Application of decision tree-based techniques to veneer processing

Suborna Shekhor Ahmed[1*] , Julie Cool[2] and Mohammad Ehsanul Karim[3,4]

Abstract

In veneer-drying facilities, controllers face many challenges to maintain desired parameters in the final product based on customer's needs. The major challenge is setting process parameters to control the temperature and humidity within the various sections in the drying machine to obtain the desired properties of the final product. The regression tree approach can be used to simplify the complex relationship among process and product variables for identifying critical factors for drying veneer and achieving the desired range of veneer temperature. In this study, we investigated veneer-drying conditions and the short-term effect of climatic variables on veneer temperature. We have shown a three-step process to develop an optimal regression tree for veneer temperature. From the developed optimal tree, we are able to identify the most important threshold points of predictor space and adjustment for the climatic variables on the temperature of veneer sheets. The findings of this study were further investigated in an industrial setting and the desired veneer temperatures were attained for the final product. This application shows that we can follow the path of the optimal tree to pinpoint the most desired veneer temperature outcome. The developed optimal tree is relatively easy to use and interpret to estimate the average response of veneer temperature.

Keywords: Decision tree, Regression tree, Veneer, Data mining, Climatic variables, ANOVA, Cross-validation

Introduction

Processing veneer sheets in a drying machine involves many process parameters that need to be set by expert personnel to control the temperature and humidity within the various sections. Drying speed, amount of gas flow, air flow, etc., are some of the process parameters, while product parameters are thickness of veneer sheet, types of wood and species [1]. In a veneer-drying facility, process parameters are adjusted to a certain level to control the final average moisture content, the temperature of veneer sheets and the average ultrasonic propagation time (UPT), which is correlated with the modulus of elasticity [2–4]. The final moisture content, temperature and UPT level of veneers are very much dependent on the type of wood and the thickness of veneer sheets [5, 6]. Climatic variables also influence the final product quality

since, after peeling or slicing, the veneers are often stored outdoors before being sent to a veneer-drying machine. During that time, climatic variables may affect the moisture content and temperature of veneer sheets, which can degrade the final product if the process parameters are not adjusted accordingly.

Veneer temperature is an important response factor that can be used to evaluate the quality of the product. While exiting the dryer, a veneer sheet having a temperature ranging from 77 to 93 °C meets the quality requirements. Otherwise, the higher-temperature veneer sheet can indicate an over-dry problem and even increase the risk of fire occurrences inside the drying machine, whereas a low-temperature veneer sheet needs re-drying, which increases the drying cost. To maintain the quality of veneer, we need to understand the process of drying veneer. We focused on the veneer temperature as the outcome of the current study because it is related to the veneer moisture content and UPT, which influence chemical adhesion and plywood strength. The goal of this study is to use a data mining approach to understand the

*Correspondence: suborna.ahmed@ubc.ca
[1] Department of Forest Resources Management, The University of British Columbia, 2045-2424 Main Mall, Vancouver, BC V6T 1Z4, Canada
Full list of author information is available at the end of the article

process of veneer drying and to interpret the effects of the predictor variables (see further details in the "Methods" section) on the veneer temperature. For that purpose, we apply a decision tree approach; specifically, a regression tree which is a commonly used data mining method [7]. The basic idea of a regression tree approach is to develop a flowchart to show the structure of data [8]. Compared to a regular regression model used popularly, the regression tree approach has several advantages, e.g., allowing for linear or nonlinear relationships, can handle complex relationship among predictors, overlook prior knowledge of functional form [9, 10]. Given the complex relationship of predictor variables on the outcome, we want to be able to interpret the veneer-drying system in such a way that would be accessible by a wider audience and non-experts using graphical tools and outputs [11] associated with these regression trees.

In this study, we have used a dataset from industrial veneer dryers fitted with sensors. The goal was to identify a suitable range of potential predictor variables to dry veneer and control outcomes to maximize the production of high-quality products while reducing energy consumption. Due to the lack of detail and uncertainty about the combination of process parameters, a large percentage of the product fails to meet the quality requirements. For drying veneer, one of the difficulties is to find an optimal setting to dry at a certain level so that the resulting moisture content of veneers is not more or less than what is desired. Ideally, the industry would like to minimize the occurrence of fire due to extreme heat and/or relative humidity, which causes loss of product.

