ONLINE ISSN 2069-7430
ISSN-L 1841-4737

PRO LIGNO
www.proligno.ro

Vol. 9 N° 4 2013
pp. 22-27

THE EVOLUTION OF KNOWLEDGE IN FOREST PRODUCTS MANUFACTURING

Timothy M. YOUNG
Prof. Dr. Statistics - University of Tennessee - Center for Renewable Carbon
Address: 2506 Jacob Drive, Knoxville, TN 37996-4570 USA
E-mail: tmyoung1@utk.edu

Marius C. BARBU
Prof. Dr. - Salzburg University of Applied Sciences - Department of Forest Products Technology and Wood Constructions
Address: FACHHOCHSCHULE SALZBURG GmbH Markt 136a | 5431 Kuchl | Austria
E-mail: marius.barbu@fh-salzburg.ac.at

Alexander PETUTSCHNIGG
Prof. Dr. and Head - Salzburg University of Applied Sciences - Department of Forest Products Technology and Wood Constructions
Address: FACHHOCHSCHULE SALZBURG GmbH Markt 136a | 5431 Kuchl | Austria
E-mail: alexander.petutschnigg@fh-salzburg.ac.at

Abstract:

The manuscript is conceptual; it assesses the use of real-time process data by forest products manufacturers as related to data discovery and the co-evolutionary existence of data mining inductive algorithms. Forest products companies have intellectual latency concerning process variation and such latency leads to higher than necessary costs of manufacturing, inferior product quality, and inaccurate prediction. Strength properties from the laboratory are unreliable for real-time prediction given their historic temporal orientation. Induction of real-time process data when aligned with destructive test data from the laboratory improves manufacturing latency and results in discovery of unknown sources of process variation. This discovery of unknown sources of process variation allows for quantification of variation and defines the relationship with product attributes which improves accuracy of predictions.

The manuscript outlines the importance of the real-time relational database and data quality verification as essential steps in the data discovery journey. These essential steps are often overlooked and lead to antidotal analyses that limit process improvement and predictability. Statistical methods for inductive data mining are discussed, e.g., multiple linear regression analysis (MLR), regression trees (RT), principal component analysis (PCA), and quantile regression. The strengths and weaknesses of each method with appropriate citations are noted.

Keywords: *real-time relational database; data quality; induction; data mining.*

INTRODUCTION

Forest products industries are important to the global economy and are essential sources of export revenues for many countries. Many companies have recently faced unprecedented international competition, rising costs of energy and raw materials, and price erosion from the global economic recession and substitution to low cost non-wood products. Providing value to the customer at a low cost of manufacturing has never been more essential for this industry.

Knowledge discovery for manufacturers is evolving at an unprecedented rate. A co-evolutionary force for knowledge discovery has been created with the advent of low-cost, high-speed, large database storage and the existence of inductive data mining algorithms. However, the knowledge threshold in manufacturing is yet to be attained*, i.e.,* unknown sources of process variation are yet to be discovered which sustains inferior product quality and higher than necessary costs of production (Bernardy and Scherff 1998, 1999).

A key barrier that must be overcome to improve the intellectual latency of forest products companies is aligning real-time process sensor data with product quality data derived from destructive testing. The development of a relational dataset is a key first step in the improvement process. Strength and other product attributes obtained from the lab, yet essential are of a historic temporal nature and limit decisions in the modern realm of continuous processing, high speed throughput, and real-time predictability. This "historic temporal view" of the process from the testing laboratory leads to higher than necessary operating targets, costs, pressing schedule reruns from poor strength properties, and are not sustainable from a business perspective.

Another barrier that must be overcome is poor data quality. Inductive data mining algorithms assume the relational data are of high quality and error-free. Often, induction using algorithms proceeds without the data quality verification. This may result in erroneous analytical conclusions, incorrect manufacturing decisions, and poor business performance. Induction from statistical and heuristic algorithms can lead to data discovery and quantify key relationships between process variance and product variance. Discovery is gained from induction which leads to the evolution of knowledge in manufacturing and business.

INFORMATION PARADOX

Information technology is the largest single capital investment for many enterprises (Thorpe 1998). However, many companies struggle with making use of the vast amount of data that are acquired at increasingly faster rates. Thorpe (1998) called this phenomenon the "information paradox" where companies invest increasing amounts of money on information acquisition but cannot demonstrate a connection between the money spent and business results. This paradox exists from the lack of "real-time relational databases" that are of sufficient design and organization where statistical and heuristic methods can be used for investigating correlation and constructing knowledge. As Harding *et al.* (2006) noted, "*Knowledge is the most valuable asset of a manufacturing enterprise, as it enables a business to differentiate itself from competitors and to compete efficiently and effectively to the best of its ability*." The relational database in many ways is the nucleus or core-reactor for knowledge discovery in manufacturing.

