To the Graduate Council:

I am submitting herewith a thesis written by Dillon Matthew Carty entitled "An Analysis of Boosted Regression Trees to Predict the Strength Properties of Wood Composites." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Statistics.

Timothy M. Young, Major Professor

We have read this thesis
and recommend its acceptance:

Frank M. Guess

Russell L. Zaretzki

Accepted for the Council:

Carolyn R. Hodges
Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

An Analysis of Boosted Regression Trees to Predict the Strength Properties of Wood Composites

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Dillon Matthew Carty
August 2011

Copyright © 2011 by Dillon Matthew Carty
All rights reserved.

DEDICATION

This thesis would never have been possible if it were not for the love and support of my family. They have provided an incredible amount of support through this entire process. I know that it has been just as much of an adventure for them as it has been for me. When I was weary during these two years, they kept me focused with their words and humor. To Mom, Dad, and Devon, I cannot thank you enough. I love each one of you very much.

To my fellow graduate students and friends, thank you for helping me maintain my focus during these two years. Also, thank you for convincing me to take a break from work and making me laugh while always showing me a good time, even when I felt that I did not have the time to step away from my research work. To my fellow graduate students and friends Evan, Paul, Bahar, Maria, and Yan, I want to thank each of you for making me a better student of Statistics. Together we built a support group and made our two years of graduate school a fun two years that each of us will remember forever. To my closest friends, words cannot express how grateful I am that each one of you is a part of my life. I can honestly say that I have the best friends that any person can have. There is always a smile on my face when I am around each one of you, or even better all of you at once.

To each of you, especially my family, I dedicate this thesis. Each one of you has played a valuable role in my life and has shaped me into the person that I am today.

ACKNOWLEDGEMENTS

I gratefully acknowledge both the University of Tennessee Agricultural Experiment Station McIntire-Stennis Grant and the USDA Special Grant program for Wood Utilization Research for financially supporting my graduate research assistantship. I especially thank Professor Dr. Timothy M. Young for his diligent efforts in bringing in funding that has supported my graduate research assistant position. For the administering of funding, I would also like to thank Dr. Keith Belli, Professor and Head of the Department of Forestry, Wildlife and Fisheries, and Dr. Timothy G. Rials, Professor and Director of the Center for Renewable Carbon.

I express much appreciation to the Committee Co-chairs and Professors Dr. Timothy M. Young, Dr. Frank M. Guess, and Dr. Russell L. Zaretzki, for their overwhelming and endless amounts of support and guidance throughout the two years spent on this research project. It has been a tremendous pleasure and an incredible two years working with these great individuals and mentors. Each of these individuals have provided valuable input and intuition on the techniques discussed in this thesis. I cannot express in words how grateful I am of the help these individuals provided me with during the process of this research. None of the research documented in this thesis could have been completed without the overwhelming help, support, and insight of these three tremendous individuals. It truly has been a privilege to be a part of this excellent research team.

I also want to thank three impressive colleagues. Dr. Nicolas André provided the real-time, temporal process data set from the U.S. particleboard manufacturer as well

as valuable information on the manufacturing process and the data set itself. Mr. Yan Zeng contributed parts of his own research on imputed (missing) data that helped make the analysis provided in this thesis possible. Ms. Xia Huang provided insight and support throughout this process.

Finally, I thank my professors and fellow graduate students (past and present) in the Department of Statistics, Operations, and Management Science. I thank my professors for challenging me to become a better student of Statistics and providing me with knowledge that will be forever useful in my future endeavors. I thank my fellow graduate students (past and present) for also challenging me to become a better student of Statistics and providing the support needed that only other graduate students can supply.

ABSTRACT

The forest products industry is a significant contributor to the U.S. economy contributing six percent of the total U.S. manufacturing gross domestic product (GDP), placing it on par with the U.S. automotive and plastics industries. Sustaining business competitiveness by reducing costs and maintaining product quality will be essential in the long term for this industry. Improved production efficiency and business competitiveness is the primary rationale for this work. A challenge facing this industry is to develop better knowledge of the complex nature of process variables and their relationship with final product quality attributes. Quantifying better the relationships between process variables (e.g., press temperature) and final product quality attributes plus predicting the strength properties of final products are the goals of this study. Destructive lab tests are taken at one to two hour intervals to estimate internal bond (IB) tensile strength and modulus of rupture (MOR) strength properties. Significant amounts of production occur between destructive test samples.

In the absence of a real-time model that predicts strength properties, operators may run higher than necessary feedstock input targets (e.g., weight, resin, etc.). Improved prediction of strength properties using boosted regression tree (BRT) models may reduce the costs associated with rework (i.e., remanufactured panels due to poor strength properties), reduce feedstocks costs (e.g., resin and wood), reduce energy usage, and improve wood utilization from the valuable forest resource.

Real-time, temporal process data sets were obtained from a U.S. particleboard manufacturer. In this thesis, BRT models were developed to predict the continuous

response variables MOR and IB from a pool of possible continuous predictor variables. BRT model comparisons were done using the root mean squared error for prediction (RMSEP) and the RMSEP relative to the mean of the response variable as a percent (RMSEP%) for the validation data set(s). Overall, for MOR, RMSEP values ranged from 0.99 to 1.443 MPa, and RMSEP% values ranged from 7.9% to 11.6%. Overall, for IB, RMSEP values ranged from 0.074 to 0.108 MPa, and RMSEP% values ranged from 12.7% to 18.6%.

TABLE OF CONTENTS

CHAPTER I. Introduction 1

CHAPTER II. Literature Review 7
- Wood Composites 7
- Data Mining 9
- Decision Trees 11
- Boosting 13
- Boosted Decision Trees 15
- Predictive Modeling of Engineered Wood Products 17

CHAPTER III. Methods 19
- Regression Trees 19
- Boosted Regression Trees 20
- Stochastic Gradient Boosting 22
- Software and Parameters 26
- Imputation of the Data Set and BRT Models 27
- Model Comparison 29

CHAPTER IV. Predicting the Strength Properties of Wood Composites Using Boosted Regression Trees 31
- Data Set 31
- BRT and CART Regression Tree Models for MOR 32
- BRT and CART Regression Tree Models for IB 42
- Remarks 50

CHAPTER V. A Comparison of Several Imputation Methods Using Boosted Regression Trees to Predict Strength Properties of Wood Composites 53
- Data Sets 53
- Imputation Method Results Using BRT Models for MOR 55
- Imputation Method Results Using BRT Models for IB 61
- Remarks 64

CHAPTER VI. Concluding Remarks and Future Research 68

List of References 73

Appendix 80

Vita 91

LIST OF TABLES

Table 1: Statistics obtained from analysis performed on the validation data set for 70 different BRT models predicting MOR, with a value of three being used for the parameter *mnn*.* **33**

Table 2: Statistics obtained from analysis performed on the validation data set for 70 different BRT models predicting MOR, with a value of five being used for the parameter *mnn*.* **36**

Table 3: Statistics obtained from analysis performed on the validation data set for 70 different BRT models predicting IB, with a value of three being used for the parameter *mnn*.* **43**

Table 4: Statistics obtained from analysis performed on the validation data set for 70 different BRT models predicting IB, with a value of five being used for the parameter *mnn*.* **45**

Table 5: RMSEP (and RMSEP%) values obtained on the validation data sets for each of the four different imputation methods are shown. Each cell contains three RMSEP (and RMSEP%) values obtained on three different validation data sets for BRT models predicting MOR and for the given BRT parameter settings. A value of three for the parameter *mnn* was used.* **56**

Table 6: RMSEP (and RMSEP%) values obtained on the validation data sets for each of the four different imputation methods are shown. Each cell contains three RMSEP (and RMSEP%) values obtained on three different validation data sets for BRT models predicting MOR and for the given BRT parameter settings. A value of five for the parameter *mnn* was used.* **58**

Table 7: Descriptive statistics on the RMSEP (and RMSEP%) values obtained on the validation data sets for the 288 BRT models (i.e., 144 BRT models when the *mnn* parameter was equal to three and 144 BRT models when the *mnn* parameter was equal to five) predicting MOR. **59**

Table 8: RMSEP (and RMSEP%) values obtained on the validation data sets for each of the four different imputation methods are shown. Each cell contains three RMSEP (and RMSEP%) values obtained on three different validation data sets for BRT models predicting IB and for the given BRT parameter settings. A value of three for the parameter *mnn* was used.* **62**

Table 9: RMSEP (and RMSEP%) values obtained on the validation data sets for each of the four different imputation methods are shown. Each cell contains three RMSEP (and RMSEP%) values obtained on three different validation data sets for BRT models predicting IB and for the given BRT parameter settings. A value of five for the parameter *mnn* was used.* **63**

Table 10: Descriptive statistics on the RMSEP (and RMSEP%) values obtained on the validation data sets for the 288 BRT models (i.e., 144 BRT models when the *mnn*

parameter was equal to three and 144 BRT models when the *mnn* parameter was equal to five) predicting IB. **65**

LIST OF FIGURES

Figure 1: The relationship between learning rate (*lr*) and MOR RMSEP% for the 70 different BRT models with seven values chosen for the number of additive terms (*nat*) and a value of three chosen for the maximum number of nodes (*mnn*). **34**

Figure 2: The relationship between learning rate (*lr*) and MOR RMSEP% for the 70 different BRT models with seven values chosen for the number of additive terms (*nat*) and a value of five chosen for the maximum number of nodes (*mnn*). **37**

Figure 3: Scatterplot of the observed values versus the predicted values of the validation data set for the BRT model that best predicts MOR. **38**

Figure 4: Regression tree model for MOR. **40**

Figure 5: Scatterplot of the observed values versus the predicted values of the validation data set for the regression tree model that predicts MOR. **41**

Figure 6: The relationship between learning rate (*lr*) and IB RMSEP% for the 70 different BRT models with seven values chosen for the number of additive terms (*nat*) and a value of three chosen for the maximum number of nodes (*mnn*). **44**

Figure 7: The relationship between learning rate (*lr*) and IB RMSEP% for the 70 different BRT models with seven values chosen for the number of additive terms (*nat*) and a value of five chosen for the maximum number of nodes (*mnn*). **46**

Figure 8: Scatterplot of the observed values versus the predicted values of the validation data set for the BRT model that best predicts IB. **48**

Figure 9: Regression tree model for IB. **49**

Figure 10: Scatterplot of the observed values versus the predicted values of the validation data set for the regression tree model that predicts IB. **51**

Figure 11: For each imputation method, a scatterplot of the observed MOR values and the predicted MOR values of the validation data set for the BRT model that best predicts MOR. **60**

Figure 12: For each imputation method, a scatterplot of the observed IB values and the predicted IB values of the validation data set for the BRT model that best predicts IB. **66**

CHAPTER I. INTRODUCTION

Wood composite products (e.g., particleboard) is a name given to a material that is manufactured by binding the strands, fibers, particles, or veneers of wood, together with adhesives, to form composite materials. These products are engineered to precise design specifications, which are tested to meet national and international standards. Wood composite products are growing in popularity with the forest products industry because design can be done to meet application-specific performance requirements, products have improved wood recovery (i.e., can be made from waste wood or other non-merchantable wood with defects), increased product reliability, etc. These products are growing in popularity with consumers because they provide the consumer with the natural warmth and beauty of wood, versatility, and availability in a wide variety of thicknesses, sizes, and grades, etc. Wood composite products are used in a variety of applications, such as home construction, commercial buildings, and industrial products.

The forest products industry is an important contributor to the U.S. economy. The U.S. forest products industry accounts for approximately six percent of the total U.S. manufacturing gross domestic product (GDP), placing it on par with the U.S. automotive and plastics industries (American Forest and Paper Association 2010). The industry generates more than $200 billion a year in sales and employs approximately 900,000 people earning $50 billion in annual payroll (American Forest and Paper Association 2010). The industry is among the top 10 manufacturing employers in 42 states (American Forest and Paper Association 2010).

Sustaining business competitiveness by reducing costs and maintaining product quality will be essential in the long-term for this industry. One of the challenges facing this industry is to develop better knowledge of the complex nature of process variables and their relationship with final product quality attributes. Quantifying the relationships between process variables (e.g., line speed, press temperature, etc.) and final product quality attributes (e.g., internal bond tensile strength, modulus of rupture, etc.) is the goal of this thesis. The final quality attributes of interest in this study are internal bond (IB) tensile strength and modulus of rupture (MOR) (i.e., flexural strength). IB tensile strength is measured as the tensile strength perpendicular to the surface. Importantly, increasing the amount of smaller particles (i.e., density) and/or resin used to create a board increases IB tensile strength (Gamage 2007). MOR strength is defined as the ultimate bending strength of the board, which is generally determined after a static bending test (Gamage 2007). MOR strength is an extremely important property that "controls the usability" of a board as a building element (Gamage 2007). In this thesis, both IB and MOR are measured in megapascals (MPa).

Of at least as great an importance as the number of process variables and the often complex inter-relationships as related to final product attributes, is the delay between the time at which a destructive test sample is taken at the output end of the production line and the time at which the strength characteristics of this sample (e.g., IB, MOR, etc.) have been determined in a testing laboratory. This delay can be one to three hours in particleboard. In the absence of a real-time model that predicts mechanical properties, it is difficult to optimize production and correct for possible poor

mechanical properties of the final manufactured product. Improved prediction using boosted regression tree (BRT) models can directly influence minimizing the risk of producing hours of defective or off-grade product, or hours of production that is unnecessarily over-engineered and of higher cost. Prediction using BRT models can possibly reduce the costs associated with rework (i.e., remanufactured panels due to poor strength properties), reduce feedstocks costs (e.g., resin and wood), reduce energy usage, and improve wood utilization from the valuable forest resource. Improved production efficiency and business competitiveness are essential for the wood composites industry and are the primary rationale for this work.

The remaining thesis is organized into the following chapters. Chapter II of this thesis is a literature review focusing on the main topics of this thesis. The literature review begins with a brief history and description of wood composites. A specific subgroup of wood composites known as “engineered wood products” is discussed. Since the data sets analyzed in this thesis are from a U.S. particleboard manufacturer, some discussion and background information on particleboard production, which is an engineered wood product, is provided. Next, the research literature related to the topic of data mining is reviewed. The development of data mining from “knowledge discovery in databases” (KDD) is summarized. The large number of the methodologies that are used in data mining and the evolution of the machine learning research community as part of the statistical research community are encapsulated. An emphasis is given on data mining practices as related to large data sets. Several data mining tasks are mentioned, but more detail is presented on predictive modeling is the only task given its

relevance to the research of this thesis. “Decision trees,” which is a machine learning and/or data mining predictive modeling technique, is reviewed. A synopsis of decision tree methods and advantages of decision trees for predictive modeling are given. “Boosting” is elaborated on and the genesis of boosting is discussed, e.g., the ensemble method of boosting allows for the combination of “base classifiers” (also known as weak learners) to produce a committee whose performance can be significantly better than that of any of the base classifiers (Freund 1995). The history of the AdaBoost algorithm (Freund and Schapire 1997), one of the first and most popular boosting algorithms, and various research applications of the AdaBoost algorithm are mentioned. Next, the idea of boosted decision trees, which combines the ideas of boosting and decision trees, as well as various studies using boosted decision trees for predictive modeling are discussed. Finally, the relevance of real-time predictive modeling (using boosted decision trees) for engineered wood product strength properties is mentioned. Also, numerous studies on real-time predictive modeling of final product quality characteristics of wood composites using statistical methods are cited.

In Chapter III, a summary of the statistical and data mining methods used to predict MOR and IB is given. More information is provided on regression trees for prediction and the CART algorithm (Breiman et al. 1984) in particular. The general construction of a regression tree is mentioned. More importantly, since the CART algorithm for regression trees is used in Chapter IV of this thesis, the process of this algorithm is discussed in detail. Next, the topic of BRT is also introduced. Given that regression trees generally having poor predictive power, the boosting technique is used

to enhance the predictive performance of regression trees. The combination of these two techniques (i.e., boosting and regression trees) and the design for producing an improved predictive a model is introduced. After mentioning the topic of BRT, the Stochastic Gradient Boosting algorithm (Friedman 2002) is explored. This is the algorithm that is used to build the hundreds of BRT models discussed throughout this thesis. The software (*STATISTICA*) used to perform all of the statistical analysis for this study is presented. Information on the various parameters used to control the BRT algorithm of *STATISTICA* is discussed in detail. The basic of idea of missing value imputation and the imputation methods used in this thesis are briefly mentioned in this chapter. Further discussion on these techniques is provided in Chapter V. The statistical methods used in Chapter IV are slightly different than the statistical methods used in Chapter V, and these differences are highlighted.

Chapter IV summarizes a manuscript that was developed as part of the thesis which is an analysis of BRT and CART regression tree models for predicting MOR and IB strength metrics. The particleboard data set obtained from a U.S. particleboard manufacturer is discussed in detail.

Chapter V summarizes an analysis that compares several different missing value imputation methods using BRT models for predicting MOR and IB. Three different imputation methods (i.e., median, expectation-maximization, and last observation carried forward) and one non-imputation method are compared. Four different data sets corresponding to each of these four methods are discussed. These data sets are similar to the data set used in Chapter IV in that they were provided by the same U.S.

particleboard manufacturer, but have been imputed due to missing values. An analysis comparing the different imputation methods using BRT for predicting MOR is done, and then the same is done for IB. Some general remarks regarding the analysis are provided at the end of the chapter.

In Chapter VI, a summary of the overall research and thesis findings is given. A discussion of future research possibilities is given. Expansion of the thesis work to predictive modeling for other manufacturing systems other than word composites and particleboard is given.

CHAPTER II. LITERATURE REVIEW

This chapter provides a brief description of wood composites and particleboard in particular plus succinct basic information on data mining, decision trees, boosting, boosted decision trees, and predictive modeling of engineered wood products. It is assumed that the reader has an elementary knowledge of wood composites but some prior knowledge of statistical methodology and various applications.

Wood Composites

"Wood composites" is a term that refers to many different materials that have been developed by using small wood particles together with resins/glues or other elements to create larger materials. Most of today's wood composites "have an origin within the past [60] years" (Maloney 1996). Many wood composites today use what was once considered waste wood residues (e.g., wood particles, wood chips, and waste fibers) from infrequently used tree species or noncommercial species, which may also include agricultural residues (LeVan-Green and Livingston 2001; Maloney 1996). According to LeVan-Green and Livingston (2001), the fact that wood composites are assembled from small pieces of wood, or agricultural residues, provides a technology that is easily adaptable to a changing resource base.

A certain assemblage of composites is grouped together and referred to as engineered wood products, for example, plywood, oriented strand board, and glued laminated timber. Some composites such as particleboard, medium density fiberboard (MDF), and oriented strand board (OSB) have been described as composite panel products (LeVan-Green and Livingston 2001; Maloney 1996). Most wood composites

can be engineered to assorted specifications, thus taking advantage of wood's inherent properties and at the same time improving these properties by using materials science and technology (Maloney 1996). Many wood composites can be used structurally (e.g., roofing, floors, and structural panels in buildings). Importantly, these technologies make it possible to use small-diameter and underutilized material (LeVan-Green and Livingston 2001). Small-diameter and underutilized material refers to timber that is left out in the woods because it is not economical to remove, or local capacity to process it does not exist (LeVan-Green and Livingston 2001).

Particleboard is a term for a multi-layer panel that is manufactured from lignocellulosic materials (generally custom-made softwood that is mostly in the form of discrete pieces or particles) combined with a blended resin or other binder and then bonded together under heat and pressure in a hot press (Gamage 2007; Maloney 1996).[1] Particleboard includes different panel types called flakeboard and chipboard, where the size and shape of the wood particles used to make the board are varied (Wood Handbook 1999). According to Maloney (1996), particleboards are defined by the "method of pressing" that is used in the manufacturing process. When pressure is applied in the direction perpendicular to the faces, they are defined as flat-platen pressed; and when pressure is applied parallel to the faces, they are defined as extruded (Maloney 1996).

The first platen-pressed particleboard plant started operation in Dubuque, Iowa, in 1933, but this plant was only operational until 1942, while a larger commercial plant started operation in Germany in 1941 (Maloney 1996). Extruded particleboard was

[1] Some examples of softwood are Douglas fir, southern pines, etc. and some low-value wood sources such as aspen or cottonwood.

developed in Germany between 1947 and 1949 (Maloney 1996). Numerous plants manufacturing this type of particleboard were built throughout the world and the U.S., but low production capacities and some board physical property limitations kept extruded particleboard from becoming a major product line (Maloney 1996).

The great increase in particleboard production started in the 1950s (Maloney 1996). To this day, particleboard remains one of the world's dominant furniture panels, but significant amounts of its production also go into structural applications such as manufactured home floors, roof sheathing, wall panels, and stair treads (Gamage 2007; Maloney 1996).

Data Mining

Hand et al. (2001) noted that the science of mining information from large data sets or databases is known as data mining. Data mining is a huge subject area and a large amount of literature exists on the topic. The discussion of data mining in this literature review is not meant to be extensive, but it is intended to be an antecedent to the methods discussed later in this thesis.

Hand et al. (2001) stated, "Data mining is often set in the broader context of knowledge discovery in databases, or KDD," which originated in the artificial intelligence research field. According to Giudici (2003), "the term 'data mining' was used to describe the component of the KDD process where the learning algorithms were applied to the data." The KDD process involves several stages: "selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to

extract patterns and relationships, and then interpreting and assessing the discovered structures" (Hand et al. 2001).

Many of the methodologies that are used in data mining come from two branches of research, one developed in the machine learning community and the other developed in the statistical community, particularly in multivariate and computational statistics (Giudici 2003). Hence, a mastery of data mining requires both an understanding of statistical and computational issues. Importantly, the methodologies in both of these research communities are essential when dealing with large data sets. Statistically, data mining can be viewed as computer automated exploratory data analysis of large data sets (Friedman 2001). Friedman (2001) notes, "Statistics can potentially have a major influence on Data Mining."

Data mining can be fittingly categorized into types of tasks, which correspond to the distinctive objectives of the individual that is studying the data (Hand et al. 2001). According to Hand et al. (2001), some of these tasks are exploratory data analysis, descriptive modeling, discovering patterns and rules, predictive modeling, etc. Predictive modeling is the task of interest in this thesis and will be briefly discussed here.

The aim in predictive modeling is to construct a model that will permit one value of a variable to be predicted (estimated) from the known values of other variables (Friedman 2001). In classification, the variable being predicted is categorical; whereas, in regression, the variable being predicted is quantitative (Hand et al. 2001). There have a been a large number of methods that have been developed in the fields of

machine learning and statistics that tackle predictive modeling, and work in this area has led to "significant theoretical advances and improved understanding of deep issues of inference" (Hand et al. 2001).

Decision Trees

A machine learning and/or data mining technique that uses a decision tree as a predictive model, which maps observations about a certain item to conclusions about a certain item's target value, is known as "decision tree learning." Decision tree learning is also known as "decision trees."[2] Decision tree methodology has roots in both the statistics and computer science literature as cited in the following references and comments. A precursor to current tree methodology was "Automatic Interaction Detector" (AID) developed by Morgan and Sonquist (1963), Kass (1975), and Fielding (1977). Breiman et al. (1984) were the first to introduce the main ideas of tree methodology to statistics. Quinlan (1993) provided an overview of how tree methodology was developed in machine learning. Hastie et al. (2009) described decision trees from a statistical perspective.

As Young (2007) noted, "Decision trees are one of the most popular predictive learning methods used in data mining." A single decision tree model can be represented by a two-dimensional graphic, which can be plotted and easily interpreted, no matter how high the dimensionality of the predictor space or how many variables are used for prediction (Friedman 2001; Hastie et al. 2009). This ease of interpretation from two-dimensional plots makes decision trees a powerful tool for the practitioner (Young 2007).

[2] http://en.wikipedia.org/wiki/Decision_tree_learning referenced on 06/01/10

In the simplest of forms, tree-based methods work by partitioning the predictor space into rectangular regions, using a series of rules to identify regions having the most homogeneous responses to predictors, and then assigning a simple model (e.g., a constant, regression model, etc.) to each region (Bishop 2006; Elith et al. 2008; Loh 2008). The growing of a tree involves recursive binary splits implying a binary split is repeatedly applied to its own output until some stopping specification is obtained. Decision trees are insensitive to outliers and can accommodate missing data in predictor variables by using surrogates (Breiman et al. 1984). As well, decision trees have the advantage of seldom selecting irrelevant predictor variables (i.e., the recursive tree building algorithm estimates the optimal variable on which to split at each step implying predictors not related to the response variable(s) tend not to be selected for splitting) (Breiman et al. 1984; Elith et al. 2008; Young 2007). Also, Elith et al. (2008) noted, "The hierarchical structure of a tree means that the response to one input variable depends on values of inputs higher in the tree, so interactions between predictors are automatically modeled." Decision trees are liked for these aforementioned reasons. In all, decision trees are "conceptually simple yet powerful" (Hastie et al. 2009).