Methods

Data description

The dataset was collected from a veneer dryer over a period of 6 months (February–July, 2017). Because the equipment is in operation every day for 24 h, data for 3,464,518 veneer sheets were recorded. For each veneer, temperature, moisture content and UPT were recorded as output variables. Additionally, veneer thickness levels and wood types were recorded as input variables. There were three veneer thickness levels ("Thik") dried in the facility (2.540 mm, 3.175 mm, and 3.632 mm) and three wood types ("Prod") categorized as (i) sap; (ii) light sap ("Ls"); and (iii) heartwood ("Hrt").

Information on the process variables, also considered as input variables, were frequently collected. In particular, information regarding (i) gas usage (giga Joules at 11 Psi); (ii) drying time (drying speed) (minute); (iii) zone temperatures (°C); and (iv) chain side temperatures (°C). The dryer was divided into three zones (Zone 1, Zone 2 and Zone 3) with sub-divisions within the first two zones. The first zone was sub-divided into three zones (Zone 1a, Zone 1b and Zone 1c), while the second one was sub-divided into two zones (Zone 2a and Zone 2b). Temperature from each zone and chain side were collected from sensors and drying time along with the temperature of veneer sheet while exiting the dryer (Fig. 1).
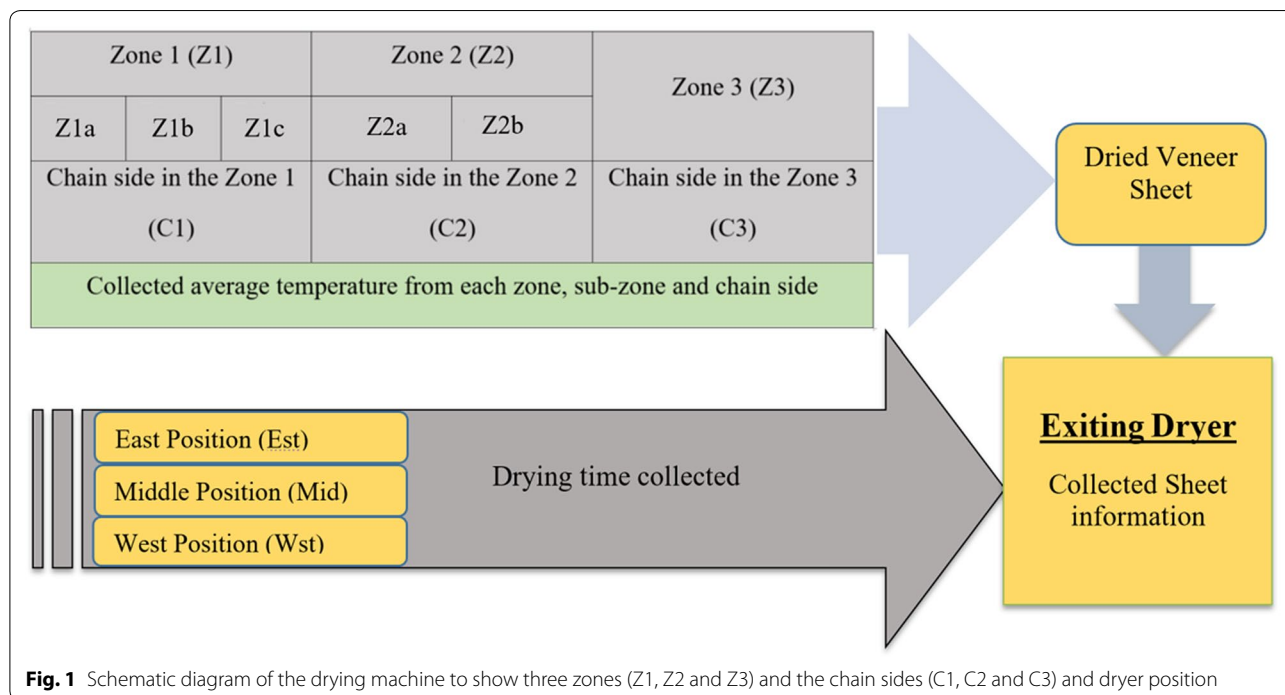

Fig. 1 Schematic diagram of the drying machine to show three zones (Z1, Z2 and Z3) and the chain sides (C1, C2 and C3) and dryer position

Ahmed *et al. J Wood Sci* (2020) 66:54

Page 3 of 8

Average chain side temperature was collected for each zone and named as (i) C1 (average chain side temperature in the zone 1) (°C); (ii) C2 (average chain side temperature in the zone 2) (°C); and (iii) C3 (average chain side temperature in the zone 3) (°C). The drying machine also had three drying positions ("DP"): (i) East ("Est"), (ii) West ("Wst"), and (iii) Middle ("Mid") and four deck levels ("DL") divided into two groups: (i) top (upper two decks) and (ii) bottom ("Bot": lower two decks).

