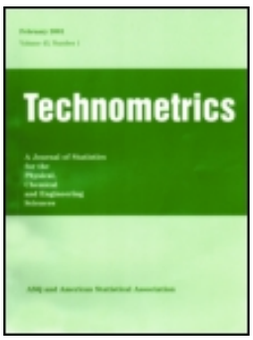

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: http://www.tandfonline.com/loi/utch20

Quantifying Uncertainty in Lumber Grading and Strength Prediction: A Bayesian Approach

Samuel W.K. Wong, Conroy Lum, Lang Wu & James V. Zidek

To cite this article: Samuel W.K. Wong, Conroy Lum, Lang Wu & James V. Zidek (2016) Quantifying Uncertainty in Lumber Grading and Strength Prediction: A Bayesian Approach, Technometrics, 58:2, 236-243, DOI: 10.1080/00401706.2015.1033108

To link to this article: https://doi.org/10.1080/00401706.2015.1033108

Supplementary materials for this article are available online. Please go to http://www.tandfonline.com/r/TECH

Quantifying Uncertainty in Lumber Grading and Strength Prediction: A Bayesian Approach

Samuel W. K. Wong

Department of Statistics
University of Florida
Gainesville, FL
(*swkwong@stat.ufl.edu*)

Conroy Lum

FPInnovations
Vancouver, BC, Canada
(*Conroy.Lum@fpinnovations.ca*)

Lang Wu and James V. Zidek

Department of Statistics
University of British Columbia
Vancouver, BC, Canada
(*lang@stat.ubc.ca; jim@stat.ubc.ca*)

This article presents a joint distribution for the strength of a randomly selected piece of structural lumber and its observable characteristics. In the process of lumber strength testing, these characteristics are ascertained under strict grading protocols, as they have the potential to be strength reducing. However, for practical reasons, only a few such selected characteristics among the many present, are recorded. We present a data-generating mechanism that reflects the uncertainties resulting from the grading protocol. A Bayesian approach is then adopted for model fitting and construction of a predictive distribution for strength that accounts for the unrecorded characteristics. The method is validated on simulated examples, and then applied on a sample of specimens tested for bending and tensile strength. Use of the predictive distribution is demonstrated, and insights gained into the grading process are described. Details of the lumber testing experiments can be found in the online supplementary materials.

KEY WORDS: Bayesian predictive model; Sawn structural lumber; Structural lumber testing; Visual stress grading; Wood strength-reducing characteristics.

1. INTRODUCTION

Concern about the effect of climate change on the growth of trees (Andalo, Beaulieu, and Bousquet 2005) combined with technologically induced changes in the way they are grown and processed to make lumber, has led to the establishment of long-term monitoring programs (Kretschmann, Evans, and Brown 1999). These concerns and changes have in turn led a renewed interest in ways of assessing lumber properties, including innovative analytical approaches that exploit modern statistical theory.

A property of great importance is lumber strength since construction is a primary use of this product. It must be strong enough to meet future demands in the form of both dynamic and static loadings. However, strength is highly variable. So a system for classifying lumber into grades has been developed to reduce that variability within grade classes; design values are then set on the basis of the in-grade conditional distributions. Consumers select a grade of lumber appropriate for their anticipated loadings.

Grading is done in accordance with grading rules, which involve characteristics that can be determined without destructive testing, especially ones that relate to strength. This process, which yields the requisite conditional distributions and hence design values, has worked well and withstood the test of time. The possible long-term effects of changes in climate and lumber production technology on the strength of lumber, however, anticipate a future need to modify the grading rules to preserve the design values. Thus, we seek to leverage modern statistical methods and computational power for constructing the conditional strength distributions based on strength determining characteristics, and for evaluating the efficacy of current grading rules.

The task is to predict lumber strength based on its recorded characteristics from the grading process. There are a few broad types of characteristics; knots, for example, are the most commonly occurring type of characteristic. A piece of lumber often has multiple characteristics, and the difficulty in our context is that not all characteristics present are recorded; most are nonrandomly censored, which distinguishes this work from classical regression and missing-data problems. We must rely on records that have much missing information, while accounting for the processes described in the next section that generate the experimental data. Together these factors merit a Bayesian approach (Section 3) that ensures a coherent hierarchical framework for linking the elements of the process while reflecting that uncertainty. The next section describes the relevant basic features of our problem and thereby lays the groundwork for the remainder of the article. Using the framework, we can quantify the uncertainty involved in generating the data on the characteristics, and study how the characteristics affect the two major

TECHNOMETRICS, MAY 2016, VOL. 58, NO. 2
DOI: 10.1080/00401706.2015.1033108

strength properties, namely the "modulus of rupture" (MOR) and the "ultimate tensile strength" (UTS). To that latter end, a predictive distribution for strength can be built from the fitted model and we illustrate its application to the prediction of strengths for future pieces of lumber.

The article is laid out as follows. Section 2 describes the methods prescribed for testing the strength of lumber. Section 3 describes the data-generating mechanism and proposes a Bayesian approach for handling the censored characteristics and constructing the distribution of interest. That model is validated through a series of simulation examples in Section 4. The method is then applied in Section 5 on an experimental dataset generated in a wood products testing laboratory. Section 6 gives our concluding remarks.

