*$ statistic values, for factors 6 to 9, obtained from the permutations of the 2D data set is shown in Fig. 6.3. As with PCA, the first plot (factor 6) corresponds to the dimensionality suggested by Eq. 6.1. It is clear from the figure that CV failed to consider additional latent variables that were important in explaining the observed variation in MFA, for example factors 7, 8, and 9. Similar plots were generated for the other two data sets (not shown). As described in the PCA section, when the observed test statistic exceeded all the simulated values, the probabilities were calculated using the inverse Gaussian distribution approach.

The coefficients of determination for the calibration models ($R^2_c$) obtained by CV were very good, ranging from 0.90 to 0.93, with cross-validation errors (RMSECV) varying from 2.38 to 2.51 deg. It was noted however, that the MFA data had a skewed distribution toward high values (not shown), probably due to an over-representation of juvenile and transition wood related to the age of the sampled trees. The skewness was reflected in some nonlinearities observed in the models (Fig. 6.4), especially for the RAW data. PLS regression is able to handle nonlinearities by adding extra factors into the models (Næs et al., 2002) and that may indicate that the “punish” factor of 0.01 used in Eq. 6.1 (Esbensen, 2002) was too restrictive causing problems of underfitting when CV was used to define model dimension-
ality. Wiklund et al. (2007) point out that this factor may require to be adjusted depending on the data under analysis.

Figure 6.4 shows that by increasing the PLS model dimensionality, as suggested by the permutation tests, better fit statistics were obtained and the non-linearity problems were reduced. $R^2_c$ were increased, ranging from 0.93 to 0.95 and RMSECV were reduced, varying from 2.20 to 2.45 deg. To assess the impact of the increased dimensionality on the predictive ability of the regression models, MFA of the radial strips left out for validation were predicted. The results are shown in Fig. 6.5.

Using the cross-validated models, the coefficients of determination ($R^2_p$) for the prediction sets ranged from 0.55 to 0.81 and when the models fitted using the number of factors defined by permutation were utilized, $R^2_p$ increased from 0.75 to 0.84. Similarly, the prediction errors (RMSEP) were reduced from a range of 3.31 to 5.15 deg for the cross-validated models from 2.97 to 3.62 deg in the models with increased dimensionality, and consequently the regression line between predicted and observed MFA values was closer to the line of equivalence as can be seen in Fig. 6.5.

The permutation test for PLS assumes that the factors enter the model in a natural order, that is according to their relevance for describing the $y$ vector. According to Wiklund et al. (2007) this is not always the case with real data and it can, for example, be distorted by an improper pre-treatment of the NIR data. Therefore, one additional advantage of the permutation test, besides the significance test, is that erratic behavior on the significance of factors may be an indication that the pre-processing algorithm used is not the most appropriate for the spectral data.

6.5 Conclusions

PCA and PLS are the two most commonly used techniques for the study of wood properties by NIR. The results obtained from these two methodologies depend exclusively on their model dimensionality or rank.
In this study, the analysis of variance-covariance matrix of spectral data by PCA, using CV and permutation, resulted in models of similar dimensions. This suggests that the criterion used by CV to define the number of PC’s to retain in the final model works well when the NIR data does not show severe problems of anomalous spectra. However, the possibility to assign a probability value to each PC, makes the permutation test an intuitive and helpful tool for reducing the uncertainty in determining the correct PCA model rank.

The CV criterion, when applied to PLS models, may lead to severe underfitting problems as shown by the results of this work. The calibration models fitted to predict MFA of 10-mm sections were improved, compared to CV, when the additional factors recommended from permutation were added to the regression models. More importantly, the predictive ability of the models was improved, reaffirming the idea that, in this work, the CV criterion was too restrictive to define the appropriate model dimensionality.

In general, MFA calibration models were similar for the different data sets used especially after including the extra factors suggested by permutation and 2D and SNV corrected some of the nonlinearities initially observed in the data. The differences observed in the prediction statistics are due in part to the pre-treatment used and in part to the radial strips assigned to the validation set. Of the 15 radial strips used for prediction, RAW and pre-treated data had only 7 strips in common, and the statistics probably reflect these differences.

The inclusion of permutation tests into wood properties analysis by NIR spectroscopy coupled with other techniques such as multivariate figures of merit (Mora and Schimleck, 2008a) will give a better understanding of the results obtained from the fitting process. This leads to more robust calibration models, especially considering that the nature of the permutation tests makes them appropriate to detect not only underfitting but also overfitting problems.
6.6 References

Barnes, R. J., Dhanoa, M. S., Lister, S. J., 1989. Standard normal variate transformation and

782A–791A.

Denham, M. C., 2000. Choosing the number of factors in partial least squares regression:
estimating and minimizing the mean squared error of prediction. J. Chemometr. 14 (4),
351–361.

parameters for endangered species. Ecol. Monogr. 61 (2), 115–143.

Dijksterhuis, G. B., Heiser, W. J., 1995. The role of permutation tests in exploratory multi-
variate data analysis. Food Qual. Prefer. 6 (4), 263–270.

AS, Oslo.

Evans, R., 1999. A variance approach to the x-ray diffractometric estimation of microfibril

Evans, R., 2006. Wood stiffness by X-ray diffractometry. In: Stokke, D., Groom, L. (Eds.),
Characterization of the Cellulosic Cell Wall. Blackwell Publishing, Ames, Ch. 11, pp. 138–
146.

the optimum number of partial least squares components for the calibration of attenuated
585 (2), 253–265.


R Development Core Team, 2008. \texttt{R}: A language and environment for statistical computing. Available from \url{http://www.r-project.org}.


Table 6.1. Individual tree mean characteristics on each test site. Standard errors are given in parenthesis.

<table>
<thead>
<tr>
<th>Test</th>
<th>Location</th>
<th>Physiographic region</th>
<th>Age (years)</th>
<th>n</th>
<th>DBH (cm)</th>
<th>HT (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GA</td>
<td>Piedmont</td>
<td>15</td>
<td>10</td>
<td>18.8 (0.6)</td>
<td>13.8 (0.2)</td>
</tr>
<tr>
<td>1</td>
<td>GA</td>
<td>Piedmont</td>
<td>16</td>
<td>10</td>
<td>19.0 (0.5)</td>
<td>15.7 (0.1)</td>
</tr>
<tr>
<td>2</td>
<td>SC</td>
<td>Coastal plain</td>
<td>14</td>
<td>10</td>
<td>22.7 (0.7)</td>
<td>20.1 (0.4)</td>
</tr>
<tr>
<td>2</td>
<td>SC</td>
<td>Coastal plain</td>
<td>15</td>
<td>10</td>
<td>23.9 (0.5)</td>
<td>20.2 (0.5)</td>
</tr>
<tr>
<td>3</td>
<td>NC</td>
<td>Coastal plain</td>
<td>18</td>
<td>10</td>
<td>23.9 (1.2)</td>
<td>19.0 (0.4)</td>
</tr>
<tr>
<td>3</td>
<td>NC</td>
<td>Coastal plain</td>
<td>19</td>
<td>10</td>
<td>22.2 (1.4)</td>
<td>19.7 (0.6)</td>
</tr>
</tbody>
</table>

**Note:** DBH, diameter at breast height (1.4 m); HT, total height.
Table 6.2. Descriptive statistics of microfibril angle (MFA) measured in degrees (deg) for the calibration and prediction sets. $n_{rs}$, number of radial strips; $n_c$, total number of 10-mm sections used for calibration; $n_p$, total number of 10-mm sections used for prediction.

<table>
<thead>
<tr>
<th>Data</th>
<th>$n_{rs} = 45$</th>
<th></th>
<th></th>
<th></th>
<th>$n_{rs} = 15$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n_c$</td>
<td>Min.</td>
<td>Max.</td>
<td>Mean</td>
<td>Std</td>
<td>$n_p$</td>
<td>Min.</td>
<td>Max.</td>
</tr>
<tr>
<td>RAW</td>
<td>382</td>
<td>9.97</td>
<td>45.49</td>
<td>28.82</td>
<td>7.48</td>
<td>131</td>
<td>11.23</td>
<td>41.77</td>
</tr>
<tr>
<td>2D &amp; SNV</td>
<td>382</td>
<td>9.97</td>
<td>45.49</td>
<td>29.09</td>
<td>7.61</td>
<td>131</td>
<td>11.23</td>
<td>45.14</td>
</tr>
</tbody>
</table>

Note: RAW, raw data; 2D, second derivative; SNV, standard normal variate.
Figure 6.1. Distribution of the simulated values $\hat{\lambda}$ for PC’s 11 to 14, generated by 1000 permutations of the 2D data set. The observed $\lambda$ values from the initial PCA model are indicated with arrows. Prob indicate the probability that the simulated values exceed the observed $\lambda$ for each component.
**Figure 6.2.** Influence plots obtained from the PCA models for second derivative (2D) and standard normal variate (SNV) data sets. Extreme samples (10-mm sections) are identified with their labels (samples well beyond the confidence limits for $Q$- and $T^2$-statistic). CV, cross-validation; PT, permutation test.
Figure 6.3. Distribution of the simulated $t^*$ values for factors 6 to 9, generated by 1000 permutations of the 2D data set. The observed $t$ values from the initial PLS model are indicated with arrows. Prob indicate the probability that the simulated values exceed the observed $t$ for each factor. The curves represent the inverse Gaussian distribution fitted to calculate Prob for some of the factors.
Figure 6.4. MFA calibration models for raw (RAW), second derivative (2D), and standard normal variate (SNV) data. $R^2_c$, coefficient of determination for calibration; NF$_{CV}$, number of factors determined by cross-validation; NF$_{PT}$, number of factors determined by permutation; SEC, standard error of calibration; RMSECV, root mean square error of cross-validation.
Figure 6.5. MFA predictions for raw (RAW), second derivative (2D), and standard normal variate (SNV) data. $R^2_p$, coefficient of determination for calibration; $N_{FCV}$, number of factors determined by cross-validation; $N_{PT}$, number of factors determined by permutation; SEP, standard error of prediction; RMSEP, root mean square error of prediction.
Chapter 7

Kernel regression methods for the prediction of wood properties of *Pinus taeda* using near infrared (NIR) spectroscopy\(^1\)

7.1 Abstract

Near infrared (NIR) diffuse reflectance spectra collected in 10-mm sections were used for the estimation of air-dry density (AD), microfibril angle (MFA), stiffness (MOE), tracheid coarseness (COARS), and tracheid wall thickness (WTHICK) in wood radial strip samples obtained at breast height (1.4 m) from sixty *Pinus taeda* trees. Calibration models were developed using traditional partial least squares (PLS) and kernel regression. The kernel methods included radial basis functions-partial least squares (RBF-PLS) and least-squares support vector machines (LS-SVM). RBF-PLS and LS-SVM models outperformed PLS-CV calibrations in terms of fit statistics. MFA and MOE, two properties that exhibited nonlinearity, showed the most significant improvements compared to PLS. In terms of predictive ability RBF-PLS performed better than PLS for the prediction of MFA, MOE, and COARS. LS-SVM showed better prediction statistics in all cases, except for WTHICK that gave similar statistics compared to PLS, and was superior to RBF-PLS. By adding statistically significant factors to the PLS regressions it was possible to capture some of the nonlinear features of the data and improve the predictive ability of the PLS models.