As Inmom and Hackathorn (1994) noted, a data warehouse is the main repository of the organization's historical data, its corporate memory.[1] The central concept of a data warehouse is that it is a collection of records. Data warehouses usually consist of one or more databases of volumes of records. The structure of a database is known as a 'schema.' The schema describes the objects that are represented in the database, and the table relationships among them. Multiple related tables each consisting of rows and columns is the most common form of schema (White 2002). Schema design is a critical factor in ensuring optimum storage and data compression, and to ensuring the overall usefulness of the data during retrieval for analysis.

Real-time Relational Database

A real-time relational database is defined as the alignment of real-time process sensor data with other product quality data (*e.g.,* destructive data developed from the testing laboratory) or elementary distributed data fusion (also called track-to-track fusion) where data from multiple diverse sensors is combined in order to make inferences about a physical event, activity or situation Hall (1992).[2] Intellectual latency is the most significant issue in successfully developing a real-time relational database. Intellectual latency results from improper time alignment of process sensor data with product quality data. Young and Huber (2004) developed an automated relational database for medium density fiberboard (MDF) and oriented strand board (OSB) that successfully overcame some of the issues of intellectual latency. Clapp *et al*. (2007) and Young

[1] The origin of the data warehouse can be traced to studies at MIT in the 1970s which were targeted at developing an optimal technical architecture. At the time, the craft of data processing was evolving into the profession of information management. The MIT work led to the modern concept of the Information Center
http://www.damanconsulting.com/company/articles/dwrealtime.htm 2007.

[2] Data fusion or information fusion are names that have been given to a variety of interrelated expert system problems which have arisen primarily in military applications (Goodman *et al.* 1997). Other applications of data fusion include remote sensing, medical diagnostics and robotics (Blackman and Broida 1990, Hovanessian 1980).

et al. (2013) used the eigenvalues from principal component analysis to identify improper time alignment of real-time process sensors with destructive test data for MDF.

A real-time data warehouse contains just data. The key to is to convert data into knowledge. However, many forest products manufacturers are unsuccessful in converting data into knowledge because the task of developing an automated real-time relational database is not successfully developed. Data mining (DM) is the process of automatically searching large volumes of data for patterns. Data mining is a complex topic that involves many core disciplines such as computer science, statistics, and machine learning (Singh *et al.* 2010).

IMPORTANT STATISTICAL METHODS FOR DATA MINING

An underlying basis of statistical methods is the study of variance (σ^2). In forest products manufacturing, process variance leads to high operational targets (*e.g.,* weight, thickness, density, raw materials inputs, energy use *etc.*) which are non-competitive as a business strategy.

Regression Methods

One of the most traditional and popular statistical methods for data mining methods is multiple linear regression (MLR). For situations where the data are drawn from reasonably homogeneous populations, traditional methods such as MLR can yield insightful analyses. The usefulness of MLR in data mining can breakdown quickly if the stringent assumptions associated with MLR are not met.

Most practitioners use the first-order multiple linear regression (MLR) model:

$$Y = \alpha + \beta X + \varepsilon \quad (1)$$

Where: **Y** is an (n × 1) vector of observations, **X** is an (n × p) matrix of known form, **β** □is an (p× 1) vector of parameters, **ε** □is an (n× 1) vector of errors. From a practical perspective second - and third - order MLR models with interaction terms may be more helpful. The least squares method is common to MLR and is used to find an affine function that best fits a given set of data. The least squares method is very defendable by minimizing the sum of the *n* squared errors (SSE) of the predicted values on the fitted line $(\hat{y}_i)$ and the observed values (y). Kim *et al.* (2012) used ridge regression with time-lagged dependent variables to improve predictions relative to MLR methods.

Regression Trees (RT)

The machine learning technique for inducing a decision tree from data is called decision tree learning, or (colloquially) decision trees. Decision Tree (DT) models have grown into a powerful class of methods for examining complex relationships with many types of data (Kim *et al.* 2007). DT models are more useful than MLR models when data are not homogeneous (Fig. 1). DT methods can be applied either to continuous or categorical data. DT methods for continuous data are often called Regression Trees while DT methods for categorical data are often called Classification Trees (CT). Since “Classification and Regression Trees” (CART) were first introduced, many adaptations and extensions have been proposed.