2. ASSESSING THE STRENGTH OF LUMBER

Grading assesses the features of each piece of lumber (specimen) that are likely to affect its strength or utility. "Knots" (formed when branches or limbs are incorporated in a piece of lumber), "shake" (a lengthwise separation of the wood), and "slope of grain" (the deviation of the line of fibers from a line parallel to the sides of the piece) are examples of characteristics that often, but not always, limit the strength of a piece of lumber when tested to failure.

Destructive strength tests are also carried out on a limited but representative sample of lumber for calibration purposes on a chosen population. Such a population might consist of all pieces currently available of a specified grade, size, and species. In a destructive test, the piece of lumber is subjected to an increasing load until it fails. The maximum stress reached immediately before failure is called its "strength."

We now outline the protocol used for carrying out a laboratory experiment on a chosen population with the aid of a human grader, to obtain such calibration data. The steps of testing each piece of lumber proceed in the following order:

1. *Grader confirms grade (piece is admissible)*. Wood processed at a mill is sorted into grades before being bundled for sale. This proceeds in accordance with particular standards involving (see, e.g., Green, Ross, and McDonald 1994) "grade controlling characteristics." Some characteristics, which may be purely cosmetic in nature, are included because they affect the commercial value of the piece. Other characteristics are also controlled for, as they are deemed to be strength reducing. In many modern mills, automated technology is used to classify sawn lumber into "grades." Therefore, in a bundle to be used for calibration purposes, a human grader first examines each piece to confirm that it indeed belongs in the population of interest. Specimens found to have been misclassified are marked as "off-grade" and excluded.
2. *Grader selects and records M*. The visual characteristic thought most likely to cause a piece to fail under a destructive test is called its "maximum strength-reducing characteristic" (MSRC, or M for short). Practical considerations have meant that M is recorded in a coded form for each piece. The record of M includes the category of the characteristic (e.g., knot, shake; see Table 4 in Section 5 for the full list of codes used by our grader), as well as a description. It is not unusual for failure to occur not at M, but at some other characteristic, as wood is highly unpredictable. Traditionally, M is visually selected by a human grader who has been trained to meet industry standards that "a maximum of 5% below grade as an allowable variation between agency qualified graders" can be maintained (NIST 2010, p. 10). More recently, machine vision or automated visual grading is being used to improve the efficiency of selecting M, designed to apply the same selection rules used by human graders.
3. *Measure and record MOE*. Methods have been developed to assess strength without destructive testing. The most widely used characteristic for this purpose is the "modulus of elasticity" (MOE), which can be measured in various ways, each of which quantifies the stiffness of a piece of lumber (Ross et al. 1991). A simple model of lumber strength could use MOE as a linear predictor of MOR (Kretschmann, Evans, and Brown 1999), as the two tend to be positively correlated. For the data in this study, MOE measurements are obtained by the transverse vibration method (Pellerin 1965). Measurements of MOE are typically given in millions of pounds per square inch (psi).
4. *Perform destructive test and record breaking strength*. Two common destructive tests used in the industry are the "modulus of rupture" (MOR, or R for short), and "ultimate tensile strength" (UTS, or T for short). R is found by bending a specimen until it breaks, while T is found by stretching the piece longitudinally until it separates. The test to be performed on the specimen will be chosen in advance; in this step, the specimen is broken according to the chosen test and its breaking strength is recorded. Measurements of R and T are typically given in thousands psi.
5. *Grader examines failure and records C*. The characteristic at which the specimen broke is called the cause of failure (C for short). The grader examines the broken specimen and records C in coded form. The possible categories of C come from the same list as M. Ideally, the C will be the same characteristic as M, but it is not unusual for M and C to turn out being different characteristics. This shows the inherent difficulty of correctly identifying M beforehand, and this is one aspect we wish to quantify with the modeling approach that follows.

The use of such visual and physical characteristics to predict breaking strengths of individual pieces of lumber has been a much-studied subject. For example, Taylor et al. (1992) considered tensile strength models that treat a piece of lumber as consisting of a set of smaller contiguous segments. Lei, Zhang, and Jiang (2005) used a regression approach to predict MOR based on tree and stand characteristics. Divos and Tanaka (1997) also used a multiple regression approach to predict both bending and tensile strength. Significant regressors found by that study were MOE and modified knot diameter ratio, and it was found that both machine stress grading and appropriate visual grading are important for lumber. However, we have not found

previous studies that attempt to coherently capture the entire grading process and the uncertainties involved.

3. FRAMEWORK FOR MODELING STRENGTH

3.1 A Data-Generating Mechanism

The simplest model for lumber strength relates a destructive strength property Y to the MOE v according to the linear regression model (Kretschmann, Evans, and Brown 1999)

$$Y = \beta_0 + \beta_1 v + \epsilon',$$

for regression coefficients β_0, β_1 and normal error ϵ'. This section extends this model in a way that accounts for the additional information recorded in the maximum strength-reducing characteristic M and the cause of failure C, and captures a meaningful description of the underlying process that generated the data.

Of interest are the visual characteristics deemed to potentially cause the failure in a destructive test. In this study, we work with three major categories of such visual characteristics: knots (k), shakes and grain deviations (s), other (o). The number of significant knots on a lumber specimen is random; some pieces have many large knots, while others have none at all. A plausible model for the number of significant knots N_k would be a Poisson distribution. Some specimens have a long crack (shake) or a major grain deviation, while others do not. We shall indicate the presence of shake by $N_s = 1$, and absence by $N_s = 0$. Finally, for simplicity we always allow one characteristic in the category "other" to act as a baseline and cover all the remaining varieties of miscellaneous defects that are neither knots nor shake, and which may not be fully evident until the piece is broken.

