Article

Improving Innovation from Science Using Kernel Tree Methods as a Precursor to Designed Experimentation

Timothy M. Young ^{1,*} , Robert A. Breyer ², Terry Liles ³ and Alexander Petutschnigg ⁴¹ Center for Renewable Carbon, The University of Tennessee, 2506 Jacob Drive, Knoxville, TN 37996-4570, USA² Georgia-Pacific Chemicals, 2883 Miller Rd., Decatur, GA 30035, USA; rabreyer@gapac.com³ Huber Engineered Wood, 1442 State Rd 334, Commerce, GA 30530, USA; Terry.Liles@huber.com⁴ Fachhochschule Salzburg, GmbH, Salzburg University of Applied Sciences, Holztechnologie & Holzbau, Markt 136a, 5431 Kuchl, Austria; alexander.petutschnigg@fh-salzburg.ac.at

* Correspondence: tmyoung1@utk.edu; Tel.: +1-865-946-1119

Received: 8 April 2020; Accepted: 8 May 2020; Published: 14 May 2020

Abstract: A key challenge in applied science when planning a designed experiment is to determine the aliasing structure of the interaction effects and selecting the appropriate levels for the factors. In this study, kernel tree methods are used as precursors to identify significant interactions and levels of the factors useful for developing a designed experiment. This approach is aligned with integrating data science with the applied sciences to reduce the time from innovation in research and development to the advancement of new products, a very important consideration in today's world of rapid advancements in industries such as pharmaceutical, medicine, aerospace, etc. Significant interaction effects for six common independent variables using boosted trees and random forests of $k = 1000$ and $k = 10,000$ bootstraps were identified from industrial databases. The four common variables were related to speed, pressing time, pressing temperature, and fiber refining. These common variables maximized tensile strength of medium density fiberboard (MDF) and the ultimate static load of oriented strand board (OSB), both widely-used industrial products. Given the results of the kernel tree methods, four possible designs with interaction effects were developed: full factorial, fractional factorial Resolution IV, Box–Behnken, and Central Composite Designs (CCD).

Keywords: kernel tree methods; interaction effects; aliasing structure; design experimentation; response surface methods

1. Introduction

Data science is evolving rapidly in the world, and the precipitous needs from innovation to adoption from industries such as the pharmaceutical, aerospace, food, etc., have never been greater. A key challenge is to shorten the time span from innovation to adoption while maintaining scientific inference. Many applied scientists rely on formal experimentation during innovation development. Scientists have budgetary and time constraints that limit cyclical experimentation. The study outlined in the paper presents a methodology of using data science kernel tree methods as a precursor for designed experimentation, to reduce the time from innovation in applied sciences to adoption for product production. This combination of induction and deduction is aligned with data science by combining contemporary methodologies with more classical methods to enhance scientific inference.

Many designed experiments in research and development (R&D) contain two or more factors, with two or more levels, e.g., a 2^k design (low [−] and high [+] levels) with $k = 3$ with a replication equate to $n = 16$ runs of experimentation. If three levels are desired, a 3^k design (low [−], medium [0], and high [+] levels) with $k = 3$ with a replication will have $n = 54$ experimental runs. In the applied

sciences, experimental runs are expensive and the number of total runs is typically restricted by budget constraints. Survey or screening designs with fractional factorials are helpful to alleviate the number of experimental runs to minimize costs. However, conducting multiple screening designs require time, labor, and cost, e.g., 2^3 design with replication requires at least 16 runs and one-half fraction of this design requires eight runs with replication and are not recommended, i.e., given the aliasing of main effects with two level interactions where two-level interactions are typically significant, as noted by [1].

The authors propose the use of kernel tree methods (e.g., regression trees, random forests, boosted trees, etc.) to minimize the time from innovation to product adoption while sustaining scientific inference. Kernel tree methods identify unknown interactions as a hierarchy in the data [2–5]. Models from these techniques have high explanatory value [6–11]. The interactions from such models may also identify unknown aliasing structures and avoid aliasing highly significant factors during experimentation. Combining bootstrapping [12] with kernel tree methods identifies the most common set of reoccurring variables that influence a response variable. Another challenge in designed experimentation is the selection of the ‘levels’ of the factors. If ‘levels’ are too narrow, knowledge gained from the experiment may be restricted. If ‘levels’ are too wide, the regions of the data space may not be feasible, which may negate experimental runs and create unbalanced design results. The ‘split-points’ of the factors in the tree-based models may guide researchers to feasible, but yet unexplored ‘levels’ within the data space. The objective of this study was to use kernel tree methods to quantify significant interactions as a hierarchy before conducting a formal experiment.

2. Material and Methods

2.1. Dataset Descriptions

Datasets for the study were obtained from medium density fiberboard (MDF) and oriented strand board (OSB) manufacturers as part of a research confidentiality agreement. Variable names in the datasets were altered to respect the terms of the confidentiality agreement. MDF is a nonstructural biomaterial that is used as a substrate for furniture, kitchen cabinets, desks, tabletops, etc. OSB is a structural biomaterial that is used for residential and non-residential construction of dwellings. Tensile strength is the primary strength metric (dependent variable) for MDF. Ultimate static load is an important strength metric for OSB and is the certification metric for use in the marketplace. Data from the destructive testing labs were obtained from two different manufacturers in the United States of America (USA). The MDF dataset had 408 records and 184 regressors which represented destructive tests over a three-month time period for a nominal product of 15.88 mm in thickness. MDF destructive tests are typically taken from the production line at one-hour periodic intervals or when the product-type is changed. The tensile strength dependent variable for MDF had a normal pdf based on the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) (Table 1) [13,14].