A decision tree for numerical data is known as a "regression tree," and a decision tree for categorical data is known as a "classification tree." These two types of decision trees will not be discussed any further here. Since this thesis uses numerical data from industrial processes, the analyses performed throughout this thesis are based around

regression trees and BRT. Refer to Chapter III for more discussion on regression trees and BRT, in particular.

Boosting

Boosting has its roots in the theoretical framework for studying machine learning called the “Probably Approximately Correct” (PAC) learning model. The PAC learning model is due to Valiant (1984). Kearns and Vazirani (1994) provided a nice introduction into the PAC learning model. Kearns and Valiant (1994) were the first to pose the question of whether a weak learning algorithm that is moderately better than random guessing in the PAC model can be “boosted” into an arbitrarily accurate “strong” learning algorithm. Schapire (1990) derived the first provable polynomial-time boosting algorithm in 1989. One year later, Freund (1995) developed a more efficient boosting algorithm. Drucker et al. (1993) carried out the first experiments with these early boosting algorithms.

Boosting is a powerful technique for combining several “base classifiers” (also known as weak learners) to produce a committee whose performance can be significantly better than that of any of the base classifiers (Freund 1995). Sutton (2005) provided an example of weak learners as being a simple classifier such as a two-node decision tree (also known as a stump). Boosting can yield good results even if the weak learners have performance that is only slightly better than random guessing. Since random guessing has an error rate equal to 0.5, a weak learner just has to predict correctly a little more than 50% of the time (Sutton 2005). Boosting was originally designed for classification problems, but it can be “profitably be extended to regression”

as well (Hastie et al. 2009). According to Bishop (2006), the main difference between boosting and the committee methods such as bagging is that the weak learners are trained in sequence. Each weak learner is trained using a weighted form of the data set in which the weighting coefficient associated with each data point is conditional upon the performance of the previous classifiers. Points that are misclassified by one of the base classifiers are given greater weight when used to train the next classifier in the sequence.

Freund and Schapire (1997) commented that they introduced, in 1995, one of the first and most popular boosting algorithms named AdaBoost, which is short for "adaptive boosting." AdaBoost solved many of the practical difficulties of boosting algorithms that came before (Freund and Schapire 1999). Much like the original boosting algorithms, AdaBoost was developed for two-class classification problems. Importantly, the most basic theoretical property of AdaBoost concerns its ability to reduce the training error, i.e., the fraction of mistakes on the training set (Schapire 2003). One can refer to Freund and Schapire (1999) for the precise form of their AdaBoost algorithm.

According to Freund and Schapire (1999), the AdaBoost algorithm has been tested empirically by many researchers. As an example, Freund and Schapire (1996) tested AdaBoost on a set of UCI benchmark datasets[3] using an algorithm that finds the best "decision stump" or single-best decision tree. Also, Schapire and Singer (2000) used the idea of boosting for text categorization tasks. In this set of experiments, base classifiers were used that test for a word or phrase being present or absent. Schapire (2003) noted that boosting has also been applied to problems in fields dealing with text

[3] http://archive.ics.uci.edu/ml/ referenced on 06/20/2011

filtering (Schapire et al.1998) and document routing (Iyer et al. 2000), "ranking" problems (Freund et al. 2004), medical diagnosis (Merler et al. 2001), and many more problems in various fields. As well, Schapire and Singer (2000) used a generalization of AdaBoost that provides an interpretation of boosting as a gradient-descent method. For additional information on boosting and other "learning methods," refer to Valiant (1984), Schapire (1990), Drucker et al. (1993), Kearns and Vazirani (1994), Kearns and Valiant (1994), Freund (1995), and a lecture given by Dr. Rich Caruana.[4]

Boosted Decision Trees

Boosted decision trees are a relatively new technique that has within the past decade burst onto the scene of predictive modeling. In this chapter, the discussion of boosted decision trees and BRT in particular will be brief but will mention early discussions about the idea of boosted decision trees and studies in which boosted decision trees were used for prediction. A more in depth discussion of the methodology behind boosted decision trees will occur in Chapter III. Importantly, the boosting technique helps to improve the predictive performance of decision trees.

Freund and Schapire (1997) provided a suggestion as to how boosting might produce regression model using their algorithm "AdaBoost.R." Breiman (1998) discussed applying boosting to CART to create a classifier (see Olshen 2001). Breiman (1999) suggested how boosting might apply to regression problems using his algorithm "arc-gv." Sutton (2005) mentioned the concept of a "weak learner" as being a simple classifier such as a two-node decision tree (i.e., tree stump).

[4] http://videolectures.net/solomon_caruana_wslmw/ referenced on 12/15/2010

There have been studies published in various fields that used boosted decision trees for prediction, and some of these studies are discussed below. Drucker (1997) was the first to actually implement and experiment with boosting regression models. Drucker (1997) used regression trees as the fundamental building blocks in “boosting committee machines.” Drucker (1997) applied an ad hoc modification of “AdaBoost.R” (Freund and Schapire 1997) to regression problems and obtained promising results. Ridgeway et al. (1999) brought together ideas from boosting, naïve Bayes learning, and additive modeling, to create a “BNB.R” algorithm that fit a boosted naïve Bayes regression model, which they compared the performance of their “BNB.R” algorithm to three other interpretable multivariate regression procedures. Zhou et al. (2005) assessed the performance of boosted decision trees on publicly available email data in filtering out unsolicited bulk emails, while comparing this method to various other methods. Deconinck et al. (2007) evaluated BRT for the modeling and predicting blood-brain barrier passage of drugs. De’ath (2007) proposed a form of boosted trees, “aggregated boosted trees,” and through a simulation study on regression data showed that this form of boosted trees reduced prediction error relative to other forms of boosted trees. Robinson (2008) assessed the ability of “regression tree boosting” to risk-adjust health care cost predictions, and he used diagnostics groups and demographic variables as inputs. Robinson (2008) used BRT because it is a method that systematically searches the data for consequential interactions, which it automatically incorporates into a risk-adjustment model. Elith et al. (2008) demonstrated the practicalities and advantages of using BRT through a distributional

analysis of the short-finned eel, a native freshwater fish of New Zealand. Abeare (2009) evaluated the performance of BRT as a potential tool for catch-rate standardization of Yellow Fin tuna. Carslaw and Taylor (2009) used BRT to analyze air pollution data at a mixed-source location. They used BRT to draw inferences concerning the source characteristics at a location of high source complexity. Li et al. (2010) used BRT to identify modern processor configurations between a key processor structure's "architectural vulnerability factor" and various performance metrics. These aforementioned studies are just a few studies that used boosted decision trees for prediction. As evident from the previous citations, boosted decision tree modeling for prediction is a contemporary statistical-based data mining topic applied to many fields, most notably medicine and ecology.

Predictive Modeling of Engineered Wood Products

A major challenge for engineered wood products manufacturers is developing better knowledge of the complex nature of process variables and their relationship with engineered wood product strength properties (e.g., IB, MOR, etc.). Some key process variables might be line speed, press temperature, wood chip dimensions, etc. At the time of production, the quality of engineered wood is not known, i.e., strength properties of the samples are determined at a later time in a laboratory through destructive testing. The time span between destructive tests can be two to six hours depending on the type of product (Young 2007). The delay can be two to three hours in particleboard. Hours of producing defective or off-grade product, or hours of production that is unnecessarily over-engineered and of higher cost could take place during the hours between these

destructive tests. Improved production efficiency and business competitiveness are essential for the wood composites industry.

Real-time predictive modeling of strength properties can reduce costs (e.g., rework costs, feedstocks costs, etc.), reduce energy usage, and improve wood utilization. With today's economy, the reduction of costs while continuing to manufacture the same (or better) quality product is crucial. Numerous studies on real-time predictive modeling of final product quality characteristics of wood composites using statistical methods have been published (Young 1996; Cook and Chiu 1997; Bernardy and Scherff 1998; Greubel 1999; Erilsson et al. 2000; Cook et al. 2000; Young and Guess 2002; Young et al. 2004; Lei et al. 2005; Xing et al. 2007; André et al. 2008; Clapp et al. 2008; Young et al. 2008; Mora and Schimleck 2010). Greubel (1999) showed how the use of "off-line" first-order statistical models led to medium density fiberboard manufacturing cost savings of five to ten percent. Erilsson et al. (2000) discussed the potential for statistical models in engineered wood manufacturing. André et al. (2008) presented new data mining-based multivariate calibration models for predicting IB strength from medium density fiberboard process variables. Clapp et al. (2008) used a modified principal components analysis to develop an empirical model to predict the IB of medium density fiberboard based on a selected subset of process variables. It is evident from the literature the engineered wood products manufacturing industry could benefit greatly from improved real-time predictive modeling using BRT.

CHAPTER III. METHODS

This chapter provides, in detail, the statistical and data mining methods that were used as analysis tools for this research. There will be more discussion than in the previous chapter on regression trees (i.e., decision trees for numerical data) and specifically the CART algorithm (Breiman et al. 1984). BRT methodology, as well as the Stochastic Gradient Boosting algorithm (Friedman 2002) used by *STATISTICA* 10 to build BRT models, will be presented and discussed. A brief description/background of the software package *STATISTICA* will be provided followed by some discussion of the relevant parameters used to control the stochastic gradient boosting algorithm in *STATISTICA* 10. The topic of imputation will be briefly mentioned but needs to be presented. Statistical techniques used to compare BRT models will be explained.

Regression Trees

A decision tree for numerical/continuous data is known as a "regression tree." A regression tree is a piece-wise linear estimate of a regression function, which is constructed by the recursive partitioning of the data and the sample space (Loh 2002). The simplest form of regression tree fits the mean response for the observations in a partitioned region (Elith et al. 2008). If skewness exists in the data, Loh (2002) suggested using the sample median as the constant. Young (2007) noted that the construction of a regression tree generally consists of the following four steps performed iteratively: (1) partition the data, (2) fit a model to the data after each partition, (3) stop when the residuals of the model are approximately zero, or there are only a few observations left, and (4) pruning the tree (if the tree overfits). Not all regression tree

algorithms agree on the first and the second step above, and some software packages possessing regression tree algorithms lack the capability to automatically perform step four.

The AID ("Automatic Interaction Detector") algorithm by Morgan and Sunquist (1963), Kass (1975), and Fielding (1977) was the first implementation of the decision tree method. The CART ("Classification and Regression Trees") algorithm by Breiman et al. (1984) followed the AID algorithm and is one of the most popular decision tree algorithms. The CART algorithm uses a backward-elimination strategy to develop the decision tree (Loh 2002). The algorithm works by growing an overly large tree and pruning away some of the branches using a test sample or v-fold cross-validation to estimate the total sum of squared errors. For regression, CART builds a piecewise-constant model with each leaf node fitted with a mean function (Loh 2008). The CART algorithm is the decision tree method used by *STATISTICA* 10 to fit regression tree models discussed in Chapter IV of this thesis. Even though the two-dimensional hierarchical interactions displayed by regression trees provide very good explanatory value, a limitation of regression trees is poor predictive power (Hastie et al. 2009).

Boosted Regression Trees

Boosting is a technique used to enhance the predictive performance of regression trees. Again, the main point of boosting is to sequentially apply the weak learning algorithm to repeatedly modified versions of the data, hence creating a sequence of weak learners (i.e., base classifiers). The predictions from all of the weak learners are combined through a weighted majority vote to produce the final prediction

(Hastie et al. 2009). In boosting, models are fit iteratively to the training data using methods to increase emphasis on observations that are modeled poorly by the existing collection of models (Elith et al. 2008).

As related to regression problems, boosting is a form of functional gradient descent (Elith et al. 2008). Take a loss function that represents the loss in predictive performance due to a suboptimal model. Boosting is a numerical optimization technique for minimizing the loss function by adding, at each step, a new model (e.g., a regression tree) that best reduces, or steps down the gradient of, the loss function (Elith et al. 2008). The boosting approach used in BRT methodology places its origins within the machine learning community (Schapire 2003), but more recent developments in the statistical community interpret it as an advanced form of regression (Friedman et al. 2000).

Elith et al. (2008) explained BRT as follows. The initial regression tree is the one that reduces the loss function the most. At each iteration the focus is on the residuals and root mean square error reduction. In the second step, a regression tree, which can contain different variables and split points than the first tree, is fit to the prediction residuals of the first tree. The overall model now contains two trees (i.e., two terms), and the residuals from this two-term model are estimated. The process is stage-wise, i.e., existing trees are left unchanged as the model grows increasingly larger. Only the fitted value for each observation is re-estimated at each step to reflect the contribution of the newly added tree. In the end, the final BRT model is a linear combination of

numerous trees and can be thought of as a regression model with each term being a tree.

Stochastic Gradient Boosting

Stochastic gradient boosting is just one of numerous algorithms for modeling BRT. Friedman (2002) stated "gradient boosting constructs additive regression models by sequentially fitting a simple parameterized function (i.e., base learner) to current 'pseudo'-residuals by least-squares at each iteration." With respect to the model values at each training data point, the "pseudo" residuals are the gradient of the loss function being minimized (Friedman 2002). Friedman (2002) explained the gradient boosting procedure as follows. In the function estimation problem, one has a response variable y and a set of random explanatory values $\mathbf{x} = \{x_1, ..., x_n\}$. Friedman (2002) noted, given a training sample $\{y_i, \mathbf{x}_i\}_1^N$ of known $(y, \mathbf{x})$ values, the objective is to find a function $F^*(\mathbf{x})$ that maps $\mathbf{x}$ onto y, such that over the joint distribution of all $(y, \mathbf{x})$ values, the expected value of some loss function $\Psi(y, F(\mathbf{x}))$ is minimized

$$F^*(\mathbf{x}) = \arg\min_{F(\mathbf{x})} E_{y,\mathbf{x}} \Psi(y, F(\mathbf{x})). \qquad [1]$$

Boosting approximates $F^*(\mathbf{x})$ by an additive expansion of the form

$$F(\mathbf{x}) = \sum_{m=0}^{M} \beta_m h(\mathbf{x}; \mathbf{a}_m), \qquad [2]$$

where functions $h(\mathbf{x}; \mathbf{a})$ (i.e., "base learner") are generally simple functions of $\mathbf{x}$ with parameters $\mathbf{a} = \{a_1, a_2, ...\}$, see Friedman (2002). In a forward stage-wise manner, the expansion coefficients β_m and the parameters $\mathbf{a}_m$ are jointly fit to the training data.

According to Friedman (2002), one starts with a preliminary guess $F_0(\mathbf{x})$, and then for $m = 1,2,...,M$

$$(\beta_m, \mathbf{a}_m) = \arg\min_{\beta,\mathbf{a}} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x};\mathbf{a})) \quad [3]$$

and

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \beta_m h(\mathbf{x};\mathbf{a}_m). \quad [4]$$

Gradient boosting approximately solves [3] for arbitrary differentiable loss functions $\Psi(y, F(\mathbf{x}))$ with a two-step procedure (Friedman 2001). First, the function $h(\mathbf{x};\mathbf{a})$ is fit using least squares

$$\mathbf{a}_m = \arg\min_{\mathbf{a},\rho} \sum_{i=1}^{N} \left[\tilde{y}_{im} - \rho h(\mathbf{x}_i;\mathbf{a})\right]^2 \quad [5]$$

to the current "pseudo" residuals

$$\tilde{y}_{im} = -\left[\frac{\partial \Psi(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x}) = F_{m-1}(\mathbf{x})}. \quad [6]$$

Second, given $h(\mathbf{x};\mathbf{a}_m)$, the optimal value for the coefficient β_m is calculated to be

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x};\mathbf{a}_m)). \quad [7]$$

According to Friedman (2002), gradient tree boosting concentrates this technique to the specific case where $h(\mathbf{x};\mathbf{a})$ is an L-terminal node regression tree. Thus, L is a meta-parameter of the whole boosting procedure, and L is to be adjusted to maximize estimated performance for the data at hand (Hastie et al. 2009). Friedman (2002) stated, at each iteration m a regression tree partitions the $\mathbf{x}$-space into L disjoint regions $\{R_{lm}\}_{l=1}^{L}$ and predicts a different constant value in each one

$$h(\mathbf{x};\{R_{lm}\}_1^L) = \sum_{l=1}^{L} \bar{y}_{lm} 1(\mathbf{x} \in R_{lm}). \quad [8]$$

In [8], $\bar{y}_{lm} = mean_{\mathbf{x}_i \in R_{lm}}(\tilde{y}_{im})$ is the mean of [6] in each separate region R_{lm} (Friedman 2002). Now, with regression trees, [7] can be solved separately within each region R_{lm} defined by the related terminal node l of the mth tree (Friedman 2002). Friedman (2002) noted that because the tree [8] predicts a separate constant value $\bar{y}_{lm}$ within each region R_{lm}, the solution to [7] diminishes to a simple "location" estimate based on Ψ

$$\gamma_{lm} = \arg\min_{\gamma} \sum_{\mathbf{x}_i \in R_{lm}} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma). \quad [9]$$

Friedman (2002) clarified the existing approximation $F_{m-1}(\mathbf{x})$ is separately updated in each related region

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + v * \gamma_{lm} 1(\mathbf{x} \in R_{lm}). \quad [10]$$

The "shrinkage" parameter[5] $0 < v \leq 1$ controls the learning rate of the boosting procedure (Friedman 2002). In other words, the simplest implementation of shrinkage in the context of boosting is to scale the contribution of each tree by a factor of v when it is added to the current approximation (Hastie et al. 2009). Empirically it has been found that small values (v ≤ 0.1) favor better test error (i.e., lead to models with better predictive validity), and require a corresponding larger number of boosting iterations (Friedman 2001, 2002).

[5] *STATISTICA* 10 *Boosted Trees* module will compute a weighted "additive" expansion of simple regression trees. The specific weight with which consecutive simple trees are added into the prediction equation is usually a constant, and referred to as the learning rate or shrinkage parameter.

With motivation from the hybrid bagging-boosting[6] procedure(i.e., "adaptive bagging"), a minor modification was made by Friedman (2002) to the gradient boosting algorithm to include randomness as an essential part of the procedure which led to Friedman's (2002) Stochastic Gradient Boosting algorithm. At each iteration a subsample of the training data is drawn at random without replacement from the full training data set. The subsample is used to fit the base learner and compute the updated model for the existing iteration (Friedman 2002). The randomization helps protect against overfitting, reduce the variance of the final model, and improve predictive performance (Friedman 2002; Hastie et al. 2009).

Friedman (2002) explained the stochastic procedure as it is applied to the Gradient Boosting algorithm as follows. Take $\{y_i, \mathbf{x}_i\}_1^N$ to be the entire training data set and $\{\pi(i)\}_1^N$ to be a random permutation of the integers $\{1,...,N\}$, and then a random subsample of size $\tilde{N} < N$ is given by $\{y_{\pi(i)}, \mathbf{x}_{\pi(i)}\}_1^{\tilde{N}}$ (Friedman 2002). Now, Friedman's (2002) stochastic gradient boosting algorithm is identical to the gradient boosting algorithm except that $\{\pi(i)\}_1^N = rand_perm\{i\}_1^N$ is inserted after the second step and is used throughout the algorithm. Refer to Friedman (2002) for the step-by-step procedure associated with the generalized Gradient Boosting and Stochastic Gradient Boosting algorithms.

[6] Refer to Leo Breiman's 1999 University of California at Berkeley technical report titled "Using adaptive bagging to debias regressions." http://www.stat.berkeley.edu/tech-reports/547.pdf referenced on 06/20/2011

Software and Parameters

STATISTICA 10[7] was used in this thesis to estimate the BRT models discussed in Chapter IV and V as well as the CART regression tree models discussed in Chapter IV. *STATISTICA* is a statistics and analytics software package developed by StatSoft. *STATISTICA* provides the user with data analysis, data management, data mining, and data visualization procedures. The first DOS version of *STATISTICA* was released in 1991.[8]

The BRT algorithm of *STATISTICA* 10 is a "full featured implementation of the stochastic gradient boosting method" (Friedman 2002; Hastie et al. 2009). Five key parameters used to control the Stochastic Gradient Boosting algorithm were manipulated in the BRT analysis. First, the "learning rate," or the shrinkage parameter, (*lr*) specified the specific weight with which consecutive simple regression trees are added into the prediction equation, i.e., *lr* specified the shrinkage applied to each tree in the final boosted regression tree model (Elith et al. 2008). For example, a BRT model with 500 trees fitted and *lr* equal to 0.01 will produce predictions that are the sum of predictions from each of the 500 trees multiplied by 0.01. Second, the "number of additive terms" (*nat*) specified the number of simple regression trees (i.e., additive terms) to be computed in successive boosting steps. According to Elith et al. (2008), a smaller *lr* and larger *nat* are preferable, conditional on the number of observations and available computation time. Since smaller values for *lr* (i.e., more shrinkage) result in larger training risk for the same *nat*, both *lr* and *nat* control the prediction risk on the

[7] http://www.statsoft.com/ referenced on 06/20/2011
[8] http://en.wikipedia.org/wiki/STATISTICA referenced on 06/15/2011

training data (Hastie et al. 2009). Third, the "maximum number of nodes" (*mnn*) specified the maximum number of nodes allowed for each individual tree in the boosting sequence. This is used as a stopping parameter in a sense that each time a parent node is split, the total number of nodes in the tree is examined, and the splitting is stopped if this number exceeds the number specified by *mnn*. Setting *mnn* equal to three (i.e., single split regression trees or stumps) produced BRT models with only main effects. Setting *mnn* equal to five produced BRT models with main effects and two-variable interactions, and so on. For illustration, three and five were the values used for *mnn* throughout this research. As noted by Elith et al. (2008), BRT modeling regularization involved jointly optimizing *nat*, *lr*, and *mnn*. Fourth, the "subsample proportion" (*sp*) was used for selecting the random learning sample for consecutive boosting steps. Given the work of Hastie et al. (2009) and Elith et al. (2008), a reasonable, balanced *sp* value of 0.5 was used to perform the BRT modeling analysis. Fifth, a value of 0.2 (or 20%), which is standard, was used for the "random test data proportion" parameter. This implies 80% of the observations were randomly selected used for the training (i.e., modeling) sample and the other 20% of the observations were used for the testing (i.e., validation) sample. Such an 80% and 20% split is quite typical in practice.

Imputation of the Data Set and BRT Models

In data mining and statistics, imputation is the substitution of some value for a missing data point or a missing component of a data point.[9] Three different imputation methods were used to replace missing predictor variable records with values, and a

[9] http://en.wikipedia.org/wiki/Imputation_(statistics) referenced on 06/12/2011

non-imputation method (NI) was used that does not replace missing predictor variable records with values. The three different imputation methods used are “Expectation-Maximization” (EM), “Last Observation Carried Forward” (LOCF), and “Median” (MED). EM imputation replaces missing values in a (predictor) variable using an expectation and maximum likelihood estimation procedure. LOCF imputation replaces missing values in a predictor variable with the last known value for the predictor variable. MED imputation replaces missing values in a predictor variable with the median value of the predictor variable. Zeng (2011) provided more discussion of these imputation methods and an explanation of these different algorithms. The literature on missing data in manufacturing industry applications is sparse. For additional information on using these methods and other forms of imputation, refer to Little (1992), Schafer (1997), Enders (2001), Faraway (2005), Truxillo (2005), Gelman and Hill (2007), Horton and Kleinman (2007), and Hamer (2009). In Chapter V, the discussion is on comparing the different imputation methods in terms of the predictive performance of BRT models. Importantly, the implementation of the Stochastic Gradient Boosting Algorithm (Friedman 2002) in *STATISTICA* 10 can easily incorporate missing data in the predictors. During BRT model building, when missing data are encountered for a particular observation (i.e., case), then the prediction for that observation is made based on the last preceding (non-terminal) node in the respective tree. For example, if at a particular point in the sequence of trees a predictor variable is selected at the root (or other non-terminal) node for which some cases have no valid data, then the prediction for those cases is simply based on the overall mean at the root (or other non-terminal) node.

Model Comparison

To compare the predictive performance of the models, the root mean squared error for prediction (RMSEP) and the RMSEP relative to the mean of the response variable as a percent (RMSEP%) were used as performance measures for the validation data set. The RMSEP value was given by

$$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \qquad [11]$$

where y_i was the observed value, $\hat{y}_i$ was the predicted value, and n was the total number of records in the validation data set. The RMSEP% value was given by

$$\left(\frac{RMSEP}{\hat{\mu}_y}\right) * 100\%, \qquad [12]$$

where $\hat{\mu}_y$ was the mean of the observed response values in the validation data set. The use of $\hat{\mu}_y$ allowed for easier comparison of each BRT model to a similar baseline. Importantly, single values of RMSEP and RMSEP% (on the validation data set for each model) will be used to compare BRT models to CART regression tree models in Chapter IV; whereas, a repeated random sub-sampling validation technique to compare BRT models across imputation methods will be used in Chapter V.

