Wood Sci Technol (2008) 42:521–534
DOI 10.1007/s00226-008-0204-7

ORIGINAL

Prediction of internal bond strength in a medium density fiberboard process using multivariate statistical methods and variable selection

Nicolas André · Hyun-Woo Cho · Seung Hyun Baek · Myong-Kee Jeong · Timothy M. Young

Received: 18 May 2007 / Published online: 17 July 2008

Abstract This paper presents new data mining-based multivariate calibration models for predicting internal bond strength from medium density fiberboard (MDF) process variables. It utilizes genetic algorithms (GA) based variable selection combined with several calibration methods. By adopting a proper variable selection scheme, the prediction performance can be improved because of the exclusion of non-informative variable(s). A case study using real plant data showed that the calibration models based on the process variables selected by GA produced better performance than those without variable selection, with the exception of the radial basis function (RBF) neural networks model. In particular, the calibration model based on supervised probabilistic principal component analysis (SPPCA) yielded better performance (only when using GA) than partial least squares (PLS), orthogonal-PLS (O-PLS), and radial basis function neural networks models. The SPPCA model benefits most from the use of GA-based variable selection in this case study.

Introduction

The wood composites industry is undergoing changes in the form of corporate divestures and consolidation, increases in the cost of raw material and energy, and strong international competition. The forest products industry is an important

N. André and H.-W. Cho equally contributed to this work.

N. André (✉) · T. M. Young
Forest Products Center, The University of Tennessee, Knoxville, TN 37996-4570, USA
e-mail: nandre@utk.edu

H.-W. Cho · S. H. Baek · M.-K. Jeong
Department of Industrial and Information Engineering, The University of Tennessee, Knoxville, TN 37996, USA

contributor to the US economy. In 2002, this sector contributed more than $240 billion to the economy and employed more than 1,000,000 Americans in 22,231 primary wood products manufacturing facilities (US Census Bureau 2004). Sustaining business competitiveness by reducing costs and optimizing product quality will be essential for this industry. One of the challenges facing this industry is to develop a more advanced knowledge of the complex nature of process variables and quantify the causality between process variables and final product quality characteristics. Improved production efficiency and business competitiveness of medium density fiberboard (MDF), a key product produced by the wood composites industry, can be realized from the methods presented in this paper.

Large-scale production of MDF began in the 1980s. MDF is an engineered wood product formed by breaking down wood chips into wood fibers, often in a defibrator (i.e., "refiner"), combining it with resin, and forming panels by applying high temperature and pressure. MDF has become one of the most popular composite materials in recent years. MDF is uniform, dense, smooth, and free of knots and grain patterns, and is an excellent substitute for solid wood in many applications. MDF's name derives from the distinction in densities of fiberboard. MDF typically greater than 12 mm in thickness has a density of 600–800 kg/m^3. High-density fiberboard (less than 12 mm thickness) has a density of 500–1,450 kg/m^3.

Some work has been initiated in data mining and predictive modeling of final product quality characteristics of forest products using statistical methods (Young 1996; Bernardy and Scherff 1998; Greubel 1999; Erilsson et al. 2000; Cook et al. 2000; Young and Guess 2002; Young et al. 2004). A study investigated the performance of RBF, backpropagation and counterpropagation neural networks for the prediction of internal bond from particleboard process variables (Cook and Chiu 1997). Each neural network was trained with a set of 152 observations (each observation was composed of 26 process variables and their corresponding panel internal bond value) and was validated with a set of 30 observations. The average prediction error was 12.5% for the RBF network, 40% for the backpropagation network and 30% for the counterpropagation network.

Most of the papers using theoretical models to predict final product quality characteristics have been published since 2001 (Barnes 2001; Gupta et al. 2007; Lee et al. 2007; Painter et al. 2005; Pereira et al. 2006; Thömen and Humphrey 2003). We are not aware of any published literature that investigates the use of supervised probabilistic regression and variable selection for modeling the MDF process.

The objective of a calibration model is to predict quality characteristics of interest from experimental or historical data. The high dimensionality and collinearity of the data, in general, makes it difficult to construct good calibration models (Qin 2003). The need to model such data led to the use of multivariate calibration models such as partial least squares (PLS). PLS has been shown to be a powerful technique for multivariate calibration of noisy, collinear, high-dimensional, and ill-conditioned data (Qin 2003). A commercial application (BoardModel^{TM}, owned by Casco Adhesives) actually merges near-infrared spectroscopic data of the raw material (particleboard furnish) with the plant process variables. PLS is used to build a relationship between the merged data ($\mathbf{X}$) and the internal bond or the modulus of rupture ($\mathbf{Y}$) of the manufactured panels

(Sjöblom et al. 2004). Recently, new calibration methods have been developed such as orthogonal-PLS (O-PLS) and supervised probabilistic principal component analysis (SPPCA) (Trygg and Wold 2002; Yu et al. 2006).