The effect of climatic variables on output variables was also investigated. Historical daily weather station data for the 6-month (February–July, 2017) period were extracted from the Environment and Natural Resources of Canada database [12]. The Vancouver International Airport's weather station was selected because it is the closest weather station from both the veneer peeling and drying facilities. Mean daily temperature (MDT, °C) in a week and total precipitation in a week (TWP, mm) were calculated from the daily weather station data.

All input and output variables were validated using summary statistics and known operational ranges. It was observed that some of the values recorded for the process parameters were erroneous, so they were removed from the database. For example, drying speed cannot be less than 5 min or more than 15 min. Drying speed outside this range was removed from the data.

Decision-based approach

In this study, we chose to focus on the output variable "veneer temperature". Since it is a continuous variable, we selected a regression tree approach to develop an optimal decision tree. Regression tree is a very popular technique used in remote sensing, ecology [13] and in various disciplines where relationships among response and predictor variables are not certain and mathematical expression of the relationship is difficult to identify [10]. A single tree-based approach finds the mean response of all observations and then partitions the data into two groups by selecting a predictor variable from the predictor space. In this study, the analysis of variance method (ANOVA) was used to partition the data into two homogeneous groups based on a single predictor variable. Along the way, data were partitioned into homogeneous groups based on the previously used predictor variable (or another predictor variable), and hence reducing the data. Data partitioning or splitting was done to maximize the homogeneity of the output variable "veneer temperature". Each homogeneous group shows the summary statistics of average temperature and the percentage of data belongs to that group.

A three-step process was used to determine the optimal regression tree: (i) grow the first tree to understand its underlying structure; (ii) grow a big tree-based to assess the optimal tree size based on the complexity parameter

(C_p) values and relative errors; and (iii) prune the big tree by adding cross-validation and obtain an optimal tree size. In this paper, the first two steps are discussed as they directly lead to the third one. The 'rpart' package in the R (R version 3.5.0) [14] was used to develop all regression trees.

Results and discussion

Comprehending the basic structure of a regression tree

The first step in developing a regression tree to determine the impact of process variables on the continuous dependent variable veneer temperature was to grow a tree to understand its basic structure. The fitted first regression tree, including splits along with root (top of the tree), nodes (terminal and internal) and branches, is presented in Fig. 2. In each node (inside the circle in Fig. 2), the average value of the dependent variable (veneer temperature) and the percentage of observations is shown (Fig. 2). For example, the root node has the entire dataset, and the average of the continuous response variable (veneer temperature) is 91 °C. The data are then divided into two homogeneous groups based on the C3 temperature which is called a sub-node. Using the
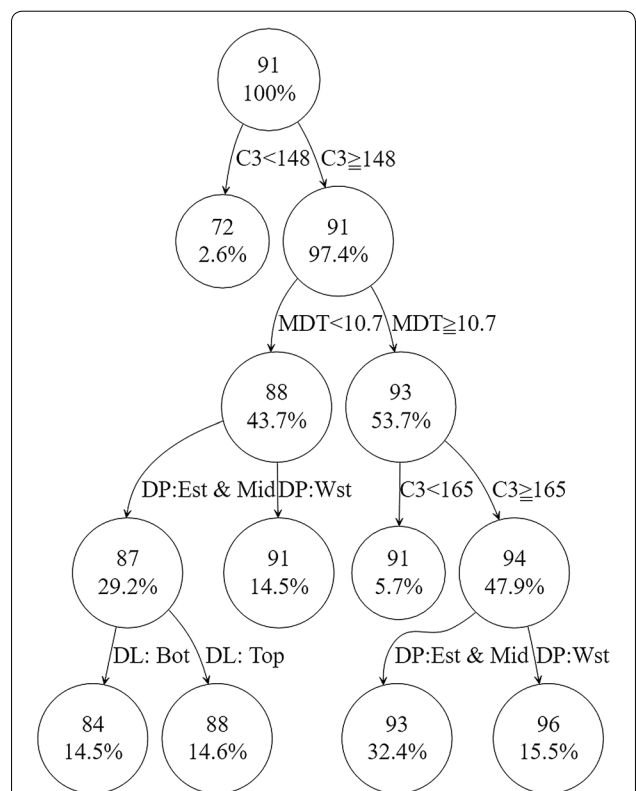

