8-2011

A Study of Missing Data Imputation and Predictive Modeling of Strength Properties of Wood Composites

Yan Zeng
yzeng1@utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by Yan Zeng entitled "A Study of Missing Data Imputation and Predictive Modeling of 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.)

To the Graduate Council:

I am submitting herewith a thesis written by Yan Zeng entitled "A Study of Missing Data Imputation and Predictive Modeling of 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.)

A Study of Missing Data Imputation and Predictive Modeling of Strength Properties of Wood Composites

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

Yan Zeng

August 2011

Dedication

This thesis would have never become possible and the pursuit of my goals would have never gone this far without my incredibly great parents. My mother read me bed stories every night and taught me English when I was very little. She was my first mentor who not only inspired my passion yearning for knowledge, but also showed me how to become a good person. She obtained her PhD when she was 40 years old. At that time she was writing her dissertation while taking care of a family and her patients as a physician. Remembering her story I always felt strength and power whenever facing challenges throughout my academic career. My father raised me almost by himself at my early age when my mother was pursuing her academic career overseas. He taught me responsibility and honesty, two of most important qualities for a good scholar. Mom and dad, thank you so very much for your unconditional and consistent love.

To my dearest and closest friend, Lin, thank you for helping me maintain my focus and balance my life well during these two years. Also thank you for always stepping ahead in pursuit of your own dreams, which gave me tremendous encouragement, made me better prepared, and helped me become a better person. I feel indebted to you for your love and support.

Also to my fellow graduate students and friends, Evan, Dillon, Nancy, and Xia, I want to thank each of you for enriching my two years' life at Knoxville. Together we have written a most beautiful page in our life that each of us will always remember.

To each of you, especially my family, I want to say "thank you" and I dedicate this thesis.

Acknowledgements

Funding sources for this project and thesis were USDA CSREES 093415820217 Special Wood Utilization research agreement R110515041 with The University of Tennessee and USDA McIntire-Stennis research project TENOOMS-101. I greatly acknowledge Center for Renewable Carbon and the Department of Statistics of the University of Tennessee for providing me with this most valuable opportunity pursuing my academic career.

Especially I would like to express my deep gratitude to professors Dr. Frank M. Guess and Dr. Timothy M. Young. Dr. Guess guided me with his energetic spirit of teaching and research, his prudence and committed support, and most important, his good heart. He is a model advisor and I chose statistics as part of my career because of him. Dr. Young taught me how to become a good researcher and helped me develop good working habits, which is a most valuable asset for my future career. They both provided me with consistent support, financially and mentally. I felt tremendous honor and very fortunate to have opportunity working and researching under their guidance for two years.

Also I would like to thank professors Dr. Russell L. Zaretzki and Dr. Robert W. Mee. They not only taught me knowledge, but inspired my research interests and provided intellectual support when I was working on this project. Sometimes I even learnt more via my conversation with them than simply reading a book or paper. Similarly, Dr. Nicolas Andrè shared with me lots of his insightful thoughts in our

discussion and sparked my ideas by his challenging but friendly questions. Most importantly, he helped me prepare the original data sources of this thesis.

Finally I want to acknowledge every person in statistics department and center for renewable carbon for their help during my research on this project.

Abstract

Problem: Real-time process and destructive test data were collected from a wood composite manufacturer in the U.S. to develop real-time predictive models of two key strength properties (Modulus of Rupture (MOR) and Internal Bound (IB)) of a wood composite manufacturing process. Sensor malfunction and data "send/retrieval" problems lead to null fields in the company's data warehouse which resulted in information loss. Many manufacturers attempt to build accurate predictive models excluding entire records with null fields or using summary statistics such as mean or median in place of the null field. However, predictive model errors in validation may be higher in the presence of information loss. In addition, the selection of predictive modeling methods poses another challenge to many wood composite manufacturers.

Approach: This thesis consists of two parts addressing above issues: 1) how to improve data quality using missing data imputation; 2) what predictive modeling method is better in terms of prediction precision (measured by root mean square error or RMSE). The first part summarizes an application of missing data imputation methods in predictive modeling. After variable selection, two missing data imputation methods were selected after comparing six possible methods. Predictive models of imputed data were developed using partial least squares regression (PLSR) and compared with models of non-imputed data using ten-fold cross-validation. Root mean square error of prediction (RMSEP) and normalized RMSEP (NRMSEP) were calculated. The second presents a series of comparisons among four predictive modeling methods using imputed data without variable selection.

Results: The first part concludes that expectation-maximization (EM) algorithm and multiple imputation (MI) using Markov Chain Monte Carlo (MCMC) simulation achieved more precise results. Predictive models based on imputed datasets generated more precise prediction results (average NRMSEP of 5.8% for model of MOR model and 7.2% for model of IB) than models of non-imputed datasets (average NRMSEP of 6.3% for model of MOR and 8.1% for model of IB). The second part finds that Bayesian Additive Regression Tree (BART) produced most precise prediction results (average NRMSEP of 7.7% for MOR model and 8.6% for IB model) than other three models: PLSR, LASSO, and Adaptive LASSO.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1
Background and Introduction

Predictive modeling has become an important technique to improve the quality of wood composites and is also used in other manufacturing systems. This paper studies two aspects of predictive modeling: 1) how to impute missing values to improve data quality; and 2) how to choose predictive models to improve prediction precision. The rest of this chapter and remainder of the thesis are organized as follows.

Chapter 1 describes the background of wood composite manufacturing and how the data were assessed and processed. Chapter 2 demonstrates the process of variable selection and dimension reduction of the original data. Chapter 3 is composed of two parts: a) six missing data imputation methods were compared and the best methods were selected to generate imputed datasets; and b) partial least square regression (PLSR) was applied to both the non-imputed and imputed data to demonstrate the impact of missing data imputation on predictive modeling results. In Chapter 4 data were re-imputed to further study and compare four predictive modeling methods. Chapter 5 concludes the findings presented in Chapters 3 and 4. In Chapter 6 future studies and research originated from current results are discussed.

1.1. Background

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 automotive and

plastics industries. The industry generates more than $200 billion a year in sales and employs approximately 900,000 people earning $50 billion in annual payroll. The industry is among the top 10 manufacturing employers in 42 states (American Forest and Paper Association (2010)).

This thesis describes a study that was performed for a large-capacity wood composite panel manufacturing factory in the southeastern U.S. The factory produces particleboard wood composite panels which are used in furniture, cabinetry, shelving, etc. Two key product quality metrics are: Modulus of Rupture (MOR) and Internal Bound (IB). MOR and IB are obtained via destructive test where samples of final product are taken every one to two hours from the production line and cut from the cross sections of the master-panel of particleboard. The average MOR and IB strengths are measured in kilopascal (kPa) and stored in a data warehouse. However, the time span between destructive tests may be as long as two hours, during which a significant amount of production occurs. Given the time gap between consecutive destructive tests, hours of particleboard could be manufactured out of specification resulting in rework and scrap. This time gap and lack of real-time knowledge of strength properties may also lead to higher than necessary operational targets of resin and wood, which are non-competitive from a business perspective.

The factory is exploring "real-time" prediction of MOR and IB as a remedy to maintain product specification and minimize costs. The data for the potential predictor variables (e.g., fiber moisture, line speed, mat temperature, press pressure, etc.) come from the hundreds of sensors on the production line that are typically used for process

control. All process data are simultaneously transported and merged with MOR and IB records into one database at the time of sampling from the production line. A real-time, automated relational database was created for this study that aligned real-time process sensor data with the destructive test data of the laboratory for the 118 possible product types (e.g., 16mm regular, ¾" high strength, etc.). Lag times of the sensor data were established that corresponded to the time required for the wood fibers to travel through the process passing by multiple sensors before finally reaching the outfeed of the continuous press. One calibration model was built from the sensor data to predict MOR and IB for all of the 118 product types.

A major problem for all manufacturers that collect real-time data from sensors is missing data. The missing data problem is usually caused by the malfunctioning sensors that require replacement, or data send/retrieve errors to the data warehouse across the programmable logic controller (PLC) Ethernet network. A common approach is to explore statistical models without the missing records associated with the null fields in the database. By default, many statistical packages (e.g., SAS® and R) will exclude the entire record if there is a missing field for a predictor variable. In real-time data warehousing in a manufacturing environment, there may be hundreds of possible predictor variables with many null fields, the exclusion of records can result in substantial information loss. The validity and precision (measured by root mean squares error) of predictive models can be adversely affected from such information loss.

1.2. Data introduction

The first step in this study was to assess the quality of collected data. After removing the non-recorded sensor data in the database, there were 237 predictor variables and two response variables (MOR and IB). In total there were 4,522 records. There were 11 records with response variables missing and the records were deleted. Every predictor variable had at least 2.4% of data missing. The missing rate varied from 2.4% to 81%. Six percent of predictor variables had more than 20% of data missing. Every record also had missing fields with missing rate ranging from less than 0.5% to 90%. Only six records had a missing rate above 20%. According to previous empirical studies (UCLA Statistical Consulting Service (2011)), an approximate 10% missing rate for a variable is considered to be suitable for an analysis without imputation, a strategy sometimes referred to as "complete case analysis." Also, with an increase in the missing rate, the accuracy and robustness of imputation, especially multiple imputation (MI), will be adversely affected (Ni et al. (2005) and Soullier et al. (2010)). An increase in the missing rate also requires relatively more iterations of imputation under MI, which elongates the computation time. Thus, predictor variables and observations with more than 20% missing rate were excluded. This resulted in 222 predictor variables and 4,411observations. There were 3,647 records with at least two or more fields missing.

Predictors were highly correlated as suggested by the variance-covariance matrix and variation-inflation factors (VIF). Given the large differences in the scales of the predictors, all predictors were standardized before model development, i.e., $\frac{x_i-\bar{x}}{\hat{\sigma}}$, where $\bar{x}$ is the average of non-missing values.

Chapter 2
Variable Selection

After initial data quality assessment, we continued to select predictor variables to impute. There are two reasons for performing variable selection before imputing missing data and proceeding with predictive modeling. The first reason is due to the constraint of computation resources required by iterated computation (e.g., maximum-likelihood based method and MI). Truxillo (2005) and Lanning et al. (2003) noted that SAS® or R can become slow or may not converge on imputation results. We also tried imputation without selecting variables beforehand. Both the expectation-maximization (EM) algorithm and MI failed to converge after lengthy iterations. A second reason for variable selection is to exclude non-informative variables. Variable selection for large highly correlated multi-dimensional or even high-dimensional process data are becoming increasingly important in statistical analysis for modern manufacturing environment, as documented by Wang and Jiang (2009), Gonźalez and Sánchez (2010). They all proposed or performed relevant variable selection methods prior to analyzing datasets of tens and hundreds of production process variables. Guyon and Elisseeff (2003) also concluded that variable selection could reduce the calibration modeling training time and improve prediction performance.

2.1. Literature Review

Variable selection and data dimension reduction have been performed quite often and become an essential step in predictive modeling in wood composites (Young and

Guess (2002), Young et al. (2004), Young et al. (2008), Clapp et al. (2008), Andréet al. (2008)). Methods such as multiple linear regression (MLR), correlation analysis, principal component analysis (PCA), and genetic algorithm (GA) were used in earlier studies. The existence of multicollinearity in this study's dataset rendered MLR ineffective because the computation of covariance matrices of predictor variables could be ill conditioned (Soh et al. (2005)). Clapp et al. (2008) analyzed the correlation matrix of predictor variables with the response variable internal bond (IB). The ten variables with the highest (absolute value) correlation with the response variable were selected for further predictive modeling. However, in this study for the purpose of imputing missing data, we needed to incorporate as much information as possible from the available data while reducing the number of variables to make the imputation feasible for computation. It would be difficult to select variables based solely on the correlation matrix. In previous studies modeling wood composite manufacturing process, Principal component analysis (PCA) has been considered as a most popular method to select variables for large multi-dimensional data (see, for instance, Parsons et al. (2004), Gonźalez and Sánchez (2010)). Clapp et al. (2008) also adopted PCA in a previous study of predictive modeling in wood composites. However, most of those studies also noted the limit of PCA on industrial applications. In PCA, the new variables (components) are linear combinations of the original weighted variables. While often very useful, PCA doesn't reduce the effective number of variables and information from irrelevant variables is still preserved.

Another newly adopted variable selection method in predicting wood composites is genetic algorithms (GA). André et al. (2008) used GA to exclude non-informative variables for the improvement of predictive models. However, they excluded records that contained null fields instead of imputing missing data. The GA method turned out to be useful in their study but the cost of computation time was high and limited the capabilities real-time predictive modeling. Other potential problems of using GA for industrial applications include extra spurious variables, computation convergence, and inappropriate control parameter setup in the GA (e.g., Draper and Fouskakis (2000), Soh (2005), Zhu and Chipman (2006)).

2.2. LASSO Method

Based on the literature reviewed related to variable selection, the LASSO method (Tibshirani (1996)) was selected in this study for variable selection. Let $\mathbf{y} = (y_1, \dots, y_n)^T$ represent the response vector and $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^T, j = 1, \dots, p,$ denote the linear independent predictors. Suppose that $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ is the predictor matrix. We assume that data are centered. The LASSO estimates can be expressed as:

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \left\| \mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (1)$$

where $\lambda \sum_{j=1}^{p} |\beta_j|$ is the penalty term and the solution could shrink towards zero, with which variables with zero coefficients are dropped. For details refer to Hastie et al. (2009).

As indicated above, LASSO was developed to achieve shrinkage and variable selection simultaneously, initially as a constrained version of the ordinary least squares (OLS) estimator. LASSO has been widely used in medicine, economics, and other scientific fields which have large multi-dimensional data (Hastie et al. (2009) and Salvin (2010)). Recently it has been adopted more in modern manufacturing environments. Wang and Jiang (2009) used a family of "penalized regression" methods, which includes LASSO, for out-of-control variable selection for statistical monitoring of high-dimensional manufacturing process data. They used a method of the same family with LASSO (except for the minor difference in the penalty term) to develop a variable-selection-based multivariate control chart and applied it to a group of high-dimensional data from a wood composites factory.

2.3. Variable Selection Results

In this study we performed LASSO using the "lars" package of R (Efron et al. (2004)). Two linear models for the two datasets with MOR and IB as the response variables were built using LASSO regression. Variables with extremely small estimators ($\beta \leq 10^{-5}$) were removed. After this selection process, we obtained two non-imputed datasets. The dataset for MOR as the response had 107 predictor variables and 1,073 complete observations out of 4,411 observations; the dataset for IB as the response had 86 predictor variables and 1,194 complete observations out of 4,411 observations.

Chapter 3
Missing data imputation

Missing data imputation depends on the pattern of missing data. Different missing pattern have different imputation methods. This chapter introduces the background of missing data patterns and the process of determining the missing mechanism of data in this study. Accordingly six appropriate imputation methods were compared. The objective of this portion of the thesis study was to determine whether the additional information generated by imputation improves the predictive modeling results.

3.1. Missing Data Pattern

Since different missing data patterns may require different imputation methods, we studied the missing pattern of the datasets before selecting an appropriate imputation method. As first introduced by Little and Rubin (1987), there are three major patterns: 1) missing completely at random (MCAR); 2) missing at random (MAR); and 3) non-ignorable missing data (MNAR). MCAR occurs when the missing values on variable Y are independent of all other observed variables and the values of Y itself, which is a very strong assumption and can be impractical for real-life data (Muthén et al. (1987)). MAR assumes that the probability that an observation is missing on variable Y depends on other observed variables but not on the values of Y itself, which is more plausible than MCAR. Under MNAR a missing value no longer occurs "at random" since the probability that an observation is missing on variable Y depends on other unobserved variables.

In this study we assumed MAR for each dataset. Many researchers have noted that in many situations MCAR can be rejected empirically in favor of MAR, especially by including a relatively rich set of predictors in the model (e.g. Rubin (1996), Schafer (1997), and Collins et al. (2001)). In this study we used sensors to include as many predictor variables as possible during the manufacturing process and retained almost half of the predictors after variable selection. We didn't assume MNAR because some studies suggested that presence or absence of NMAR can hardly be demonstrated using only the observed data (King et al. (2001) and Yarandi (2002)).

After assuming MAR, another important concept to decide was monotone "missingness." If the dataset can be rearranged in such a way that there is a hierarchy of "missingness," namely when a variable Y_j is missing for the individual observation i, implies that all subsequent variables Y_k , $k>j$, are all missing for the individual observation i. Otherwise the missing pattern of MAR is said to be arbitrary. The reason for checking monotone "missingness" is that it influences the imputation method. Simpler methods can be used if the pattern is monotone, for instance, the propensity score method using logistic regression (e.g. Lavori et al. (1995), Yuan (2000)). However, monotone assumption is uncommon in most realistic settings (Horton and Kleinman (2007)). For datasets with large dimensionality such as is the case in this study, it's also unrealistic to check the monotone "missingness" via plotting the missing pattern. Therefore, we assumed the datasets had arbitrary instead of monotone missing values.

3.2. Imputation Methods

The focus of previous studies on missing data is to use imputation to better estimate the parameters of the statistical model instead of filling missing fields (e.g., Little and Rubin (2002)). The literature on missing data in manufacturing industry applications is sparse. Substituting missing values with summary statistics such as mean or median are used often (e.g., UCLA Statistical Consulting Service (2011)). A study by Jensen et al. (2008) noted the impact of missing data on statistical process control applications. In addition to mean/median substitution, we also included two widely used methods: simple random imputation and the last-value-carried-forward (LOCF); along with two imputation methods involving iterated computation: maximum likelihood method using expectation-maximization or the EM algorithm and MI using Markov chain Monte Carlo (MCMC). A brief review for those methods and their usages in this study are provided.

Mean/median substitution replaces missing fields with the mean/median of the corresponding variable. Mean substitution is also called "unconditional mean" imputation (Little (1992)). There is also "conditional mean" imputation which considers data values from other predictor variables (Schafter (1997)). A common approach of "conditional mean" imputation is to use a regression model, i.e., replacing missing values with predicted values from a regression analysis using other predictor variables (Fetter (2001), McGee and Bergasa (2005), and Faraway (2005)). The last observation carried forward (LOCF) method replaces missing values with the last known value of the variable in a time-ordered data set (the MOR and IB datasets in our study were time-

ordered). Programs in R were written to perform LOCF. However, Horton and Kleinman (2007), Gelman and Hill (2007), and Hamer et al. (2009) have all noted bias introduced by the LOCF method.

The simple random imputation method ("hot-deck method") replaces the missing value with a randomly selected value from another observation in the same variable. This method is common and efficient for survey studies of large datasets with a low fraction of missing values and is beneficial for possible real-time predictive modeling (Altmayer (2002), Lanning and Berry (2003)). However, it has a fundamental flaw of underestimating the variability caused by uncertainty of missing values (Little and Rubin (2002), Gelman and Hill (2007)).

Programs in R were written for the simple imputation methods (mean/median/single random imputation). These methods may be considered to be accurate if the proportion of missing values is small (e.g., less than 5%) (Yarandi (2002)). They could be easy to implement for the univariate missing case but may be difficult to implement for the multiple missing variable case (Sinharay et al. (2001)). Comparatively, imputation methods based on maximum-likelihood and MI take all of the variables into account using iterated computation.

The EM algorithm is usually used for maximum-likelihood method and has theoretical benefits (Little and Rubin (1986)). Simulation studies have suggested that EM may be superior to traditional mean/median substitution and hot-deck methods (Enders (2001)). In general EM iterates through two steps to obtain estimates. The first step is the "expectation" or E step, in which missing values are filled-in with a guess, i.e.,

an estimate of the missing value, given the observed values in the data. The second step is "maximization" or M step, in which the completed data from the "E step" are processed using maximum-likelihood (ML) estimation as though they were complete data, and the mean and the covariance estimates are updated. Using the updated mean and covariance matrix, the "E step" is repeated to find new estimates of the missing values. The E and M steps are repeated until the maximum change in the estimates from one iteration to the next does not exceed a convergence criterion (Truxillo (2005)). In this study we used PROC MI in SAS® (Version 9.2) for the EM algorithm and the default convergence criterion of SAS.

In addition to the aforementioned imputation methods which replace each missing value with one value, the multiple imputation (MI) by Rubin (1986) replaces each missing value with a set of plausible values that represent the uncertainty of the correct value. The key steps can be concluded as follows (Allison (2000)):

- First, impute missing values using an appropriate model that incorporates random variation;
- Second, repeat this process for M times (e.g., 3 to 5 times), producing M "complete" data sets;
- Third, perform the desired analysis on each data set using standard complete-data methods;
- Fourth, average the values of the parameter estimates across the M samples to produce a single point estimate.

The direct approach for MI is MCMC (Schafer (1997)). In general, MCMC is based on pseudo-random draws which allows researchers to obtain several imputed data sets. In MCMC simulation, one constructs a long Markov Chain to establish a stationary distribution, which is the distribution of interest. By repeatedly simulating steps of the chain, the method draws imputed estimates from the distribution. Thus, a series of complete datasets are generated (Patterson and Yeh (2007)). In this study, we plotted auto-correlation functions (ACF) to test stationarity and MCMC model convergence.

Regarding the choice of "M" times, historically and in practice, the recommendation is to generate three to five imputed datasets (Rubin (1986), Yuan (2000)). This suggestion is most appropriate when the proportion of missing values is relatively low, which was the case of this study, e.g., most variables had a missing fraction of less than 20% after initial data quality assessment and variable selection. It has also been recommended that with a larger value of M the estimates could be more consistent (Schafer and Olsen (1998), UCLA Statistical Consulting Services (2001)). There is a trade-off between "M" and computation time. We chose to generate five imputed complete datasets for practicality.

When performing MI for each dataset using PROC MI in SAS, we put all variables into calculation including the response variables of MOR or IB. The advantage of doing so was that the imputation model could use all of the information (UCLA Statistical Consulting Service (2011)). After iterated calculation, we averaged the M imputed datasets directly to get a final complete dataset while keeping all of the observed values intact. For example, imputation for the missing value Q:

$$\overline{Q} = \frac{1}{M}\sum_{i=1}^{M} \widehat{Q}_i \quad (2)$$

where $\widehat{Q}_i$ denotes the imputed value of missing Q in each of M datasets. Ni et al. (2005) and Datta et al. (2007) adopted similar methods to combine and process the multiple imputed data.