Variable selection in multivariate analysis is quite an essential step because the exclusion of non-informative variables will produce better results even with simpler models. The predictor variable **X**, in general, contains unwanted variation that is unrelated (or orthogonal) to the response variable **Y**. In such a case, the unwanted variation may give rise to the degradation of prediction ability of a calibration model. Several approaches to variable selection in calibration have been developed such as iterative variable selection (Lindgren et al. 1994), uninformative variable elimination (Centner et al. 1996), and iterative predictor weighting (Forina et al. 1999). Recently, it has been shown that genetic algorithm (GA) can be successfully used as a very efficient variable selection technique for PLS (Leardi 2001; Gourvénec et al. 2004; Esteban-Díez et al. 2006). The selection of variables for calibration can be considered as an optimization problem. In this respect, GA is a very efficient technique for variable selection because the size of the search domain is large and there exist many local optima.

The objective of this study is to evaluate the predictive performance of several well established statistical methods such as PLS and RBF neural network and some more recent methods such as SPPCA and O-PLS, with and without GA-based variables selection. A typical sampling and testing scheme for internal bond allows for two to three hours of production between two consecutive tests. Predicting the internal bond or another indicator of the panel quality during that time interval can help identifying problems in the production, improving the process and controlling the number of "reject" panels. Moreover, throughput can be improved when internal bond strength values are higher than needed. With an effective process control system, the decrease of quality can be avoided and costs for raw material and energy could be reduced.

Materials and methods

PLS and O-PLS

PLS was developed to model the relation between a predictor matrix **X** and a response matrix **Y**. Many incomplete and correlated variables in X can be handled in a simple and efficient way. The objective of PLS is to maximize the covariance between a PLS score vector **t** and **y**:

$$\begin{aligned} \mathrm{Max}(\mathbf{t}_T\mathbf{y})^2 &= \mathrm{Max}(\mathbf{w}^T\mathbf{X}^T\mathbf{y})^2 \\ &= \mathrm{Max}(\mathbf{w}^T\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}\mathbf{w})^2 \end{aligned} \tag{1}$$

where $\|\mathbf{w}\| = 1$. It linearly transforms the **X** matrix into a new set of orthogonal vectors **T**, whose inverse $(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}$ exists and is diagonal. The PLS regression model is expressed as

$$\mathbf{Y} = \mathbf{X}\mathbf{B}_{\mathrm{PLS}} + \mathbf{F}_{\mathrm{PLS}} \tag{2}$$

where $\mathbf{B}_{\mathrm{PLS}}$ are the PLS regression coefficients of $\mathbf{X}$ for the prediction of $\mathbf{Y}$. By defining $\mathbf{K}_{\mathrm{PLS}}^T = (\mathbf{B}_{\mathrm{PLS}}^T\mathbf{B}_{\mathrm{PLS}})^{-1}\mathbf{B}_{\mathrm{PLS}}^T$ and rearranging Eq. (2) it is possible to obtain identical predictions of $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{X}\mathbf{K}_{\mathrm{PLS}}(\mathbf{K}_{\mathrm{PLS}}^T\mathbf{K}_{\mathrm{PLS}})^{-1} + \mathbf{F}_{\mathrm{PLS}}. \tag{3}$$

O-PLS is a recently developed hybrid method combining traditional PLS and orthogonal signal correction (OSC). The basic idea of O-PLS is to separate the variation in $\mathbf{X}$ and $\mathbf{Y}$ into three parts: one contains the variation common in $\mathbf{X}$ and $\mathbf{Y}$ and the other two the specific variation for $\mathbf{X}$ and $\mathbf{Y}$, called structured noise. An O-PLS model of $\mathbf{X}$ and $\mathbf{Y}$ is given by

$$\mathbf{X} = \mathbf{T}\mathbf{W}^T + \mathbf{T}_{\mathbf{Y}}\mathbf{P}_{\mathbf{Y}}^T + \mathbf{E} \text{ and } \mathbf{Y} = \mathbf{T}\mathbf{C}^T + \mathbf{F}. \tag{4}$$

In Eq. (4), $\mathbf{W}$ and $\mathbf{C}$ are joint orthonormal loading matrices, $\mathbf{T}_{\mathbf{Y}}$ represents the score matrix orthogonal to $\mathbf{Y}$, and $\mathbf{P}_{\mathbf{Y}}$ the corresponding loading. Consequently, the prediction of Y can be obtained by $\mathbf{Y}_{\mathrm{hat}} = \mathbf{T}\mathbf{C}^T$. Such an O-PLS model provides improved interpretation of the models because the structured noise can be modeled separately from the common variation. Note that traditional PLS models the structured noise and the correlated variation between $\mathbf{X}$ and $\mathbf{Y}$.

