Classification of Dry-Milled Maize Grit Yield Groups Using Quadratic Discriminant Analysis and Decision Tree Algorithm

Kyung-Min Lee,[1] Timothy J. Herrman,[1,2] Scott R. Bean,[3] David S. Jackson,[4] and Jane Lingenfelser[5]

Cereal Chem. 84(2):152–161

ABSTRACT

A genetically and environmentally diverse collection of maize (*Zea maize* L.) samples was evaluated for physical properties and grit yield to help develop a standard set of criteria to identify grain best suited for dry-milling. Application of principal component analysis (PCA) reduced a set of approximately 500 samples collected from six states to 154 maize hybrids. Selected maize hybrids were placed into seven groups according to their dry-milled grit yields. Regression analysis explained only 50% of the variability in dry-milling grit yield. Patterns of differences in the physical properties for the seven grit yield groups implied that the seven yield groups could be placed into two or three groups. Using two pattern recognition techniques for improving classification accuracy, quadratic discriminant analysis and the classification and regression tree (CART) model, dry-milled grit yield groups were predicted. The estimated correct classification rates were 69–80% when the samples were divided into three yield groups and 81–90% when samples were divided into two yield groups. The results indicated the comparable success of both techniques and the superiority of the decision tree algorithm to quadratic discriminant analysis by offering higher accuracy and clearer classification rules in differentiating among dry-milled grit yield groups.

Corn dry-milling separates maize kernels into three main components (endosperm, germ, and pericarp), producing numerous product streams for use in food, animal feeds, and industrial products (Duensing et al 2003). Corn grits, meals, and flour are the primary products obtained from the endosperm. Of these, higher recovery of larger grits is desirable for dry-millers because of the greater economic value. Dry-milled products are different in proximate composition and physical properties (Duensing et al 2003). Previous research indicates that dry-milling yield determinants include maize genetics, bulk and kernel density, hardness, breakage susceptibility, protein content, drying temperature, weather, post-harvest conditions, and other influencing factors (Kirleis and Stroshine 1990; Wu and Bergquist 1991; Peplinski et al 1992; Duensing et al 2003).

Maize producers and corn dry-millers often use unofficial grades and tests to predict maize hybrid end-use processing performance because maize hybrids ranked with higher official grade do not always guarantee better suitability for the customers' use (Paulsen et al 2003). This means that in certain cases the intended end-use performance of maize determines the traits and tests that are important for the evaluation of maize quality. The early segregation of maize using quality-associated properties will increase its economic value. Little standardization exists among such unofficial physical kernel tests, analytical techniques, and "reference" processing methods. For example, the strength of various relationships (correlations) between hardness measurements and end-use processing performance reported in scientific literature varies tremendously (Paulsen and Hill 1985; Peplinski et al 1992; Pan 1996; Shandera et al 1997). The grain samples used to establish these relationships are critical. Samples from diverse genetic and environmental backgrounds will help establish which practical hardness tests are most useful and provide a foundation for establishing the fundamental physicochemical basis for those grain hardness traits important in predicting end-use performance.

Pattern recognition techniques have been recognized as useful tools for interpretation and classification of complex data with many variables. Two recognition techniques, discriminant analysis and decision tree algorithm, were employed in this study. Discriminant analysis requires two basic assumptions: a multivariate normal distribution and equal variance of data in every variable (Johnson 1998), whereas the decision tree is not based on a statistical procedure but formulates a searching process to find the solution (Witten and Frank 2000). Decision tree algorithm is often used in a variety of classification problems and can visualize classification rules perspicuously by splitting the given data set into branches. The tree continues to grow until it is terminated by predetermined stopping rules (Witten and Frank 2000).

A need exists for improved determination of maize physical properties associated with processing performance and easy-to-use predictive laboratory measures. The two objectives of this study were to establish a sample set of diverse genetic and environmental backgrounds, and to develop dry-milling classification and prediction models using pattern recognition techniques based on selected maize kernel physical properties. This will enable classification of predefined dry-milled grit yield groups and rapid prediction of unknown samples into predefined grit yield groups. By focusing on more relevant physical properties, the resulting classification rule and approach for differentiating maize samples for dry-milling would assist producers and processors by identifying maize lots most appropriate for shipping or dry-milling unit operations at any given point in the grain marketing system.

MATERIALS AND METHODS

Sample Selection

Over 500 maize hybrids with a broad genetic background of known pedigrees were planted at different locations in Illinois, Indiana, Iowa, Kansas, Kentucky, Missouri, and Nebraska in 2003. Harvested maize hybrids were tested using several rapid physical and spectral property measurements. A group of 114 samples was then identified from these hybrids using a multivariate statistical technique described in Lee et al (2005). In this procedure, the spectral data of maize samples were mathematically converted into principal component scores using near-infrared software. The first four principal components accounted for ≈95% of variability in

[1] Office of the Texas State Chemist, Texas Agricultural Experiment Station, College Station, TX 77841-3160.
[2] Corresponding author. Phone: 979-845-1121. Fax: 979-845-1389. E-mail: tjh@otsc.tamu.edu
[3] USDA-ARS, Grain Marketing and Production Research Center, Manhattan, KS 66502. Names are necessary to report factually on available data; however, the USDA does not guarantee the standard of a product, nor does the use of the name by the USDA imply any approval of the product to the exclusion of others that may also be suitable.
[4] Department of Food Science and Technology, University of Nebraska-Lincoln, Lincoln, NE 68583-0919. A contribution of the University of Nebraska Agricultural Research Division, supported in part by funds provided through the Hatch Act. Mention of a trade name, proprietary product, or company name is for presentation clarity and does not imply endorsement by the authors or the University of Nebraska.
[5] Department of Agronomy, Kansas State University, Manhattan, KS 66506.

doi:10.1094/CCHEM-84-2-0152

the original data. Therefore, four principal component scores were subsequently used in cluster analysis to group maize samples into several different groups consisting of spectrally analogous maize hybrids. Ward's minimum variance method appeared to perform better than other algorithms with respect to grouping maize hybrids naturally, resulting in nine total clusters. The samples for the first year of the study were randomly and proportionally selected from each cluster. The selected 114 maize hybrids were subjected to a laboratory dry-milling process. To retain the genetic and environmental diversity, while reducing the number of samples, principal component analysis (PCA) and cluster analysis were performed to further reduce the 114 maize hybrids for planting in 2004, resulting in 40 maize hybrids that were planted in Illinois, Indiana, Kansas, and Nebraska. The planted locations, states, and the number of hybrids in the regions during 2003 and 2004 are displayed on the map in Fig. 1.