7.2 Introduction

Several studies have demonstrated that near infrared (NIR) spectroscopy coupled with multivariate calibration methods provides a rapid alternative for estimating many wood properties (see Schimleck and Workman Jr., 2004; Tsuchikawa, 2007). Most of these studies have assumed a linear relationship between the spectra and the property under analysis, with partial least squares (PLS) being the most commonly used algorithm to fit the calibration models.

While PLS can be used to derive good solutions in most cases, in some situations a nonlinear model may be preferred, especially when the analytical form of relationship between the spectral data and the property of interest is not known (Walczak and Massart, 1996b)
or when there is little or no basis for linearity in Beer’s law (Cogdill et al., 2004). Among the approaches that have been suggested for such situations are the use of spectroscopic and/or statistical transformations of the data, the addition of extra terms to the calibration model, the option of splitting the data into linear subsets, or the use of nonlinear calibration methods (Næs et al., 2002).

The modeling of nonlinear systems can be done using semi- or non-parametric nonlinear methods and machine learning, a subfield of statistical learning (Hastie et al., 2001), which represents an interesting alternative for the development of nonlinear NIR calibration models. The major focus of machine learning is to automatically produce models that perform better in the future (prediction) based on what was experienced in the past (training).

In the context of NIR analysis, learning methods based on kernels have been proposed as alternatives to PLS. According to Cogdill and Dardenne (2004), the traditional PLS method can be considered as a variable-centric approach where the solution is obtained by evaluating the distribution of data for each variable or wavelength. The kernel approach instead is sample-centric, where the relationship between every sample is characterized by a kernel function that maps the data into a higher dimensional feature space where linear regression is performed. Simply explained, the idea is to characterize the relationship between $X$ (spectra) and $Y$ (property of interest) in a hyperspace where nonlinearity is no longer a problem.

One of the kernel-based methods that can be applied to NIR data is a hybrid algorithm that combines radial basis functions networks (RBFN) with PLS known as radial basis functions partial least squares (RBF-PLS) (Walczak and Massart, 1996b). RBFNs are artificial neural networks that use radial basis functions (RBF) as activation functions (a function that defines the output of a node given an input or set of inputs). The characteristic feature of RBFs is that their response decreases or increases symmetrically and monotonically with distance from the origin or any other center (Buhman, 2003).
RBFNs typically have three components: an input layer, a single hidden feedforward structure layer with a non-linear RBF activation function, and an output layer. The Gaussian RBF is the most widely adopted activation function since it can be used to model functions of arbitrary nonlinearity (Buhman, 2003; Cogdill and Dardenne, 2004).

The Gaussian RBF has the form \( \exp \left( - \| x_j - x_i \|^2 / \sigma^2 \right) \), where \( x_i \) represents the \( i \)th sample, \( x_j \) the center of the \( j \)th RBF unit, \( \| \cdot \| \) the Euclidean distance, and \( \sigma^2 \) the width of the RBF, a parameter that controls the degree of nonlinearity that can be modeled. After transforming \( X \) using the RBF, a so-called activation matrix \( A \) is obtained and PLS is utilized to model \( Y \) as a function of \( A \) instead of \( X \). Readers are referred to Walczak and Massart (1996a,b) or Wilson et al. (1997) for further details.

Few published studies, related to NIR modeling, have used the RBF-PLS approach. Walczak and Massart (1996a) applied RBF-PLS for classification of process faults and concluded that the performance and robustness of this method was comparable with other neural network approaches with the advantage of leading to a linear PLS methodology that can handle nonlinear processes. Wilson et al. (1997), using a slightly different methodology, combined RBFs with PLS for solving a variety of nonlinear problems, including fault detection, and in all cases the training and predictions errors were significantly reduced.

Recently, a different hybrid model combining RBFs and PLS was presented by Wang et al. (2008) and, according to the authors, more parsimonious network structures were developed reducing the risks of overfitting. However, its applicability to NIR calibration problems has not been explored yet.

Least-squares support vector machines (LS-SVM) (Suykens et al., 2002) is another class of kernel machines and corresponds to a reformulation of the principles of support vector machines (SVM) used for regression and classification problems (Steinwart and Christmann, 2008). In LS-SVM a linear estimation is done in a kernel-induced feature space minimizing an augmented “cost function” that contains terms related to the complexity of the model and a training error, both balanced by a regularization constant (\( \gamma \)).
As with RBF-PLS, the Gaussian RBF function is commonly used for transforming the data; therefore, in LS-SVM the user must optimize two parameters: $\sigma^2$ and $\gamma$. According to Cogdill and Dardenne (2004), $\gamma$ plays a similar role to number of factors in PLS and $\sigma^2$ is similar to adjusting the neighborhood size of a neural network model. LS-SVM substitutes the matrix $X$, in this case by a kernel matrix $K$ that contains the sample-sample relationships, and a regression equation is derived solving a set of linear equations instead of a quadratic programming problem such as in SVM. Further details can be found in Suykens et al. (2002) and Cogdill and Dardenne (2004).

LS-SVM has been employed in several studies related to spectroscopy recently. Cogdill and Dardenne (2004) compared the performances of LS-SVM with PLS, LOCAL (Berzaghi et al., 2000; Shenk et al., 2008) and artificial neural networks (ANN) for regression and classification using NIR spectra and different properties measured in apples, meat, corn, and animal feed. They found that LS-SVM outperformed the other methodologies in terms of predictive ability. For the study of wood properties in *Pinus taeda*, Cogdill et al. (2004) compared PLS with LS-SVM and found that, for a comparable number of factors, LS-SVM calibrations were superior although the predictive performance of the models was not significantly improved.

Thissen et al. (2004b) applied LS-SVM to NIR data collected from mixtures of ethanol, water, and 2-propanol measured at different temperatures and found that for relating temperature-affected NIR spectra to other properties of interest, LS-SVM was superior to PLS. Similarly, Borin et al. (2006) used NIR spectroscopy and LS-SVM for the quantification of common adulterants found in powdered milk samples and found that LS-SVM models produced lower prediction errors compared to PLS. Fernández-Ahumada et al. (2008), working with NIR for the prediction of ingredient composition in a large data set of intact compound feeds, reported a reduction of 49% in the standard error of prediction values for LS-SVM models compared to PLS models.
To our knowledge, there is only one study published using kernel regressions for the prediction of wood properties by NIR (Cogdill et al., 2004). For this reason, the objectives of this study were: (1) to develop nonlinear NIR calibration models for the prediction of wood properties in *P. taeda* using RBF-PLS and LS-SVM, and (2) to compare the calibration and prediction performance of the kernel regression methods with the traditional PLS approach.

7.3 Material and methods

7.3.1 Data set

Wood samples were obtained from *P. taeda* progeny tests established by members of the North Carolina State University Cooperative Tree Improvement Program between 1987 and 1992 in the south-eastern United States. The field tests were located in three regions (central Georgia (GA), south-west South Carolina (SC), and North Carolina (NC) Atlantic coast). A 6-tree disconnected diallel mating design was utilized to produce progeny of 12 parents and a randomized complete block design was used on each test location. Thirty full-sib families plus one check lot were available for sampling on each test site.

A total of 20 trees representing the different full-sib families were destructively sampled in each region (Table 7.1). From each selected tree, a 4 cm thick wood disk was extracted at breast height (1.4 m from the base) and a 12.5 × 12.5 mm bark-to-bark sample was obtained from the center of the disk.

Each bark-to-bark sample was divided at the pith and one half was prepared for data collection. The selected half was dried at 50°C for 24 hours, glued into wood holders and sectioned along its longitudinal axis using a twin-blade saw to produce a radial longitudinal strip approximately 2 mm thick from the center of each pith-to-bark sample.

7.3.2 Near infrared spectroscopy

Near infrared (NIR) diffuse reflectance spectra were obtained in consecutive 10-mm sections from the radial longitudinal face of each 2-mm strip using a FOSS NIRSystems (Laurel MD,
USA) Model 5000 spectrophotometer. The radial strips were held in a custom-made holder and a $5 \times 10$ mm mask was used to ensure that a constant area was tested (Schimleck et al., 2001b). NIR spectra were collected over the range of 1100-2498 nm with a resolution of 2 nm. The instrument reference used was a ceramic standard. The estimated instrumental noise, obtained by replicate measurements of the instrument’s reference, was 1.2E-5 AU (absorbance units). The measurements were made in a conditioned atmosphere at 40% relative humidity (RH) and 20°C.

7.3.3 SilviScan data

Smaller radial strips, 2 mm tangentially $\times$ 7 mm longitudinally, were cut from the strips scanned for NIR analysis and used for air-dry density (AD; kg/m$^3$), microfibril angle (MFA; deg), stiffness (MOE; GPa), tracheid coarseness (COARS; µg/m), and tracheid wall thickness (WTHICK; µm) determination using SilviScan-3 (Paprican, Vancouver BC). The length of the SilviScan strips varied depending on the pith-to-bark length of the samples used for NIR.