Fig. 2 Fitted first regression tree of temperature of veneer. The value inside the circles indicates the veneer temperatures in °C and the percentage of the observations

Ahmed *et al. J Wood Sci*   (2020) 66:54

Page 4 of 8

ANOVA method, the regression tree procedure determined that a temperature (C3 temperature) of 148 °C maximized between groups sum of squares among all variables. It is possible to show a vector of summary statistics in each node, but the only average of veneer temperature and percentage of observation were used to reduce the complexity of Fig. 2.

The ANOVA splitting method was used to increase the R-squared value at each step (or split), while reducing the C_p values to improve the prediction ability of the model. In this specific regression tree, a threshold C_p of 0.01 was selected to enhance group homogeneity. The C_p values at each split of the fitted regression tree along with their corresponding error values are presented in Table 1, and indicate that six splits were necessary to reach the threshold value of 0.01. Whenever a splitting occurred in the regression tree method, it improved the resulting fitted tree by reducing its C_p value. C_p value not reducing further indicates that there was no improvement, and the tree was trimmed off at that particular split. Table 1 shows all splitting steps while developing the tree. At the initial stage of the fitting, there were only the observed data and summary statistics without any split. The height of the tree was getting bigger by allowing more splits until it reaches C_p of 0.01. In Table 1, a cross-validation error was generated from a tenfold cross-validation (used as the default in the implementation of 'rpart' function) to minimize the error and evaluate the fitted tree. In each split, this validation approach was performed to quantify the validation error. In this validation approach, the entire dataset was divided into ten randomly selected parts and fitted the regression tree onto the nine folds and calculated validation error from the left-out fold. We compared cross-validation error with C_p value and number of splits. Cross-validation error and C_p value both reduced with the increase in split.

Optimizing the size of the tree

A smaller C_p value (0.0001), a minimum number of splits (5), and a minimum of observations per node (5) were used to develop a larger tree. Also, for fitting the larger tree, the number of cross-validation was set at 10, which corresponds to a tenfold cross-validation. As expected, increase in tree size improved the C_p value, while reducing the relative error (Fig. 3). The rate of improvement in the C_p value, evaluated using the cross-validation relative error, is more substantial as tree size increased from 1 to approximately 10 nodes. As the tree became larger in size, less improvement was noticed, which corresponded to a cross-validation relative error of 0.70. The challenge in optimizing the size of the tree consists in identifying the number of splits that minimize overfitting. In other words, it is essential to determine where the decrease in relative error is negligible in comparison to the increase in splits. To achieve this, the cross-validation relative error was compared with the sum of the relative error and the cross-validation standard error. If the sum was less than the former, the tree could be pruned at its corresponding split. In this study, a tree having 30 nodes was selected as optimal because the C_p value was no longer improving, which
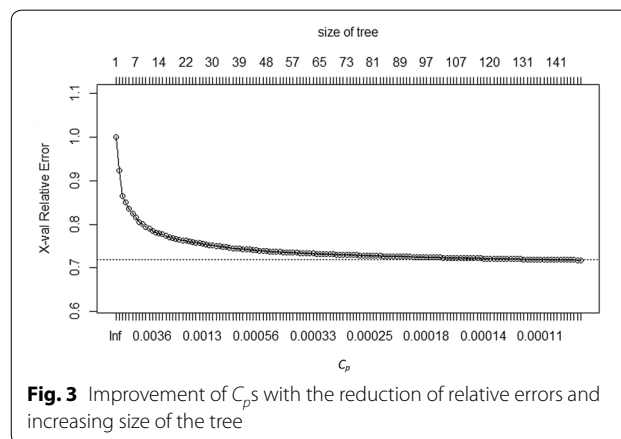