A review of the literature suggested that MI has been widely adopted in areas such as survey polling, psychology, agriculture economics, and clinical research, but no literature was published for manufacturing applications (King et al. (2001), Fetter (2001), Lin (2010)). The advantage of MI over traditional simple imputation methods is that the uncertainty of the correct value to impute is represented by replacing each missing value with a set of plausible values (Allison (2000)). There are not many direct comparisons between maximum-likelihood method and MI imputed data precision. Enders (2001) and Newman (2003) used simulated data and concluded that under the assumption of MAR both methods worked best outperformed other methods in terms of the smaller error of the estimated parameters based on the imputed data. Lin (2010) performed empirical studies and simulation studies comparing EM and MCMC and suggested no significant difference between the two when estimating parameters. Lin (2010) also noted that the number of imputations in MCMC and proportion of missing data had little impact on error.

3.3. Imputation Comparison

Ten-fold cross-validation was used on imputed values for the MOR and IB datasets to compare the precision of imputed values for each of above six methods. The portion of complete observations of each dataset were randomly partitioned as matrix into

ten subsamples, i.e., for the MOR dataset the 1,073 complete observations had $1{,}073 \times 107 = 114{,}811$values; each subsample had 11,481 values. Each subsample of IB dataset had sample size of 10,268, one tenth of $1{,}194 \times 86 = 102{,}684$ values.

Take IB dataset for example, for each of ten subsamples, one subsample of 10,268 was retained as validation data. This subsample of complete records had all known values purposively removed as missing, which was used for validation. This imputation process was repeated ten times for each of the six imputation methods for each subsample data set. The statistic for imputation validation was root mean square error (RMSE):

$$\text{RMSE}\ (x, \hat{x}) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}} \qquad (3)$$

where x is the true value of predictors assigned as missing and $\hat{x}$ is its imputed value; n represents the number of values in one subsample (validation dataset).

We present the cross-validation results calculated from data for MOR in Table 3.1 and data for IB in Table 3.2. Since the RMSEs were calculated from data of different standardized predictor variables, the RMSEs in Table 3.1 and Table 3.2 are unit less. We denoted the smallest RMSE of each imputation in bold.

As shown in Table 3.1, the averaged RMSE for data imputed with the EM method and MCMC method are the lowest among all, 0.43 and 0.41 respectively, much lower than results of other methods. Table 3.2 indicates the same results. EM imputed data produced RMSE of 0.42 and MCMC imputed data produced RMSE of 0.53. For MOR dataset, EM and MCMC also outperformed other methods across the ten times of validation. For the IB dataset, RMSE was lower for eight out of ten imputations for the EM or MCMC methods. Surprisingly, LOCF had the lowest RMSE twice. However, the

Table 3.1 RMSEs from Imputations for Standardized Dataset with MOR as Response

RMSE	Mean Substitution	Median Substitutio	Single Random Imputation	LOCF	EM	MCMC
1	1.92	0.17	1.87	2.14	0.14	**0.09**
2	4.54	2.28	5.01	1.84	0.70	**0.37**
3	4.43	1.92	2.66	1.41	0.92	**0.59**
4	3.47	1.39	3.14	0.96	**0.07**	0.26
5	2.16	0.27	2.52	0.54	0.12	**0.07**
6	2.01	0.40	2.60	0.93	**0.24**	0.48
7	2.18	0.74	0.86	0.84	0.27	**0.25**
8	4.08	1.58	3.04	2.54	0.87	**0.86**
9	3.63	1.58	5.12	1.48	**0.19**	0.28
10	5.23	2.87	2.62	1.83	**0.79**	0.84
Average	3.37	1.32	2.94	1.45	0.43	**0.41**

Table 3.2 RMSEs from Imputations for Standardized Dataset with IB as Response

RMSE	Mean Substitution	Median Substitutio	Single Random Imputation	LOCF	EM	MCMC
1	3.08	0.77	4.92	0.08	**0.33**	**0.33**
2	4.55	1.15	1.44	0.92	**0.65**	0.79
3	2.48	1.46	0.57	1.70	0.27	**0.25**
4	4.16	1.08	2.84	2.31	0.98	**0.90**
5	3.92	0.34	2.61	1.43	**0.12**	1.41
6	2.47	1.09	2.55	2.06	0.10	**0.02**
7	2.24	1.63	1.99	0.74	**0.11**	0.15
8	2.26	0.79	1.48	0.75	**0.43**	0.66
9	2.96	0.51	1.46	1.79	0.64	**0.45**
10	3.49	0.05	5.04	**0.05**	0.59	0.37
Average	3.16	0.89	2.49	1.18	**0.42**	0.53

median substitution method produced the third best results for both datasets. There does not appear to be any consistent difference between the RMSEs for either the EM and MCMC methods. However, there is evidence that the EM and MCMC methods greatly outperform imputation using the average or median methods. Given these results, the EM and MCMC methods were used to impute the MOR and IB datasets.

3.4. Partial Least Squares Regression

Partial least squares regression (PLSR) was developed to model the relation between predictor variable matrix $\boldsymbol{X}$ and a response matrix $\boldsymbol{Y}$ (Wold et al. 1984 Tobias 1997). PLSR decomposes $\boldsymbol{X}$ into orthogonal component scores $\boldsymbol{T}$ and loadings $\boldsymbol{P}$ using singular value decomposition (SVD):

$$\boldsymbol{X} = \boldsymbol{TP} \tag{4}$$

$\boldsymbol{Y}$ is not regressed on $\boldsymbol{X}$ but on the first a columns of the component scores $\boldsymbol{T}$.

In this study, the "pls" package of R (Version 2.11.1) was used to perform PLSR (Mevik and Wehrens 2007). The components (latent variables) in PLSR are obtained iteratively. We demonstrate the algorithm of PLSR in steps as follows:

We start with the singular value decomposition (SVD) of the cross-product matrix $\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{Y}$ to include information on variation in both $\boldsymbol{X}$ and $\boldsymbol{Y}$, and on the correlation between them. The first left and right singular vectors, w and q, are used as weight vectors for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, to obtain scores t and u:

$$t = \boldsymbol{X}w = \boldsymbol{E}w \tag{5}$$

$$u = \boldsymbol{Y}q = \boldsymbol{F}q \tag{6}$$

where $\boldsymbol{E}$ and $\boldsymbol{F}$ are initialized as $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The $\boldsymbol{X}$ scores t are often normalized:

$$t = \frac{t}{\sqrt{t^{\mathrm{T}}t}} \tag{7}$$

Next $\boldsymbol{E}$ and $\boldsymbol{F}$ loadings are obtained by regressing on the same vector t:

$$p = \boldsymbol{E}^{\mathrm{T}}t \tag{8}$$

$$q = \boldsymbol{F}^{\mathrm{T}}t \tag{9}$$

Finally, the data matrices are "deflated": the information related to this component (latent variable), in the form of the outer products tp^{T} and tq^{T}, is subtracted from the (current) data matrices $\boldsymbol{E}$ and $\boldsymbol{F}$.

$$\boldsymbol{E}_{n+1} = \boldsymbol{E}_n - tp^{\mathrm{T}} \tag{10}$$

$$\boldsymbol{F}_{n+1} = \boldsymbol{F}_n - tq^{\mathrm{T}} \tag{11}$$

The estimation of the next component then can start from the SVD of the cross-product matrix ${\boldsymbol{E}_{n+1}}^{\mathrm{T}}\boldsymbol{F}_{n+1}$. After each iteration, vectors w, t, p, and q are saved as columns in matrices $\boldsymbol{W}$, $\boldsymbol{T}$, $\boldsymbol{P}$, and $\boldsymbol{Q}$, respectively. We use $\boldsymbol{W}$ and $\boldsymbol{P}$ to form a new matrix $\boldsymbol{R}$ to relate to the original matrix predictor variable $\boldsymbol{X}$ for further regression analysis: $\boldsymbol{R} = \boldsymbol{W}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W})^{-1}$ and $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{R}$.

Instead of regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$ using ordinary least squares (OLS) regression, we use component scores $\boldsymbol{T}$ to obtain the regression coefficients $\boldsymbol{B}$, where

$$\boldsymbol{B} = \boldsymbol{R}(\boldsymbol{T}^{\mathrm{T}}\boldsymbol{T})^{-1}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{Q}^{\mathrm{T}} \tag{12}$$

In the above regression, we do not use all but the first a columns of the component scores $\boldsymbol{T}$. We use 10 fold cross-validation in R (version 2.11.1) to decide the

optimal number of *a*. For the specific settings in statistical package R please refer to CRAN online document of package "pls" (Wehrens and Mevik (2007)).

After imputation, both EM and MCMC imputed data had 4,411 complete observations. Ten-fold cross-validation was used to evaluate model performance of imputed data and non-imputed data.[1] We again used the root mean square error for prediction (RMSEP) to measure the precision of PLSR model predictions in validation

$$\text{RMSEP}\ (y, \hat{y}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (13)$$

where y is the true value of MOR and $\hat{y}$ is the predicted value; n represents the number of values in one validation dataset. We also computed the normalized RMSEP or NRMSEP as follows

$$\text{NRMSEP} = \frac{\text{RMSEP}}{y_{max} - y_{min}} \qquad (14)$$

where the denominator represents the data range of validation dataset.

3.5. Impact of Imputation on PLSR Prediction

To evaluate the model fitting for imputed and non-imputed datasets, we first plotted predicted MOR and IB against actual MOR and IB from ten-fold cross-validation respectively. We demonstrate the forth validation for PLSR models of MOR and IB data in following Figure 3.1 and Figure 3.2. As can be seen less dispersion of the plot

[1] For MOR, 1,073 complete observations of non-imputed data were randomly partitioned into ten subsamples, each of 107 observations. We retained one single subsample of 107 as the validation dataset and used the other nine subsamples to impute missing values using EM and MI. Then we used the nine subsamples (for the non-imputed case) and the nine subsamples along with imputed data (for imputed case) to train (build) the calibration model of PLSR respectively. This process was repeated ten times.

represents better precision of the model. We also calculated the correlation coefficient (r) between the true values and predicted values to demonstrate the level of the correlation between the two. The correlation coefficient (r) serves to express the strength and direction of a linear relationship between the predicted values and true values of the validation dataset. Values of r closer to 1 indicate stronger positive linear relationship, which represents better predictive performance of PLSR models.

In Figure 3.1, PLSR models of imputed datasets of MOR with EM ($r = 0.93$) and MI ($r = 0.93$) generate less spread plots than the model of non-imputed dataset does ($r = 0.91$). Figure 3.2 indicates that plots for models of EM-imputed dataset of IB ($r = 0.92$) and MI-imputed dataset of IB ($r = 0.92$) show less dispersion than the plot for the model of non-imputed dataset ($r = 0.89$) does. More plots are available in Appendix A. Those plots also illustrate that PLSR models based on imputed datasets provide better predictive ability.

We further compared and presented RMSEPs (in kPa) and NRMSEPs (%) for PLSR modeling of non-imputed and imputed datasets for MOR in following Table 3.3. PLSR models of datasets imputed with MCMC outperformed PLSR models of EM imputed datasets in seven of the ten cross-validations. The average RMSEP from models of the MCMC-imputed datasets is 678.24 kPa and 680.76 kPa from models of the EM-imputed datasets. In all cross-validations, non-imputed datasets consistently generated the highest RMSEPs and NRMSEPs with the average of 733.05 kPa and 6.3%, respectively. The average NRMSEP for models of MCMC-imputed datasets and EM-imputed datasets were both 5.8%.

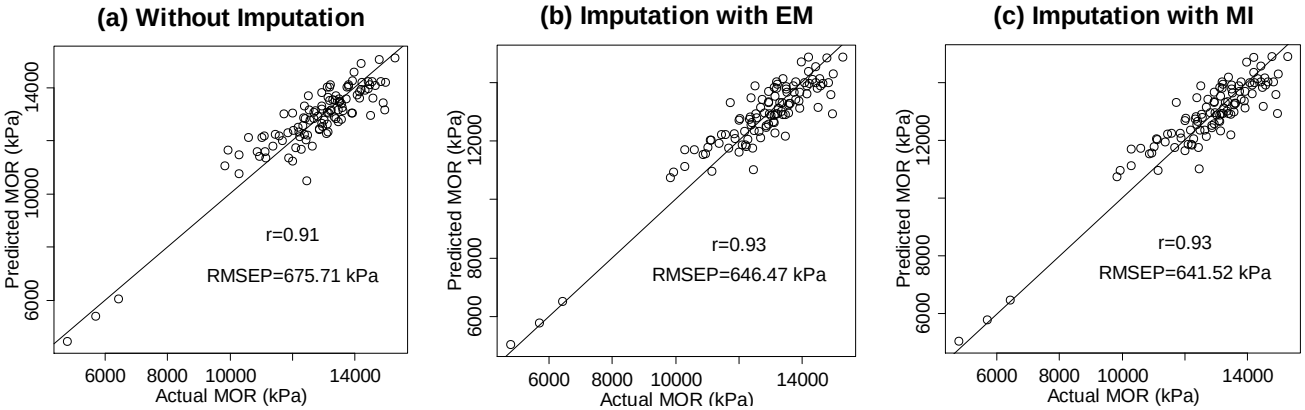


Figure 3.1 Plots of Predicted MOR (kPa) versus Actual MOR (kPa) from the Forth Validation out of Ten-fold Cross-validation.

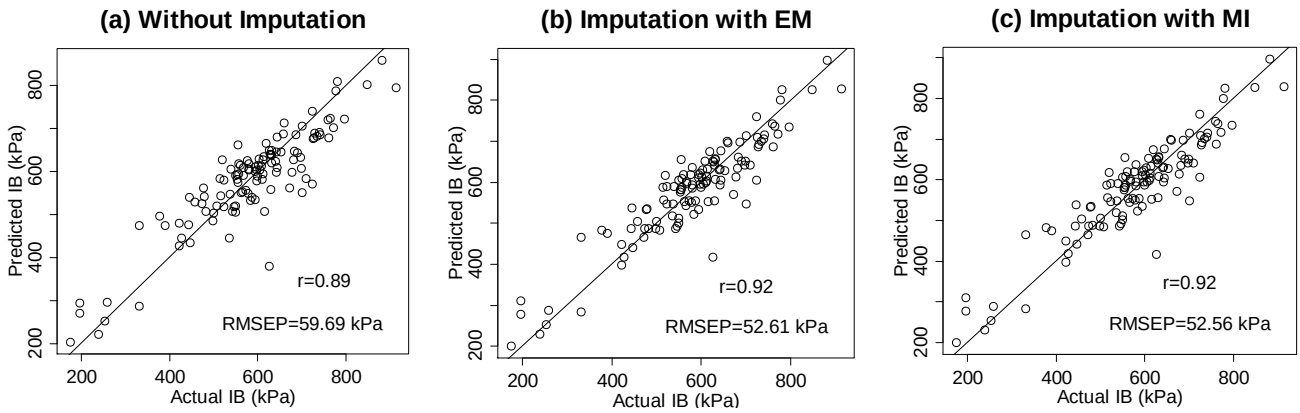


Figure 3.2 Plots of Predicted IB (kPa) versus Actual IB (kPa) from the Fourth Validation out of Ten-fold Cross-validation.

These results demonstrate the usefulness and benefits of imputation in predictive modeling as applied to manufacturing. The results also illustrate the impact of information loss on PLSR predictive models for the MOR strength of particleboard.

Comparative results of PLSR models for imputed and non-imputed datasets for IB are summarized in Table 3.4. Similar to the results for MOR (Table 3.3), EM imputed datasets had best prediction results in six of the ten cases when compared to the PLSR models developed from MCMC-imputed datasets. There is no notable difference between the average RMSEP from models of EM-imputed data (RMSEP = 51.10 kPa) and the one (51.20 kPa) from models of MCMC-imputed data. The NRMSEPs produced by models of imputed datasets (average NRMSEP = 7.2% for both EM and MCMC cases) are also consistently smaller than NRMSEP from models of non-imputed datasets (average NRMSEP = 8.1%). This demonstrates the potential benefits of imputation for PLSR predictive models of IB particleboard strength and further illustrates the impact of information loss.

Study results for both MOR and IB illustrate the benefit of imputation on PLSR model performance, i.e., smaller RMSEPs and NRMSEPs. Improved precision of predictive models using imputation may help practitioners better diagnose sources of variation and provide early detection signals of potential strength failures which result in potential customer claims. Such imputed datasets and predictive models may also reduce the practice of unnecessarily over-engineering the strength of wood composite panels. Over-engineering of particleboard by operations personnel are the result of a lack of real-

Table 3.3 RMSEPs (kPa) and NRMSEP(%) for PLSR on Non-imputed and Imputed Datasets for MOR Strength

RMSEP(NRMSEP)	PLSR[a] for Non-Imputed Dataset		PLSR[b] for Dataset Imputed with EM		PLSR[b] for Dataset Imputed with MCMC	
1	675.71	(6.4%)	646.47	(6.1%)	**641.52**	**(6.1%)**
2	669.52	(5.7%)	599.75	(5.1%)	**595.21**	**(5.0%)**
3	653.17	(5.5%)	**631.33**	**(5.3%)**	632.85	(5.3%)
4	813.40	(6.7%)	683.28	(5.6%)	**680.62**	**(5.6%)**
5	602.40	(5.5%)	569.42	(5.2%)	**565.45**	**(5.1%)**
6	848.90	(6.8%)	757.78	(6.0%)	**753.09**	**(6.0%)**
7	760.74	(7.0%)	716.93	(6.6%)	**712.18**	**(6.5%)**
8	768.12	(6.2%)	738.68	(6.0%)	**736.05**	**(6.0%)**
9	797.44	(6.7%)	772.14	(6.5%)	**768.40**	**(6.4%)**
10	741.10	(6.1%)	**691.85**	**(5.7%)**	697.00	(5.8%)
Average	733.05	(6.3%)	680.76	(5.8%)	**678.24**	**(5.8%)**

Table 3.4 RMSEPs (kPa) and NRMSEP(%) for PLSR on Non-imputed and Imputed Datasets for IB Strength

RMSEP(NRMSEP)	PLSR for Non-Imputed Dataset		PLSR for Dataset Imputed with EM		PLSR for Dataset Imputed with MCMC	
1	49.64	**(6.8%)**	46.59	**(6.4%)**	**45.95**	(**6.3**%)
2	59.69	(8.1%)	52.61	(7.1%)	**52.56**	**(7.1%)**
3	55.45	(8.2%)	**49.75**	**(7.3%)**	49.78	(7.3%)
4	55.52	(9.1%)	47.72	(7.8%)	**47.66**	**(7.8%)**
5	62.84	(8.9%)	57.14	(8.1%)	**56.95**	**(8.1%)**
6	53.36	(6.9%)	49.63	(6.4%)	**49.57**	**(6.4%)**
7	61.91	(9.0%)	53.56	(7.8%)	**53.44**	**(7.8%)**
8	62.00	(8.7%)	53.72	(7.5%)	**53.67**	**(7.5%)**
9	51.72	(6.5%)	45.18	(5.7%)	**45.13**	**(5.7%)**
10	60.87	(8.3%)	**56.02**	**(7.7%)**	56.15	(7.7%)
Average	57.30	(8.1%)	51.20	(7.2%)	**51.10**	**(7.2%)**

time knowledge of quality strength metrics which lead to higher than necessary targets of weight and resin, which also result in higher manufacturing costs and higher energy usage. Higher than necessary weight targets of wood-fiber composite panels also lead to lower wood utilization efficiency and an unwise use of the valuable forest resource.

Chapter 4
Predictive Modeling Method

From the perspective of data quality, as introduced and demonstrated above, we showed that missing value imputation methods EM and MI produced similar imputation results and were the best of six potential candidate methods. When we further used PLSR for prediction, models of EM-imputed and MI-imputed data produced more precise prediction results than models of non-imputed data did. In this chapter, we conducted comparison studies from the second perspective of predictive modeling. We selected four methods to model EM-imputed datasets.

4.1. Data Preparation and Assessment

Since EM method produced similar imputation results as MI using MCMC and took comparatively shorter time to compute, we imputed the aforementioned datasets using EM. The objective was to compare predictive modeling methods in order to achieve better modeling results utilizing as much information as possible. We imputed the original incomplete IB and MOR datasets without variable selection which had 4411 observations and 222 predictor variables, respectively. The reason why we chose EM over MCMC was because MCMC wouldn't converge when imputing the original dataset without variable selection. After imputation, we calculated VIF and created a correlation matrix for two complete datasets respectively to do preliminary analysis as was previously conducted in the study. Many of the predictor variables were highly correlated. Under multicollinearity, classical linear regression using ordinary least square (OLS) was not applicable for predictive modeling.

4.2. Modeling Methods

In earlier studies, we used PLSR to perform predictive modeling to study the impact of imputation. In this section we compare PLSR with three other modeling methods: Bayesian Additive Regression Trees (BART); LASSO; and Adaptive LASSO. When selecting from the above three modeling methods, we considered two major categories of methods in predictive modeling: parametric models and non-parametric methods (Muñoz and Felicísimo (2004)). LASSO and Adaptive LASSO are two parametric methods with constraints. As introduced in Chapter 2, LASSO is relatively new and has become a popular technique for variable selection and predictive modeling. Adaptive LASSO is a newer version of the LASSO, where adaptive weights are used to address certain problems of LASSO. BART is the latest non-parametric Bayesian regression approach developed on the basis of traditional "regression tree" method. All three methods have gained wide attention and popularity. Since the details of LASSO and PLSR have been introduced in Chapters 2 and 3 respective, we will focus on BART and Adaptive LASSO in the following section.

4.2.1. Bayesian Additive Regression Trees (BART)

BART was proposed and developed by Chipman et al. (2006, 2010), combining recent advances in Bayesian modeling with regression tree idea from machine learning to sensibly search the potentially high-dimensional space of possible models relating the response to a high-dimensional predictor variables. The BART method was suitable for the large-dimensional datasets of this study.

As suggested by the name of Bayesian Additive Regression Tree, BART consists of sum-of-trees model (additive model) and regularization prior (Bayesian method). In the following paragraphs we briefly introduce BART from three aspects: 1) sum-of-trees model; 2) regularization prior; and 3) MCMC algorithm. The following exposition closely follows Hill (2010) and Chipman et al. (2010).

