International Journal of Reliability and Applications
Vol. 3, No. 4, pp. 155-164, 2002

Mining Information in Automated Relational Databases for Improving Reliability in Forest Products Manufacturing

Timothy M. Young*
Tennessee Forest Products Center, 2506 Jacob Drive,
University of Tennessee, Knoxville, TN 37996-4570 USA

Frank M. Guess
Department of Statistics, 338 Stokely Management Center,
University of Tennessee, Knoxville, TN 37996-0532 USA

Abstract. This paper focuses on how modern data mining can be integrated with real-time relational databases and commercial data warehouses to improve reliability in real-time. An important issue for many manufacturers is the development of relational databases that link key product attributes with real-time process parameters. Helpful data for key product attributes in manufacturing may be derived from destructive reliability testing. Destructive samples are taken at periodic time intervals during manufacturing, which might create a long time-gap between key product attributes and real-time process data. A case study is briefly summarized for the medium density fiberboard (MDF) industry. MDF is a wood composite that is used extensively by the home building and furniture manufacturing industries around the world. The cost of unacceptable MDF was as large as 5% to 10% of total manufacturing costs. Prevention can result in millions of US dollars saved by using better information systems.

Key Words : *improving reliability, automated relational database, data warehouse, destructive reliability tests, medium density fiberboard, forest products, information quality.*

1. INTRODUCTION

Data Mining and Knowledge Discovery have become active areas of research attracting people from many different disciplines, e.g., computer science, industrial engineering, statistics, operations research, etc. Data mining toolkits are now

---

* Corresponding Author
E-mail address: tmyoung1@utk.edu

commercially available, and such toolkits are emerging more in industrial applications. Many organizations have been labeled "data-rich" and "knowledge-poor" (Chen 2001). Data mining enables complex manufacturing processes to be better understood by examining the patterns in data related to the previous behavior of a manufacturing process (Chen 2001).

This paper addresses the challenge of how data mining can be integrated with real-time relational databases and commercial data warehouses to improve reliability in real time. We believe that for data mining to be more widely adopted by the forest product industry, and other industries globally, data mining toolkits must be developed in the context of real-time relational databases, using affordable commercial data warehouses. The cost benefits, plus increased product reliability, from data mining techniques may be highly significant and offer improved, stronger competitiveness for companies in any industrial sector.

An important issue for many manufacturers is the development of relational databases that link key product attributes with real-time process parameters (e.g., Yang 1986). Needed data for key product attributes in forest product manufacturing may be derived from destructive reliability testing. Most destructive samples are taken at periodic time intervals during manufacturing which creates a time-gap between key product attributes and real-time process data. This time-gap may hinder the real-time decision-making capabilities of operators.

A case study is briefly summarized for the medium density fiberboard (MDF) industry. MDF is a wood composite that is used extensively by the home building and furniture manufacturing industries around the world. The objective of the case study was to develop an automated relational database from which statistical data mining methods could be used to better predict the internal bond (strength) of MDF and improve reliability. The cost of unacceptable MDF was as large as 5% to 10% of total manufacturing costs. Better information systems can prevent unacceptable reliability and result in millions of US dollars saved.

2. METHODS

2.1 Mass Data Storage

Wonderware© Industrial SQL Server 7.1 was used as the data warehouse of process data. Other software packages could be used. This was picked due to its relatively lower cost; thus, liker adoption by forest product manufactures concerned with cost factors and international competition. A smaller subset of 230 process variables was collected from the data warehouse of approximately 2,850 process variables.

All process data are stored in the data warehouse based on "delta" change, i.e., any change in the process variable is written to the data warehouse based on leading-edge detection. All process data were stored on a PC server, which was separate from the PC server that stored destructive test data.

2.2 Automated Relational Database

The relational database was the Cartesian product of two sets S_1 and S_2, consisting of ordered pairs (a, e_j) of a in S_1 and e_j in S_2, i.e.,

$$S_1 \times S_2 = \{(a, e_j) \mid a \in S_1 \text{ and } e_j \in S_2, 1 \leq j \leq 230\} \quad (1)$$

where S_1 is the destructive test element a, namely internal bond (in pound per square inches: p.s.i.). S_2 are the process elements e_j where $1 \leq j \leq 230$. Examples of process data elements e_j were: fiber-mat moisture, fiber-mat weight, line speed, press position-1 temperature, press position-1 pressure, product type, MDF thickness, etc. The relation in the Cartesian product of the two sets was the time-stamp associated with the event of a destructive sample. All data were time-ordered using the time-stamp of the destructive test event. Partitions of the Cartesian product of the two sets were developed based on product type. Product types were based on MDF thickness (inches), panel width (inches), panel length (inches), and panel density (lbs/ft^3). An example of the relational databases is illustrated in Figure 1.