An advantage of regression trees is the high explanatory value and ability to detect multiple levels of interactions. However, sometimes regression tree models do not predict well. Boosting creates a model with hundreds to thousands of potential smaller trees, which has been demonstrated to improve prediction quality. However, interpretation of the final BRT model can be challenging. One way to help understand

or interpret the BRT model is to assess the predictor importance values or contribution of each input variable in predicting the response (Hastie et al. 2009). Often, only a few of the predictor variables have a substantial influence on the response variable, and the vast majority of predictor variables are extraneous and could have just as well not been included in the analysis for predicting the response (Hastie et al. 2009). *STATISTICA* 10 has a feature that estimates "predictor importance values." During the building of each tree, for every split, predictor statistics (i.e., sums of squares regression) are computed for each predictor variable. The best predictor variable, which yields the best split at the respective node, is chosen as a split. The particular predictor variable chosen is the one that gives "maximal estimated improvement in squared error risk over that for a constant fit over the entire region" (Hastie et al. 2009). An average of the predictor statistic for all variables over all splits and over all trees in the boosting sequence is then computed. The final predictor importance values are computed by normalizing those averages so that the highest average is assigned the value of one, and the importance of all other predictors is expressed in terms of the relative magnitudes of the average values of the predictor statistic, relative to the most important predictor. Because of "shrinkage" (i.e., learning rate) the masking of important variables by others with which they are highly correlated is less of a problem (Hastie et al. 2009). Predictor importance values will only be discussed in Chapter IV.

CHAPTER IV. PREDICTING THE STRENGTH PROPERTIES OF WOOD COMPOSITES USING BOOSTED REGRESSION TREES

In this chapter, an analysis of BRT and CART regression tree models for predicting two different strength property metrics (i.e., IB and MOR) of particleboard wood composite will be discussed. First, the U.S. particleboard manufacturer data set used for the BRT and CART regression tree models is reviewed. Second, the analysis of BRT and CART regression tree models for predicting MOR and IB strength metrics of particleboard wood composite is discussed. Third, remarks in regards to the analyses for MOR and IB are given.

Data Set

A time-ordered data set was obtained from a U.S. particleboard (wood composite) manufacturer. The key quality strength metrics (i.e., response variables) for this manufacturer's product were MOR and IB. The data set consisted of 4,307 records,[10] which spanned the time period from March 2009 to June 2010. There were 189 possible continuous predictor variables.[11] There were 118 different particleboard product types manufactured by the producer within the 4,307 records. Product types were not differentiated in the overall BRT model predictions of MOR and IB. Of the 4,307 observations in the data set, 3,449 observations were used for training and 858 observations were randomly selected for validation.

[10] The 4,307 records were derived after 104 records were removed from the original data set given incomplete cell data in the records.

[11] The 189 continuous predictor variables were obtained after 33 predictor variables were removed from the initial set of 222 predictor variables. The 33 predictor variables were removed due to the fact that at least 2.5% of each variable's records were null values or a single constant value was represented in the variable for the whole data set.

BRT and CART Regression Tree Models for MOR

Table 1 provides statistics obtained from analysis performed on the validation data set for 70 different BRT models predicting MOR. Each of these 70 aforementioned BRT models used a *mnn* parameter value equal to three, but the values for the *lr* and *nat* parameters were not held constant. The range of RMSEP% values resulting from BRT models for MOR with various combinations of the *lr* and *nat* parameters and the value three for the *mnn* parameter is given in Figure 1. The 70 different BRT models represented in Table 1 and Figure 1 were the product of testing 10 different levels of *lr* values (0.005, 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, and 0.5), a *mnn* value equal to three, and seven different levels of *nat* values (100, 200, 300, 400, 500, 600, and 1,000). Table 1 and Figure 1 both provide RMSEP% values on the validation data set for the 70 different BRT models predicting MOR. Plus, Table 1 provides RMSEP values and the Pearson correlation coefficient values between the predicted and observed MOR values of the validation data set for the 70 different BRT models predicting MOR. The Pearson correlation coefficient measures the correlation (linear dependence) between two variables X and Y, which is a value between -1 and +1 inclusive.[12] For MOR and a parameter value of three for *mnn*, the lowest RMSEP value obtained was 1.051 MPa (refer to Table 1), while the other parameter settings were 0.15 for *lr* and 1,000 for *nat* (refer to Table 1 and Figure 1). For the BRT model with a RMSEP value equal to 1.051 MPa, the Pearson correlation coefficient value between the predicted and observed MOR values was equal to 0.91, and the RMSEP% value was equal to 8.5% (refer to Table 1).

[12] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient referenced on 06/19/2011

Table 1: Statistics obtained from analysis performed on the validation data set for 70 different BRT models predicting MOR, with a value of three being used for the parameter *mnn*.*

lr	***nat* = 100**	***nat* = 200**	***nat* = 300**	***nat* = 400**	***nat* = 500**	***nat* = 600**	***nat* = 1,000**
0.005	0.83, 1.443, 11.6%	0.83, 1.42, 11.5%	0.84, 1.403, 11.3%	0.84, 1.39, 11.2%	0.84, 1.379, 11.1%	0.85, 1.368, 11%	0.85, 1.336, 10.8%
0.01	0.83, 1.419, 11.4%	0.84, 1.39, 11.2%	0.85, 1.37, 11%	0.85, 1.352, 10.9%	0.85, 1.337, 10.8%	0.86, 1.323, 10.7%	0.87, 1.281, 10.3%
0.05	0.85, 1.339, 10.8%	0.87, 1.281, 10.3%	0.87, 1.243, 10%	0.88, 1.212, 9.8%	0.89, 1.187, 9.6%	0.89, 1.17, 9.4%	0.9, 1.119, 9%
0.1	0.86, 1.294, 10.4%	0.88, 1.221, 9.9%	0.89, 1.177, 9.5%	0.89, 1.151, 9.3%	0.9, 1.137, 9.2%	0.9, 1.116, 9%	0.91, 1.077, 8.7%
0.15	0.87, 1.267, 10.2%	0.89, 1.184, 9.6%	0.89, 1.145, 9.2%	0.9, 1.115, 9%	0.9, 1.092, 8.8%	0.91, 1.081, 8.7%	0.91, 1.051, 8.5%
0.2	0.88, 1.238, 10%	0.89, 1.168, 9.4%	0.9, 1.129, 9.1%	0.9, 1.104, 8.9%	0.9, 1.1, 8.9%	0.91, 1.084, 8.7%	0.91, 1.059, 8.5%
0.25	0.88, 1.216, 9.8%	0.9, 1.14, 9.2%	0.9, 1.097, 8.8%	0.91, 1.085, 8.8%	0.91, 1.072, 8.7%	0.91, 1.07, 8.6%	0.91, 1.062, 8.6%
0.3	0.88, 1.199, 9.7%	0.9, 1.134, 9.1%	0.9, 1.099, 8.9%	0.91, 1.084, 8.7%	0.91, 1.079, 8.7%	0.91, 1.078, 8.7%	0.91, 1.07, 8.6%
0.4	0.89, 1.163, 9.4%	0.9, 1.12, 9%	0.9, 1.105, 8.9%	0.9, 1.101, 8.9%	0.9, 1.091, 8.8%	0.91, 1.09, 8.8%	0.91, 1.089, 8.8%
0.5	0.89, 1.187, 9.6%	0.89, 1.15, 9.3%	0.9, 1.13, 9.1%	0.9, 1.129, 9.1%	0.9, 1.129, 9.1%	0.9, 1.129, 9.1%	0.9, 1.129, 9.1%

*A total of 10 different values for the parameter *lr* and seven different values for the parameter *nat* are shown here. The statistics provided in Table 1 from top to bottom of each cell are: (1) the Pearson correlation coefficient (e.g., 0.87) between the observed and predicted MOR values for the validation data set, (2) the RMSEP value (e.g., 1.243) obtained for the validation data set, and (3) the RMSEP% value (e.g., 10%) obtained for the validation data set.

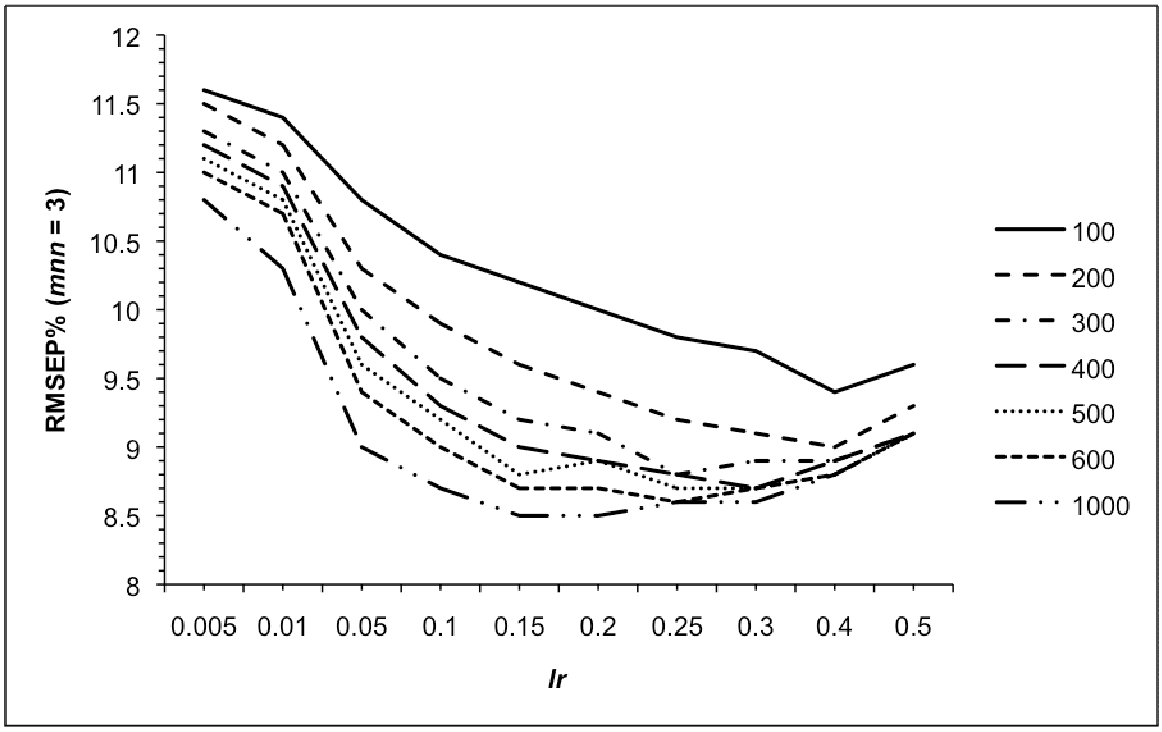


Figure 1: The relationship between learning rate (*lr*) and MOR RMSEP% for the 70 different BRT models with seven values chosen for the number of additive terms (*nat*) and a value of three chosen for the maximum number of nodes (*mnn*).

Table 2 and Figure 2 have similarities to Table 1 and Figure 1, respectively. Just like Table 1, Table 2 provides Pearson correlation coefficient, RMSEP, and RMSEP% values obtained from analysis performed on the validation data set for 70 different BRT models predicting MOR. Just like Figure 1, the range of RMSEP% values resulting from BRT models for MOR with various combinations of the *lr* and *nat* parameters is given in Figure 2. As well, The 70 different BRT models represented in Table 2 and Figure 2 were the product of testing the same 10 different levels of *lr* values and same seven different levels of *nat* values as above. The difference between the tables and the figures was that the *mnn* parameter value changed from three to five. For MOR and a parameter value of five for *mnn*, the lowest RMSEP value obtained was 1.056 MPa (refer to Table 2), while the other parameter settings were 0.1 for *lr* and 1,000 for *nat* (refer to Table 2 and Figure 2). For the BRT model with a RMSEP value equal to 1.056 MPa, the Pearson correlation coefficient value between the predicted and observed MOR values was equal to 0.91, and the RMSEP% value was equal to 8.5% (refer to Table 2). So, the parameter settings for the lowest RMSEP for MOR of 1.051 MPa were an *lr* value of 0.15, an *mnn* value of three, and a *nat* value of 1,000. Importantly, the optimal number of trees obtained for these 1,000 iterations was 943 (i.e., the smallest average squared error for the validation sample was obtained at 943 trees for these 1,000 boosting steps). The RMSEP% for this BRT model was 8.5%. A scatterplot of the observed MOR values and the predicted MOR values for the validation data set is given in Figure 3. The Pearson correlation coefficient value between the observed MOR values and the predicted MOR values was 0.91. Overall, RMSEP

Table 2: Statistics obtained from analysis performed on the validation data set for 70 different BRT models predicting MOR, with a value of five being used for the parameter *mnn*.*

lr	***nat* = 100**	***nat* = 200**	***nat* = 300**	***nat* = 400**	***nat* = 500**	***nat* = 600**	***nat* = 1,000**
0.005	0.84, 1.384, 11.2%	0.85, 1.36, 11%	0.85, 1.342, 10.8%	0.86, 1.326, 10.7%	0.86, 1.313, 10.6%	0.86, 1.302, 10.5%	0.87, 1.262, 10.2%
0.01	0.85, 1.36, 11%	0.86, 1.327, 10.7%	0.86, 1.303, 10.5%	0.87, 1.28, 10.3%	0.87, 1.263, 10.2%	0.87, 1.249, 10.1%	0.88, 1.202, 9.7%
0.05	0.87, 1.272, 10.3%	0.88, 1.208, 9.7%	0.89, 1.168, 9.4%	0.9, 1.138, 9.2%	0.9, 1.114, 9%	0.9, 1.101, 8.9%	0.91, 1.062, 8.6%
0.1	0.88, 1.211, 9.8%	0.9, 1.128, 9.1%	0.9, 1.101, 8.9%	0.91, 1.088, 8.8%	0.91, 1.077, 8.7%	0.91, 1.068, 8.6%	0.91, 1.056, 8.5%
0.15	0.89, 1.172, 9.5%	0.9, 1.113, 9%	0.9, 1.102, 8.9%	0.91, 1.087, 8.8%	0.91, 1.079, 8.7%	0.91, 1.076, 8.7%	0.91, 1.066, 8.6%
0.2	0.89, 1.172, 9.5%	0.9, 1.126, 9.1%	0.9, 1.112, 9%	0.9, 1.108, 8.9%	0.9, 1.102, 8.9%	0.9, 1.1, 8.9%	0.9, 1.1, 8.9%
0.25	0.9, 1.143, 9.2%	0.91, 1.086, 8.8%	0.91, 1.076, 8.7%	0.91, 1.076, 8.7%	0.91, 1.076, 8.7%	0.91, 1.076, 8.7%	0.91, 1.076, 8.7%
0.3	0.89, 1.15, 9.3%	0.9, 1.137, 9.2%	0.9, 1.124, 9.1%	0.9, 1.124, 9.1%	0.9, 1.124, 9.1%	0.9, 1.124, 9.1%	0.9, 1.124, 9.1%
0.4	0.89, 1.156, 9.3%	0.9, 1.139, 9.2%	0.9, 1.138, 9.2%	0.9, 1.135, 9.2%	0.9, 1.135, 9.2%	0.9, 1.135, 9.2%	0.9, 1.135, 9.2%
0.5	0.89, 1.162, 9.4%	0.9, 1.141, 9.2%	0.9, 1.138, 9.2%	0.9, 1.138, 9.2%	0.9, 1.138, 9.2%	0.9, 1.138, 9.2%	0.9, 1.138, 9.2%

*A total of 10 different values for the parameter *lr* and seven different values for the parameter *nat* are shown here. The statistics provided in Table 2 from top to bottom of each cell are: (1) the Pearson correlation coefficient (e.g., 0.89) between the observed and predicted MOR values for the validation data set, (2) the RMSEP value (e.g., 1.168) obtained for the validation data set, and (3) the RMSEP% value (e.g., 9.4%) obtained for the validation data set.

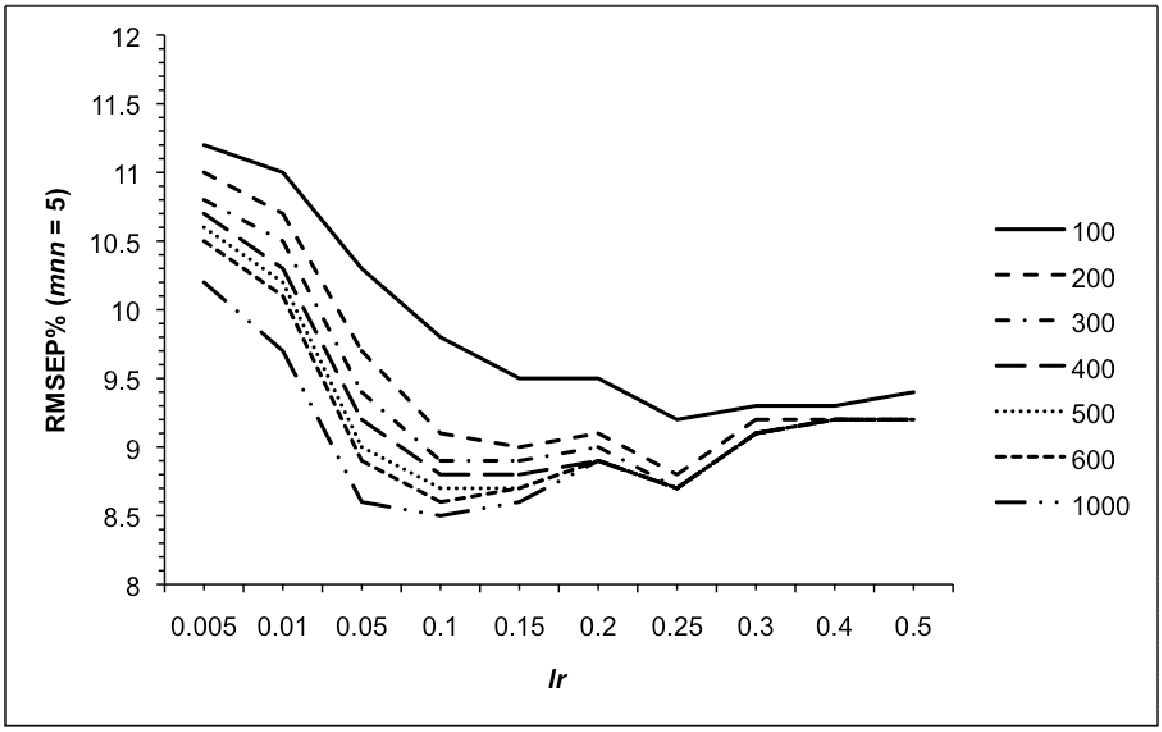


Figure 2: The relationship between learning rate (*lr*) and MOR RMSEP% for the 70 different BRT models with seven values chosen for the number of additive terms (*nat*) and a value of five chosen for the maximum number of nodes (*mnn*).

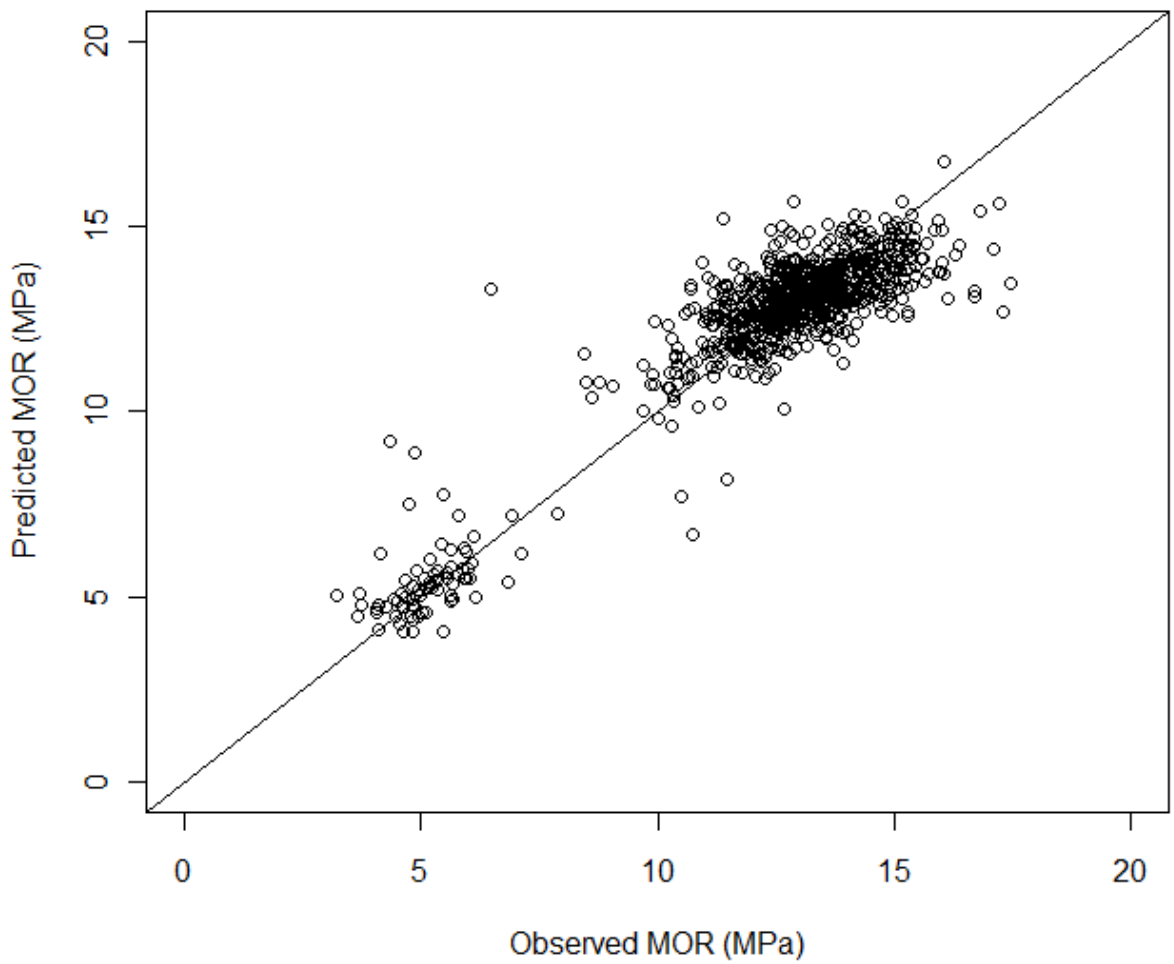


Figure 3: Scatterplot of the observed values versus the predicted values of the validation data set for the BRT model that best predicts MOR.

values ranged from 1.051 to 1.443 MPa, and RMSEP% values ranged from 8.5% to 11.6%.

For the sake of comparison, a regression tree model with MOR as the response variable was fit to the same data using the CART algorithm of *STATISTICA* 10 (Breiman et al. 1984). The regression tree model can be seen in Figure 4. The RMSEP and RMSEP% values obtained for the validation data set of the regression tree model were 1.263 MPa and 10.2%, respectively. A scatterplot of the observed MOR values and the predicted MOR values can be seen in Figure 5. The Pearson correlation coefficient value between the observed MOR values and the predicted MOR values was 0.87. Comparing the "observed vs. predicted" scatterplots revealed the predictive modeling weakness of regression tree models (compare Figures 3 and 5). For the regression tree model, the scatterplot showed "step-like" predictions, and the number of these "steps" had to do with the number of terminal nodes in the tree; whereas, the scatterplot for the BRT model showed more of the desired linear correlation. The BRT model for MOR with the lowest RMSEP and RMSEP% predicted better than the regression tree model for MOR on the validation data set.

It was important to examine the top-five predictor importance values in an attempt to understand or interpret the BRT model for MOR. For MOR, the predictors with the top five importance values were related to particleboard "pressing temperature zones," "thickness of pressing," and "pressing pressure." The predictor importance values for these predictor variables ranged from 0.97 to one.