BART is developed from regression or classification tree models. Tree models consist of binary trees with root nodes and child nodes. As illustrated in Figure 4.1, a tree model of data starts from a root node "A" consisting of entire data which are defined by two predictor variables (X_1 and X_2) and one response variable (Y). Then based on whether a predictor variable $X_1 \geq a$ *or* $X_1 < a$, the root node is split into two branches of the tree, generating two child nodes (B and C). Each child node represents a subset of the original data. Child node (C) can be split again using splitting rule ($X_2 \geq b$ *or* $X_2 < b$) based on another predictor variable X_2 into two branches to produce two more child nodes (D and E). Child node (B) becomes a terminal node without any child nodes. When a child node becomes terminal, a parameter value (e.g.,μ_1 which represents the mean of the subset of data that fall in this terminal) is returned. If the tree model is for prediction purpose, this returned parameter value plays the role of response (output) in a regression model. The growth of this single tree eventually partitions the original dataset into three different regions associated with three terminal nodes and three parameter values (μ_1, μ_2, and μ_3). When this single tree model is used for prediction, depending on which region the values of $X_1$and X_2 fall, a relevant parameter value μ will be returned as prediction result. If the μ is a real number, then this single tree is called regression tree;

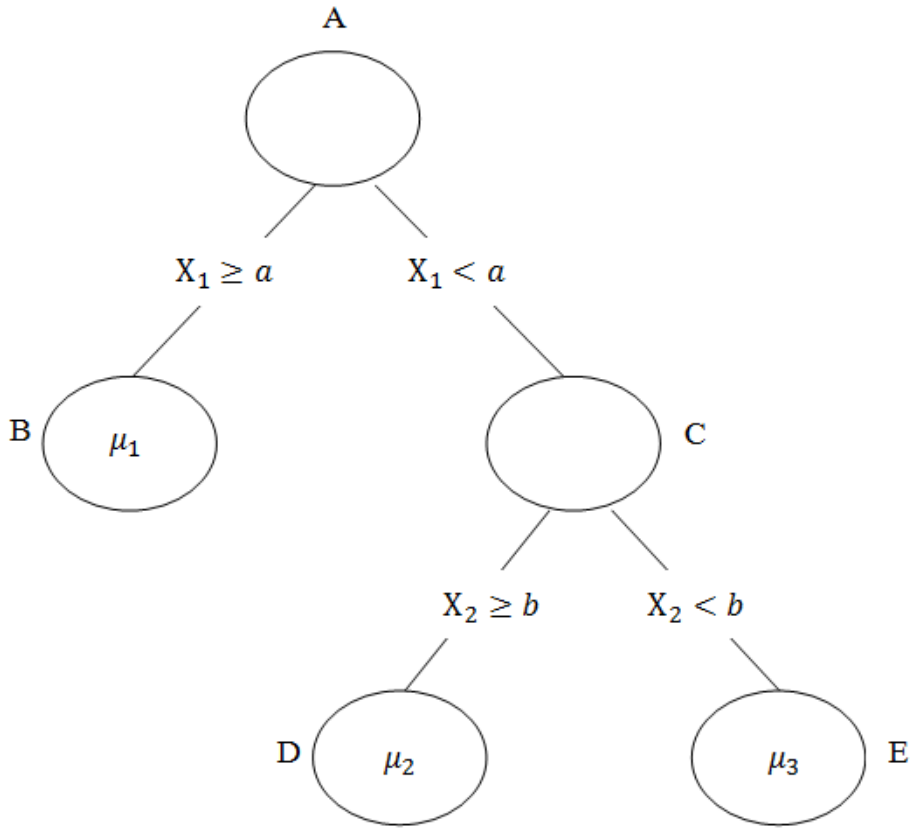


Figure 4.1 Illustration of a Single Tree Model.

if the μ is the class to which the data belongs, then the tree is a classification tree. For details of splitting rules of tree models, please refer to Hastie et al. (2009).

As a non-parametric method, the single-tree model is not limited to any parametric assumptions but suffers from a number of drawbacks (Green and Kern (2010)), so various methods have been developed to combine a set of single tree models. This combination idea is called ensemble learning in data mining/statistical learning area, i.e.,a predictive model is built by combining the strengths of a collection of simpler base models. Ensemble learning has been embedded into various methods such as bagging

and boosting (Hastie et al. (2009)). The development of BART was inspired by the boosting method (Chipman et al. (2010)). For more detail on the boosting method, please refer to Friedman (2001).

To further elaborate on BART, a series of concepts and notations are defined as follows. The "weak learner" refers to a "weak tree" that contributes a small amount to the predictive capability of the overall mode in terms of low R^2 and large RMSEP. The probability that a "weak tree" makes correct classification on some predictor *X* is not significantly different or better than random guessing. Let T_i denote a single tree of *B* terminal nodes, $M_i = \{\mu_1, \mu_2, \dots, \mu_B\}$ represent a set of parameter values associated with the *B* terminal nodes of T_i. Also instead of fitting a single tree, there are now *m* trees in the model. Function $g(x;\ T_i,\ M_i)$ is defined corresponding to specific $(T_i,\ M_i)$ which assigns a $\mu \in M$ to a subset of *X* associated with some terminal node. Now values of response variable *Y* can be expressed as

$$Y = \sum_{i=1}^{m} g(x;\ T_i, M_i) + \epsilon\,, \qquad \epsilon \sim N(0, \sigma^2) \tag{15}$$

The motivation behind above "sum-of-trees model" is to improve the precision of modeling/prediction results and combine the results of all "weak trees" to produce a "powerful committee." After building a single "weak tree" $g(x;\ T_1, M_1)$ to fit data, residuals are taken as the difference between the fit from the first tree and the observed response values *y*. Then a second "weak tree" is built to fit residuals. New residuals are formed and a third "weak tree" is built to fit them. This process is iterated until more of those "weak trees" eventually add up to a sum-of-trees model. The final modeling result

is the sum of all trees' terminal node parameters (M_i) assigned to an observation with predictors X defined by $g(x;\ T_i, M_i)$.

However, while building more trees to improve the accuracy of modeling, the probability of over-fitting increases. So the Bayesian approach using regularization priors is adopted to regularize the model fitting, i.e., a prior is placed on each of three major parameters in equation (15): the T_i, its terminal node parameters M_i, and σ. To simplify the prior specification, all T's are assumed to be independent on prior and identically distributed (i.i.d); all μ's of M are i.i.d given all T's; and σ are independent of all T's and μ's. For full details on prior setting, please refer to Chipman et al. (2010). There is also a more concise introduction of this method in Chipman et al. (2007).

After placing priors on parametersT_i, M_i, and σ, a posterior distribution $p\big((T_1, M_1), (T_2, M_2), \dots, (T_m, M_m), \sigma \big| y\big)$ is computed and sampled using Markov Chain Monte Carlo (MCMC). To be specific, Gibbs sampling is used. To demonstrate this, the same notation are used as in Chipman et al. (2010): let $T_{(i)}$ denote all other trees in the sum-of-trees except T_i; $M_{(i)}$ denote all other terminal node parameters of other trees except parameters M_i of tree T_i. As for sampling of (T_i, M_i), a sequence of samples from the joint distribution of (T_i, M_i) conditional on ($T_{(i)}$, $M_{(i)}$, σ) are drawn. Variances ($\sigma's$) are sampled from a distribution of σ conditional on all m (trees) of T_i and M_i, i.e., $(\sigma|T_1, \dots, T_m,\ M_1, \dots, M_m)$. Because σ^2 follows inverse chi-square distribution, i.e., $\sigma^2 \sim \frac{\nu\lambda}{\chi_\nu{}^2}$, where degrees of freedom ν and scale parameter λ are defined in prior setting, $\sigma's$ can be sampled using routine methods. For details on sampling of (T_i, M_i), again please refer to Chipman et al. (2010). Obviously, σ could be identified during each

MCMC iteration, however, (T_i, M_i) are mingled together. As summarized in Hill (2010), during MCMC iterations sampling parameters fitting Y in equation (15), the size of each of m trees may vary from iteration to iteration. The (T_i, M_i) pair for one tree maybe switched to another tree in next time's iteration; the contribution of a particular tree cannot be identified. However due to this lack of identification, MCMC in BART could lead to stable and rapid computation results. After a series of MCMC iterations (e.g., K iterations), a sample of K fitting results of Y, $\{\hat{Y}_1, \ldots, \hat{Y}_k\}$ are formed. To estimate or predict Y, the average of the sample is taken as

$$\hat{Y} = \frac{1}{K}\sum_{i=1}^{K} \hat{Y}_i \qquad (16)$$

4.2.2. Adaptive LASSO

As introduced before, LASSO is a regularization technique that through shrinking some coefficients of predictors to zero to simultaneously achieve variable selection and estimation (prediction). However, as mentioned in several previous studies such as Zou (2006), Leng et al. (2006), there are certain scenarios where the LASSO is inconsistent for variable selection. This is especially true when there are superfluous variables and when prediction accuracy (e.g., using cross-validation) are used as the criteria to choose tuning parameter (parameter λ in LASSO algorithm). A thorough discussion and literature review on this topic can be found in Huang (2006). To address this potential problem, Zou (2006) proposed Adaptive LASSO. For details of deriving Adaptive LASSO and the proof of how Adaptive LASSO improves the consistency of variable selection, please refer to the original paper by Zou (2006). The basic idea of Adaptive

LASSO is to apply adaptive weight vector $\widehat{\boldsymbol{w}}$ for penalizing different coefficients in the penalty term (l_1 penalty) in original LASSO algorithm as:

$$\hat{\beta}^{adaptive\ lasso} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \widehat{w_j}|\beta_j| \right\}. \quad (17)$$

In next section we discuss how the weight vector $\widehat{\boldsymbol{w}}$ and chose the tuning parameter λ are obtained.

4.3. Model Development

Ten-fold cross-validation of the aforementioned EM-imputed data was performed to develop and compare model quality. Take the dataset with MOR as response for example, the dataset was first randomly portioned into ten subsets of equal sample size 441. Every time one subset was retained as the validation dataset and remaining nine subsets were used to train the model. This process was repeated for ten times using PLSR, BART, LASSO, and Adaptive LASSO, respectively. These models were realized with self-written code and libraries in R (version 2.11.1). For the details of R-code please refer to the coding illustrated in Appendix C

4.3.1. BART Modeling

We used library "BayesTree" in R (Chipman et al. (2006)) to develop the BART model. There were two key aspects when modeling: 1) prior parameter set up; and 2) MCMC computation.

When setting parameters, we used the default settings recommended by Chipman et al. (2010). When performing MCMC computation, iterations were repeated until satisfactory convergence was reached. We assessed convergence by monitoring σ drawsover time. Time series plots of σ in MCMC iterations of IB and MOR data for the fifth validation of ten-fold cross-validation are given in Figure 4.2. The initial parts where the σ draws reduce rapidly represent the "burn-in" period of the Markov Chain. This period was not included in the average of MCMC iterations as in equation (16). After the σ draws level off and flatten (latter part of two plots), improvement declines and convergence is assumed. Based on the convergence situation of this study, 1,500 iterations for the MOR data and 1,000 iterations for the IB data were used.

When predicting the response in the validation dataset, each MCMC iteration generates a sum-of-trees model. There were a sample of 1,500 predicted results for one observation in the MOR validation data and a sample of 1,000 predicted results for one observation in IB validation data. An example of MCMC iteration results from the fifth validation for both MOR and IB data is presented in Figure 4.3. To get the "point estimator" for predicted response of MOR and IB the average of each sample was taken.

4.3.2. Adaptive LASSO Modeling

Similar to the previous variable selection using LASSO (Chapter 2), we used the same library "lars" in R to perform the computation for the LASSO modeling process. The tuning parameter "λ" in equation (3) was estimated using cross-validation. The key part of this technique was to apply the appropriate weight "$\widehat{w_j}$" in equation (3). When

doing computation, we followed the following algorithm (which is more detail than the one suggested in Zou (2006)):

- First "decenter" and "scale" all predictor variables using the method detailed in Appendix C coding part;
- Perform ridge regression (refer to Hastie et al.(2009) for details on ridge regression) to original data to obtain model coefficients "$\widehat{w_j}$" except for the intercept;
- Apply above coefficients to "standardized" predictor variables from the first step;
- Perform regular LASSO using "lars" function in R to above weighted predictor variables and do cross-validation for tuning parameter "λ" in LASSO regression;
- Since above LASSO estimators are based on processed predictor variables, those estimators are "scaled" back.
- Predictor variables with non-zero estimators as coefficients are chosen for prediction.

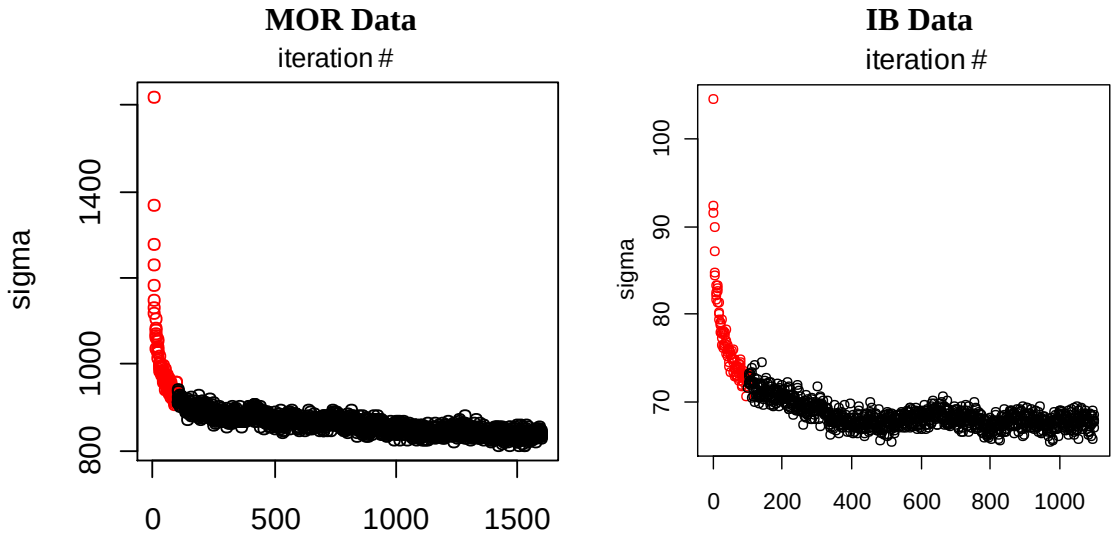


Figure 4.2 Plots of σ against Iteration Number in MCMC Computation of BART from the fifth validation.

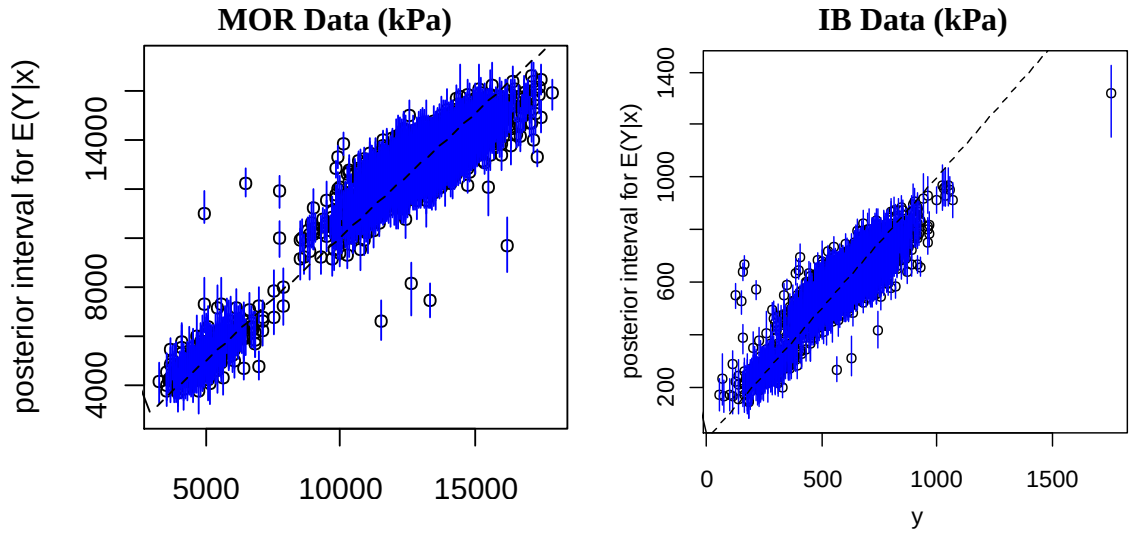


Figure 4.3 Plots of Iteration Results in MCMC Computation of BART from the fifth validation (Each Individual Vertical Line Represents 1,000/1,500 Predicted Results for One Individual Response Y in Validation Dataset; Each Dot Represents the Average of 1,000/1,500 Iteration Results).

In the original paper about Adaptive LASSO, Zou (2006) suggested using coefficients of ordinary least squares (OLS) regression as adaptive weights for most of occasions. However, considering the multicollinearity of our data, we chose ridge regression as another alternative approach to obtain weights.

4.4. Model Comparison

In comparing modeling results, we used RMSEP, Normalized RMSEP (NRMSEP), and Coefficient of Variation of RMSEP or CV (RMSEP). To evaluate the model fitting of BART, PLSR, LASSO, and Adaptive LASSO for EM-imputed data of MOR and IB, we also plotted predicted MOR and IB against actual MOR and IB from ten-fold cross-validation for the four methods, respectively. The third validation for four models of MOR and IB data are given in Figures 4.4 and 4.5. The remaining plots of MOR and IB data are presented in Appendix B.

Less dispersion of the plots of BART models of both MOR and IB data represents better prediction precision than the other three models. We also calculated the correlation coefficient (r) between the true values and predicted values to demonstrate the level of the correlation between the two. As noted in figures of plots, values of r larger and closer to 1 indicate stronger positive linear relationship, which implies that BART models or MOR data (r = 0.92) and IB data (r = 0.83) outperformed other three methods. Smaller RMSEPs of BART models in both figures are also exemplified.

A complete comparison of the four methods in terms of RMSEP, NRMSEP, CV (RMSEP), and correlation coefficient (r) across ten validations is given in Tables 4.1, 4.2, 4.3, and 4.4..

The comparison results for models of MOR data in Tables 4.1 and 4.2 suggest that BART produce best prediction results for eight of the ten validations. BART had the smallest average RMSEP (965.75 kPa), average NRMSEP (7.1%), and average CV (RMSEP) (7.7%). Smallest NRMPSEP and CV imply that prediction performance of BART has the least variation and is more consistent than other methods. Correspondingly, the BART model had the largest correlation coefficient for eight of the ten validations (average r = 0.91). Although there is notable gap compared with the prediction performance of BART, PLSR is the second best method with the smallest RMSEP (average 1005.83 kPa), NRMSEP (average 7.3%), and CV (RMSEP) (average 8.1%) for two of the ten validations. PLSR also has largest correlation coefficient (average r = 0.90) for two of the validations. There is no apparent difference in prediction performance between LASSO and Adaptive LASSO models. LASSO models appear to be slightly better (average RMSEP = 1,073.89 kPa) than Adaptive LASSO (average RMSEP = 1,077.19 kPa).

The comparison results for models of IB data in Tables 4.3 and 4.4 show similar results. Overall, BART has the best prediction performance, with the smallest RMSEP (average of 76.79 kPa), NRMSEP (average of 8.6%), CV (RMSEP) (13.1%) and the largest correlation coefficient (average of 0.83). For seven of the ten validations, BART outperformed the other three modeling methods.

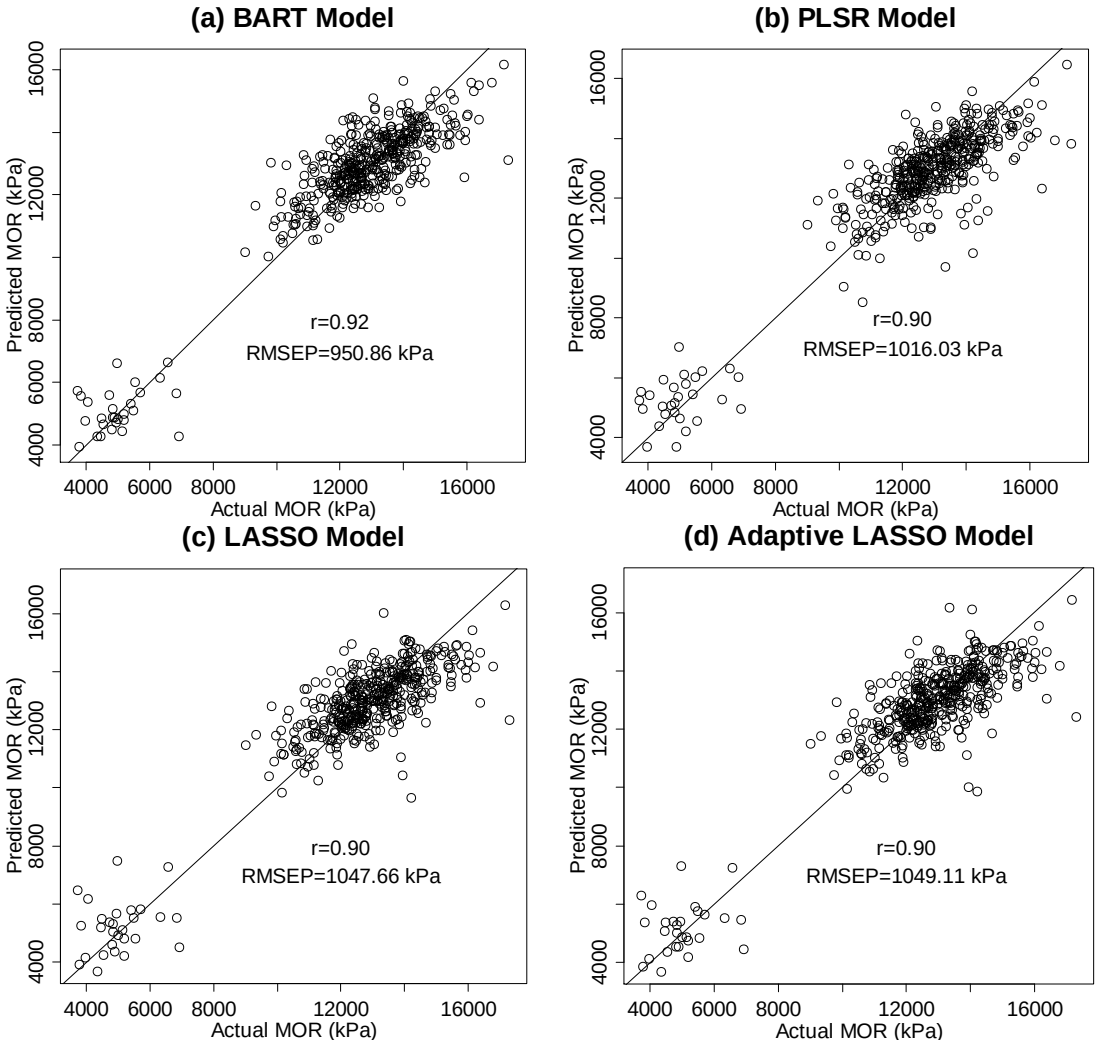


Figure 4.4 Plots of Predicted MOR (kPa) versus Actual MOR (kPa) for Four Models from the 3rd Validation of 10-fold Cross-Validation.