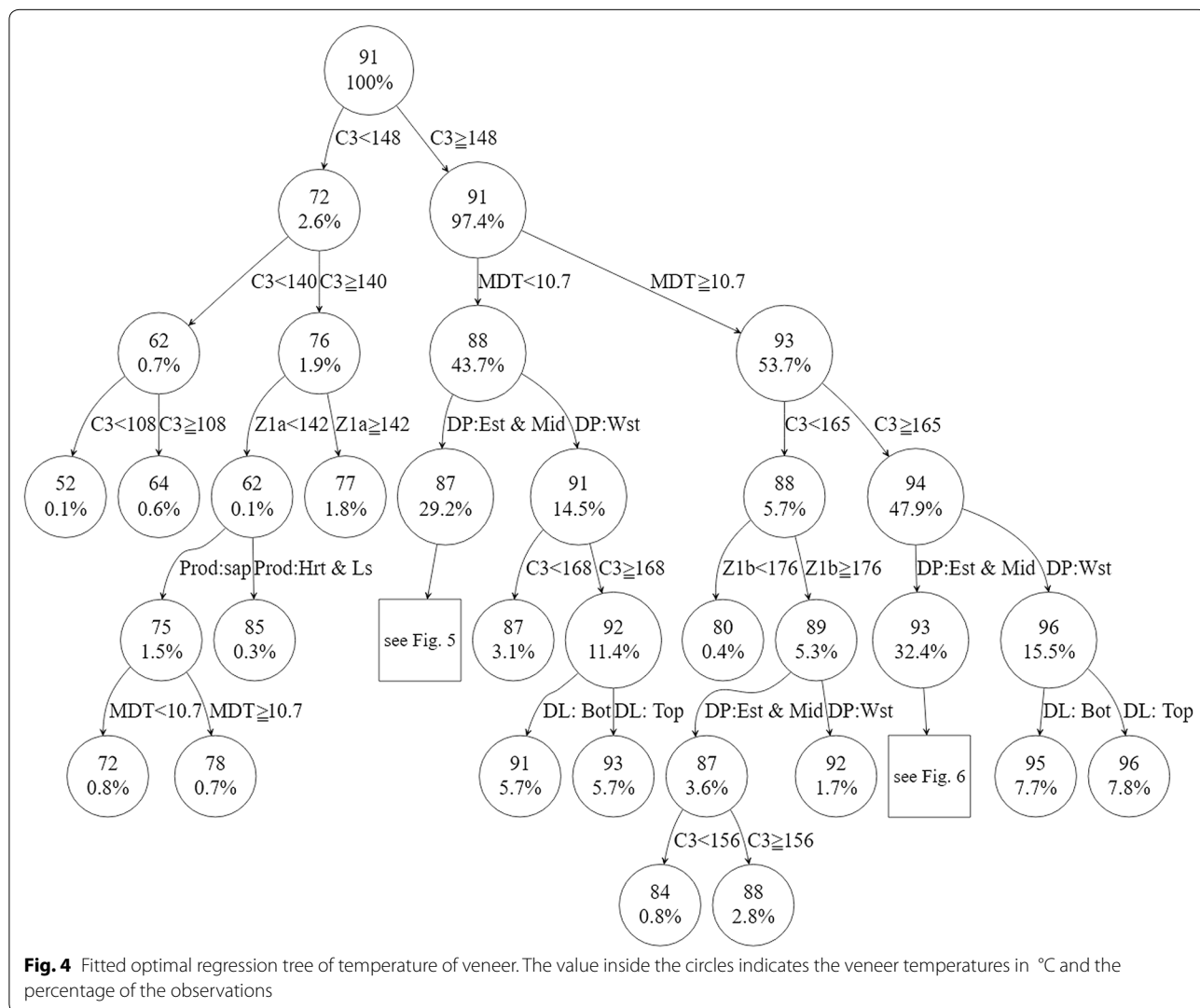**Fig. 3** Improvement of C_ps with the reduction of relative errors and increasing size of the tree

Table 1 Splitting criteria with validation statistics in the fitted regression tree of temperature of veneers

Complexity parameter (C_p)	Split number	Relative error	Cross-validation error	Standard error of cross-validation error
0.077	0	1.000	1.000	0.0012
0.057	1	0.923	0.923	0.0011
0.015	2	0.865	0.866	0.0011
0.015	3	0.850	0.850	0.0011
0.011	4	0.836	0.836	0.0011
0.011	5	0.825	0.825	0.0011
0.010	6	0.814	0.819	0.0010

Ahmed *et al. J Wood Sci*   (2020) 66:54

Page 5 of 8

corresponds to a minimum cross-validation error of 0.75 and a relative error of 0.75.

