IDENTIFICATION OF THE RELATIONSHIP BETWEEN EQUILIBRIUM MOISTURE CONTENT, DRY BULB TEMPERATURE, AND RELATIVE HUMIDITY USING REGRESSION ANALYSIS

Charles D. Ray†

Assistant Professor
205 Forest Resources Building
School of Forest Resources
The Pennsylvania State University
University Park, PA 16802

Neelesh Gattani

Sr. Systems Analyst
Planmatics, Inc.
1375 Piccard Drive, Suite 225
Rockville, MD 20850

Enrique del Castillo

Professor
357 Leonhard Building
The Harold and Inge Marcus Department of Industrial Engineering
The Pennsylvania State University
University Park, PA 16802

and

Paul R. Blankenhorn†

Professor
201 Forest Resources Building
School of Forest Resources
The Pennsylvania State University
University Park, PA 16802

(Received September 2005)

ABSTRACT

This paper evaluates the performance of equilibrium moisture content (EMC) predictions using the least squares regression equation given in The Dry Kiln Operator's Manual (Simpson 1991). The fit of the regression equation in The Manual was found to be adequate only when the dry bulb temperature is below 110°F. At temperatures above 110°F, it generally overestimates EMC. A new polynomial regression equation is presented in this paper to predict EMC at dry bulb temperatures above 110°F. Comparisons between the old and new regression equations show an improvement in the root mean squared error of the predictions of about 44% when using the new equation. The proposed equation facilitates better control of the drying process in computer-controlled kiln applications using prediction equations for EMC estimates.

Keywords: Regression, EMC prediction, dry kiln control.

† Member of SWST.

INTRODUCTION

The most important and most widely used method of kiln control in drying hardwoods is to keep the dry bulb temperature (T_{db}) and relative humidity (RH) as close as possible to the specified value in hardwood kiln schedules. At any particular constant setting of T_{db} and RH, wood eventually reaches a stable state where the moisture content (MC) is constant and the wood neither loses nor gains moisture. This is the equilibrium moisture content (EMC). Equilibrium moisture content (EMC) is one of the critically important factors associated with kiln-drying. It has been demonstrated that effective control of EMC can lead to significant reduction in kiln-drying time and moisture content variability over the drying period (Gattani et al. 2005), potentially creating tremendous value to the drying operation through increased kiln throughput. In kiln-drying, the surface moisture content of green wood quickly approaches the EMC conditions in the kiln atmosphere, setting up a moisture gradient between the exterior and the interior of the wet wood. As the EMC conditions inside the kiln are changed, with each T_{db} and RH change in the kiln schedule, the surface MC is also changed, promoting moisture movement and reducing the average MC of the wood. EMC conditions inside the kiln are essential to establish the moisture gradient in wood. If the wet wood is allowed to remain in these EMC conditions for a long period of time, the average MC of the wood will continue to approach the EMC value for the kiln atmosphere. The empirical tables of the relationship between EMC, relative humidity, and dry bulb temperature are given in the dry kiln operator's manual; this publication also provides a least-squares regression equation that can be used to derive EMC for a given RH and dry bulb temperature (Simpson 1991).

With the advent of computer-controlled drying, statistical equations are becoming more critical in providing the opportunity to calculate and interpolate the EMC values for various RH and T_{db} values not in the table (Simpson 1991). Various automatic and semi-automatic control systems are available commercially that control the various physical parameters of the kiln in order to keep kiln conditions as close as possible to targeted set points.

Multiple regression is a very well-known statistical method to represent the relationship mathematically between a set of k independent variables, also called predictor variables, and a dependent variable or criterion variable (for example, see Draper and Smith 1998, Montgomery 2001, and Neter at al. 1985).

A simple multiple linear regression equation is of the form:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots . \beta_k x_k + \varepsilon \quad (1)$$

where y is the dependent variable or response, the x_i are the independent variables or regressors, and ε denotes a random error used to model other sources of variability in the response not explained by changes in the regressors. The errors are usually assumed to be Normal $(0, \sigma^2)$ i.i.d. random variables.