The OSB dataset had 150 records and 98 regressors which represented destructive tests over a six-month time period for the nominal product of 11.11 mm in thickness. OSB destructive tests are taken from the production line at four-hour periodic intervals or at product change. The ultimate static load dependent variable for OSB had a normal pdf based on the AIC and BIC criteria (Table 1). The regressors for both MDF and OSB were taken from the process data warehouses and were fused with the dependent variables from the destructive testing labs. Eighty percent of the data for both OSB and MDF datasets were used for training and 20% of the data were used for validation. For MDF 184 regressors, and in OSB, 98 regressors were in the training datasets. Ten-fold cross validation was used for both MDF and OSB models.

Table 1. Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) statistics for tensile strength pdfs of medium density fiberboard (MDF) and ultimate static load of oriented strand board (OSB).

Tensile Strength-Model Comparisons		
Distribution	AIC	BIC
Normal	4966.5326	4975.3464
Generalized Gamma	4967.6229	4980.8336
Log Generalized Gamma	4967.6304	4980.8411
Lognormal	4970.1264	4978.9402
Logistic	4978.8342	4987.6480
Loglogistic	4981.3085	4990.1222
Weibull	5028.7795	5037.5932
LEV	5044.7598	5053.5736
SEV	5087.5301	5096.3439
Frechet	5110.6015	5119.4153
Exponential	7255.9066	7260.3168
Ultimate Static Load-Model Comparisons		
Distribution	AIC	BIC
Normal	−524.1410	−518.2014
Generalized Gamma	−522.2338	−513.3663
Log Generalized Gamma	−522.0893	−513.2218
Lognormal	−521.9522	−516.0125
Logistic	−518.1157	−512.1760
Loglogistic	−516.7510	−510.8113
Weibull	−516.6517	−510.7120
SEV	−507.4300	−501.4904
LEV	−504.8098	−498.8702
Frechet	−491.4876	−485.5480
Exponential	30.1596	33.1432

2.2. Description of Predictor Variables

Process data are from sensors on the production line and are related to line speed, pressing speed, pressing pressure, pressing temperature, fiber moisture, etc. Both the MDF and OSB processes have a variation in all of the variables during the normal manufacturing process. Line speed typically is product specific, i.e., faster speeds for thinner and lower density products and slower speeds for thicker, higher density products. For example the predictor variable for OSB called ‘wet bin speed’ is directly related to line speed given that the bin that holds the flakes are emptied is a function of line speed. Line speed also changes because of moisture content changes in the fibers during the pressing stage. Pressing occurs under pressure and high temperature for the curing of the bonding of fibers and adhesives during the pressing stage. Fiber moisture may change during the manufacturing process due to natural variation in the feedstocks and variations in the temperatures during the drying process. The predictor variables reported in this manuscript are given descriptive names as related to these processes, e.g., ‘flake moisture content’ is the moisture content of the OSB flakes, ‘total press time’ is the time the material resides in the pressing stage, ‘mat weight’ is the actual weight of the formed fiber mats before the pressing stage, etc. The MDF process is different from the OSB process primarily in the early stages of the process, where wood is refined to small fibers with lengths of 1.34–1.84 mm. The predictor variables named ‘face plate position’ refer to the gap in grinding plates during the wood to fiber refining stage and ‘face steam flow’ refers to how much steam is injected into the refiners that create the fibers. Descriptions of the MDF and OSB processes are documented in [15,16].

2.3. Kernel Tree Methods

Decision trees as applied to continuous data are known as regression trees (RT) [17–19]. Given that documentation on the methodologies of kernel tree methods is extensive, only a summary is presented. As Hand et al. [20] noted, “linear regression is a global model, where there is a single predictive formula holding over the entire data-space. An alternative approach is to sub-divide, or partition, the space into smaller regions, where the interactions are more manageable.” The ‘regression tree’ approach to modeling identifies a hierarchy of interactions that were previously unknown. The regression tree approach creates recursive partitions or cells within the entire data space (i.e., terminal nodes or leaves) and the cells are modeled separately. The cells of regression trees are typically ‘pruned’ during the model validation phase in identifying the best predictive model. One strength of this method is that data do not need to be imputed, given that partitions will be made in the presence of missing data (Georges, 2009). RTs are represented as two-dimensional graphics, which makes it easily understood and interpreted [6,7].

RTs are quite popular as an exploratory modeling technique, and are commonly associated with data mining techniques [21–28]. RTs are very resistant to irrelevant regressors given that the recursive tree-building algorithm estimates the optimal variable on which to split at each step, regressors unrelated to the response are not chosen for splitting [29] (pp. 199–215). In theory, a regression tree partitions the data space of all joint regressor values $\mathbf{X}$ into J -disjoint regions $\{R_j\}_1^J$ [6]. For a given set of joint regressor values $\mathbf{X}$, the tree prediction $\hat{Y} = T_j(\mathbf{X})$ assigns as the response estimate, the value assigned to each region containing $\mathbf{X}$:

$$\mathbf{X} \in R_j \Rightarrow T_j(\mathbf{X}) = \hat{y}_j \quad (1)$$

Given a set of regions, the optimal response values associated with each region minimize the prediction error in that region:

$$\hat{Y}_j = \operatorname{argmin}_{y'} E_y[L(y, y') | \mathbf{X} \in R_j] \quad (2)$$

Unlike some kernel methods, RTs use the data to estimate a good partition instead of relying on a predefined model by the analyst.