AD was determined by X-ray densitometry at a resolution of 25 µm; MFA was measured by X-ray diffractometry at 5-mm resolution using the variance method developed by Evans (1999); MOE estimates were obtained using a combination of X-ray densitometry and X-ray diffractometry data collected at 5-mm resolution (Evans, 2006); while COARS and WTHICK were obtained by image analysis using the relationships given in Evans (1994):

$$\text{PERIM} = 2 \times (\text{RAD} + \text{TAN}) \quad (7.1)$$

$$\text{COARS} = \text{RAD} \times \text{TAN} \times \text{AD} \quad (7.2)$$

$$\text{WTHICK} = \frac{\text{PERIM}}{8} - \frac{1}{2} \times \sqrt{\frac{\text{PERIM}^2}{16} - \frac{\text{COARS}}{d}} \quad (7.3)$$

where RAD and TAN are the radial (pith-to-bark direction) and tangential (parallel to the rings) tracheid dimensions (µm), respectively, PERIM is the external perimeter of rectangular tracheid cross-section (µm), and $d$ is the tracheid wall density, approximately 1500 kg/m$^3$ for all softwoods (Kellogg et al., 1975).
The SilviScan strips were not resin extracted and, as with NIR analysis, the measurements were made in a controlled environment of 40% RH and a temperature of 20°C.

7.3.4 NIR DATA PRE-PROCESSING AND SELECTION OF THE TRAINING AND PREDICTION SETS

Raw NIR data (RAW) was pre-treated using first (1D) and second derivatives (2D), standard normal variate (SNV), SNV plus detrending (SNVD), and multiplicative scatter correction (MSC). Each pre-treatment was applied individually to the raw data and they were not combined. 1D and 2D spectra were obtained by the Savitzky-Golay method (Savitzky and Golay, 1964) using a nine-point window as described by Wentzell and Brown (2000). SNV and SNVD pre-treatments were applied using the methodology given in Barnes et al. (1989). For MSC, the methodology described by Geladi et al. (1985) was utilized.

The mean-centered raw and pre-processed NIR matrices were compressed by principal component analysis (PCA) using the eigenvalue decomposition of the variance-covariance matrix (Jolliffe, 2002). The final number of components (PC’s) to retain was determined by randomization (Ledauphin et al., 2004). After defining the number of PC’s to use for each data set, the $Q$- and Hotelling’s $T^2$-statistic of each 10-mm section was calculated to identify potential outliers in the calibration phase (Mora and Schimleck, 2008b). Radial strips with any 10-mm sections identified as outliers were assigned to the prediction set.

Three radial strips were consistently classified as outliers in the different data sets (RAW and pre-processed) and therefore were removed from the analysis. The number of wood strips for prediction was fixed to 12 (2 per region-age combination), therefore a second selection of prediction samples was carried out by uniformly selecting radial strips from the training sets to complete the number of samples required for prediction. The number of training and prediction samples and the sample statistics of each data set is given in Table 7.2.
7.3.5 PLS calibration models

Nonlinear iterative partial least squares (NIPALS) models (Martens and Næs, 1989) were used to fit a set of initial NIR calibration models for the different wood properties with a maximum of 20 factors. The final number of factors (NF) to retain was determined by cross-validation (Esbensen, 2002) and by randomization (Wiklund et al., 2007). Models fitted using the cross-validation criteria will be referred to in the text as PLS-CV and models fitted using the randomization test to as PLS-RT.

Calibration performance was assessed using the coefficient of determination for calibration ($R^2_c$) and the root mean square error of cross-validation (RMSECV). Prediction performance was evaluated using the coefficient of determination for prediction ($R^2_p$) and the root-mean square error of prediction (RMSEP). In addition, the ratio of performance to deviation was calculated for calibration ($RPD_c$) and prediction ($RPD_p$). RPD is a standardized expression of the calibration and prediction errors with respect to the standard deviation of the training and prediction sets, respectively. RPD values greater than 2.5 are considered satisfactory for screening (Williams and Sobering, 1993).

Based on the prediction results obtained from the PLS models for the individual wood properties using all the data sets, the pre-processing technique that gave the lowest prediction errors (RMSEP) was selected for further analysis with kernel regressions (not shown), hence reducing the number of nonlinear models to fit. For AD and MOE, first derivative (1D) gave the best results. For COARS and WTHICK, second derivative (2D) resulted in the lowest RMSEP, while SNV was the pre-treatment that showed the best results for MFA.

7.3.6 RBF-PLS calibration models

The RBF-PLS approach developed by Walczak and Massart (1996b) was one of the nonlinear alternatives adopted in this work. RBF-PLS replaces the $X_{[m \times n]}$ matrix of spectral data by a symmetrical activation matrix $A_{[m \times m]}$ of the form:
where $\Theta_i$ is the Gaussian RBF characterized by a center and a width ($\sigma$), for each training sample $i = 1, 2, \ldots, m$. Centers of the RBFs are defined by the coordinates of the training samples determined by PCA. PLS is then applied to model $Y$ and $A$ as:

$$Y = TC^T + F$$

where $Y$ is the vector of wood properties, $T$ is the scores matrix of $A$, $C$ is the matrix of loadings, and $F$ is matrix of residuals. The superscript $T$ denotes transpose.

The tuning of $\sigma$ was carried out by Monte Carlo cross-validation (MCCV; Gourvéneec et al., 2003; Xu et al., 2004b). Values ranging from $\sigma = 1$ to $\sigma = 100$ with a step size of 1 unit were tested, using twice the size of the training data sets as number of iterations and 35 randomly selected samples left out for each iteration. Before analysis, the data sets were scaled to the $[0,1]$ interval as suggested by Walczak and Massart (1996b). RBF-PLS calibration and prediction performances were evaluated using the same fit and prediction statistics used for PLS models.

### 7.3.7 LS-SVM Calibration Models

The second alternative for the development of nonlinear calibration models used in this work was the LS-SVM approach of Suykens et al. (2002). For regression problems, the LS-SVM model function estimation is given by the expression:

$$y(x) = \sum_{i=1}^{m} \alpha_i K(x, x_i) + b$$
where $\alpha_i$ with $i = 1, 2, \ldots, m$ is the $i$th element of the vector of parameters $\alpha$, $b$ is a bias term, and $K(x, x_i)$ is the kernel function that measures the closeness of a new spectrum $x$ to the spectrum of the $i$th sample ($x_i$) in the training set. Similar to RBF-PLS, the kernel adopted for LS-SVM was the Gaussian RBF. The resulting kernel matrix $K_{[m \times m]}$, of elements $K(x, x_i)$, is similar to the activation matrix given in Eq. 7.4.

Under this formulation, the $\alpha_i$’s are called support values and training samples with support values significantly larger than zero are called support vectors. Following Fernández-Ahumada et al. (2008), Eq. 7.6 can be regarded as a linear combination of $K(x, x_i)$ with coefficients $\alpha_i$ which are estimated by modified least squares using the training samples. The modification is a penalty on the square length of the coefficient vector controlled by the tuneable regularization parameter $\gamma$. For prediction purposes, Eq. 7.6 can be viewed as a linear combination of $\alpha_i$’s with weights $K(x, x_i)$. Therefore, training samples that are spectrally similar to the unknown samples will have a larger influence on the predicted values.

The $\sigma^2$ and $\gamma$ parameters for all the LS-SVM models were optimized by cross-validation along with a search over a range of values of 1 to 20,000 for both parameters. Subregions within the grid search were tested exhaustively to locate the final parameter values as suggested by Cogdill et al. (2004). LS-SVM model performance was evaluated as in the previous approaches.

7.3.8 Software

PLS calibrations were fitted using R version 2.8.1 (R Development Core Team, 2008) and the pls package (Mevik and Wehrens, 2007). MATLAB (version 7.5, The MathWorks, Inc.) was used to fit the kernel regressions. RBF-PLS models were obtained using the MATLAB functions implemented in the TOMCAT toolbox (Daszykowski et al., 2007b) and the LS-SVM models were fitted using the functions provided in the LS-SVMLab (version 1.5) toolbox (Suykens et al., 2002).
7.4 Results and discussion

7.4.1 PLS calibrations

The fit and prediction statistics of the PLS calibration models are given in Table 7.3. These values are consistent with different results reported in the literature (e.g. Schimleck et al., 2003c, 2004c; Cogdill et al., 2004; Jones et al., 2005b,a, 2007; Mora et al., 2008).

In general, model dimensionality obtained by cross-validation and randomization were in agreement, except for MFA and MOE. The criterion used by cross-validation to define NF seemed to be too restrictive for these two properties compared to the approach utilized by the randomization test that added extra (statistically significant) factors to account for nonlinearities in the data sets. Figure 7.1 shows the plot of raw residuals obtained from PLS-CV for AD, MFA, MOE, and COARS. Residuals for both, MFA and MOE, show some degree of nonlinearity at low values that were not corrected by the pre-processing techniques applied to the spectral data (SNV and 1D, respectively).

The coefficient of determination for calibration ($R^2_c$) varied from moderately strong (AD, COARS, WTHICK) to very strong (MFA and MOE) and was not affected by the method used to determine model dimensionality. However, the effects of different NF was clearly reflected in the RPD$_c$ values, which varied from 2.3–3.7 for PLS-CV to 2.3–5.1 for PLS-RT. Although the RPD values are acceptable for screening, it must be noted that wood property NIR calibrations rarely display RPD $> 4–5$, which suggests that the limits given by Williams and Sobering (1993) must be interpreted differently when wood is analyzed.

The predictive ability of some of the PLS models was improved by the addition of factors. In the case of COARS, by augmenting the model dimensionality by 1 factor, the $R^2_p$ increased from 0.78–0.84 and the RMSEP reduced from 46.24–37.97 µg/m. The reduction in RMSEP led to an increase in the RPD$_p$ values from 2.0–2.4.

For MFA, by adding 5 factors in the model, $R^2_p$ increased from 0.81–0.91, RMSEP was reduced from 3.04–2.37 deg, and RPD$_p$ increased from 2.4–3.0. For MOE, the change in $R^2_p$
was less dramatic after the addition of 5 factors, but the reduction in RMSEP from 1.01–0.79 GPa was translated into a significant increase in RPD from 4.1–5.3.

These results show that the addition of significant factors does not necessarily cause model overfitting, as it might be expected from the results obtained from cross-validation, and in some cases it is required to adequately capture nonlinear features of the data. Although the addition of PLS factors can be used to overcome nonlinearities, according to Cogdill et al. (2004) it is expected that PLS regressions with high NF would fail to generalize due to overfitting and that the validity of including a large number of factors needs to be demonstrated.