A least squares estimation technique is usually employed to estimate the coefficients or the parameters of the model. It is achieved by minimizing the sum of squared errors (SS_E) of the observed values for the dependent variable from those predicted by the model.

The least squares estimators of the parameters β_i are obtained by minimizing SS_E with respect to all β_i:

$$SS_E = \Sigma(y_i - \hat{y}_i)^2 \quad (2)$$

where $\hat{y}_i$ = predicted value of y from the fitted model.

If the response is suspected to be curvilinear in nature, polynomial regression and its variations can be used to represent the function. Any polynomial regression model with interactions can also be modeled using least squares estimation technique. A polynomial regression model can be represented as:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \quad (3)$$

which is still linear in the parameters.

The polynomial model can be easily fitted by transforming the variables $x_1^2 = x_3$, $x_2^2 = x_4$, $x_1 x_2 = x_5$, etc., using the techniques to fit a

multiple linear regression model. One of the potential drawbacks of polynomial regression can be the extrapolation, especially with higher order terms in the model.

Matrix algebra for multiple regression

Suppose we have collected n observations of the response at different conditions of independent variables. Let

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ .. \\ .. \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & .. & x_{1k} \\ 1 & x_{21} & x_{22} & .. & x_{2k} \\ 1 & .. & .. & .. & .. \\ 1 & .. & .. & .. & .. \\ 1 & .. & .. & .. & .. \\ 1 & x_{n1} & .. & .. & x_{nk} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ .. \\ .. \\ \beta_k \end{bmatrix} \quad \text{and} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ .. \\ .. \\ \varepsilon_n \end{bmatrix} \tag{4}$$

Thus, in matrix notation, the regression model can be represented as

$$y = X\beta + \varepsilon \tag{5}$$

To obtain the least squares estimators $\hat{\beta}$, we need to minimize the sum of squared errors (SS_E), which in vector notation, is given by the equation:

$$SS_E = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta) \tag{6}$$

The function is convex; thus after minimizing the above function by differentiating it with respect to parameters β, we get the following ordinary least squares (OLS) parameter estimates of the model:

$$\hat{\beta} = (X'X)^{-1} X'y \tag{7}$$

Residual diagnostics

Several statistics and plots are utilized to test the adequacy of the regression model fitted. The explanation given here is not meant to be exhaustive, yet it will provide the reader a brief overview of residual diagnostics.

While building a regression model, various assumptions are made. Residual diagnostics is a procedure to test the validity of those assumptions. Minor violations may have little effect on the efficacy of the model, but major departures can seriously hamper the model's performance.

Some of the important statistics used in statistical literature are explained below:

1) *R^2 Statistic*: R^2 is also referred to as coefficient of determination. It is the ratio of variability explained by the model and the total variability of the observations. Its value can be between 0 and 1. A high value of R^2 statistic is desirable. Mathematically, it can be defined as:

$$R^2 = 1 - \frac{SS_{ERROR}}{SS_{TOTAL}} \tag{8}$$

2) *Adjusted R^2 Statistic*: One problem with the R^2 statistic is that it can be misleading when comparing models with different number of parameters. Adding a new parameter to a model always increases R^2 without consideration for statistical significance.. To obviate this problem, the adjusted R^2 statistic is used. Adjusted R^2 does not increase if an insignificant variable is added; on the contrary, it often decreases. Algebraically,

$$\text{Adj } R^2 = 1 - \frac{SS_{ERROR}/df_{ERROR}}{SS_{TOTAL}/df_{TOTAL}} \tag{9}$$

3) *PRESS (Prediction Sum of Squares):* PRESS is defined as the sum of squares of residuals for each observation resulting from dropping out that observation and predicting it on the basis of all other observations.

$$\text{Predicted Residual} = \text{PRESID}_i = \frac{\text{RESID}_i}{1 - h_i}$$

$$\text{PRESS} = \sum_{i=1}^{n} \text{PRESID}_i^2 \tag{10}$$

where

$$h_i = x_i(X'\ X)^{-1}\ x_i'$$

Assumptions of regression modeling

1. *Normality assumption*: If the model fit is adequate, then the residuals should be normally distributed across the mean 0.
2. *Non-constant variance*: Observations or data points should have the same underlying average squared error, or *variance*, for the regression model to be valid. If the residuals or errors have unequal variance for different values of *X*'s, the assumption of constant variance is violated.
3. *Independence*: If the errors are independent, there should not exhibit pattern or structure. If the residuals depart from 0 in systematic fashion, assumption of independence is violated.