Considering an individual specimen, it has a total $J = N_k + N_s + 1$ independent characteristics, where

$$N_k \sim \text{Pois}(\lambda_k)$$
$$N_s \sim \text{Bern}(p_s)$$

are independent. Each of the J characteristics has a latent effect on the strength, listed in the vector

$$\mathbf{X^o} = [X_1^o, X_2^o, \ldots, X_J^o].$$

Each characteristic in $\mathbf{X^o}$ belongs to one of the three major categories, as indicated in the following vector corresponding to $\mathbf{X^o}$,

$$\mathbf{t} = [t_1, t_2, \ldots, t_J].$$

For instance, a specimen with four total characteristics—composed of two knots, the presence of shake, and other—would have $\mathbf{t} = [k, k, s, o]$. The distribution of a latent effect will depend on the category of the characteristic, which we model using independent Normals for $i = 1, \ldots, J$,

$$X_i^o | t_i = k \sim N\left(\mu_k, \sigma_k^2\right) \quad (1)$$
$$X_i^o | t_i = s \sim N\left(\mu_s, \sigma_s^2\right) \quad (2)$$
$$X_i^o | t_i = o \sim N\left(\mu_o, \sigma_o^2\right). \quad (3)$$

Only the most severe one of the J characteristics will be the cause of failure, namely $C = \left\{i : X_i^o = \max(\mathbf{X^o})\right\}$. Therefore, we postulate that a regression model including $\mathbf{X^o}$ as a predictor will only depend on X_C^o, namely

$$Y = \beta_0 + \beta_1 v + \gamma X_C^o + \epsilon, \quad (4)$$

where the coefficient γ acts as a scaling factor for relating the latent effect to a corresponding change in the strength Y. The remaining error term $\epsilon \sim N(0, \sigma_e^2)$ is small and intends to capture measurement uncertainty in Y. Note that it is not possible to fit the model (4) directly, as the vector $\mathbf{X^o}$ of latent effects is not observable.

A grader examining the specimen cannot know the true values $\mathbf{X^o}$, but instead observes the noise-contaminated vector of subjective assessments of effects

$$\mathbf{X} = [X_1, X_2, \ldots, X_J],$$

where $X_i | X_i^o, \mathbf{b} \sim N(X_i^o + b_{t_i}, \sigma_x^2)$ are independent for $i = 1, \ldots, J$ and $\mathbf{b} = [b_k, b_s, b_o]$. For an unbiased grader, we would have $\mathbf{b} = \mathbf{0}$, while a nonzero $\mathbf{b}$ can describe scenarios where the grader has a systematic bias in over/underestimating the effects of certain categories. If grader bias is suspected, we set $b_k = 0$ as a baseline and include b_s, b_o as parameters to be estimated.

Prior to destructive testing, the grader selects the MSRC based on $\mathbf{X}$, namely $M = \{i : X_i = \max(\mathbf{X})\}$. After destructive testing, the grader visually determines the characteristic at which the specimen failed and thusly identifies C, albeit without knowing the value of X_C^o. The frequency with which M and C agree depends on the size of σ_x^2; smaller values of σ_x^2 correspond to a larger probability that $M = C$ for an unbiased grader.

Recall that grading rules stipulate that a coded category and description of the M and C be recorded by the grader. These must be converted to a quantitative measurement for modeling purposes. The coded description of a knot allows the grader's subjective assessment of its effect to be computed based on its size and location. We use these calculated quantities as values for elements of $\mathbf{X}$ in the case of knots, which is the most common characteristic. For shake and other, not enough detail is available to calculate a corresponding effect, and hence their corresponding elements in $\mathbf{X}$ will be treated as missing.

In summary, the data for a single test specimen consist of the following. The values $[y, v, m, c]$ are always observed. In addition, if $t_m = k$, then we observe the value x_m. Likewise, if $t_c = k$, we observe x_c. The remaining elements of $\mathbf{x}$ are missing. Furthermore, the entire vector $\mathbf{x^o}$ is missing, and the number of characteristics $j \geq 1$ is missing.

3.2 Bayesian Inference

The parameters to be inferred are $\boldsymbol{\theta} = [\beta_0, \beta_1, \gamma, \mu_k, \sigma_k^2, \mu_s, \sigma_s^2, \mu_o, \sigma_o^2, \sigma_x^2, b_s, b_o]$ from a sample of independent test specimens. Expert knowledge provides values of λ_k, p_s, σ_e^2, which we assume to be known constants throughout the analysis. We adopt a Bayesian approach to handle this inference problem, with emphasis on the posterior predictive distribution of Y based on v, M, X_M that is obtained by integrating over the missing characteristics and the posterior distribution of the parameters.

