IIE Transactions (2006) **38**, 13–23

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/07408170500346378

Selection and validation of predictive regression and neural network models based on designed experiments

CHANG-XUE JACK FENG[1,*], ZHI-GUANG (SAMUEL) YU[1] and ANDREW KUSIAK[2]

[1]*Department of Industrial & Manufacturing Engineering & Technology, Bradley University, Peoria, IL 61625, USA*
E-mail: cfeng@bradley.edu
[2]*Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, IA 52242-1527, USA*

Received October 2002 and accepted March 2005

Model selection and validation are critical in predicting the performance of manufacturing processes. The correct selection of variables minimizes the model mismatch error whereas the selection of suitable models reduces the model estimation error. Models are validated to minimize the model prediction error. In this paper, the relevant literature is reviewed and a procedure is proposed for the selection and cross-validation of predictive regression analysis and neural network models. Specifications on surface roughness and tolerances impact on manufacturing process plans, and differentiate product quality, and ultimately the product cost and lead times. Experimental data from a turning surface roughness study is used to demonstrate the developed concepts with regression and neural network techniques being used for the purpose of predictive rather than descriptive modeling.

1. Introduction

The performance of a manufacturing process can be predicted using models that can originate from either introspection or observation (or both) (Gershenfeld, 1999). Although developing an analytical model is feasible in some simplified situations, most manufacturing processes are complex and therefore empirical models that are less general, more practical and less expensive than analytical models are of interest.

In any data-driven empirical modeling effort, the two central tasks are: (i) to select the functional form of the model; and (ii) using the data to determine the adjustable parameters of the model (Gorman and Torman, 1966; Gershenfeld, 1999). These two issues are closely related but result in different types of errors. *Model mismatch errors* arise from a model that is unable to represent the data, and *model estimation errors* arise from using incorrect values for the model parameters. Decreasing one type of error is likely to increase the other type. This is the so-called bias/variance trade-off as shown in Fig. 1; less bias in the estimate of a model parameter normally leads to more variance (Twomey and Smith, 1998; Gershenfeld, 1999). A more flexible model that can better represent the data may also be more easily led astray by noise in the data. The above two tasks are usually completed during two different, but related, phases of the empirical modeling. Determining the right form of the model in order to reduce model mismatch error is accomplished during the model construction phase whereas determining the correct model parameters is achieved at the model selection and validation phase.

Data mining has found many early successful applications in marketing, sales, finance, and health sciences (Groth, 1998), and it has also been applied to engineering design and manufacturing applications (Kusiak, 2000a, 2000b; Kusiak and Kurasek, 2001; Feng *et al.*, 2002; Feng and Wang, 2002a, 2003; Feng and Yu, 2003). Both Regression Analysis (RA) and Neural Network (NN) approaches have been used in empirical modeling studies and they have recently been termed as data-mining tools (Groth, 1998; Witten and Frank, 2000). One of the most important issues in data mining as applied to predictive modeling is model assessment and selection (Mitchell, 1997; Hastie *et al.*, 2001; Witten and Frank, 2000). A brief comparison of the pros and cons of the RA and the NN methods as applied to empirical modeling can be found in Feng and Wang (2003). For a comparison of the pros and cons of analytical and empirical modeling techniques as applied to machining surface roughness modeling, also refer to Feng and Wang (2003).

A number of criteria and methods have been used to develop alternative models, make trade-offs, select the best model alternative, and assess the credibility of the selected model. This paper reviews the critical issues involved in the construction, selection, and validation of predictive RA and NN models. Data from a turning surface roughness study are used to illustrate the methodology and allow a

*Corresponding author

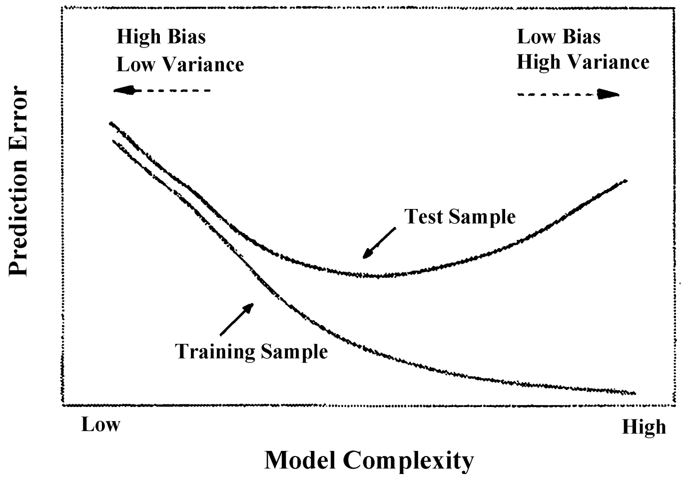

Fig. 1. The behavior of test sample and training sample errors as the model complexity varies.

comparison between the RA and NN models. A rigorous procedure is proposed for the construction, selection, and validation of the predictive RA and NN models.

The reminder of the paper is organized as follows. Section 2 reviews and evaluates the issues and methodologies involved in model development and selection in regression and model validation in RA studies. Section 3 presents the issues and techniques involved in the model development and selection in NN studies. The computational studies presented in Section 4 illustrate the methodology and procedures outlined in Section 2 and 3. Finally, conclusions are drawn in Section 5.

2. Development and selection of RA models

The development and selection of RA models follow somewhat different steps from those in NN modeling. Construction of a RA model normally begins with the correct selection of variables. A model builder faces a pool of candidate regressors that should include all the influential factors. The actual subset of regressors that should be used in the model needs to be determined. Finding an approprite subset of regressors for the model is called the *variable selection problem* (Montgomery *et al.*, 2001). Building a RA model that includes only a subset of the available regressors involves two conflicting objectives. On the one hand, one attempts to include as many regressors as possible in the model so that the "information content" in these factors can influence the predicted value of y. On the other hand, one attempts to include as few regressors as possible in the model because the variance of the prediction $\hat{y}$ increases as the number of regressors increases, resulting in the so-called bias/variance trade-off. Also, the cost of data collection and model maintenance increases with the number of regressors in the model. The process of finding a model that is a compromise between these two objectives is called selecting the "best" subset regression equation. Unfortunately, there is no single definition for what constitutes the best equation. Furthermore, several algorithms are available for variable selection, and these procedures frequently specify different subsets of the candidate regressors as the best one. These two issues will be discussed in Sections 2.1 and 2.2, respectively, and illustrated with computational examples in Section 4.

2.1. *Selection of the best subset RA models*

The best subset of RA models is usually selected using some of the following criteria: (i) the value of the coefficient of multiple determination R^2; (ii) the value of the residual mean square s^2; (iii) Mallows' C_p-statistic (Draper and Smith, 1998; Montgomery *et al.*, 2001); and (iv) Akaike's information criterion (Burnham and Anderson, 2002; Miller, 2002). Of particular value for predictive model selection is the C_p-statistic proposed in Mallows (1973, 1995, 1997). This criterion is related to the mean-square error of a fitted value as follows:

$$C_p = \frac{RSS_p}{s^2(p)} - n + 2p, \tag{1}$$

where p is the number of terms in the model with $p-1$ regressors, RSS_p the residual sum of squares, and n the number of data used for fitting the subset regression model.