E8 = 232.149993896484

	A	B	C	D	E	F	G	H	I	J	K
1	DateTime	EventLogKey	Internal_Bond	FIC3224SP	FT3224	Mat_Density	PO_LENGTH	PO_DENSITY	PO_THICK	PO_WIDTH	TARGET_W
2	3/19/02 3:35 PM	30622	112.3999939	13.30000019	276.75	3.485566854	220	42	0.75	61	
3	3/19/02 5:04 PM	30623	188.3999939	16	346.1499939	3.507122278	293	48	0.75	61	
4	3/19/02 8:13 PM	30625	168.8999939	16	357.9500122	3.599458218	293	48	0.75	61	
5	3/19/02 10:33 PM	30626	186.6000061	16	326.6499939	3.524864435	293	48	0.75	61	
6	3/20/02 1:15 AM	30627	171.3999939	15.5	346.7000122	3.469734669	293	48	0.75	61	
7	3/20/02 4:51 AM	30628	190.8999939	15.5	335.5499878	3.482540131	293	48	0.688000023	61	
8	3/20/02 6:25 AM	30629	131.6999969	11	232.1499939	3.305812359	254	46	0.688000023	61	
9	3/20/02 9:27 AM	30631	198.7999878	16	342.2999878	3.432584286	220	48	0.625	61	
10	3/20/02 12:11 PM	30632	196.8000031	16	364.5499878	3.416193247	220	48	0.625	61	
11	3/20/02 2:01 PM	30633	199.5	16	355.8999939	3.395559072	220	48	0.625	61	
12	3/20/02 4:00 PM	30634	129	10.5	230.3999939	3.128822088	258	46	0.625	61	
13	3/20/02 6:28 PM	30635	120.3999939	10.5	239	3.125454664	258	46	0.625	61	
14	3/20/02 8:40 PM	30637	124.3000031	10.5	231.3999939	3.077786684	244	46	0.625	61	
15	3/20/02 11:17 PM	30638	132.6999969	10.5	243.3500061	3.096590281	244	46	0.625	61	
16	3/21/02 2:20 AM	30639	122.6999969	10.5	234.9499969	3.121264458	293	46	0.625	61	
17	3/21/02 5:23 AM	30640	127.1999969	10	230.9499969	3.08416748	221	46	0.625	61	
18	3/21/02 8:24 AM	30642	134.1999969	10	235.8500061	3.128793716	221	46	0.625	61	
19	3/21/02 11:34 AM	30643	108.1999969	10	230.9499969	3.030071974	220	46	0.625	61	
20	3/21/02 2:34 PM	30644	129.5	10.5	233.8500061	3.01809597	220	46	0.625	61	
21	3/21/02 5:12 PM	30645	127.5999985	10.5	230.0600031	3.005338192	220	46	0.625	61	
22	3/21/02 8:03 PM	30647	117.5	10	230.5500031	3.021934509	220	46	0.625	61	
23	3/21/02 10:34 PM	30648	119.3000031	10.5	236.3000031	2.979324579	292	46	0.625	61	
24	3/22/02 1:42 AM	30649	137.2000122	10.5	238.8500061	3.03383708	292	46	0.625	61	
25	3/22/02 4:59 AM	30650	198.1000061	15.80000019	356.8500061	3.170253515	293	48	0.625	61	
26	3/22/02 7:30 AM	30652	190.3999939	15.5	361.8999939	3.263065338	293	48	0.625	61	
27	3/22/02 10:06 AM	30653	173.2000122	15.30000019	295.75	3.347020864	293	48	0.563000023	61	
28	3/22/02 12:19 PM	30654	137	10.5	212.6499939	3.179245949	220	48	0.5	61	
29	3/22/02 4:03 PM	30655	124.0999985	10.69999981	228.1000061	2.949615717	293	48	0.375	61	
30	3/22/02 11:01 PM	30657	123.9000015	10.80000019	158.25	2.999330044	293	48	0.5	49	
31	3/23/02 12:57 AM	30658	137.2000122	10.80000019	196.1499939	3.037169456	293	46	0.625	49	
32	3/23/02 3:14 AM	30659	85.80000305	15	216.6499939	3.074455738	293	40	0.75	49	
33	3/23/02 4:14 AM	30660	102.1999969	15.5	220.75	3.018988848	293	40	0.75	49	
34	3/23/02 5:28 AM	30661	126.8000031	11.80000019	209.3999939	3.047281265	293	46	0.75	49	

InSQLPivotData7142002003

Figure 1. Example of 34 records and 10 variables in relational database.