The likelihood of a single test specimen based on a complete data vector $\mathbf{x}$ is

$$p(y, m, c, \mathbf{x}|v, \boldsymbol{\theta}) \propto \sum_{n_k}\sum_{n_s} p(n_s)p(n_k) \times p\left(m, c|\mu_k, \sigma_k^2, \mu_s, \sigma_s^2, \mu_o, \sigma_o^2, \sigma_x^2, n_k, n_s, \mathbf{t}, \mathbf{b}\right) \quad (5)$$

$$\times \int p\left(y|x_c^o, c, v, \beta_0, \beta_1, \gamma\right) p\left(x_c^o, \mathbf{x}|m, c, \mu_k, \sigma_k^2, \mu_s, \sigma_s^2, \mu_o, \sigma_o^2, \sigma_x^2, n_k, n_s, \mathbf{t}, \mathbf{b}\right) dx_c^o. \quad (6)$$

Letting $Z_i \sim N(b_{t_i}, \sigma_x^2)$, iid for $i = 1, \ldots, J$, we can represent X_i by $X_i = X_i^o + Z_i$ with all the X_i and Z_i independent. Then the events $M = m$ and $C = c$ can be expressed equivalently in terms of the X_i's and Z_i's, as

$$\{M = m\} = \bigcap_{i \neq m}\{X_i < X_m\},$$

$$\{C = c\} = \bigcap_{i \neq c}\{X_i - Z_i < X_c - Z_c\}.$$

With this representation, it follows that the term $p(m, c|\boldsymbol{\theta}, n_k, n_s)$ can be obtained from the CDF of the multivariate normal distribution.

Next, for computational purposes we expand the following joint probability,

$$p\left(x_c^o, \mathbf{x}, m, c|\boldsymbol{\theta}, n_k, n_s, \mathbf{t}\right) = P\left(M = m, C = c|x_c^o, \mathbf{x}, \boldsymbol{\theta}, n_k, n_s\right)$$

$$p\left(x_c^o|x_c, \boldsymbol{\theta}\right) \prod_{i=1}^{n_k+n_s+1} p\left(x_i|t_i, \boldsymbol{\theta}\right) = \prod_{i \neq m} I(x_m > x_i) \prod_{i \neq c}$$

$$P\left(Z_i > x_i - x_c^o\right) \times p\left(x_c^o|x_c, \sigma_x^2\right) \prod_{i=1}^{n_k+n_s+1} p(x_i|t_i, \boldsymbol{\theta}). \quad (7)$$

Any missing elements in $\mathbf{x}$ for each specimen must be integrated out in the likelihood computation.

Priors $\pi(\theta)$ are required to complete the specification of the posterior. While further expert knowledge concerning the effects of characteristics could be infused at this point, in our analysis we assume independent flat priors on the regression parameters as well as the μ's, and independent noninformative Inv-Gamma(0.001,0.001) priors on the σ^2's. If grader bias is to be estimated, we give b_s and b_o Normal prior distributions with mean zero.

Denote the observed data from n specimens by D, and the observed and missing parts of $\mathbf{x}$ for an individual specimen by $\mathbf{x}_{\text{obs}}$ and $\mathbf{x}_{\text{mis}}$, respectively. Then, after obtaining the posterior distribution $\pi(\boldsymbol{\theta}|D) = \pi(\boldsymbol{\theta}) \prod_{l=1}^{n} p(y_l, m_l, c_l, \mathbf{x}_{l,\text{obs}}|v_l, \boldsymbol{\theta})$, this framework readily provides the strength predictive distribution of interest for a future piece of lumber. Its quantities v_f, m_f, x_{m_f} can be recorded without destructive testing, and using the likelihood function as expanded in (5) we have

$$p(y|v_f, m_f, x_{m_f}, D) \propto \sum_{n_k}\sum_{n_s}\sum_{c=1}^{n_k+n_s+1} \iint p(y, m_f, c, \mathbf{x}, n_k, n_s|v_f, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|D)\, d\mathbf{x}_{\text{mis}}\, d\boldsymbol{\theta}. \quad (8)$$

In cases where $t_{m_f} \in \{s, o\}$, then we must also integrate over the missing x_{m_f}.

3.3 Computational Methods

Computation of the likelihood of each specimen requires integration over a multidimensional vector consisting of x_c^o and the missing parts of $\mathbf{x}$, for each potential combination of values n_k and n_s. To achieve usable computational speeds, a Monte Carlo integration procedure was used based on the expansion (7). Also, the maximum number of knots n_k is truncated beyond six, which is a reasonable approximation when $\lambda_k \leq 3$. As a concrete illustration, consider the likelihood evaluation for $n_k = 1, n_s = 1$ where $m = 1, c = 2, \mathbf{t} = [k, s, o]$. In this case, x_c^o and $[x_2, x_3]$ are missing. The likelihood (suppressing the parameters $\boldsymbol{\theta}$ for simplicity) is

$$p(y, m = 1, c = 2, x_1|v, \mathbf{t}) \propto \iiint p\left(y|x_c^o, c, v\right) p(x_c^o, x_1, x_2, x_3, m, c|\mathbf{t})\, dx_2^o\, dx_2\, dx_3 = p(x_1|t_1) P\left(X_2 < x_1, X_3 < x_1\right) E_{X_2^o, U_2, U_3}\left[p\left(y|x_2^o, v\right) P\left(Z_1 > x_1 - x_2^o\right) P\left(Z_3 > x_3 - x_2^o\right)\right],$$

where

$$\begin{pmatrix} X_2^o \\ U_2 \\ U_3 \end{pmatrix} \sim \text{Trunc} - \text{MVN}\left[\begin{pmatrix} \mu_s \\ \mu_s + b_s \\ \mu_o + b_o \end{pmatrix}, \begin{pmatrix} \sigma_s^2 & \sigma_x^2 & 0 \\ \sigma_x^2 & \sigma_s^2 + \sigma_x^2 & 0 \\ 0 & 0 & \sigma_o^2 + \sigma_x^2 \end{pmatrix}\right] : U_2 < x_1, U_3 < x_1.$$

Thus, the expectation is approximated by drawing samples of $[X_2^o, U_2, U_3]$ and averaging the values of the function within. Sample draws from the truncated multivariate normal are obtained via the R package tmvtnorm (Wilhelm and Manjunath 2014); drawing from the truncated distribution is a natural importance weight adjustment for handling the indicator functions $I(x_2 < x_1)I(x_3 < x_1)$ from (7) and increasing the efficiency of the Monte Carlo estimate. It was found empirically that 2000 draws is sufficient to reliably compute the total log-likelihood over $n = 200$ specimens. The likelihood computation for other values of n_k, n_s and other patterns of missingness in $\mathbf{x}$ proceeds in an analogous fashion.