7.4.2 RBF-PLS calibrations

The NF and $\sigma^2$ parameters for the different calibration models fitted using the RBF-PLS approach are given in Table 7.4.

In this methodology, NF can be large contrary to classical PLS where usually a small NF is preferred to avoid overfitting. The optimal $\sigma^2$, obtained by MCCV, was selected for each model based on the smallest RMSECV for a given combination of $\sigma$ and NF (Fig. 7.2).

It is difficult to make an interpretation of the parameter values as they are adjusted simultaneously and probably vary depending on the pre-processing technique applied to the spectral data. Table 7.4 suggests that the higher the degree of nonlinearity the higher the NF. High $\sigma^2$ values may also indicate that differences between similar training samples are less resolved and the models are confined to be more “linear” (Cogdill and Dardenne, 2004).

One difficulty of using RBF-PLS is the lack of standards to compare with; therefore, the final evaluation of the quality of the models must be based on the predictive ability of the calibrations only.

RBF-PLS calibrations were superior to PLS models in terms of fit statistics (Table 7.5). $R_c^2$ ranged from moderately strong for COARS to very strong ($>0.90$) for AD, WTHICK, MFA, and MOE. RMSECV errors were significantly decreased, with reductions ranging from
9.5% (COARS) to 35.5% (MFA) compared to PLS-CV, and from 7.8% (COARS) to 25.8% (MFA) compared to the errors obtained from PLS-RT. These reductions were translated into higher RPD$_c$ values ranging from 2.5–5.7.

Although the calibrations were significantly improved, the prediction statistics derived from the RBF-PLS models were similar to those obtained with PLS-CV and, in almost all cases, were inferior compared to the values derived from the PLS-RT models (Table 7.5). Similar results were found in the studies by Walczak and Massart (1996b) and Cogdill and Dardenne (2004).

MFA, MOE, and COARS were properties that benefited from using the RBF-PLS methodology compared to PLS-CV. For the MFA data, $R^2_p$ increased from 0.88–0.91, RMSEP was reduced from 3.04–2.39 deg, and RPD$_p$ was increased from 2.4–4.4. For MOE, the most important result was the reduction of RMSEP from 1.01–0.92 GPa and the increase of RPD$_p$ from 4.1–4.5. In the case of COARS, $R^2_p$ increased from 0.78–0.84, RMSEP decreased from 46.24–37.49 µg/m, and RPD$_p$ increased from 2.0–2.5.

The results indicate that RBF-PLS performed better than PLS for modeling nonlinear features of the data sets. This was also confirmed by the plot of raw residuals (not shown). Considering that the prediction statistics were similar or better than those obtained by the traditional approach of PLS-CV, RBF-PLS represents a good alternative for the development of NIR calibration models for the estimation of wood properties. However, the same or even better predictive ability can be achieved by PLS models that consider alternative approaches for determining the model dimensionality, such as the randomization test proposed by Wiklund et al. (2007).

7.4.3 LS-SVM calibrations

The optimized parameter values ($\gamma$ and $\sigma^2$) for the LS-SVM models are given in Table 7.4. The optimization process was carried out by cross-validation using a custom-written MATLAB function provided by Dr. Robert Cogdill previously utilized in the work by Cogdill
and Dardenne (2004) and Cogdill et al. (2004). An example subregion of the optimization surface obtained for the AD data set is shown in Fig. 7.3.

Compared to the parameters obtained by Cogdill et al. (2004), Table 7.4 shows lower $\gamma$ and higher $\sigma^2$ values. Following the interpretation given by Cogdill and Dardenne (2004), these results may indicate a low risk of overfitting (an increase in $\gamma$ is analogous to an increase in NF) and, as in the case of RBF-PLS, the high $\sigma^2$ values suggest that the LS-SVM did not take into account small differences between similar training samples in the fitting process (recall that $\sigma^2$ controls the degree of nonlinearity that can be modeled).

According to the same authors, the optimum level of $\gamma$ and $\sigma^2$ seems to be related to the density of the training data and the signal-to-noise ratio and nonlinearity of the data sets, respectively, i.e., the values of the parameters are data-dependent.

LS-SVM calibrations exhibited superior fit statistics for all wood properties compared to PLS-CV and PLS-RT (Table 7.5). In all cases, $R^2_c$ values were very strong ($\geq 0.90$) and RMSECV errors were significantly reduced, from 9.4% (COARS) to 33.6% (MOE) when compared to PLS-CV and from 7.6% (COARS) to 25.8% (MFA) compared to the errors from PLS-RT. RPD$_c$ values ranged from 2.7–5.6. Fig. 7.4 shows an example of calibration models obtained for MFA where RBF-PLS and particularly LS-SVM resulted in tighter fitted and predicted values compared to PLS.

It is interesting to note in Fig. 7.4 that, although the nonlinear trend in MFA was corrected, the residuals tend to increase with values of MFA $> 30$ deg. This phenomenon is related to the methodology used by SilviScan-3 to measure this property and not necessarily result of the fitting algorithm. As MFA increases its determination by X-ray diffractometry becomes more difficult owing to the broadening and weakening of the 002 reflections (Evans, 1999). The signal-to-noise ratio in the diffraction pattern also decreases and the precision of the MFA measurements is reduced (Cogdill et al., 2004).

Compared to PLS-CV calibrations, LS-SVM resulted in models with better predictive abilities for all properties, except WTHICK that gave the same statistics (Table 7.5). $R^2_p$
varied from 0.85 (COARS) to 0.95 (MOE) with reductions in RMSEP up to 22.8% in the case of COARS. Depending on the property considered, \( RPD_p \) values varied between 2.6–4.7.

With respect to the previous kernel method (RBF-PLS), LS-SVM performed similarly in terms of fit statistics but was slightly superior for prediction. Cogdill and Dardenne (2004) reported a similar finding. As with RBF-PLS, when LS-SVM models are compared against PLS-RT, the differences in performance between kernel methods and traditional PLS tend to disappear. Although the prediction statistics derived from the PLS-RT models were superior in some cases to LS-SVM and in most cases to RBF-PLS, the plot of residuals (not shown) revealed the presence of nonlinearities, especially at low values, that were not corrected by augmenting the models rank. In this sense, there seems to be a compromise between having good predictions and more robust calibration models.

However, the performance of the kernel regression methods could be affected by the pre-processing techniques applied to the NIR data. In this work, the spectral data used for modeling the wood properties was pre-processed differently according to the prediction statistics derived from the PLS-CV models. In addition, the samples used for training and prediction were selected based on the results from the PCA analysis performed on the NIR data. Alternatively, the selection of samples could have been done considering the kernel matrix, calculated from spectral data pre-processed using different algorithms.

7.5 Conclusions

The main purpose of this study was to establish whether the use of nonlinear calibration methodologies were able to improve the results obtained by PLS-CV for the prediction of wood properties in \( P. \) taeda by NIR spectroscopy. The nonlinear methodologies selected were a class of kernel regression, a non-parametric estimation technique used to find a nonlinear relation between a pair of random variables in an induced feature space.

The two kernel methods evaluated, RBF-PLS and LS-SVM, gave substantial improvements over PLS-CV in terms of fit and prediction statistics. The reductions in RMSECV
resulted in calibration models with higher $\text{RPD}_c$ values. From the properties analyzed, MFA and MOE showed the most significant improvements in terms of calibration statistics, indicating that nonlinear methodologies are a good alternative for the development of calibration models when the properties exhibit a nonlinear relationship between the NIR-fitted and measured values.

Although RBF-PLS and LS-SVM performed similar in terms of calibration statistics, the predictive ability of the LS-SVM calibrations were superior to both, RBF-PLS and PLS-CV models. RBF-PLS performed better than PLS-CV for MFA, MOE, and COARS, all properties with some degree on nonlinearity. However, LS-SVM predictions were similar or better in cases were the properties were considered to be “linear”, such as AD and WTHICK.

The predictive ability of the nonlinear models is directly related to the value of the parameters that must be optimized for each method which control how “global” or “local” the resulting model is. This step is critical for assessing the quality of the models.

7.6 References


Table 7.1. Individual tree mean characteristics on each test site. Standard errors are given in parenthesis.

<table>
<thead>
<tr>
<th>Test</th>
<th>Location</th>
<th>Physiographic region</th>
<th>Age (years)</th>
<th>n</th>
<th>DBH (cm)</th>
<th>HT (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GA</td>
<td>Piedmont</td>
<td>15</td>
<td>10</td>
<td>18.8 (0.6)</td>
<td>13.8 (0.2)</td>
</tr>
<tr>
<td>1</td>
<td>GA</td>
<td>Piedmont</td>
<td>16</td>
<td>10</td>
<td>19.0 (0.5)</td>
<td>15.7 (0.1)</td>
</tr>
<tr>
<td>2</td>
<td>SC</td>
<td>Coastal plain</td>
<td>14</td>
<td>10</td>
<td>22.7 (0.7)</td>
<td>20.1 (0.4)</td>
</tr>
<tr>
<td>2</td>
<td>SC</td>
<td>Coastal plain</td>
<td>15</td>
<td>10</td>
<td>23.9 (0.5)</td>
<td>20.2 (0.5)</td>
</tr>
<tr>
<td>3</td>
<td>NC</td>
<td>Coastal plain</td>
<td>18</td>
<td>10</td>
<td>23.9 (1.2)</td>
<td>19.0 (0.4)</td>
</tr>
<tr>
<td>3</td>
<td>NC</td>
<td>Coastal plain</td>
<td>19</td>
<td>10</td>
<td>22.2 (1.4)</td>
<td>19.7 (0.6)</td>
</tr>
</tbody>
</table>

DBH, diameter at breast height (1.4 m); HT, total height.
Table 7.2. Range and standard deviation of the wood properties used for training and prediction.