Supervised probabilistic principal component analysis (SPPCA)

SPPCA was first introduced by (Yu et al. 2006) and is based on latent variable models. The key concept of supervised probabilistic PCA considers that all the observations are conditionally independent given the latent variables. In SPPCA, the observed data (x, y) is generated from a latent variable model as:

$$\begin{aligned} \mathbf{x} &= \mathbf{W}_x\mathbf{d} + \boldsymbol{\mu}_x + \varepsilon_x \\ \mathbf{y} &= \mathbf{W}_y\mathbf{d} + \boldsymbol{\mu}_y + \varepsilon_y \end{aligned}. \tag{5}$$

The latent variable $(\mathbf{d})$ and the error terms $(\boldsymbol{\varepsilon}_x,\boldsymbol{\varepsilon}_y)$ are defined as isotropic Gaussians distribution: $\mathbf{d} \sim N(0,\mathbf{I}), \varepsilon_x \sim N(0,\sigma_x^2\mathbf{I}), \varepsilon_y \sim N(0,\sigma_y^2\mathbf{I})$. It is shown that the maximum likelihood estimate of $\mathbf{W}_x$ and $\mathbf{W}_y$ are given by (Yu et al. 2006):

$$\begin{aligned} \tilde{\mathbf{W}}_x &= \sigma_x\mathbf{U}_M(\mathbf{D}_P - \mathbf{I}_P)^{1/2}\mathbf{R} \\ \tilde{\mathbf{W}}_y &= \sigma_y\mathbf{U}_N(\mathbf{D}_P - \mathbf{I}_P)^{1/2}\mathbf{R} \end{aligned} \tag{6}$$

where $\mathbf{U}_M \cdot \mathbf{U}_N$ contains the first M (or last N) rows of the eigenvectors of the normalized sample covariance matrix $\mathbf{S}$ for centered observations $\{(x_i, y_i)\}_{i=1}^l$,

$$\mathbf{S} = \begin{pmatrix} \frac{1}{\sigma_x^2}\mathbf{S}_{xx} & \frac{1}{\sigma_x\sigma_y}\mathbf{S}_{xy} \\ \frac{1}{\sigma_y\sigma_x}\mathbf{S}_{yx} & \frac{1}{\sigma_y^2}\mathbf{S}_{yy} \end{pmatrix}. \tag{7}$$

$\mathbf{D}_P \in \Re^{p\times p}$ is the diagonal matrix of the corresponding eigenvalues, $\mathbf{I}_P \in \Re^{p\times p}$ is the p-dimensional identity matrix, and $\mathbf{R}$ is an arbitrary $p \times p$ orthogonal matrix. The projected latent variable $\mathbf{d}^*$ for centered new input $\mathbf{x}^*$ is given by

$$\mathbf{d}^* = \frac{1}{\sigma_x}\mathbf{R}^T(\mathbf{D}_p - \mathbf{I}_p)^{1/2}[\mathbf{U}_M^T\mathbf{U}_M + (\mathbf{D}_p - \mathbf{I}_p)^{-1}]^{-1}\mathbf{U}_M^T\mathbf{x}^*. \quad (8)$$

When $N > 0$, where PCA only explains the covariance of inputs ($\mathbf{S}_{xx}$), and where PLS considers the covariance between inputs and outputs ($\mathbf{S}_{xy}$), SPPCA explains not only the covariance of inputs ($\mathbf{S}_{xx}$) and outputs ($\mathbf{S}_{yy}$), respectively, but also the inter-covariance between inputs and outputs, ($\mathbf{S}_{xy}$).

RBF neural networks

The radial basis function neural network has two layers, a hidden and an output layer. The hidden layer is a non-linear, local mapping layer and the output layer is composed of standard linear neurons (Buhmann 2003). The activation function (also called transfer function) of the hidden layer is a radial basis function called Gaussian activation function ($\mathbf{f}(\mathbf{x})$).These functions are centered over receptive fields, which activate the local radial basis neurons. The activation function of the output layer is a standard linear function and performs a linear transformation of the output of the hidden nodes. The weights of this function are solved using a least square algorithm. The Gaussian activation function for a radial basis neuron is:

$$\mathbf{f_i}(\mathbf{x}) = \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\mu}_i)^2}{\sigma_i^2}\right] \quad (9)$$

where $\mathbf{x}$ is the input vector, $\boldsymbol{\mu}_i$ is the center of a region called a receptive field, σ_i is the width of the receptive field, $\mathbf{f_i}(\mathbf{x})$ is the output of the i^{th} neuron, and i is the number of neurons.