Various procedures and methods are available in the statistical literature to eliminate the violations. One such method is transformation of the dependent variables. The various assumptions of regression can be assessed with different plots and numerical methods.

Multicollinearity

Multicollinearity is also one of the profound problems found in polynomial regression. Multicollinearity arises due to a lack of independent variation in the predictor variables (*X*'s). For regression models, the addition of more explanatory variables does not necessarily increase the goodness-of-fit of the regression. In fact, linear combinations of the existing variables may approximately predict the added variable and hence the added variable provides little new information. Multicollinearity has to be considered when the modeler is interested in relative importance of effects of predictor variables on dependent variable and the magnitude of effect of predictor variable.

There are various ways in which the effects of multicollinearity can be reduced. Centering is one of the procedures used in conjunction with polynomial regression to reduce its effects. When predictor variables are centered, it implies that *X* is represented by deviations across its mean *X*. This is done by obtaining a new set of predictor variables by subtracting the mean ($\overline{X_{org}}$) from all the X_{org}. When multicollinearity exists in a data set, an estimation method other than OLS is suggested, such as ridge regression. For more details about ridge regression and other methods, please refer to Hoerl and Kennard (1970).

STATEMENT OF THE PROBLEM

There are a total of 826 values of RH, EMC, and T_{db} in Table 1–6 given in the Dry Kiln Operators Manual (Simpson 1991). The observations were tabulated for dry bulb temperatures of 60°F, 65°F and so on. The problem presented to the research team was the degree of error associated with the published approximation of these 826 observations. The equation given in Simpson (1991), page 40, is an ordinary least squares regression fit of the data in the table, and yields:

$$\mathrm{EMC} = \frac{1800}{W}\left[\frac{kh}{1-kh} + \frac{k_1kh + 2k_1k_2k^2h^2}{1 + k_1kh + k_1k_2k^2h^2}\right] \tag{11}$$

where EMC is equilibrium moisture content (percent), *h* relative vapor pressure, and

$$W = 330 + 0.452\mathrm{T_{db}} + 0.00415\mathrm{T_{db}^2}$$

$$k = 0.791 + 0.000463\mathrm{T_{db}} - 0.000000844\mathrm{T_{db}^2}$$

$$k_1 = 6.34 + 0.000775\mathrm{T_{db}} - 0.0000935\mathrm{T_{db}^2}$$

$$k_2 = 1.09 + 0.0284\mathrm{T_{db}} - 0.0000904\mathrm{T_{db}^2}$$

This model has 12 parameters. If the fit of this equation is good, then the calculated error or *residual* between the actual and predicted value of EMC should be small, with a mean of zero and a constant and small variance. An additional regression assumption used when the model is used in statistical inference (e.g., in building confidence intervals for the predictions), is normality of the distribution of the errors.

For each of the 826 values of the regressors, the error between actual EMC (as determined from Table 1–6 of Simpson 1991) and predicted

values (as calculated from Eq. 11) was calculated. Figure 1 shows a plot of the error vs. the dry bulb temperature. These errors represent the same error as illustrated in the example given in Simpson (1991), page 40, last paragraph.

Figure 1 clearly suggests that Eq. (11) violates the assumption of independence and zero mean of the errors. There is a systematic pattern in the errors and they are not normally distributed. From Fig. 1, the fit of the equation as seen from the graph is adequate only while dry bulb temperature values are below 110°F. Above 110°F, the regression equation predicts a higher estimate of EMC than the empirical values. The differences in actual and predicted value of EMC are as large as 1.5 under (positive error) and 1.0 over (negative error) actual EMC.