Assuming that $E[s^2(p)] = \sigma^2$, it is true approximately that the ratio RSS_p/s^2 has an expected value $(n-p)\sigma^2/\sigma^2 = n-p$. Thus, approximately, $E(C_p) = p$. The best subset model is chosen after inspecting the C_p against p plot. Using the C_p-statistic to determine the best subset, one would look for a RA model with a low C_p value about equal to p. Draper and Smith (1998) summarized the precautions in using the C_p-static. Some modifications of Mallows' C_p-statistic have been made over the years (e.g., see Gilmour (1996). For an in-depth treatment of this topic see Miller (2002). Since the square-root of S instead of $s^2(p)$ is reported in Minitab, only S will be used instead of $s^2(p)$ later in this paper.

2.2. *Computational techniques for variable selection*

For the purpose of prediction, the model builder should use RA models that employ a subset of the candidate regressor variables. To find the subset of variables to be used in the final equation, the model builder needs to fit models with various combinations of the candidate regressors. Computational techniques to generate subset models can be broadly divided into two categories: (i) all possible regressions including the so-called best subset regression; and (ii) stepwise regression. The latter includes three general techniques; (i) forward selection; (ii) backward elimination; and (iii) Efroymson's stepwise regression. These techniques are widely covered in classic texts including Draper and Smith (1998), Montgomery *et al.* (2001) and Miller (2002).

The disadvantage of stepwise regression is that it does not produce a subset of models that are necessarily best with respect to any standard criterion, and furthermore it is oriented towards producing a single final equation. On the other hand, all possible regression will identify the best subset of models with respect to whatever criterion the analyst selects. For up to 20 or 30 candidate regressors, the computational cost is approximately the same as the stepwise-type procedure (Montgomery *et al.*, 2001). The best subset regression technique is used in this paper.

2.3. *Hypothesis testing*

Hypothesis testing is used to validate that the predictions from any learning scheme have the same probability distribution (equal means and variances) as that of the experimental data. For normally distributed data, the t-test and F-test can be employed to compare the means and variances of predictions and the corresponding observations reserved for cross-validation. Of particular interest is the paired t-test for data following normality, since one is always interested in comparing the predictions and observations in pairs given a set of input conditions. For nonparametric data or data with unknown distributions, the Mann-Whitney test instead of the t test and Levene's test instead of the F-test are used to test the medians and variances, respectively.

Similarly, the t-test and F-test or their counterpart nonparametric tests can be employed to compare the predictions from two learning schemes or models, such as the RA model and the NN model. To the surprise of the authors, the normality tests, the variance test, and the nonparametric tests do not appear to be recommended in the available literature that includes Mitchell (1997) and Witten and Frank (2000) on validating a learning scheme or comparing multiple learning schemes. Feng *et al.* (2002), Feng and Wang (2002a, 2002b, 2003, 2004) and Feng *et al.* (2004) have applied the above procedure in qualifying models and comparing two competitive learning schemes. Additional approaches for comparing two linear regression models are discussed in Miller (2002, p. 97).

2.4. *Cross-validation*

In engineering practice, it is common to withold one-third of the data for validation and use the remaining two-thirds for training and testing (Witten and Frank, 2000). This *threefold Cross-Validation* (CV) is used in this paper. Balancing the data set is important so that the data in each of the training, testing, and validation sets are representative. In each of the CV iterations, a certain proportion, say two-thirds of the data, is randomly selected for training (in NNs, 90% of the two-thirds for training and 10% of the two-thirds for testing), possibly with stratification, and the remainder is used for validation. The error rates on the different iterations are averaged to yield an overall error rate. In this train-test-validate procedure, the test set of data is used to optimize the generalization ability of the model, and the final CV set is used in a final test that evaluates the prediction error rated (Swingler, 1996). This procedure will be used throughout this paper. The difference is that the leave-one-out estimate is used in RA for testing to take advantage of the existing function in the Minitab statistics package, whereas 10% of the data in the entire training-test set will be used to test the NN models.

Although Witten and Frank (2000) indicated that tenfold CV is the standard way of evaluating the error rate of a learning technique given a single, fixed sample of data, Breiman and Spector (1992) and Zhang (1993) did not reveal any statistical advantages created by using tenfold CV rather than fivefold CV. Breiman and Spector (1992) concluded that when the predictors **X** are randomly selected, the tenfold or fivefold CV and bootstrap error estimation perform almost the same, and they both perform better than the leave-one-out method in terms of the prediction error rates. The bootstrap method uses sampling with replacement which is especially useful for a small data set. Refer to Efron and Tibshirani (1998) for more coverage of this topic. Furthermore, no evidence is available to support the contention that the tenfold CV performs better than the fivefold CV. Zhang (1993) concluded that a twofold CV would lead to the worst prediction errors. These are the reasons why we used threefold CV in this reserach.

2.5. *Prediction error evaluation*

RA has four possible uses: (i) data description; (ii) prediction; (iii) parameter estimation; and (iv) control (Montgomery *et al.*, 2001). Since the focus of this research is to construct and select the best model(s) for the prediction of future observations, we would like to select regressors such that the mean-square error of the prediction is minimized. The PRESS (PREdiction Sum of Squares) statistic was suggested by Miller (1974), and is essentially a leave-one-out CV statistic for model selection based on this consideration. Denote by $\hat{y}_{ip}$ the predicted value for y_i. The *PRESS* statistic is computed from Equation (2) which is taken from Montgomery *et al.* (2001, p. 301):

$$PRESS_p = \sum_{i=1}^{n} [y_i - \hat{y}_{ip}]^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2 \qquad (2)$$

where h_{ii} is the diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Modern statistics software packages usually report the above value. If $n \gg p$, then according to Miller (2002, p. 146) we can write:

$$PRESS_p \approx \sum_{i=1}^{n} \frac{e_i^2}{(1 - p/n)^2} = RSS_p \frac{n^2}{(n-p)^2}. \qquad (3)$$

A related criterion is the prediction R^2-like statistic which according to Montgomery *et al.* (2001, p. 535) can be written as:

$$R^2_{\text{Prediction}} = 1 - \frac{PRESS_p}{SS_{\text{T}}} \quad (4)$$

that measures, in an approximate sense, how much of the variability in new observations the model might be expected to explain, where SS_{T} is the total sum of squares of the model.