Genetic algorithms

Genetic algorithms (GAs) are problem-solving methods inspired by evolution theory that simulate a natural evolution process (Hibbert 1993). GAs have been used successfully in solving many problems such as molecular modeling, curve fitting, classification, and variable selection for calibration (Blommers et al. 1992; DeWeijer et al. 1994; Leardi and Gonzalez 1998; Hunger and Huttner 1999; Kemsley 2001; Gourvénec et al. 2004; Esteban-Díez et al. 2006). In particular, GA has been shown to be a very efficient tool for variable selection in calibration (Leardi 2001). GA consists of four basic steps. First, each variable is represented by a binary code in a vector, with one cell for each variable. The original chromosome is then perturbed randomly to create the initial population. For each chromosome, the response associated with the experiment is evaluated, which is the criterion for guiding the GA to the global optimum. The reproduction step generates a new population of the next generation by recombining original chromosomes based on cross-over between two chromosomes. The pair of chromosomes chosen for cross-over is controlled by a cross-over probability. Finally, the mutation step, which alters each gene independently with a mutation probability, is required to overcome some problems: if a variable should not be selected in any of the original

chromosomes, it would never be selected in the coming generation (Leardi et al. 1992). During the genetic algorithm procedure, the cross validation is performed based on deletion groups. The deletion group is selected by every *k*th sample out of *n* sample in the data (Ghasemi and Ahmadi 2006).

Process and quality variables

To obtain the strength to failure data, workers test at least six 50 mm × 50 mm sample blocks from the cross sections of MDF panels and measure the tensile strength perpendicular to the surface (ASTM D1037 2006), also called internal bond strength (IB), in kilopascal (kPa) via destructive testing at different time periods. The IB strength is an indicator of the cohesion of the panel in the direction perpendicular to the plane of the panel. A special measuring device is utilized that pulls the cross section apart and stresses the specimen until failure. Unlike most of the process variables, IB cannot be determined online.

MDF has a large number of differing, but interdependent process variables that have complex functional forms that influence final IB quality characteristics. Raw material passes through many processing stages that may influence physical properties. Key parameters may include wood chip dimensions, fiber dimension, fiber-resin formation, mat-forming consistency, line speed, press closing characteristics, etc. At the time of production, the quality of the board is unknown, i.e., samples are analyzed as an event using destructive tests in the lab. The time span between destructive tests may be as long as two hours, during which a significant amount of production occurs. Given the time gap between consecutive destructive tests, hours of unacceptable MDF could be manufactured. Cost savings from accurately predicting the real-time IB may be realized from a reduction in waste, reduced customer claims, faster press cycles, lower wood usage, lower resin usage and lower energy usage.

Relational database

A real-time, automated relational database was created which aligned real-time process sensor data with IB (Young and Guess 2002). Real-time process data were collected using Wonderware Industrial SQL 8.0 (developed by Wonderware, a sub-business unit of Invensys PLC) and were combined with IB by product type at the instant when a panel was extracted from the production line for testing. The process data were collected using a median value from the last 100 sensor values (e.g., for most of the different sensor variables this represents a two to three minute time interval). Any process variable's value falling outside of its minimum–maximum range (set by the electrical engineer) would not be recorded in the relational database. Lag times, corresponding to the time required for the fibers or particles to travel through the process from the point where a given parameter had an influence to the point where the panel was extracted for IB destructive testing, were taken into account when collecting process data with Industrial SQL. A unique number was generated when the panel was extracted

from the process, and was later used to match process data with corresponding IB results.

Results and discussion

A large data set (495 time-ordered records) has been extracted from the relational database, representing several months worth of process data collection (**X**: 164 variables) and IB destructive testing (**Y**). This data set was split into two sets, a training set (first 445 records) to build calibration models using several multivariate methods (with or without variable selection) and a test set (50 remaining records) to validate the calibration models.

Preliminary tests and determination of optimal parameters

Outliers can significantly reduce the performance of data-based prediction models. There are two possible approaches to detect outliers. One can construct classical regression models and then remove the outliers using regression diagnostics. The second approach consists in using statistics such as residuals obtained from robust multivariate techniques (Rousseuw and Leroy 1987). In this work, outliers were analyzed based on supervised probabilistic principal component analysis (SPPCA). The second approach was selected because regression diagnostics are often affected by the outliers they have to detect.

GA-based variable selection was successfully applied to various data (Leardi and Gonzalez 1998; Leardi 2001). Leardi and Gonzalez (1998)'s algorithm was adopted in this work to select the influent variables of a MDF manufacturing process. The parameters used in implementing GA-PLS variable selection for the MDF process data were as follow: population size: 30 chromosomes, 30 maximum number of variables selected in the same chromosome, average 5 variables per chromosome in the original population, 5 deletion groups, 1% mutation probability, 50% cross-over probability, and 100 runs.