The creation and updating of the relational database was automated. The destructive database and process data were on separate PC servers on a common LAN. Creation and updating were performed using Microsoft SQL 7.0 encoding and Microsoft SQL 7.0 automated functionality, e.g., DTS Package and SQL Enterprise Manager "Jobs." The Microsoft SQL 7.0 DTS Package was used to link PC servers. Microsoft SQL 7.0 SQL Enterprise Manager "Jobs" were used to schedule and execute Transact SQL encoding. The Transact SQL encoding was used to create the relational database in a Microsoft SQL 7.0 table structure. The Microsoft SQL table was updated within 15 minutes after the destructive tests. The destructive test results were merged with a temporary Microsoft

SQL Table that contained data for 230 process variables corresponding to the time-stamp from the most recent destructive test.

2.3 Data Quality

Current commercial relational database management systems such as Microsoft SQL Server 7.0 and their underlying relational model are based on the assumption that data stored in the databases are correct (Ballou et al. 1998, Wang et al. 2001). The assumption that the join operation in the SQL encoding queries produced correct data was validated using the Attribute-Based Model (Wang et al. 2001). The Attribute-Based Model was used to facilitate cell-level tagging of data quality. Integrity rules were used during the SQL query process which relied on quality indicators characteristic of the data, e.g., cells with null values noted, internal bond greater than zero and less than 300 p.s.i., line speed greater than zero and less than 150 ft./min., fiber mat moisture great than zero and less than 20%, etc. An important issue related to data quality was the synchronizing of time clocks on the destructive testing PC server and the process data server. The servers were checked everyday at 12:00 a.m. for proper time synchronization.

2.4 Predictive Modeling

The linear regression model is of the form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where $\mathbf{Y}$ is an $(n \times 1)$ vector of observations, $\mathbf{X}$ is an $(n \times p)$ matrix of known form, $\boldsymbol{\beta}$ is an $(p \times 1)$ vector of parameters, $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of errors.

A pre-selected F_{OUT} of $\alpha = 0.05$ was use as a critical value (Myers 1989). Forward selection multiple linear regression methods were used to develop the models, but backward and mixed yielded the same final "best" models. The coefficient of determinations (R^2) ranged from 0.77 to 0.97. The predictive models were developed in the spirit of maximizing R^2_a (adjusted R^2), maximum R^2 subject to the principle of parsimony (i.e., fewer predictor variables being used as possible), minimum mean square residual, VIFs < 10 (Variance Inflation Factor), no pattern in residuals, minimum PRESS (Prediction Sum of Squares), Mallow's $C_p \approx p$, and residual plots with homogeneous variance (Draper and Smith 1981). This approach produced very practical, easily implemented first order approximations. Predictive models were not possible for all product types in the case study. However, predictive models were developed for product types that comprised about 65% of the producer's annual production. See the next section for more details.

3. EXPLORATORY CASE STUDY AND RESULTS

The case study was conducted at one of the larger producers of medium density fiberboard (MDF) with a capacity in excess of 100 million ft^2. The producer has a continuous press which has a wide variety of MDF products ranging in thickness from ½" to 1" and densities from 40 lbs/ft^3 to 48 lbs/ft^3. The producer is considered an industry leader in quality and productivity.

Multiple linear regression models were developed for three products that represented 65% of the manufacturer's annual production. A multiple linear regressions for 5/8" thick MDF is as follows (Figure 2):

Y = Internal Bond = 706.10 (3)
- 12.40 (Actual Position Press Frame 14 Left Side)
+ 0.51 (Actual Position Press Frame 03 Left Side)
- 1.02 (Inside Mat Temperature)
- 1.00 (Actual Position Press Frame 19 Left Side)
+ 0.30 (Resin Flow)
- 0.75 (Actual Position Press Frame 05 Right Side)
- 10.59 (Moisture Content at Forming Out-feed)
+ 0.12 (Pre-Press Outlet Right Pressure).

where, $R_a^2 = 0.96$, $R^2 = 0.97$, PRESS = 220.91, Root MSE = 2.36, F = 60.48 and n = 27.