The fitted optimal tree (Figs. 4, 5 and 6) simplifies the complex relationship among temperature of veneers and predictor variables by dividing data into nonoverlapping homogeneous groups and sub-groups. The advantage of this approach lies in the fact that to estimate an average response of veneer temperature, one merely has to follow the path. It also highlights the variables of importance. In the regression tree, variables that were used in splitting into homogenous groups were listed according to their importance to the fitted tree (Table 2). The variable importance list was determined in a more complex way than fitting the regression tree while partitioning into a homogeneous group. To obtain the importance of a variable in the regression tree, total goodness of split measures were used and scaled up to 100 and rounded to omit decimals

for all variables. For rounding issue, the total is a little bit more than 100. Variable importance values less than one are usually ignored. We found that the three most important variables were C3, C1 and mean daily temperature climatic variable (MDT). However, in the variable importance list, we found C1 as an important variable, but on the optimal tree, this variable did not split any nodes because in our fitted regression tree, C1 variable appeared as a surrogate variable. In the absence of splitting variable in a node to predict the actual split, surrogate variables are accounted in the variable importance plot [15, 16]. As seen in the literature, it is possible that this surrogate variable may contribute a larger portion in the variable importance list, but in the optimal fitted regression tree, that surrogate variable may not split any node [17]. As such, the other two variables C3 and MDT were used as top nodes in the fitted regression tree. The fact that MDT
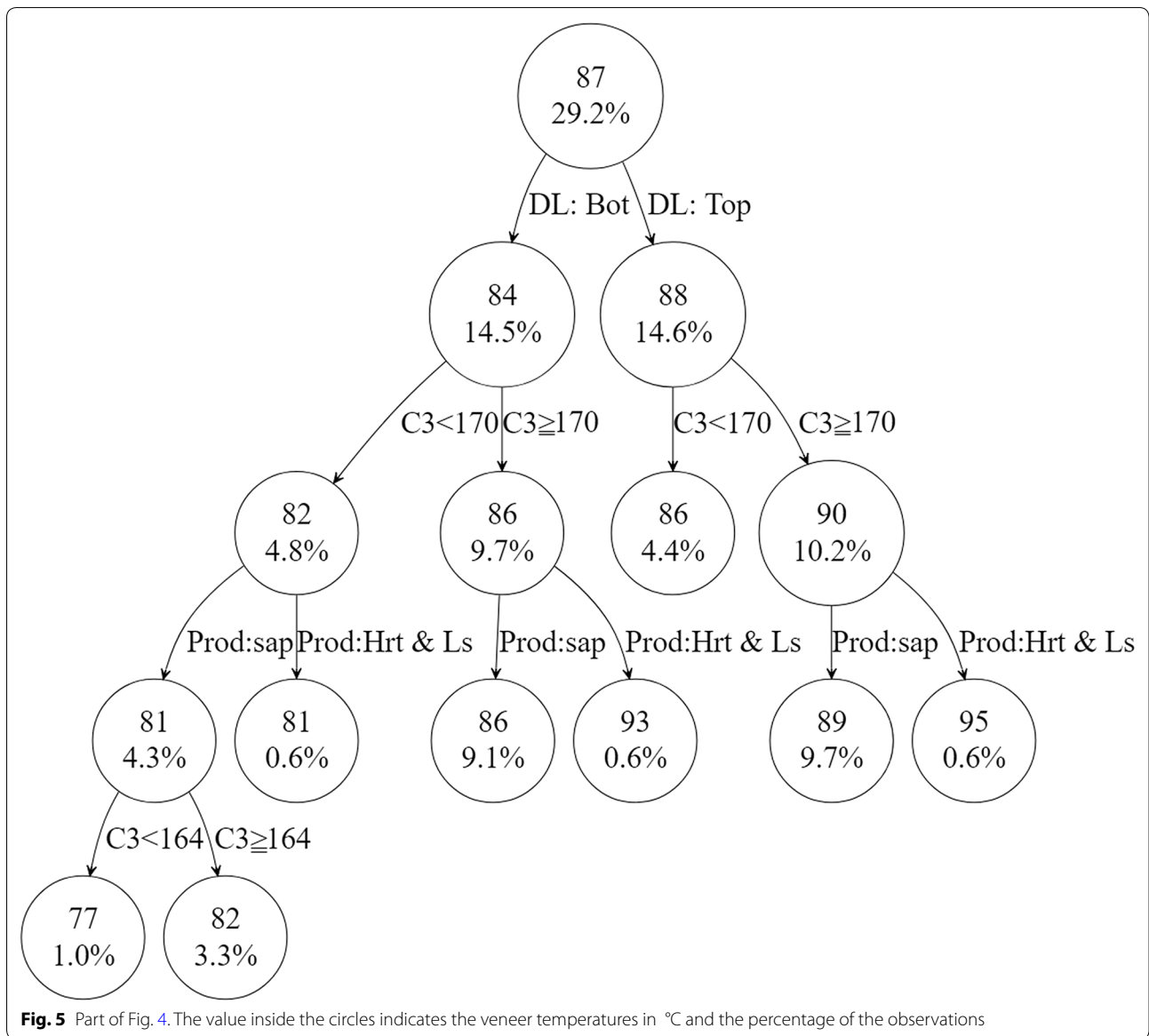