First, a randomization test was performed to access the potential overfitting problem in a GA-based variable selection of the data (Leardi and Gonzalez 1998). This test provides a metric to assess the risk of overfitting and is executed by randomizing **Y** relative to **X** and building a calibration model. The order of the elements of **Y** is randomized so that each row of **X** corresponds to the randomized **Y**, not to its own element of **Y**. Thus there is no information in such a dataset so that the calibration model constructed from a randomization test should have no significant prediction ability. If such prediction ability exists, this finding indicates the presence of overfitting. Figure 1 shows the results of the randomization tests. Here, 50 GA runs were performed on randomized datasets, and in each run, 100 chromosomes were evaluated. The fitness function of this randomization test is the average percentage of explained variances in cross-validation. The average value from the randomization test is 2.13, where, in general, the better or more reliable a dataset, the lower this average value is. As a rule of thumb, GA can be applied

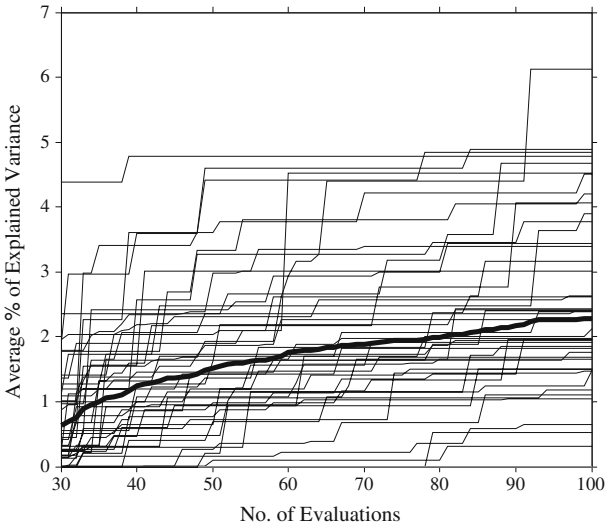

Fig. 1 Randomization test plot

safely without overfitting problem when an average value is less than 10 (Leardi and Gonzalez 1998). A critical decision in implementing GA is to determine when to stop a GA run. Thus one needs to select the optimal number of evaluations to perform in each GA run to obtain a good calibration model without an overfitting problem. Performing too many evaluations in each GA run means that noise in data is modeled. A total of 40 runs (each with a maximum of 200 evaluations) were performed. That is, the first 20 runs were based on real Y values and the last 20 runs were based on randomized Y values. By investigating the difference between the averages of real and randomized runs as a function of the number of evaluations, the optimal number of evaluations was chosen as 200, after which no significant increase in the degree of overfitting was observed.

Variable selection results

GA was performed on the training data set using the optimal number of evaluations (i.e., 200 evaluations). In this work, genetic algorithms are used as an optimization tool to determine variables that provide better prediction of internal bond strength of medium density fiberboard. To verify the robustness of the variable selection results, a GA-based variable selection procedure was performed with 100 runs five times to produce five different sets of selected variables. The main reason for this procedure is that the variables selected from each of five GA models will not be identical because of the stochastic nature of GA. Thus common information was extracted from the results of five GA models to select the most informative variables. This information can be obtained from the frequency with which each variable is selected. In summary, variables were selected that are consistently selected (i.e., variables having high selection frequency) in the five GA models.

As an example, Fig. 2 shows several plots obtained from two of the five GA models. Figure 2a and c are two bar plots of the cumulative frequency of selections. By investigating them, it becomes clear which variables are selected most often and which ones are rarely or never selected. The horizontal lines of Fig. 2a and c represent the cutoff values for selecting informative variables. They are calculated by an F test. For more details, refer to Leardi and Gonzalez (1998). In addition, as shown in Fig. 2b and d, the number of selected variables can be easily recognized by referencing two threshold values, denoted as asterisk points, given at 80 and 92 in the root mean squared error of cross validation (RMSECV) plots. Consequently, a total of 56 variables was chosen out of 164 original variables. A description of the top ten selected process variables is shown in Table 1. Not surprisingly, the top ten selected process variables are related to several of the refiner's parameters, fibers moisture contents, resin percentages and line speed (main motor forming line power).

Prediction results

Several calibration models were built using a training data set and were tested on a validation (test) data set to evaluate the performance for predicting the internal bond