A number of other methods are available to estimate the prediction errors. It is generally acceptable that some kind of CV method should be used to obtain a satisfactory prediction error estimate. Prediction error estimation is a much more mature area in RA than in any other of the learning schemes including NNs. For example, in addition to the PRESS statistic, the following three traditional estimates from forecasting techniques are often used: (i) the Mean Absolute Deviation (MAD); (ii) the Mean Absolute Percent Error (MAPE); and (iii) the Mean-Root-Squared Error (MRSE). They can be defined mathematically as:

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|, \quad (5)$$

$$MAPE = \left(\frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right|\right) \times 100\%, \quad (6)$$

$$MRSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \quad (7)$$

Additional formulations for the prediction error rate are available in Witten and Frank (2000, p. 148), and Burnham and Anderson (2002).

To evaluate the prediction error rates of a RA model, the so-called Mean-Squared Error of Prediction (MSEP) is also recommended. Thompson (1978a, 1978b) suggested two models that differ in that the regressors in the RA model can be fixed or random. The values of regressors are fixed or controllable if they are from a controlled experiment; otherwise they are random. Miller (2002) summarized the computation of the two kinds of MSEP. Since our data are from designed experiments, only the MSEP model for fixed variables will be discussed. The MSEP is defined mathematically as (Miller, 2002, p. 116):

$$E(y - \hat{y}(x))^2 \quad (8)$$

If the regression coefficients are unbiased Least-Square (LS) estimates, then according to Miller (2002, p. 116) we can write:

$$MSEP \approx (RSS_p + 2p\sigma^2)/n. \quad (9)$$

Substituting σ^2 by an unbiased estimate $\hat{\sigma}^2 = RSS_k/(n - k)$:

$$MSEP \approx \frac{1}{n}\left(RSS_p + \frac{2pRSS_k}{n-k}\right) \quad (10)$$

To obtain unbiased LS-estimates of the regression coefficients, the data set used for model estimation or construction must be independent from that used for model validation. Also, the regressors must be fixed to apply the above equation to estimate the prediction error rate. By comparing the forms of the C_p-statistic in Equation (1) and of the MSEP in Equation (10), Miller (2002, p. 116) has shown that minimizing the MSEP is equivalent to minimizing Mallows' C_p-statistic. He cautioned that Mallows' C_p-static has been widely used as a criterion in subset selection despite the requirement for unbiased regression coefficients. Most of the applications have been to situations with predictors that are random variables, whereas the derivation of Equations (1) and (10) requires that the variables be fixed or controllable. Montgomery *et al.* (2001, p. 537) cautioned that the validation data set should contain at least 15 points in order to obtain a reasonable assessment of model prediction performance.

The computational studies presented in Section 4 report all of the above prediction error statistics for the RA model selection, whereas only the MAD, MAPE, and MRSE statistics will be reported for the NN models, since the other statistics are not readily available from the NN models.

3. NN modeling

NNs possess a number of attractive properties for modeling complex manufacturing process and systems: (i) a universal function approximation capability; (ii) resistance to noisy or missing data; (iii) the accommodation of multiple nonlinear variables for unknown interactions; and (iv) a good generalization capability (Coit *et al.*, 1998; Twomey and Smith, 1998). For a justification as to why NNs are used to model the machining surface roughness, refer to Feng and Wang (2003). This research uses a commercial package BrainMaker (Lawrence and Fredrickson, 1998) with a back-propagation-based feed-forward model from California Scientific Software for NN modeling. For an introduction to NNs see Wasserman (1989), Lawrence (1994) and Swingler (1996).

Determining the number of hidden layers and the number of neurons in each hidden layer is a considerable task. The number of hidden layers is usually determined first. The rule of thumb is to start with a one-hidden-layer network (Lawrence, 1994, p. 201). If the one-hidden-layer does not train well, then the number of neurons or the training and testing tolerances or both can be changed. Adding more layers should be the last option. Here, a one-hidden-layer network is used.

The number of hidden neurons is problem dependent. For example, any network that requires data compression must have a hidden layer that is smaller than the input layer (Swingler, 1996, p. 52). A conservative approach is to select a number between the number of input neurons and the number of output neurons (Kusiak, 2000a, p. 364).

Hecht-Nelson (1987) used Kolmogorov's theorem (which states that any function of i variables may be represented by the superposition of a set of $2i + 1$ univariate functions) to derive the upper bound for the required number of hidden neurons, where i is the number of inputs. Therefore, Hecht-Nelson (1987) determined that $2i + 1$ should be used as the upper bound on the number of hidden neurons required for a one-hidden-layer back-propagation network.

Girossi and Poggio (1989) pointed out that Kolmogorov's theorem requires that the component functions are chosen to fit each particular case, and not as is the case with NNs, that the functions are fixed and parameterized. They also pointed out that a NN requires a smooth function in order to be able to generalize its results and that Kolmogorov's theorem does not guarantee this. Kurkova (1991), however, stated that the fact that a NN is only an approximation eliminates both of these difficulties. Kurkova restated Kolmogorov's theorem in terms of a set of sigmoid functions.

Lawrence and Fredrickson (1998, pp. 11–12) suggested that a best estimation for the number of hidden neurons is half of the sum of the inputs and outputs, i.e.,:

$$h = (i + o)/2 \tag{11}$$

where o is the number of output neurons, and h is the number of hidden neurons. An alternative proposed in Lawrence and Fredrickson (1998, pp. 11–13) is to relate the training data size to the number of hidden neurons as follows:

$$\frac{N}{10} - i - o \leq h \leq \frac{N}{2} - i - o, \tag{12}$$

where N is the number of training facts or data. Lawrence and Fredrickson (1998, pp. 11–13) further suggested that in order to choose between the above wide ranges of numbers, the training tolerance must be considered. In general, a tighter tolerance requires that the network be trained with fewer hidden neurons.

Marchandani and Cao (1989) derived the following relationship:

$$h = i \log_2 P, \tag{13}$$

where P is the number of training patterns. Baum and Haussler (1989) presented various expressions related to training patterns, neurons, and weights under the consideration of various statistical confidence levels. Lipmann (1987) indicated that the maximum number of hidden neurons for a one-hidden-layer network is $o(i + 1)$.

An overfitted model could approximate the training data well but poorly generalize the validation data set. On the other hand, an underfitted model could well generalize the validation data set but poorly approximate the training data. To avoid overfitting and underfitting, the best number of training facts must be determined. There is no general guideline available on how to achieve this number. However, Lawrence and Fredrickson (1998, pp. 11–12) suggested the following relationship to estimate the best number of training facts:

$$2(i + h + o) \leq N \leq 10(i + h + o). \tag{14}$$

A back-propagation network is sensitive to the initial values of the weights (Kolen and Pollack, 1990). Properly selected initial weights obtained through the selection of training tolerances can shorten the learning time and result in stable weights. Initial weights that are too small increase the learning time, which may cause difficulties in the convergence to an optimal solution (Kusiak, 2000a, p. 365). If the initial weights (training tolerances) are too large then the network may end up with unstable weights (Wasserman, 1989). The BrainMaker software allows the user to change the weights by changing the training and testing tolerances, and it has a default training tolerance of 0.1 and a default testing tolerance of 0.4.