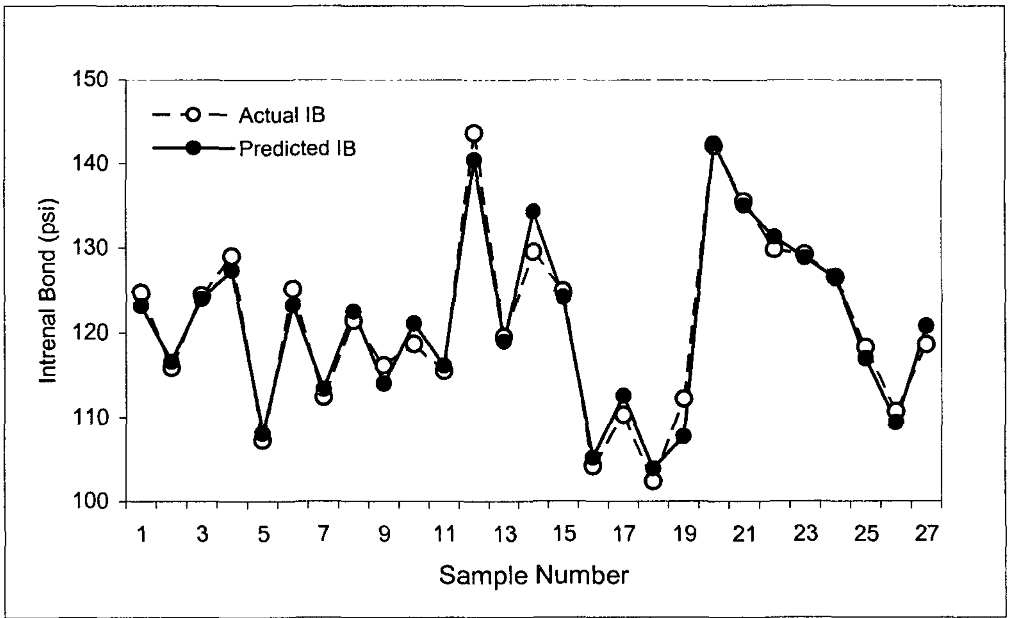

Figure 2. Actual and predicted internal bond for model in equation (3)

Summary of fit and analysis of variance statistics are presented in Table 1. The parameter estimates, standard errors, t-ratios, p-values and VIF statistics are presented in Table 2.

Table 1. Summary of fit and analysis of variance statistics for equation (3)

R-square		0.9661		
R-square adjusted		0.9550		
Root Mean Square Error		2.3623		
Mean of Response		121.25		
Observations		27		
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	2700.3755	337.547	60.4862
Error	17	94.8696	5.581	Prob > F
Total	25	2795.2451		< 0.0001

Table 2. Parameter estimates for equation (3)

Term	Estimate	Std Error	T ratio	Prob>|t|	VIF
Intercept	706.10	52.19	13.53	< 0.0001	--
Actual Position Press Frame 03 Left Side	0.51	0.07	7.77	< 0.0001	1.90
Inside Mat Temperature	-1.02	0.14	-7.33	< 0.0001	1.20
Actual Position Press Frame 19 Left Side	-1.00	0.15	-6.59	< 0.0001	3.07
Resin Flow	0.30	0.05	5.49	< 0.0001	1.72
Actual Position Press Frame 05 Right Side	-0.76	0.22	-3.37	0.0036	1.33
Moisture Content at Forming Out-feed	-10.59	4.04	-2.62	0.0179	1.69
Pre-Press Outlet Right Pressure	0.12	0.07	1.76	0.0965	1.86

Residuals for equation (3) are presented in Figure 3. There was no apparent systematic pattern in the residuals for equation (3). The predicted values for internal bond for test validation values are presented in Figure 4. The predictions of test validation values indicated that the model in equation (3) is helpful for predicting the internal bond for MDF based on the listed terms.