OBJECTIVE

The objective of this study is to specify a more accurate, precise, and therefore more useful general model for prediction of EMC, and demonstrate its performance versus the Simpson 1991 model over the entire range of normal hardwood drying practice.

METHODOLOGY

Since the published Eq. (11) is shown to be ineffective for accurate predictions of EMC above 110°F, the objective of developing a new equation that would demonstrate significantly better fit above 110°F was undertaken. The SAS statistical software package was used to perform the analysis. The following steps summarize our methodology:

1. A polynomial model was developed to adequately describe and capture the complex relationship between EMC, RH, and T_{db}. Only the 432 entries for which dry bulb temperature is greater than 110°F were used.
2. After fitting the specified polynomial model

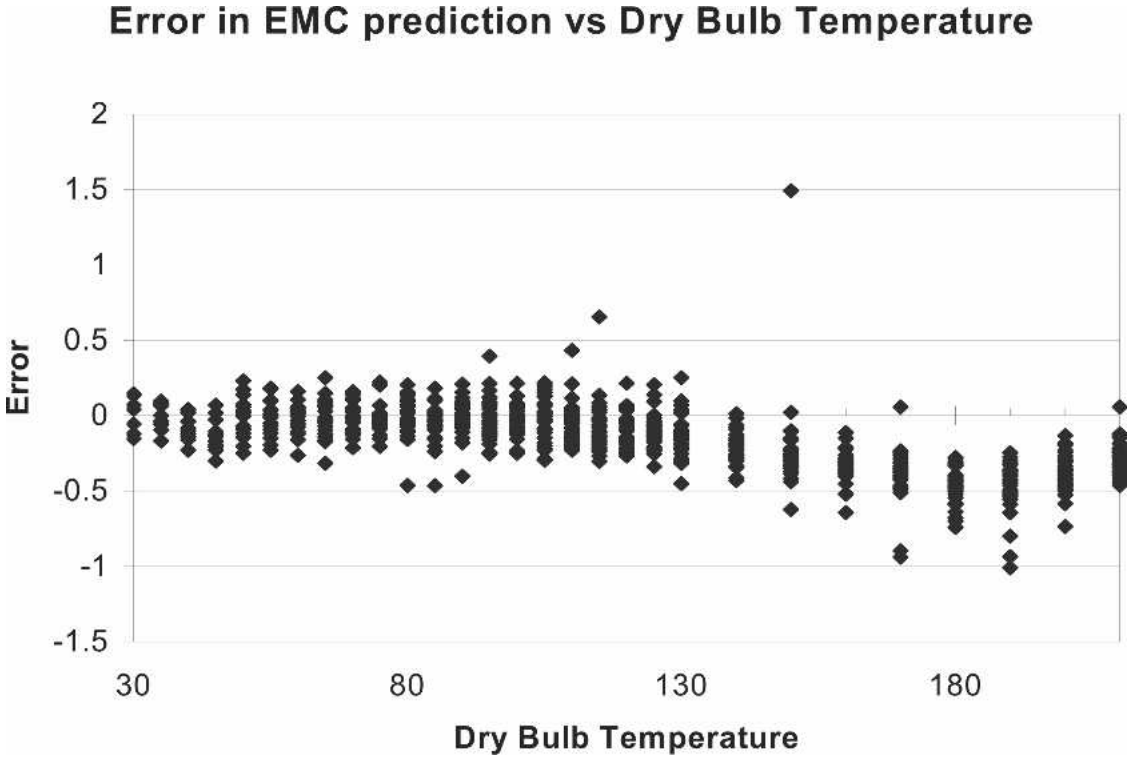

FIG. 1. Error distribution vs. dry bulb temperature of predictions of regression Eq. (11).

using SAS, various goodness-of-fit statistics were observed to check the accuracy of the model. If the model fit was found to be inadequate, then step 1 was repeated with a newly specified polynomial model.

3. Residual diagnostics were analyzed to confirm that none of the underlying assumptions of regression analysis stated above were violated.