4. Computational study

4.1. *Experimental conditions and data*

The theoretical and practical issues reviewed in Sections 2 and 3 are now illustrated with computational studies using data obtained during from a turning surface roughness study. For a recent survey of machining surface roughness modeling, refer to Feng (2001), and also Feng and Wang (2002b, 2003). Fundamentals of machining surface roughness issues are widely covered in classic machining texts, including Shaw (1984), Boothroyd and Knight (1989), Groover (2002) and also Kalpakjian and Schmid (2003). The screening experiment is a 2^5 full factorial design with two replicates. The order of the 64 experiments was randomized, and they were performed on a production-type CNC turning center. The two materials under consideration were steel 8620 and aluminum 6061T. In order to quantify the material, the Rockwell B hardness measure HRB is used.

After analysis of the screening experiment, a follow-up confirmation experiment was conducted based on a 2^{5-2} fractional factorial design with two replicates. The new design added two levels of feed, depth of cut, and speed but maintained the material and cutter nose radius at the same two levels. Execution of these 16 confirmation experiments also followed a randomized order. Each of the 80 samples was measured three times with each measurement taken about 120° apart along the axis by using the Mitutoyo surface profilometer SJ-301. More details about the design of experiments and data are available in Kapse (2001).

The 80 data sets were divided into three sets in order to apply the threefold CV method for model selection and validation. An attempt was made to balance the data in each set based on the speed level so that each set would be representative of the process parameter combination. Sets 1 and 2 have 27 data sets each and set 3 has 26 data sets since 80 data sets were available in total. Dividing the total data into three sets ensures that each validation set will have at least 15 data sets for a reasonable assessment of prediction errors recommended in Montgomery *et al.* (2001).

4.2. *Computational results obtained using a RA approach*

Three best subset regressions were performed based on the threefold CV scheme. For example, the first one uses sets 1 and 2 for model construction, the second one uses sets 1 and 3 for model construction, and the third uses sets 2 and 3 for model construction. The authors' past experience shows that the relationship between the surface roughness and the process parameters is highly nonlinear, and a logarithmic transformation of the data would yield a good result (Feng and Wang, 2003).

Borrowing the modified Taylor tool life equation in metal cutting (Kalpakjian and Schmid, 2003, p. 430), the surface roughness model can be postulated as follows:

$$R_a = Ch^{b_1} f^{b_2} r^{b_3} d^{b_4} s^{b_5}. \quad (15)$$

Taking the natural logarithm on both sides leads to.

$$\ln R_a = \ln C + b_1 \ln h + b_2 \ln f + b_3 \ln r + b_4 \ln d + b_5 \ln s, \quad (16)$$

where, R_a is the arithmetic average surface roughness in μm, h is material hardness in HRB, f is the feed in mm/rev, r is the cutter nose radius in mm, d is the finishing depth of cut in mm, and s is the speed in m/min. Denote $y = \ln R_a$, $b_0 = \ln C$, $x_1 = \ln h$, $x_2 = \ln f$, $x_3 = \ln r$, $x_4 = \ln d$, and $x_5 = \ln s$. The Equation (16) can be rewritten as:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5. \quad (17)$$

A more general form of the above equation is a polynomial model that includes the quadratic and interaction terms as follows (considering only the two-factor interaction terms):

$$\begin{aligned} y = {} & b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5 + b_{11} x_1^2 + b_{22} x_2^2 \\ & + b_{33} x_3^2 + b_{44} x_4^2 + b_{55} x_5^2 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{14} x_1 x_4 \\ & + b_{15} x_1 x_5 + b_{23} x_2 x_3 + b_{24} x_2 x_4 + b_{25} x_2 x_5 + b_{34} x_3 x_{34} \\ & + b_{35} x_3 x_5 + b_{45} x_4 x_5 + e. \end{aligned} \quad (18)$$

After data transformation, the above polynomial RA model can be used to fit the data. See Montgomery *et al.* (2001, ch. 7) for a good treatment of fitting a polynomial RA model. Thus, the three sets of data were transformed into their logarithmic counterparts before the best subset regression starts. A summary of related statistics from Minitab for one of the three best subset regressions is provided in Table 1, where "*X1*" means x_1, and "*X1X2*" means $x_1 x_2$, for example.

Using the principle for variable selection outlines in Section 2.1, the C_p against p plot for the first of the subset regressions was generated and is presented in Fig. 2. The other two tables and plots are not reported here due to space limitations, but they are available in Yu (2003). This kind of plot can be used to select model candidates worthy of further consideration. From the third subset regression (Table 1), the following three subsets with $p = 5$ were selected based on Fig. 2: (i) x_2, x_3, x_1x_2, x_1x_4; (ii) x_2, x_1x_2, x_1x_4, x_3x_5; and (iii) x_2, x_3, x_1x_2, x_2x_4.

A simple RA was completed for each of the above selected subsets, and a summary of the relevant fitting statistics is presented in Table 2. Notice that in addition to the three criteria for model selection discussed in Section 2.1, the PRESS statistic and prediction R^2 values are also reported. More importantly, a paired t-test was performed to compare the means of the fitted values and their corresponding observed values, and an F-test was conducted to compare the variances of the fitted values with those of their observed values in order to check the goodness of fit of each candidate model.

It turns out that models I-1, I-2 and I-3 in Table 2 have a marginal P value for the F-test based on a 5% significance level or 95% confidence level, and models III-1, III-2, and III-3 are among the best qualified candidates. Models II-1, II-2, and II-3 also have a satisfactory goodness of fit based on the two kinds of hypothesis testing. If the other fitting statistics are considered, models III-1 and III-2 have almost the same quality of fit, since their residual squared error, adjusted R^2, *PRESS* value, and prediction $R^2_{\text{Prediction}}$ are all among the best.

Finally, the prediction error statistics from CV were computed for the above model candidates and are provided in Table 3. No units are given in Table 3 since these prediction error statistics are based on the logarithmic transformation. These statistics should be used to select the best model candidate (because it is the prediction performance that matters the most. The hypothesis test compared the predictions with the observations reserved for the purpose of validation. Their results are provided in the last two columns of Table 3. It shows that each of the model candidates was able to provide on average a prediction that is statistically the same as its respective observation. However, the four prediction error measures picked different winners. For example, MAD would pick model II-3, MAPE model II-1, MRSE model II-3, and MSEP models III-1, III-2, and III-3.