<table>
<thead>
<tr>
<th>Property</th>
<th>Training set</th>
<th></th>
<th></th>
<th>Prediction set</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
<td>Std</td>
<td>m</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>AD</td>
<td>327.2</td>
<td>810.5</td>
<td>98.2</td>
<td>385</td>
<td>328.7</td>
<td>754.2</td>
</tr>
<tr>
<td>MFA</td>
<td>10.8</td>
<td>45.5</td>
<td>7.55</td>
<td>384</td>
<td>9.9</td>
<td>38.9</td>
</tr>
<tr>
<td>MOE</td>
<td>1.6</td>
<td>20.3</td>
<td>3.9</td>
<td>376</td>
<td>1.4</td>
<td>20.1</td>
</tr>
<tr>
<td>COARS</td>
<td>260.2</td>
<td>633.3</td>
<td>81.3</td>
<td>381</td>
<td>263.7</td>
<td>645.8</td>
</tr>
<tr>
<td>WTHICK</td>
<td>1.7</td>
<td>4.3</td>
<td>0.6</td>
<td>384</td>
<td>1.7</td>
<td>4.5</td>
</tr>
</tbody>
</table>

AD, Air-dry density (kg/m³); MFA, Microfibril angle (deg); MOE, Stiffness (GPa); COARS, Coarseness (µg/m); WTHICK, Wall thickness (µm); m, number of samples (10-mm sections) used for training; m<sub>p</sub>, number of samples (10-mm sections) used for prediction.
Table 7.3. Fit and prediction statistics for the PLS models.

<table>
<thead>
<tr>
<th>Property</th>
<th>PLS + cross-validation (PLS-CV)</th>
<th>PLS + randomization (PLS-RT)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NF  $R^2_c$ RMSECV RPD$_c$ $R^2_p$ RMSEP RPD$_p$</td>
<td>NF  $R^2_c$ RMSECV RPD$_c$ $R^2_p$ RMSEP RPD$_p$</td>
</tr>
<tr>
<td>AD</td>
<td>4 0.83 41.83 2.3 0.85 42.18 2.6</td>
<td>4 0.83 41.83 2.3 0.85 42.18 2.6</td>
</tr>
<tr>
<td>MFA</td>
<td>9 0.92 2.46 3.1 0.88 3.04 2.4</td>
<td>14 0.95 2.29 3.3 0.91 2.37 3.0</td>
</tr>
<tr>
<td>MOE</td>
<td>6 0.94 1.07 3.7 0.94 1.01 4.1</td>
<td>14 0.98 0.77 5.1 0.96 0.79 5.3</td>
</tr>
<tr>
<td>COARS</td>
<td>3 0.83 35.69 2.3 0.78 46.24 2.0</td>
<td>4 0.84 35.02 2.3 0.84 37.97 2.4</td>
</tr>
<tr>
<td>WTHICK</td>
<td>4 0.87 0.23 2.7 0.90 0.21 3.2</td>
<td>4 0.87 0.23 2.7 0.90 0.21 3.2</td>
</tr>
</tbody>
</table>

AD, Air-dry density (kg/m$^3$); MFA, Microfibril angle (deg); MOE, Stiffness (GPa); COARS, Coarseness (µg/m); WTHICK, Wall thickness (µm); NF, number of factors; $R^2_c$, coefficient of determination for calibration; RMSECV, root mean square error of cross-validation; RPD$_c$, ratio of performance to deviation for calibration; $R^2_p$, coefficient of determination for prediction; RMSEP, root-mean square error of prediction; RPD$_p$, ratio of performance to deviation for prediction.
Table 7.4. Kernel regression parameters.

<table>
<thead>
<tr>
<th>Property</th>
<th>RBF-PLS</th>
<th>LS-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NF</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>AD</td>
<td>29</td>
<td>3969</td>
</tr>
<tr>
<td>MFA</td>
<td>34</td>
<td>5476</td>
</tr>
<tr>
<td>MOE</td>
<td>32</td>
<td>2809</td>
</tr>
<tr>
<td>COARS</td>
<td>9</td>
<td>1225</td>
</tr>
<tr>
<td>WTHICK</td>
<td>15</td>
<td>7569</td>
</tr>
</tbody>
</table>

AD, Air-dry density (kg/m$^3$); MFA, Microfibril angle (deg); MOE, Stiffness (GPa); COARS, Coarseness (µg/m); WTHICK, Wall thickness (µm); NF, number of factors.
Table 7.5. Fit and prediction statistics for the kernel regression models.

<table>
<thead>
<tr>
<th>Property</th>
<th>RBF-PLS</th>
<th>LS-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2_c$</td>
<td>RMSECV</td>
</tr>
<tr>
<td>AD</td>
<td>0.91</td>
<td>35.73</td>
</tr>
<tr>
<td>MFA</td>
<td>0.96</td>
<td>1.70</td>
</tr>
<tr>
<td>MOE</td>
<td>0.98</td>
<td>0.69</td>
</tr>
<tr>
<td>COARS</td>
<td>0.85</td>
<td>32.29</td>
</tr>
<tr>
<td>WTHICK</td>
<td>0.91</td>
<td>0.19</td>
</tr>
</tbody>
</table>

AD, Air-dry density (kg/m$^3$); MFA, Microfibril angle (deg); MOE, Stiffness (GPa); COARS, Coarseness (µg/m); WTHICK, Wall thickness (µm); $R^2_c$, coefficient of determination for calibration; RMSECV, root mean square error of cross-validation; RPD$_c$, ratio of performance to deviation for calibration; $R^2_p$, coefficient of determination for prediction; RMSEP, root-mean square error of prediction; RPD$_p$, ratio of performance to deviation for prediction.
Figure 7.1. Raw residuals for the PLS-CV calibration models. AD, air-dry density; MFA, microfibril angle; MOE, stiffness; COARS, coarseness.
Figure 7.2. RMSECV versus $\sigma$ (kernel width) obtained from cross-validation using the MFA data and the RBF-PLS approach. The arrow shows the value of the parameters used in the final model. RMSECV, root mean square error of cross-validation; NF, number of factors associated to each $\sigma$ value.
Figure 7.3. Subregion of the parameter optimization surface for AD. RMSECV, root mean square error of cross-validation; $\sigma^2$, kernel width; $\gamma$, regularization parameter.
Figure 7.4. NIR calibration models for MFA. PLS-CV, PLS and number of factors determined by cross-validation; PLS-RT, PLS and number of factors determined by randomization; RBF-PLS, radial basis functions PLS; LS-SVM; least-squares support vector machines. Training samples are shown as gray circles. Prediction samples are in black. Fit and prediction statistics are given in Tables 7.4 and 7.5. The segmented line represents the line of equivalence.
Chapter 8

Determination of within-tree variation of *Pinus taeda* wood properties by
near infrared spectroscopy: Whole-tree wood property maps

8.1 Abstract

The application of three spatial interpolation algorithms (Akima, universal kriging, and semiparametric regression methods) to the study of within-tree variation of air-dry density and microfibril angle predicted by near-infrared models in *Pinus taeda* trees is presented. The theory behind each algorithm is briefly reviewed and its application to the study of the within-tree variation of wood properties is addressed. The use of predicted wood properties as control data points for spatial interpolation was successful and allowed the generation of maps showing within-tree variability at high resolution and quality. The maps obtained by the different methods showed similar features indicating that air-dry density and microfibril angle were highly predictable as a function of their positions within the trees.

8.2 Introduction

The successful application of near infrared (NIR) spectroscopy to the study of the within-tree variation of wood properties relies on two basic aspects: the provision of accurate data and the use of adequate modelling techniques. The incorporation of SilviScan (Evans, 1994) for the acquisition of high resolution data has facilitated the development of NIR calibration models to predict a range of wood properties for different species. Strong relationships between SilviScan data and NIR estimates have been reported for *Pinus radiata* (Schimleck and Evans, 2004), *Pinus taeda* (Schimleck et al., 2003c; Jones et al., 2005b), and *Eucalyptus nitens* (Schimleck et al., 2006a).

Traditionally, NIR studies based on SilviScan measured wood properties have been limited to evaluations at breast height. An exception is the work by Schimleck et al. (2009) that developed multiple height NIR calibration models using SilviScan data for the determination of within-tree variation of several wood properties in *P. taeda*. According to So et al. (2004b), when the study involves sampling points at different heights and distances from the pith,
the range and variation of the wood properties within a tree can be better visualized and
described in the form of wood property maps.

Several examples of wood property maps exist in the literature but, with few exceptions,
little detail is given about the methodology employed to obtain them. The development of
any kind of map requires the utilisation of data interpolation algorithms and the nature of
the algorithm selected will determine their interpretability and validity.

Evans et al. (1993) presented one of the first known approaches to generate digital maps
of wood density and fibre property variation within-trees based on data combined from
image analysis and X-ray densitometry (SilviScan in its earliest stages of development).
Some years later, Ona et al. (1995) obtained maps of polysaccharides content for *Eucalyptus
camaldulensis* and *Eucalyptus globulus* based on increment cores collected at different heights.
The samples were measured in sections and this information was used to create a simple
representation of the internal variation of various polysaccharides within the trees. Tian
et al. (1995b) developed a model-based approach to describe the patterns of basic density
variation within *P. radiata* trees using a logistic regression model with ring number and stem
height as covariates. The model was then used to interpolate the values to any section of
wood within the trees but the authors failed to account for the spatial dependence of the
data in their approach.

More recently, Hudson et al. (2000) developed the concepts of fibre morphology and
oblique axes for developing fibre length and coarseness maps for *E. globulus* using a simple
contour interpolation algorithm. Similarly, Turner et al. (2001) obtained maps showing the
within-tree variation of collapsibility (a ratio that determines the ability of the fibre to
defibrate without being destroyed) for *Pinus patula* based on kriging analysis with little
discussion on the assumptions made to interpolate the data.

For *P. taeda*, whole-tree maps showing the internal variation of mechanical properties
within a tree have been reported in the literature (So et al., 2002; Mott et al., 2002; Groom
et al., 2002b) with no mention of the methodology used to generate them. A more detailed
analysis based on X-ray densitometry data was presented by Daniels et al. (2002). These authors used a modified non-linear mixed model to obtain an average map showing the changes of latewood specific gravity within a *P. taeda* tree. The application of this methodology, although statistically sound, is difficult to implement in the case of NIR studies on solid wood samples where the number of measured data points is usually more limited.

A different approach was recently presented by Grabner and Wimmer (2006) who used a correlation analysis to create wood property maps for *Picea abies* using each terminal shoot as a reference point. According to the authors the information displayed by such maps is the best way to represent wood quality in the entire tree. However, this analysis does not consider the spatial nature of the data.