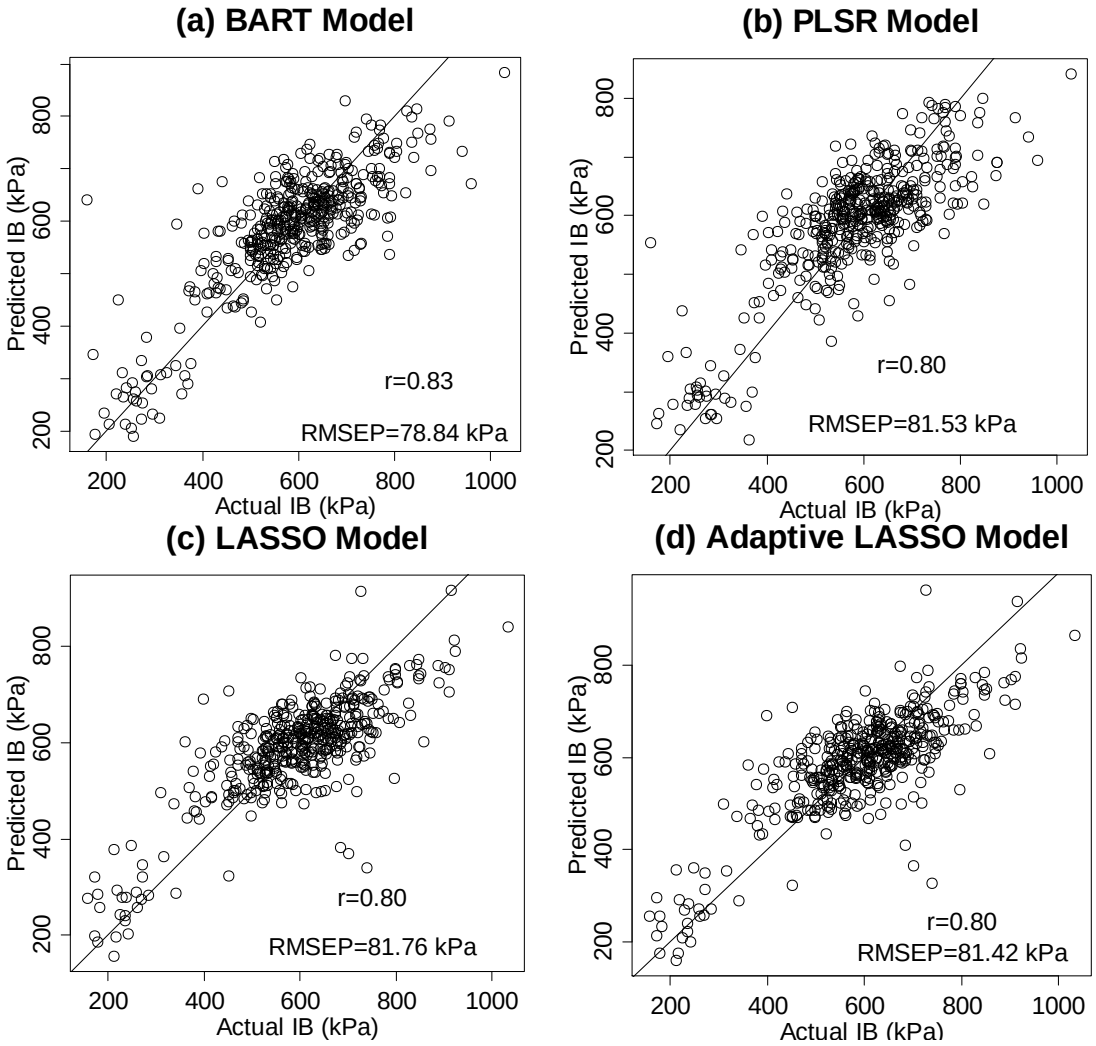


Figure 4.5 Plots of Predicted IB (kPa) versus Actual IB (kPa) for Four Models from the 3rd Validation of 10-fold Cross-Validation.

However, BART is only slightly better than PLSR model with average RMSEP of 77.07 kPa, NRMSEP of 8.9%, CV (RMSEP) of 13.2%. The PLSR model also has the same average correlation coefficient 0.83 as the BART model. Comparatively, prediction results of BART and PLSR models are apparently better than results of LASSO and Adaptive LASSO models. There is still no notable performance difference between LASSO (average RMSEP = 84.11) and Adaptive LASSO (average RMSEP = 83.79). Adaptive LASSO is slightly more precise. Since the BART method is relatively new, the literature on the application of BART is sparse. We noted similar results and remarks on the superior performance of BART than for other methods as noted in Chipman et al. (2006), Hill (2010), Green and Kern (2010).

Adaptive LASSO and LASSO didn't improve model quality. Their performance may suffer from the multicollinearity of the datasets. Zou (2006) brought up the concern on collinearity and suggested trying ridge regression when fitting Adaptive LASSO for more stable results. Although we did so to compensate for better consistency of prediction, the precision of models did not improve. The method of ten-fold cross-validation that we used to choose the tuning parameter (λ) may also have affected the results of LASSO and Adaptive LASSO, as documented in Leng et al. (2006), and Martinez et al. (2010).

Table 4.1 RMSEPs (kPa) and NRMSEP(%) of 10-fold Cross-validation of BART, PLSR, LASSO, and Adaptive LASSO Modeling of MOR Strength

RMSEP (NRMSEP)	BART		PLSR		LASSO		Adaptive LASSO	
1	**1069.15**	**(8.4%)**	1096.71	(8.6%)	1165.56	(9.1%)	1195.96	(9.4%)
2	**920.30**	**(7.0%)**	1106.70	(8.2%)	1060.02	(7.9%)	1160.76	(8.6%)
3	**950.86**	**(7.0%)**	1016.03	(7.5%)	1047.66	(7.7%)	1049.11	(7.7%)
4	1032.70	(7.6%)	**1010.63**	**(7.5%)**	1166.05	(8.6%)	1086.44	(8.0%)
5	**971.58**	**(7.2%)**	1016.04	(6.8%)	1048.19	(7.7%)	1047.62	(7.7%)
6	**917.05**	**(6.7%)**	925.14	(6.8%)	963.87	(7.1%)	941.71	(6.9%)
7	**956.86**	**(6.9%)**	970.35	(7.0%)	1079.19	(7.8%)	1078.68	(7.8%)
8	**881.62**	**(6.3%)**	967.43	(7.0%)	998.49	(7.2%)	1045.33	(7.6%)
9	**897.59**	**(6.2%)**	952.96	(6.6%)	1093.91	(7.6%)	1056.30	(7.4%)
10	1059.80	(7.8%)	**996.30**	**(7.3%)**	1115.92	(8.2%)	1109.93	(8.2%)
Average	**965.75**	**(7.1%)**	1005.83	(7.3%)	1073.89	(7.9%)	(1077.19)	(7.9%)

Table 4.2 Correlation and CV(RMSEP) of 10-fold Cross-validation of BART, PLSR, LASSO, and Adaptive LASSO Modeling of MOR Strength

Correlation (CV)	BART		PLSR		LASSO		Adaptive LASSO	
1	**0.90**	**(8.6%)**	0.89	(8.8%)	0.88	(9.3%)	0.88	(9.6%)
2	**0.92**	**(7.3%)**	0.88	(8.8%)	0.90	(8.8%)	0.88	(9.2%)
3	**0.92**	**(7.6%)**	0.90	(8.1%)	0.90	(8.4%)	0.90	(8.4%)
4	0.91	(8.2%)	**0.91**	**(8.1%)**	0.88	(9.3%)	0.90	(8.7%)
5	**0.91**	**(7.8%)**	0.90	(8.1%)	0.90	(8.4%)	0.90	(8.4%)
6	**0.90**	**(7.2%)**	0.90	(7.3%)	0.89	(7.6%)	0.90	(7.4%)
7	**0.91**	**(7.6%)**	0.91	(7.9%)	0.88	(8.6%)	0.88	(8.6%)
8	**0.93**	**(7.1%)**	0.91	(7.8%)	0.91	(8.0%)	0.90	(8.4%)
9	**0.93**	**(7.1%)**	0.92	(7.6%)	0.89	(8.7%)	0.90	(8.4%)
10	0.91	(8.5%)	**0.92**	**(7.9%)**	0.90	(8.9%)	0.90	(8.9%)
Average	**0.91**	**(7.7%)**	0.90	(8.1%)	0.89	(8.6%)	0.89	(8.90%)

Table 4.3 RMSEPs (kPa) and NRMSEP(%) of 10-fold Cross-validation of BART, PLSR, LASSO, and Adaptive LASSO Modeling of IB Strength

RMSEP(NRMSEP)	BART		PLSR		LASSO		Adaptive LASSO	
1	74.9	(7.8%)	**73.6**	**(7.7%)**	80.06	(8.4%)	80.15	(8.4%)
2	80.91	(7.0%)	**75.50**	**(7.8%)**	85.08	(8.7%)	83.34	(8.5%)
3	78.64	(8.6%)	**75.49**	**(8.5%)**	85.63	(9.4%)	86.65	(9.5%)
4	**76.19**	**(8.4%)**	77.53	(8.5%)	84.11	(9.2%)	82.69	(9.0%)
5	**77.22**	**(9.2%)**	78.19	(9.3%)	86.84	(10.0%)	86.18	(10%)
6	**81.90**	**(11%)**	82.35	(11%)	87.27	(12%)	89.14	(12%)
7	**73.59**	**(9.3%)**	75.63	(9.6%)	80.69	(10.2%)	80.27	(10.1%)
8	**70.63**	**(7.6%)**	74.69	(8.0%)	84.57	(10.9%)	84.11	(10.6%)
9	**75.14**	**(8.6%)**	76.19	(8.7%)	85.08	(9.7%)	83.97	(9.6%)
10	**78.84**	**(9.1%)**	81.53	(9.4%)	81.76	(9.4%)	81.42	(9.3%)
Average	**76.79**	**(8.6%)**	77.07	(8.9%)	84.11	(9.8%)	83.79	(9.7%)

Table 4.4 Correlation and CV(RMSEP) of 10-fold Cross-validation of BART, PLSR, LASSO, and Adaptive LASSO Modeling of IB Strength

Correlation (CV)	BART		PLSR		LASSO		Adaptive LASSO	
1	0.82	(12.6%)	**0.83**	**(12.4%)**	0.79	(13.4%)	0.79	(13.5%)
2	0.82	(13.9%)	**0.85**	**(13.0%)**	0.80	(14.6%)	0.81	(14.3%)
3	0.82	(13.6%)	**0.83**	**(13.5%)**	0.78	(14.9%)	0.78	(15%)
4	**0.85**	**(13.2%)**	0.84	(13.4%)	0.81	(14.7%)	0.82	(14.3%)
5	**0.84**	**(13.1%)**	0.83	(13.3%)	0.80	(14.6%)	0.80	(14.6%)
6	**0.82**	**(14.3%)**	0.81	(14.4%)	0.79	(15.2%)	0.78	(15.6%)
7	**0.83**	**(12.5%)**	0.82	(12.9%)	0.79	(13.7%)	0.80	(13.6%)
8	**0.85**	**(11.8%)**	0.83	(12.5%)	0.80	(13.0%)	0.80	(12.9%)
9	**0.85**	**(12.7%)**	0.84	(12.9%)	0.79	(14.4%)	0.80	(14.2%)
10	**0.82**	**(13.4%)**	0.80	(14.0%)	0.80	(14.0%)	0.80	(13.9%)
Average	**0.83**	**(13.1%)**	0.83	(13.2%)	0.80	(14.3%)	0.80	(14.2%)

Chapter 5
Conclusion and Recommendation

This thesis research addressed two aspects of predictive modeling of strength properties of wood composites. The first was the missing data problem (data quality) and the second was selection of predictive modeling methods.

Study results for the first part of thesis on missing data imputation indicated that data imputation and variable selection based on the LASSO method prior to the development of partial least squares regression (PLSR) predictive models for MOR and IB strength properties greatly improved model performance. The LASSO method for variable selection prior to data imputation has certain advantages over other methods (e.g., principal component analysis (PCA) and genetic algorithm (GA)) as indicated by study results. In this study multiple imputation (MI) with MCMC and maximum likelihood method using the EM algorithm outperformed other imputation methods such as mean/median substitution, simple random imputation, or last-observation-carried-forward method. PLSR models developed from EM and MI imputed data had better model performance than PLSR models developed from non-imputed data.

The second part of thesis compared four predictive modeling methods: Bayesian Additive Regression Tree (BART), PLSR, LASSO, and Adaptive LASSO. The BART method provided best performance over other three methods in predicting MOR and IB. The relative short computation time (20 to 25 minutes in CPU time) spent on BART demonstrated the applicability of the method in real-time industrial settings. The PLSR

method was the second best prediction method. There was no notable difference in model performance between the LASSO and Adaptive LASSO methods.

Given the difficulties associated with the missing data problem and predictive model selection common to industrial settings, this thesis study demonstrated the value of LASSO variable selection with data imputation for PLSR predictive models of the strength quality metrics for wood composites. Such predictive models with imputation may also help practitioners understand sources of process variation and reduce overall manufacturing costs.

Predictive models of strength properties using BART for manufacturing settings may help practitioners maintain product quality specification and prevent claim costs. Accurate real-time predictive models may also discourage operational practices of running higher than necessary feedstock targets.

Chapter 6
Future Research

Future research topics originating from current study may be plentiful. Three future possible research topics are presented as part of this thesis.

First, missing data imputation and predictive modeling using BART may be expanded beyond wood composite manufacturing and also a real-time data test in a manufacturing mill seems plausible. Applications of this study for other manufacturing areas (e.g., paper, aluminum, steel, etc.) appear to be a logical extension of the research. Such an extension would greatly benefit practitioners interested in continuous improvement using statistical methods. New avenues of research may include the use of BART with larger datasets than the ones used in this research.

A second research topic will be to improve the interpretation of BART models. Currently some literature is available on BART model interpretation (Chipman et al. (2010)). However, for BART models to be more beneficial for the practitioner for continuous improvement, prioritizing key predictor variables of the process are needed.

A third potential research topic is the choice of the tuning parameter λ for the LASSO and Adaptive LASSO methods. The choice of "fold" in *v*-cross-validation also affects LASSO and Adaptive LASSO modeling results and requires further investigation. Further extensive simulation studies to explore these areas are planned.

LIST OF REFERENCES

Andr é, N., Cho, H.W., Baek, S.H., Jeong, M.K. and Young, T.M. (2008). "Enhanced Prediction of Internal Bond Strength In A Medium Density Fiberboard Process Using Multivariate Methods And Variable Selection." *Wood Science and Technology*. 42, pp.521-534.

Allison, P. (2000). "Multiple Imputation For Missing Data: A Cautionary Tale." *Sociological Methods and Research*, 28, pp.301-309.

Altmayer, L. (2002). "Hot-Deck Imputation: A Simple DATA Step Approach." *Proceedings of the 2002 Northeast SAS User's Group*. Buffalo, NY: Northeast SAS User's Group. pp.773–780

Barnes, D. (2001). "A Model of the Effect of Strand Length and Strand Thickness on the Strength Properties of Oriented Wood Composites." *Forest Product Journal*, 51, 9, pp.36–46.

Chipman, H.A., George, E.I., and McCulloch, R.E. (2006). "Bayesian Ensemble Learning." *NIPS* 2006.

Chipman, H.A., George, E.I., and McCulloch, R.E. (2010). "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics*, 4, pp. 266-298.

Chipman, H.A., George, E.I., Lemp, J., and McCulloch, R.E. (2010). "BART: Bayesian Fleixble Modelling of Trip Durations." *Transportation Research* B, 44, pp. 686-698.

Clapp, N.E., Jr., Young, T.M., and Guess, F.M. (2008). " Predictive Modeling The Internal Bond of Medium Density Fiberboard Using A Modified Principal Component Analysis." *Forest Products Journal*, 58, 4, pp.49-55.

Collins, L.M., Schafer, J.L. and Kam, C.M. (2001). "A Comparison Of Inclusive And Restrictive Strategies In Modern Missing-Data Procedures." *Psychological Methods*, 6, 330-351.

Draper, D. and Fouskakis, D. (2000). "A Case Study of Stochastic Optimization in Health Policy: Problem Formulation and Preliminary Results." *Journal of Global Optimization*, 18, pp.399-416.

Datta, S., Le-Rademacher, J., and Datta, S. (2007). "Predicting Patient Survival From Microarray Data By Accelerated Failure Time Modeling Using Partial Least Squares And LASSO." *Biometrics*, 63, 1, pp.259-271.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "A New Method For Variable Subset Selection, with The Lasso and "Epsilon" Forward Stagewise Methods As Special Cases. LARS Software For R And Splus." *Annals of Statistics (with discussion*, 32, 2, pp.407-499.

Enders, C.K. (2001). "The Impact Of Nonnormality On Full Information Maximum-Likelihood Estimation For Structural Equation Models With Missing Data." *Psychological Methods*, 6, pp.352–370.

Faraway, J.J. (2005). *Linear Models with R*, Chapman & Hall, New York.

Fetter, M. (2001). "Mass imputation of agricultural economic data missing by design: a simulation study of two regression based techniques", *Federal Conference on Survey Methodology*, http://www.fcsm.gov/01papers/Fetter.pdf.

Friedman, J.H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 29, 5, pp.1189-1232.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Gonźalez, I. and Sánchez, I. (2010). "Variable Selection in Multivariate Statistical Process Control." *Journal of Quality Technology*, 42, 3, pp. 242-259.

Green, D.P. and Kern, H.L. (2010). "*Modeling Heterogeneous Treatment Effects in Large-scale Experiments using Bayesian Additive Regression Trees*." http://andrewgelman.com/movabletype/mlm/Green%20and%20Kern%20BART.pdf (Last Accessed: July 29, 2011)

Guyon, I. and Elisseeff, A. (2003). "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research*, 3, pp.1157-1182.

Hamer, R.M. (2009). "Last Observation Carried Forward Versus Mixed Models in the Analysis of Psychiatric Clinical Trials." *American Journal of Psychiatry*, 166, pp.639-641.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd ed.* Springer-Verlag, New York.

Hill, J. (2011). "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics*, 20, 1, pp.217-240.

Horton, N.J., Kleinman, K.P. (2007). "Much Ado About Nothing: A Comparison Of Missing Datamethods And Software To Fit Incomplete Data Regression Models." *American Statistician*, 61, 1, pp.79-90.

Jensen, W A., Jeffrey, B.B., and Woodall, W.H. (2008). "Monitoring Correlation within Linear Profiles Using Mixed Models." *Journal of Quality Technology*, 40, pp. 167-183.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review*, 95, 1, pp.49-69.

Kourti, T., Lee, J., and MacGregor, J.F. (1996). "Experiences with Industrial Applications of Projection Methods for Multivariate Statistical Process Control." *Computers in Chemical Engineering*, 20, pp.745-750.

Lanning, D. and Berry, D. (2003). "An Alternative to PROC MI for Large Samples." *SAS Users Group International* (SUGI) 28, Seattle, Washington.

Lavori, P.W., Dawson, R., and Shera, D. (1995). "A Multiple Imputation Strategy For Clinical Trials with Truncation Of Patient Data." *Statistics in Medicine*, 14, pp.1913–1925.

Leng, C. and Zhang, H. (2006). "Model Selection in Nonparametric Hazard Regression." *Nonparametric Statistics*, 18, 7, pp.417-429.

Lin, T.H. (2010). "A Comparison Of Multiple Imputation With EM Algorithm And MCMC Method for Quality Of Life Missing Data." *Qual Quant*, 44, pp.277-287.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* John Wiley, New York, NY.

Little, R.J.A. (1992). "Regression with Missing X's: A Review." *Journal of the American Statistical Association*, 87, pp.1227-1237.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. John Wiley, New York, NY.

Mevik, B.H. and Wehrens, R. (2007). "The pls Package: Principal Component and Partial Least Squares Regression in R." *Journal of Statistical Software*, 18, 2, pp.1-24.

McGee M. and Bergasa, N.V. (2005). "*Imputing Missing Data in Clinical Pilot Studies.*" *Technical Report*. Dallas, TX: Southern Methodist University.

Muñoz, J. and Felicísimo, A.M. (2004). "Comparison of Statistical Models Commonly Used in Predictive Modeling." *Journal of Vegetation Science*, 15, 2, pp. 285-292.

Muthen, B., Kaplan, D., Hollis, M. (1987). "On Structural Equation Modeling With Data That Are Not Missing Completely At Random." *Psychometrica*, 52, pp.431-462.

Newman, D.A. (2003). "Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques." *Organizational Research Methods*, 6, 3, pp.328-362.

Ni, D., Leonard, J.D., II, Guin, A., and Feng, C. (2005). "Multiple Imputation Scheme for Overcoming the Missing Values and Variability Issues in ITS Data." *Journal of Transportation Engineering*, 131, 12, pp. 931-938.

Parsons, L., Haque, E., and Liu, H. (2004). "Evaluating Subspace Clustering Algorithms." *SIAM International Conference on Data Mining*, pp. 48–56.

Patterson, S. and Yeh, S-T. (2007). "SAS-based MCMC Modelling and Bayesian Statistics. *Pharma SAS User Group*

Petersen, J.J., Snee, R.D., McAllister, P.R., Schofield, T.L, and Carella, A. (2009). "Statistics In Pharmaceutical Development And Manufacturing." *GlaxoSmithKline Pharmaceuticals BDS Technical Report.*

Rubin, D.B. (1996). "Multiple imputation after 18+ years (with discussion)." *Journal of the American Statistical Association*, 91, pp.473-489.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Schafer, J. L. and Olsen, M. K. (1998). "Multiple Imputation For Multivariate Missing-Data Problems: A Data Analyst's Perspective." *Multivariate Behavioral Research*, 33, 4, pp.545-571.