Fig. 4 Fitted optimal regression tree of temperature of veneer. The value inside the circles indicates the veneer temperatures in °C and the percentage of the observations

Ahmed *et al. J Wood Sci*   (2020) 66:54

Page 6 of 8

Fig. 5 Part of Fig. 4. The value inside the circles indicates the veneer temperatures in °C and the percentage of the observations

is one of the top nodes indicates the process parameters need to be adjusted based on the previous week's climate. Although this was expected, it was not foreseen that MDT would be the second most important variable. From the optimal fitted tree, it was concluded that the veneer temperature also depends on the dryer position (Figs. 4, 5 and 6). Specifically, East and Middle positions provided similar outcomes, whereas the West position yielded warmer veneer temperatures. In similar weather conditions, the top deck levels resulted in warmer veneers when compared to the bottom deck levels (Figs. 4, 5 and 6). This finding implies that sorting the raw material prior to drying could result in a more uniform veneer temperature and potentially

product quality. This information is confirmed by the fact that the heartwood and light sapwood types were on average − 12.22 °C warmer than the sapwood wood type (Figs. 4, 5 and 6). Interestingly, veneers processed through the West position when C3 temperature was kept greater than 165 °C seemed to minimize differences between the top and bottom dryer levels.

Based on the developed optimal tree, we can identify all important threshold points of predictor space and evaluate the effects of process parameter settings, dryer levels, positions and climatic variables on the temperature of veneer sheets. Based on the findings and threshold values of important predictor variables, it is possible to get some idea of the final temperature
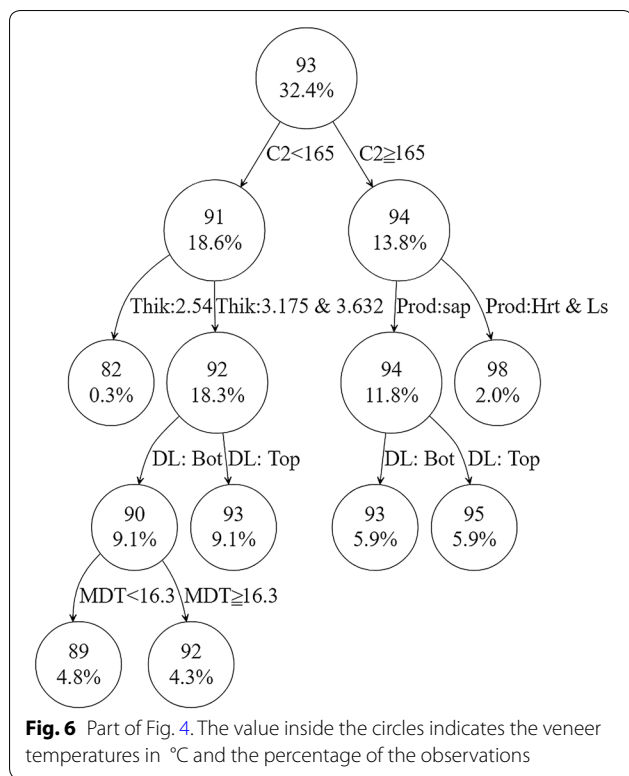
Ahmed *et al. J Wood Sci*   (2020) 66:54

Page 7 of 8

Fig. 6 Part of Fig. 4. The value inside the circles indicates the veneer temperatures in °C and the percentage of the observations

Table 2 Variable importance in the fitted regression tree of temperature of veneers

Variable	Importance (out of 100)
C3	19
C1	11
MDT	9
Temperature in Z1b	9
Temperature in Z3	9
Temperature in Z1a	9
Temperature in Z1c	8
TWP	6
DP	4
C2	4
Temperature in Z2a	3
DL	3
Temperature in Z2b	3
Prod	2
Thik	1
Drying time	1

of the veneer sheet while exiting the dryer. However, regression trees do not have a similar predictive ability as the classical predictive models [11]. In the future analysis, we aimed to use the knowledge gained from the regression tree approach in this study to develop a predictive model using a tree-based approach.