‘Boosted trees’ (BT) relies on the philosophy that a small number of simple trees of weak learners that are combined as one model outperform the predictions of one large RT [6,30,31]. ‘Boosting’ builds trees sequentially such that each new tree improves the predictive power of the ensemble [32–34]. The result grows new trees specifically aimed at accommodating observations that an existing ensemble predicts poorly, i.e., overall improved predictive performance of the final BRT model. BRT approximates a solution to the problem of fitting a sum-of-trees by adding new trees one at a time, while keeping all existing trees unchanged [35]. As stated by Schapire [31] and Elith et al. [35], BT is a model to enhance the model accuracy and the key step in boosting is to consecutively apply the algorithm to constantly modified data, i.e., it minimizes the loss function through adding a regression tree in each iteration step [35].

‘Random forests’ (RF) developed by Breiman [36] and as summarized by Fawagreh et al. [37] combines Breiman’s bagging sampling approach, and the random selection of features, introduced independently [38,39] in order to construct a collection of decision trees with controlled variation. Each tree in the ensemble acts as a base classifier to determine the class label of an unlabeled instance.” Key advantages of RF over BT are robustness to noise and less overfitting [35,38–43].

There were five key parameters used for model calibration in this study. A minimum size split of 10 was used for both the boosted tree and random forests to prevent the program from splitting any node with below a specified number of cases [44]. ‘Early stopping’ stopped the additive boosting process if further boosting failed to improve the fit in the validation dataset. A minimum split size of five was used with a learning rate (r) of 0.1 where ($0 < r \leq 1$) which cued the program to build

separate boosted trees for every combination of splits. This permitted the boosted tree to try various combinations of parameters in order to find the one that maximizes the fit. The tree was grown with 50 layers and splits per tree ranged from five to 10. An overfit penalty ensured against having any cases with predicted probabilities equal to zero, where higher values will result in less overfitting. The probability is:

$$Prob_i = \frac{n_i + prior_i}{\sum (n_i + prior_i)} \quad (3)$$

where the summation is across all response levels and n_i is the number of observations at the node for the i th response level. $Prior_i$ is the prior probability for the i th response level and is calculated as follows:

$$prior_i = \lambda p_i + (1 - \lambda) P_i \quad (4)$$

where p_i is the $prior_i$ from the parent node, P_i is the $prob_i$ from the parent node, and λ is a weighting factor which was set at 0.9.

3. Fractional Factorials and Aliasing

Statistically designed experimentation (i.e., design of experiments or DOE) is a formal ‘deductive’ methodology, where independent variables (factors) are manipulated at different settings (levels) in a controlled fashion to explore optimization problems for key response variables. R.A. Fisher [45] is the father of DOE and expanded on the classic analysis of variance (ANOVA) methodology. A ‘full factorial’ DOE has experimental runs with replicates at all possible levels for all of the factors. Even though full factorial designs are the most informative designs, such designs are expensive.

George E.P. Box was influenced immensely by R.A. Fisher’s work and studied under Egon Pearson at the University of London [46]. He developed a series of designs known as ‘response surface methods’ (RSM) which allows researchers to use a fraction of the total experimental runs (fractional factorials). Box’s popular ‘central composite’ and ‘Box–Behnken’ RSM designs [1] minimize the number of experimental runs, while sustaining inference. A key consideration for RSM and any type of fractional factorial design is the aliasing structure of the design. Resolution III designs are designs where two-level interactions are aliased with main effects and other two-level interactions, and are typically avoided given that two-level interactions are typically significant. For example, a $\frac{1}{2}$ fraction 2^{k-1} , $k = 4$ with a replicate has $n = 16$ experimental runs is a Resolution III design with aliases: $A = BC$, $B = AC$, $C = AB$, $AB = CD$, $BC = AD$, and $AC = BD$. Generally, analysts select the highest possible design resolution (e.g., resolutions $\geq IV$) when choosing a design to avoid confounding main effects. If the analyst conducts a fractional factorial design and does not have a detailed knowledge of the phenomenon under investigation, an aliasing structure is assumed. By developing RT models as a possible first step for the factors under investigation, unknown interactions and split-points may be discovered. This may accelerate innovation by choosing statistically significant interactions for a DOE that lead to lower costs of experimentation.