Sinharay, S., Stern, H.S., and Russell, D. (2001). "The Use of Multiple Imputation for the Analysis of Missing Data." *Psychological Methods*, 6, 4, pp.317-329.

Sjöblom, E., Johnsson, B., and Sundström, H. (2004). "Optimization Of Particleboard Production Using Nirspectroscopy And Multivariate Techniques." *Forest Product Journal*, 54, 6, pp.71–75.

Soh, C.S., Ong, K.M., and Raveendran, P. (2005). "Variable Selection Using Genetic Algorithm for Analysis of Near-infrared Spectral Data Using Partial Least Squares." *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China.*

Soullier, N., Rochebrochard, E., and Bouyer, J. (2010). "Multiple Imputation for Estimation of an Occurrence Rate in Cohorts with Attrition and Discrete Follow-up Time Points: A Simulation Study." *BMC Medical Research Methodology*, 10, 79, pp.1-7.

Tibshirani, R. (1996). "Regression Shrinkage And Selection Via The *Lasso." Journal of the Royal Statistical Society*. 58, 1, pp.267-288.

Truxillo, C. (2005). "Maximum Likelihood Parameter Estimation with Incomplete Data." *Proceedings of the Thirtieth Annual SAS(R) User Group International Conference.*

UCLA: Academic Technology Services, Statistical Consulting Group. "Multiple Imputation in SAS, Part1." http://www.ats.ucla.edu/stat/sas/seminars/missing_data/part1.htm (Accessed April 24, 2011)

Wang, K. and Jiang, W. (2009). "High-Dimensional Process Monitoring and Fault Isolation via Variable Selection." *Journal of Quality Technology*, 41, 3, pp.247-258.

Yarandi, H.N. (2002). "Handling Missing Data with Multiple Imputation Using PROC MI in SAS." *Proceedings of the Southeast SAS User Group, Savannah, GA*

Yuan, Y.C. (2000). "Multiple Imputation for Missing Data: Concepts and New Development." *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, pp. 267-25.

Young, T.M., Andrė, N., and Huber, C.W. (2004). "Predictive Modeling Of The Internal Bond Of MDF Using Genetic Algorithms with Distributed Data Fusion*." Proceedings of the Eighth European Panel Products Symposium*, pp.45–59.

Zhu, M., Chipman, H.A. (2006). "Darwinian Evolution in Parallel Universes: A Parallel Genetic Algorithm for Variable Selection." *Technometrics*, 48, 4, pp. 491-502.

Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties." *Journal of American Statistical Association*, 101, 476, pp. 1418-1429.

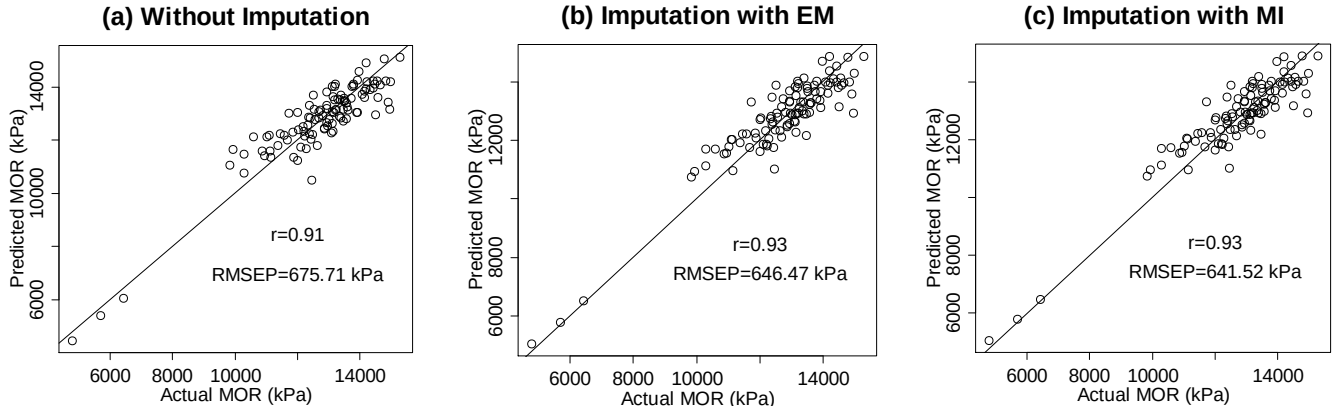
APPENDIX

Appendix A

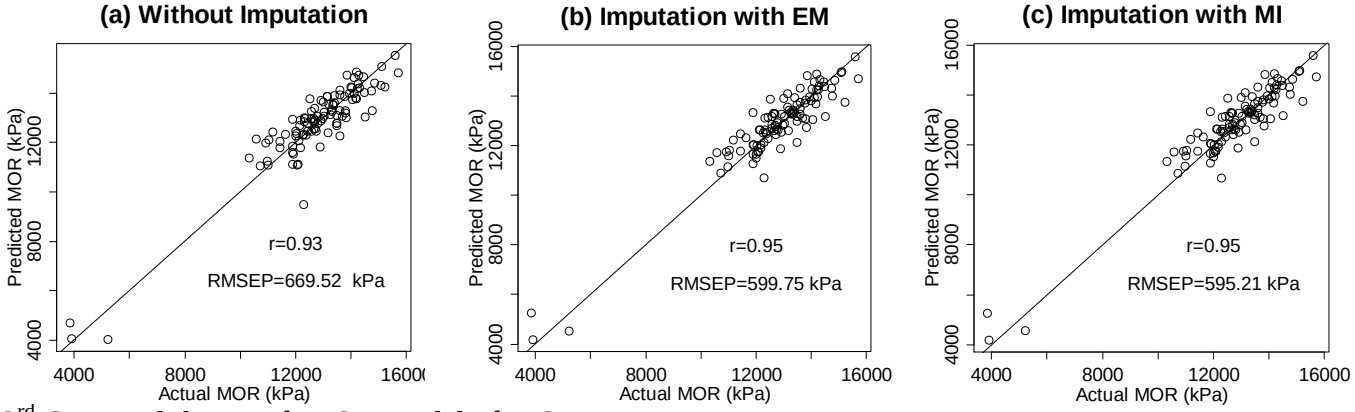
Appendix A.1.

Ten-fold Cross-validation of PLRS Modeling of Imputed and Non-imputed MOR Data

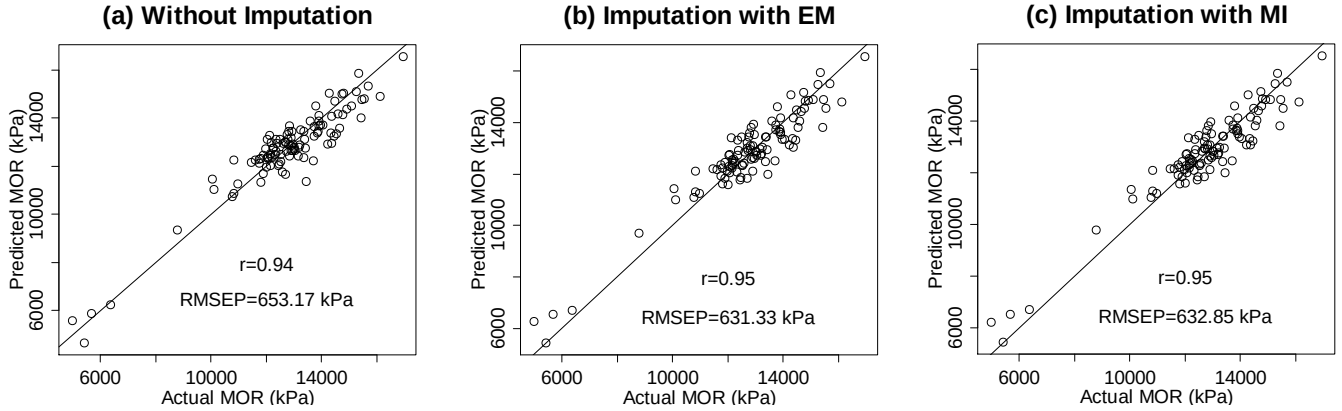
1st Cross-validation of PLSR Model of MOR Data



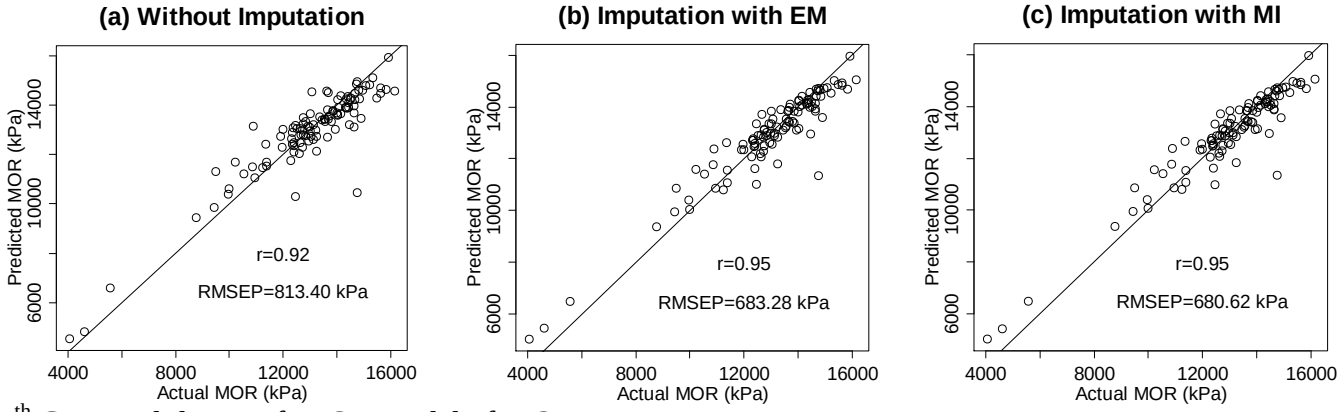
2nd Cross-validation of PLSR Model of MOR Data



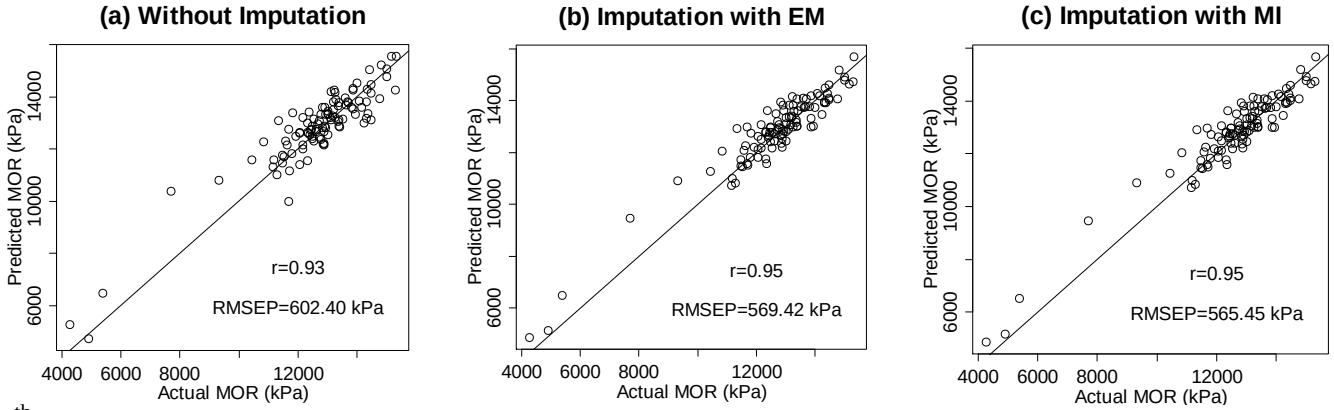
3rd Cross-validation of PLSR Model of MOR Data



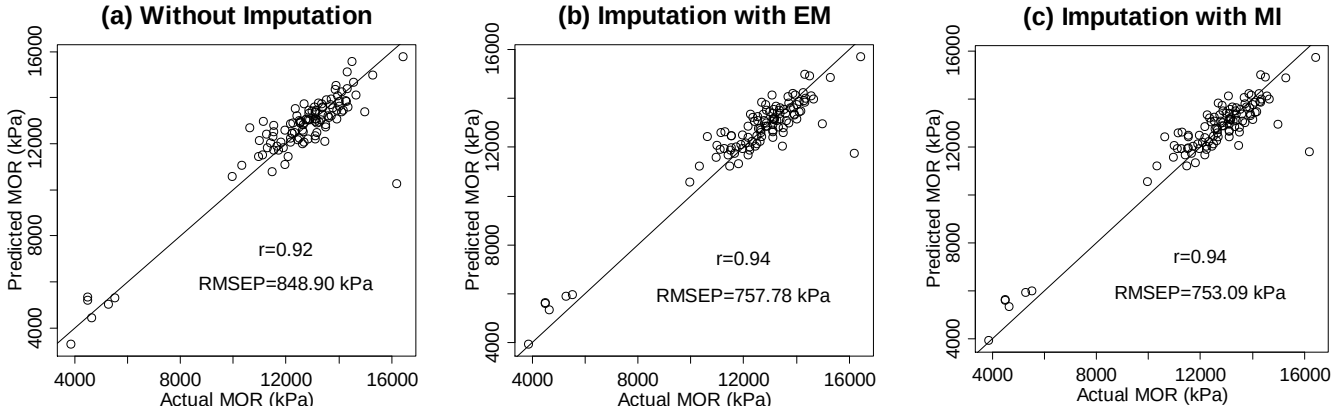
4th Cross-validation of PLSR Model of MOR Data



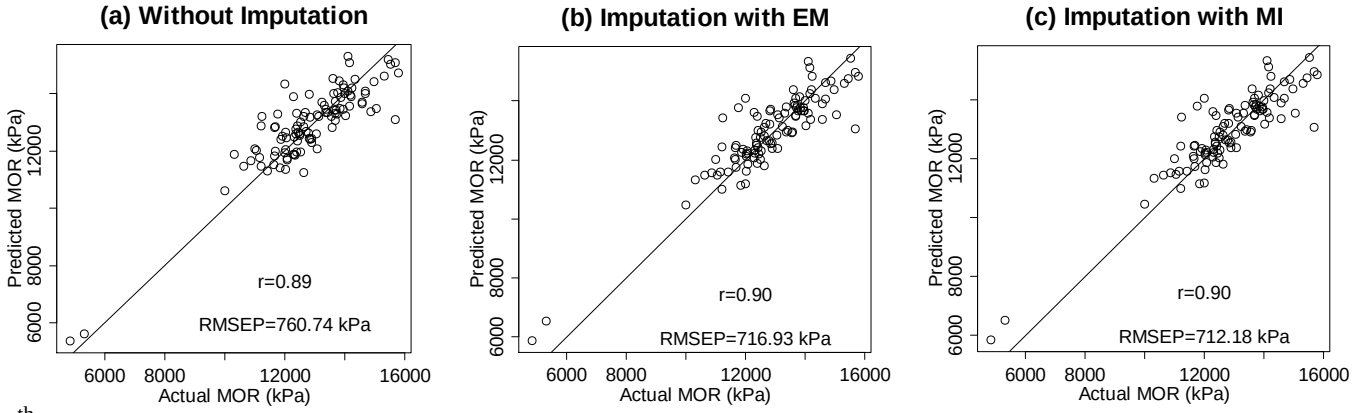
5th Cross-validation of PLSR Model of MOR Data



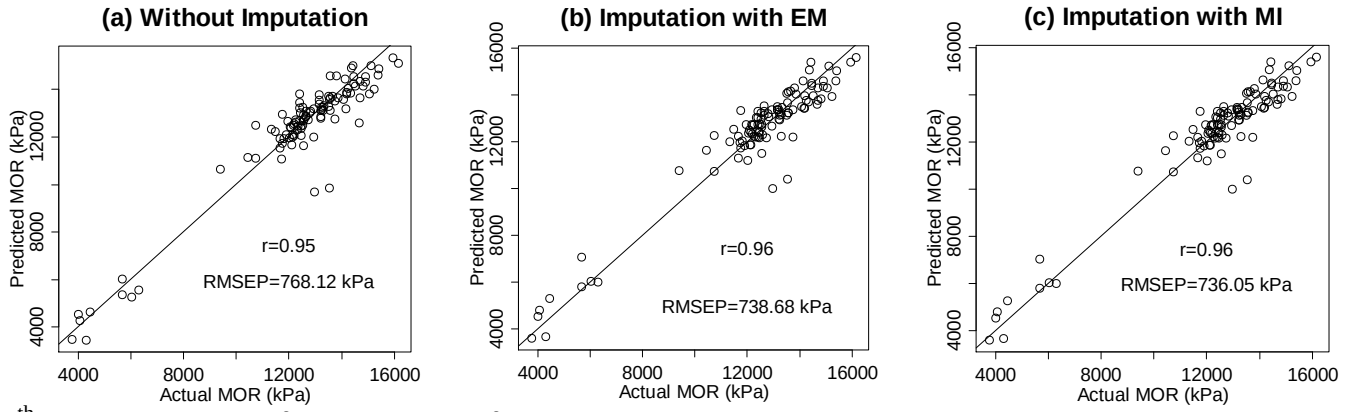
6th Cross-validation of PLSR Model of MOR Data



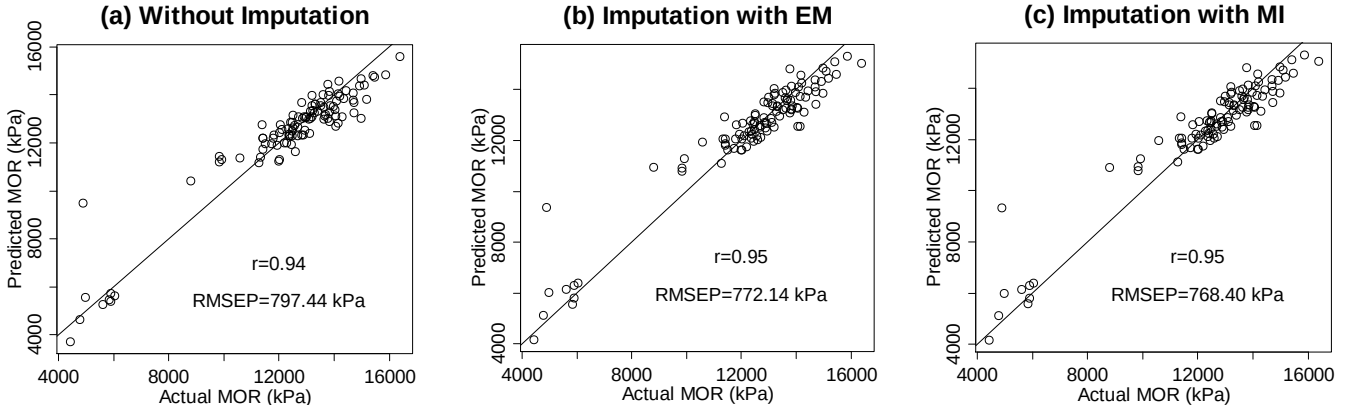
7th Cross-validation of PLSR Model of MOR Data



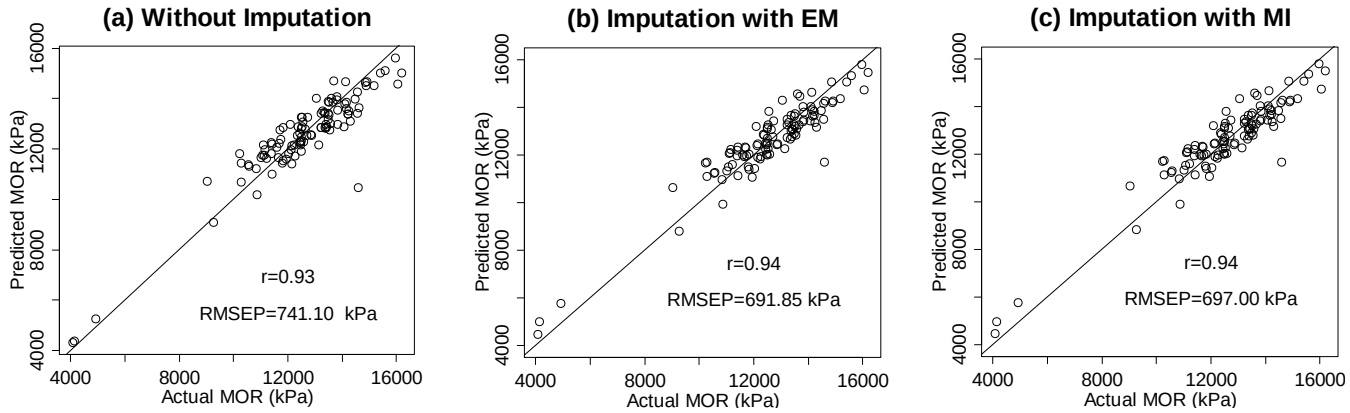
8th Cross-validation of PLSR Model of MOR Data



9th Cross-validation of PLSR Model of MOR Data



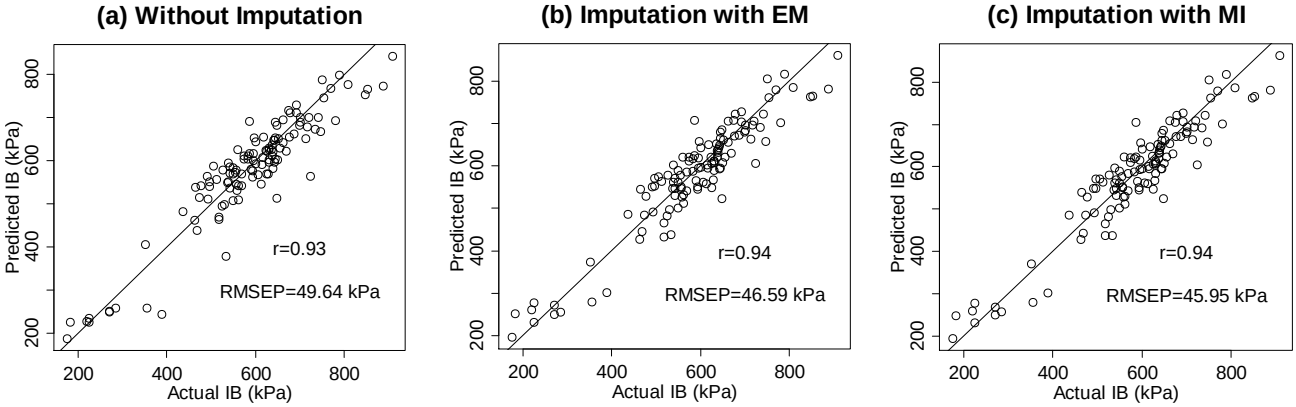
10th Cross-validation of PLSR Model of MOR Data



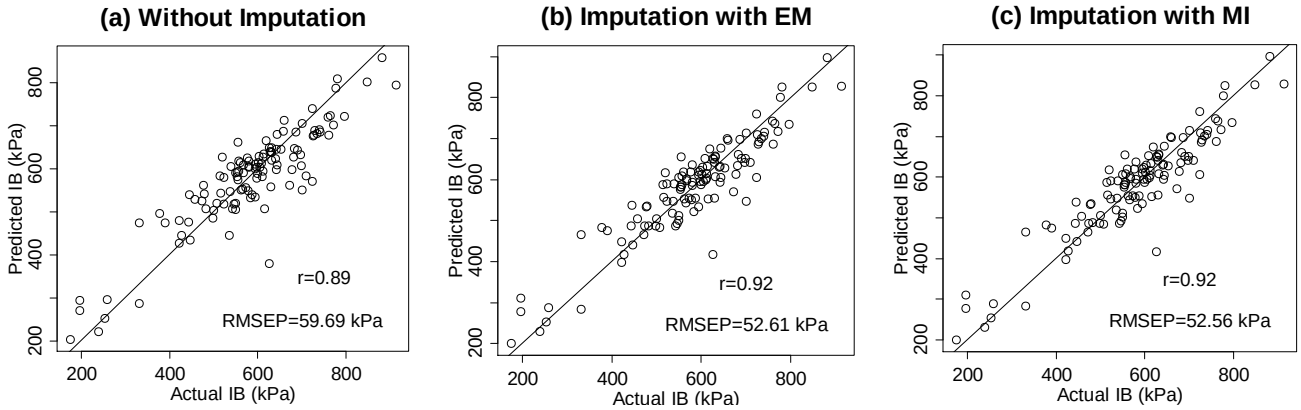
Appendix A.2.