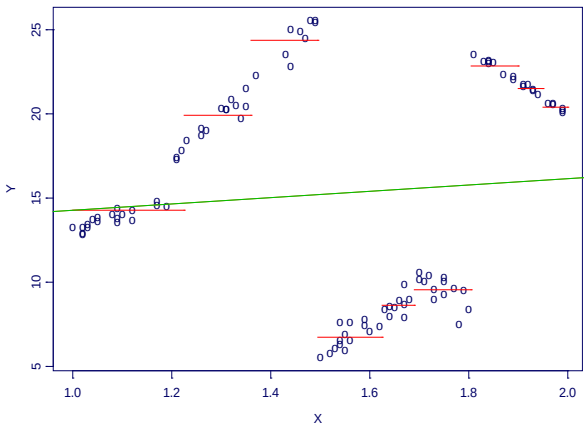

Fig. 1
Illustration of MLR fit and RT fits for non-homogeneous data (Kim et al. 2007).

Kim and Loh (2001) provide insight and comparisons on many DT procedures ranging from the classical CART, C4.5, FACT, CHAID, FIRM, GUIDE, QUEST, to their newly proposed CRUISE (Kim *et al*. 2007). Construction of a regression tree consists of the following three steps performed iteratively, ending with step four:

- Partition the data,
- Fit a model to the data in each partition,

- Stop when the residuals of the model are near zero or a small fraction of observations are left,
- Prune the tree if it over fits.

Most of the new developments in regression trees differ on steps 1 and 2. For example, in step 2, the CART regression tree fits a mean function in each partition (also called a piecewise constant regression tree). Quinlan's (1992) M5 method constructs an ordinary regression tree with a stepwise linear regression model fitted to each node at every stage. As noted in Kim *et al.* (2007), Chaudhuri *et al.* (1994) chose a residual-based approach from MLR models. This approach selects the variable with the signs of the residuals which appear most non-random, as determined by the significance probabilities of two-sample *t*-tests.

The GUIDE algorithm (Loh 2002) extended the idea of Chaudhuri *et al.* (1994) by means of "curvature tests." A curvature test is a chi-square test of association for a two-way contingency table where the rows are determined by the signs of the residuals (positive versus non-positive) from a fitted regression model. The idea is that if a model fits well, its residuals should have little or no association with the values of the predictor variable. As Kim *et al.* (2007) noted, GUIDE has five properties that make it desirable for the analysis and interpretation of large datasets: (1) negligible bias in split variable selection, (2) sensitivity to curvature and local pairwise interactions between predictor variables, (3) applicability to quantitative and categorical variables, (4) choice of multiple or simple linear regression models, and (5) choice of three roles for each numerical predictor variable: split selection only, regression modeling only, or both. Decision trees represent a powerful tool for forest products manufacturers interested in data mining. Decision trees are superior to MLR models when data are non-homogeneous.

Principal Components Analysis

Principal components analysis (PCA) is a powerful and common multivariate statistical method of identifying patterns in data.[3] PCA is also called the (discrete) Karhunen-Loève transform (or KLT) or the Hotelling transform (Hotelling 1933). An advantage of PCA in data mining is its ability to compress large data sets and reduce dimensionality of the data set and identify underlying meaningful variables (Jackson 1991). Principal component analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible (André *et al.* 2008).

A key requirement for PCA to work properly is to standardize the data given that it uses the covariance and variance in the covariance matrix. To standardize the data, the mean of each independent variable is subtracted from each independent value and mean of the response variable is typically subtracted from each response value. This produces a data set with an overall mean of zero. The mathematical technique used in PCA is called Eigen analysis. PCA uses the unit eigenvalues and eigenvectors of the covariance matrix on the standardized data set. In Eigen analysis, we solve for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component and so forth. PCA has a number of different interpretations. The simplest is that it is a projection method finding projections of maximal variability, *i.e.,* it seeks linear combinations of the columns of X with maximal variance (Venables 2002). A key step in PCA is transforming the original data back to its original state. PCA is a common in data mining and multivariate analyses with data sets that have more than one response variable for each record, *e.g.,* IB, Parallel EI, Perpendicular EI and Thickness Swell for OSB. Clapp *et al.* (2007) used PCA to model and predict the IB of MDF. The research by Clapp *et al.* (2007) identified common process variables for different MDF product types that influenced IB variability, *e.g.,* core refiner steam pressure. The study by Clapp *et al.* (2007) was further refined in a study by Young *et al.* (2013); also see Reigler *et al.* (2013).