Physical Properties

Maize hybrids harvested in 2003 and 2004 were evaluated for test weight (USDA 1990); time (sec) required to grind kernels measured by the Stenvert Hardness Tester (SHT) (Glenmills model V with a 2-mm screen) run at 3,600 rpm (Pomeranz et al 1985); specific density in a helium compression pycnometer (model 930, Beckman Instruments, Fullerton, CA) (Pomeranz et al 1984); 100 kernel weight using sound whole kernels free from defects (Dorsey-Redding et al 1990); and abrasiveness using the Tangential Abrasive Dehulling Device (TADD) (Venables Machine Works Ltd., Saskatoon, SK, Canada) (Wehling et al 1996). Near-infrared transmittance (NIT) (Grainspec, Multispec Ltd., Wheldrake, NY) was calibrated by the manufacturer and the Grain Quality Laboratory at Iowa State University against chemical methods and used to determine density, moisture, protein, starch, and oil content. Kernel size distribution was determined by a strand size shaker (Seedburo Equipment, Chicago, IL) and expressed as the percentage of samples/initial amount of sample over a grain dockage sieve with 6.75-mm diameter round holes. NIT spectroscopic data in a log [1/T] were collected using Infratec 1229 (Foss North America) with a 30-mm path sample holder. Ten individual scans were averaged for the sample spectrum. The range of collected spectra was 850–1,048 nm in 2-nm increments. Collected spectra data were converted into principal component scores using WINSI II software (v. 1.0, Foss NIRSystems, Infrasoft International, Silver Spring, MD).

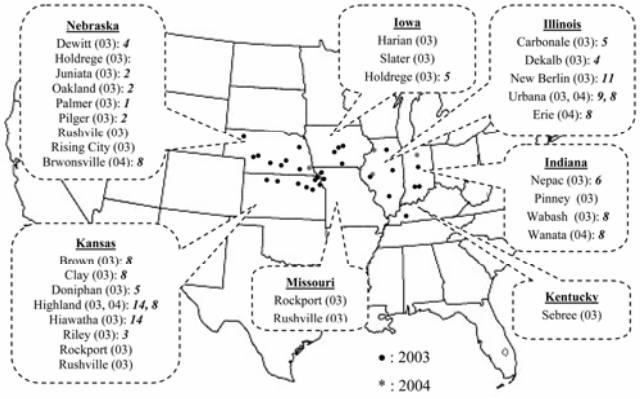

Fig. 1. Locations across several states where genetically diverse maize hybrids were planted during 2003 and 2004. Numbers in parentheses indicate planting year. Number of different hybrids selected for dry-milling during the two years denoted after the planting year.

Dry-Milling

All maize samples were cleaned with the MCI Kicker dockage tester (Mid-Continent Industries, Newton, KS) before dry-milling. The moisture content of a 1,000-g sample was determined with near-infrared transmittance (NIT) (Grainspec). Samples were shaken vigorously in a plastic jar, initially tempered to 16% moisture by the addition of water, then set aside for 30 min. After the first tempering, additional water was added to bring the sample moisture to 18%, followed by a 15-min rest period. The second tempered sample was milled using Allis experimental roll stands with a long-flow procedure that yielded snack grits with <1% fat in the grit extraction (Reddy 1996). Roll gaps, roll corrugations, roll differentials, and test sifter sieves in the dry-milling flow are illustrated in Fig. 2. The milling stages used in this study consist of 1 break (1BK), 2 break (2BK), 3 break (3BK), germ, 1 sizing (1SIZ), chunk, and 2 sizing (2 SIZ). The products produced from this dry-milling procedure are #1 grits (–10+14 Mesh), #2 grits (–14+26

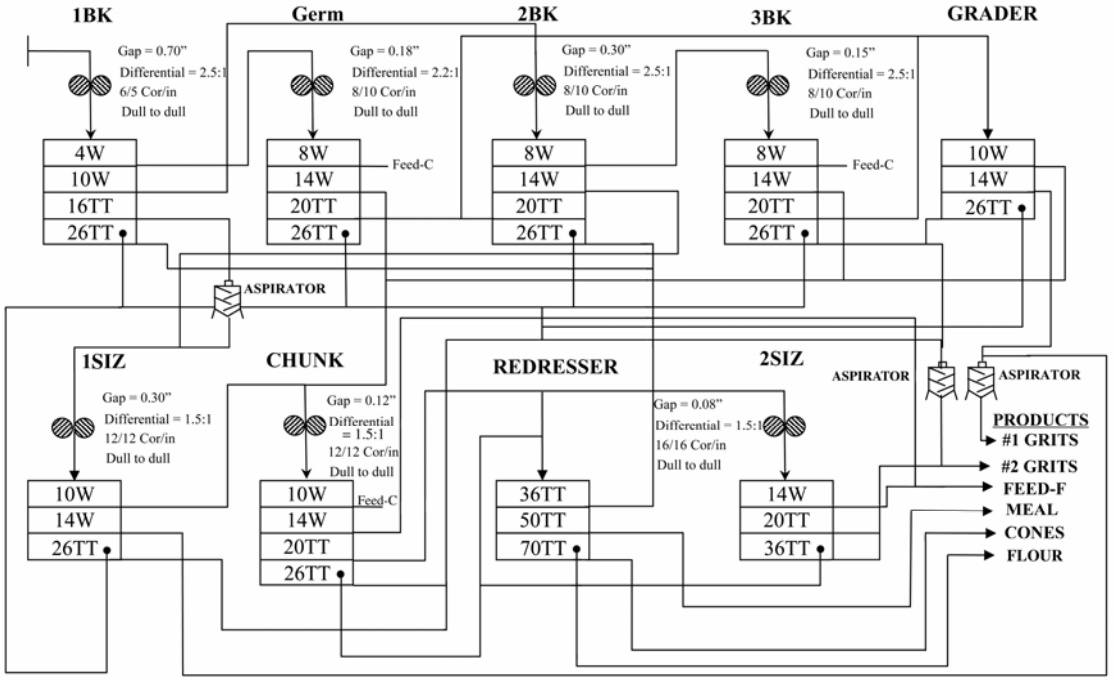

Fig. 2. Experimental maize dry-milling flow. Roll gaps, roll corrugations, roll differentials, test sifter sieves, and final milled products are illustrated.