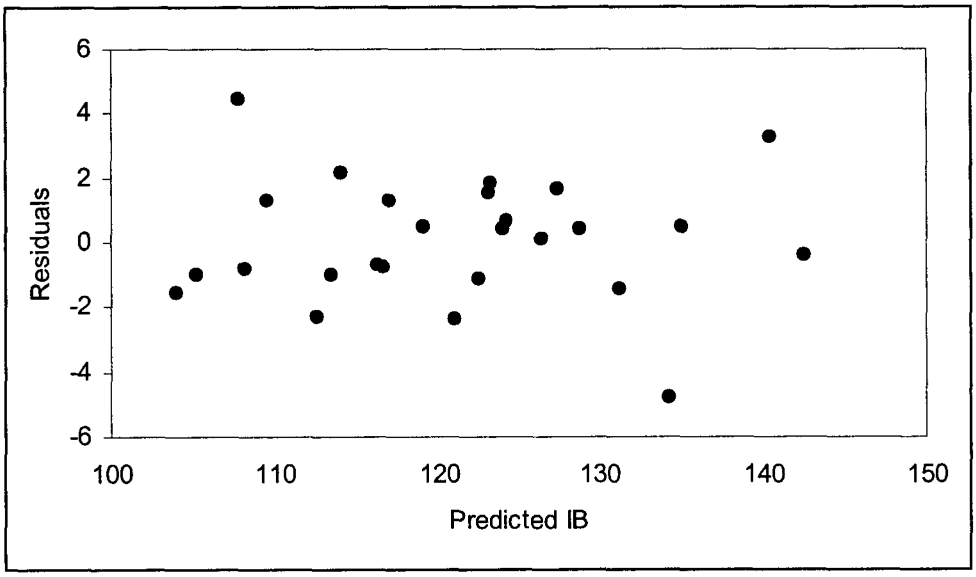

Figure 3. Residuals for equation (3)

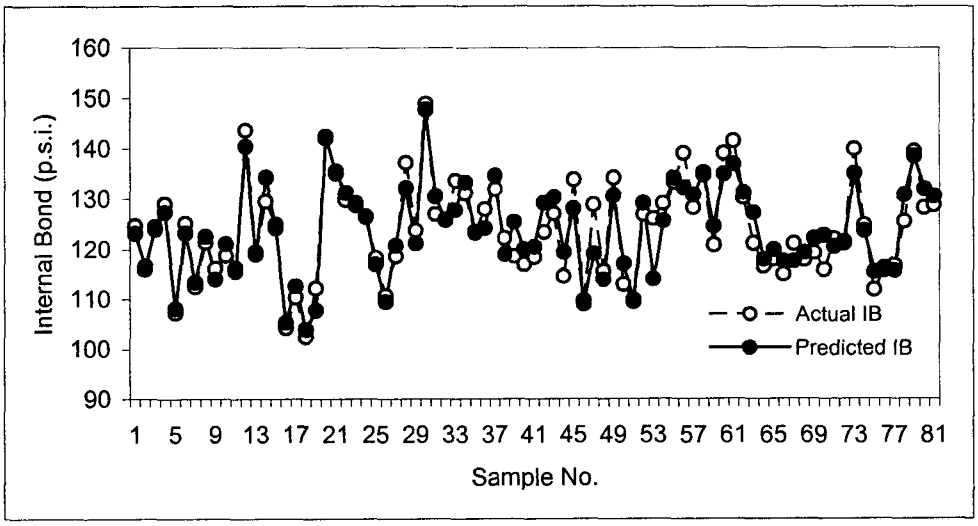

Figure 4. Predicted internal bond for data set and test values for equation (3)

Benefits of these first order predictive models for MDF manufacture are preventing manufacturing defects, minimizing manufacturing costs and maximizing throughput. The key and most obvious benefit of a predictive model is its ability to prevent a future failing internal bond, i.e., an internal bond below the customer's minimum specification. Given the two to three hour time period between destructive tests in most MDF manufacturing situations, a substantial amount of failing MDF production may be avoided if a predictive model can predict the failures. The second benefit of a predictive model may be realized by raw material minimization and process optimization. If a predictive model can accurately predict internal bond, raw material inputs in the manufacture of MDF may be minimized. Examples of raw materials in MDF manufacture that are significant manufacturing costs are wood fiber and resin (Suchsland and Woodson 1986, Maloney 1993). Process optimization may also be realized from the use of a predictive model by maximizing process throughput. Line speed is directly related to MDF throughput. Predicting internal bond can allow machine operators to maximize line speed.