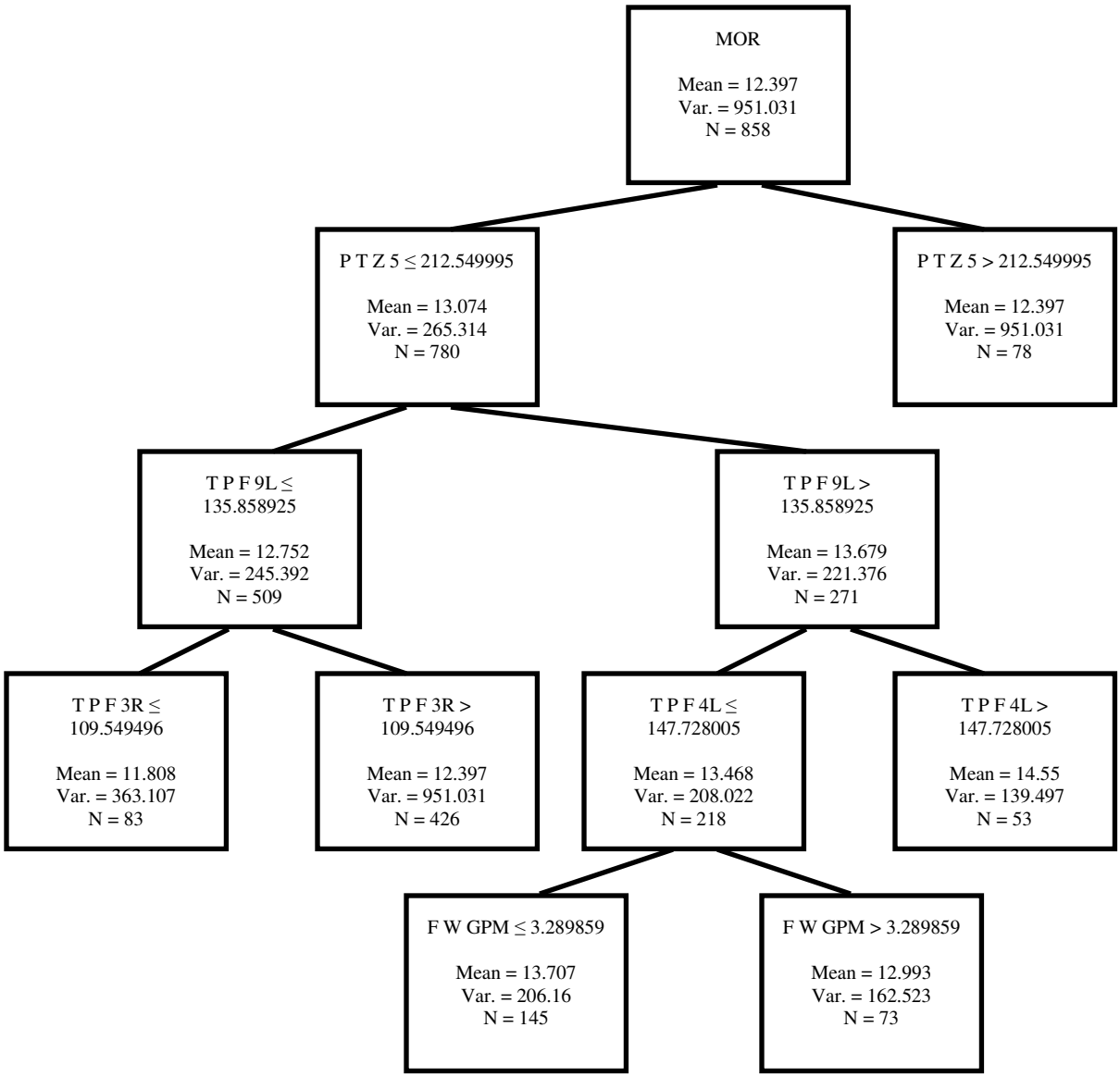


Figure 4: Regression tree model for MOR.

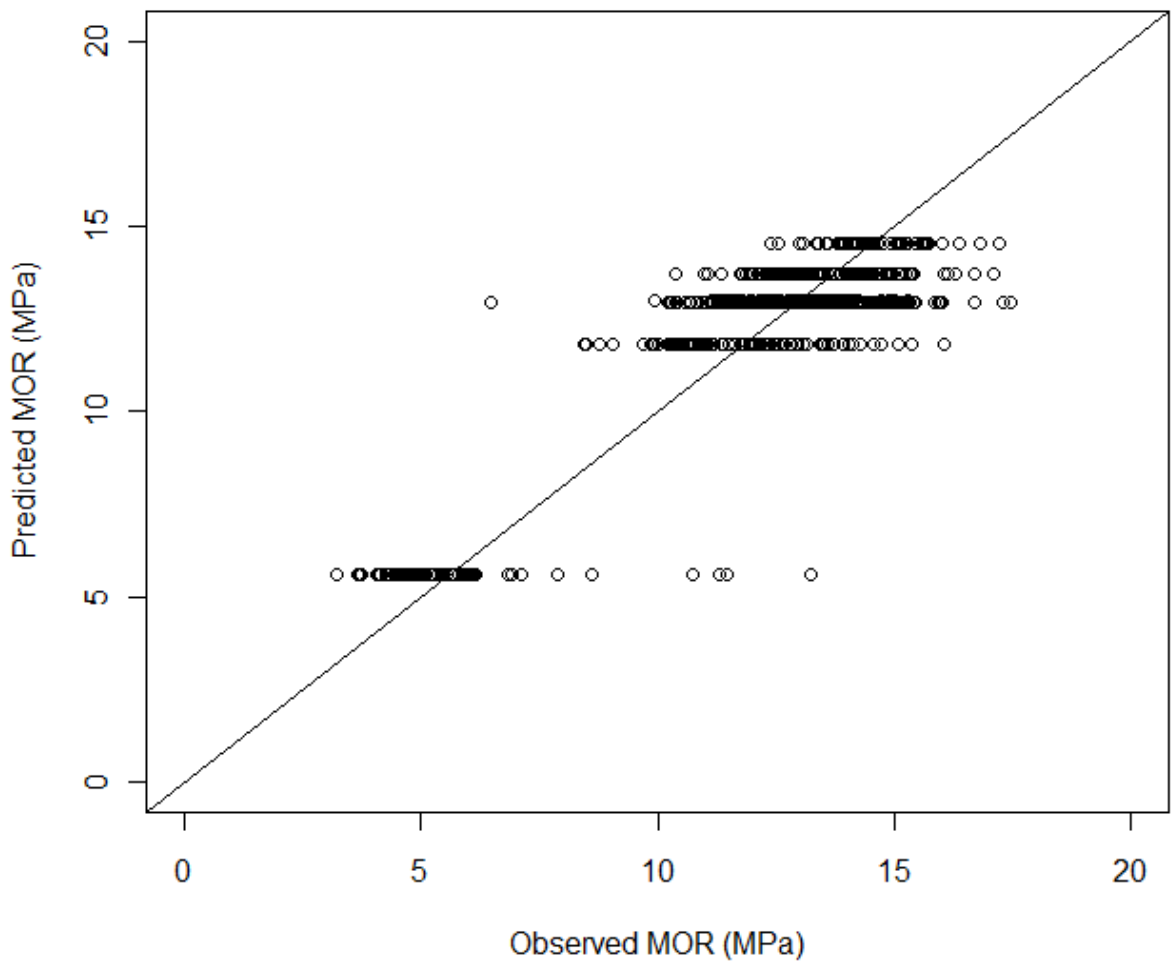


Figure 5: Scatterplot of the observed values versus the predicted values of the validation data set for the regression tree model that predicts MOR.

BRT and CART Regression Tree Models for IB

Table 3 provides statistics obtained from analysis performed on the validation data set for 70 different BRT models predicting IB. These 70 BRT models used a *mnn* parameter value equal to three, but the values for the *lr* and *nat* parameters were again not held constant. The range of RMSEP% values resulting from BRT models for IB with the various combinations of the *lr* and *nat* parameters and the value three for the *mnn* parameter is given in Figure 6. The 70 BRT models represented in Table 3 and Figure 6 were the product of testing the aforementioned 10 different levels of *lr* values and seven different levels of *nat* values. Table 3 and Figure 6 both provide RMSEP% values on the validation data set for the 70 BRT models predicting IB. Plus, Table 3 provides RMSEP values and the Pearson correlation coefficient values between the predicted and observed IB values of the validation data set for the 70 BRT models predicting IB. For IB and a parameter value of three for *mnn*, the lowest RMSEP value obtained was 0.076 MPa (refer to Table 3), while the other parameter settings were 0.15 for *lr* and 1,000 for *nat* (refer to Table 3 and Figure 6). For the BRT model with a RMSEP value equal to 0.076 MPa, the Pearson correlation coefficient value between the predicted and observed IB values was equal to 0.85, and the RMSEP% value was equal to 13% (refer to Table 3).

Much like the case in MOR, for IB, Table 4 and Figure 7 have similarities to Table 3 and Figure 6, respectively. Just like Table 3, Table 4 provides Pearson correlation coefficient, RMSEP, and RMSEP% values obtained from analysis performed on the validation data set for 70 different BRT models predicting IB. Figure 7 is similar to

Table 3: Statistics obtained from analysis performed on the validation data set for 70 different BRT models predicting IB, with a value of three being used for the parameter *mnn*.*

lr	***nat* = 100**	***nat* = 200**	***nat* = 300**	***nat* = 400**	***nat* = 500**	***nat* = 600**	***nat* = 1,000**
0.005	0.66, 0.108, 18.6%	0.69, 0.105, 18%	0.7, 0.103, 17.6%	0.71, 0.101, 17.4%	0.72, 0.1, 17.2%	0.73, 0.099, 17%	0.75, 0.095, 16.4%
0.01	0.69, 0.105, 18%	0.71, 0.101, 17.3%	0.73, 0.099, 17%	0.74, 0.097, 16.7%	0.75, 0.096, 16.4%	0.76, 0.094, 16.2%	0.78, 0.091, 15.6%
0.05	0.75, 0.095, 16.4%	0.78, 0.091, 15.6%	0.79, 0.088, 15.1%	0.8, 0.086, 14.8%	0.81, 0.085, 14.6%	0.82, 0.084, 14.4%	0.84, 0.079, 13.6%
0.1	0.77, 0.092, 15.8%	0.8, 0.087, 15%	0.82, 0.083, 14.3%	0.83, 0.081, 14%	0.83, 0.08, 13.8%	0.84, 0.079, 13.6%	0.85, 0.076, 13.1%
0.15	0.79, 0.089, 15.3%	0.81, 0.084, 14.5%	0.83, 0.081, 13.9%	0.84, 0.079, 13.7%	0.84, 0.078, 13.5%	0.84, 0.078, 13.3%	0.85, 0.076, 13%
0.2	0.8, 0.087, 14.9%	0.83, 0.081, 14%	0.84, 0.079, 13.5%	0.84, 0.078, 13.4%	0.84, 0.077, 13.3%	0.85, 0.077, 13.2%	0.85, 0.077, 13.1%
0.25	0.81, 0.085, 14.6%	0.84, 0.079, 13.6%	0.84, 0.077, 13.3%	0.84, 0.077, 13.3%	0.84, 0.077, 13.2%	0.85, 0.077, 13.2%	0.85, 0.077, 13.2%
0.3	0.82, 0.083, 14.3%	0.84, 0.078, 13.4%	0.84, 0.078, 13.4%	0.84, 0.078, 13.4%	0.84, 0.077, 13.3%	0.85, 0.077, 13.2%	0.85, 0.077, 13.2%
0.4	0.82, 0.083, 14.2%	0.83, 0.079, 13.7%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%
0.5	0.82, 0.082, 14.1%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%

*A total of 10 different values for the parameter *lr* and seven different values for the parameter *nat* are shown here. The statistics provided in Table 3 from top to bottom of each cell are: (1) the Pearson correlation coefficient (e.g., 0.79) between the observed and predicted MOR values for the validation data set, (2) the RMSEP value (e.g., 0.088) obtained for the validation data set, and (3) the RMSEP% value (e.g., 15.1%) obtained for the validation data set.

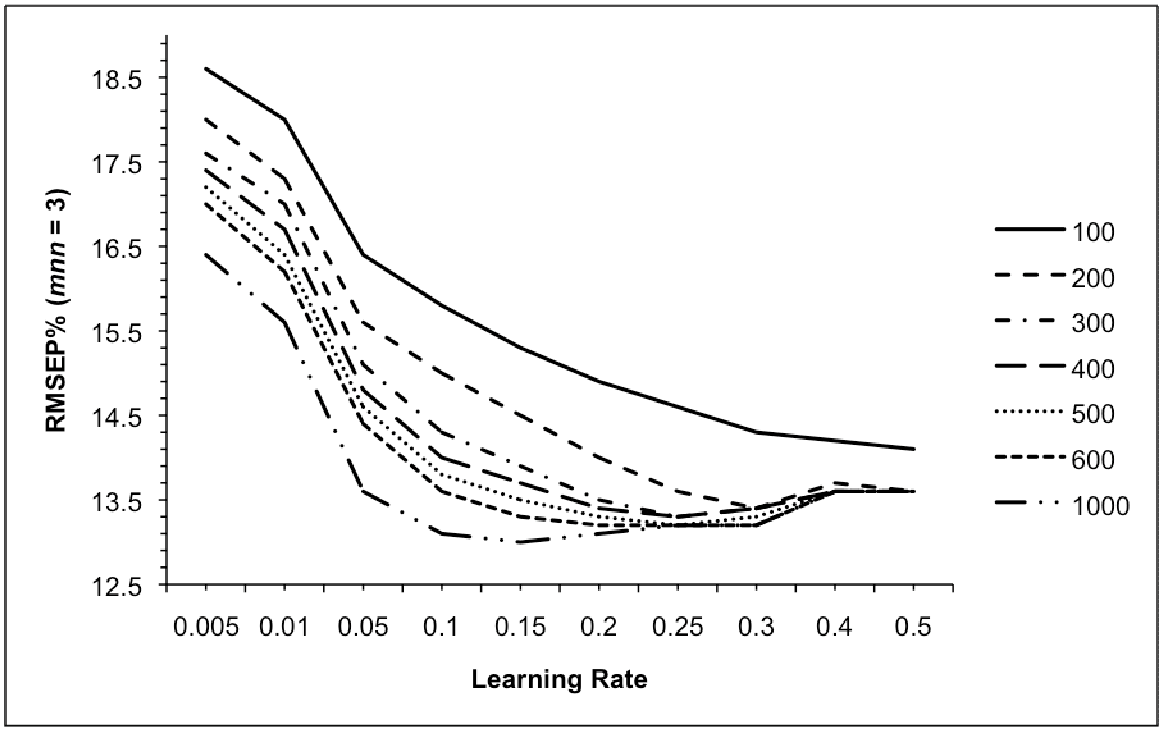


Figure 6: The relationship between learning rate (*lr*) and IB RMSEP% for the 70 different BRT models with seven values chosen for the number of additive terms (*nat*) and a value of three chosen for the maximum number of nodes (*mnn*).

Table 4: Statistics obtained from analysis performed on the validation data set for 70 different BRT models predicting IB, with a value of five being used for the parameter *mnn*.*

lr	***nat* = 100**	***nat* = 200**	***nat* = 300**	***nat* = 400**	***nat* = 500**	***nat* = 600**	***nat* = 1,000**
0.005	0.71, 0.102, 17.4%	0.73, 0.099, 17.1%	0.74, 0.097, 16.7%	0.75, 0.096, 16.5%	0.75, 0.095, 16.3%	0.76, 0.094, 16.1%	0.78, 0.091, 15.6%
0.01	0.72, 0.099, 17.1%	0.74, 0.096, 16.5%	0.76, 0.094, 16.2%	0.77, 0.093, 15.9%	0.77, 0.091, 15.7%	0.78, 0.09, 15.4%	0.81, 0.085, 14.6%
0.05	0.78, 0.091, 15.6%	0.81, 0.085, 14.6%	0.82, 0.082, 14.1%	0.83, 0.08, 13.7%	0.84, 0.078, 13.5%	0.85, 0.077, 13.2%	0.86, 0.075, 12.8%
0.1	0.8, 0.086, 14.7%	0.83, 0.08, 13.7%	0.84, 0.077, 13.3%	0.85, 0.076, 13.1%	0.85, 0.075, 12.9%	0.86, 0.075, 12.8%	0.86, 0.074, 12.7%
0.15	0.81, 0.084, 14.4%	0.84, 0.078, 13.5%	0.85, 0.077, 13.2%	0.85, 0.076, 13.1%	0.85, 0.075, 12.9%	0.85, 0.075, 12.8%	0.86, 0.075, 12.8%
0.2	0.83, 0.081, 14%	0.84, 0.078, 13.5%	0.84, 0.078, 13.3%	0.85, 0.077, 13.2%	0.85, 0.077, 13.2%	0.85, 0.077, 13.2%	0.85, 0.077, 13.2%
0.25	0.83, 0.081, 13.9%	0.84, 0.078, 13.4%	0.84, 0.078, 13.4%	0.84, 0.078, 13.4%	0.84, 0.078, 13.4%	0.84, 0.078, 13.3%	0.84, 0.078, 13.3%
0.3	0.83, 0.08, 13.7%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%
0.4	0.83, 0.081, 13.9%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%
0.5	0.83, 0.081, 13.9%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%	0.84, 0.079, 13.6%

*A total of 10 different values for the parameter *lr* and seven different values for the parameter *nat* are shown here. The statistics provided in Table 4 from top to bottom of each cell are: (1) the Pearson correlation coefficient (e.g., 0.82) between the observed and predicted MOR values for the validation data set, (2) the RMSEP value (e.g., 0.082) obtained for the validation data set, and (3) the RMSEP% value (e.g., 14.1%) obtained for the validation data set.

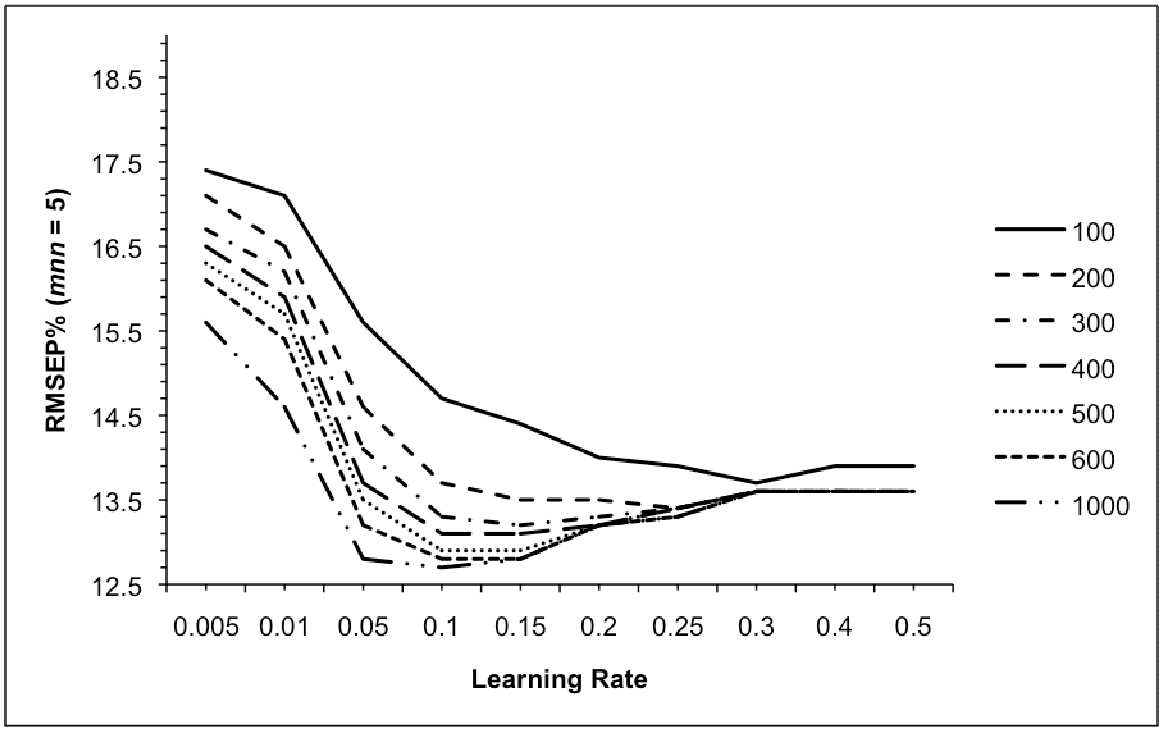


Figure 7: The relationship between learning rate (*lr*) and IB RMSEP% for the 70 different BRT models with seven values chosen for the number of additive terms (*nat*) and a value of five chosen for the maximum number of nodes (*mnn*).

Figure 6 indicating the range of RMSEP% values resulting from BRT models for IB with various combinations of the *lr* and *nat* parameters can be seen. As well, The 70 BRT models represented in Table 4 and Figure 7 were the product of testing the same 10 levels of *lr* values and same seven levels of *nat* values as above. The difference was that the *mnn* parameter value changed from three to five. For IB and a parameter value of five for *mnn*, the lowest RMSEP value obtained was 0.074 MPa (refer to Table 4), while the other parameter settings were 0.1 for *lr* and 1,000 for *nat* (refer to Table 4 and Figure 7). For the BRT model with a RMSEP value equal to 0.074 MPa, the Pearson correlation coefficient value between the predicted and observed IB values was equal to 0.86, and the RMSEP% value was equal to 12.7% (refer to Table 4).

So, the parameter settings for the lowest RMSEP for IB of 0.074 MPa were an *lr* value of 0.1, an *mnn* value of five, and a *nat* value of 1,000. Importantly, the optimal number of trees obtained for these 1,000 iterations was 957 (i.e., the smallest average squared error for the validation sample was obtained at 957 trees for these 1,000 boosting steps). The RMSEP% for this BRT model was 12.7%. A scatterplot of the observed IB values and the predicted IB values for the validation data set is given in Figure 8. The Pearson correlation coefficient value between the observed IB values and the predicted IB values was 0.86. Overall, RMSEP values ranged from 0.074 to 0.108 MPa, and RMSEP% values ranged from 12.7% to 18.6%.

As was the case for MOR, for comparison, a regression tree model with IB as the response variable was fit to the same data using the CART algorithm of *STATISTICA* 10 (Breiman et al. 1984). The regression tree model can be seen in Figure 9. The

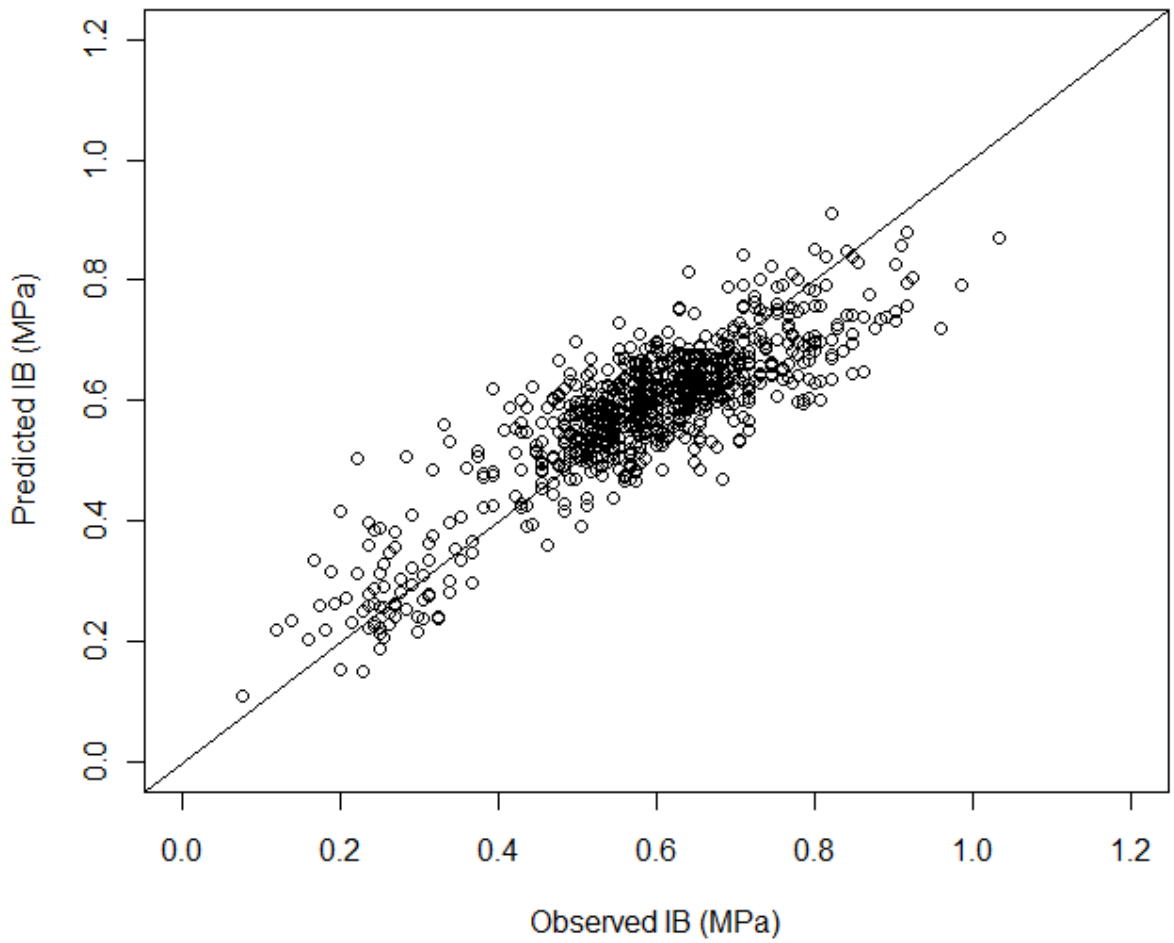


Figure 8: Scatterplot of the observed values versus the predicted values of the validation data set for the BRT model that best predicts IB.