Mesh and –14+36 Mesh), meal (–36+56 Mesh), cones (–50+70 Mesh), flour (–70 Mesh), and feeds (hulls, tipcap, and germ). A total dry-milled grit yield was calculated as the percentage of the sum of #1 and #2 grits/100 g of total products. The reproducibility standard deviation of this milling method was approximately <1% between individual results. Each maize hybrid was then assigned to one of seven predefined groups according to the calculated total dry-milled grit yield: <46, 46–50, 50–54, 54–58, 58–62, 62–66, and >66%. The seven grit yield groups were defined after creating a histogram of grit yields (in 1% increments) and frequency. The range of 54–56% total grit yield was in the center of the grit yield distribution. The seven defined yield groups at 42–70% allowed for a relatively sufficient and balanced number of observations among these yield groups, presumably providing more accurate and dependable information on the pattern distribution of the grit yield groups relative to grain physical properties.

Discriminant Analysis

Discriminant analysis was conducted to classify maize samples into a sample set of two grit yield groups and another of three grit yield groups. Linear discriminant analysis assumes the normal distribution of measurements and the equal variance and covariance matrices for groups to be separated (Johnson 1998). However, the application of quadratic discriminant analysis does not require the assumption of an identical covariance for each grit yield group. In this study, sample sizes of different dry-milled grit yield groups varied considerably, implying the unequal dispersion of measurements of the group. In addition, Bartlett's modification of the likelihood ratio test suggested using nonpooling covariance matrices. Therefore, a quadratic discriminant analysis was utilized for classification and prediction purposes. For the appropriate selection of the physical properties as input variables, univariate statistics of discriminant analysis were determined (SAS Institute, Cary, NC). After determining the most relevant sets of the variables, discriminant analysis model derived from a total of 154 samples was estimated using a jackknifed cross-validation method. Thereafter, 154 samples were separated randomly into a training data set (106 observations) and a test data set (48 observations). The training data set was used to build a classification model, whereas the test data set was used to estimate the predictive ability of the model. This model's classification ability was also evaluated using a jackknifed cross-validation method.

Decision Tree Algorithm

A decision tree algorithm was applied to classify and predict either two or three dry-milled grit yield groups based on measured physical properties. The best binary split was searched using the classification and regression tree (CART) method (Breiman et al 1984) which is available in SAS Enterprise Miner software. A cross-validation method was used to build the tree model that best fits the observed data. First, a decision tree model for grit yield groups was developed using the cross-validation method with all the samples. After classification of the entire data set, the data set was partitioned into a training data (consisting of 70% of the samples) and a test data set (consisting of the remaining samples) as in discriminant analysis. The training data was used to generate a tree model whose predictive ability was evaluated with the test data. In the search for a split point, the Gini index (or Gini impurity) was used as the splitting criterion (Breiman et al 1984). The Gini index is defined as

$$Gini(D) = 1 - \sum_{j}^{k} (p_j)^2$$

where p_j is the relative frequency of the group j in the whole dataset D. If all data at the node can be classified into one group, the Gini index is computed as zero. The data at each node (mother node) is split into two split nodes (child nodes) in which the data become more homogeneous. A goodness of split criterion was evaluated at the node using Gini indices. The split point is selected to maximize the decrease in deviance of $\Delta i(s,t)$ of a split s at node t. If the split s sends a data point to the left child node with a proportion P_L and to the right child node with a proportion P_R, the decrease in impurity of the split s is computed as

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R)$$

where $i(t)$ is a measure of a Gini index of the split s at the parent node t. This procedure is repeated for a new child node to grow the tree. When the tree level meets one of the predetermined stopping rules, such as a maximum tree level depth and a minimum number of data in child nodes, the node becomes a terminal node. Created classification trees were pruned by finding and eliminating weak links after building the larger tree models. Tree pruning prevents the decision tree model from overfitting the training data, which usually does not help fit other independent data. Different combinations of the tree splitting criteria were tested to find the best decision tree model that includes the most relevant physical properties at split points. Although all measured physical properties were initially used as input variables to create a decision tree model, only two or three of them were used in the final decision tree models. The detailed decision tree algorithm methodology used in this study was described in Breiman et al (1984) and Witten and Frank (2000). The results from decision tree analysis were compared with those from discriminant analysis relative to each technique's ability to determine correct classifications.

Statistical Analysis

All statistical analyses were conducted using SAS software. Mixed model analysis using the Proc MIXED procedure was performed to evaluate all variance components for each estimated physical property and to characterize the seven grit yield groups within a location and between locations. Total grit yield group,

TABLE I
Significance of Random and Fixed Effects by Physical Properties

	Random Effects[a]		Fixed Effects		
Physical Properties	**Year**	**Year-by-Location**	**Group (*G*)**	**Location (*L*)**	***G* × *L***
Protein content	<0.001	0.003	0.004	0.372	0.648
Test weight	0.999	0.999	<0.001	0.155	0.232
NIT density	<0.001	<0.001	<0.001	0.849	0.246
Pycnometer density	<0.001	<0.001	<0.001	0.187	0.738
Time to grind in SHT[b]	<0.001	<0.001	<0.001	0.959	0.002
TADD[c]	<0.001	<0.001	<0.001	0.414	0.103
100 Kernel weight	<0.001	<0.001	0.843	0.485	0.999
Kernel size distribution	0.011	<0.001	0.611	0.993	0.986

[a] Likelihood ratio test used to evaluate variance of random effects.
[b] Stenvert Hardness Tester (SHT).
[c] Tangential Abrasive Dehulling Device (TADD).

location, and their interaction were considered as fixed effects, while year and year-by-location were regarded as random effects by which the environment influences on the physical properties were incorporated into the mixed model. The mean difference of the seven grit yield groups was examined using least significant differences (LSD) at $\alpha = 0.05$. A likelihood ratio test was used to test the variance of the random effects. Correlation coefficients were determined between physical properties and total grit yields. Stepwise regression analysis was performed to find significant independent variables for the total grit yield using the Stepwise selection method implemented in the Proc REG procedure ($\alpha = 0.05$). Multiple linear regression analysis was conducted with the selected independent variables to predict the total grit yield.

RESULTS AND DISCUSSION

Random and Fixed Effects

The results from a likelihood ratio test for random effects, year and year-by-location, for the physical properties is presented in Table I. The effects of year and the year-by-location interaction were significant for all physical properties except test weight. Seasonal variation in grain end-use performance and its expression by location is common and has been documented in previous studies (Dombrink-Kurtzman and Bietz 1993; Shandera et al 1997).