Interestingly, none of them picked those models disqualified by the hypothesis tests during model fitting shown in Table 2. Actually, the models picked by the three traditional measures, MAD, MAPE, and MRSE, are not significantly different, because the MAPE value of model II-3 is only slightly greater than that of model II-1. Model II-3 can be used, since the other two traditional prediction error measures also favor this model. Notice that it is not unusual that a selection method will lead to several models with a similar performance. The problem finally boils down to how to resolve the difference between the winners selected by the traditional measures and the MSEP, respectively. This is where domain knowledge comes into play. The literature contains numerous reports on turning-surface roughness models; they are surveyed in Feng (2001), Feng and Wang (2003) and Kalpakjian and Schmid (2003, p. 436) and these

Table 1. The third subset regression statistics (sets 2 and 3 used for model construction)

Vars	*p*	R^2	R^2 *(adj)*	C_p	$p - C_p$	*S*	*X1*	*X2*	*X3*	*X4*	*X5*	*X1X2*	*X1X3*	*X1X4*	*X1X5*	*X2X3*	*X2X4*	*X2X5*	*X3X4*	*X3X5*	*X4X5*
1	2	21.1	19.6	53.2	51.2	0.457 15		X													
1	2	20.7	19.1	53.8	51.8	0.458 48												X			
1	2	19.3	17.7	55.5	53.5	0.462 35	X														
2	3	43.3	41.0	26.5	23.5	0.391 57		X				X									
2	3	42.0	39.7	28.1	25.1	0.395 97	X											X			
2	3	42.0	39.6	28.2	25.2	0.396 04	X	X													
3	4	54.0	51.2	14.6	10.6	0.356 15		X	X			X									
3	4	54.0	51.2	14.6	10.6	0.356 18		X				X								X	
3	4	53.1	50.3	15.7	11.7	0.359 47		X				X	X								
4	**5**	**65.2**	**62.3**	**2.1**	**2.9**	**0.313 09**		**X**	**X**			**X**		**X**							
4	**5**	**65.2**	**62.3**	**2.1**	**2.9**	**0.313 19**		**X**				**X**		**X**						**X**	
4	**5**	**65.0**	**62.1**	**2.3**	**2.7**	**0.313 74**		**X**	**X**			**X**				**X**					
5	6	68.0	64.6	0.5	5.5	0.303 34		X	X			X	X	X							
5	6	67.6	64.2	0.9	5.1	0.3051		X	X	X		X	X								
5	6	67.4	63.9	1.3	4.7	0.306 35		X	X			X	X								X
6	7	69.2	65.1	0.9	6.1	0.300 98		X	X			X	X	X							X
6	7	69.0	65.0	1.1	5.9	0.3016		X	X			X	X	X		X					
6	7	68.8	64.7	1.5	5.5	0.3029		X	X	X		X	X	X							
7	8	70.0	65.3	1.9	6.1	0.300 18		X	X			X	X	X					X		X
7	8	70.0	65.3	1.9	6.1	0.3004		X	X			X	X	X							X
7	8	69.7	65.0	2.3	5.7	0.301 72		X	X			X	X	X		X			X		
8	9	71.0	65.8	2.5	6.5	0.298 31		X	X			X	X	X		X			X		X
8	9	70.5	65.2	3.2	5.8	0.300 93		X	X	X		X	X	X		X			X		
8	9	70.4	65.0	3.3	5.7	0.301 47	X	X	X			X	X	X		X					X
9	10	71.3	65.3	4.2	5.8	0.300 24	X	X	X			X	X	X		X			X		X
9	10	71.0	65.0	4.5	5.5	0.301 66		X	X			X	X	X		X		X	X		X
9	10	71.0	65.0	4.5	5.5	0.301 69		X	X			X	X	X		X	X		X		X
10	11	71.3	64.5	6.1	4.9	0.303 68	X	X	X	X		X	X	X		X			X		X
10	11	71.3	64.5	6.1	4.9	0.303 68	X	X	X			X	X	X		X	X		X		X
10	11	71.3	64.5	6.2	4.8	0.303 78	X	X	X			X	X	X	X	X			X		X
11	12	71.4	63.7	8.1	3.9	0.307 17	X	X	X		X	X	X	X	X	X			X		X
11	12	71.4	63.7	8.1	3.9	0.307 22	X	X	X	X		X	X	X		X	X		X		X
11	12	71.3	63.7	8.1	3.9	0.307 33	X	X	X	X		X	X	X		X			X	X	X
12	13	71.4	62.8	10	3	0.310 81	X	X	X	X	X	X	X	X	X	X			X		X
12	13	71.4	62.8	10	3	0.310 85	X	X	X		X	X	X	X	X	X	X		X		X
12	13	71.4	62.8	10.1	2.9	0.310 98	X	X	X		X	X	X	X	X	X			X	X	X
13	14	71.4	61.9	12	2	0.314 63	X	X	X	X	X	X	X	X	X	X	X		X		X
13	14	71.4	61.9	12	2	0.314 76	X	X	X	X	X	X	X	X	X	X			X	X	X
13	14	71.4	61.9	12	2	0.314 77	X	X	X	X	X	X	X	X	X	X		X	X		X
14	15	71.4	60.9	14	1	0.318 73	X	X	X	X	X	X	X	X	X	X	X	X	X		X
14	15	71.4	60.9	14	1	0.318 73	X	X	X	X	X	X	X	X	X	X	X		X	X	X
14	15	71.4	60.9	14	1	0.318 87	X	X	X	X	X	X	X	X	X	X		X	X	X	X
15	16	71.4	59.9	16	0	0.322 99	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

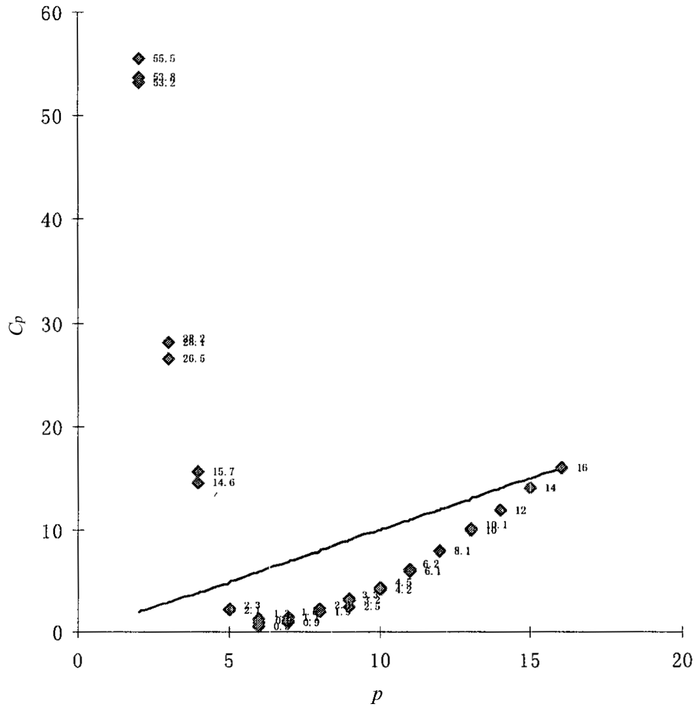

Fig. 2. The plot of C_p against p obtained for the first of the subset regressions.

studies highlight that the feed and the cutter nose radius are the two critical process parameters. It happens that models III-1 and III-3 both contain these two variable terms as can be seen in Table 1, and they are among these picked by MSEP. Between these two models, MAD, MAPE, and MRSE all favor model III-1. Notice that requiring the use of domain knowledge should not be considered to be a drawback of a modeling method.