The statistically significant variables in a predictive model for the internal bond of MDF may be important starting points for implementing better statistical process control (SPC). One of the most important decisions in implementing SPC is the selection of important process variables that may influence key product attributes (Deming 1986, 1993). Predictive modeling of the key product attribute of the internal bond of MDF may indicate which "critically-few" process variables need to be analyzed and monitored in a SPC framework. Reduction in special-cause and common cause variation of the "critically-few" process variables may reduce variation in the internal bond of MDF. Long-term reduction in the internal bond of MDF may result in future process and product optimization.

4. SUMMARY

A data warehouse was developed for 230 process variables. The automated relational database was developed by linking the Microsoft© SQL 7.0 lab database with Wonderware© Industrial SQL 7.1 process data warehouse. Transact SQL encoding with Microsoft SQL DTS and automated JOBS were used to automate the updating of the database. The relational database was updated automatically when a destructive test was completed. The crucial product attribute was the internal bond of MDF measured in p.s.i.

An exploratory case study was conducted on MDF from a large international manufacturing facility. The purpose of the case study was to develop an automated relational database using commercial software and to develop first order approximate predictive models for the internal bond of MDF from the relational database using data mining methods.

Three predictive models were developed for three distinct MDF products. We have presented a sample of one of these models and related analysis in this paper. Forward selection multiple linear regression methods were used to develop the models, but backward and mixed yielded the same final "best" models. The coefficient of determinations (R^2) ranged from 0.77 to 0.97. The predictive models were developed in the spirit of maximizing R^2_a (adjusted R^2), maximum R^2 subject to the principle of parsimony (i.e., fewer predictor variables being used as possible), minimum mean square residual, VIFs < 10 (Variance Inflation Factor), no pattern in residuals, minimum PRESS (Prediction Sum of Squares), Mallow's Cp $\approx$ p, and residual plots with homogeneous variance. This approach produced very practical, easily implemented first order approximations. In another paper, we plan to discuss some improvements using other techniques. Predicting values of the internal bond of MDF validated these first order models and their important usefulness in real time manufacturing. For two of the three models there was no significant increase in the variance of the residuals for test validation predicted values. The model with the lowest R^2 had the largest variance in residuals for test validation predicted values.

The ability to predict the internal bond of MDF helps prevent the manufacture of defective product. Prediction of internal bond may also lead to process optimization by minimizing raw material inputs and maximizing production throughput. An additional benefit of predictive modeling may be in the successful implementation of better statistical process control and continuous improvements strategies.

ACKNOWLEDGEMENT

We both gratefully acknowledge funding from the United States Department of Agriculture in 2002 calendar year. Dr. Frank M. Guess appreciates funding from the competitive Scholarly Research Grant Program of the College of Business Administration at the University of Tennessee during academic year 2002-2003 and the Department of Statistics at the University of Tennessee. We, also, thank students Tim Pickrell, Alicia Christman, and David McGinnis for helpful proof reading of our final draft before submission.

REFERENCES

Ballou, D. P., Wang, R. Y., Pazer, H. and Tayi, G. K. (1998). Modeling information manufacturing systems to determine information product quality, *Management Science,* **44** (4), 462-484.

Chen, Z. (2001). *Data Mining and Uncertain Reasoning – An Integrated Approach,* John Wiley and Sons, Inc., New York, NY.

Deming, W. E. (1986). *Out Of The Crisis,* Massachusetts Institute of Technology's Center for Advanced Engineering Study, Cambridge, MA.

Deming, W. E. (1993). *The New Economics,* Massachusetts Institute of Technology's Center for Advanced Engineering Study, Cambridge, MA.

Draper, N. and Smith, H. (1981). *Applied Regression,* second edition, John Wiley and Sons, Inc., New York, NY.

Maloney, T. M. (1993). *Modern Particleboard and Dry-Process Fiberboard Manufacturing,* Miller Freeman Inc., San Francisco, CA.

Myers, R. H. (1989). *Classical and Modern Regression with Applications,* PWS-Kent Publishing Company, Boston, MA.

Suchsland, O. and Woodson, G. E. (1986). *Fiberboard Manufacturing Practices in the United States,* U.S. Department of Agriculture Forest Service's Agriculture Handbook No. 640, Washington, D.C.

Wang, R. C., Ziad, M. and Lee, Y.W. (2001). *Data Quality,* Kluwer Academic Publishers Norwell, MA.

Yang, C. C. (1986). *Relational Databases,* Prentice-Hall, Englewood Cliffs, NJ.