4. Results

4.1. Boosted Tree Models and Bootstrap Forest

The boosted tree (BT) and bootstrap forest (BF), with $k = 1000$ and $k = 10,000$ bootstraps [47] for each method, indicated a set of significant variables ($\alpha = 0.05$). The recursive partitions using smaller additive trees, which are unique to BT and BF methods identified a set of predictor variables that were common to both modeling types. Four common predictor variables in these bootstrapped additive trees for MDF were: ‘face plate position’, ‘press position time’, ‘face steam flow’, and ‘adhesive percent’ (Table 2). The R^2 in training and validation for the BT was 0.758 and 0.552, respectively. The R^2 in training and R^2 in validation for the BF was 0.628 and 0.484, respectively. Even though the R^2 in validation is not high, these common variables explain a reasonable proportion of the variation influencing tensile

strength. Screening designs may not have led to a similar result in a short amount of time. The means and variances of the dependent variable tensile strength (kPa) in the training and validation datasets were similar for the 10-fold cross validation with an $\bar{x} = 946.7$, $\bar{s} = 95.6$, $\bar{M} = 944.6$, $\bar{CV} = 10.1\%$ in the training dataset; and $\bar{x} = 946.4$, $\bar{s} = 91.2$, $\bar{M} = 945.9$, $\bar{CV} = 9.6\%$ in the validation dataset.

Table 2. Common predictors variables for 10-fold cross validation for both bootstrap forest and boosted tree models for MDF after $k = 1000$ bootstraps and $k = 10,000$, ranked by highest sums of squares.

Bootstrap Forest ($k = 1000$ Bootstraps)			Boosted Tree ($k = 1000$)		
Variable	Number of Splits	Sums of Squares	Variable	Number of Splits	Sums of Squares
Face Plate Position	25	2771.68	Face Plate Position	9	26,525.05
Core Dust Speed	13	1052.24	Press Position Time	6	17,736.67
Press Position Time	12	796.26	Core Dust Speed	7	14,952.18
Total Press Time	11	594.18	Face Steam Flow	4	9565.28
Face Steam Flow	15	592.02	Total Press Time	2	8360.52
Adhesive Percent	14	469.34	Adhesive Percent	4	7688.29
Bootstrap Forest ($k = 10,000$ Bootstraps)			Boosted Tree ($k = 10,000$)		
Variable	Number of Splits	Sums of Squares	Variable	Number of Splits	Sums of Squares
Face Steam Flow	20	2379.71	Face Plate Position	7	28,640.33
Adhesive Percent	29	2142.70	Swing Plate Position	7	16,677.42
Face Plate Position	87	1571.86	Face Steam Flow	8	15,481.28
Press Position Time	15	1470.78	Adhesive Percent	4	9565.11
Swing Plate Position	45	1319.93	Press Position Time	6	9553.90
Resin Temperature	20	1211.34	Face Steam Flow	6	8170.58

The boosted tree (BT) and bootstrap forest (BF) with $k = 1000$ and $k = 10,000$ bootstraps for the OSB data revealed a set of five common predictor variables: ‘wet bin speed’, ‘flake moisture content’, ‘dryer inlet temperature’, ‘dryer outlet temperature’, and ‘wood weight’ (Table 3). The R^2 in training and R^2 in validation for the BT was 0.675 and 0.410, respectively. The R^2 in training and validation for the BF were 0.534 and 0.304, respectively. The means and variances of the dependent variable ultimate static load (kg) in the training and validation datasets were similar across the 10-fold cross validation with an $\bar{x} = 213.7$, $\bar{s} = 28.1$, $\bar{M} = 210.6$, $\bar{CV} = 5.9\%$ in the training dataset; and $\bar{x} = 204.8$, $\bar{s} = 27.7$, $\bar{M} = 204.0$, $\bar{CV} = 6.1\%$ in the validation dataset. Again, even though the R^2 in validation is not high, these five variables explain a reasonable proportion of the variation influencing ultimate static load. In order to find reasonable split points and levels for factors to be tested in designed experiment regression, tree models were developed using the five common independent variables previously mentioned.

Table 3. Common predictors variables for both bootstrap forest and boosted tree models for OSB after $k = 1000$ bootstraps, ranked by highest sums of squares.

Bootstrap Forest ($k = 1000$ Bootstraps)			Boosted Tree ($k = 1000$ Bootstraps)		
Variable	Number of Splits	Sum of Squares	Variable	Number of Splits	Sum of Squares
Wet Bin Speed	7	0.117	Wet Bin Speed	4	0.050
Dryer Inlet Temperature	2	0.023	Flake Moisture Content	4	0.049
Flake Moisture Content	2	0.019	Dryer Inlet Temperature	3	0.035
Dryer Outlet Temperature	1	0.015	Dryer Outlet Temperature	2	0.017
Mat Weight	1	0.012	Mat Weight	2	0.017
Wood Weight	1	0.009	Wood Weight	2	0.016
Bootstrap Forest ($k = 10,000$ Bootstraps)			Boosted Tree ($k = 10,000$ Bootstraps)		
Variable	Number of Splits	Sum of Squares	Variable	Number of Splits	Sum of Squares
Dryer Outlet Temperature	5	0.172	Dryer Inlet Temperature	8	0.062
Dryer Inlet Temperature	5	0.093	Wet Bin Speed	5	0.045
Wood Weight	4	0.076	Wood Weight	3	0.034
Flake Moisture Content	6	0.035	Press Closing Time	6	0.026
Dry Bin Speed	2	0.026	Dryer Outlet Temperature	3	0.020
Wet Bin Speed	1	0.024	Flake Moisture Content	4	0.016

4.2. RT Models as a Precursor for Response Surface Methods

The following examples for designed experiments are presented to demonstrate that kernel tree results can be helpful in designing a DOE without time consuming prescreening designs. Assuming that experimental runs are costly, several response surface designs are presented that minimize the number of runs while maintaining an $\alpha = 0.05$ while maximizing the power of the experiment.