Although speed seems to be a statistically significant process factor, neither of these two models contains speed. Therefore, we recommend model III-1, and it is given in Table 4. Model III-1 can be converted into its original form in Equation (19) following the form of Equation (15)

$$R_a = \frac{3.248 f^{2,364}}{r^{0.172}} e^{-\ln h\,(0.476 \ln f + 0.116 \ln d)} \tag{19}$$

Table 3. Summary of prediction error statistics from the candidate regression models.

	MAD	*MAPE*	*MRSE*	*MSEP*	*t-test (P value)*	*F-test (P value)*
Model I-1	0.289	1.680	0.071	0.011	0.334	0.560
Model I-2	0.290	1.649	0.072	0.011	0.407	0.548
Model I-3	0.287	1.630	0.071	0.011	0.376	0.546
Model II-1	0.281	**0.830**	0.070	0.010	0.316	0.422
Model II-2	0.338	1.049	0.079	0.010	0.304	0.418
Model II-3	**0.278**	0.832	**0.069**	0.010	0.353	0.401
Model III-1	0.350	1.659	0.090	**0.007**	0.755	0.086
Model III-2	0.345	1.587	0.089	**0.007**	0.597	0.091
Model III-3	0.366	1.707	0.094	**0.007**	0.672	0.093

Without considering the logarithmic terms, it resembles the form provided in Groover (2002, p. 570) and also in Kalpakjian and Schmid (2003, p. 436) as shown in Equation (20):

$$R_a = \frac{f^2}{32r}. \tag{20}$$

4.3. *Computational results obtained using NN models*

4.3.1. *Design of the NN experiments*

Based on the principles discussed in Section 3, a NN with one hidden layer is considered. Two levels of each of the following parameters were investigated in the construction and training of the network: (i) the hidden neuron size (3 versus 10); (ii) the training tolerance (10% versus 15%); and (iii) the testing tolerance (20 versus 40%). A hidden neuron size of three was determined based on Equation (11), and a neuron size of 10 was based on Equation (13). Selection of the above two neuron sizes also satisfies the lower and upper bound requirements for the training data size of 52 or 53. For example, given a neuron size of 10, the lower bound of training facts is $2(5 + 10 + 1) = 32 < 52$, and the upper bound is $10(5 + 10 + 1) = 160 > 52$.

A training tolerance of 10% and a testing tolerance of 40% are the default values from BrainMaker. Therefore, the 2^3 full factorial or eight neural nets are required to cross-examine the impact of changing the above parameters

Table 2. Summary of statistics from the candidate regression models

	s	R^2*(%)*	R^2*(adj)(%)*	*PRESS*	R^2*(pred)(%)*	*t-test(P value)*	*F-test(P value)*
Model I-1	0.3868	57.6	54.1	8.851 23	48.81	0.980	0.048
Model I-2	0.3873	57.5	54.0	8.882 27	48.63	0.874	0.047
Model I-3	0.3876	57.4	53.9	8.880 89	48.64	0.926	0.046
Model II-1	0.3905	59.1	55.7	8.918 37	50.15	0.908	0.062
Model II-2	0.3905	59.1	55.7	8.890 86	50.30	0.930	0.064
Model II-3	0.3916	58.9	55.4	8.988 07	49.76	0.986	0.059
Model III-1	**0.3131**	65.2	**62.3**	**5.872 72**	**56.54**	0.784	0.119
Model III-2	**0.3132**	65.2	**62.3**	5.873 17	**56.54**	0.815	0.130
Model III-3	0.3137	65.0	62.1	6.040 59	55.30	0.931	0.122

Table 4. The best predictive regression model (model III-1 from Table 1) users

$y = 1.18 + 2.36x_2 - 0.172x_3 - 0.467x_1x_2 - 0.116x_1x_4$ i.e., ($Ln(R_a) = 1.18 + 2.36Ln(Feed) - 0.172Ln(Radius) - 0.467 Ln(Hardness) Ln(Feed) - 0.116Ln(Hardness) Ln(Depth)$)

Predictor	*Coef*	*Se coef*	*T*	*P*
Constant	1.1781	0.1725	6.83	0.000
*X*2	2.3635	0.3692	6.40	0.000
*X*3	−0.172 14	0.041 61	−4.14	0.000
*X*1*X*2	−0.466 66	0.84 43	−5.53	0.000
*X*1*X*4	−0.116 25	0.29 52	−3.92	0.000

$S = 0.3131$ R-Sq = 65.2% R-Sq (adj) = 62.3%
PRESS = 5.87272 R-Sq (pred) = 56.54%

on the NNs prediction performance. In a threefold CV, a total of 24 NNs were trained and evaluated.

4.3.2. *Computational results obtained for the NN*

No data transformation was applied here since the NNs handle nonlinearity well. A summary of the prediction error statistics based on Equations (5)–(7) from the 24 NN models is given in Table 5. In addition, the *P* values of the paired *t*-test and the *F*-test are also provided in Table 5. Notice that the prediction error statistics in Table 5 were obtained when the 24 models from Table 3 were applied for prediction with the respective independent validation data set. Among the free parameters in Table 5, the first two parameters are the used training and testing tolerances, respectively.

On the basis of the *P* values of the paired *t*-test and the *F*-test in Table 5, each of the 24 models provides predictions that are statistically the same as the observations reserved for validation, since their *P* values are all greater than 0.05. Therefore, all the 24 models have a satisfactory goodness of prediction. When evaluating these 24 models based on the three prediction error statistics in Table 5, model I-6 was selected as the best model since it produces the smallest prediction errors in all of three measures: MAD, MAPE, and MRSE. The second best model is model I-1, which has the second smallest prediction errors. Consequently, we recommend the use of NN model I-6 to predict turning surface roughness values.

4.4. *Comparison of two learning schemes*

As discussed in Section 2.3, the statistical tests can also be applied to compare multiple learning schemes. Due to