The ability of SilviScan to measure several wood properties at high spatial resolution has facilitated the generation of high quality whole-tree wood properties maps without the need of data interpolation. Examples of these are given in the work of Evans et al. (2000) and Lundqvist (2002); however, maps of within-tree variation based on SilviScan data can be costly if a large number of samples are analysed. Ideally, by using an appropriate method we could generate similar maps using fewer measurements. This alternative was presented by Schimleck and Michell (1998) who developed whole-tree pulp yield maps for *E. nitens* trees based on NIR predicted values. The maps were created using a custom interpolation program but no additional details are given.

Considering that the importance of using maps to present the results of wood quality studies has been recognized for years, the lack of an adequate description of the methods employed has certainly limited their use. For this reason, the aim of this work is to show how the utilization of three different spatial interpolation algorithms can assist in developing air-dry density and microfibril angle whole-tree maps for *P. taeda* using, in this study, wood property data predicted by NIR calibration models. The methods presented are not restricted to NIR data but can be applied to wood property data collected in any form. The theory behind each method is briefly reviewed and its application to the study of within-tree
variation of wood properties is addressed. The discussion is focused on the interpolation techniques and not on the description of the within-tree variation of wood properties which can be found elsewhere. Details about the NIR model used to predict the wood properties can be found in the first article of this series.

8.3 Material and methods

8.3.1 Data set

A subsample of the data set analysed by Schimleck et al. (2009) was used to create the maps showing the within-tree variation of wood properties. Briefly, eighteen 13-year-old *P. taeda* trees from a half-sib progeny trial established near Bainbridge, Georgia, U.S.A. were destructively sampled for wood properties analysis. From each tree, 2.5 cm thick wood disks were extracted at 1.5 m intervals from the base of each tree up to a height of 2.5 cm top diameter. Radial sections of 12.5 by 12.5 mm were obtained from the middle portion of the disks and prepared to give radial-longitudinal samples of approximately 2 mm thick from the centre of the sections. The samples were not resin-extracted.

NIR spectra were collected from the radial-longitudinal face of the samples at intervals of 10 mm using a FOSS NIRSystems 5000 scanning spectrometer. Air-dry density measured in kg/m$^3$ (DEN) and microfibril angle (MFA) measured in degrees (deg) were predicted on each 10-mm segment using the NIR calibration model fitted with raw data given in Schimleck et al. (2009). The predicted values were used as control points for the spatial analysis. Prior to analysis, the trees were grouped in 3 classes of 6 trees each based on the length of the basal radial strips (group 1: <100 mm; group 2: 100 to 110 mm; group 3: >110 mm) to ensure that trees of similar size were used in the interpolation of the wood properties data.
8.3.2 Spatial interpolation

Akima’s method

This is a method of bivariate interpolation and smooth surface fitting, originally developed for control points in rectangular grids (Akima, 1974) and later extended to control points irregularly distributed in a two-dimensional space (Akima, 1978).

The first step consists of dividing the two-dimensional space defined by the bivariate predictor \( \mathbf{x} \in \mathbb{R}^2 \) (coordinates) into a number of triangular cells, each having projections of three data points in the plane as its vertices. The triangle vertices, as shown in Figure 8.1 for the trees in group 3, are located at the control points. Details of the triangulation method can be found in Akima (1978) and Ripley (1981).

Interpolation of the \( y \) values (DEN or MFA) is obtained by fitting a fifth-order surface in \( \mathbf{x} = (x_1, x_2) \) within each triangle. Such a polynomial surface has 21 coefficients to be estimated and 18 conditions are imposed to fit the surface with first and second derivatives values given at the three vertices. The remaining 3 conditions are that the derivative of the surface perpendicular to each side should be a cubic function of the distance from the side. The net effect is to achieve a continuously differentiable surface within the triangulation.

The local polynomials used to obtain the values of the function in a triangle are given by the expression:

\[
y_i(x_{1i}, x_{2i}) = \sum_{j=0}^{5} \sum_{k=0}^{5-j} (q_{jk} x_{1i}^j x_{2i}^k) \quad \text{(8.1)}
\]

where \( y_i \) is the interpolated wood property for the \( i \)th point in a triangle, \( x_{1i} \) and \( x_{2i} \) are coordinate distances relative to some local origin, in this work measured in centimetres from the pith in the radial direction and in meters from the base of the tree in the vertical direction, \( q_{jk} \) are the coefficients of each local polygon to be determined, and \( j \) and \( k \) are dummy variables.
Universal kriging method

Following Schabenberger and Gotway (2005), suppose that the data $Z(s_1), \ldots, Z(s_n)$ are partial realizations of a random process $\{Z(s) : s \in D\}$ where $s$ denotes location and $D$ is a fixed domain. The objective of this method is to predict the value $Z(s_0)$ at the unsampled location $s_0$ from the control points at locations $s_1, \ldots, s_n$.

In this method it is assumed that $Z(s) = X(s)\beta + e(s)$, where $e(s)$ represents a zero mean second-order stationary random field with variogram $2\gamma(h)$ and $h = s_i - s_j$ denotes the distance vector (lag) between locations $s_i$ and $s_j$. By allowing a non-constant mean, i.e. $E\{Z(s)\} = X(s)\beta$, it is possible to account for trends in DEN or MFA values related to the position within the tree(s).

It is also assumed that $Z(s_0) = x(s_0)^T\beta + e(s_0)$, where $x(s_0)$ is a $p \times 1$ known vector of explanatory values associated with the unobserved location $s_0$. The interpolated wood properties at $s_0$ can be expressed as a linear combination of all the control points in the form:

$$\hat{Z}(s_0) = \sum_{i=1}^{n} a_i Z(s_i)$$

(8.2)

where the coefficients $a_i$ are chosen in order to minimize the mean-squared prediction error $\text{MSE} = E\{|\hat{Z}(s_0) - Z(s_0)|^2\}$

For the calculation of the vector $a$ in Equation 8.2, a variogram is required. A variogram is a measure of the spatial continuity of the data and given two sample locations $h$ units apart and the difference for the variable of interest, the semivariogram $\gamma(h)$, is defined as half the variance of this difference. The reader is referred to Schabenberger and Gotway (2005) and Cressie (1991) for details about fitting empirical semivariogram models to the data.

Semiparametric bivariate smoothing method

Smoothing methodology offers a means by which non-linear relationships can be handled without the restrictions of parametric models. The term smoothing usually involves the use
of regression models that contain at least one function being modelled nonparametrically, also known as semiparametric regression models. Bivariate smoothing, in particular, is concerned with the interactions between two continuous variables free of any structural assumptions on the way in which they affect the mean response.

Following Ruppert et al. (2003), consider the situation where the data are available on a response $y$ with a bivariate predictor $\mathbf{x} \in \mathbb{R}^2$ and we want to fit the model $y_i = f(x_i) + \varepsilon_i$, where $f$ is a smooth bivariate function and $x_i$ usually represents a geographical location or any two predictors for which additivity is not reasonably assumed.

In this particular case, $y_i$ represents the wood property of interest (DEN or MFA) observed at the control point $i$ ($i = 1, \ldots, n$), $x_i$ is the location of the point $i$, and $\varepsilon_i$ are uncorrelated errors with common variance $\sigma^2_\varepsilon$. The estimation of $f$ can be done by using radial basis functions with the family of basis functions corresponding to the thin-plate spline family (higher dimension approximate smoothing splines) and expressed as a linear mixed model of the form:

$$y = \mathbf{X}\beta + \mathbf{Z}u + \varepsilon, \quad \text{Cov}
\begin{bmatrix}
u \\
\varepsilon
\end{bmatrix}
= \begin{bmatrix}
\sigma^2_u I & 0 \\
0 & \sigma^2_\varepsilon I
\end{bmatrix}
$$

(8.3)

where $y$ is the response vector, $\mathbf{X} = \begin{bmatrix} 1 & x_i^T \end{bmatrix}_{1 \leq i \leq n}$ and $\mathbf{Z} = \mathbf{Z}_K \Omega^{-1/2}$ are design matrices, $\beta$ is the vector of coefficients associated to the fixed effects, $u$ is the vector of random effects, and $\sigma^2_u$ and $\sigma^2_\varepsilon$ are variance components obtained via restricted maximum likelihood (REML).

The matrix $\mathbf{Z}_K$ is expressed as:

$$\mathbf{Z}_K = \left[ \| x_i - \kappa_k \|^2 \log \| x_i - \kappa_k \| \right]_{1 \leq i \leq n, 1 \leq k \leq K}
$$

(8.4)

where $\kappa_1, \ldots, \kappa_K \in \mathbb{R}^2$ are a set of $K$ knots that cover the space defined by $\mathbf{x}$ and:
\[ \Omega = \left[ \sum_{1 \leq k, k' \leq K} \left( \| \kappa_k - \kappa_{k'} \|^2 \log \| \kappa_k - \kappa_{k'} \| \right) \right] \] (8.5)

The choice of the bivariate knots \( \kappa_k \) is done via an efficient space filling algorithm using the criterion \( K = \max\{20, \min(n/4), 150\} \). Figure 8.2 shows the position of the selected knots for the trees in group 3.

The interpolation of \( y \) to unobserved locations is obtained by creating a grid of \( x_0 \) values in the area of \( \mathbb{R}^2 \) of interest and obtaining the appropriate matrices \( X_0 \) and \( Z_0 \) which are then plugged in the expression \( \hat{y}(x_0) = X_0\hat{\beta} + Z_0\hat{u} \).

8.3.3 Software

R version 2.7.1 (http://www.r-project.org) was used for the analysis. The Akima interpolation method is part of the R system and its Fortran code is available elsewhere. The universal kriging models were fitted using the geoR package (Ribeiro Jr. and Diggle, 2001) and the semiparametric regression models were obtained using the SemiPar package (Wand et al., 2005).

8.4 Results and discussion

8.4.1 Akima’s method

The within-tree wood property maps generated with the Akima’s interpolation are shown in Figure 8.3.

The DEN maps showed, as expected, an increasing trend of DEN from pith to bark and the absence of a significant trend in the vertical direction. In \( P. \) taeda trees, density increases from the pith outward at all heights mainly as a result of an increase in latewood density, but the rise occurs faster and levels out sooner in the lower 2 m of examined trees (Megraw, 1985). This pattern of variation was adequately captured by the method which
also reflected that the NIR model used to predict DEN was appropriate (good quality of the control points).