Ten-fold Cross-validation Results of PLRS Modeling of Imputed and Non-imputed IB Data

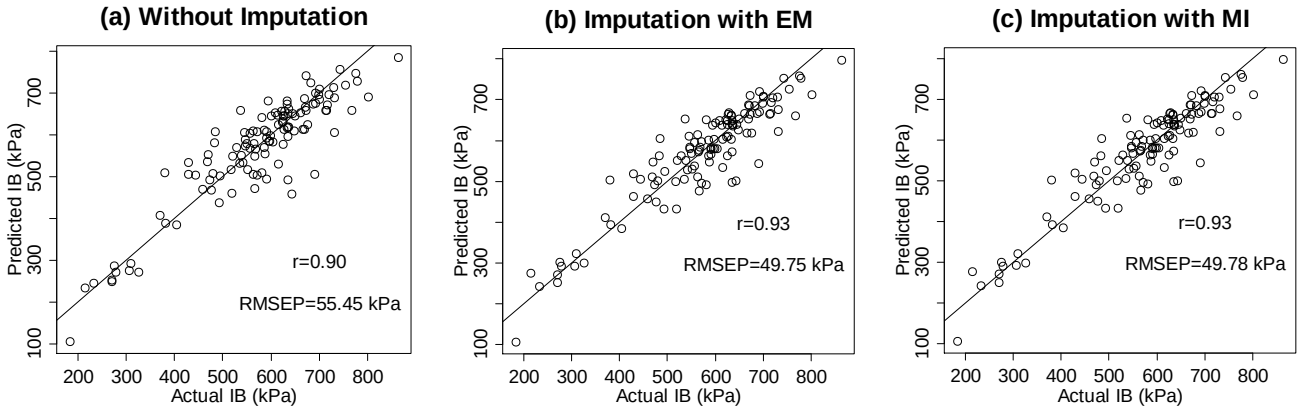
1st Cross-validation of PLSR Model of IB Data



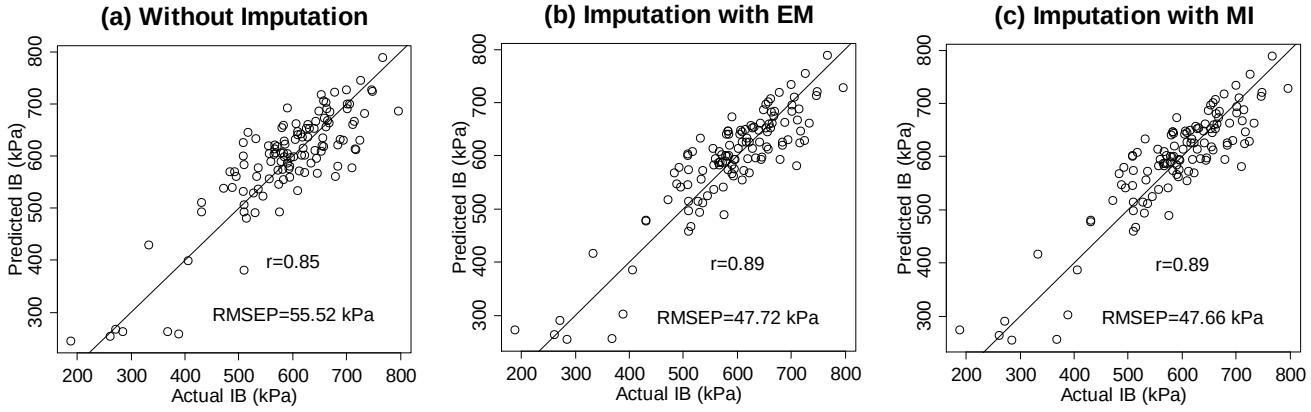
2nd Cross-validation of PLSR Model of IB Data



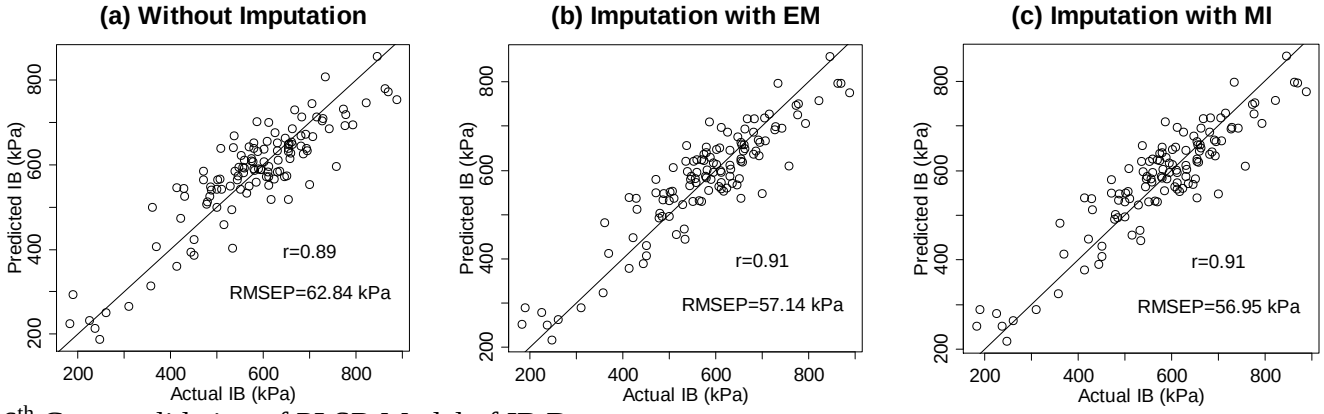
3rd Cross-validation of PLSR Model of IB Data



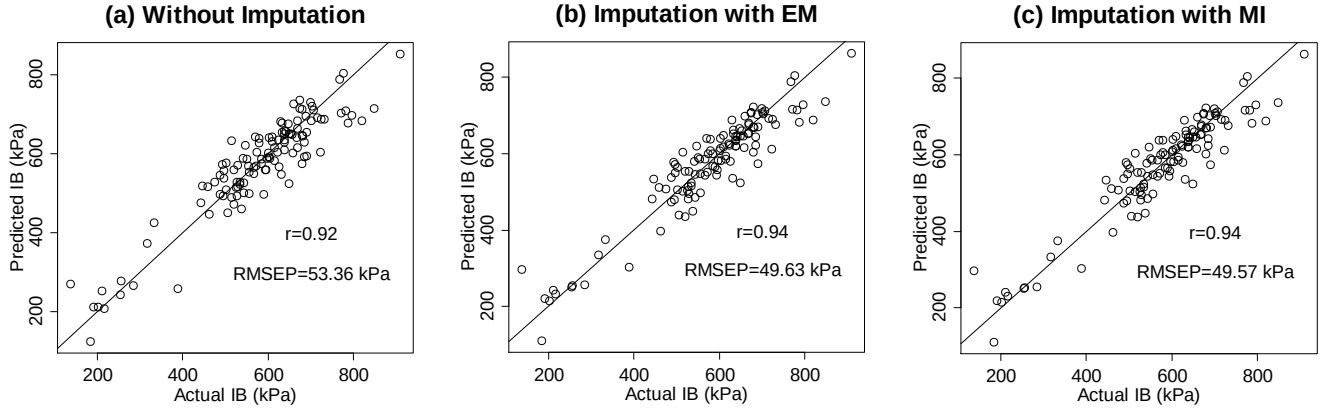
4th Cross-validation of PLSR Model of IB Data



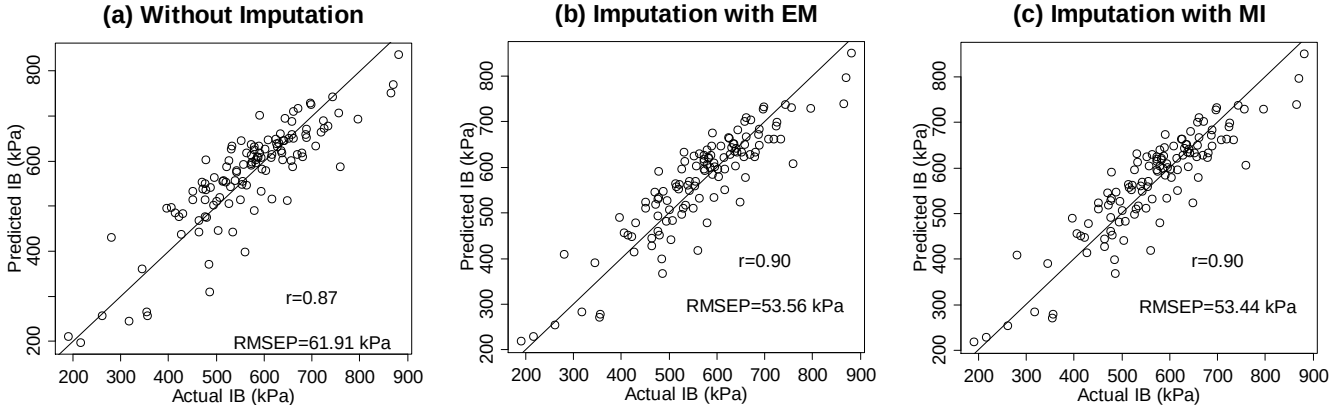
5th Cross-validation of PLSR Model of IB Data



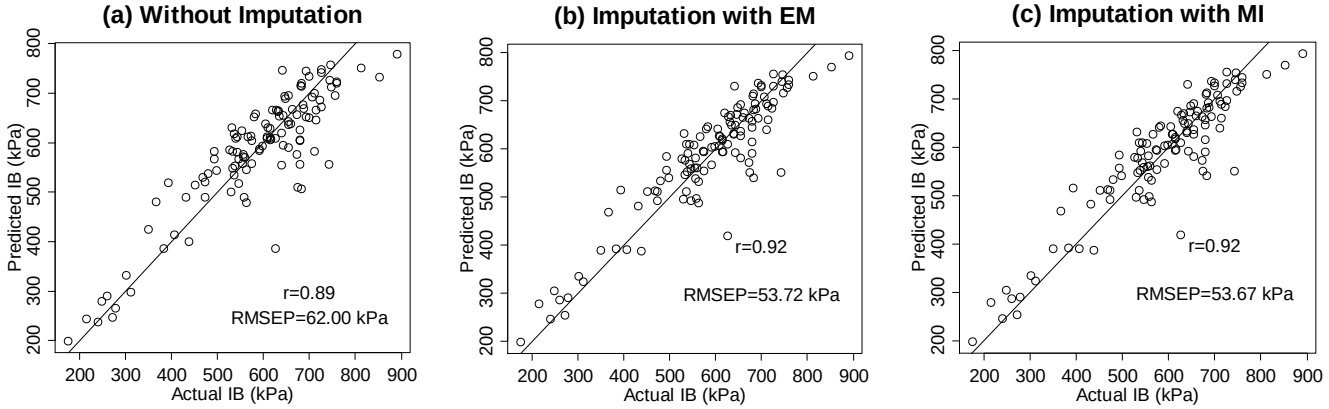
6th Cross-validation of PLSR Model of IB Data



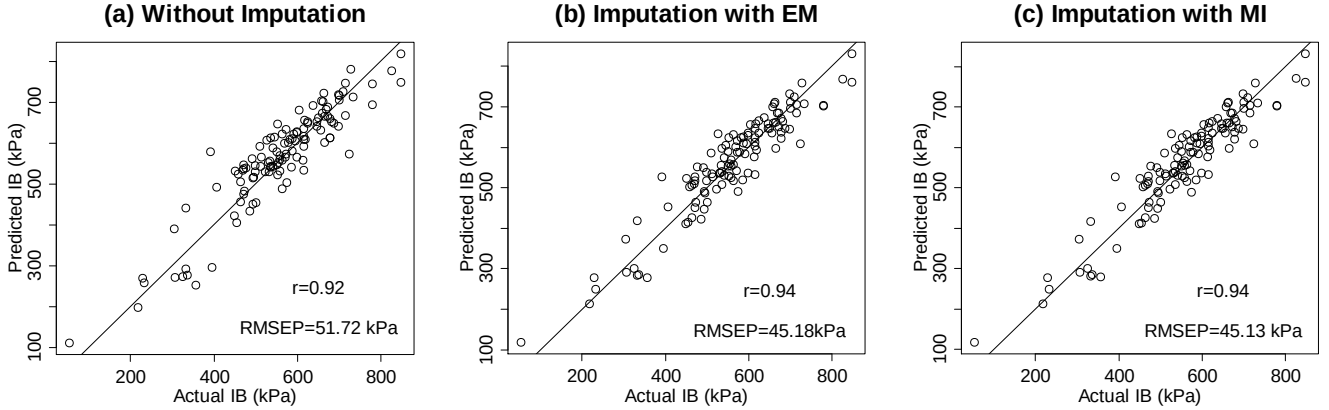
7th Cross-validation of PLSR Model of IB Data



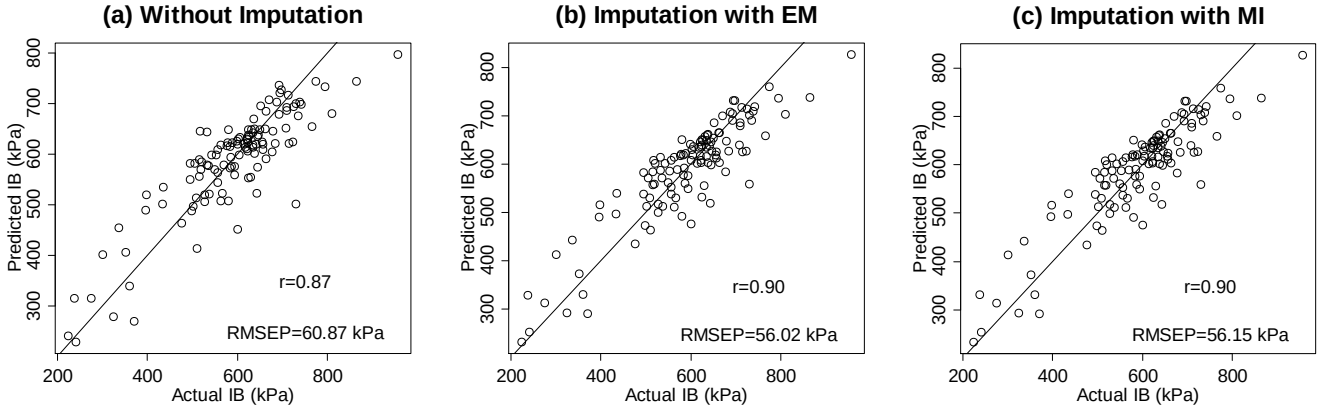
8th Cross-validation of PLSR Model of IB Data



9th Cross-validation of PLSR Model of IB Data



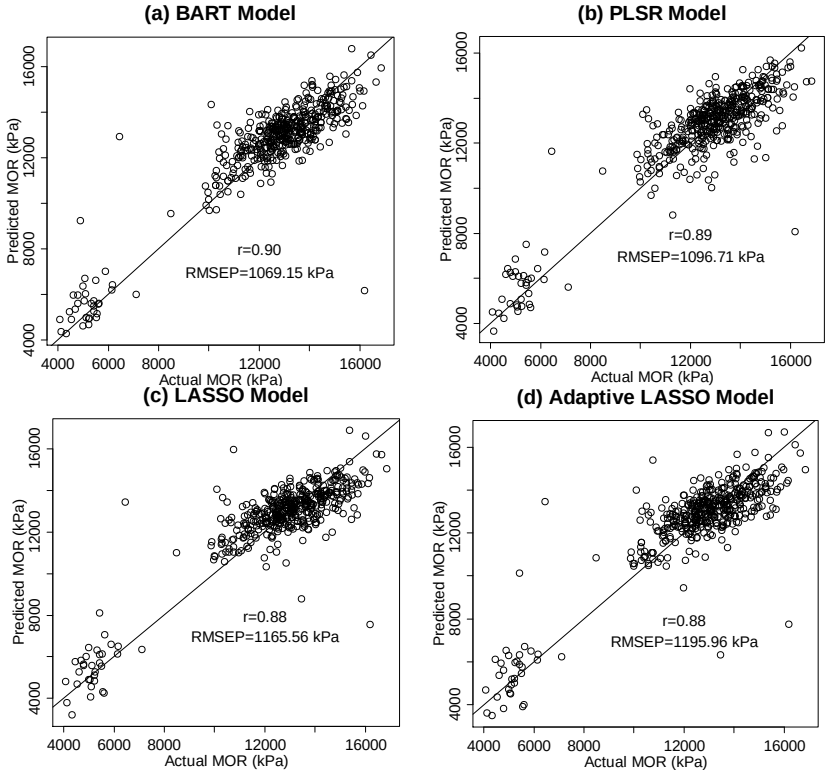
10th Cross-validation of PLSR Model of IB Data



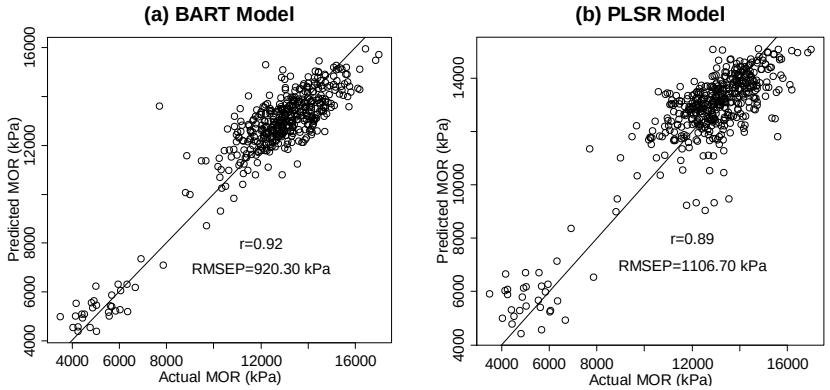
Appendix B
Appendix B.1.