Designs for MDF from the RT model. From the RT model for MDF, a fractional factorial Resolution IV design with 2^{k-1} , $k = 4$ with a replicate would require $n = 16$ runs. This design has a power of 82% for the main effects ($\alpha = 0.05$), given a standard deviation of 95 kPa with the ability to detect a difference in the mean tensile strength of 950 kPa (Table 4). From the RT model results (Figure 1), the factor notation is: A = ‘face plate position’, B = ‘adhesive percent’, C = ‘press position time’, and D = ‘face steam flow.’ The aliasing structure for this fractional factorial design is A = BCD, B = ACD, C = ABD, D = ABC, AB = CD, AC = BD, and AD = BC. Even though this may be a feasible design for the analyst, it only explores the corner points of the data space. Box and Behnken [1] proposed the Box–Behnken design when points at the corners are costly [48]. A Box–Behnken design with $n = 27$ runs is presented in Figure 3, with the run order given in Table 5. If the analyst can afford three more experimental runs and corner points may be risky for the process, a spherical circumscribed central composite (CCD) RSM, $k = 4$, with four axial points is an option (Figure 3). A full factorial design would have only the corner points examined and requires $n = 32$ runs which includes a replicate.

Table 4. Descriptive statistics for the dependent and independent variables in the regression tree (RT) model for MDF.

Tensile Strength (kPa) Quantiles			Face Plate Position (mm) Quantiles			Core Dust Speed (m/min) Quantiles		
100.0%	maximum	1275.575	100.0%	maximum	10.459	100.0%	maximum	24.542
99.5%		1226.483	99.5%		10.456	99.5%		24.195
97.5%		1137.675	97.5%		10.422	97.5%		22.788
90.0%		1067.346	90.0%		9.720	90.0%		19.928
75.0%	quartile	1020.46	75.0%	quartile	9.140	75.0%	quartile	16.468
50.0%	median	951.51	50.0%	median	8.324	50.0%	median	10.921
25.0%	quartile	882.56	25.0%	quartile	5.492	25.0%	quartile	8.245
10.0%		827.4	10.0%		4.160	10.0%		7.094
2.5%		767.4135	2.5%		3.135	2.5%		6.128
0.5%		704.1174	0.5%		2.271	0.5%		3.539
0.0%	minimum	668.815	0.0%	minimum	2.180	0.0%	minimum	2.851
Summary Statistics			Summary Statistics			Summary Statistics		
Mean		950.0881	Mean		0.321	Mean		41.069
Std Dev		94.6977	Std Dev		0.068	Std Dev		16.341
Std Err Mean		3.831056	Std Err Mean		0.003	Std Err Mean		0.772
Upper 95% Mean		957.6118	Upper 95% Mean		0.327	Upper 95% Mean		42.587
Lower 95% Mean		942.5644	Lower 95% Mean		0.315	Lower 95% Mean		39.552
N		408	N		408	N		408
Quantiles			Face Steam Flow (bar) Quantiles			Adhesive Percent (%) Quantiles		
100.0%	100.0%	100.0%	100.0%	maximum	292.618	100.0%	maximum	14.743
99.5%	99.5%	99.5%	99.5%		285.132	99.5%		14.583
97.5%	97.5%	97.5%	97.5%		231.635	97.5%		14.089
90.0%	90.0%	90.0%	90.0%		190.579	90.0%		13.687
75.0%	75.0%	75.0%	75.0%	quartile	160.880	75.0%	quartile	9.555
50.0%	50.0%	50.0%	50.0%	median	114.143	50.0%	median	9.468
25.0%	25.0%	25.0%	25.0%	quartile	69.911	25.0%	quartile	8.579
10.0%	10.0%	10.0%	10.0%		60.558	10.0%		7.662
2.5%	2.5%	2.5%	2.5%		51.778	2.5%		5.985
0.5%	0.5%	0.5%	0.5%		49.570	0.5%		5.679
0.0%	0.0%	0.0%	0.0%	minimum	48.733	0.0%	minimum	5.671
Summary Statistics			Summary Statistics			Summary Statistics		
Mean	Mean		Mean		121.76	Mean		9.50
Std Dev	Std Dev		Std Dev		53.10	Std Dev		1.93
Std Err Mean	Std Err Mean		Std Err Mean		2.36	Std Err Mean		0.09
Upper 95% Mean	Upper 95% Mean		Upper 95% Mean		126.39	Upper 95% Mean		9.68
Lower 95% Mean	Lower 95% Mean		Lower 95% Mean		117.13	Lower 95% Mean		9.31
N		408	N		408	N		408

Figure 1. Regression tree for MDF tensile strength derived from boosted tree and random forest models.

Table 5. Box–Behnken design for tensile strength ($n = 27$) for 2^k , $k = 4$ with three center points based on the results of the RT for MDF.