The parameter vector $\boldsymbol{\theta}$ has 12 dimensions, so a naive Markov chain Monte Carlo (MCMC) Metropolis sampling algorithm would be very inefficient. We first applied Nelder-Mead iterations on the posterior, from a set of crude parameter values. This yields a set of local modes from which to use as starting points for MCMC exploration. To improve the speed and reliability of subsequent convergence, we applied parallel tempering (Swendsen and Wang 1986) over 15 computing nodes for a range of temperatures geometrically spaced from 1 to 50. Metropolis-Hastings draws are used within the individual chains, with proposals adapted to the temperature of the chain. With this setup, the computational time required for 1000 posterior draws for a specimen size $n = 200$ is about 2 hr.

To obtain the posterior predictive distribution, $p(y|v, m_f, x_{m_f})$ is evaluated according to (8) on a grid of y values using the same techniques as in the likelihood computation. The approximate density at each value y is obtained by averaging (8) for MCMC samples that had been drawn from $\pi(\boldsymbol{\theta}|D)$, after a suitable burn-in.

Table 1. Frequencies of M and C under the different simulation scenarios for a sample of $n = 500$ specimens

Scenarios 1 and 2				Scenario 3			
$M\backslash C$	k	s	o	$M\backslash C$	k	s	o
k	265	8	12	k	288	10	27
k'	114	—	—	k'	132	—	—
s	10	13	3	s	2	8	3
o	25	1	49	o	0	1	29

4. SIMULATION STUDIES

To explore the inference and prediction procedures in a controlled setting, we simulated data under three different scenarios. The parameter values were chosen to roughly mimic the proportions of (k, s, o) in the real dataset, while expert knowledge provides the values $\sigma_e^2 = 0.04$, $\lambda_k = 3$, $p_s = 0.28$.

Parameters:

- $\beta_0 = 6, \beta_1 = 4, \gamma = -0.25, \mu_k = 25, \sigma_k = 5, \mu_s = 20, \sigma_s = 6, \mu_o = 15, \sigma_o = 9, \sigma_x = 5.$
- Sample size $n = 500$.

To produce a simulated dataset according to given parameters, for each specimen we apply the following steps:

1. Draw $n_k \sim \text{Pois}(\lambda_k)$, $n_s \sim \text{Bern}(p_s)$ to obtain $\mathbf{t}$.
2. Draw $\mathbf{x^o}$ according to (1). Set $c = \{i : x_i^o = \max(\mathbf{x^o})\}$.
3. Draw $v \sim N(1.6, 0.2^2)$, to mimic the distribution of MOE in our dataset.
4. Draw $y \sim N(\beta_0 + \beta_1 v + \gamma x_c^o, \sigma_e^2)$.
5. Draw $\mathbf{x}$ and set $m = \{i : x_i = \max(\mathbf{x})\}$. Keep x_m if $t_m = k$. Keep x_c if $t_c = k$.

To study the properties of the method, the following scenarios were considered:

1. n_k and n_s known, and no grader bias.
2. n_k and n_s unknown, and no grader bias.
3. n_k and n_s unknown, and data generated with $b_s = b_o = -10$. We then set fairly diffuse priors $N(0, 10^2)$ for b_s and b_o.

Table 3. Mean-squared prediction errors for fitted models under the different scenarios, based on a simulated test set of 100 pieces

Scenario 1	Scenario 2	Scenario 3
1.08	1.17	1.24

The same dataset is used for Scenarios 1 and 2; the difference is in the model-fitting treatment of n_k and n_s. For the different scenarios, the counts of M and C are tabulated and shown in Table 1. Since there can be more than one knot, there are two cases to distinguish when $t_m = t_c = k$. The cell (k, k) indicates that M and C were the same knot, while (k', k) indicates that M and C turned out to be different knots. Scenario 3 is intended to mimic a pattern seen in the real data. The grader has a stronger preference for selecting knots as M, leading to asymmetry between the cells (k, o) and (o, k) for example.

Models were fitted to the three scenarios using the techniques described in the previous section. A summary of the posterior distributions of the parameters is shown in Table 2. The fits appear acceptable. The estimates relating to shake and other in Scenario 3 have comparatively more uncertainty, as grader bias results in relatively fewer pieces having those characteristics recorded; these must also be estimated together with grader biases, nonetheless fairly reasonable estimates were obtained.

Next, additional test sets of 100 pieces were generated to study the performance of the posterior predictive distribution $p(y|v, m, x_m)$ under the scenarios. The summaries of the mean-squared prediction errors (MSPEs) based on the posterior predictive modes are listed in Table 3. The difference between Scenarios 1 and 2 shows that the number of characteristics being considered J, if always recorded, has a role in improving predictive performance. In Scenario 3, grader bias skews the M that is recorded for new data, and increases prediction error slightly although bias terms have been fitted.