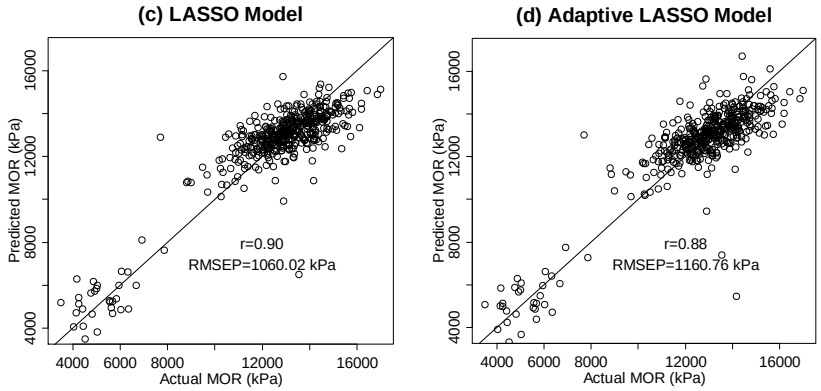
Ten-fold Cross-validation of BART, PLSR, LASSO, and Adaptive LASSO Models Using MOR Data

1st Cross-validation of Four Models of MOR data

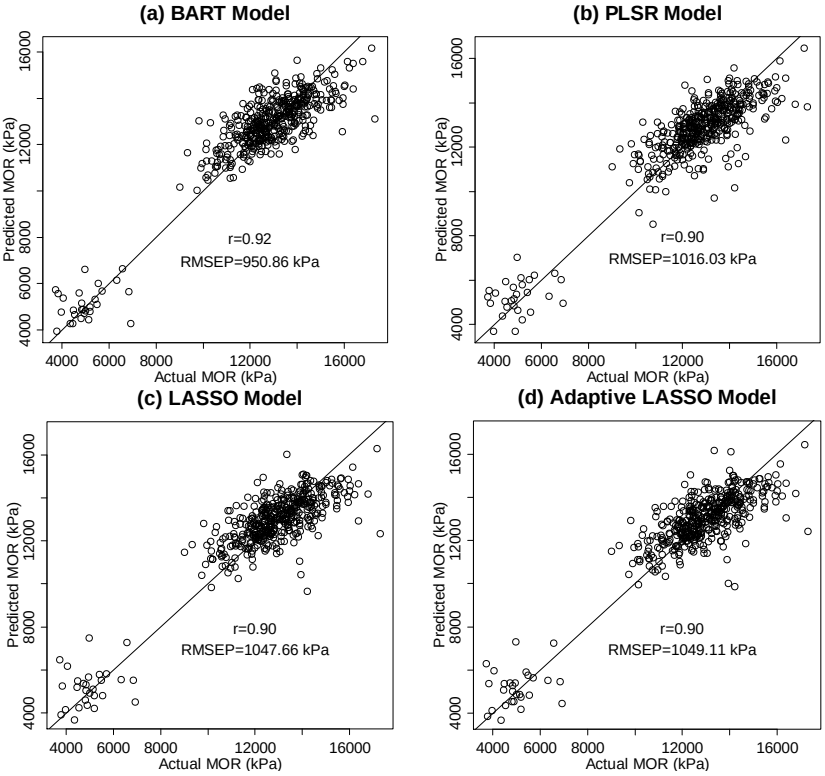


2nd Cross-validation of Four Models of MOR data

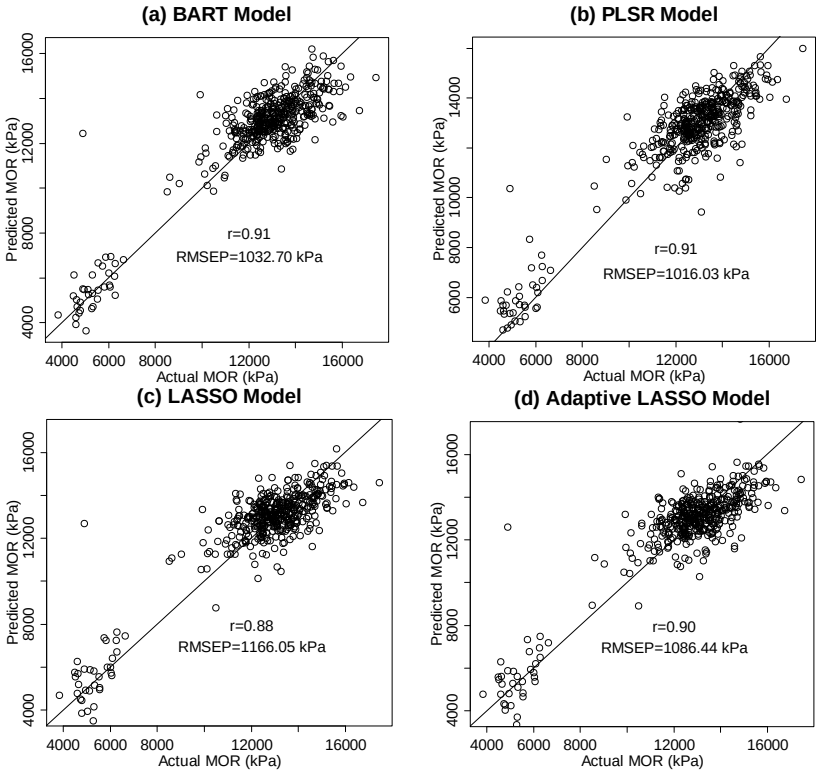




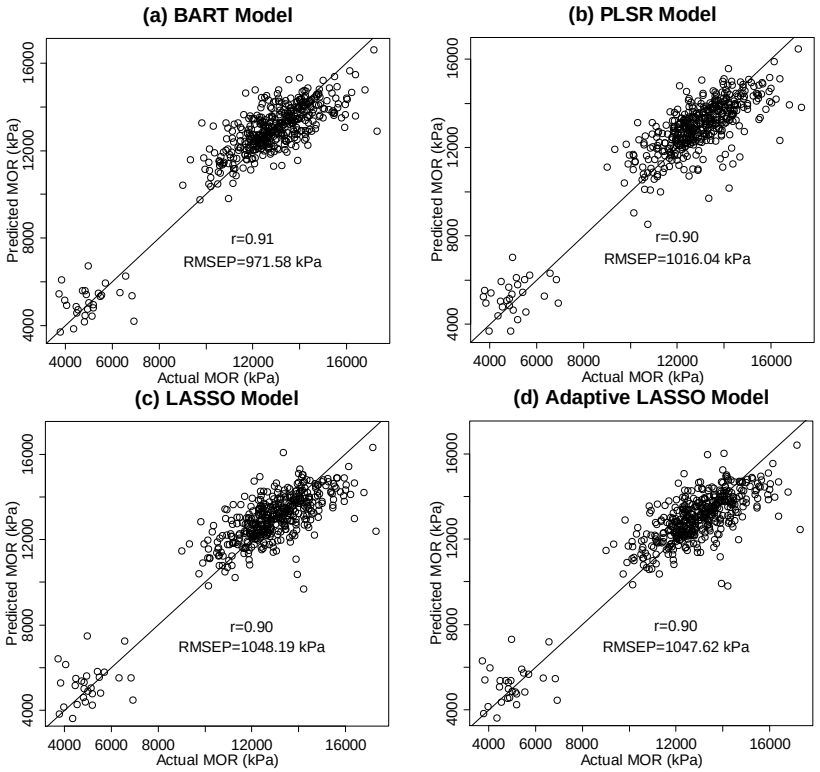
3rd Cross-validation of Four Models of MOR data



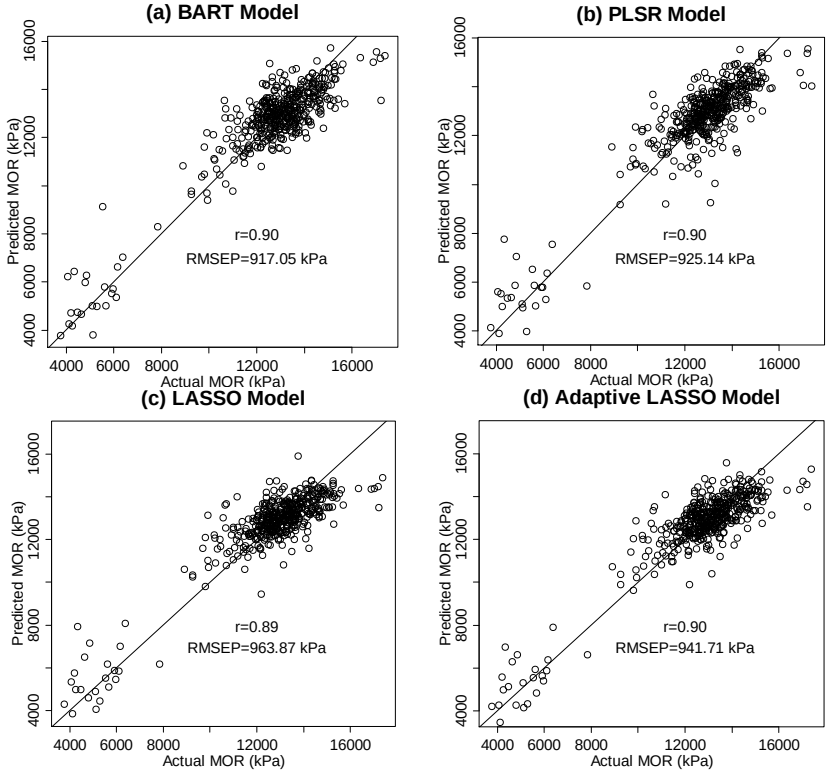
4th Cross-validation of Four Models of MOR data



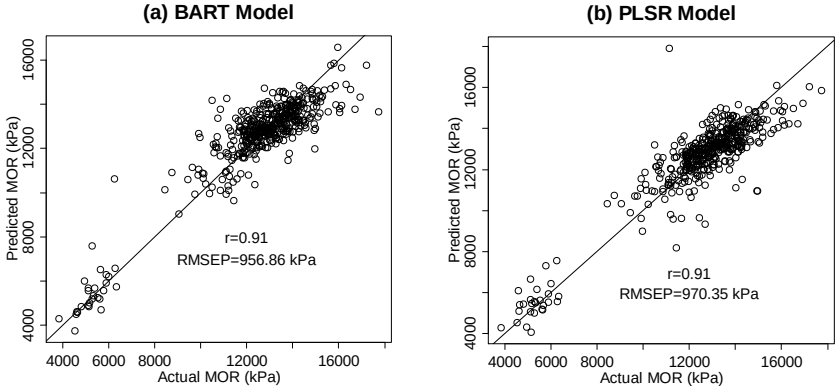
5th Cross-validation of Four Models of MOR data

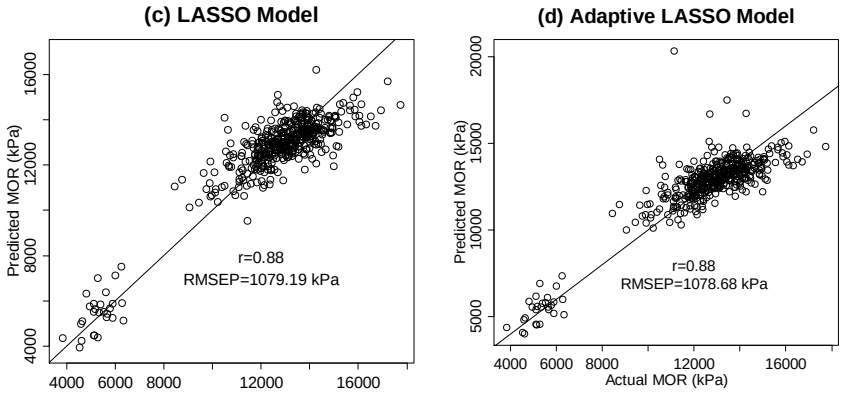


6th Cross-validation of Four Models of MOR data

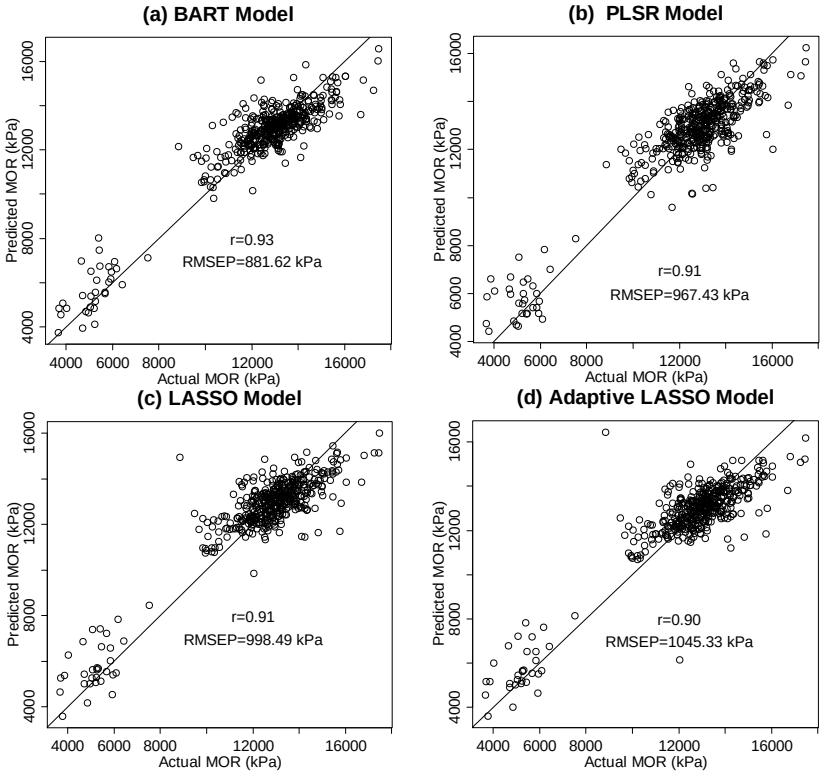


7th Cross-validation of Four Models of MOR data

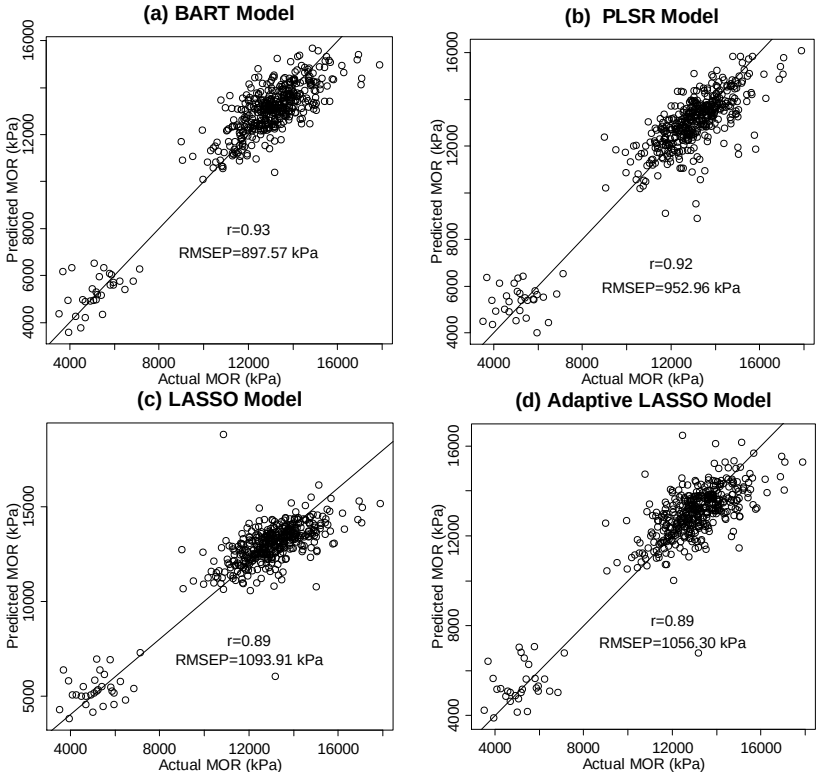




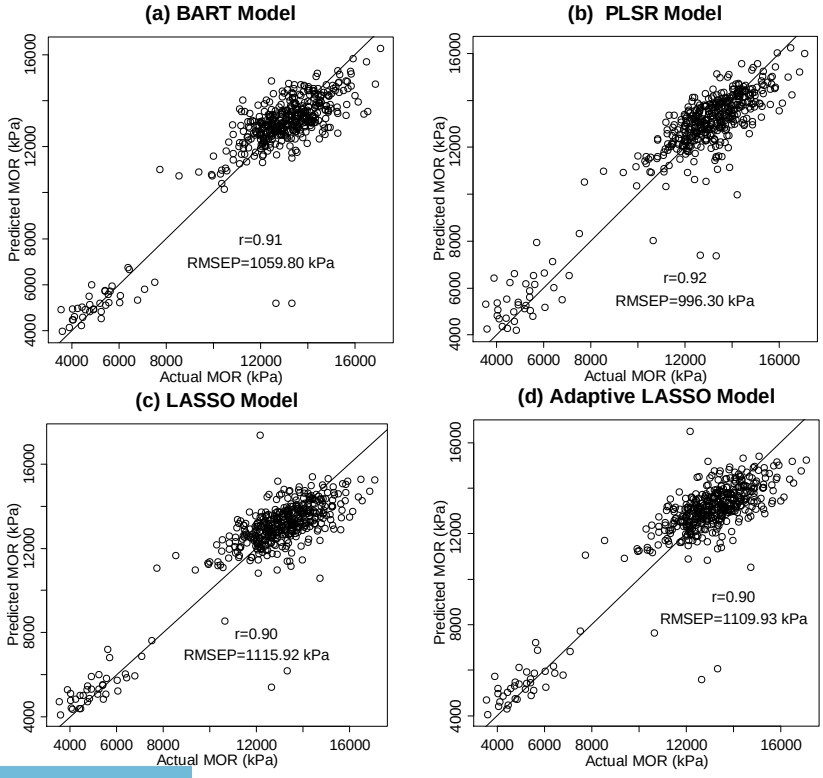
8th Cross-validation of Four Models of MOR data



9th Cross-validation of Four Models of MOR data



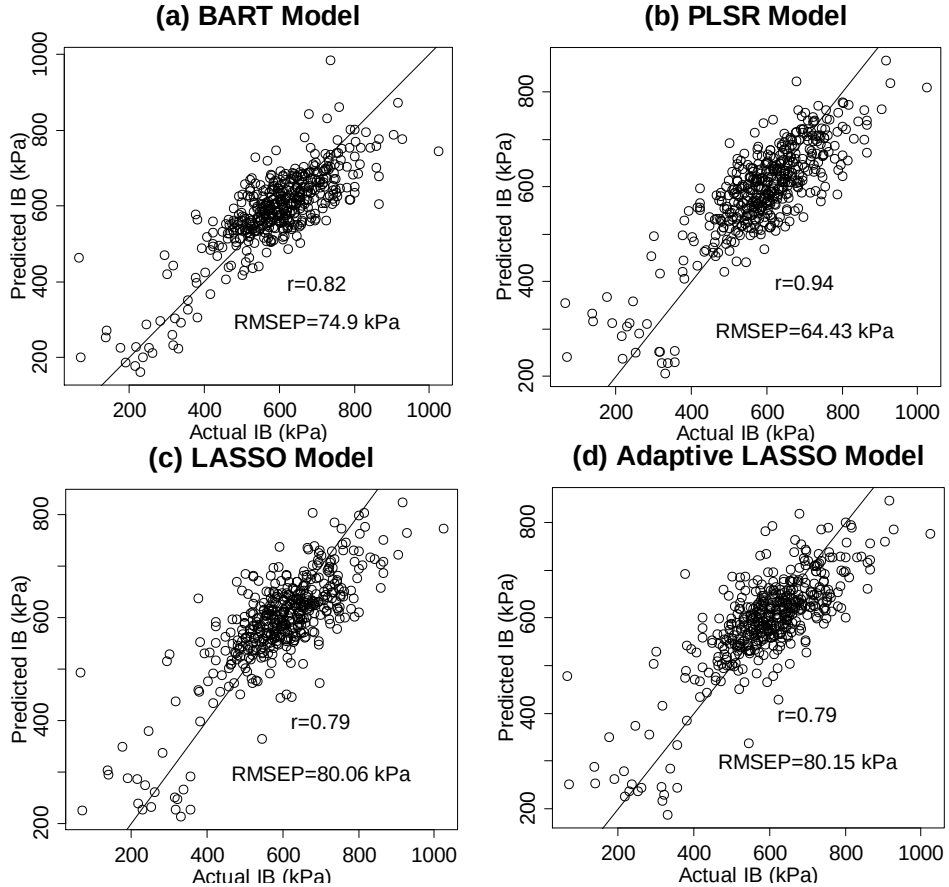
10th Cross-validation of Four Models of MOR data



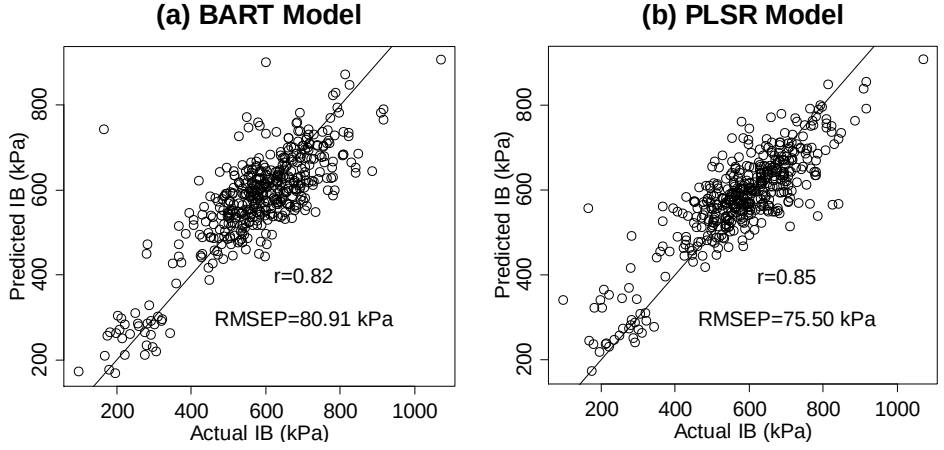
Appendix B.2.