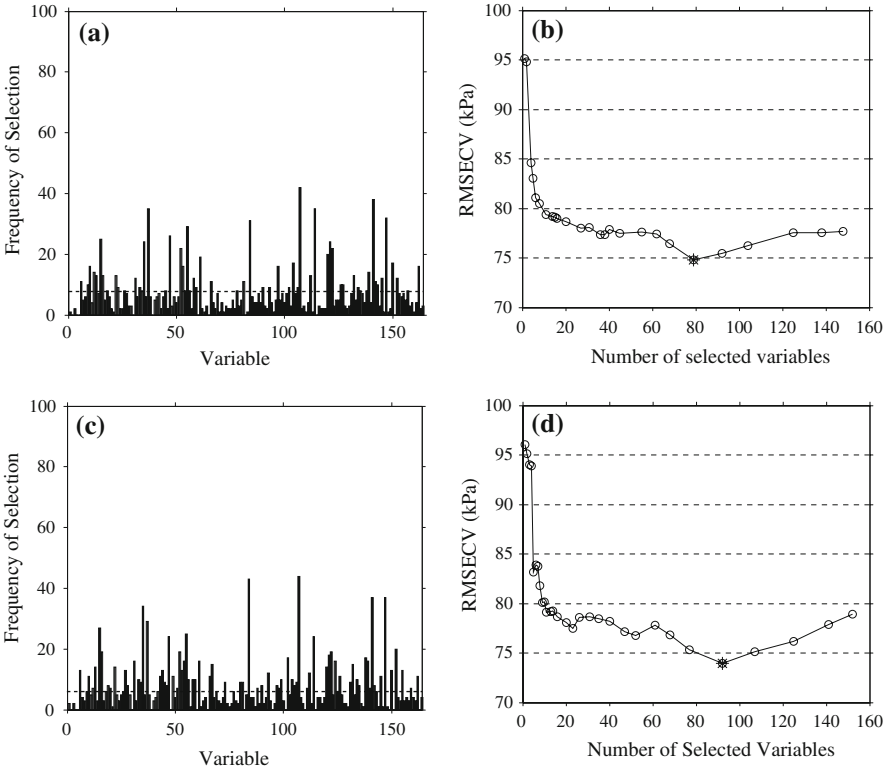

Fig. 2 Variable selection plots for two different GA models: **a** and **c** cumulative frequency of selection after 100 runs and (**b**) and (**d**) RMSECV plots

Table 1 Top ten ranked MDF process variables among the 56 process variables selected by the GA-PLS method (out of 164 original process variables)

Top ten selected variables' description	Frequency of selection by GA-PLS
01. Press movement time to pre-position set point	41
02. Core fiber moisture percent	40
03. Face fiber moisture percent	36
04. Resin to wood ratio for refiner X as percent	32
05. Core refiner blow line steam pressure	31
06. Main motor forming line power	30
07. Core fiber mat weight without moisture	28
08. Face refiner grinding steam flow	26
09. Face resin to wood for face refiner percent	24
10. Core resin to wood for face refiner percent	23

strength. The models were based on RBF neural networks, PLS, O-PLS and SPPCA. The potential advantage of performing a GA-based selection of the influent process variables that will be employed to construct the calibration models was investigated too. Four "full" models were constructed using a full set of **X** variables without performing GA-based variable selection. On the other hand, "Variable Selected" prediction models were built using 56 **X** variables selected from GA. For the validation of the calibration models, a total of 495 samples was split into a training data set of 445 samples and a test data set of 50 samples. GA-based variable selection was based on the 445 training samples. To compare the predictive abilities of the models, the root mean squared error for prediction (RMSEP) was used as a performance measure for the test data set. The RMSEP value is given by

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}, \tag{10}$$

where y_i is the actual value, $\hat{y}_i$ the predicted value, and n is the total number of test samples (records).

To perform PLS and O-PLS, the SIMCA-P software (Umetrics, Sweden) was run whilst RBF neural networks and SPPCA were implemented in MATLAB 6.5 (MathWorks Inc., Natick, MA, USA). A critical parameter of a calibration model is the number of latent variables retained for model-building. It should be determined by considering both the curse of dimensionality and the loss of data information. In this work, a cross-validation method was used to select the number of latent variables for the calibration models (Wold 1978), which are based on the predicted residual error sum of squares (PRESS). The number of latent variables with the minimum PRESS value was chosen as the optimum.

The prediction results for the MDF process data are summarized in Table 2. Here, the "Full Model" represents the validation results of the four types of calibration models based on a full set of variables. On the other hand, "variable selected model" represents the validation results of the four kinds of calibration

Table 2 Prediction errors of the calibration models with and without variable selection on the validation (test) data set

	RMSEP (kPa) (Mean-normalized RMSEP)	
	Full model[a]	Variable selected model[b]
RBF neural networks	75.26 (6.81%)	93.77 (8.20%)
Traditional PLS	72.88 (6.80%)	66.74 (6.18%)
Orthogonal PLS (O-PLS)	67.91 (6.37%)	66.95 (6.20%)
Supervised probabilistic PCA (SPPCA)	77.70 (7.23%)	65.22 (5.89%)

[a] Calibration model built without GA variable selection using the training set (445 records,164 variables)

[b] Calibration model built with GA variable selection using the training set (445 records, 56 selected variables)

models based on the 56 variables selected by GA. For the RBF neural network, 21 spread constants were used for the best fitted full model while 8 spread constants were used in the variable selected model. As shown in Table 2, in the case of "Full Model", the O-PLS produced the best prediction performance for the test data with the minimum RMSEP value of 67.91 kPa and an average error rate of 6.37%. The full O-PLS model shows a slightly better predictive ability than the traditional PLS model with a RMSEP value of 72.88 kPa. It should be noted that O-PLS requires fewer latent components (i.e., one component in this case) than traditional PLS (i.e., four components) to predict the internal bond strength with a lower RMSEP. It may be due to the fact that O-PLS offers the advantage of separating the non-relevant or **Y**-orthogonal variation and the relevant or predictive variation in **X**. Table 2 also shows that the RBF neural network and the PLS-based calibration models produced slightly lower RMSEP than the SPPCA-based one.