Table 5. Summary of the prediction error statistics from the 24 NN models

	Free parameters Training/Testing/ number of neurons	*MAD* (μm)	*MAPE* (%)	*MRSE* (μm)	*P value*	
					Paired t-test	*F-test*
		Sets 1 and 2 for training and testing				
Model I-1	0.1/0.2/3	0.323	0.247	0.077	0.231	0.348
Model I-2	0.1/0.4/3	0.336	0.240	0.083	0.611	0.174
Model I-3	0.15/0.2/3	0.330	0.260	0.079	0.596	0.100
Model I-4	0.15/0.4/3	0.380	0.291	0.088	0.602	0.158
Model I-5	0.1/0.2/10	0.332	0.247	0.083	0.709	0.509
Model I-6	**0.1/0.4/10**	**0.311**	**0.236**	**0.074**	**0.522**	**0.519**
Model I-7	0.15/0.2/10	0.361	0.261	0.090	0.930	0.412
Model I-8	0.15/0.4/10	0.343	0.241	0.089	0.921	0.458
		Sets 1 and 3 for training and testing				
Model II-1	0.1/0.2/3	0.435	0.320	0.116	0.496	0.339
Model II-2	0.1/0.4/3	0.436	0.306	0.114	0.892	0.577
Model II-3	0.15/0.2/3	0.476	0.367	0.116	0.684	0.887
Model II-4	0.15/0.4/3	0.551	0.431	0.132	0.228	0.573
Model II-5	0.1/0.2/10	0.550	0.481	0.163	0.942	0.136
Model II-6	0.1/0.4/10	0.613	0.511	0.173	0.649	0.063
Model II-7	0.15/0.2/10	0.579	0.491	0.164	0.864	0.123
Model II-8	0.15/0.4/10	0.663	0.547	0.178	0.539	0.069
		Sets 2 and 3 for training and testing				
Model III-1	0.1/0.2/3	0.503	0.284	0.161	0.312	0.112
Model III-2	0.1/0.4/3	0.429	0.291	0.143	0.488	0.060
Model III-3	0.15/0.2/3	0.472	0.290	0.150	0.381	0.059
Model III-4	0.15/0.4/3	0.471	0.265	0.158	0.224	0.082
Model III-5	0.1/0.4/10	0.452	0.259	0.155	0.251	0.151
Model III-6	0.1/0.4/10	0.470	0.299	0.159	0.455	0.173
Model III-7	0.15/0.2/10	0.403	0.235	0.147	0.409	0.206
Model III-8	0.15/0.4/10	0.406	0.242	0.147	0.330	0.115

Table 6. Hypothesis testing statistics of predictions from the two competitive models

	P value		
95% C. I.	*Prediction*	*Error*	*Relative error*
Mann-Whitney test for median	0.618	0.027	0.081
Variance via Levene's test	0.356	0.129	0.623

non-normality in the data, the Mann-Whitney test was used to compare the means of the two competitive models and Levene's test was applied to compare the variances of the two models, one each from the RA and NN models. In addition to comparing the predictions from the two models, the absolute error and relative error to the respective observed value were also compared to obtain a better understanding of the schemes. The statistics from these three sets of median and variance tests are summarized in Table 6.

Inspection of the *P* values of the statistical tests listed in Table 6 reveals some differences. The *P* values that result from comparing either the prediction or the relative error between predictions to the corresponding observations are greater than 0.05, indicating that the best RA and NN models performed statistically the same in terms of these two error evaluators. However, the performance of these two models differs statistically when the absolute error between the prediction and the observation is used as the criterion, since the *P* values from the variance test are smaller than 0.05. A close examination of the hypothesis test results reveals that the mean absolute error and standard deviation of these absolute errors for the RA approach are statistically greater than those to the NN approach.

Which criterion should be used to pick the best technique? Results from the surface roughness study indicate that the relative error is more important than the absolute error. Viewed this way, the two learning schemes perform statistically the same. The next question then is which model should be used in practice when predicting turning surface roughness values. The RA model is preferred, since the NN model is a black-box technique and therefore does not provide any insights into potential process improvements. However, if this study had shown that the NN models were superior to the RA models then this recommendation would not have been made.

5. Conclusions

This paper has reviewed and evaluated important issues and techniques involved in model construction, selection, and validation for two competing data modeling schemes: RA models and NNs. Data obtained from designed experiments on a turning surface roughness study were used to illustrate the model selection and validation procedures. Although different approaches were used to develop the RA and NN models, the same techniques of CV, prediction error statistics, factorial design of physical and computational experiments and hypothesis tests were used to evaluate the candidate models and select the best one. Hypothesis testing was used to compare the two learning schemes based on the prediction performances of the individual best models of the two competing learning schemes. The outlined procedures in this paper for selection and validation of predictive models may be valuable for similar applications in the analysis and modeling of complex manufacturing processes and systems.

Acknowledgements

The authors are grateful to the reviewers for their insights in revising the paper. This research has been partially funded by the Bradley University Heuser Research Awards grant 25-13-755 and No. 25-13-757 and Caterpillar Fellowship grant 25-11-154 awarded to Jack Feng. We are grateful to Professor Alan Miller for his valuable comments on the first draft and for verifying our models with his Fortran codes, some of which are located at http://users.bigpond.net.au/amiller.

References

Baum, E.B. and Haussler, D. (1989) What net size gives valid generalization? *Neural Computation*, **1**(1), 151–160.

Boothroyd, G. and Knight, W.A. (1989) *Fundamentals of Machining and Machine Tools*, Marcel Dekker, New York, NY.

Breiman, L. and Spector, P. (1992) Submodel selection and evaluation in regression: the ***X***-random case. *International Statistics Review*, **60**(3), 291–319.

Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Inference: A Practical Information—Theoretic Approach*, 2nd edn., Springer-Verlag, New York, NY.

Coit, D.W., Jackson, B.T. and Smith, A.E. (1998) Static neural network process models: considerations and case studies. *International Journal of Production Research*, **36**(11), 2953–2967.

Draper, N.R. and Smith, H. (1998) *Applied Regression Analysis*, 3rd edn., Wiley, New York, NY.

Efron, B. and Tibshirani, R. (1998) *An Introduction to the Bootstrap*, Chapman & Hall/CRC, Boca Raton, FL.

Feng, C.-X. (2001) Experimental study of the effect of turning parameters on surface roughness in finish turning, in *Proceedings of the 2001 Industrial Engineering Research Conference*, Institute of Industrial Engineers, Norcross, GA, paper 2036.

Feng, C.-X. and Wang, X.-F. (2002a) Digitizing uncertainty modeling for reverse engineering applications: regression vs. neural networks. *Journal of Intelligent Manufacturing*, **13**(3), 189–199.

Feng, C.-X. and Wang, X.-F. (2002b) Development of empirical models for surface roughness prediction in finish turning. *International Journal of Advanced Manufacturing Technology*, **20**(5), 348–356.

Feng, C.-X. and Wang, X.-F. (2003) Surface roughness predictive modeling: neural networks versus regression. *IIE Transactions*, **35**(1), 11–27.

Feng, C.-X., Wang, X.-F. and Yu, Z. (2002) Neural networks modeling of honing surface roughness parameters defined by ISO13565. *SME Journal of Manufacturing Systems*, **21**(5), 398–408.

Feng, C.-X. and Yu, Z. (2003) Neural networks modeling of turning surface roughness parameters defined by ISO13565. *Transactions* *of the NAMRI/SME*. Technical Paper No. MS03-202, Society of Manufacturing Engineers, Dearborn, MI, pp. 467–474.

Feng, C.-X., Yu, Z. and Wang, J.-H. (2004) Validation and data splitting in predictive regression modeling of honing surface roughness data. *International Journal of Production Research*, **43**(8), 1555–1571.

Gershenfeld, N. (1999) *The Nature of Mathematical Modeling*, Cambridge University Press, Cambridge, UK.