Runs	Pattern	Face Plate Position	Press Position Time	Adhesive Percent	Face Steam Flow	Random Run Order
1	--00	9.3	8.0	7.75	95.5	9
2	-0-0	9.3	8.2	7.60	95.5	5
3	-00-	9.3	8.2	7.75	90.5	1
4	-00+	9.3	8.2	7.75	100.5	18
5	-0+0	9.3	8.2	7.90	95.5	23
6	-+00	9.3	8.4	7.75	95.5	26
7	0--0	9.5	8.0	7.60	95.5	16
8	0-0-	9.5	8.0	7.75	90.5	17
9	0-0+	9.5	8.0	7.75	100.5	2
10	0-+0	9.5	8.0	7.90	95.5	25
11	00--	9.5	8.2	7.60	90.5	21
12	00-+	9.5	8.2	7.60	100.5	15
13	0000	9.5	8.2	7.75	95.5	7
14	0000	9.5	8.2	7.75	95.5	11
15	0000	9.5	8.2	7.75	95.5	3
16	00+-	9.5	8.2	7.90	90.5	4
17	00++	9.5	8.2	7.90	100.5	27
18	0+-0	9.5	8.4	7.60	95.5	12
19	0+0-	9.5	8.4	7.75	90.5	10
20	0+0+	9.5	8.4	7.75	100.5	19
21	0++0	9.5	8.4	7.90	95.5	8
22	+--00	9.7	8.0	7.75	95.5	13
23	+0-0	9.7	8.2	7.60	95.5	24
24	+00-	9.7	8.2	7.75	90.5	14
25	+00+	9.7	8.2	7.75	100.5	6
26	+0+0	9.7	8.2	7.90	95.5	20
27	+ +00	9.7	8.4	7.75	95.5	22

Designs for OSB from the RT model. Given that the OSB RT model resulted in three significant factors, A = ‘dryer inlet temperature’, B = ‘mat weight’, and C = ‘wet bin speed’ at an $\alpha = 0.05$, a full factorial design is feasible, e.g., 2^k , $k = 3$, $n = 16$ with a power of 84% for significant main effects (assuming an $\alpha = 0.05$ for the experiment) (see Table 6 and Figure 2). Even though this design has no aliasing assumptions, it only explores corner points. A Box–Behnken RSM design would require two fewer runs, $n = 15$, with three center points (Figure 3). An alternative design would be a spherical circumscribed CCD RSM design with $n = 17$, $k = 3$, with four axial points (Figure 3), where the run order is presented in Table 7. Even though there are many other feasible designs, the aforementioned designs were developed to minimize the number of experimental runs while using the helpful results of the RT models.

Table 6. Descriptive statistics for the dependent and independent variables in the RT model for OSB.

Ultimate Static Load (kg) Quantiles			Dryer Inlet Temperature (°C) Quantiles		
100.0%	maximum	283.04	100.0%	maximum	145.00
99.5%		283.04	99.5%		145.00
97.5%		269.20	97.5%		133.33
90.0%		250.84	90.0%		127.22
75.0%	quartile	232.35	75.0%	quartile	125.00
50.0%	median	209.11	50.0%	median	120.28
25.0%	quartile	189.71	25.0%	quartile	112.22
10.0%		174.22	10.0%		103.33
2.5%		168.53	2.5%		96.67
0.5%		147.87	0.5%		93.33
0.0%	minimum	147.87	0.0%	minimum	93.33
Summary Statistics					
Mean		211.96	Mean		117.92
Std Dev		28.23	Std Dev		8.36
Std Err Mean		2.31	Std Err Mean		1.38
Upper 95% Mean		216.52	Upper 95% Mean		119.44
Lower 95% Mean		207.41	Lower 95% Mean		116.40
N		150	N		150
Wood Weight (kg) Quantiles			Wet Bin Speed (m/min) Quantiles		
100.0%	maximum	19.66	100.0%	maximum	9.75
99.5%		19.66	99.5%		9.75
97.5%		19.66	97.5%		9.45
90.0%		5.68	90.0%		9.14
75.0%	quartile	4.38	75.0%	quartile	9.14
50.0%	median	3.20	50.0%	median	8.53
25.0%	quartile	2.19	25.0%	quartile	7.92
10.0%		0.10	10.0%		7.32
2.5%		0.10	2.5%		6.40
0.5%		0.10	0.5%		6.10
0.0%	minimum	0.10	0.0%	minimum	6.10
Summary Statistics					
Mean		3.54	Mean		8.45
Std Dev		4.19	Std Dev		0.84
Std Err Mean		0.34	Std Err Mean		0.07
Upper 95% Mean		4.21	Upper 95% Mean		8.59
Lower 95% Mean		2.86	Lower 95% Mean		8.32
N		150	N		150

Figure 2. Regression tree for OSB ultimate static load derived from boosted tree and random forest models.

Figure 3. Cont.

Small colored blocks are sample points with number of runs, no number next to block indicates one run.

Figure 3. Designed experimental models for MDF and OSB using the outcomes of the boosted tree and random forest models.

Table 7. Central composite design for ultimate static load ($n = 17$) for 2^k , $k = 3$ with three center points based on the results of the RT for OSB.