To quantify how much prediction uncertainty is due to parameter estimation, in Scenario 2 we compared the posterior predictive $p(y|v, m, x_m)$ and $p(y|v, m, x_m, \boldsymbol{\theta})$, where the latter case refers to $\boldsymbol{\theta}$ set at the true values. We found that using the latter for prediction decreases the MSPE for Scenario 2 from 1.17 to 1.06. To illustrate this graphically, Figure 1 plots

Table 2. Summary of the posterior distributions of the parameters under simulation scenarios

	Simulation 1 Posterior quantiles			Simulation 2 Posterior quantiles			Simulation 3 Posterior quantiles		
	50%	2.50%	97.50%	50%	2.50%	97.50%	50%	2.50%	97.50%
β_0	5.3	4.6	6.2	5.9	5.1	6.5	5.6	4.5	6.6
β_1	4.0	3.7	4.4	4.1	3.8	4.3	3.7	3.3	4.2
μ_k	25.8	25.2	26.5	28.3	27.6	29.0	27.7	27.1	28.4
σ_k	5.9	5.4	6.6	4.7	4.2	5.3	5.0	4.3	5.8
μ_s	19.2	15.9	21.2	18.6	15.9	22.3	17.9	12.7	22.4
σ_s	7.4	6.1	12.8	7.5	4.9	10.9	7.9	4.9	16.0
μ_o	14.7	12.1	16.9	14.1	10.3	17.3	9.7	4.7	14.0
σ_o	9.8	8.4	11.5	10.3	9.0	13.1	11.7	9.2	16.0
γ	−0.22	−0.24	−0.21	−0.24	−0.26	−0.22	−0.21	−0.24	−0.19
σ_x	4.7	4.5	5.0	5.0	4.7	5.3	5.0	4.7	5.5
b_s	—	—	—	—	—	—	−4.0	−9.1	0.4
b_o	—	—	—	—	—	—	−10.3	−17.6	−5.9

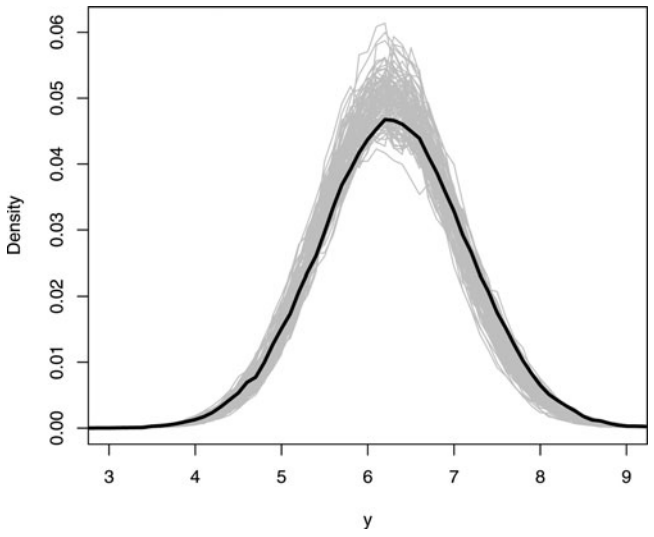

Figure 1. Comparison of estimated predictive density for MCMC draws from $\pi(\boldsymbol{\theta}|D)$ (gray) compared to $\boldsymbol{\theta}$ at true values (black).

predictive density functions of a specimen computed for each of 100 posterior draws from $\pi(\boldsymbol{\theta}|D)$ in gray and some spread is evident. The black curve shows the density based on the true values of $\boldsymbol{\theta}$.

5. DATA ANALYSIS

5.1 Lumber Strength Dataset

We now illustrate the methodology proposed in this article on a real data example, obtained from a set of destructive experiments run in a wood products testing laboratory. The chosen population of interest for this study was 12-ft 1650f-1.5E Spruce-Pine-Fir (SPF) 2x4, with species composition as listed in appendix A of NIST (2010) and listed in the online supplementary materials for our sample bundles. This population is an example of machine-stress-rated lumber, which controls for the expected variability in MOE and R within the grade. Members of this population also satisfy certain restrictions on allowable visual characteristics, such as maximum allowable knot size and maximum length of shake. In other words, the grade of this population has been assigned based on machine-stress rating together with visual screening.

A sample of 710 pieces from this population was destructively tested, to collect data on the R (496 pieces) and T (214 pieces) strength properties. The steps involved in the preparation and testing of these samples are described in the supplementary materials available online, along with commentary on some examples of tested pieces. The M and C for these samples were coded by one human grader, following a system developed by a wood products firm, which lists 37 possible codes for characteristics. To test the effectiveness of the predictive model, the last set of R measurements (116 samples) and the same proportion of T measurements (50 samples) were not used for model fitting.

Figure 2 shows the histograms of MOE (in millions psi), R (in thousands psi), and T (in thousands psi). Table 4 shows the frequency distribution of M and C over the raw codes, tabulated separately for R and T. The final column shows the category (k, s, o) that each raw code was assigned to for this analysis. Table 5 shows the frequencies of M and C after collapsing the raw codes into the three categories. The cell (k, k) indicates cases where the grader-selected knot as M is correct and turns out to be C. The cell (k, k') indicates the cases where the grader selected one knot as M, but the specimen failed at a different knot C. The few specimens that were deemed by the grader to be "off-grade" (i.e., in this case not belonging to the grade 1650f-1.5E) were removed from the analysis.