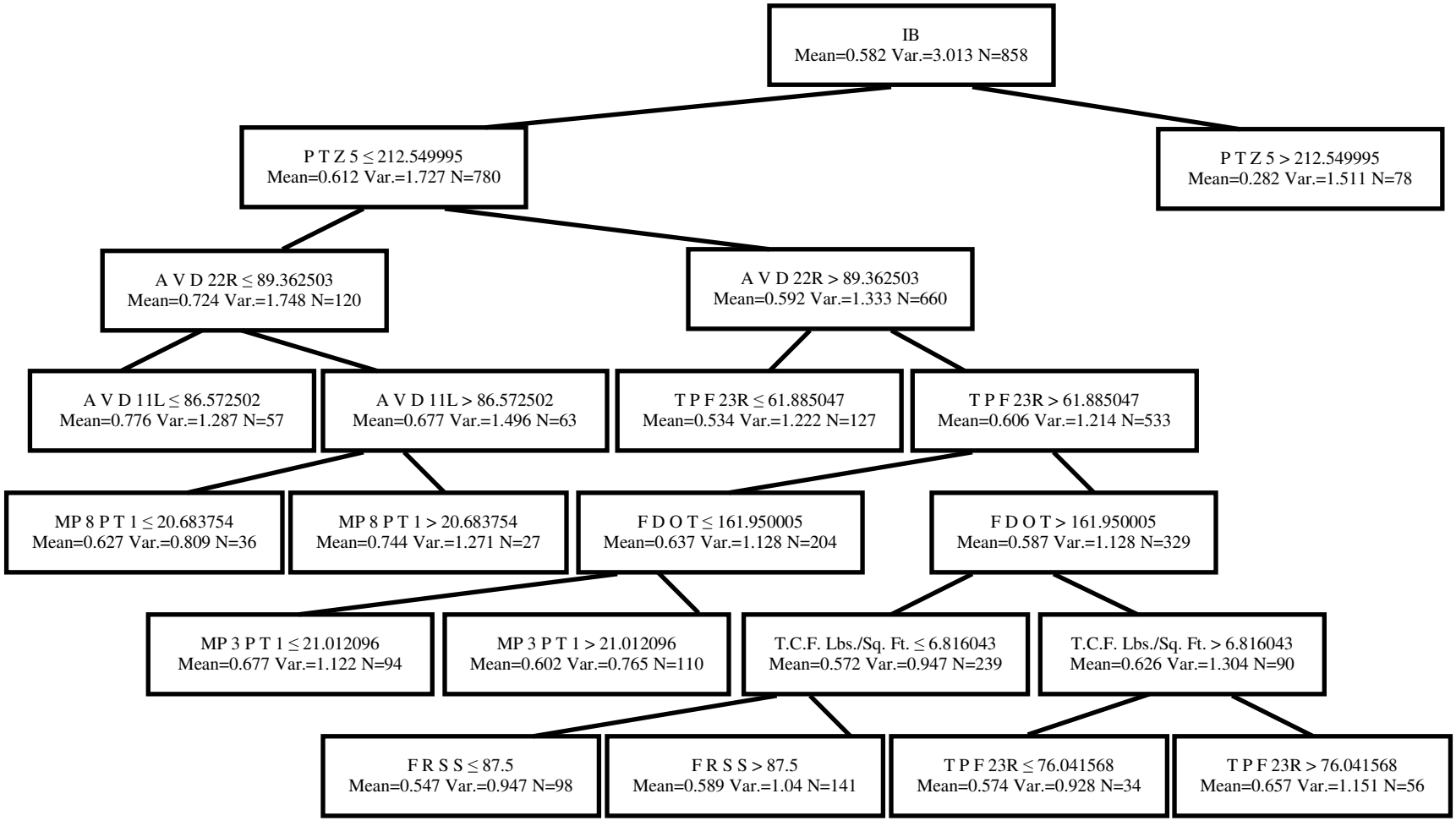


Figure 9: Regression tree model for IB.

RMSEP and RMSEP% values obtained for the validation data set of the regression tree model were 0.085 MPa and 14.6%, respectively. A scatterplot of the observed IB values and the predicted IB values can be seen in Figure 10. The Pearson correlation coefficient value between the observed IB values and the predicted IB values was 0.81. Again, comparing the "observed vs. predicted" scatterplots revealed the predictive modeling weakness of regression tree models (compare Figures 8 and 10). As expected, for the regression tree model, the scatterplot showed "step-like" predictions, and the number of these "steps" had to do with the number of terminal nodes in the tree; whereas, the scatterplot for the BRT model showed more of the desired linear correlation. The number of terminal nodes in the regression tree model for IB was larger than the number in the regression tree model for MOR, but the "step-like" predictions were still evident. The BRT model for IB with the lowest RMSEP and RMSEP% predicted better than the regression tree model for IB on the validation data set.

It was important to examine the top-five predictor importance values in an attempt to understand or interpret the BRT model for IB. For IB, the predictors with the top five importance values were related to particleboard "thickness of pressing." The predictor importance values for these predictor variables ranged from 0.967 to one.

Remarks

Overall, for the parameter settings considered for this BRT analysis, BRT models predicted the MOR measurement more accurately than IB (i.e., the smaller RMSEP% values are associated with BRT models for MOR). For both MOR and IB, the predictive

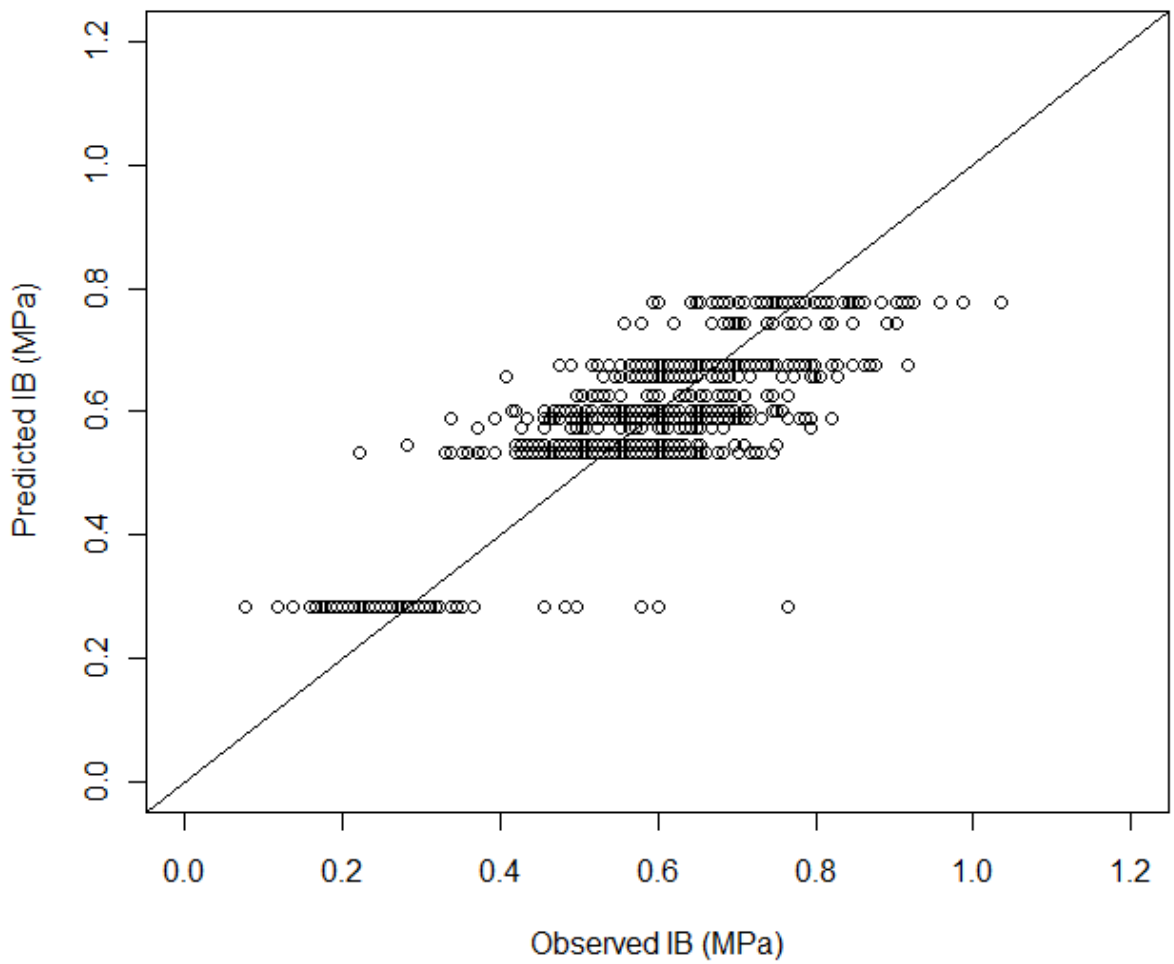


Figure 10: Scatterplot of the observed values versus the predicted values of the validation data set for the regression tree model that predicts IB.

performance of the BRT models was for most parameter settings better than the predictive performance of the standard CART regression tree models. A key finding of this study was the low RMSEP% value obtained on the validation data set when modeling all 118 product types across 16 months of production data (n = 4,307). The research provided within this chapter on the relationships between the BRT parameter settings and model predictive performance will be expanded upon in Chapter V by studying higher values for the *nat* parameter and fewer values for *lr*, but the same values for the *mnn* parameter in the context of comparing non-imputed and imputed data sets.

CHAPTER V. A COMPARISON OF SEVERAL IMPUTATION METHODS USING BOOSTED REGRESSION TREES TO PREDICT STRENGTH PROPERTIES OF WOOD COMPOSITES

In this chapter, a comparison of several different imputation methods will be done using BRT models for predicting MOR and IB. First, the four different data sets used in this study will be discussed. The four different data sets were similar (i.e., provided by the same U.S. particleboard manufacturer, very similar predictor variables, and same March 2009 to June 2010 time frame) to the data set used in Chapter IV, but there were differences (e.g., number of observations and predictor variables) because of the imputation methods used on the data sets. Each of the data sets was provided by Zeng (2011) after the imputation methods were performed to replace missing values. Second, a comparison of the imputation methods using BRT models for predicting MOR and IB was conducted. Third, some general remarks on the analyses are presented.

Data Sets

First, in the “median” imputed data set, missing values in a predictor variable were replaced with the median value of that variable. The data set contained a total of 4,411 observations and a total of 222 continuous predictor variables.

Second, in the “last observation carried forward” imputed (LOCF) data set, missing values in a predictor variable were replaced with the last known value of that variable. One predictor variable was removed entirely because more than 2.5% of the variable’s values were missing from the beginning of the data set. Since no value could be “carried forward” to impute these missing values, they could not be replaced using this method. So, instead of removing all of these observations from the entire data set,

the predictor variable itself was removed. A separate predictor variable had six values that were missing and could not be imputed using this method because no value could be “carried forward” to replace the missing values. But, instead of removing this variable as well, these few observations were deleted from the entire data set. Taking all of this into account, the LOCF data set contained a total of 4,405 observations and a total of 221 continuous predictor variables.

Third, in the “expectation-maximization” imputed (EM) data set, missing values in a predictor variable were replaced using an expectation and maximum likelihood estimation procedure. Initially, the missing values were estimated using the other values of the predictor variable (Zeng 2011). Next, the missing values were processed using maximum-likelihood estimation as though they were complete data (Zeng 2011). This process continued until the change in the estimates for each consecutive iteration did not exceed a convergence criterion developed by Zeng (2011). The EM data set contained a total of 4,411 observations and a total of 222 continuous predictor variables.

Finally, in the “non-imputed” data set, missing values in a predictor variable were not replaced with another value. Again, a total of 25 predictor variables were removed because more than 2.5% of the variable’s values were missing. Even after the deletion of these variables, a total of 278 values were still missing in numerous different predictor variables. These 278 observations were carefully removed from the entire data set. The reason for the deletion of these observations will be discussed after the analysis for MOR and IB is done. After these changes to the data set were completed,

the non-imputed data set contained 4,133 observations and 197 continuous predictor variables.

Lastly, the model comparison techniques used in this chapter were different than the model comparison techniques used in Chapter IV. In this chapter, a repeated random sub-sampling validation technique was used to compare BRT models across imputation methods. In other words, for each imputation method and each set of parameters, three different BRT models for predicting MOR and IB were developed using different randomly selected validation data sets (and training data sets), and a RMSEP and RMSEP% value was calculated for each validation data set. Each BRT model was developed using 80% of the observations for the training data set and 20% of the observations for the validation data set. Since the imputed and non-imputed data sets vary in size, not every validation data set will be the same in size. The random sub-sampling technique should help assure a better comparison between imputation methods using BRT models for MOR and IB.

Imputation Method Results Using BRT Models for MOR

For each of the four different imputation methods used, Table 5 provides RMSEP (and RMSEP%) values obtained on the validation data sets for 144 BRT models predicting MOR. Each of these 144 aforementioned BRT models were obtained using a *mnn* parameter value equal to three, but the values for the *lr* and *nat* parameters were not held constant. The 144 BRT models represented in Table 5 were the product of testing four different levels of *lr* values (0.01, 0.05, 0.1, and 0.15), a *mnn* value equal to three, three different levels of *nat* values (1,000, 1,500, and 2,000), and three different

Table 5: RMSEP (and RMSEP%) values obtained on the validation data sets for each of the four different imputation methods are shown. Each cell contains three RMSEP (and RMSEP%) values obtained on three different validation data sets for BRT models predicting MOR and for the given BRT parameter settings. A value of three for the parameter *mnn* was used.*

lr	***nat***	**Median**	**LOCF**	**EM**	**Non-imputed**
0.01	1,000	1.408 (11.2%), 1.267 (10%), 1.333 (10.6%)	1.443 (11.6%), 1.23 (9.9%), 1.361 (10.9%)	1.405 (11.2%), 1.262 (10%), 1.319 (10.5%)	1.336 (10.6%), 1.197 (9.5%), 1.286 (10.2%)
0.01	1,500	1.364 (10.9%), 1.234 (9.7%), 1.299 (10.3%)	1.394 (11.2%), 1.19 (9.6%), 1.316 (10.5%)	1.364 (10.4%), 1.232 (9.7%), 1.288 (10.2%)	1.3 (10.4%), 1.172 (9.3%), 1.235 (9.8%)
0.01	2,000	1.326 (10.6%), 1.21 (9.5%), 1.273 (10.1%)	1.361 (10.9%), 1.16 (9.3%), 1.28 (10.2%)	1.328 (10.2%), 1.21 (9.5%), 1.264 (10%)	1.276 (10.2%), 1.16 (9.2%), 1.197 (9.5%)
0.05	1,000	1.204 (9.6%), 1.147 (9.1%), 1.187 (9.4%)	1.239 (9.9%), 1.072 (8.6%), 1.17 (9.4%)	1.207 (9.5%), 1.141 (9%), 1.193 (9.5%)	1.196 (9.5%), 1.122 (8.9%), 1.084 (8.6%)
0.05	1,500	1.161 (9.3%), 1.12 (8.8%), 1.168 (9.3%)	1.192 (9.5%), 1.041 (8.4%), 1.132 (9.1%)	1.167 (9.3%), 1.114 (8.8%), 1.173 (9.3%)	1.16 (9.3%), 1.104 (8.7%), 1.035 (8.2%)
0.05	2,000	1.137 (9.1%), 1.099 (8.7%), 1.157 (9.2%)	1.166 (9.3%), 1.023 (8.2%), 1.111 (8.9%)	1.143 (9.1%), 1.098 (8.7%), 1.16 (9.2%)	1.135 (9%), 1.088 (8.6%), 1.009 (8%)
0.1	1,000	1.149 (9.2%), 1.108 (8.7%), 1.155 (9.2%)	1.177 (9.4%), 1.024 (8.2%), 1.113 (8.9%)	1.16 (9%), 1.112 (8.8%), 1.172 (9.3%)	1.128 (9%), 1.101 (8.7%), 1.031 (8.2%)
0.1	1,500	1.13 (9%), 1.091 (8.6%), 1.152 (9.1%)	1.152 (9.2%), 1.005 (8.1%), 1.091 (8.7%)	1.143 (8.9%), 1.091 (8.6%), 1.159 (9.2%)	1.111 (8.9%), 1.087 (8.6%), 1.003 (8%)
0.1	2,000	1.121 (8.9%), 1.085 (8.6%), 1.149 (9.1%)	1.141 (9.1%), 1.001 (8%), 1.079 (8.6%)	1.134 (8.7%), 1.083 (8.5%), 1.153 (9.2%)	1.096 (8.7%), 1.076 (8.5%), 0.997 (7.9%)
0.15	1,000	1.135 (9.1%), 1.113 (8.8%), 1.148 (9.1%)	1.149 (9.2%), 1.019 (8.2%), 1.113 (8.9%)	1.143 (8.9%), 1.122 (8.9%), 1.165 (9.2%)	1.116 (8.9%), 1.081 (8.5%), 1.006 (8%)
0.15	1,500	1.119 (8.9%), 1.107 (8.7%), 1.147 (9.1%)	1.136 (9.1%), 1.013 (8.1%), 1.095 (8.8%)	1.139 (8.9%), 1.111 (8.8%), 1.153 (9.1%)	1.111 (8.9%), 1.077 (8.5%), 0.998 (7.9%)
0.15	2,000	1.117 (8.9%), 1.104 (8.7%), 1.147 (9.1%)	1.132 (9.1%), 1.013 (8.1%), 1.086 (8.7%)	1.139 (8.8%), 1.108 (8.7%), 1.151 (9.1%)	1.106 (8.8%), 1.077 (8.5%), 0.998 (7.9%)

*A total of four different values for the parameter *lr* and three different values for the parameter *nat* are shown here. The RMSEP (and RMSEP%) values provided in Table 5 from top to bottom of each cell are from: (1) the validation data set obtained with the random number generator seed set to one, (2) the validation data set obtained with the random number generator seed set to three, and (3) the validation data set obtained with the random number generator seed set to five.

random number generator seeds (one, three, and five). For MOR and a parameter value of three for *mnn*, the lowest RMSEP value obtained on a validation data set was 0.997 MPa, while the other parameter settings were 0.1 for *lr* and 2,000 for *nat* (refer to Table 5). The RMSEP% value obtained on the validation data set for this model was 7.9% (refer to Table 5). The BRT model with these aforementioned model comparison statistics was developed from the non-imputed data.

Table 6 is similar to Table 5. The only difference is that the RMSEP (and RMSEP%) values provided in Table 6 were obtained using a *mnn* parameter value equal to five. For MOR and a parameter value of five for *mnn*, the lowest RMSEP value obtained on a validation data set was 0.99 MPa, while the other parameter settings were 0.05 for *lr* and 2,000 for *nat* (refer to Table 6). The RMSEP% value obtained on the validation data set for this model was 7.9% (refer to Table 6). Again, the BRT model with these aforementioned model comparison statistics was developed from the non-imputed data.

For each of the four imputation methods, Table 7 provides descriptive statistics on the RMSEP (and RMSEP%) values obtained on the validation data sets for the 288 BRT models (i.e., 144 BRT models when the *mnn* parameter value was equal to three and 144 BRT models when the *mnn* parameter values was equal to five) predicting MOR. Finally, for each of the four imputation methods, a scatterplot of the observed MOR values and the predicted MOR values of the validation data set for the one BRT model that best predicts MOR is given in Figure 11. The BRT model for each imputation method that best predicts MOR can be determined by referring to Table 5

Table 6: RMSEP (and RMSEP%) values obtained on the validation data sets for each of the four different imputation methods are shown. Each cell contains three RMSEP (and RMSEP%) values obtained on three different validation data sets for BRT models predicting MOR and for the given BRT parameter settings. A value of five for the parameter *mnn* was used.*

lr	***nat***	**Median**	**LOCF**	**EM**	**Non-imputed**
0.01	1,000	1.292 (10.3%), 1.188 (9.4%), 1.253 (9.9%)	1.334 (10.7%), 1.146 (9.2%), 1.232 (9.9%)	1.293 (10%), 1.181 (9.3%), 1.255 (10%)	1.253 (10%), 1.145 (9.1%), 1.172 (9.3%)
0.01	1,500	1.24 (9.9%), 1.158 (9.1%), 1.215 (9.6%)	1.278 (10.2%), 1.108 (8.9%), 1.18 (9.4%)	1.242 (9.6%), 1.15 (9.1%), 1.217 (9.7%)	1.206 (9.6%), 1.123 (8.9%), 1.114 (8.9%)
0.01	2,000	1.205 (9.6%), 1.132 (8.9%), 1.188 (9.4%)	1.238 (9.9%), 1.079 (8.7%), 1.146 (9.2%)	1.21 (9.4%), 1.127 (8.9%), 1.19 (9.4%)	1.176 (9.4%), 1.103 (8.7%), 1.077 (8.6%)
0.05	1,000	1.134 (9%), 1.079 (8.5%), 1.13 (9%)	1.153 (9.2%), 1.02 (8.2%), 1.076 (8.6%)	1.136 (8.8%), 1.073 (8.5%), 1.146 (9.1%)	1.099 (8.8%), 1.066 (8.4%), 1.01 (8%)
0.05	1,500	1.122 (9%), 1.066 (8.4%), 1.119 (8.9%)	1.129 (9%), 1.008 (8.1%), 1.062 (8.5%)	1.123 (8.6%), 1.063 (8.4%), 1.136 (9%)	1.083 (8.6%), 1.049 (8.3%), 1 (8%)
0.05	2,000	1.115 (8.9%), 1.059 (8.4%), 1.117 (8.9%)	1.11 (8.9%), 1.007 (8.1%), 1.056 (8.4%)	1.113 (8.5%), 1.058 (8.3%), 1.13 (9%)	1.069 (8.5%), 1.044 (8.2%), 0.99 (7.9%)
0.1	1,000	1.114 (8.9%), 1.055 (8.3%), 1.107 (8.8%)	1.14 (9.1%), 1.008 (8.1%), 1.053 (8.4%)	1.121 (8.6%), 1.063 (8.4%), 1.119 (8.9%)	1.079 (8.6%), 1.046 (8.3%), 0.999 (8%)
0.1	1,500	1.114 (8.9%), 1.054 (8.3%), 1.103 (8.8%)	1.123 (9%), 1.008 (8.1%), 1.052 (8.4%)	1.116 (8.6%), 1.059 (8.4%), 1.113 (8.8%)	1.075 (8.6%), 1.045 (8.3%), 0.999 (8%)
0.1	2,000	1.114 (8.9%), 1.054 (8.3%), 1.099 (8.7%)	1.117 (8.9%), 1.008 (8.1%), 1.051 (8.4%)	1.114 (8.6%), 1.059 (8.4%), 1.109 (8.8%)	1.074 (8.6%), 1.038 (8.2%), 0.999 (8%)
0.15	1,000	1.102 (8.8%), 1.082 (8.5%), 1.124 (8.9%)	1.13 (9%), 1.018 (8.2%), 1.076 (8.6%)	1.137 (8.8%), 1.077 (8.5%), 1.15 (9.1%)	1.103 (8.8%), 1.051 (8.3%), 0.996 (7.9%)
0.15	1,500	1.102 (8.8%), 1.079 (8.5%), 1.12 (8.9%)	1.13 (9%), 1.018 (8.2%), 1.076 (8.6%)	1.137 (8.8%), 1.077 (8.5%), 1.144 (9.1%)	1.103 (8.8%), 1.051 (8.3%), 0.996 (7.9%)
0.15	2,000	1.102 (8.8%), 1.079 (8.5%), 1.115 (8.8%)	1.13 (9%), 1.018 (8.2%), 1.076 (8.6%)	1.137 (8.8%), 1.077 (8.5%), 1.141 (9.1%)	1.103 (8.8%), 1.049 (8.3%), 0.996 (7.9%)

*A total of four different values for the parameter *lr* and three different values for the parameter *nat* are shown here. The RMSEP (and RMSEP%) values provided in Table 6 from top to bottom of each cell are from: (1) the validation data set obtained with the random number generator seed set to one, (2) the validation data set obtained with the random number generator seed set to three, and (3) the validation data set obtained with the random number generator seed set to five.

Table 7: Descriptive statistics on the RMSEP (and RMSEP%) values obtained on the validation data sets for the 288 BRT models (i.e., 144 BRT models when the *mnn* parameter was equal to three and 144 BRT models when the *mnn* parameter was equal to five) predicting MOR.