After several iterations and the trial of various methods to remove problems of multicollinearity and non-constant variance, the following additional measures were taken.

1. To reduce the problem of multicollinearity, a new model was built after centering the predictor variables.
2. The non-constant variance was removed by using the square-root transformation on EMC values. So, the model was fitted on $\sqrt{\text{EMC}}$ instead of EMC values.

RESULTS AND DISCUSSION

The resulting regression model for predicting the values of EMC above 110°F dry bulb temperature is:

$$\begin{aligned}\sqrt{\text{EMC}} = {} & 7.30548 + 11.64339h - 0.00792\text{T}_{\text{db}} \\ & - 0.37436w_1 - 0.39562w_3 \\ & + 0.06902w_2^2 + 0.00518w_3^2 \\ & + 0.00129w_1w_4 - 0.00048153w_2w_4 \\ & + 0.61135x_1 \end{aligned} \quad (12)$$

where

$$\text{EMC} = \text{Equilbrium moisture content}$$

$$h = \text{Centered relative vapor pressure} = h_{\text{original}} - .58537037$$

$$\text{T}_{\text{db}} = \text{Centered dry bulb temperature} = \text{T}_{\text{db}_{\text{original}}} - 157.5$$

$$w_1 = 0.0001 + 0.0025\text{T}_{\text{db}} + 0.0007\text{T}_{\text{db}}^2$$

$$w_2 = 0.2 + 0.06\text{T}_{\text{db}} - 0.00004\text{T}_{\text{db}}^2$$

$$w_3 = 14 + 35.5h + 20.7h^2$$

$$w_4 = 1 + .01h\text{T}_{\text{db}} + 0.1h^2\text{T}_{\text{db}}^2$$

$$x_1 = \frac{(1 + w_2 + w_3)}{(w_1^2 + w_3^2)}$$

The ANOVA statistics for this model are presented in Table 1.

From the ANOVA Table 1, we see a high adjusted R^2 (.9990), low root mean square error,

TABLE 1. *ANOVA table of model Eq. (12).*

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	203.89720	22.65524	45785.4	<.0001
Error	422	0.20881	0.00049481		
Corrected Total	431	204.10601			

Root MSE	0.02224	R-Square	0.9990
Dependent Mean	2.89835	Adj R-Sq	0.9990
Coeff Var	0.76748		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.30548	0.07950	91.89	<.0001
RH	1	11.64339	0.16366	71.15	<.0001
Tdb	1	-0.00792	0.00021799	-36.34	<.0001
w1	1	-0.37436	0.06540	-5.72	<.0001
w3	1	-0.39562	0.00666	-59.40	<.0001
w22	1	0.06902	0.01237	5.58	<.0001
w33	1	0.00518	0.00007184	72.10	<.0001
w15	1	0.00129	0.00021354	6.04	<.0001
w25	1	-0.00048153	0.00010509	-4.58	<.0001
x1	1	0.61135	0.02252	27.14	<.0001

and highly significant parameter estimates. Figure 2 demonstrates that the fit of model (12) is adequate, showing an excellent approximation of normality of residuals and adequate dispersion of residuals versus predicted values.

Comparison of performance of old and new regression equations

Figure 3 shows the error distribution (actual-predicted) of the old and new regression equations above 110°F. For comparison purposes, the scale on the plots was kept consistent. As clearly seen by comparing the two graphs in Fig. 3, the new model is a superior predictor of EMC. While the old model shows an obvious systematic bias (i.e., the errors differ from zero on average), the new model exhibits zero mean errors (no bias), normality in dispersion of EMC predictions within a tighter range of values. Table 2 shows some of the comparison statistics.

There is a 44% reduction in root mean square error for model (12) as compared to model (11). This implies that the new model is predicting significantly more precisely. The distribution of errors is normal in Eq. (12), showing no systematic bias, indicating that contrary to model (11), the new model (12) is accurate. We point out that the new model has 10 parameters, as opposed to 12 parameters in model (11), and hence is less complex.