The rapid radial increase of DEN below 2 m was more evident in the group of smaller trees where for a given distance from the pith they averaged higher DEN than trees in groups 2 and 3. For trees of the same age, this is usually the result of a different number of rings in a given diameter section. The maximum DEN values in group 1 were in the range of 800 to 850 kg/m$^3$ while in group 3 maximum DEN was less than 725 kg/m$^3$.

According to Megraw (1985), the wood of *P. taeda* for the first 10 rings from the pith has, on average, a 10 to 15% higher density near the base of the tree than at a height of 5 m. Above this height, the density of this inner zone stays essentially constant and, excluding the butt log, average density can be thought of as a simple function of ring position from the pith regardless of height. The maps in Figure 3 showed a similar trend where DEN bands running parallel to the pith were observed suggesting that height played a minor role in describing the internal variation of DEN compared to the effects of radial position, especially above a height of 3 m.

The plots on the bottom part of Figure 8.3 show the within-tree variation of MFA for the same groups of trees. The MFA values were similar across groups, ranging from a low of approximately 15 to a high of 40 deg. High values were observed below breast height (1.5 m) near the pith. The transition to low MFA values, toward the periphery of the trunk, was more rapid above 3 m than at the base of the trees.

The within-tree MFA variation shown in the maps coincides with the description given by Jordan et al. (2005b). These authors report that MFA in *P. taeda* is large near the pith and decreases rapidly to approximately 10 rings from the pith, and then continues dropping, regardless of height, but at a much slower rate until it essentially stabilizes. They also report that the decrease in MFA with ring position takes place at a slower rate near the base of the tree compared to higher in the tree and that this results in higher MFA values for a given
ring number from the pith at the butt and breast height regions than at locations higher in the tree.

In summary, the maps generated by Akima’s interpolation method captured the expected trends of DEN and MFA in *P. taeda* well. Because this method does not smooth the data, the resulting surface passed through all the control points (exact interpolator). The principal advantage of the algorithm was that only straightforward procedures were required and therefore, no problems concerning computational stability or convergence existed.

This method is especially useful for control points that are irregularly distributed in the space. Although there was no problem with interpolating from a regular grid, the triangulation-based methods, such as the one showed in Figure 1, are not isotropic as the triangulation is not unique (degeneracy in the Delaunay triangulation, refer to Ripley (1981) page 41 for details). Other limitations of this method are that higher order derivatives are not continuous across the sides of the triangles, and that if the property changes drastically over a small distance, there tends to be some oscillations in the estimated values close to the vertices.

8.4.2 Universal kriging method

The maps obtained with the universal kriging interpolation method are shown in Figure 8.4. A higher degree of smoothing, compared to Akima’s method, was observed in Figure 8.4 with better defined contour bands showing less oscillation around the control points, especially for MFA. Universal kriging is also an exact interpolator which means that predicted values at the control points are the same as their known values. The air-dry density maps showed the increasing trend in DEN from pith to bark and the lack of variation with height. Similarly, the MFA maps showed high values below breast height near the pith and a transition to low values that depended on height level.

As mentioned before, universal kriging assumes that spatial variation in the *y* values has a drift or trend in addition to the spatial correlation between control points. Typically,
this method incorporates a first-order (plane surface) or second-order (quadratic surface) polynomial in the kriging process. In this work, the non-constant mean ($\mu$) was assumed to be a second-order polynomial on the coordinates ($x$) of the form:

$$\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$ (8.6)

The objective of Equation 8.6 is to remove the trend related to the spatial location so the residuals appropriately represent a zero mean second-order stationary random field. Equation 8.6 is the main difference between ordinary kriging, used for instance in the approach presented by Turner et al. (2001), and the universal kriging interpolation used in this work.

Next, a semivariogram that represents the spatial continuity of the data must be provided. The process of fitting an empirical semivariogram is difficult and often controversial but essential to estimate the semivariance at any given distance. Usually it requires the observations to be grouped by distance and direction (binning) to calculate the semivariances. In this work, a maximum of 30 bins and the Hawkins-Cressie robust estimator for calculating the semivariances were used (refer to Cressie (1991) page 75 for additional details). Because the spatial locations were measured in different units (centimetres in the radial direction and meters in the vertical direction) the empirical semivariograms were obtained for the radial and vertical directions separately. This allowed the incorporation of anisotropy into the analysis, a term used for describing the existence of directional differences in spatial dependence, in this case induced by the sampling scheme used.

Several mathematical functions to fit the empirical semivariograms exist. In this work, we adopted the more general of these functions, the Matérn class of covariance models given by the expression:

$$2\gamma(h) = c_0 + c_1 \left(1 - \frac{(h/2\alpha)^\nu}{2\Gamma(\nu)} K_\nu(h/\alpha)\right)$$ (8.7)

where $K_\nu(\cdot)$ is a modified Bessel function, $\nu$ is the parameter that controls the smoothness of the random field, $c_0$ is the nugget effect, $c_1$ is the scale parameters so that $c_0 + c_1$ represents
the sill, and $\alpha$ is the practical range. As an example, the empirical semivariograms for the MFA data and the models fitted to them for both the radial and vertical directions are shown in Figure 8.5. Once fitted, the directional models were combined to obtain a single isotropic model (refer to Wackernagel (2003)).

Figure 8.5 shows that the patterns of spatial variation in the radial direction were similar across groups with only small differences in the sill. On the contrary, the variation of MFA in the vertical direction was more erratic and the semivariances calculated at different lags showed higher variability. For DEN (not shown) the situation was similar for the radial direction and practically no correlation was observed among measurements at different heights explaining the observed vertical contour DEN bands. This lack of a strong spatial association in the vertical direction can be explained, excluding physiological reasons, by the large spacing between sampling heights (1.5 m). We may expect to observe some higher correlation between consecutive measurements if the sampling spacing is reduced.

The within-tree wood property maps obtained by universal kriging were appropriate and adequately reflected the expected trends in DEN and MFA. Although this method, compared to Akima’s interpolation, required a more detailed analysis, the nature of the spatial dependency among measurements was explicitly modelled and represented on the maps. It is also possible to expand the results by including maps of the standard errors associated with the interpolated values, something that it is not possible with Akima’s method. By knowing the steps needed to generate the maps the user can certainly make a better interpretation of them and evaluate their validity.

8.4.3 SEMIPARAMETRIC BIVARIATE SMOOTHING METHOD

The results from the semiparametric smoothing method are shown in Figure 8.6.

The maps showed an increase of DEN from pith to bark markedly faster at lower heights (below 2 to 3 m) than at upper heights. The DEN gradient ranged from approximately 450 kg/m$^3$ near the pith to 850 kg/m$^3$ in the outermost part of the trunk. The vertical contour
DEN bands were better defined with fewer oscillations than in the previous maps. Trees in group 3 showed a zone of high DEN similar to the other two groups, something that was not observed with the previous interpolation methods. The resulting whole-tree MFA maps showed less variation in relation to the preceding maps. Low MFA values were observed toward the periphery of the trunks, especially above 3 m with values ranging between 22 and 15 deg, and high MFA were observed in the bottom-part of the stems with values greater than 30 deg. The transition from high to low MFA appeared to be more rapid higher in the tree.

Splines for spatial interpolation are conceptually similar to splines for line generalization, except that in spatial interpolation they apply to surfaces rather than lines. Thin-plate splines create a surface that passes through the control points and has the least possible change in slope at all points. In other words, thin-plate splines fit the control points with a minimum curvature surface.

The whole-tree maps obtained by semiparametric regression are dependent on the choice of knots and degrees of freedom. In the full-rank case the knots correspond to the predictors, but in the low-rank case, such the models used in this work, a set of reduced knots $K < n$ needs to be chosen. For simplicity, we adopted the default settings and the bivariate knots were automatically chosen based on the FUNFITS function implemented in R (refer to Wand (2003) for details) as shown in Figure 8.2 for the trees in group 3. This figure shows that the knots are positioned in order to cover as much of the bivariate space as possible. It is possible to use a different set of knots based on some other criterion (e.g. selection from a regular grid) but there is no guarantee that the selection will be efficient and to some extent appropriate.

No changes were made to the degrees of freedom of the fit, defined by Ruppert et al. (2003) as the trace of the matrix $S_{n \times n}$ for which $\hat{y} = Sy$. In this work, the degrees of freedom ranged from 5.3 to 9.3 indicating that REML choose to fit a nonlinear surface on
the coordinates (a planar fit uses 3 degrees of freedom). As we may conclude from the maps in Figure 8.6, the nonlinearity was more important for MFA than for DEN.

The benefit of using semiparametric models for analysing spatial data is related to the inherent flexibility these models have in allowing the incorporation of geographic location and any other covariate in a linear or non-linear fashion, generalizing the form of the covariance structure and the option to be easily implemented in most available statistical software. For example, if the definition of the design matrix $Z$ given by Equations 4 and 5 is changed to some other proper covariance functions such as those from the Matérn class, the algorithm would be equivalent to the kriging interpolation used in the previous section. As a matter of fact, kriging is a subset of a family of radial smoothers that also includes smoothing splines.

8.5 Conclusions

Air-dry density and microfibril angle predicted by NIR calibration models were successfully used as control data points for spatial interpolation allowing the generation of wood property maps of high resolution and quality. The use of NIR data and spatial statistics proved to be a useful tool for describing the internal variation of $P. taeda$ wood properties within trees.

The whole-tree maps obtained by the three interpolation methods showed similar features indicating that air-dry density and microfibril angle were highly predictable as a function of their positions within the tree. However, the trees used in this work shared several characteristics in common such as age, site, and silvicultural treatments. Extension of these results to trees growing under different conditions may require the consideration of other covariates in addition to the spatial coordinates.

From the three interpolation methods analysed, Akima’s method is recommended for a rapid exploration of the within-tree variation of wood properties. This method does not require major assumptions and the code of the algorithm is freely available on the internet and, in the case of the R system, is already incorporated as part of the $\text{base}$ package.
To control how the spatial dependencies among the observations are modelled or to incorporate additional covariates into the analysis, universal kriging or semiparametric regression are the recommended options. Semiparametric regression is suggested to avoid the semivariogram fitting step of the universal kriging or when the new covariates are nonlinearly related with the variable of interest.