Gilmour, S.-G. (1996) The interpretation of Mallows C_p-statistic. *The Statistician*, **45**(1), 49–56.

Girossi, F. and Poggio, T. (1989) Representation qualities of neural networks: Kolmogorov's theorem is irrelevant. *Neural Computation*, **1**(4), 465–469.

Gorman, J.-W. and Torman, R.-J. (1966) Selection of variables for fitting equations to data. *Technometrics*, **8**(1), 27–51.

Groover, M.-P. (2002) *Fundamentals of Modern Manufacturing*, 2nd edn., Wiley, New York, NY.

Groth, R. (1998) *Data Mining: A Hands on Approach for Business Professionals*, Prentice Hall, Upper Saddle River, NJ.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY.

Secht-Nelson, R. (1987) Kolmogorov's mapping neural network existence theorem, in *Proceedings of the 1st IEEE Annual International Conference on Neural Networks*, IEEE Press, Piscataway, NJ, pp. III.11–III.14.

Kalpakjian, S. and Schmid, S.R. (2003) *Manufacturing Processes for Engineering Materials*, 4th edn., Prentice Hall, Upper Saddle River, NJ.

Kapse, P. (2001) The effect of turning parameters on surface roughness in finish turning. MS Project Report, Department of Industrial and Manufacturing Engineering, Bradley University, Peoria, IL.

Kolen, J.-F. and Pollack, J.B. (1990) Backpropagation is sensitive to initial conditions. *Complex Systems*, **4**(3), 269–280.

Kurkova, V. (1991) Kolmogorov's theorem is relevant. *Neural Computing*, **3**(4), 617–622.

Kusiak, A. (2000a) *Computational Intelligence in Design and Manufacturing*, Wiley, New York, NY.

Kusiak, A. (2000b) Decomposition in data mining: an industrial case study. *IEEE Transactions on Electronics Packaging Manufacturing*, **23**(4), 345–353.

Kusiak, A. and Kurasek, C. (2001) Data mining of printed circuit boards. *IEEE Transactions on Robotics and Automation*, **17**(2), 191–196.

Lawrence, J. (1994) *Introduction to Neural Networks: Design, Theory, and Applications*, 6th edn., California Scientific Software, Nevada City, CA.

Lawrence, J. and Fredrickson, J. (1998) *BrainMaker User's Guide and Reference Manual*, 7th edn., California Scientific Software, Nevada City, CA.

Lippmann, R.P. (1987) An introduction to computing with neural nets. *IEEE Acoustics, Speech and Signal Processing Magazine*, April, **3**(4), 4–22.

Mallows, C.L. (1973) Some comments on C_p. *Technometrics*, **15**(4), 661–675.

Mallows, C.L. (1995) More comments on C_p. *Technometrics*, **37**(4), 362–372.

Mallows, C.L. (1997) C_p and prediction with many regressors: comments on Mallows (1995). *Technometrics*, **39**(1), 115–116.

Marchandani, G. and Cao, W. (1989) On hidden nodes for neural nets. *IEEE Transactions on Circuits and Systems*, **36**(5), 661–664.

Miller, A.J. (2002) *Subset Selection in Regression*, 2nd edn., Chapman & Hall, Boca Raton, FL.

Miller, R.-G. (1974) The jackknife—a review. *Biometrika*, **61**(1), 1–15.

Mitchell, T.-M. (1997) *Machine Learning*, McGraw-Hill, New York, NY.

Montgomery, D.-C., Peck, E.-A., and Vining, G.-G. (2001) *Introduction to Linear Regression Analysis*, 3rd edn., Wiley, New York, NY.

Shaw, M. (1984) *Metal Cutting Principles*, Oxford University Press, Oxford, UK.

Swingler, K. (1996) *Applying Neural Networks: A Practical Guide*, Morgan Kaufmann, San Francisco, CA.

Thompson, M.-L. (1978a) Selection of variables in multiple regression: part I—a review and evaluation. *International Statistical Review*, **46**(1), 1–19.

Thompson, M.-L. (1978b) Selection of variables in multiple regression: part II—chosen procedures, computations and examples. *International Statistial Review*, **46**(2), 129–146.

Twomey, J.-M. and Smith, A.E. (1998) Bias and variance of validation methods for function approximation neural networks under conditions of sparse data. *IEEE Transactions on Systems, Man and Cybernetics*, Part C (Applications and Reviews), **28**(3), 417–430.

Wasserman, P.D. (1989) *Neural Computing: Theory and Practice*, Van Nostrand Reinhold, New York, NY.

Witten, I.-H. and Frank, E. (2000) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, CA.

Zhang, P. (1993) Model selection via multifold cross-validation. *Annals of Statistics*, **21**(1), 299–311.

Yu, Z.-G. (2003) Selection and validation of predictive regression and neural networks models for experimental data from machining surface roughness studies. MS thesis, Department of Industrial and Manufacturing Engineering, Bradley University, Peoria, IL.

Biographies

Chang-Xue Jack Feng is a Professor of Industrial and Manufacturing Engineering at Bradley University in Peoria, IL. He has been President of the IIE Central Illinois Chapter since 1999. He has won the College of Engineering Faculty Excellence Award for Research and Scholarship and the Caterpillar Inc. New Faculty Achievement Award for Scholarship, both from Bradley University. He was Assistant Professor of Industrial and Manufacturing Engineering between 1995 and 1998 at Penn State University (Berks), where he developed, named, and directed the William & Mary Hintz Manufacturing Technology Laboratory. His current research focuses on data mining, logistics engineering, and lean production. He is a senior member of ASQ, IIE, and SME. He has published extensively and over 25 of his papers have appeared in the *Annals of Operations Research, ASME Transactions, Computer-Aided Design, IEEE Transactions, IIE Transactions, International Journal of Production Research, Journal of Manufacturing Systems*, and other international journals. He currently serves on the Editorial Board of the *International Journal of Production Research*.

Zhi-Guang (Samuel) Yu is a Project Manager at Supply Chain Services International in Peoria, IL. He obtained his MS in Manufacturing Engineering from Bradley University and a BS in Material Science and Engineering from Tsinghua University, Beijing, China. He has co-authored several papers published in *IIE Transactions, SME Journal of Manufacturing Systems, International Journal of Production Research*, and *Transactions of the NAMRI/SME.*

Andrew Kusiak is a Professor in the Department of Mechanical and Industrial Engineering at the University of Iowa in Iowa City, IA. He is interested in the application of computational intelligence to automation, manufacturing, product development, and healthcare. He has published numerous books and technical papers in journals sponsored by professional societies, such as AAAI, ASME, IEEE, IIE, ESOR, IFIP, IFAC, INFORMS, ISPE, and SME. He speaks frequently at international meetings, conducts professional seminars, and consults for industrial corporations. He serves on the editorial boards of numerous journals, edits book series, and is the Editor-in-Chief of the *Journal of Intelligent Manufacturing*.

Contributed by Enterprise Computing Department