Statistic	Median	LOCF	EM	Non-imputed
Minimum	1.054 (8.3%)	1.001 (8%)	1.058 (8.3%)	0.99 (7.9%)
Maximum	1.408 (11.2%)	1.443 (11.6%)	1.405 (11.2%)	1.336 (10.6%)
Median	1.127 (9%)	1.113 (8.9%)	1.141 (9%)	1.083 (8.6%)
Mean	1.151 (9.1%)	1.125 (9%)	1.157 (9.1%)	1.095 (8.7%)
Standard Deviation	0.075 (0.61%)	0.102 (0.81%)	0.072 (0.57%)	0.081 (0.65%)

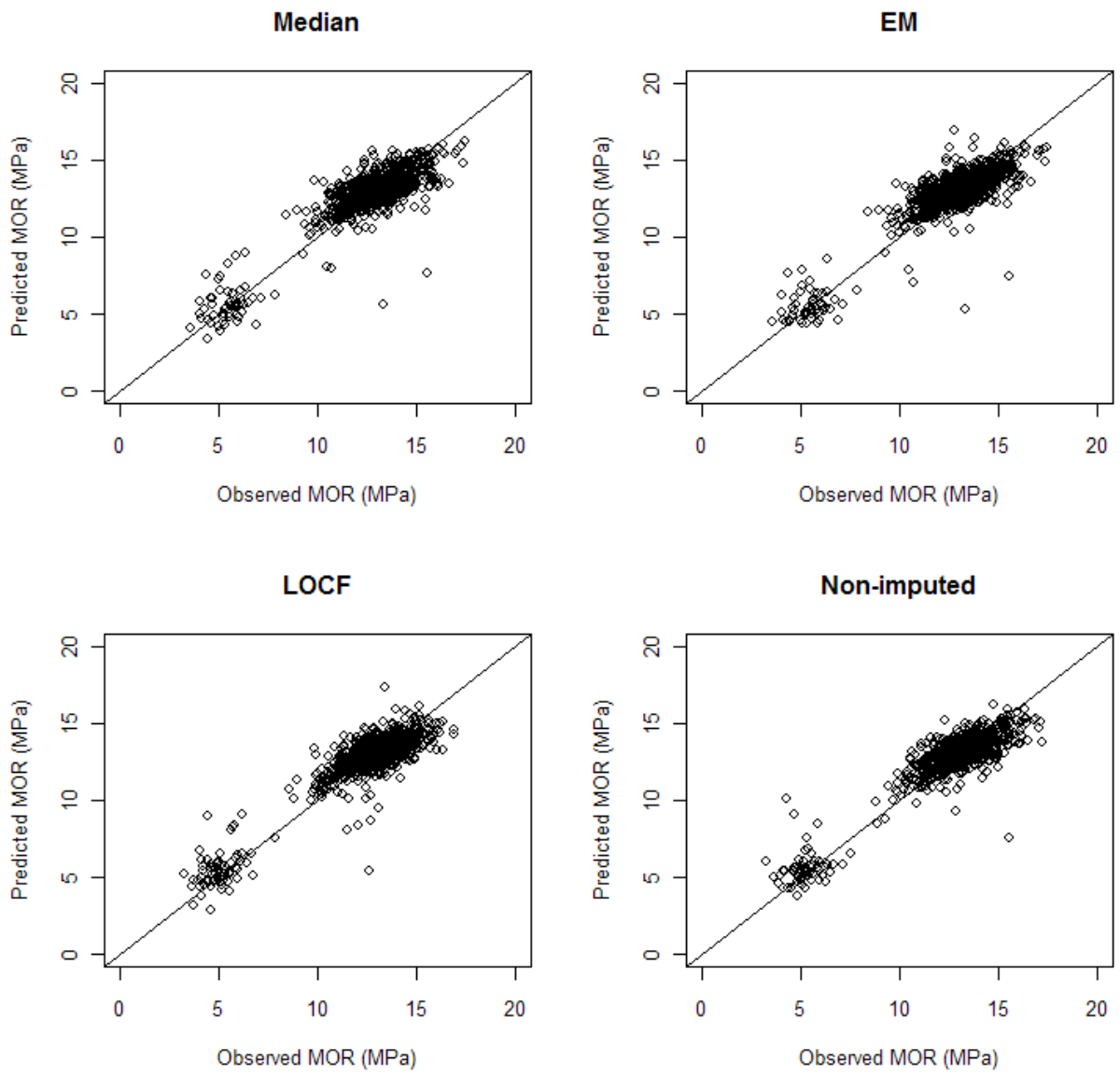


Figure 11: For each imputation method, a scatterplot of the observed MOR values and the predicted MOR values of the validation data set for the BRT model that best predicts MOR.

and/or Table 6. The Pearson correlation coefficient values between the observed MOR values and the predicted MOR values for the four imputation methods Median, LOCF, EM, and Non-imputed, are 0.88, 0.91, 0.88, and 0.92, respectively.

Imputation Method Results Using BRT Models for IB

For each of the four different imputation methods used, Table 8 provides RMSEP (and RMSEP%) values obtained on the validation data sets for 144 BRT models predicting IB. Each of these 144 aforementioned BRT models were obtained using a *mnn* parameter value equal to three, but the values for the *lr* and *nat* parameters were not held constant. The 144 BRT models represented in Table 8 were the product of testing four different levels of *lr* values (0.01, 0.05, 0.1, and 0.15), a *mnn* value equal to three, three different levels of *nat* values (1,000, 1,500, and 2,000), and three different random number generator seeds (one, three, and five). For IB and a parameter value of three for *mnn*, the lowest RMSEP value obtained on a validation data set was 0.081 MPa, while the other parameter settings were 0.15 for *lr* and 2,000 for *nat* (refer to Table 8). The RMSEP% value obtained on the validation data set for this model was 13.7% (refer to Table 8). The BRT model with these aforementioned model comparison statistics was developed from the non-imputed data. Importantly, when the *mnn* parameter value was equal to three, several BRT models obtained this low RMSEP value of 0.081, but the aforementioned model obtained the lowest RMSEP% value on a validation data set.

As was the case for MOR, Table 9 is similar to Table 8. The only difference is that the RMSEP (and RMSEP%) values provided in Table 9 were obtained using a *mnn*

Table 8: RMSEP (and RMSEP%) values obtained on the validation data sets for each of the four different imputation methods are shown. Each cell contains three RMSEP (and RMSEP%) values obtained on three different validation data sets for BRT models predicting IB and for the given BRT parameter settings. A value of three for the parameter *mnn* was used.*

lr	***nat***	**Median**	**LOCF**	**EM**	**Non-imputed**
0.01	1,000	0.098 (16.6%), 0.092 (15.5%), 0.099 (16.9%)	0.1 (17.1%), 0.095 (16.4%), 0.092 (15.9%)	0.098 (16.6%), 0.092 (15.5%), 0.098 (16.8%)	0.096 (16.6%), 0.091 (15.6%), 0.099 (17%)
0.01	1,500	0.096 (16.3%), 0.09 (15.2%), 0.097 (16.3%)	0.097 (16.6%), 0.093 (16%), 0.09 (15.5%)	0.096 (16.3%), 0.09 (15.1%), 0.096 (16.4%)	0.094 (16.2%), 0.089 (15.2%), 0.097 (16.6%)
0.01	2,000	0.094 (16%), 0.089 (15%), 0.095 (16.1%)	0.095 (16.3%), 0.092 (15.8%), 0.088 (15.3%)	0.094 (16%), 0.088 (14.9%), 0.095 (16.1%)	0.093 (16%), 0.088 (15%), 0.095 (16.3%)
0.05	1,000	0.09 (15.2%), 0.085 (14.4%), 0.091 (15.3%)	0.089 (15.2%), 0.087 (14.9%), 0.084 (14.6%)	0.089 (15.2%), 0.084 (14.2%), 0.09 (15.3%)	0.088 (15.1%), 0.084 (14.3%), 0.091 (15.5%)
0.05	1,500	0.088 (15%), 0.084 (14.2%), 0.089 (15%)	0.087 (14.8%), 0.085 (14.6%), 0.082 (14.2%)	0.088 (15%), 0.083 (14%), 0.088 (15.1%)	0.086 (14.8%), 0.082 (14%), 0.088 (15.2%)
0.05	2,000	0.088 (14.9%), 0.083 (14%), 0.088 (14.8%)	0.085 (14.6%), 0.084 (14.5%), 0.081 (14.1%)	0.088 (14.9%), 0.083 (13.9%), 0.088 (14.9%)	0.085 (14.7%), 0.081 (13.8%), 0.087 (15%)
0.1	1,000	0.088 (14.9%), 0.084 (14.2%), 0.088 (14.8%)	0.085 (14.6%), 0.085 (14.7%), 0.082 (14.2%)	0.088 (14.9%), 0.083 (14%), 0.088 (15%)	0.086 (14.8%), 0.082 (13.9%), 0.089 (15.2%)
0.1	1,500	0.088 (14.9%), 0.083 (14%), 0.088 (14.7%)	0.084 (14.3%), 0.085 (14.6%), 0.081 (14%)	0.088 (14.9%), 0.083 (13.9%), 0.088 (14.9%)	0.085 (14.6%), 0.081 (13.9%), 0.087 (14.9%)
0.1	2,000	0.087 (14.8%), 0.083 (14%), 0.087 (14.7%)	0.084 (14.3%), 0.084 (14.5%), 0.081 (14%)	0.087 (14.9%), 0.083 (13.9%), 0.087 (14.9%)	0.085 (14.6%), 0.081 (13.8%), 0.087 (14.9%)
0.15	1,000	0.087 (14.8%), 0.084 (14.1%), 0.088 (14.8%)	0.084 (14.3%), 0.084 (14.5%), 0.081 (14.1%)	0.088 (14.9%), 0.082 (13.8%), 0.087 (14.8%)	0.087 (15.1%), 0.081 (13.8%), 0.087 (14.9%)
0.15	1,500	0.087 (14.8%), 0.084 (14.1%), 0.088 (14.8%)	0.084 (14.3%), 0.084 (14.5%), 0.081 (14.1%)	0.088 (14.9%), 0.082 (13.8%), 0.087 (14.8%)	0.087 (15.1%), 0.081 (13.8%), 0.086 (14.8%)
0.15	2,000	0.087 (14.8%), 0.084 (14.1%), 0.088 (14.8%)	0.084 (14.3%), 0.084 (14.5%), 0.081 (14.1%)	0.088 (14.9%), 0.082 (13.8%), 0.087 (14.8%)	0.087 (15.1%), 0.081 (13.7%), 0.086 (14.8%)

*A total of four different values for the parameter *lr* and three different values for the parameter *nat* are shown here. The RMSEP (and RMSEP%) values provided in Table 8 from top to bottom of each cell are from: (1) the validation data set obtained with the random number generator seed set to one, (2) the validation data set obtained with the random number generator seed set to three, and (3) the validation data set obtained with the random number generator seed set to five.

Table 9: RMSEP (and RMSEP%) values obtained on the validation data sets for each of the four different imputation methods are shown. Each cell contains three RMSEP (and RMSEP%) values obtained on three different validation data sets for BRT models predicting IB and for the given BRT parameter settings. A value of five for the parameter *mnn* was used.*

lr	***nat***	**Median**	**LOCF**	**EM**	**Non-imputed**
0.01	1,000	0.094 (15.9%), 0.088 (14.9%), 0.095 (16.1%)	0.096 (16.3%), 0.092 (15.8%), 0.09 (15.6%)	0.094 (15.9%), 0.088 (14.9%), 0.095 (16.2%)	0.094 (16.2%), 0.089 (15.1%), 0.094 (16.1%)
0.01	1,500	0.091 (15.5%), 0.086 (14.5%), 0.093 (15.7%)	0.093 (15.8%), 0.089 (15.3%), 0.087 (15%)	0.091 (15.5%), 0.086 (14.5%), 0.093 (15.8%)	0.091 (15.7%), 0.087 (14.8%), 0.091 (15.7%)
0.01	2,000	0.09 (15.3%), 0.085 (14.2%), 0.091 (15.4%)	0.091 (15.4%), 0.087 (14.9%), 0.084 (14.6%)	0.089 (15.2%), 0.084 (14.2%), 0.091 (15.5%)	0.089 (15.4%), 0.085 (14.5%), 0.09 (15.4%)
0.05	1,000	0.086 (14.7%), 0.083 (13.9%), 0.087 (14.7%)	0.085 (14.5%), 0.081 (14%), 0.081 (14.1%)	0.087 (14.8%), 0.081 (13.7%), 0.087 (14.8%)	0.085 (14.6%), 0.081 (13.8%), 0.086 (14.7%)
0.05	1,500	0.086 (14.5%), 0.082 (13.8%), 0.087 (14.6%)	0.083 (14.2%), 0.08 (13.8%), 0.08 (13.9%)	0.087 (14.7%), 0.08 (13.5%), 0.086 (14.7%)	0.083 (14.3%), 0.08 (13.6%), 0.085 (14.5%)
0.05	2,000	0.085 (14.5%), 0.081 (13.7%), 0.086 (14.5%)	0.083 (14.1%), 0.079 (13.7%), 0.08 (13.9%)	0.086 (14.6%), 0.08 (13.5%), 0.086 (14.7%)	0.082 (14.2%), 0.08 (13.6%), 0.085 (14.5%)
0.1	1,000	0.087 (14.8%), 0.082 (13.8%), 0.087 (14.6%)	0.084 (14.3%), 0.081 (13.9%), 0.08 (13.9%)	0.086 (14.7%), 0.081 (13.7%), 0.087 (14.8%)	0.082 (14%), 0.081 (13.8%), 0.085 (14.6%)
0.1	1,500	0.087 (14.7%), 0.082 (13.8%), 0.086 (14.5%)	0.083 (14.2%), 0.08 (13.8%), 0.08 (13.9%)	0.086 (14.7%), 0.081 (13.7%), 0.087 (14.8%)	0.081 (14%), 0.081 (13.7%), 0.085 (14.5%)
0.1	2,000	0.087 (14.7%), 0.082 (13.8%), 0.086 (14.5%)	0.083 (14.2%), 0.08 (13.8%), 0.08 (13.9%)	0.086 (14.7%), 0.081 (13.7%), 0.087 (14.8%)	0.081 (14%), 0.081 (13.7%), 0.085 (14.5%)
0.15	1,000	0.087 (14.8%), 0.083 (14%), 0.087 (14.7%)	0.084 (14.3%), 0.08 (13.8%), 0.081 (14%)	0.086 (14.6%), 0.084 (14.2%), 0.087 (14.9%)	0.085 (14.6%), 0.081 (13.8%), 0.084 (14.4%)
0.15	1,500	0.087 (14.8%), 0.083 (14%), 0.087 (14.7%)	0.083 (14.2%), 0.08 (13.7%), 0.081 (14%)	0.086 (14.6%), 0.084 (14.2%), 0.087 (14.9%)	0.085 (14.6%), 0.081 (13.8%), 0.084 (14.4%)
0.15	2,000	0.087 (14.8%), 0.083 (14%), 0.087 (14.7%)	0.083 (14.2%), 0.08 (13.7%), 0.081 (14%)	0.086 (14.6%), 0.084 (14.2%), 0.087 (14.9%)	0.085 (14.6%), 0.081 (13.8%), 0.084 (14.4%)

*A total of four different values for the parameter *lr* and three different values for the parameter *nat* are shown here. The RMSEP (and RMSEP%) values provided in Table 9 from top to bottom of each cell are from: (1) the validation data set obtained with the random number generator seed set to one, (2) the validation data set obtained with the random number generator seed set to three, and (3) the validation data set obtained with the random number generator seed set to five.

parameter value equal to five. For IB and a parameter value of five for *mnn*, the lowest RMSEP value obtained on a validation data set was 0.079 MPa, while the other parameter settings were 0.05 for *lr* and 2,000 for *nat* (refer to Table 9). The RMSEP% value obtained on the validation data set for this model was 13.7% (refer to Table 9). Again, the BRT model with these aforementioned model comparison statistics was developed from the LOCF data.

For each of the four imputation methods, Table 10 provides descriptive statistics on the RMSEP (and RMSEP%) values obtained on the validation data sets for the 288 BRT models (i.e., 144 BRT models when the *mnn* parameter value was equal to three and 144 BRT models when the *mnn* parameter values was equal to five) predicting IB. Finally, for each of the four imputation methods, a scatterplot of the observed IB values and the predicted IB values of the validation data set for the one BRT model that best predicts IB can be seen in Figure 12. The BRT model for each imputation method that best predicts IB can be determined by referring to Table 8 and/or Table 9. The Pearson correlation coefficient values between the observed IB values and the predicted IB values for the four imputation methods Median, LOCF, EM, and Non-imputed, are 0.8, 0.83, 0.81, and 0.81, respectively.

Remarks

Not surprisingly, across the different imputation methods (“median,” “expectation-maximization,” “last observation carried forward,” and “non-imputation”), BRT models predicted the MOR measurement more accurately than IB. Again, the smaller RMSEP% values were associated with BRT models for MOR. Given the information

Table 10: Descriptive statistics on the RMSEP (and RMSEP%) values obtained on the validation data sets for the 288 BRT models (i.e., 144 BRT models when the *mnn* parameter was equal to three and 144 BRT models when the *mnn* parameter was equal to five) predicting IB.

Statistic	Median	LOCF	EM	Non-imputed
Minimum	0.081 (13.7%)	0.079 (13.7%)	0.08 (13.5%)	0.08 (13.6%)
Maximum	0.099 (16.9%)	0.1 (17.1%)	0.098 (16.8%)	0.099 (17%)
Median	0.087 (14.8%)	0.084 (14.3%)	0.087 (14.8%)	0.085 (14.6%)
Mean	0.088 (14.8%)	0.085 (14.6%)	0.087 (14.8%)	0.086 (14.7%)
Standard Deviation	0.004 (0.7%)	0.005 (0.79%)	0.004 (0.74%)	0.004 (0.81%)

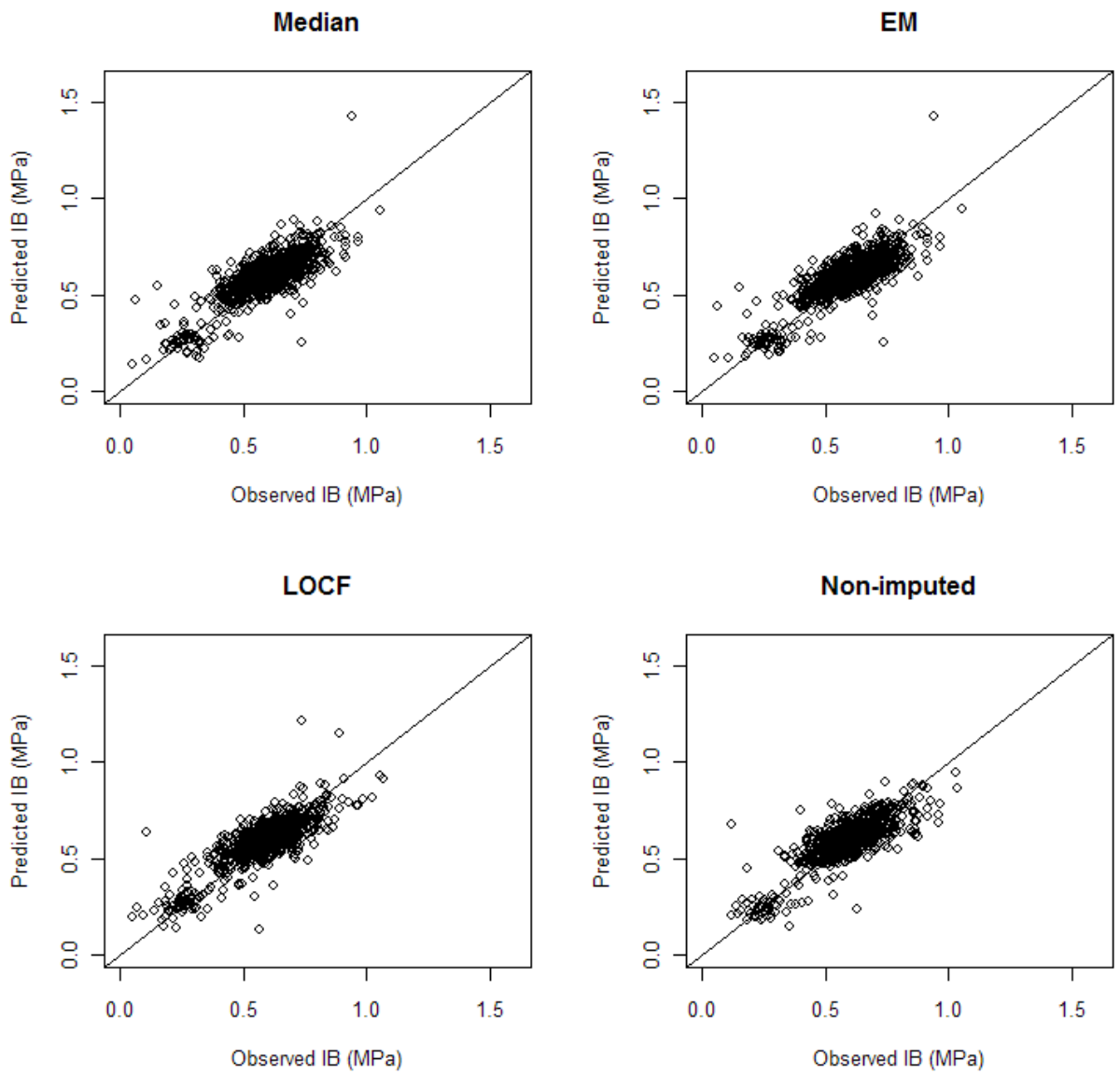


Figure 12: For each imputation method, a scatterplot of the observed IB values and the predicted IB values of the validation data set for the BRT model that best predicts IB.

learned in Chapter IV, allowed for an even more in depth study in this chapter on the relationship between BRT parameter settings and model predictive performance.

A key finding of this study was the minimal effect on BRT model prediction due to a loss of information. This may be due to the methodology of BRT. The BRT methodology will select a relatively small sample of weak learner predictor variables. If the weak learner predictor variables selected in the BRT model do not have many missing values capable of being imputed, there would be small differences in the outcomes of RMSEP% values among the imputed and non-imputed methods. Further research is required in this area, which is beyond the scope of this thesis.

CHAPTER VI. CONCLUDING REMARKS AND FUTURE RESEARCH

The use of boosted decision trees for predictive modeling started just over a decade ago. More wide-spread use of boosted trees for predictive modeling has occurred within the past decade. Boosted regression trees (BRT) are a predictive modeling technique that draw on insights and techniques from both statistical and machine learning traditions which combine boosting algorithms with regression tree methods. BRT models have the ability to select pertinent variables, fit accurate functions, and model interactions. In the study conducted for this thesis, the boosting technique enhances the predictive performance of regression trees for predicted MOR and IB of particleboard by combining a series of small regression trees in a stochastic-based gradient descent method.

A major challenge for engineered wood products manufacturers is developing better knowledge of the complex nature of process variables (e.g., line speed, press temperature, etc.) and their relationship with engineered wood product strength properties (e.g., internal bond (IB), modulus of rupture (MOR), etc.). Also, accurate real-time prediction of strength properties between long time periods associated with destructive tests would create opportunities for improved business competitiveness. Accurate real-time prediction of strength properties would prevent the production of defective or off-grade product between destructive test samples and would avert unwarranted high operating targets for wood and resin addition. Improved production efficiency may also be possible with accurate real-time prediction of strength properties. The research of this thesis has shown that accurate real-time predictions of strength

properties of particleboard are feasible with the use of BRT. This thesis is the first known published work for using BRT methods to predict the strength properties of wood composites. BRT models had far superior predictive quality of wood composites strength properties in validation when compared to regression tree models. This research advances the area of study that models real-time strength properties of wood composites. A significant contribution is the development of one BRT model for all product types. In this study, the data set contained 118 different product types for particleboard. Previous research (see André et al. 2008; Clapp et al. 2008) developed separate models for nominally manufactured product types (e.g., ¾'', 5/8'', etc.) which left a gap in predictive models for a large number of other product types.

Specifically, this thesis documented the development of a total of 140 different BRT models for both MOR and IB of particleboard. For MOR, the RMSEP values ranged from 1.051 to 1.443 MPa, and RMSEP% values ranged from 8.5% to 11.6%. For IB, RMSEP values ranged from 0.074 to 0.108 MPa, and RMSEP% values ranged from 12.7% to 18.6%. For MOR, the five most important process predictor variables were related to "pressing temperature zones," "thickness of pressing," and "pressing pressure." For IB, the five most important process predictor variables were related to "thickness of pressing."

The thesis research also examined the effect of missing values in the data set on BRT model predictive quality. Missing values were a common problem with the industrial data set use in this study, which is not atypical for any other industrial data set. Three different imputation methods ("median," "expectation-maximization," and

“last observation carried forward”) were used and the BRT model predictions for MOR and IB were compared with the BRT model predictions for a non-imputed data set. For both MOR and IB, 72 BRT models were developed for each of the four imputation methods. BRT model performance using RMSEP and RMSEP% was determined for three different validation data sets for each set of the BRT parameter settings *lr*, *nat*, and *mnn*. For MOR, the best BRT model for the non-imputed data set had RSMEP and RMSEP% values of 0.99 MPa and 7.9%, respectively. In comparison, the best BRT model for MOR for an imputed data set using “last observation carried forward” imputation had RMSEP and RMSEP% values of 1.001 MPa and 8%, respectively. For IB, the best BRT model for the non-imputed data set had RMSEP and RMSEP% values of 0.08 MPa and 13.6%, respectively. In comparison, the best BRT model for an imputed data set using “last observation carried forward” imputation had RMSEP and RMSEP% values of 0.079 MPa and 13.7%, respectively. The approximate 270 observations lacking in the non-imputed data set as compared to the other three imputed data sets did not appear to affect the predictive ability of BRT models. This may be due to the methodology of BRT. The BRT methodology will select a relatively small sample of weak learner predictive variables using a series of consecutive three or five node trees as it builds an overall model. If the predictor variables selected in the BRT models do not have many missing values, there would be small differences in the predictive quality between imputed and non-imputed data sets.