P-values for the fixed effects grit yield group, location, and group-by-location interaction are presented for each physical property in Table I. The significant group-by-location interaction effect observed for time to grind ($P = 0.002$) indicates that the amount of energy used to grind grain of a particular maize hybrid varied between production areas. The group effect was significant ($P < 0.001$) for all physical property measurements except 100 kernel weight and kernel size distribution. The significant relationship between protein content, test weight, NIT density, pycnometer density, time to grind, TADD, and the seven groupings based on 4% grit yield increments (<46, 46–50, etc.) supports the decision to use these intervals. The absence of a significant location effect indicates a consistent relationship among the seven grit yield groups and physical kernel measurements.

Physical Properties

The seven grit yield groups showed differences ($P < 0.05$) in estimated physical properties that have been considered as factors directly or indirectly associated with maize dry-milling quality (Fig. 3A–H). Pattern differences among the grit yield groups in relation to physical properties indicate that multivariate techniques will likely improve data description and prediction as compared with regression analysis (Baker et al 1999).

Protein content. Protein content increased with an increase in the grit yield, but the difference in protein contents was not large among the seven grit yield groups (Fig. 3A). Samples in the 54–58% grit yield group had significantly higher protein contents than in the other yield groups ($P < 0.05$), followed by a decrease in protein content for the 58–62% grit yield group. Previous research documents a relationship between protein and maize kernel hardness, a determinant in dry-milled product yield (Shandera et al 1997). The results in this study reveal that the geographically and genotypically diverse sample set obtained for this study does not conform closely to prior research results utilizing fewer hybrids and growing locations.

Test weight. A significant increase ($P < 0.05$) in test weight was observed as the grit yield increased (Fig. 3B). However, the increase in test weight was not very large between the 50–54% and

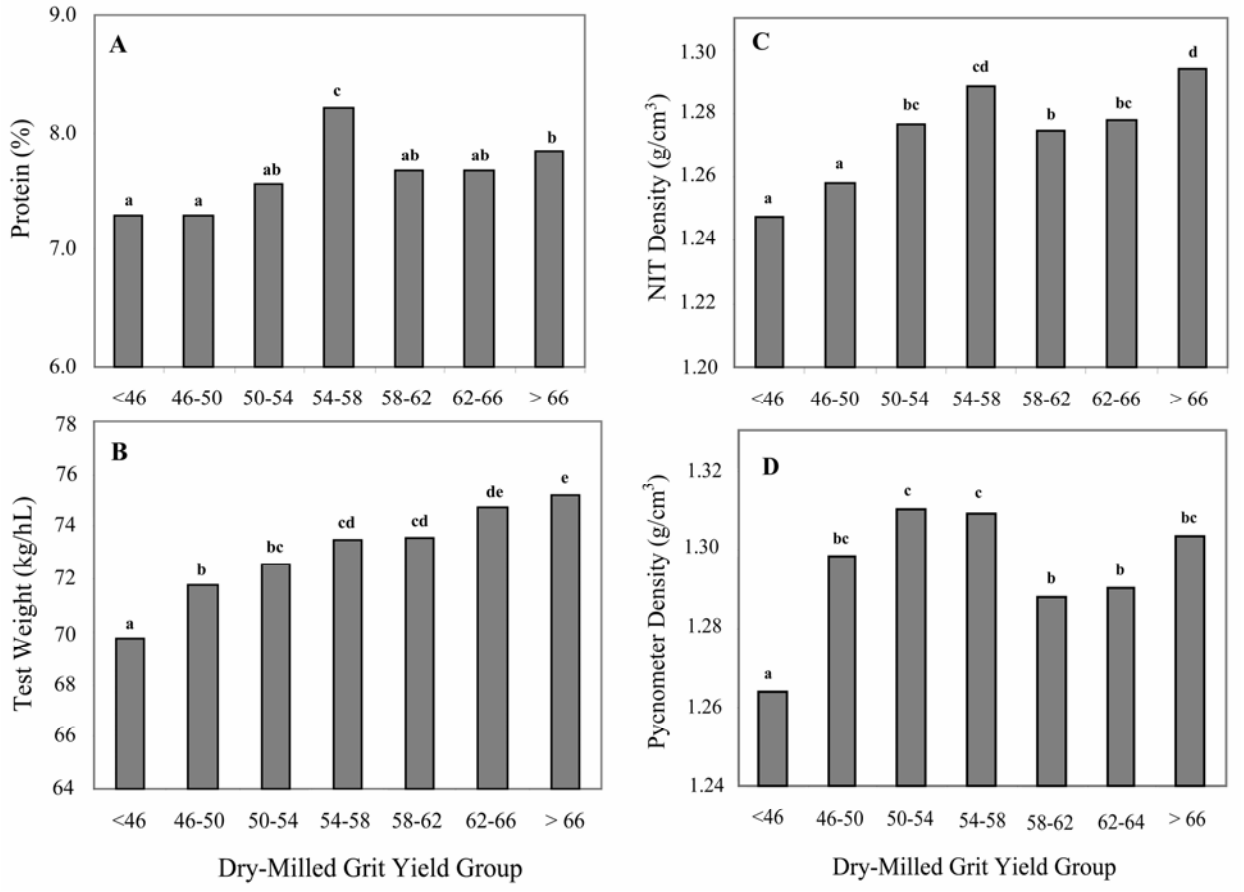

Fig. 3. Physical properties of 154 maize hybrids and significant differences among seven classified dry-milled grit yield groups. **A,** Protein (%); **B,** Test weight (kg/hL); **C,** NIT density (g/cm^3); **D,** Pycnometer density (g/cm^3). Means with the same letter are not significantly different at $\alpha = 0.05$.

58–62% yield groups, or between the 62–66% and >66% yield groups. Test weight (bulk density) is an important quality property used to determine maize grades and selling price (Duensing et al 2003). In previous studies, lower test weights resulted in lower prime and total grit yields during dry-milling processes (Paulsen and Hill 1985; Dorsey-Redding et al 1991; Peplinski et al 1992), which is consistent with our findings. The grit yield was second-best correlated with test weight ($P < 0.01$, $r = 0.423$) (Table II).

Kernel density. Maize kernel density was estimated by a helium compression pycnometer and NIT spectroscopy calibrated using the values measured with a nitrogen pycnometer (Micromeretics, WayCross, GA). Patterns of differences in kernel density among the grit yield groups appeared to be similar irrespective of the estimating methods (Fig. 3C, D). Kernel density greatly increased up to the 54–58% yield group, but decreased for the 58–62% yield group. The 50–54% and the 54–58% yield groups had significantly higher kernel densities than those of other yield groups.