CONCLUSIONS

The equation published in Simpson (1991) has significant statistical deficiencies for prediction of EMC above 110°F. Equation (12) of this paper demonstrates better prediction of EMC above 110°F. The new model for EMC has fewer parameters (12 vs. 10) than the existing model, and shows considerable improvements in the quality of the predictions when the dry bulb temperature exceeds 110°F.

An application of this work is the use of the new model developed here for process control purposes. The new fitted model provides better

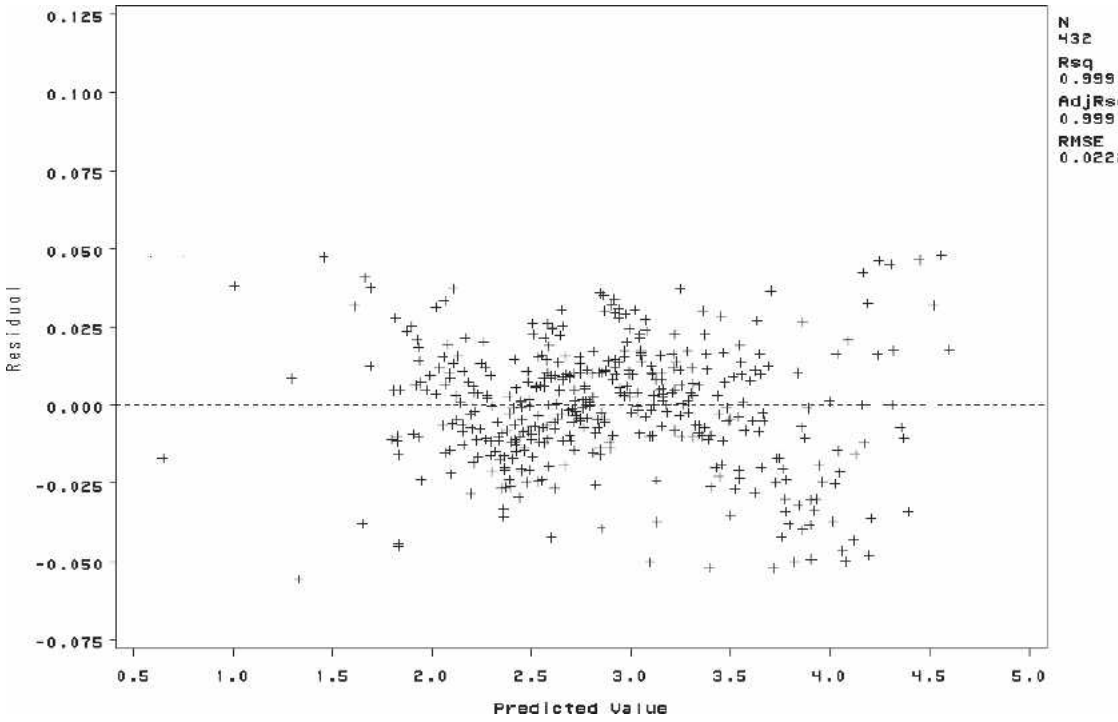

FIG. 2. Residuals vs. predicted EMC values for Eq. (12).