In our work, the previous week's climatic variables played an important role in the drying process. In other facilities, if a nearby weather station is not available, then measurements of climatic variables can be interpolated using the inverse distance weighting (IDW) of a few weather station data. Although this technique is commonly used for predicting tree growth in remote areas, it will provide an approximate estimation. Alternatively, the climatic variables (i.e., daily temperature, humidity, etc.) could be measured at the facility to control veneer temperature in the drying process.

Conclusions

The developed optimum regression provided valuable insight into the drying process and allowed us to deepen our knowledge and understanding of the science governing veneer drying. The regression tree model was validated using real industrial data as well as the expertise from dryer operators. From the regression tree approach and findings, we found the most important variables and their ranges to achieve the best possible range of final temperature of veneer. From this study, we found that the final temperature of veneer was profoundly affected by the chain side temperatures and climatic variables. To obtain the best temperatures of veneer, we have to consider previous week's climatic variables. If a climatic variable is ignored, chain and zone sides' temperature should be adjusted accordingly.

Ahmed *et al. J Wood Sci*   (2020) 66:54

Page 8 of 8

Author details
[1] Department of Forest Resources Management, The University of British Columbia, 2045-2424 Main Mall, Vancouver, BC V6T 1Z4, Canada. [2] Department of Wood Science, The University of British Columbia, Vancouver, Canada. [3] School of Population and Public Health, The University of British Columbia, Vancouver, Canada. [4] Centre for Health Evaluation and Outcome Sciences, Vancouver, BC, Canada.

References
1. Thant AA, Yee SS, Htike TT (2009) Modeling drying time during veneer drying and comparison with experimental study. In: Proceedings of the international multiconference of engineers and computer scientists, Hong Kong, 2009
2. Rippy RC, Wagner FG, Gorman TM, Layton HD, Bodenheimer T (2000) Stress-wave analysis of Douglas-fir logs for veneer properties. For Prod J 50(4):49–52
3. Zhang SY, Yu Q, Beaulieu J (2004) Genetic variation in veneer quality and its correlation to growth in white spruce. Can J For Res 34(6):1311–1318
4. Vikram V, Cherry ML, Briggs D, Cress DW, Evans R, Howe GT (2011) Stiffness of Douglas-fir lumber: effects of wood properties and genetics. Can J For Res 41(6):1160–1173
5. Lutz JF (1974) Drying veneer to a controlled final moisture content by hot pressing and steaming, USDA, No. FSRP-FPL-227, Forest Service, Forest Products Laboratory, Madison. Wisconsin, USA
6. Aydin I, Colakoglu G, Colak S, Demirkir C (2006) Effects of moisture content on formaldehyde emission and mechanical properties of plywood. Build Environ 41:1311–1316
7. Rokach L, Maimon O (2008) Data mining with decision trees: theory and applications. World Scientific Publishing, New Jersey
8. Matloff N (2017) Statistical regression and classification: from linear models to machine learning. Chapman & Hall/CRC, London
9. Moore DE, Lees BG, Davey SM (1991) A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. Environ Manage 15(1):59–71
10. Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9(2):181–199
11. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York
12. Environment and Natural Resources of Canada database. https://clima te.weather.gc.ca/historical_data/search_historic_data_e.html. Accessed 22 Mar 2018
13. Moisen GG, Freeman EA, Blackard JA, Frescino TS, Zimmermann NE, Edwards TC (2006) Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecol Model 199(2):176–187
14. R version 3.5.0. R Core Team (2018), A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. software available from http://www.R-project.org/. Accessed 20 June 2020
15. Hastie T, Tibshirani R, Friedman J (2009) An introduction to statistical learning. In: Hastie T, Tibshirani R, Friedman J (eds) Tree-based methods. Springer, New York
16. Rey T, Kordon A, Wells C (2012) Applied data mining for forecasting using SAS. In: Rey T, Kordon A, Wells C (eds) A Practitioner's guide of DMM methods for forecasting. SAS Institute, North Carolina
17. Wu X, Kumar V (2009) The top ten algorithms in data mining. In: Wu X, Kumar V (eds) Cart: classification and regression trees. Chapman and Hall/CRC, New York