In case of "variable selected model", the best prediction performance was obtained from SPPCA (i.e., RMSEP = 65.22 kPa and average error rate = 5.89%). Although the SPPCA-based full model has the highest RMSEP, the performance of the "Variable Selected SPPCA Model" significantly improved from 77.70 to 65.22 kPa. Overall, all variable selected models except RBF neural networks show better prediction performances than their full-variable models. The GA-based variable selection models for PLS and O-PLS, for example, were able to predict **Y** with RMSEP values of 66.74 and 66.95 kPa, respectively. Such results demonstrate that variable selection helps reducing the dimensionality of the data while slightly improving the prediction results for three out of four models.

The SPPCA calibration model benefits most from the use of GA-based variable selection for calibration modeling. To visualize the SPPCA prediction performances of variable selected vs. full regression models, the predicted values for the test data were plotted against the observed values, which are shown in Fig. 3. In such plots, the data should fall on the diagonal (target line), which means that the calibration model predicts new data perfectly (correlation coefficient $r = 1$). It is observed that the GA-based variable selected regression model (Fig. 3a) has a better predictive ability than the full model (Fig. 3b). This is evident when comparing the degree of dispersion of the test data and the correlation coefficients (0.71 versus 0.58) in the

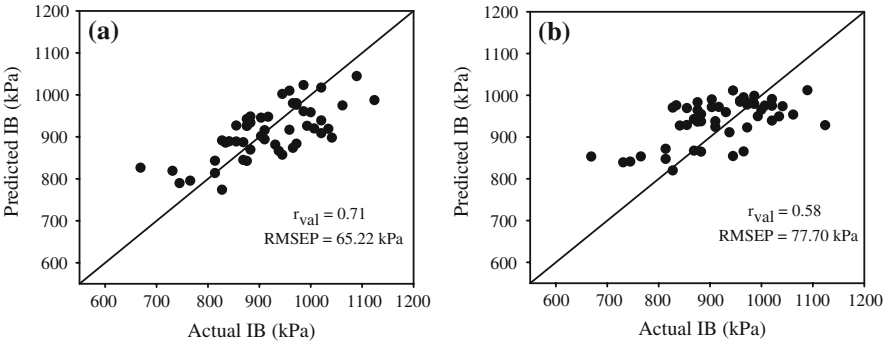

Fig. 3 Observed versus predicted plots for the validation (test) data set using SPPCA: **a** variable selected model and **b** full model

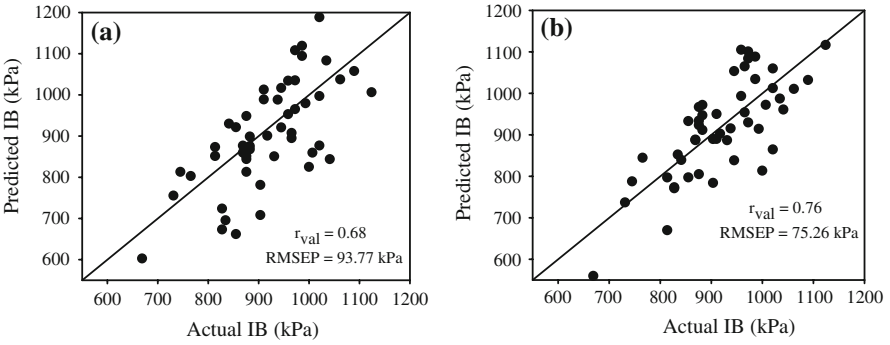

Fig. 4 Observed versus predicted plots for the validation (test) data set using RBF Neural networks: **a** variable selected model and **b** full model

two plots. The variable selected model produced reliable predicted values closer to the target line than the full model. For comparison purposes, similar plots are shown in Fig. 4, which were obtained from the RBF neural network models. It is observed that the GA-based variable selected regression model (Fig. 4a) has a lower predictive ability than the full model (Fig. 4b). The correlation coefficient for the best performing RBF model (0.76 for the full model) is higher than the one for the best performing SPPCA model (0.71 for the variable selected model). Yet, the RMSEP of the full RBF model is 15% higher than the RMSEP of the variable selected SPPCA model.

Conclusion

This work demonstrates the use of GA-based variable selection combined with several calibration methods for predicting the internal bond strength of medium density fiberboard from the plant process variables. By adopting a proper variable

selection scheme based on GA-PLS, the prediction performance of several calibration models was improved. It was made possible by removing less important variables of **X**. Among several calibration techniques employed, SPPCA was the best performer when coupled with the GA-PLS variable selection, yielding an average error rate of 5.89%. Yet, SPPCA without variable selection had the highest prediction error. The prospect of being able to predict the internal bond strength with such a small error on a longer time period could be very beneficial for the plant operators and managers. Real-time knowledge of the quality of the manufactured product will allow to run the production on tighter specifications and will help reducing some of the raw material usage and/or increasing the overall throughput.