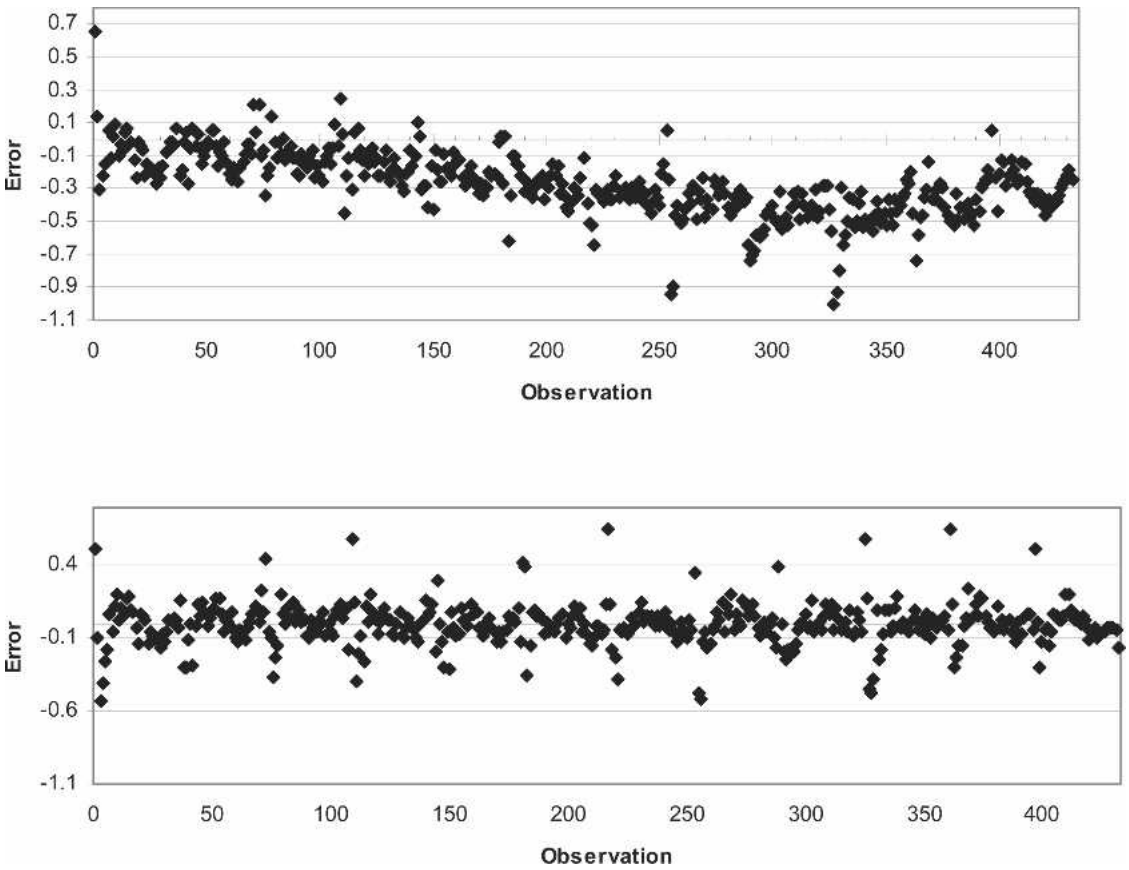

FIG. 3. Top: Error distribution above 110°F for predictions of regression Eq. (11); Bottom: Error distribution above 110°F for predictions of regression Eq. (12).

TABLE 2. *Comparison statistics of Simpson (1991) model (11) and new model (12).*

	Equation (11)	New Model (12)
Sum of Squared Errors	48.78485	9.83799
Maximum positive error	1.49	0.88
Maximum negative error	−1.008	−0.525
Mean Absolute Error	0.112928	0.0228
Number of parameters	12	10
$\sqrt{\text{MSE}} = \hat{\sigma}^2$	0.3408	0.1526

predictions which can allow a kiln operator to find better set points of the drying process. In such cases, the smaller prediction error variance of the new model can result in smaller control error variance, with the resulting savings in the operation of the kiln.

REFERENCES

DRAPER, N., AND H. SMITH. 1998. Applied regression analysis, 3rd ed. John Wiley & Sons, Inc, New York, NY.

SIMPSON, W. T. (ed.). 1991. Dry Kiln Operator's Manual. Agricultural Handbook No. 188, United States Department of Agriculture, Madison, WI. 274 pp.

GATTANI, N., E. DEL CASTILLO, C. RAY, AND P. R. BLANKENHORN. 2005. Times series analysis and control of dry kilns. Wood Fiber Sci. 37(3):472–483.

HOERL, A. E., AND R. W. KENNARD. 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(3):55–67.

MONTGOMERY, D.C. 2001. Design and analysis of experiments, 5th ed. John Wiley & Sons, Inc, New York, NY.

NETER, J. , W. WASSERMAN, AND M. KUTNER. 1985. Applied linear statistical models. 2nd ed. R. D. Irwin.