Similar findings were reported in previous studies where maize kernel density showed a positive correlation with other hardness-associated properties that are important intrinsic traits closely associated with percentage of vitreous endosperm (Kirleis and Stroshine 1990; Wu and Berquist 1991; Shandera et al 1997). A higher ratio of vitreous to floury endosperm is preferred by dry-millers because it produces a higher percentage of large flaking grits with significant economic value.

Consequently, kernel density alone and with other hardness-associated properties showed the best predictive ability for dry-milled product yields and are frequently considered useful screening properties for evaluation of dry-milling quality (Kirleis and Stroshine 1990; Wu and Bergquist 1991).

Time to grind in SHT. An increase in grit yield was accompanied by an increase in time to grind in SHT (Fig. 3E). Time to grind for maize in the 46–50% grit yield group was not extremely different from that of the 50–54% yield group. However, a significant difference ($P < 0.05$) existed among other grit yield groups. Figure 3E implies that time to grind might be a variable with good discriminating power in differentiating the dry-milled grit yield groups. Time to grind is positively correlated with total grit yield ($P < 0.01$, $r = 0.672$) (Table II) and maize endosperm texture (Pomeranz et al 1985; Watson 2003), and negatively correlated with kernel moisture content (Pomeranz et al 1986). According to findings in previous studies (Pomeranz et al 1985; Kirleis and Stroshine 1990), longer time to grind is positively correlated with smaller total volume (lower column height) and higher ratio of coarse to fine particles collected in SHT receptacles This indicates a harder endosperm texture and a larger dry-milled grit yield.

Tangential Abrasive Dehulling Device (TADD). TADD index increased as the grit yield increased for the 54–58% yield group, and then decreased for the 58–62% yield group ($P < 0.05$) (Fig. 3F). A lack of statistical difference between the 50–54% yield group and the 54–58% yield group was observed, indicating that this physical test of kernel hardness was only capable of discriminating between two or three grit yield groups. TADD index is a fast and reproducible hardness test but it is rather sensitive to moisture content (Lawton and Faubion 1989) and kernel surface area (Shandera et al 1997).

Kernel moisture content was maintained at 13 ± 1%. Higher TADD index is generally correlated to higher protein content, higher test weight and kernel density, and lower percent floaters (Shandera et al 1997).

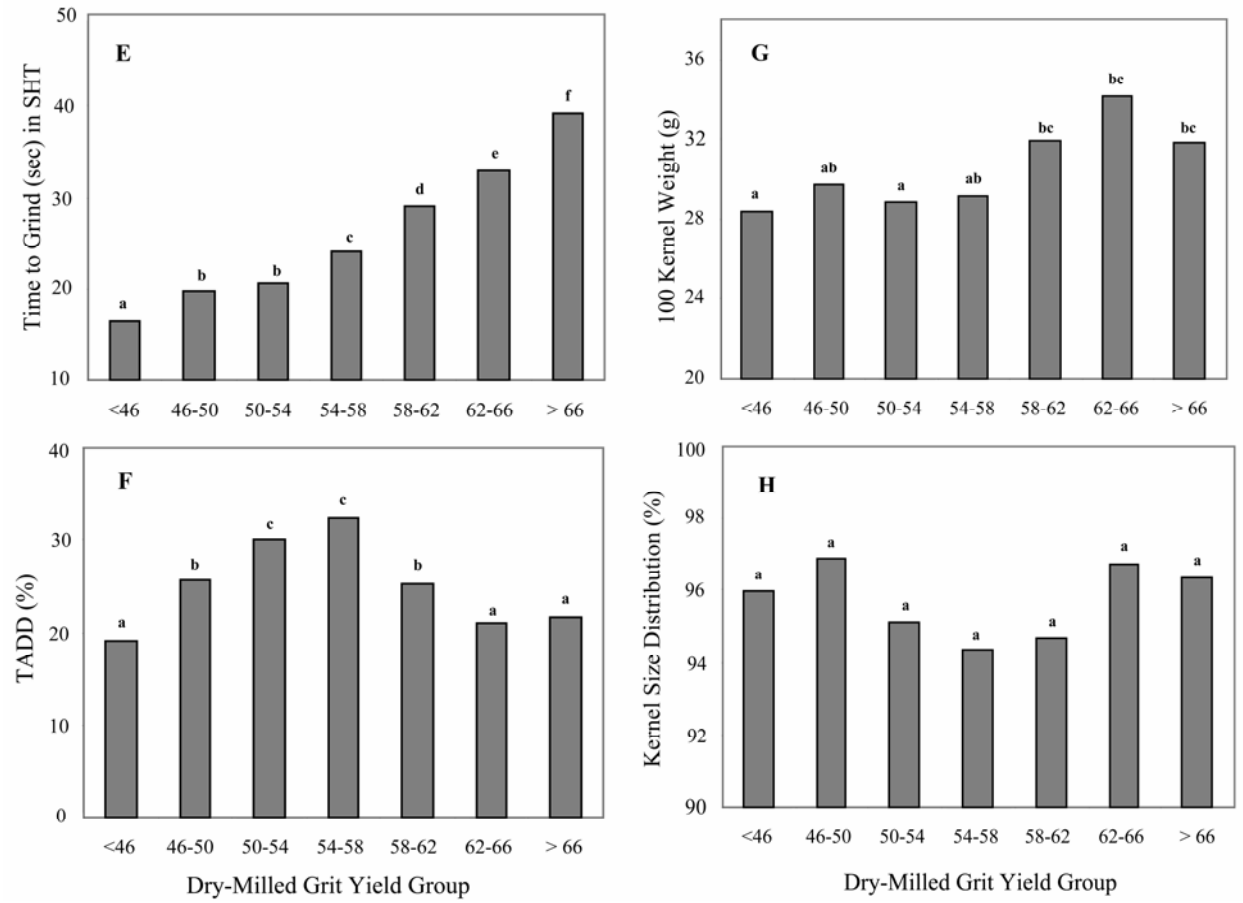

Fig. 3. (continued) Physical properties of 154 maize hybrids and significant differences among seven classified dry-milled grit yield groups. **E,** Time to grind in Stenvert Hardness Tester (SHT); **F,** Tangential Abrasive Dehulling Device (TADD, %); **G,** 100 kernel weight (g); **H,** Kernel size distribution (%). Means with the same letter are not significantly different at $\alpha = 0.05$.