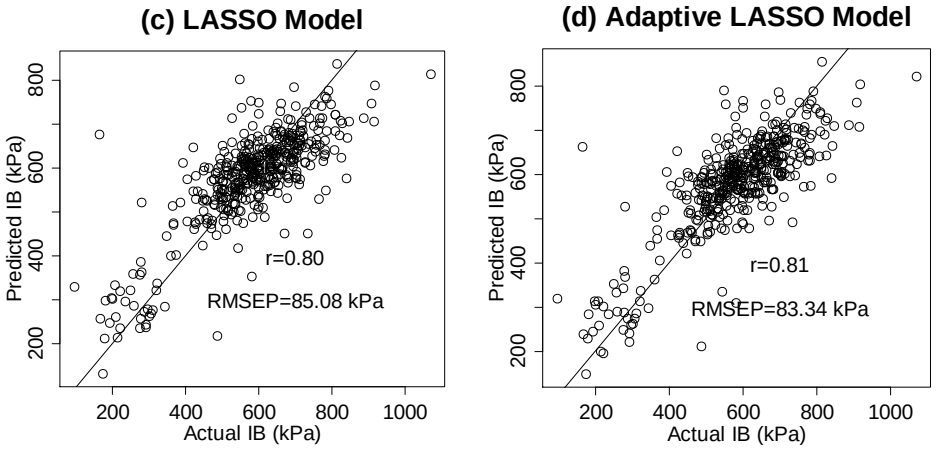
Ten-fold Cross-validation of BART, PLSR, LASSO, and Adaptive LASSO Models Using IB Data

1st Cross-validation of Four Models of IB data

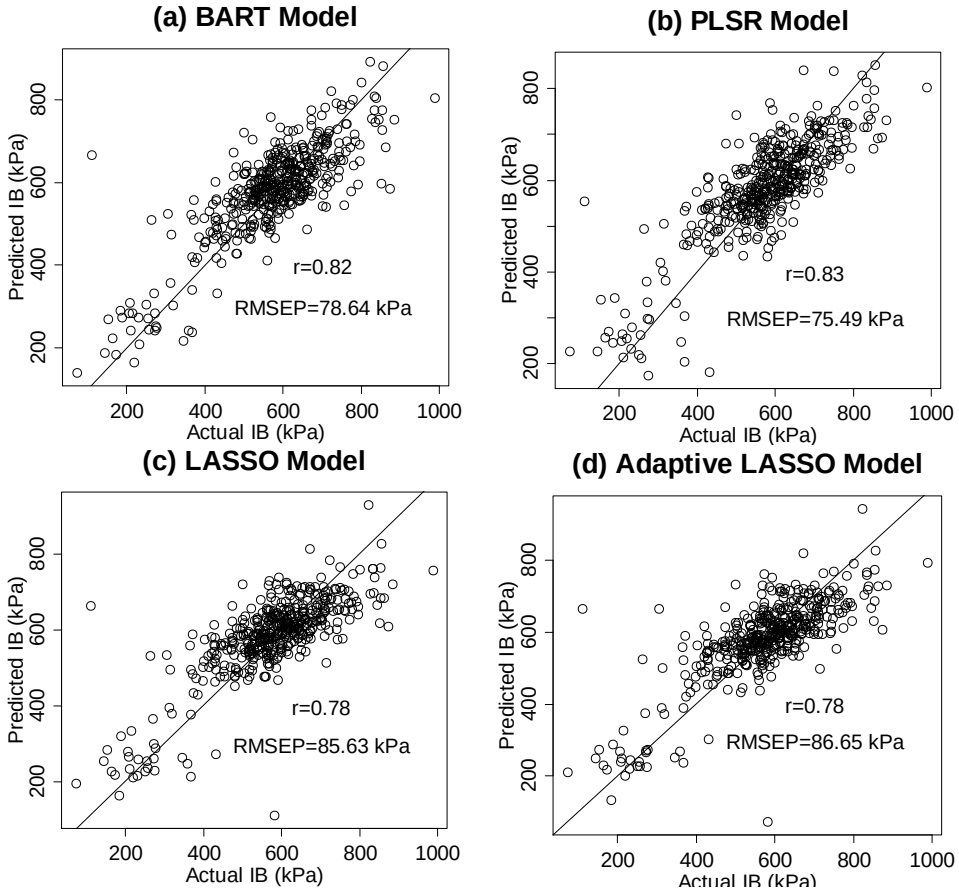


2nd Cross-validation of Four Models of IB data

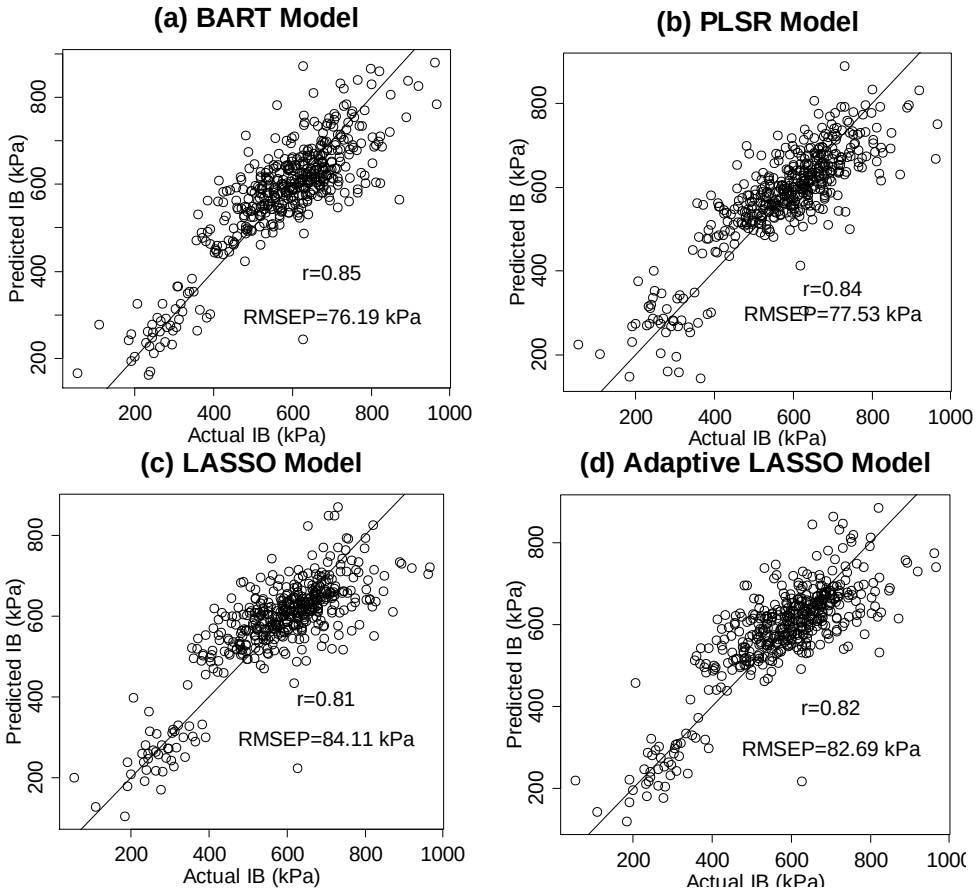




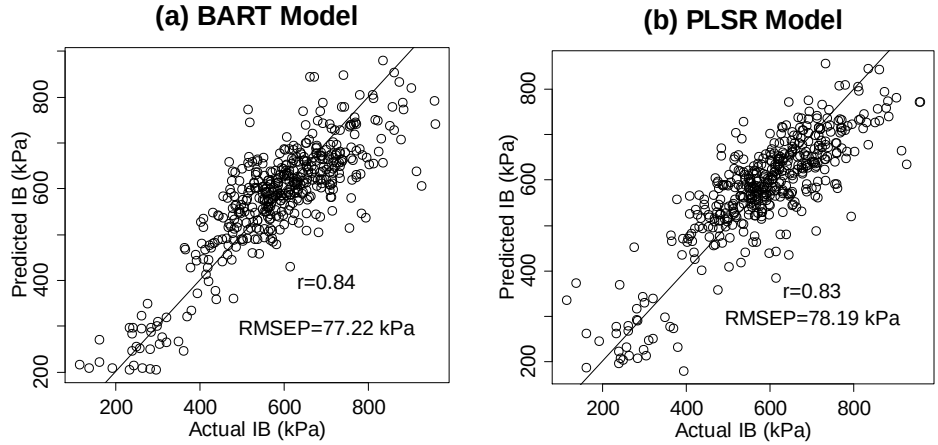
3rd Cross-validation of Four Models of IB data

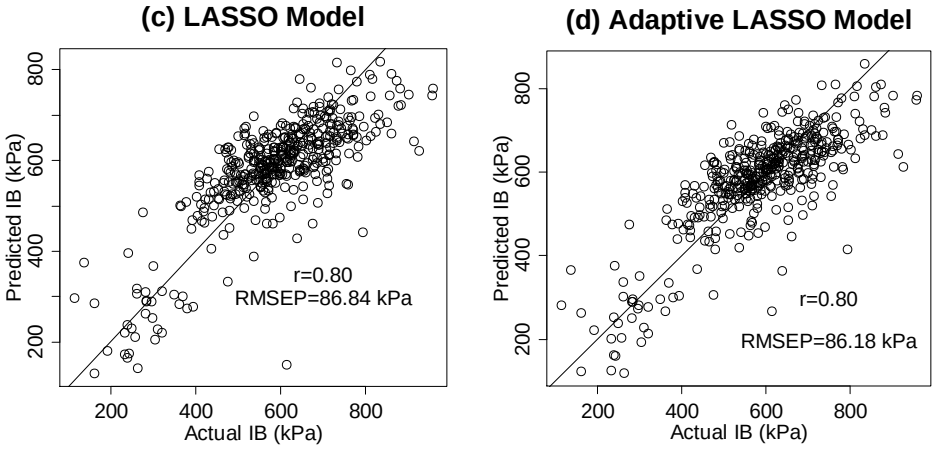


4th Cross-validation of Four Models of IB data

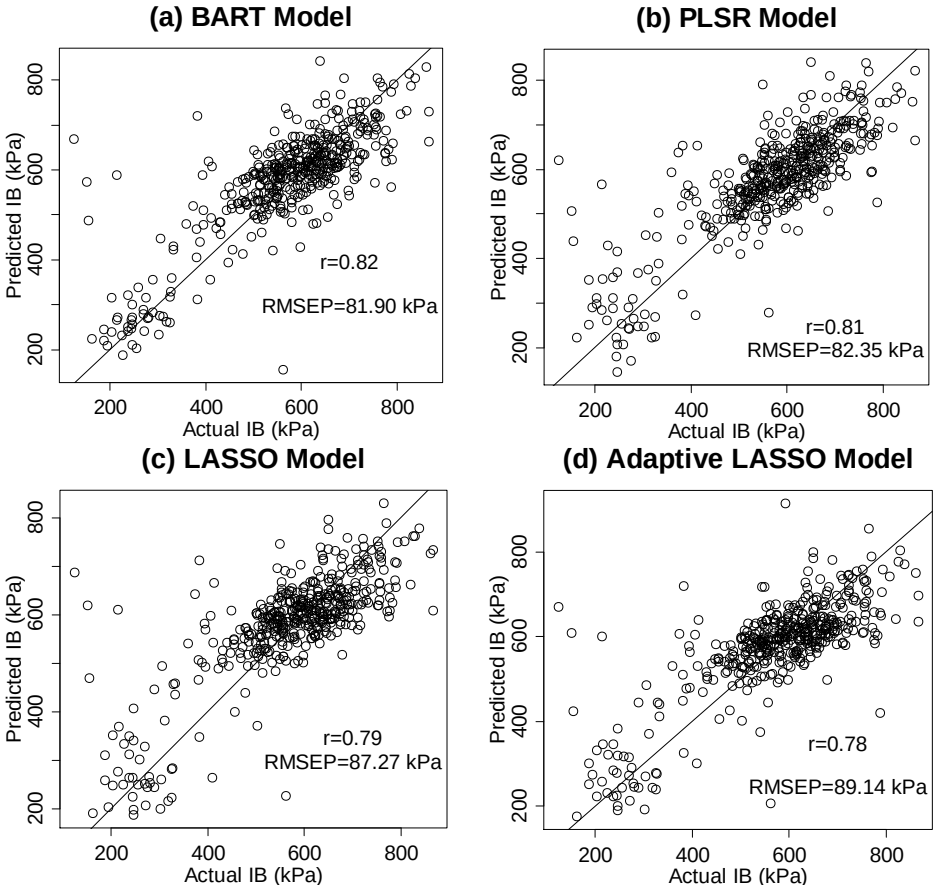


5th Cross-validation of Four Models of IB data

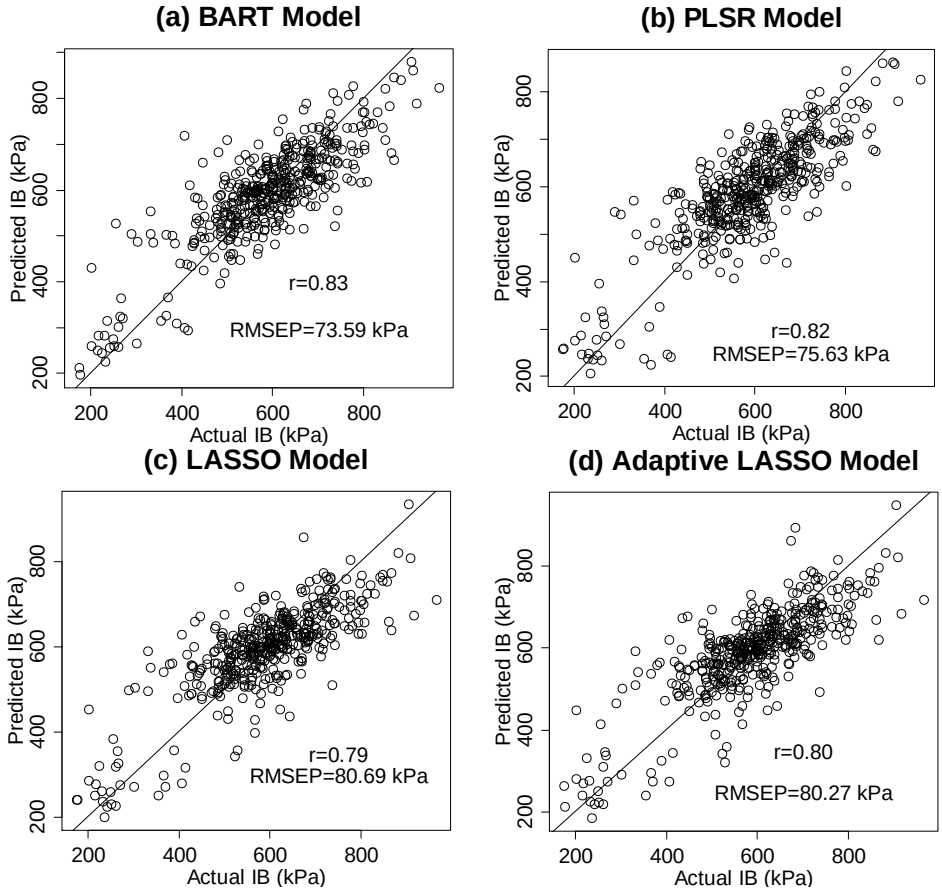




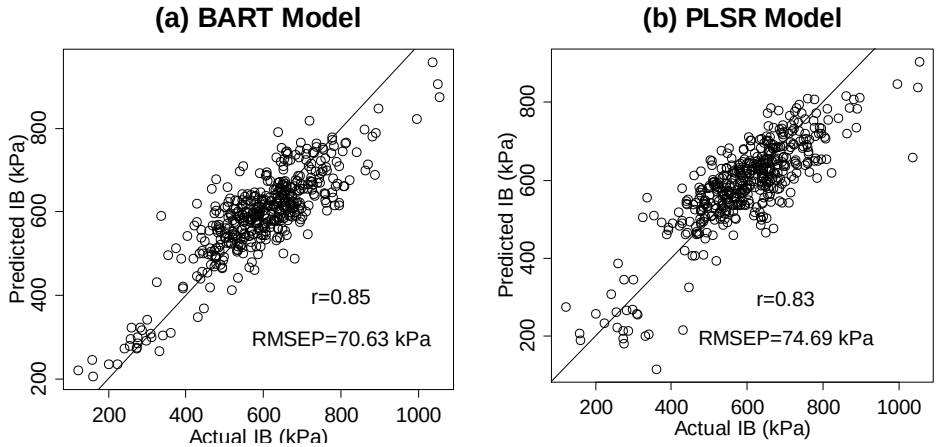
6th Cross-validation of Four Models of IB data

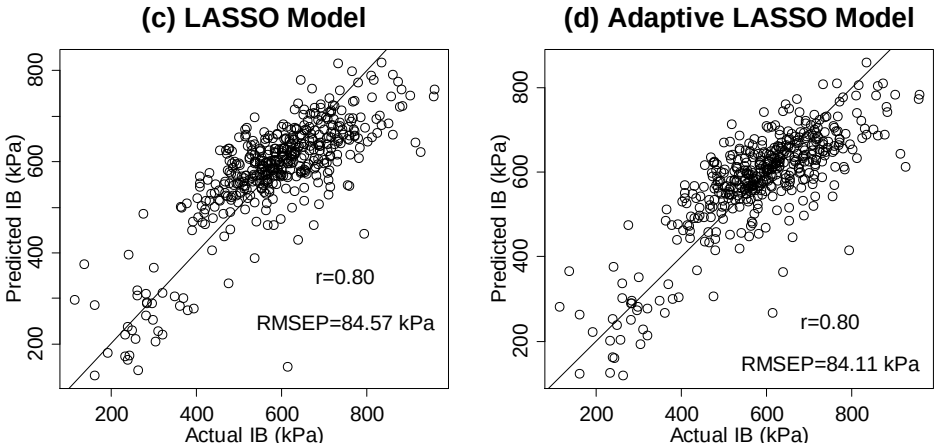


7th Cross-validation of Four Models of IB data

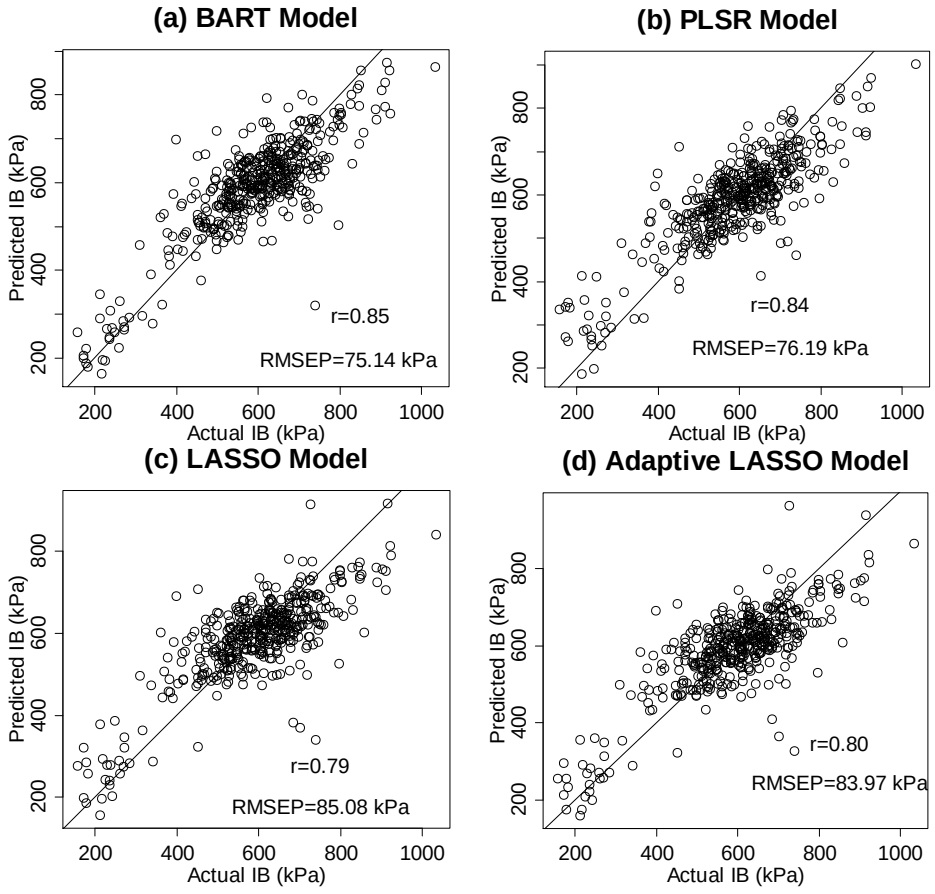


8th Cross-validation of Four Models of IB data

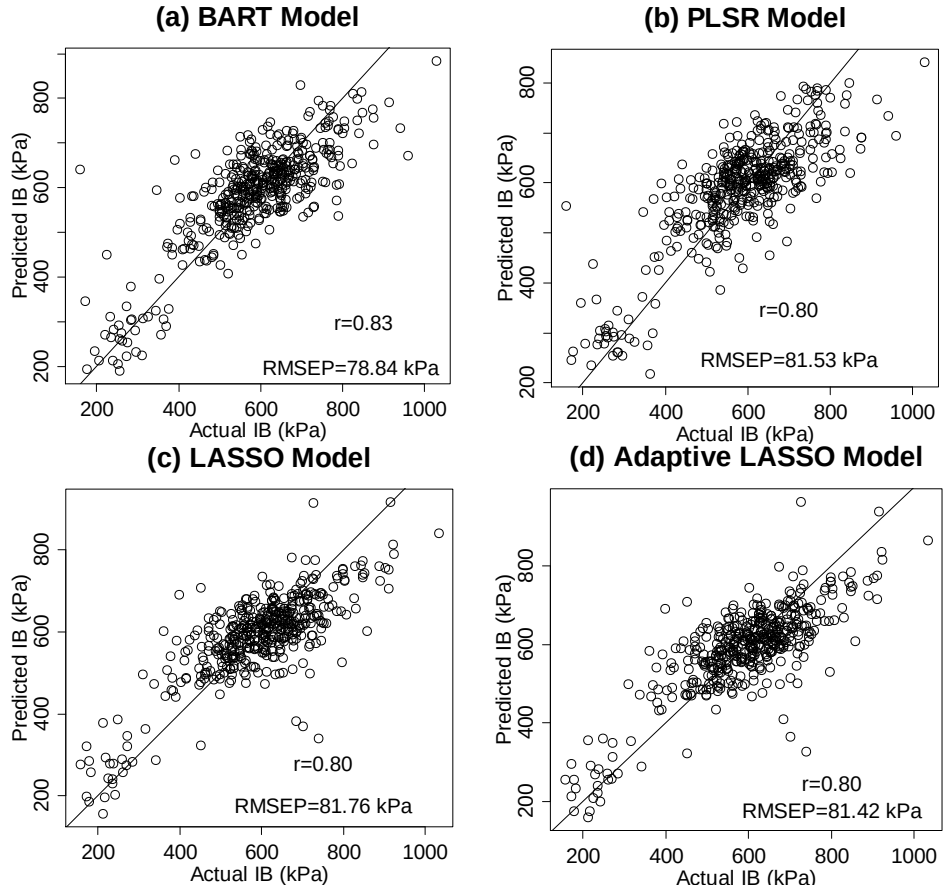




9th Cross-validation of Four Models of IB data



10th Cross-validation of Four Models of IB data



VITA

Yan Zeng is a Graduate Research Assistant under Dr. Timothy M. Young at the Center for Renewable Carbon and Graduate Teaching Assistant in the Department of Statistics, Operations, and Management Science under Dr. Frank Guess at the University of Tennessee, Knoxville. He plans to graduate from the University of Tennessee with a Master of Science degree in Statistics in August 2011. After graduation he will go on to work as a Financial Analyst at Discover Financial Service in Chicago, Illinois.

Yan received a Bachelor of Engineering degree in Mechanical Engineering with concentration in Industrial Design and Minor in Computer Engineering, from Xi'an Jiao Tong University in Xi'an, China. He was also trained as an undergraduate Research Assistant in National Tsing Hua University in Hsinchu, Taiwan.

After undergraduate studies, Yan was awarded College of Engineering Fellowship and Research Assistantship to study at the University of Nebraska-Lincoln, where he obtained his M.S in Industrial Engineering, minor in Statistics.

Currently Yan is part-time teaching and full-time pursuing his second M.S in Statistics at the University of Tennessee at Knoxville. In addition to his academic experiences, he also interned as a Healthcare Analyst at the National Committee for Quality Assurance (NCQA) in Washington, D.C.