An advantage of using BRT models is that the technique may abstain from selecting predictor variables with large numbers of missing values, which would make

the technique robust and applicable for industrial applications. Another advantage of the BRT models when compared to regression tree models is the improved predictive performance in validation.

A possible disadvantage of using BRT for predictive models is difficult interpretation of the BRT model itself and the significant predictor variables. Specifically indentifying the significant predictor variables of a process and the magnitude of influence of such variables on strength properties may limit the usefulness of the BRT technique for root cause analysis in continuous improvement efforts. Another disadvantage of using BRT for predictive modeling is that building a BRT model can be computationally time consuming depending on the power of the computer being used to build the BRT models. In this thesis study the average CPU time for one BRT model using a HP dual-processor Power Edge was approximately seven minutes. As time progresses and the power of computers (even personal computers) grow, this may reduce this disadvantage in the use of BRT models.

Suggestions for future research on the use of BRT models for predicting the strength properties of wood composites are as follows. First, the relationship between BRT parameter settings and model predictive performance should be studied in an effort to hopefully determine parameter settings that will yield an optimal model, or close to it, for MOR and IB. Second, the effect of information loss on the predictive performance of BRT models requires further investigation. Thesis results suggest BRT models may be robust to information loss given the specifics of the methodology in

selecting weak learner predictor variables. Third, BRT predictive modeling of MOR, IB, and other strength properties, should be tested at a mill site in a real-time setting.

LIST OF REFERENCES

Abeare SM (2009) Comparisons of boosted regression tree, GLM and GAM performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico Lonline Fishery. Master's Thesis, Louisiana State University

American Forest and Paper Association (2010) Forest Products Industry Technology Roadmap. http://www.agenda2020.org/PDF/ForestProductsIndustryTechRM-043010.pdf. Accessed 16 March 2011

André N, Cho H-W, Baek SH, Jeong M-K, Young TM (2008) Enhanced prediction of internal bond strength in a medium density fiberboard process using multivariate statistical methods and variable selection. Wood Sci Technol 42(7):521-534

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Bernardy G, Scherff B (1998) Saving costs with process control engineering and statistical process optimization: uses for production managers, technologists and operators. In: Proceedings of the second European panel products symposium (EPPS), pp 95-106

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International Group, Belmont

Breiman L (1998) Arcing classifiers. Ann Stat 26(3):801-824

Breiman L (1999) Prediction games and arcing algorithms. Neural Comp 11(7):1493-1517

Carslaw DC, Taylor PJ (2009) Analysis of air pollution data at a mixed source location using boosted regression trees. Atmospheric Environ 43(22-23):3563-3570

Clapp NE Jr, Young TM, Guess FM (2008) Predictive modeling the internal bond of medium density fiberboard using a modified principal components analysis. For Prod J 58(4):49-55

Cook DF, Chiu CC (1997) Predicting the internal bond strength of particleboard utilizing a radial basis function neural network. Eng Appl Artif Intell 10(2):171-177

Cook DF, Ragsdale CT, Major RL (2000) Combining a neural network with a genetic algorithm for process parameter optimization. Eng Appl Artif Intell 13:391-396

De'ath G (2007) Boosted trees for ecological modeling and prediction. Ecol 88:243-251

Deconinck E, Zhang MH, Coomans D, Heyden YV (2007) Evaluation of boosted regression trees (BRTs) and two-step BRT procedures to model and predict blood-brain barrier passage. J Chemom 21:280-291

Drucker H (1997) Improving regressors using boosting techniques. In: Fisher DH Jr (Ed.) Proceedings of the Fourteenth International Conference on Machine Learning, pp 107-115

Drucker H, Schapire RE, Simard P (1993) Boosting performance in neural networks. Int J Pattern Recognit Artif Intell 7(4):705-719

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802-813

Enders CK (2001) Impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. Psychological Methods 6:352–370

Erlisson L, Hagberg P, Johansson E, Rannar S, Whelehan O, Astrom A, Lindgren T (2000) Multivariate process monitoring of a newsprint mill. Application to modeling and predicting COD load resulting from de-inking of recycled paper. J Chemom 15:337-352

Faraway JJ (2005) Linear models with R. Chapman and Hall, New York

Fielding A (1977) Binary segmentation: the automatic interaction detector and related techniques for exploring data structure. In: O'Muircheartaigh CA, Payne C (Eds.) The analysis of survey data, Vol. 1, exploring data structures. John Wiley and Sons, New York

Freund Y (1995) Boosting a weak learning algorithm by majority. Inf Comput 121(2):256-285

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Machine Learning: Proceedings of the Thirteenth International Conference, pp 325-332

Freund Y, Schapire RE (1997) A decision-theoretic generalizat of on-line learning and an application to boosting. J Comp Syst Sci 55(1):119-139

Freund Y, Schapire RE (1999) Adaptive game playing using multiplicative weights. Games Econ Behav 29:79-103

Freund Y, Iyer RD, Schapire RE, Singer Y (2004) An efficient boosting algorithm for combining preferences. J Mach Learn Res 4(6):933-969

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189-1232

Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38(4):367-378

Friedman JH, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. Ann Stat 28(2):337-407

Gamage N (2007) Economical particleboard production using hardwood sawmill residues. Dissertation, RMIT University

Gelman A, Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge

Giudici P (2003) Applied data mining: statistical methods for business and industry. John Wiley and Sons, West Sussex

Greubel D (1999) Practical experiences with a process simulation model in particleboard and MDF production. In: Proceedings of the second European wood-based panel symposium, pp 8-10

Hamer RM (2009) Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials. Am J Psychiatry 166:639-641

Hand DJ, Mannila H, Smyth P (2001) Principles of data mining. MIT Press, Cambridge

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

Horton NJ, Kleinman KP (2007) Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Stat 61(1):79-90.

Iyer RD, Lewis DD, Schapire RE, Singer Y, Singhal A (2000) Boosting for document routing. In: Proceedings of the Ninth International Conference on Information and Knowledge Management, pp 70-77

Kass G (1975) Significance testing in automatic interaction detection (AID). Appl Stat 24(2):178-189

Kearns M, Valiant LG (1994) Cryptographic limitations on learning Boolean formulae and finite automata. J ACM 41(1):67-95

Kearns M, Vazirani U (1994) An introduction to computational learning theory. MIT Press, Cambridge

Lei YC, Zhang SY, Jiang ZH (2005) Models for predicting lumber bending MOR and MOE based on tree and stand characteristics in black spruce. Wood Sci Technol 39(1):37-47

LeVan-Green SL, Livingston J (2001) Exploring the uses for small-diameter trees. For Prod J 51(9):10-21

Li B, Duan L, Peng L (2010) Efficient microarchitectural vulnerabilities prediction using boosted regression trees and patient rule inductions. IEEE Transactions Comp 59(5):593-607

Little RJA (1992) Regression with missing X's: a review. J Am Stat Association 87:1227-1237

Loh W-Y (2002) Regression trees with unbiased variable selection and interaction detection. Stat Sin 12(2):361-386

Loh W-Y (2008) Classification and regression tree methods. In: Ruggeri F, Kenett R, Faltin FW (Eds.) Encyclopedia of statistics in quality and reliability. John Wiley and Sons, New York, pp. 315-323

Maloney TM (1996) The family of wood composite materials. For Prod J 46(2):19-26

Merler S, Furlanello C, Larcher B, Sboner A (2001) Tuning cost-sensitive boosting and its application to melanoma diagnosis. In: Multiple Classifier Systems: Proceedings of the 2nd International Workshop, pp 32-42

Mora CR, Schimleck LR (2010) Kernel regression methods for the prediction of wood properties of Pinus taeda using near infrared spectroscopy. Wood Sci Technol 44(4):561-578

Morgan J, Sonquist J (1963) Problems in the analysis of survey data, and a proposal. J Am Stat Assoc 58(302):415-434

Olshen RA (2001) A conversation with Leo Breiman. Stat Sci 16:184-198

Quinlan R (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo

Ridgeway G, Madigan D, Rishardson T (1999) Boosting methodology for regression problems. In: Heckerman D, Whittaker J (Eds.) Proceedings of Artificial Intelligence and Statistics, pp 152-161

Robinson JW (2008) Regression tree boosting to adjust health care cost predictions for diagnostic mix. Health Serv Res 43(2):755-772

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman and Hall, London

Schapire RE (1990) The strength of weak learnability. Mach Learn 5(2):197-227

Schapire RE (2003) The boosting approach to machine learning: an overview. In: Denison DD, Hansen MH, Holmes C, Mallick B, Yu B (eds.) MSRI workshop on nonlinear estimation and classification, 2002. Springer, New York

Schapire RE, Singer Y (1999) Improved boosting algorithms using confidence-rated predictions. Mach Learn 37(3):297-336

Schapire RE, Singer Y (2000) BoosTexter: a boosting-based system for text categorization. Mach Learn 39(2-3):135-168

Schapire RE, Singer Y, Singhal A (1998) Boosting and Rocchio applied to text filtering. In: Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval

Sutton CD (2005) Classification and regression trees, bagging, and boosting. In: Rao CR, Wegman EJ, Solka JL (eds.) Handbook of statistics: data mining and data visualization. Elsevier, Amsterdam

Truxillo C (2005) Maximum likelihood parameter estimation with incomplete data. In: Proceedings of the Thirtieth Annual SAS(R) User Group International Conference. http://www2.sas.com/proceedings/sugi30/111-30.pdf accessed on 12 July 2011

Valiant LG (1984) A theory of the learnable. Commun ACM 27:1134-1142

Wood Handbook (1999) Wood as an engineering material. Forest Products Laboratory, U.S. Department of Agriculture, Pullman, U.S.A

Xing C, Zhang SY, Deng J, Wang SQ (2007) Investigation of the effects of bark fiber as core material and its resin content on three-layer MDF performance by response surface methodology. Wood Sci Technol 41(7):585-595

Young TM (1996) Process improvement through “real-time” statistical process control in MDF manufacture. In: Proceedings of process and business technologies for the forest products industry. Forest Products Society, Madison, pp 50-51

Young TM (2007) Parametric and non-parametric regression tree models of the strength properties of engineered wood panels using real-time industrial data. Dissertation, University of Tennessee, Knoxville

Young TM, Guess FM (2002) Developing and mining higher quality information in automated relational databases for forest products manufacture. Int J Reliab Appl 3(4):155-164

Young TM, André N, Huber CW (2004) Predictive modeling of the internal bond of MDF using genetic algorithms with distributed data fusion. In: Proceedings of the eighth European panel products symposium, pp 45-59

Young TM, Shaffer LB, Guess FM, Bensmail H, León RV (2008) A comparison of multiple linear regression and quantile regression for modeling the internal bond of medium density fiberboard. For Prod J 58(4):39-48

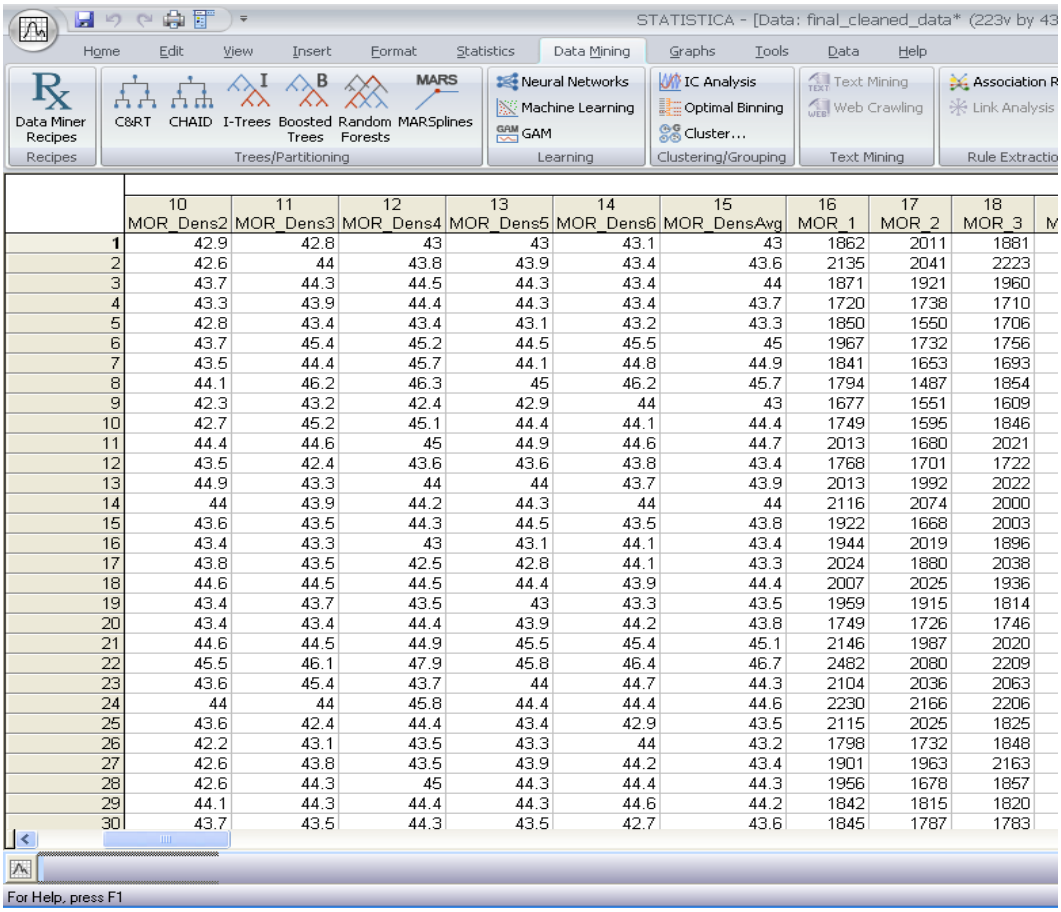
Zeng Y (2011) A study of missing data imputation and predictive modeling of strength properties of wood composites. Master's Thesis, University of Tennessee, Knoxville

Zhou Y, Mulekar MS, Nerellapalli P (2005) Adaptive spam filtering using dynamic feature space. In: Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence, pp 302-309

APPENDIX

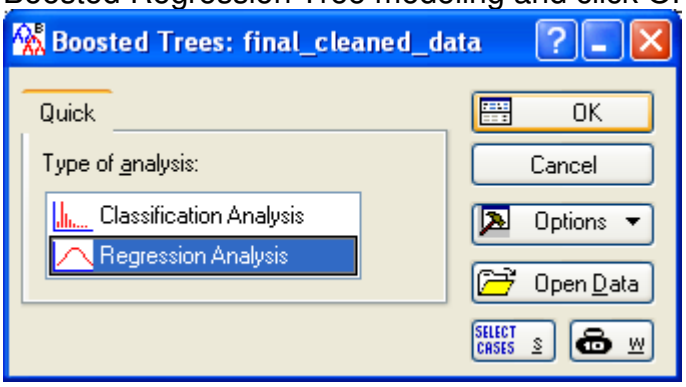
Creating Boosted Regression Tree Models in *STATISTICA* 10

Step 1: Open the data file and select the “Boosted Trees” option under the Data Mining Tab.

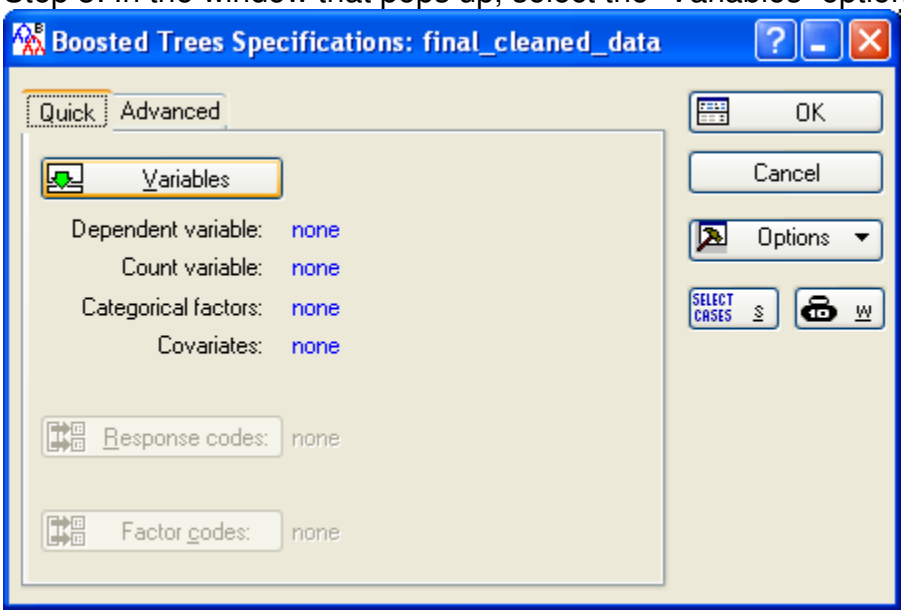


	10 MOR_Dens2	11 MOR_Dens3	12 MOR_Dens4	13 MOR_Dens5	14 MOR_Dens6	15 MOR_DensAvg	16 MOR_1	17 MOR_2	18 MOR_3
1	42.9	42.8	43	43	43.1	43	1862	2011	1881
2	42.6	44	43.8	43.9	43.4	43.6	2135	2041	2223
3	43.7	44.3	44.5	44.3	43.4	44	1871	1921	1960
4	43.3	43.9	44.4	44.3	43.4	43.7	1720	1738	1710
5	42.8	43.4	43.4	43.1	43.2	43.3	1850	1550	1706
6	43.7	45.4	45.2	44.5	45.5	45	1967	1732	1756
7	43.5	44.4	45.7	44.1	44.8	44.9	1841	1653	1693
8	44.1	46.2	46.3	45	46.2	45.7	1794	1487	1854
9	42.3	43.2	42.4	42.9	44	43	1677	1551	1609
10	42.7	45.2	45.1	44.4	44.1	44.4	1749	1595	1846
11	44.4	44.6	45	44.9	44.6	44.7	2013	1680	2021
12	43.5	42.4	43.6	43.6	43.8	43.4	1768	1701	1722
13	44.9	43.3	44	44	43.7	43.9	2013	1992	2022
14	44	43.9	44.2	44.3	44	44	2116	2074	2000
15	43.6	43.5	44.3	44.5	43.5	43.8	1922	1668	2003
16	43.4	43.3	43	43.1	44.1	43.4	1944	2019	1896
17	43.8	43.5	42.5	42.8	44.1	43.3	2024	1880	2038
18	44.6	44.5	44.5	44.4	43.9	44.4	2007	2025	1936
19	43.4	43.7	43.5	43	43.3	43.5	1959	1915	1814
20	43.4	43.4	44.4	43.9	44.2	43.8	1749	1726	1746
21	44.6	44.5	44.9	45.5	45.4	45.1	2146	1987	2020
22	45.5	46.1	47.9	45.8	46.4	46.7	2482	2080	2209
23	43.6	45.4	43.7	44	44.7	44.3	2104	2036	2063
24	44	44	45.8	44.4	44.4	44.6	2230	2166	2206
25	43.6	42.4	44.4	43.4	42.9	43.5	2115	2025	1825
26	42.2	43.1	43.5	43.3	44	43.2	1798	1732	1848
27	42.6	43.8	43.5	43.9	44.2	43.4	1901	1963	2163
28	42.6	44.3	45	44.3	44.4	44.3	1956	1678	1857
29	44.1	44.3	44.4	44.3	44.6	44.2	1842	1815	1820
30	43.7	43.5	44.3	43.5	42.7	43.6	1845	1787	1783

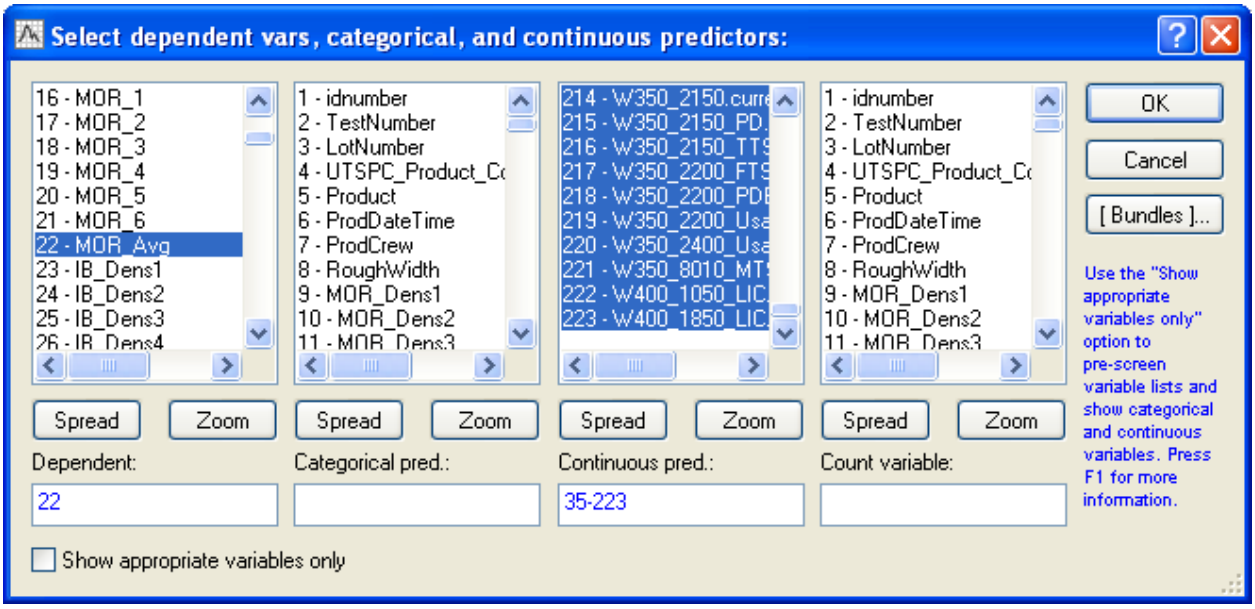
Step 2: In the window that pops up, select the “Regression Analysis” option to perform Boosted Regression Tree modeling and click OK.



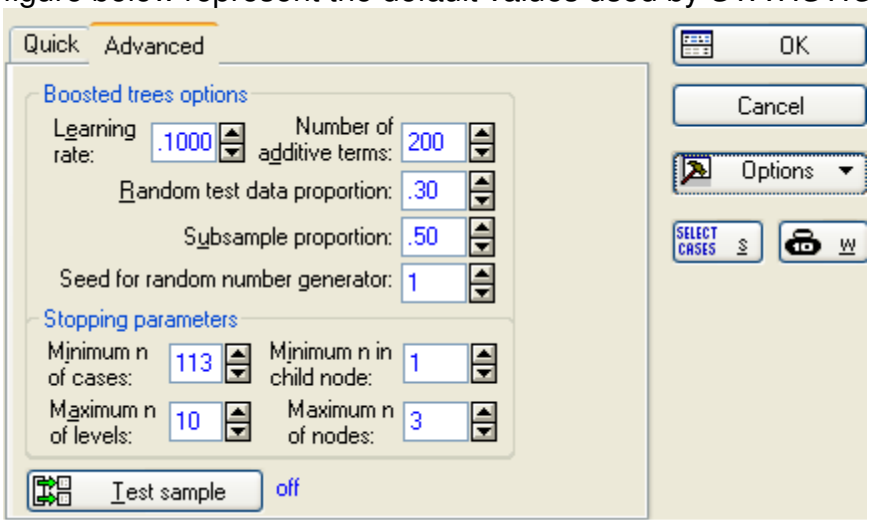
Step 3: In the window that pops up, select the “Variables” option.