100 Kernel weight and kernel size distribution. There was no significant difference in 100 kernel weight among the seven yield groups (Fig. 3G). Pomeranz et al (1985) reported a correlation between kernel weight and maize kernel hardness-associated properties such as near-infrared reflectance (NIR) at 1680 nm and SHT measurements at constant protein and ash contents, which implies an influence of kernel weight on dry-milling quality. The difference in kernel size distribution among the seven grit yield groups was apparent in Fig. 3H but not statistically significant ($P < 0.05$). The germ-endosperm ratio, oil content, hardness-associated NIT absorbance, and test weight can vary with kernel size, influencing maize endosperm texture and dry-milled products (Robutti 1995; Shandera et al 1997).

These results suggest a nonlinear relationship between physical properties and grit yield groups, and a similarity among closely located grit yield groups. Such observations may enable the original seven dry-milled grit yield groups to be reduced to a set of two yield groups (<58 and ≥58%) or a set of three yield groups (<50, 50–58, and ≥58%). Of these physical properties, time to grind in SHT, TADD index, NIT density, and test weight appeared to be important independent variables in differentiating two or three grit yield groups, as they had more conspicuously different patterns among the groups than the other variables. Improving model classification ability was attempted by using different combinations of these physical properties as input variables in pattern recognition techniques.

Regression Analysis

Stepwise regression analysis showed that significant independent variables in the selected model for the prediction of dry-milled grit yield were test weight, protein content, pycnometer density, time to grind in SHT, and kernel size distribution. Although such independent variables were most relevant to predicting dry-milled grit yield, the multinomial linear regression analysis indicated that 52.0% of the grit yield variability was explained by this regression model ($P < 0.05$, $R^2 = 0.52$)

Dry-milled grit yield (%) = 44.9 + 0.78 (test weight) + 1.33 (protein content) – 33.66 (pycnometer density) + 0.38 (time to grind in SHT) – 0.12 (kernel size distribution)

Adding other physical property variables did not improve the predictability of the model. In addition, the parsimonious variables in the model would not help breeders, producers, and processors to efficiently differentiate maize suitable for the end-use performance. This seems to support the need to explore other simple and easy-to-use techniques to identify maize hybrids best suited for dry-milling performance. The two pattern recognition techniques discussed in this study, if proven accurate, would be efficient and provide better results in classifying dry-milled grit yield by incorporating information from each measured physical property in predicting dry-milled grit yield. Implementing this approach could provide millers with simple and rapid methods to determine maize lots with superior dry-milling qualities without using an identity-preservation procurement system, and without additional costs.

Discriminant Analysis

A quadratic discriminant analysis is a general extension of a linear discriminant analysis that assumes the same variance-covariance matrix of different classes (Johnson 1998). The individual variance-covariance matrix of each class is used as a classification criterion in a quadratic discriminant analysis. Among several alternative classification rules used to discriminate among classes, the Bayes rule was used to compute the posterior probability to assign an observation x to a single class (G). According to this rule, given prior probabilities p_i and p_j, the observation x belongs to class G_i if

$$P(x/G_i)\cdot p_i > P(x/G_j)\cdot p_i \text{ for } i \neq j$$

where $P(x/G_i)\cdot$and $P(x/G_j)$ are the probability densities. A quadratic discriminant assigns the observation x to class G_i when the discriminant score $D_i(x)$, a measure of the generalized squared distance between x and class G, is minimized (Rao 1973; Johnson 1998)

$$D_i(x) = 0.5(x-\mu_i)'\sum_i^{-1}(x-\mu_i) + 0.5\log(|\sum_i|) - \log(p_i)$$

where μ_i is the mean of class i, and Σ_i is the population variance-covariance matrix of class G_i. The posterior probability for each of the possible classifications is then obtained using the computed discriminant score $D_i(x)$. An observation x is assigned to the class with the largest posterior probability. In a linear discriminant analysis, the notation Σ_i of the different population covariance matrix is replaced with Σ due to the same variance-covariance matrix assumption

$$D_i(x) = 0.5(x-\mu_i)'\sum^{-1}(x-\mu_i) - \log(p_i)$$

With all the physical properties tested in this study, the univariate and multivariate statistics for discriminant analysis of the models obtained for samples divided into either two or three grit yield groups were estimated to find the most relevant physical properties (Table III). Univariate statistics were estimated to test the equal means of the physical property for the grit yield groups, while multivariate statistics were used to investigate the main and interaction effect of the grit yield groups on all physical properties. Except for kernel size distribution, most of the physical property variables were statistically significant in sets divided into three grit yield groups by univariate statistics. As discussed previously regarding relationships between the physical properties and the grit yield groups, the most significant variables ($P < 0.001$) in univariate statistics for the three grit yield groups were time to grind in SHT ($F = 44.48$) followed by NIT density ($F = 19.56$), TADD ($F = 15.40$), and test weight ($F = 11.52$) in decreasing

TABLE II
Partial Correlation Coefficients of Maize Physical Properties at Constant Moisture Content[a]

	Test Weight	NIT Density	Pycnometer Density	Time to Grind in SHT[b]	100 Kernel Weight	Kernel Size Distribution	TADD[c]	Total Grit Yield
Protein content	–0.147	0.376**	0.181*	0.054	–0.422**	–0.204*	0.439**	0.205*
Test weight		0.561**	0.432**	0.417**	0.478**	0.098	0.212*	0.423**
NIT density			0.613**	0.291**	–0.035	–0.178*	0.634**	0.331**
Pycnometer density				–0.040	0.058	–0.066	0.610**	–0.025
Time to grind in SHT					0.333**	0.062	–0.164*	0.672**
100 Kernel weight						0.469**	–0.200*	0.215*
Kernel size distribution							–0.165*	–0.035
TADD								–0.050

[a] *, ** Significant at 0.05 and 0.01 level.
[b] Stenvert Hardness Tester (SHT).
[c] Tangential Abrasive Dehulling Device (TADD).

order of significance. Again, this result implies that time to grind in SHT might substantially contribute to the discriminating and predictive ability of the model. Univariate statistics of a quadratic discriminant model for two grit yield groups gave a slightly different result: five variables, including time to grind in SHT (F = 78.33), TADD (F = 12.16), 100 kernel weight (F = 11.16), test weight (F = 10.27), and pycnometer density (F = 6.09) were more relevant than the other variables. Time to grind in SHT had a much higher F-value than the other variables in the two grit yield groups, as well as the time to grind F-value for the three grit yield groups. NIT density in the model for the three grit yield groups appeared to be substituted with another true density measurement, pycnometer density, in the model for two yield groups. Pycnometer density, however, was not used as an input variable in the discriminant analysis because the F-value of pycnometer density was relatively lower than those of other selected variables. Two multivariate statistics in both classification models were significant ($P < 0.001$), demonstrating a difference among the grit yield groups in a set of physical properties. F-values of multivariate statistics in the model of two grit yield groups were greater than those for the three grit yield groups.