Quantile Regression

Examining causality between process variables and product quality characteristics beyond the mean of the distribution is an important issue for forest products manufacture. Most forest products manufacturers (especially wood composites manufacturers) have a strong interest in understanding the lower percentiles of the distribution of manufactured product quality. Traditionally, MLR is used to study correlation between independent variables and the average of a response variable, with an important goal of making useful predictions of the response variable. However, there are three important assumptions of this approach: 1) the MLR assumption of linearity; 2) the assumption of a normal or Gaussian distribution for the response variable and 3) MLR models the mean of the distribution of the response variable.

Quantile regression (QR) is intended to offer a comprehensive strategy for completing the regression picture (Koenker 2005). As Mosteller and Tukey (1977) note in their influential text, as cited by Koenker

[3] The method of principal components dates back to Karl Pearson in 1901, although the general procedure we use today was published in Hotelling in 1933 (Hotelling 1933, Jackson 1991).

(2005): "…*the regression curve gives a grand summary for the averages of the distributions corresponding to the set of Xs…and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions*."

Quantile Regression (QR) is an approach that allows us to examine the behavior of the target variable (Y) beyond its average of the Gaussian distribution, *e.g.,* median (50th percentile), 10th percentile, 80th percentile, 95th percentile *etc.* Examining the median and average tendencies of the distribution of a product quality characteristic may yield different conclusions. Young *et al.* (2008) noted that independent variables influencing the response variable of the IB of MDF varied dramatically by quantile. In some cases the sign of the coefficient of similar independent variables explaining IB was reversed by quantile.

The QR model does not require the product quality characteristics to be normally distributed and does not have the other rigid assumptions associated with MLR. The first-order QR model has the form:

$$Q_{y_i}\left\langle \tau \middle| x \right\rangle = \beta_0 + x_i\beta_1 + F_u^{-1}(\tau) \quad (2)$$

where, Q_{yi} is the conditional value of the response variable given τ in the i[th] trial, β_0 is the intercept, β_1 are parameters, τ denotes the quantile, x_i is the value of the independent variable in the i^{th} trial, F_u is the common distribution function (*e.g.,* normal, Weibull, lognormal, other etc.) of the error given τ, $E(F_u^{-1}(\tau)) = 0$, for i = 1 ,…, n (Koenker 2005).

Just as we can define the sample mean as the solution to the problem of minimizing a sum of squared residuals, we can define the median as the solution to the problem of minimizing a sum of absolute residuals (Koenker and Hallock 2001). The symmetry of the piecewise linear absolute value function implies that the minimization of the sum of absolute residuals must equate the number of positive and negative residuals, thus assuring that there are the same number of observations above and below the median (Koenker and Hallock 2001). By minimizing a sum of asymmetrically weighted absolute residuals yields the quantiles (Koenker and Hallock 2001). Solving

$$\min \sum_{i=1}^{n} \rho_\tau(y_i - \xi), \quad (3)$$

where the function $\rho_\tau(\,|\,)$ is a tilted absolute value function that yields the τ^{th} sample quantile as its solution (Koenker and Hallock 2001). To obtain an estimate of the conditional median function in quantile regression, we simply replace the scalar ξ in equation [7] by the parametric function $\xi(x_i, \beta)$ and set τ to ½.[4] To obtain estimates of the other conditional quantile functions, replace absolute values by $\rho_\tau(\,|\,)$ and solve,

$$\hat{\beta}(\tau) = \min \sum_{i=1}^{n} \rho_\tau(y_i - \xi(x_i, \beta)) \quad (4)$$

For any quantile $\tau \in (0,1)$. The quantity $\hat{\beta}(\tau)$ is called the τ^{th} regression quantile.

QR is an important stochastic method for data mining that will greatly surpass MLR in its usefulness for data mining in forest products manufacturing. The forest product industry can benefit greatly from examining the lower percentiles of product quality characteristics, e.g., low strength percentiles.

CONCLUSION

The forest products industries are experiencing unprecedented change and competition. The industries subsist in a co-evolutionary existence with automated data discovery. The successful and competitive forest products companies of this millennium will take advantage of this co-evolutionary existence to discover sources of variation that will lead to lower costs and improved product value.

Real-time data warehousing is common place for this industry but it is used in an antidotal fashion and the industry generally suffers from intellectual latency, *i.e.,* data rich but knowledge poor. The conversion of real-time databases to real-time relational databases is essential for data discovery. The knowledge threshold of sources of process variance is yet to be discovered by this industry. Standard statistical methods used for data mining such as multiple linear regression will evolve to algorithmic combinations of more powerful methods such as decision trees, quantile regression and principal components analysis methods (André and Young 2013).