Runs	Pattern	Dryer Inlet Temperature	Mat Weight	Wet Bin Speed	Random Run Order
1	---	105	1.4	8.4	6
2	--+	105	1.4	9.4	11
3	a00	105	1.5	8.9	15
4	-+-	105	1.6	8.4	10
5	++	105	1.6	9.4	16
6	0a0	107.5	1.4	8.9	12
7	00a	107.5	1.5	8.4	13
8	000	107.5	1.5	8.9	4
9	000	107.5	1.5	8.9	7
10	00A	107.5	1.5	9.4	17
11	0A0	107.5	1.6	8.9	2
12	+--	110	1.4	8.4	1
13	+--	110	1.4	9.4	8
14	A00	110	1.5	8.9	3
15	++-	110	1.6	8.4	14
16	+++	110	1.6	9.4	5
17	---	105	1.4	8.4	9

5. Conclusions

Identifying interaction effects is important in innovation and product development. Kernel tree methods provide an accepted data science approach for enhancing the applied sciences. Such methods quickly identify and quantify undiscovered interactions among regressors. In this research, boosted trees and random forest models were constructed from two different manufacturing systems. Common significant variables in the models that effected strength of materials were related to speed of the process, fiber refining, pressing time, and pressing pressures. Experimental designs were proposed with a Resolution $> III$, while incorporating the aliasing structure identified from the RT models. Possible designs with four factors were: fractional factorial Resolution IV; Box–Behnken RSM; and CCD spherical circumscribed designs. The above designs were selected to minimize the number of experimental runs while sustaining inference. The hierarchy of interaction effects and split-points of regressors in kernel tree models may provide applied scientists with an important foundation for planning a designed experiment while minimizing the costs during innovation development. If feasible, future studies will explore support vector machine, Bayesian additive regression trees (BART), etc., for comparative analyses.

Author Contributions: All authors contributed to the development of this paper. T.M.Y. did the primary statistical analysis and development of original draft manuscript. R.A.B. and T.L. were instrumental in procuring the industrial datasets and validating the results of the models. A.P.'s contribution was invaluable in advising on the methodology, results, and the review of the overall manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the University of Tennessee Institute of Agriculture AgResearch McIntire-Stennis project [TEN00MS-107] APC funded by [1006012].

Conflicts of Interest: No conflicts of interest.

References

- Box, G.E.P.; Behnken, D. Some new three level designs for the study of quantitative variables. *Technometrics* **1960**, *2*, 455–475. [\[CrossRef\]](#)
- Fielding, A. Binary segmentation: The automatic detector and related techniques for exploring data structure. In *The Analysis of Survey Data, Exploring Data Structures*; O'Muircheartaigh, C.A., Payne, C., Eds.; John Wiley and Sons, Inc.: New York, NY, USA, 1977; Volume I, pp. 221–257.
- Kass, G.V. Significance testing in automatic interaction detection (A.I.D.). *Appl. Stat.* **1975**, *24*, 178–189. [\[CrossRef\]](#)
- Loh, W.Y. Regression trees with unbiased variable selection and interaction detection. *Stat. Sin.* **2002**, *12*, 361–386.
- Morgan, J.N.; Sunquist, J.A. Problems in the analysis of survey data and a proposal. *J. Am. Stat. Assoc.* **1963**, *58*, 415–434. [\[CrossRef\]](#)
- Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [\[CrossRef\]](#)
- Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [\[CrossRef\]](#)
- Friedman, J.H.; Meulman, J.J. Multiple additive regression trees with application in epidemiology. *Stat. Med.* **2003**, *22*, 1365–1381. [\[CrossRef\]](#)
- Kim, H.; Loh, W.Y. Classification trees with unbiased multiway splits. *J. Am. Stat. Assoc.* **2001**, *96*, 589–604. [\[CrossRef\]](#)
- Kim, H.; Loh, W.Y. Classification trees with bivariate linear discriminant node models. *J. Comput. Graph. Stat.* **2003**, *12*, 512–530. [\[CrossRef\]](#)
- Kim, H.; Guess, F.M.; Young, T.M. Using data mining tools of decision trees in reliability applications. *IIE Trans.* **2011**, *43*, 43–54.
- Stoma, P.; Stoma, M.; Dudziak, A.; Caban, J. Bootstrap analysis of the production processes capability assessment. *Appl. Sci.* **2019**, *9*, 5360. [\[CrossRef\]](#)
- Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [\[CrossRef\]](#)
- Schwarz, G.E. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [\[CrossRef\]](#)
- Adcock, T.; Wolcott, M.P. Wood: Structural Panel Processes. In *Encyclopedia of Materials: Science and Technology*; Buschow, K.H.J., Cahn, R.W., Flemings, M.C., Ilshner, B., Kramer, E.J., Mahajan, S., Veyssire, P., Eds.; Elsevier: Amsterdam, The Netherlands, 2001; pp. 9678–9683.
- Kamke, F.A. Wood: Nonstructural panel processes. In *Encyclopedia of Materials: Science and Technology*; Buschow, K.H.J., Cahn, R.W., Flemings, M.C., Ilshner, B., Kramer, E.J., Mahajan, S., Veyssire, P., Eds.; Elsevier: Amsterdam, The Netherlands, 2001; pp. 9673–9678.
- Chaudhuri, P.; Huang, M.C.; Loh, W.Y.; Yao, R. Piecewise-polynomial regression trees. *Stat. Sin.* **1994**, *4*, 143–167.
- Dath, G.; Fabricius, K. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology* **2000**, *81*, 3178–3192. [\[CrossRef\]](#)
- Loh, W.Y.; Vanichsetakul, N. Tree-structured classification via generalized discriminant analysis. *J. Am. Stat. Assoc.* **1988**, *83*, 715–728. [\[CrossRef\]](#)
- Hand, D.J.; Mannila, H.; Smyth, P. *Principles of Data Mining (Adaptive Computation and Machine Learning)*, 3rd ed.; MIT Press: Cambridge, MA, USA, 2001.