TABLE III
Univariate and Multivariate Statistics of Discriminant Analysis for Two and Three Dry-Milled Grit Yield Groups[a]

	Two Dry-Milled Grit Yield Group	Three Dry-Milled Grit Yield Group
Univariate		
Protein content	0.03	5.40**
Test weight	10.27**	11.52***
NIT density	0.23	19.56***
Pycnometer density	6.09*	8.18***
Time to grind in SHT[b]	78.33***	44.48***
TADD[c]	12.16***	15.40***
100 Kernel weight	11.16**	5.60**
Kernel size distribution	0.00	2.15
Multivariate		
Wilks' lamda	13.01***	8.98***
Pillai's trace	13.01***	8.68***

[a] *, **, ***, Significant at 0.05, 0.01, and 0.001 level.
[b] Stenvert Hardness Tester (SHT).
[c] Tangential Abrasive Dehulling Device.

After investigating the contribution of physical property variables to the classification model based on the univariate and multivariate statistical results, time to grind, NIT density, TADD, and test weight were selected for the classification of three grit yield groups. Time to grind, TADD, 100 kernel weight, and test weight were selected for the classification of two grit yield groups. Less relevant physical property variables from the discriminant analysis point of view were eliminated in developing the model. The selected four variables in both classification models resulted in better separation among either the two or three grit yield groups. In preliminary trials, correct classification rates of discriminant models created by using the selected physical property variables were higher than those for models built by all measured physical property variables (Table IV). Furthermore, combinations of selected variables based on the results from a stepwise regression analysis did not show higher correct classification rates for two or three grit yield groups than those from the univariate statistics (Table IV). Correct classification rates were almost equal between grit yield groups with a sufficient number of measurements, regardless of the number of variables used for those models, indicating the importance of the number of observation in establishing the model. In grit yield groups with a small number of measurements, the models developed with eight variables had more accurate classification and predictive abilities than those with five variables from the stepwise regression analysis, leading to slightly better overall model accuracy.

A quadratic discriminant analysis for three grit yield groups also derived two canonical variables, the quadratic combination of the original variables. The two canonical variables were very significant ($P < 0.001$), indicating a significant contribution of these variables to the discrimination among three yield groups. The first function accounted for 72% of the variation in grit yield groups. Canonical correlation, the measure of the association between the function and dry-milled grit yields, had a rather low correlation (r = 0.44) between the second canonical discriminant function and three grit yield groups.

This suggests that better discrimination between grit yield groups might be obtained when the posterior probability criterion and the discriminant score are used. Scattered plots created by canonical discriminant scores did not show good discriminating power among the grit yield groups (data not shown).

TABLE IV
Correct Classification Rates of Two or Three Dry-Milled Grit Yield Groups Estimated Using Discriminant Analysis with All Eight Variables and Selected Five Variables by Stepwise Regression Analysis

Actual Group	8 Variables			5 Variables[a]		
	All Samples (%)	Training Set (%)	Test Set (%)	All Samples (%)	Training Set (%)	Test Set (%)
Three yield groups						
<50	53.6	36.8	55.6	64.3	47.4	33.3
50–58	72.6	74.1	76.9	66.7	74.1	80.7
≥58	69.0	72.4	69.2	59.5	31.0	61.5
Total	68.2	67.0	71.0	64.3	67.0	67.0
Two yield groups						
<58	89.3	89.6	91.3	91.1	89.6	91.4
≥58	66.7	72.4	61.5	59.5	65.5	53.9
Total	83.1	84.9	83.2	82.5	83.0	81.1

[a] Selected five variables: time to grind in Stenvert Hardness Tester (SHT); test weight; protein content; pycnometer density; kernel size distribution.

TABLE V
Correct Classification Rates for Three Dry-Milled Grit Yield Groups Estimated Using Discriminant Analysis with Time to Grind, NIT Density, TADD, and Test Weight Variables

Actual Group	Predicted Group (all samples)				Predicted Group (training set)				Predicted Group (test set)			
	<50	50–58	≥58	Total (% correct)	<50	50–58	≥58	Total (% correct)	<50	50–58	≥58	Total (% correct)
<50	15	11	2	28 (53.6)	11	7	1	19 (57.9)	4	5	0	9 (44.5)
50–58	10	69	5	84 (82.1)	5	49	4	58 (84.5)	2	21	3	26 (80.8)
≥58	1	12	29	42 (69.1)	1	7	21	29 (72.4)	1	4	8	13 (61.5)
Total	26	92	36	154 (73.4)	17	63	26	106 (76.4)	7	30	11	48 (69.0)

The samples belonging to each grit yield group were not distinctively separated, implying that selected independent variables were not best for a differentiation between grit yield groups.

With the samples divided into three grit yield groups, two quadratic discriminant models built with all the samples and a training data set exhibited similar classification abilities (Table V). Good classification ability was observed for the 50–58% yield group in both models, while samples belonging to <50% and ≥58% yield groups were often improperly classified, which appears to be associated with number of observations available in each group for model development. Sufficient numbers of observations may further improve the performance of those models. The overall estimated predictive ability of the classification model created by the training data set was 69.0% for test data set. The best predictive ability was obtained for 50–58% yield group samples, which were correctly placed for 21 of 26 samples (80.8% correct classification rate). Generalized Mahalanobis squared distances between grit yield groups are shown in Table VI. The 50–58% yield group had a greater distance value against the <50% group than against the ≥58% group. This may imply a more feasible combination of the 50–58% group with the <50% group rather than with the ≥58% group if dry-milled grit yields needed to be categorized into two yield groups. As noticed, the matrices of dry-milled yield groups are not symmetric. This is caused by different distance weighing between each pair of groups due to the unequal variance and covariance matrices of yield groups.