For any method of spatial analysis, the accuracy of interpolations is highly dependent on how precise the control points are measured. For this reason, the success of these techniques applied to other wood properties cannot be guaranteed, unless the method used to measure the properties be checked in advance.

8.6 References


Groom, L., Shaler, S., Mott, L., 2002b. Mechanical properties of individual southern pine fibers. part iii: Global relationships between fiber properties and fiber location within an individual tree. Wood Fiber Sci. 34 (2), 238–250.


Figure 8.1. Triangulation method used in the Akima’s interpolation for trees in group 3.
Figure 8.2. Position of the knots used in the semiparametric interpolation for trees in group 3.
**Figure 8.3.** Maps showing the within-tree variation of air-dry density (DEN) and microfibril angle (MFA) developed using Akima’s method. Filled circles are the control points. Tree groups are denoted by G1 to G3.
Figure 8.4. Maps showing the within-tree variation of air-dry density (DEN) and microfibril angle (MFA) developed using universal kriging. Filled circles are the control points. Tree groups are denoted by G1 to G3.
Figure 8.5. Empirical semivariogram models fitted to the microfibril angle (MFA) data in the radial (pith-bark) and vertical (bottom-top) directions. Tree groups are denoted by G1 to G3.
Figure 8.6. Maps showing the within-tree variation of air-dry density (DEN) and microfibril angle (MFA) developed using semiparametric regression. Filled circles are the control points. Tree groups are denoted by G1 to G3.
There is no simple definition for wood quality. Similarly, there is no simple way to measure wood quality. However, it is recognized that in any wood-based industry the availability and quality of the raw material should be evaluated in order to optimize its use. The extent to which these assessments can be effectively implemented depends largely on the cost-benefit ratio of the available techniques. The inherent variability of wood, within and between trees, sometimes prevent the industry from making these evaluations. Usually a level of compromise must be reached between the number of trees to evaluate and the type of samples and technique to use. In this dissertation, the use of three nondestructive techniques for the assessment of wood properties in forest plantations based on increment cores samples extracted at breast height (1.4 m), wood disks, wood chips, wood billets, and measurements collected in standing trees was studied.

The relatively recent incorporation of field-based acoustic tools in wood quality studies has positioned acoustics as the best option for assessing stiffness (MOE) in a large number of trees. Results from this research (Chapter 2), using the time-of-flight method (ToF) in standing *Pinus taeda* trees, showed that corrected velocities derived from ToF readouts can be successfully employed for the nondestructive evaluation of MOE and that the values compared well with those obtained from traditional static bending tests.

By using dynamic stiffness data measured with SilviScan, it was demonstrated that ToF reflects only the outerwood stiffness of the logs. In order to obtain the average MOE value, tree velocities must be adjusted for differences between dilatational waves (measured in trees)
and resonance waves (measured in logs) and for differences in wood moisture content between standing trees and the reference method utilized.

The correction methods imply that green density ($\rho$) and wood moisture content (MC) must be known. However, it was shown that individual tree values were not required and, therefore, $\rho$ and MC can be obtained from increment cores extracted from a subsample of the trees considered for assessment.

Along with the use of acoustics, there has been also considerable interest in utilizing NIR spectroscopy in wood quality studies, particularly for tree breeding purposes and for resource evaluation. The implementation of NIR as a rapid nondestructive tool for the study of wood allows the analysis of large numbers of samples that otherwise would be impractical, providing a wealth of information to tree growers, tree breeders, and manufacturers of forest products. However, the selection of training and validation samples, the identification of suitable pre-treatments for the NIR data, and the use of robust calibration models, among other aspects must be resolved.

It is economically unfeasible and inefficient to use all possible samples, for which NIR spectra have been collected, in the development of a regression model. It is necessary to include tools that facilitate the selection of samples for training (calibration) and prediction (validation) based on NIR spectra. The identification of suitable samples for calibration based on spectral characteristics prior to laboratory analysis would provide substantial savings in analytical costs.

Using pulp yield data measured in *Eucalyptus nitens* wood chip samples, it was shown in Chapter 3 that the use of a reduced number of training samples chosen by different sample selection algorithms resulted in NIR calibration models of similar performance compared with a model fitted using all available samples. Sample selection based on NIR spectra (CADEX, DUPLEX, SELECT) was as successful as selection based on measured pulp yields (RANKING).
Because wood chip samples involves the destructive sampling of trees, NIR spectroscopy has been increasingly used for the analysis of increment cores with spectra collected in consecutive sections from the radial longitudinal face of strips cut from the cores. As the properties are repeatedly measured within the strips, it is possible that serial correlation may affect the results obtained from multivariate modeling or prevent the use of the sample selection algorithms introduced previously.

In this research, the traditional partial least squares (PLS) algorithm for developing NIR calibration models was modified to allow the incorporation of data with a grouped structure, i.e. considering individual wood sections nested within radial strips (Chapter 4). The difference between the modified and traditional PLS methods was in the way used to estimate the loadings of the $Y$ vector (wood properties), which were obtained by generalized linear models instead of ordinary least squares (OLS). This allowed the specification of the individual variance-covariance matrices for each sample. The analysis of the normalized residuals obtained from models fitted using the modified approach showed that serial dependence was completely removed.

Wood radial strips are commonly measured in segments of certain width, e.g. 5 or 10 mm. These segments do not necessarily represent individual growth rings (which is a more natural scale on which to model the correlation) and therefore most of the expected autocorrelation was averaged out. Consequently, although serial correlation was successfully removed from the original PLS models, it was found that the analysis of increment cores by segments (in this case a width of 10-mm) did not represent a practical problem for the development of NIR calibration models for the prediction of wood properties.

Before developing a calibration model, NIR spectra are commonly transformed to reduce problems associated to the scattering of light. However, by pre-processing the NIR data it is possible that some significant information from the spectra is also removed making it difficult to evaluate model performance based on the error of cross-validation (RMSECV) and error of prediction (RMSEP) only.
A way to obtain better descriptions of the calibration models was explored in Chapter 5 by incorporating the concepts of net analyte signal (NAS) and multivariate analytical figures of merit. Statistics such as multivariate sensitivity, selectivity, and signal-to-noise ratio are part of the so-called figures of merit.

After fitting models with NIR data pre-treated by second derivative, multiplicative scatter correction, and orthogonal signal correction (three common pre-treatments techniques), using two data sets consisting of green wood samples obtained from *P. taeda* trees, it was found that the relative ranking of pre-processing algorithms was consistent for the two data sets when figures of merit were used, but they differed when RMSEP and $R^2$ were utilized. Therefore, the selection of the best pre-treatment(s) to apply must be based on multivariate figures of merit and not on the traditional fit statistics.

The interpretation of the figures of merit was not straightforward though. The values were dependent on the number of latent variables (factors) used to fit the PLS models, which are the base for the calculation of the NAS vector. For this reason, the determination of model dimensionality, another critical step in the development of NIR calibration models, was studied in Chapter 6.

The results from principal component analysis (PCA) and PLS (two of the most commonly used techniques for the study of wood properties by NIR) depend exclusively on their model dimensionality or rank. By using traditional segmented cross-validation (CV) and randomization tests (RT), it was found that PCA obtained using either approach resulted in models of similar dimensions. However, the CV criterion, when applied to PLS models for the prediction of microfibril angle (MFA), led to underfitting compared to the results from the RT. The calibration models and their predictive ability (as measured by RMSECV and RMSEP) were improved when additional statistically significant factors, as suggested by the RT, were included in the regressions.

Although the correct definition of the model dimensionality is critical for the development of NIR calibration models by PLS, there are some situations where there is insufficient basis
to assume linearity (one of the main assumptions behind PLS). In Chapter 7, two kernel methods were explored as alternative to PLS for the development of nonlinear calibration models for the prediction of different wood properties.

The two kernel methods evaluated, radial basis functions-partial least squares (RBF-PLS) and least-squares support vector machines (LS-SVM), gave substantial improvements over PLS models fitted using the CV criterion. Of the different wood properties considered, MFA and MOE showed the most significant improvements in terms of calibration statistics, indicating that nonlinear methodologies are a good alternative for the development of calibration models when the properties exhibit a nonlinear relationship between NIR-fitted and measured values.

The results from Chapter 7 showed that RBF-PLS and LS-SVM performed similar in terms of calibration statistics but the predictive ability of the LS-SVM calibrations were superior to both, RBF-PLS and PLS models. RBF-PLS performed better than PLS for MFA, MOE, and coarseness, all properties that exhibited nonlinearity. However, LS-SVM predictions were similar or better in cases were the properties were considered to be “linear”, such as air-dry density (AD) and tracheid wall thickness.

All the modeling “refinements” covered in this dissertation have one final goal: To provide rapid and reliable estimates of wood properties so they can be used for further analysis and/or decision making. Chapter 8 addresses this idea, where predicted NIR data along with interpolation methods from spatial statistics were utilized for illustrating the within-tree variation of AD and MFA in *P. taeda* trees.

The methodologies used for data interpolation were the Akima’s method, universal kriging, and semiparametric regression. It was found that AD and MFA predicted by NIR models can be successfully used as control data points for spatial interpolation, which allowed the generation of whole-tree wood property maps of high resolution and quality.

The whole-tree maps obtained by the three interpolation methods showed similar features indicating that AD and MFA were highly predictable as a function of their positions within
the tree. However the trees used for modeling shared similar growing conditions. Extrapolation of these results to trees growing under different environmental conditions or under different management regimes may require the consideration of additional covariates besides spatial coordinates.

In summary, there exist good alternatives for the rapid prediction of wood properties in standing trees. Because nondestructive sampling is often a priority, the use of standing tree acoustic tools for the segregation of trees in terms of stiffness or the use of increment cores whose wood properties are subsequently analyzed are probably the best alternatives. When a relatively small number of trees are to be analyzed and a detailed knowledge of the variation of the wood properties is required, then SilviScan is probably the most suitable option. For a number of trees in the low hundreds, NIR spectroscopy offers a more attractive approach from the point of view of cost. However as it was shown throughout this dissertation, the relative success of NIR spectroscopy relies on data of good quality to fit the models, on the appropriate selection of samples for the training and validation sets, on the appropriate treatment of the data, and on the use of the correct modeling techniques.