21. André, N.; Young, T.M. Real-time process modeling of particleboard manufacture using variable selection and regression methods ensemble. *Eur. J. Wood Wood Prod.* **2013**, *71*, 361–370. [[CrossRef](#)]
22. Carty, D.M.; Young, T.M.; Zaretzki, R.L.; Guess, F.M.; Petutschnigg, A. Predicting the strength properties of wood composites using boosted regression trees. *Forest Prod. J.* **2015**, *65*, 365–371. [[CrossRef](#)]
23. Cherkassky, V.S.; Mulier, F. *Learning from Data: Concepts, Theory, and Methods*; John Wiley & Sons, Inc.: New York, NY, USA, 1998; pp. 1–536.
24. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. *From Data Mining to Knowledge Discovery: An Overview of Advances in Knowledge Discovery and Data Mining*; The MIT Press: Cambridge, MA, USA, 1996; pp. 1–34.
25. Loh, W.Y. Classification and regression trees. *WIREs Data Min. Knowl.* **2011**, *1*, 14–23. [[CrossRef](#)]
26. Young, T.M.; León, R.V.; Chen, C.-H.; Chen, W.; Guess, F.M.; Edwards, D.J. Robustly estimating lower percentiles when observations are costly. *Qual. Eng.* **2015**, *27*, 361–373. [[CrossRef](#)]
27. Young, T.M.; Clapp, N.E., Jr.; Guess, F.M.; Chen, C.-H. Predicting key reliability response with limited response data. *Qual. Eng.* **2014**, *26*, 223–232. [[CrossRef](#)]
28. Zeng, Y.; Young, T.M.; Edwards, D.J.; Guess, F.M.; Chen, C.-H. Case studies: A study of missing data imputation in predictive modeling of a wood composite manufacturing process. *J. Qual. Technol.* **2016**, *48*, 284–296.
29. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.I. *Classification and Regression Trees*; Wadsworth: Belmont, CA, USA, 1984; pp. 199–215.
30. Luna, J.M.; Gennatas, E.D.; Ungar, L.H.; Valdes, G. Building more accurate decision trees with the additive tree. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 19887–19893. [[CrossRef](#)]
31. Schapire, R.E. The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification*; Denison, D.D., Hansen, M.H., Holmes, C., Mallick, B., Yu, B., Eds.; Springer: New York, NY, USA, 2003; pp. 113–141.
32. Feng, J.; Yu, Y.; Zhou, Z.-H. Multi-layered gradient boosting decision trees. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS), Montréal, QC, Canada, 3–8 December 2018; pp. 3555–3565.
33. Khan, Z.; Gul, A.; Perperoglou, A. Ensemble of optimal trees, random forest and random projection ensemble classification. *Adv. Data Anal. Cl.* **2020**, *14*, 97–116. [[CrossRef](#)]
34. Khuri, N. Mining environmental chemicals with boosted trees. In Proceedings of the 35th Annual ACM Symposium on Applied Computing, Brno, Czech Republic, 30 March–3 April 2020; pp. 1082–1089.
35. Elith, J.; Leathwick, J.R.; Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol.* **2008**, *77*, 802–813. [[CrossRef](#)]
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
37. Fawagreh, K.; Gaber, M.M.; Elyan, E. Random forests: From early developments to recent advancements. *J. Syst. Sci. Syst. Eng.* **2014**, *2*, 602–609. [[CrossRef](#)]
38. Amit, Y.; Geman, D. Shape quantization and recognition with randomized trees. *Neural Comput.* **1997**, *9*, 1545–1588. [[CrossRef](#)]
39. Ho, T.K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal.* **1998**, *20*, 832–844.
40. Boinee, P.; De Angelis, A.; Foresti, G.L. Meta random forests. *Int. J. Comput. Int. Syst.* **2005**, *2*, 138–147.
41. Gregorutti, B.; Michel, B.; Saint-Pierre, P. Correlation and variable importance in random forests. *Stat. Comput.* **2017**, *27*, 659–678. [[CrossRef](#)]
42. Jaiswal, J.K.; Samikannu, R. Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. In Proceedings of the World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, India, 2–4 February 2017; pp. 65–68.
43. Liaw, A.; Wiener, M. Classification and regression by randomforest. *IRNews* **2002**, *2*, 18–22.
44. Attewell, P.; Monaghan, D. *Data Mining for the Social Sciences: An Introduction*; University of California Press: Berkeley, CA, USA, 2015; pp. 1–264.
45. Fisher, R.A. *The Design of Experiments*; Hafner Publishing Company: New York, NY, USA, 1971; pp. 23–36.
46. Box, G.E.P. Science and statistics. *J. Am. Stat. Assoc.* **1976**, *71*, 791–799. [[CrossRef](#)]

47. Pattengale, N.D.; Alipour, M.; Bininda-Emonds, O.R.P.; Moret, B.M.E.; Stamatakis, A. *How Many Bootstrap Replicates Are Necessary*; Batzoglou, S., Ed.; RECOMB, LNCS 5541; Springer: Berlin/Heidelberg, Germany, 2009; pp. 184–200.
48. Box, G.E.P.; Draper, N.R. *Empirical Model Building and Response Surfaces*; John Wiley and Sons: New York, NY, USA, 1987; pp. 1–688.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).