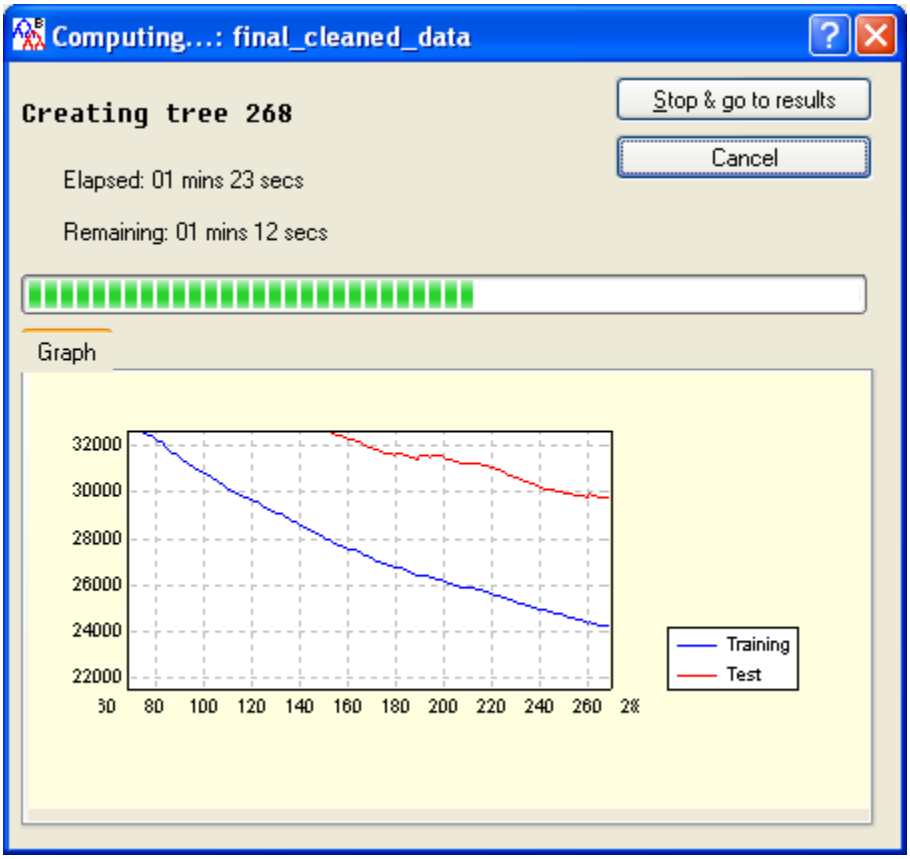
Step 4: Select the “Dependent” variable and the “Predictor” variables from the variables list. Click OK.



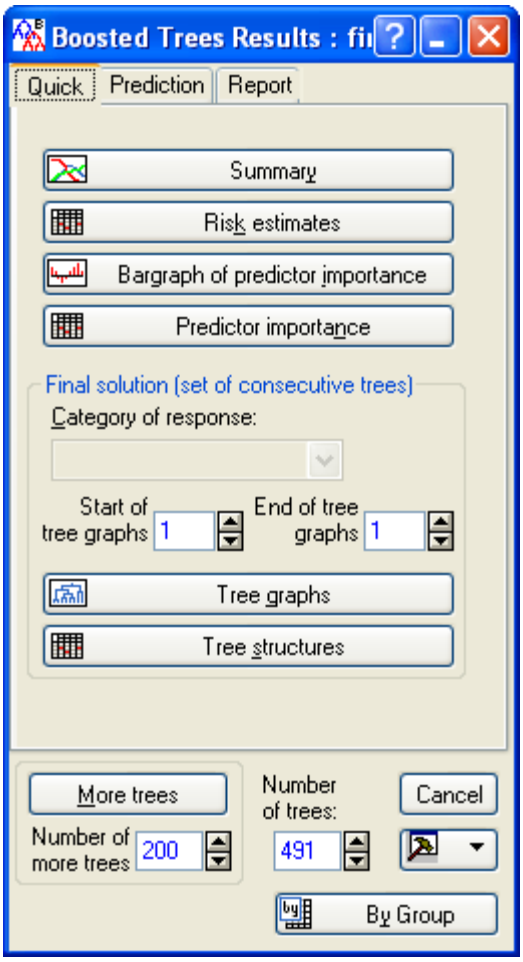
Step 5: Then click on the “Advanced” tab in the window shown in Step 4 and choose parameter settings used to fit a desired Boosted Regression Tree model. For example, set the learning rate (*lr*) to be 0.1, number of additive terms (*nat*) to be 500, subsample proportion (*sp*) to be 0.5, random test data proportion to be 0.2, and maximum number of nodes (*mnn*) to be 3.* Click OK. *STATISTICA* 10 will start building the Boosted Regression Tree model (refer to the top of the next page). (The values provided in the figure below represent the default values used by *STATISTICA* for this data set.)



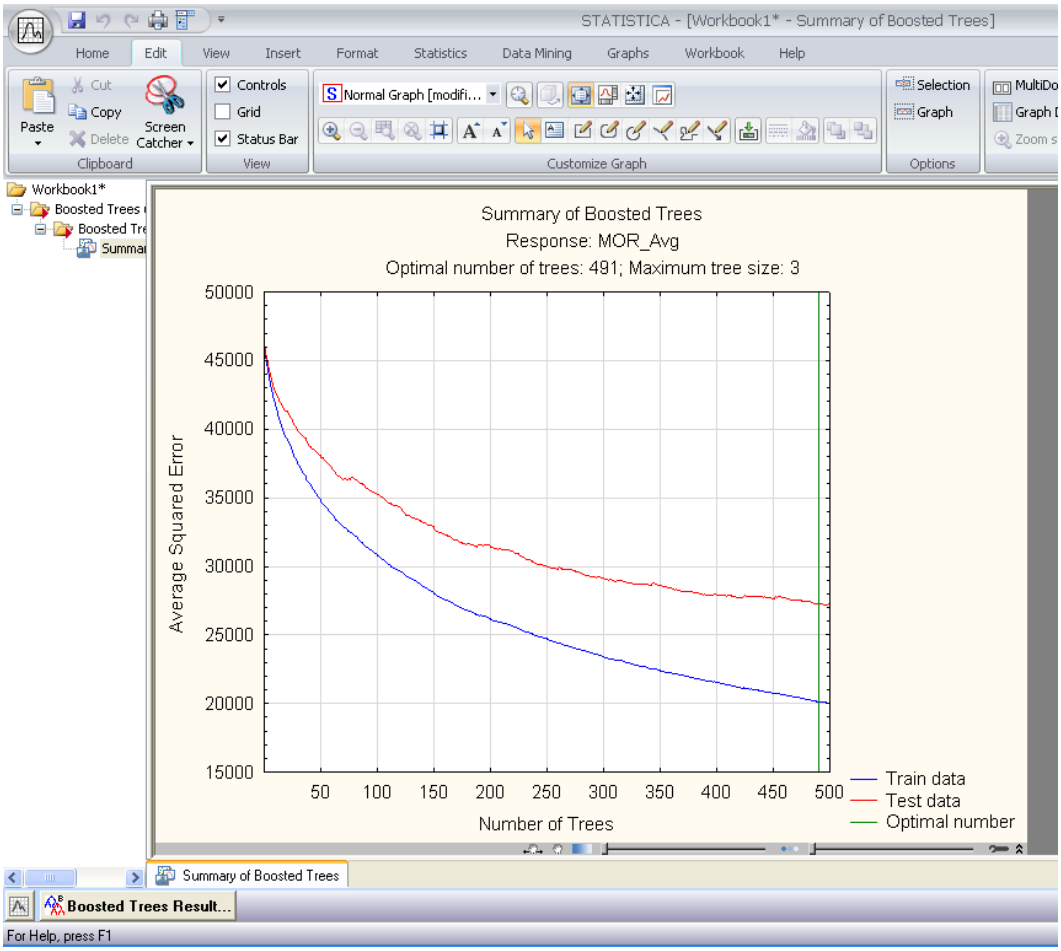
* The “Learning rate,” or the shrinkage parameter, (*lr*) specified the specific weight with which consecutive simple regression trees are added into the prediction equation. For example, a BRT model with 500 trees fitted and *lr* equal to 0.01 will produce predictions that are the sum of predictions from each of the 500 trees multiplied by 0.01. Second, the “Number of additive terms” (*nat*) specified the number of simple regression trees (i.e., additive terms) to be computed in successive boosting steps. The “Maximum number of nodes” (*mnn*) specified the maximum number of nodes allowed for each individual tree in the boosting sequence. Setting *mnn* equal to three (i.e., single split regression trees or stumps) produced BRT models with only main effects. Setting *mnn* equal to five produced BRT models with main effects and two-variable interactions, and so on. The “Subsample proportion” (*sp*) was used for selecting the random learning sample for consecutive boosting steps. The “Random test data proportion” parameter setting determines the percent of randomly selected observations to be used in the testing (i.e., validation) sample.



Step 6: After *STATISTICA* 10 has built the Boosted Regression Tree model, a “Boosted Regression Tree Results” window will open (refer to the top of the next page).



Step 7: Under the “Quick” tab in the “Boosted Regression Tree Results” window select the “Summary” option. This will show the optimal Boosted Regression Tree model for the number of additive terms designated in terms of average squared error on the validation data set (refer to the top of the next page).

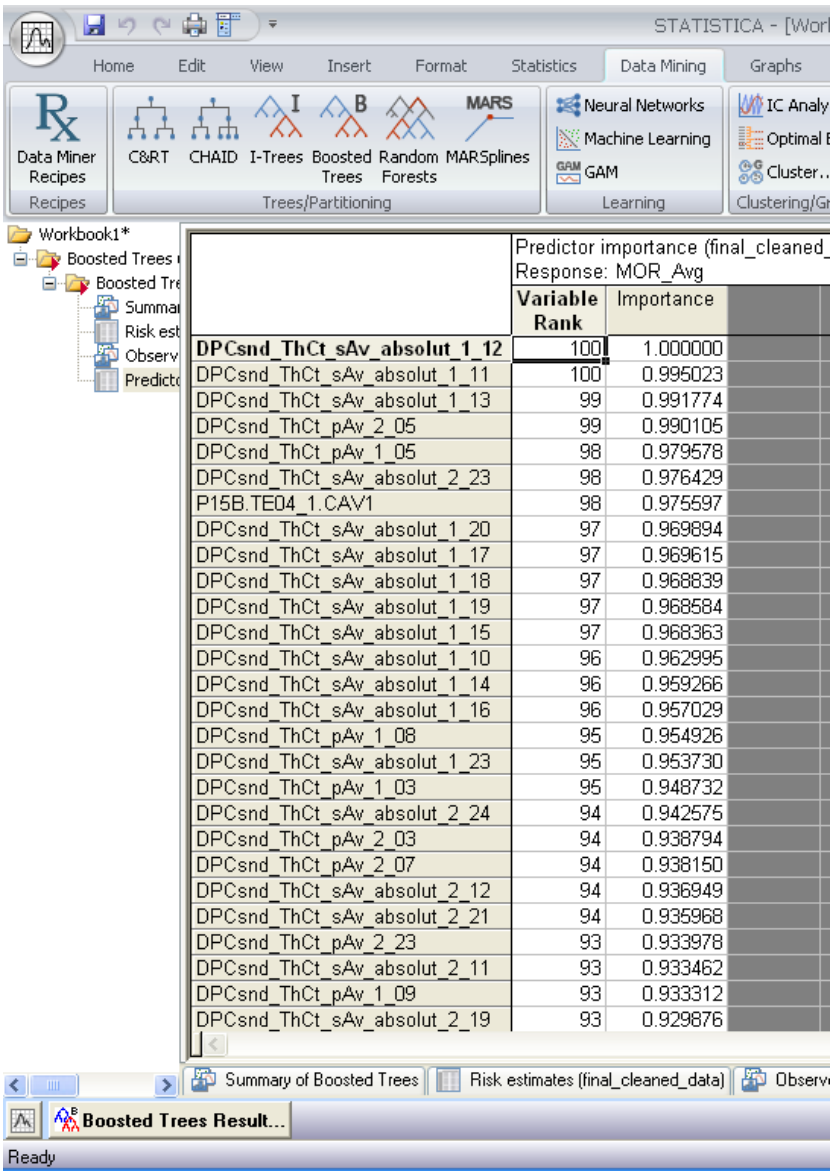


Step 8: Under the “Quick” tab in the “Boosted Regression Tree Results” window select the “Risk Estimates” option. This provides the MSE values obtained on the training and testing (i.e., validation) data sets.

Risk estimates (final_cleaned_data)
Response: MOR_Avg

	Risk Estimate	Standard error
Train	20134.51	820.625
Test	27214.85	2394.969

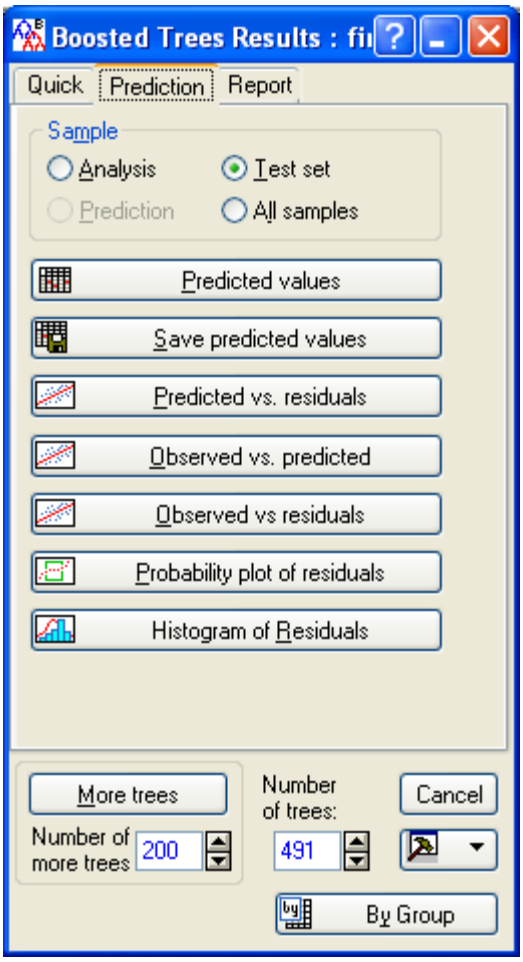
Step 9: Under the “Quick” tab in the “Boosted Regression Tree Results” window select the “Predictor Importance” option. This provides “Predictor Importance” values and rankings for predictor variables (refer to the top of the next page).



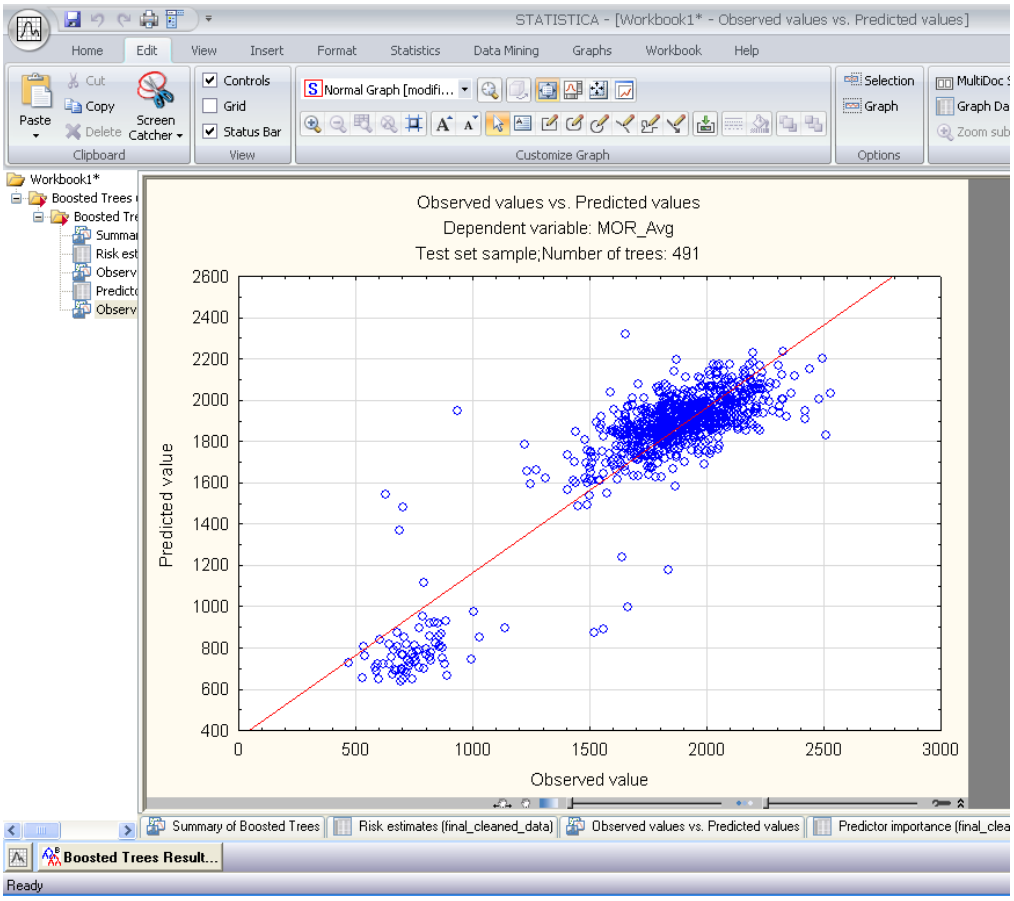
Predictor importance (final_cleaned_
Response: MOR_Avg

	Variable Rank	Importance
DPCsnd_ThCt_sAv_absolut_1_12	100	1.000000
DPCsnd_ThCt_sAv_absolut_1_11	100	0.995023
DPCsnd_ThCt_sAv_absolut_1_13	99	0.991774
DPCsnd_ThCt_pAv_2_05	99	0.990105
DPCsnd_ThCt_pAv_1_05	98	0.979578
DPCsnd_ThCt_sAv_absolut_2_23	98	0.976429
P15B.TE04_1.CAV1	98	0.975597
DPCsnd_ThCt_sAv_absolut_1_20	97	0.969894
DPCsnd_ThCt_sAv_absolut_1_17	97	0.969615
DPCsnd_ThCt_sAv_absolut_1_18	97	0.968839
DPCsnd_ThCt_sAv_absolut_1_19	97	0.968584
DPCsnd_ThCt_sAv_absolut_1_15	97	0.968363
DPCsnd_ThCt_sAv_absolut_1_10	96	0.962995
DPCsnd_ThCt_sAv_absolut_1_14	96	0.959266
DPCsnd_ThCt_sAv_absolut_1_16	96	0.957029
DPCsnd_ThCt_pAv_1_08	95	0.954926
DPCsnd_ThCt_sAv_absolut_1_23	95	0.953730
DPCsnd_ThCt_pAv_1_03	95	0.948732
DPCsnd_ThCt_sAv_absolut_2_24	94	0.942575
DPCsnd_ThCt_pAv_2_03	94	0.938794
DPCsnd_ThCt_pAv_2_07	94	0.938150
DPCsnd_ThCt_sAv_absolut_2_12	94	0.936949
DPCsnd_ThCt_sAv_absolut_2_21	94	0.935968
DPCsnd_ThCt_pAv_2_23	93	0.933978
DPCsnd_ThCt_sAv_absolut_2_11	93	0.933462
DPCsnd_ThCt_pAv_1_09	93	0.933312
DPCsnd_ThCt_sAv_absolut_2_19	93	0.929876

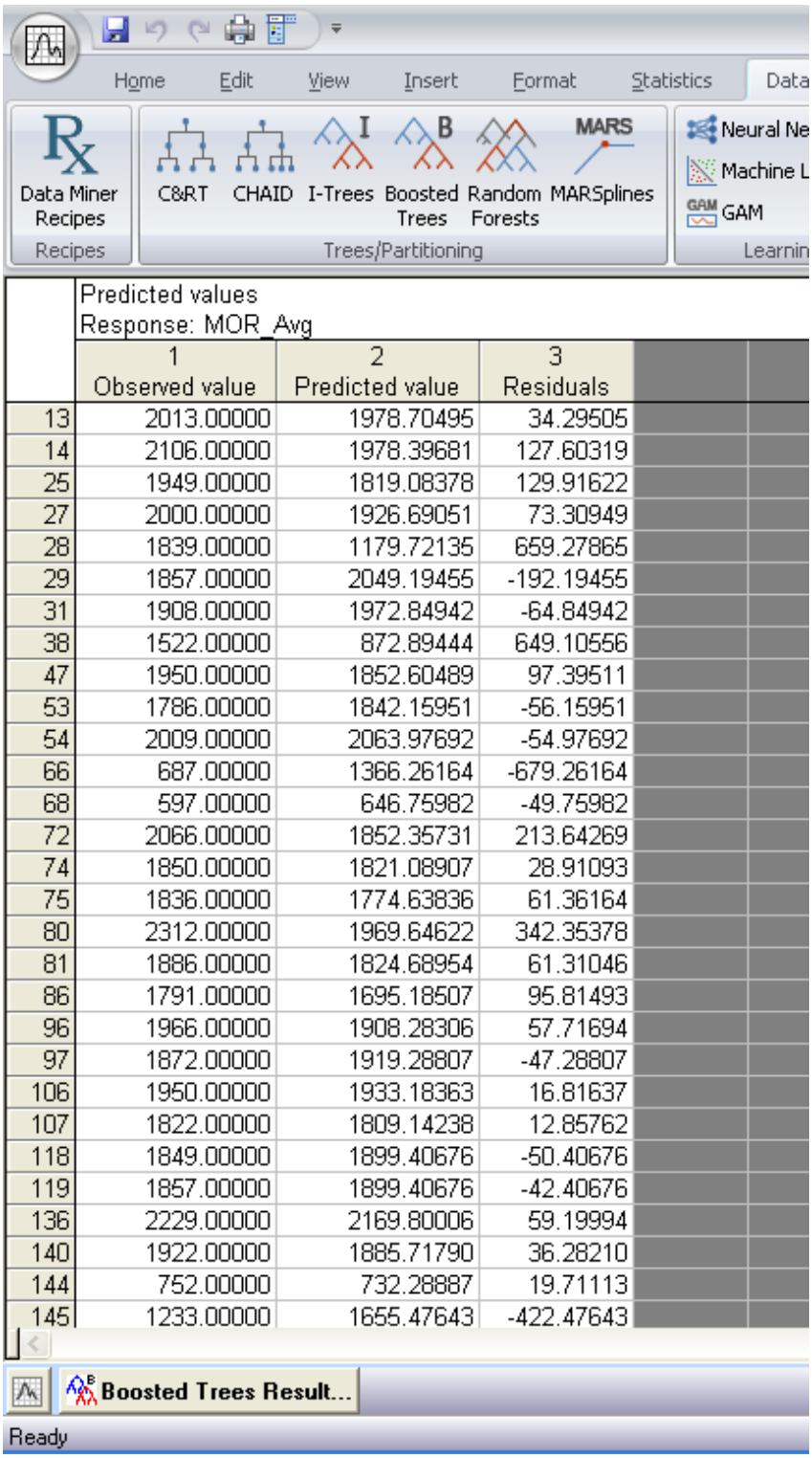
Step 10: Select the “Prediction” tab in the “Boosted Regression Tree Results” window to obtain graphs, plots, and values, for the training and validation data sets (refer to the top of the next page).



Step 11: Under the “Prediction” tab in the “Boosted Regression Tree Results” window select the Sample to be “Test set.” (One can also do this soon to be mentioned analysis on the training data set by selecting the Sample to be “Analysis” or “All Samples.”) Then select the Observed vs. Predicted option. This provides a scatterplot of the observed vs. predicted values in the validation data set (refer to the top of the next page).



Step 12: If desired, one can obtain the observed and predicted values for the different data sets by selecting the “Predicted Values” option under the “Prediction” tab in the “Boosted Regression Tree Results” window (refer to the top of the next page).



Predicted values
Response: MOR_Avg

	1 Observed value	2 Predicted value	3 Residuals
13	2013.00000	1978.70495	34.29505
14	2106.00000	1978.39681	127.60319
25	1949.00000	1819.08378	129.91622
27	2000.00000	1926.69051	73.30949
28	1839.00000	1179.72135	659.27865
29	1857.00000	2049.19455	-192.19455
31	1908.00000	1972.84942	-64.84942
38	1522.00000	872.89444	649.10556
47	1950.00000	1852.60489	97.39511
53	1786.00000	1842.15951	-56.15951
54	2009.00000	2063.97692	-54.97692
66	687.00000	1366.26164	-679.26164
68	597.00000	646.75982	-49.75982
72	2066.00000	1852.35731	213.64269
74	1850.00000	1821.08907	28.91093
75	1836.00000	1774.63836	61.36164
80	2312.00000	1969.64622	342.35378
81	1886.00000	1824.68954	61.31046
86	1791.00000	1695.18507	95.81493
96	1966.00000	1908.28306	57.71694
97	1872.00000	1919.28807	-47.28807
106	1950.00000	1933.18363	16.81637
107	1822.00000	1809.14238	12.85762
118	1849.00000	1899.40676	-50.40676
119	1857.00000	1899.40676	-42.40676
136	2229.00000	2169.80006	59.19994
140	1922.00000	1885.71790	36.28210
144	752.00000	732.28887	19.71113
145	1233.00000	1655.47643	-422.47643

VITA

Dillon Matthew Carty was born in Kingsport, TN on July 3, 1986. He attended Dobyns-Bennett High School and graduated in June 2004. He attended the University of Tennessee, Knoxville, where he received a B.S. in Mathematics, B.S. in Statistics, a Minor in Anthropology, and graduated Cum Laude in May 2009. Dillon is a Graduate Research Assistant with the Center for Renewable Carbon at the University of Tennessee, Knoxville. He is currently pursuing his M.S. degree in Statistics from the University of Tennessee and plans to graduate in August 2011. He is currently an intern at Eastman Chemical Company within the Applied Statistics Group. Dillon is a student member of the American Statistical Association (ASA) and the Institute of Mathematical Statistics (IMS).