5.2 Analysis Results

The category counts in Table 5 are strongly asymmetric, for example comparing the (k, o) and (o, k) cells in the bending table. This suggests that our grader overestimated the strength effects of knots compared to shake and other, and we include the bias terms b_s and b_o in the model fitting with $N(0, 10^2)$ priors. With the full models fitted, the posterior distributions of the parameters are summarized in Table 6. Then applying the predictive distribution on the test sets, we found an MSPE of 1.45 for R and 1.35 for T.

The fitted mean effects of the three categories are not significantly different between the two destructive tests based on these samples. However, due to having fewer samples in the

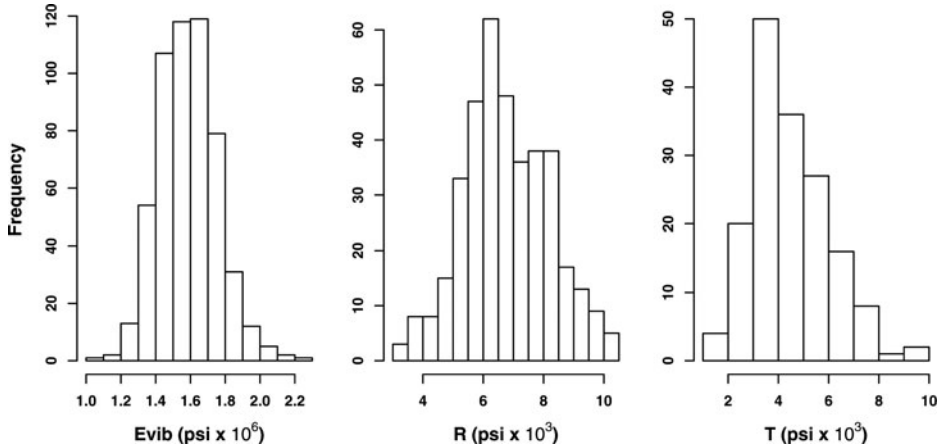

Figure 2. Histograms for the sample of strength values stiffness (MOE), bending strength (R), and ultimate tensile strength (T) based on sample sizes of 544, 380, and 164, respectively.

Table 4. Marginal frequencies of occurrence of the M and C for both the samples for bending strength (R) and ultimate tensile strength (T)*

Code	Explanation	R,M	R,C	T,M	T,C	Category
1	Single knot (type 1, tension edge)	4	9	2	4	k
2	Single knot (type 2, tension edge)	2	4	7	6	k
3	Single knot (type 3, tension edge)	38	29	30	11	k
4	Single knot (type 4, tension edge)	15	17	9	8	k
8	Single knot (type 8, tension edge)	44	28	17	19	k
9	Single knot (type 9, tension edge)	46	43	12	11	k
10	Knot combination (pith present)	106	101	10	17	k
11	Single knot (type 1, compression edge)	1	2	1	2	k
12	Single knot (type 2, compression edge)	7	2	2	2	k
13	Single knot (type 3, compression edge)	18	15	13	11	k
14	Single knot (type 4, compression edge)	3	1	2	1	k
18	Single knot (type 8, compression edge)	25	8	18	14	k
19	Single knot (type 9, compression edge)	5	5	6	11	k
20	Knot combination (no pith)	17	11	4	1	k
23	Knot cluster	9	5	2	2	k
24	Slope of grain (wide face)	4	12	1	6	s
25	Grain deviation	5	3	2	3	s
26	Cross-grain (narrow face)	—	4	—	1	s
27	Shake and checks	3	5	4	7	s
32	Pinworm holes	—	—	—	1	o
35	Bark pocket	6	5	—	—	o
37	Wane	—	—	1	1	o
45	Machine damage	5	2	3	2	o
46	Falling break	—	—	—	1	o
50	Longitudinal shear	—	3	—	—	o
52	Tension and compression failure	—	11	—	—	o
53	Tension failure	—	—	—	6	o
55	Brash	—	14	—	6	o
60	Small defects	18	44	18	10	o

NOTE: *For example, in the bending tests, 4 specimens out of the 380 in the sample were deemed by the grader to have a visual characteristic of code type 1 as the M.

"shake" and "other" cells, the estimates of μ_s and μ_o have much uncertainty. The σ_x values do appear to be significantly different between R and T. This suggests that it may intrinsically be more difficult to correctly select M in bending. From a physical perspective, the top and bottom edges of a specimen undergoing bending experience tensile forces and compressive forces, respectively; as the effect of a characteristic may depend on the type of force applied to it, additional uncertainty in the grader's effect assessments may thus ensue for bending specimens. In contrast, specimens undergoing tension experience uniform tensile forces throughout. There is significant grader bias in the selection of M for bending, as evidenced by the negative b_s and b_o values, which imply an underestimation of the effects of shake and other compared to knots. There is likely some bias for tension as well, but not as apparent. Even after accounting for grader bias, both the bending and tension σ_x values are relatively large compared to the estimated latent effects. Taken together, these results suggest that the predictive power gained from incorporating M in its current form is small at best for a high-quality grade of lumber such as this one.