REFERENCES

André N, Young TM (2013) Real-time process modeling of particleboard manufacture using variable selection and regression methods ensemble. European Journal of Wood and Wood Products (Eur. J. Wood Prod. Holz als Roh- und Werkstoff) DOI 10.1007/s00107-013-0689-0. *Current on-line*

André N, Cho HW, Baek SH, Jeong MK, Young TM (2008) Enhanced prediction of internal bond strength in a medium density fiberboard process using multivariate methods and variable selection. Wood Sci Technol 42:521-534

[4] Variants of this idea were proposed in the mid-eighteenth century by Boscovich and subsequently investigated by Laplace and Edgeworth (Koenker and Hallock 2001).

Bernardy S (1998) Saving costs with process control, engineering and statistical process optimization. Proc. 2nd European Panel Products Symposium (EPPS). Llandudno, Wales.

Bernard S (1999) Process modeling provides on-line quality control and process optimization in particle and fiberboard production. ATR Industrie-Elektronik GmbH&Co.KG – TextilstraBe. D-41751 Viersen Germany. Blackman SS, Broida TJ (1990) Multiple sensor data association and fusion in aerospace applications. Journal of Robotic Systems 7(3):445-485

Chaudhuri P, Huang MC, Loh WY, Yao R (1994) Piecewise-polynomial regression trees Statistica Sinica 4:143-167

Clapp NE Jr, Young TM, Guess FM (2007) Predictive modeling the internal bond of medium density fiberboard using principal component analysis. Forest Products Journal 58(4):49-55

Goodman IR, Mahler RPS, Nguyen HT (1997) Mathematics of data fusion. Kluwer Acadmeic Publishers Norwell MA

Hall D (1992) Mathematical techniques in multisensor data fusion. Artech House Norwood MA

Harding JA, Shahbaz M, Srinivas S, Kusiak A (2006) Data mining in manufacturing: a review. Journal of Manufacturing Science and Engineering 128:969-976

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24:417-441

Hovanessian S (1980) Introduction to synthetic array and imaging radars. Artech House. Norwood, MA

Jackson JE (1991) A user's guide to principal components. John Wiley and Sons. New York, NY

Kim H, Loh WY (2001) Classification trees with unbiased multiway splits. Journal of the American Statistical Association 96:589-604

Kim H, Guess FM, Young TM (2007) Using data mining tools of decision trees in reliability applications. IIE Transactions 43(1):43-54

Kim N, Jeong YS, Jeong MK, Young TM (2012) Kernel ridge regression with lagged dependent variable: applications to prediction of internal bond strength in a medium density fiberboard process. IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews 42(6):1011-1020

Koenker R (2005) Quantile regression. Cambridge University Press New York NY

Koenker R, Hallock, KF (2001) Quantile regression. Journal of Economic Perspective 15(4):143-156

Loh WY (2002) Regression trees with unbiased variable selection and interaction detection. Statistica Sinica 12:361-386

Mosteller F, Tukey J (1977) Data analysis and regression: a second course in statistics. Addison-Wesley. Reading, MA

Quinlan JR (1992) Learning with continuous classes. Proc. 5th Australian Joint Conference on Artificial Intelligence pp. 343–348

Riegler M, Spangl B, Weigl M, Wimmer R, Müller U (2013) Simulation of a real-time process adaptation in the manufacture of high-density fibreboards using multivariate regression analysis and feedforward control Wood Sci Technol DOI 10.1007/s00226-013-0571-6

Singh M, Guess FM, Young TM, Liu L (2010) Modeling randomness using system dynamics concepts. International Encyclopedia of Statistical Sciences Part 13 839-841. Editor: Dr. Miodrag Lovric. Springer. doi 10.1007/978-3-642-04898-2_37

Thorpe J (1998) The information paradox. McGraw-Hill Toronto

Venables WN (2002) Modern applied statistics with S. Springer-Verlag New York NY

White C (2002) Intelligent business strategies: real-time data warehousing heats up. DM Review Magazine (August)

Young TM, Huber CW (2004) Predictive modeling of the physical properties of wood composites using genetic algorithms with considerations for distributed data fusion. Proc. of the 38th International Particleboard/Composite Materials Symposium. Washington State University, Pullman, WA. pp.145-153

Young TM, Clapp NE Jr, Guess FM, Chen CH (2013) Predicting key reliability response with limited response data. Quality Engineering *In Press*

Young TM, Shaffer LB, Guess FM, Bensmail H, León RV (2008) A comparison of multiple linear regression and quantile regression for modeling the internal bond of medium density fiberboard. Forest Products Journal 58(4):39-48