References

ASTM D1037 (2006) Standard test methods for evaluating properties of wood-base fiber and particle panel materials. ASTM International, West Conshohocken

Barnes D (2001) A model of the effect of strand length and strand thickness on the strength properties of oriented wood composites. For Prod J 51(9):36–46

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

Blommers MJJ, Lucasius CB, Kateman G, Kaptein R (1992) Conformation analysis of a dinucleotide photodimer with the aid of the genetic algorithm. Biopolymers 32:45–52

Buhmann MD (2003) Radial basis functions: theory and implementations. Cambridge University Press, Cambridge

Centner V, Massart DL, de Noord OE, de Jong S, Vandeginste BM, Sterna C (1996) Elimination of uninformative variables for multivariate calibration. Anal Chem 68:3851–3858

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

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

DeWeijer AP, Lucasius CB, Buydens LMC, Kateman G, Heuvel HM, Mannee H (1994) Curve fitting using natural computation. Anal Chem 66:23–31

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

Esteban-Díez I, González-Sáiz JM, Gómez-Cámara D, Pizarro Millan C (2006) Multivariate calibration of near infrared spectra by orthogonal wavelet correction using a genetic algorithm. Anal Chim Acta 555:84–95

Forina M, Casolino C, Pizarro Millan C (1999) Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. J Chemom 13:165–184

Ghasemi J, Ahmadi S (2006) Combination of genetic algorithm and partial least squares for cloud point prediction of nonionic surfactants from molecular structures. Ann Chim 97(1–2):69–83

Gourvénec S, Capron X, Massart DL (2004) Genetic algorithms (GA) applied to the orthogonal projection approach (OPA) for variable selection. Anal Chim Acta 519:11–21

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

Gupta A, Jordan P, Pang S (2007) Modelling of the development of the vertical density profile of MDF during hot pressing. Chem Prod Process Modeling 2(2):Article2

Hibbert DB (1993) Genetic algorithms in chemistry. Chemom Intell Lab Syst 19:277–293

Hunger J, Huttner G (1999) Optimization and analysis of force field parameters by combination of genetic algorithms and neural networks. J Comput Chem 20:455–471

Kemsley EK (2001) A hybrid classification method: discrete canonical variate analysis using a genetic algorithm. Chemom Intell Lab Syst 55:39–51

Leardi R (2001) Genetic algorithms in chemometrics and chemistry: a review. J Chemom 15:559–569
Leardi R, Gonzalez AL (1998) Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemom Intell Lab Syst 41:195–207
Leardi R, Boggia R, Terrile M (1992) Genetic algorithms as a strategy for feature selection. J Chemom 6(5):267–281
Lee JN, Kamke FA, Watson LT (2007) Simulation of hot-pressing of a multi-layered wood strand composite. J Compos Mater 41(7):879–904
Lindgren F, Geladi P, Rännar S, Wold S (1994) Interactive variable selection (IVS) for PLS. Part 1: theory and algorithms. J Chemom 8:349–363
Painter G, Budman H, Pritzker M (2005) Prediction of oriented strand board properties from mat formation and compression operating conditions. Part 2: MOE prediction and process optimization. Wood Sci Technol 40:291–307
Pereira C, Carvalho LMH, Costa CAV (2006) Modeling the continuous hot-pressing of MDF. Wood Sci Technol 40:308–326
Qin SJ (2003) Statistical process monitoring: basics and beyond. J Chemom 17:480–502
Rousseuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York
Sjöblom E, Johnsson B, Sundström H (2004) Optimization of particleboard production using NIR spectroscopy and multivariate techniques. For Prod J 54(6):71–75
Thömen H, Humphrey PE (2003) Modeling the continuous pressing process for wood-based composites. Wood Fiber Sci 35(3):456–468
Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). J Chemom 16:119–128
US Census Bureau (2004) 2002 Economic census: table 1. Advance summary statistics for the United States 2002 NAICS basis. Washington, DC. http://www.census.gov/mcd/asm-as1.html
Wold S (1978) Cross-validatory estimation of components in factor and principal components models. Technometrics 20:397–405
Young TM (1996) Process improvement through "real-time" statistical process control in MDF manufacture. In: Proceedings of process and business technologies for the forest products industry. Forest Products Society, Madison, pp 50–51
Young TM, Guess FM (2002) Developing and mining higher quality information in automated relational databases for forest products manufacture. Int J Reliab Appl 3(4):155–164
Young TM, André N, Huber CW (2004) Predictive modeling of the internal bond of MDF using genetic algorithms with distributed data fusion. In: Proceedings of the eighth European panel products symposium, pp 45–59
Yu S, Yu K, Tresp V, Kriegel H, Wu M (2006) Supervised probabilistic principal component analysis. In: Proceedings of the 12th international conference on knowledge discovery and data mining (SIGKDD), pp 464–473