Table 5. Frequencies of M and C based on categories knot, shake, and other; bending (left) and tension (right)

$M\backslash C$	k_1	s	o	$M\backslash C$	k_1	s	o
k	102	19	52	k	57	10	12
k'	164	—	—	k'	54	—	—
s	5	5	4	s	1	3	4
o	4	1	19	o	6	3	10

Table 6. Summaries of posterior distributions of parameters under tension and bending

	Tension Posterior quantiles			Bending Posterior quantiles		
	50%	2.50%	97.50%	50%	2.50%	97.50%
β_0	3.4	2.2	5.8	5.0	4.4	5.9
β_1	4.0	3.1	4.7	3.3	2.6	3.8
μ_k	18.2	16.3	19.9	18.6	17.4	19.9
σ_k	4.3	2.8	5.6	6.4	5.7	7.3
μ_s	19.6	16.5	22.1	12.7	4.8	18.7
σ_s	3.6	2.2	6.5	9.0	5.7	16.9
μ_o	5.4	0.7	10.3	7.7	4.8	12.0
σ_o	11.0	7.1	15.9	9.7	6.7	11.9
γ	−0.26	−0.36	−0.21	−0.18	−0.20	−0.16
σ_x	9.3	8.3	10.2	11.1	10.5	12.5
b_s	−5.9	−15.0	0.7	−10.8	−18.4	−4.4
b_o	−3.8	−8.5	1.1	−16.3	−20.4	−12.1

6. DISCUSSION AND CONCLUSIONS

This study presented a coherent framework for understanding the uncertainty involved in the process of grading lumber and for constructing a prediction rule to apply on future specimens. This permitted the effect of unrecorded characteristics, due to the design of grading protocols and the choices of the grader, to be modeled and analyzed. A fitted model provides quantification of uncertainty in the selection of M, which in turn informs the value of M as a predictor of C and ultimately breaking strength.

The model presented could quite easily be extended to include more than three categories. This would require more extensive testing, and would ideally take place on a more variable global population of lumber. The results of this immediate study do suggest some recommendations for such future testing. First, the amount of information recorded on the characteristics in the current grading protocol may be too heavily censored to be useful for prediction purposes. For instance, the number and types of characteristics present could be recorded without too much additional effort, even if they are not individually documented in detail, and yet provide valuable information toward prediction. Second, based on these results graders (whether human or automated) might require calibration to adjust for the bias currently seen in their subjective assessments of effects among the different characteristics, or alternatively the underlying rules for selecting M could be updated to increase the probability that $M = C$. A robust analysis toward this latter objective, however, would require a full documentation of all characteristics on test specimens, which we anticipate will be achievable in the future.

SUPPLEMENTARY MATERIALS

Lumber strength testing experiment: Details about laboratory experiment and the species of lumber samples used to produce the dataset analyzed in this article. Examples of specimens are also illustrated. (PDF file)

ACKNOWLEDGMENTS

The work reported in this article was partially supported by FPInnovations and grants from the Natural Sciences and Engineering Research Council of Canada. We thank Roy Abbott for professionally grading the lumber test samples used in this study, and the staff at FPInnovations for assisting with the testing procedure. We thank Yilan Zhu, Yan Cheng, Yanling Cai, Jessica Chen, Yang Liu, Yongliang Zhai, and Chen Xu for their assistance in the FPInnovations laboratory to produce the datasets used in this study. Thanks to Lynne Zidek for entering the experimental data. Thanks to Yang Chen for some helpful discussions. Finally, an expression of appreciation to the anonymous reviewers, the associate editor, and the former editor, Hugh Chipman, for many suggestions that greatly enhanced the article's clarity.

[Received January 2012. Accepted March 2015.]

REFERENCES

Andalo, C., Beaulieu, J., and Bousquet, J. (2005), "The Impact of Climate Change on Growth of Local White Spruce Populations in Quebec, Canada," *Forest Ecology and Management*, 205, 169–182. [236]

Divos, F., and Tanaka, T. (1997), "Lumber Strength Estimation by Multiple Regression," *Holzforschung*, 51, 467–471. [237]

Green, D., Ross, R., and McDonald, K. (1994), "Production of Hardwood Machine Stress Rated Lumber," in *Proceedings of 9th lnternational Symposium on Nondestructive Testing of Wood*, pp. 141–150. [237]

Kretschmann, D., Evans, J., and Brown, L. (1999), "Monitoring of Visually Graded Structural Lumber," Technical Report, U.S. Forest Products Laboratory. [236,237,238]

Lei, Y., Zhang, S., and Jiang, Z. (2005), "Models for Predicting Lumber Bending MOR and MOE Based on Tree and Stand Characteristics in Black Spruce," *Wood Science and Technology*, 39, 37–47. [237]

NIST (2010), "Voluntary Product Standard PS20-10: American Softwood Lumber Standard," Technical Report, National Institute of Standards and Technology. [237,241]

Pellerin, R. (1965), "A Vibrational Approach to Nondestructive Testing of Structural Lumber," *Forest Products Journal*, 15, 93–101. [237]

Ross, R., Geske, E., Larson, G., and Murphy, J. (1991), "Transverse Vibration Nondestructive Testing Using a Personal Computer," Technical Report, U.S. Forest Products Laboratory. [237]

Swendsen, R. H., and Wang, J.-S. (1986), "Replica Monte Carlo Simulation of Spin-Glasses," *Physical Review Letters*, 57, 2607. [239]

Taylor, S., Bender, D., Kline, D. E., and Kline, K. S. (1992), "Comparing Length Effect Models for Lumber Tensile Strength," *Forest Products Journal*, 42, 23–30. [237]

Wilhelm, S., and Manjunath, B. G. (2014), *Tmvtnorm: Truncated Multivariate Normal and Student t Distribution*, R package version 1.4-9. [239]