Two grit yield groups improved the correct classification rates by more than 10% (Table VII). Correct classification rates using the quadratic discriminant model for the <58% yield group were close to 94%, as estimated using a jackknifed cross-validation method. The improvement in classification and predictive abilities of the models observed with the <58% group might be due to the larger number of measurements compared with models for three grit yield groups. However, the ≥58% yield group showed the same classification and predictive ability in all data sets. With a test data set, 33 of 35 samples were correctly placed for the <58% yield group, yet the ≥58% yield group samples were highly misclassified, resulting in a 85.3% overall correct classification rate. These observations suggest that the large difference among sample sizes of grit yield groups largely contributes to the variation among the correct classification rates.

Decision Tree Algorithm

A decision tree algorithm generates clear rules from the training data to classify and predict the data into different groups or categories, while simultaneously identifying the important variables. Each selected variable is assigned to a split point at each node. This study applied a decision tree algorithm for classifying and predicting maize samples into either two or three grit yield groups. With the use of all samples to develop a decision tree, data was partitioned into different ratios of training data to test data to select and determine the optimal partition of data into two sets. Partitioning data into a 70% observation training data set and a 30% observation test data set appeared to be optimal because this ratio provided a slightly better accuracy and a simpler sub-tree than other data set ratios. Among sampling methods used in data partitioning, a stratified sample method was employed because this method could sustain the ratio of the dry-milled grit yield groups in both the training data set and the test data set, improving the model's classification accuracy. The

TABLE VI
Generalized Mahalanobis Squared Distance Between Dry-Milled Grit Yield Groups

	All Samples			Training Set		
	<50	50–58		<50	50–58	
Actual Group	<58[a]		≥58	<58[a]		≥58
Three yield groups						
<50	2.54	2.69	6.82	2.75	3.15	6.40
50–58	4.97	0.76	4.77	5.27	0.49	4.48
≥58	15.32	5.23	3.29	15.85	5.37	3.02
Two yield groups						
<58	12.16		16.40	12.18		15.94
≥58	17.42		14.82	17.88		14.47

[a] Set in two dry-milled grit yield groups (<58% vs. ≥58%).

TABLE VII
Correct Classification Rates of Two Dry-Milled Grit Yield Groups Estimated Using Discriminant Analysis with Time to Grind, TADD, 100 Kernel Weight, and Test Weight Variables

	Predicted Group (all samples)			Predicted Group (training set)			Predicted Group (test set)		
Actual Group	<58	≥58	Total (% correct)	<58	≥58	Total (% correct)	<58	≥58	Total (% correct)
<58	105	7	112 (93.8)	72	5	77 (93.5)	33	2	35 (94.3)
≥58	13	29	42 (69.1)	8	21	29 (72.4)	5	8	13 (61.5)
Total	118	36	154 (87.0)	80	26	106 (87.7)	38	10	48 (85.3)

TABLE VIII
Correct Classification Rates of Three Dry-Milled Grit Yield Groups Estimated Using Decision Tree Algorithm with TADD, Time to Grind, and NIT Density Variables

	Predicted Group (all samples)				Predicted Group (training set)				Predicted Group (test set)			
Actual Group	<50	50–58	≥58	Total (% correct)	<50	50–58	≥58	Total (% correct)	<50	50–58	≥58	Total (% correct)
<50	16	11	1	28 (57.1)	11	7	1	19 (57.9)	5	4	0	9 (55.6)
50–58	5	74	5	84 (88.1)	3	53	2	58 (91.4)	2	21	3	26 (80.8)
≥58	3	11	28	42 (66.7)	1	7	21	29 (72.4)	2	4	7	13 (53.8)
Total	24	96	34	154 (76.6)	15	67	24	106 (80.2)	9	29	10	48 (68.8)

observations on the physical property variables were kept in an original form before building a decision tree model because the transformation of eight physical property variables did not improve the model's fit. The decision tree model built using all samples was generated after the average correct classification rates were obtained through a cross-validation procedure. The classification rule best fitted with the test data set was also obtained when the data was divided into training and test data sets. With the model derived from all samples, tree growth stopped for the three grit yield groups at four tree levels, while the stopping rules halted growth for two grit yield groups at three tree levels (Figs. 4 and 5).

Only three of the eight physical property variables, including time to grind, TADD, and NIT density were used in the decision tree built using all the samples to differentiate samples when divided into three grit yield groups (Fig. 4). Time to grind was the best discriminant variable in the decision tree model. This physical property has been widely accepted as a parameter associated with maize kernel hardness and dry-milling evaluation in previous studies (Pomeranz et al 1985; Shandera et al 1997); it also served as a good predictor of dry-milled grit yield in this study. Kernel density (Kirleis and Stroshine 1990; Wu and Bergquist 1991) and TADD (Wehling et al 1996; Shandera et al 1997) were also identified as the better predictors directly related to kernel hardness and dry-milling characteristics. In the resulting tree, two homogeneous terminal nodes (2 and 4) were labeled as being in the 50–58% yield group. A decision tree derived from the training and the test data sets had a similar tree structure to the decision tree using all the samples with the same time to grind threshold value at the split point, but the TADD value was 32.88% rather than 24.93%. With a decision tree model built using all 154 samples, 76.6% of samples were correctly classified when the samples were placed in three grit yield groups (Table VIII). As in discriminant analysis, samples belonging to the 50–58% yield group were placed with close to 90% accuracy, while the poorest classified maize hybrids were in the <50 and ≥58% yield groups. Similar results were observed in the classification model derived from the training and test data sets. To test the decision tree model derived using a training set with "unknown" samples, a new test data set was developed which included 30% of the original 154 samples. Only 69% samples were correctly classified with this sample set. Because the model was built with relatively sufficient samples pertaining to the 50–58% yield group and the model showed high correct classification rates, the "unknown" 50–58% yield group samples were usually correctly placed, but samples belonging to the other groups were highly misclassified. Compared with the accuracy in discriminant analysis, decision tree models typically had higher correct classification rates for <58% and 50–58% groups in all data sets, but slightly lower correct classification rates for the ≥58% group for both the full and the test data sets.

Decision tree models were also developed to classify and predict total grit yield when the samples were divided into two grit yield groups: <58 and ≥58%. Decision tree models built using all samples for differentiation into two grit yield groups were built only with time to grind and TADD variables, presumably because fewer groups were classified (Fig. 5). The first physical property variable used to split the original group was time to grind, the same variable as the tree model built for differentiating three grit yield groups. With a decision tree built using the training and the test data sets, the variable and its threshold value at each split point

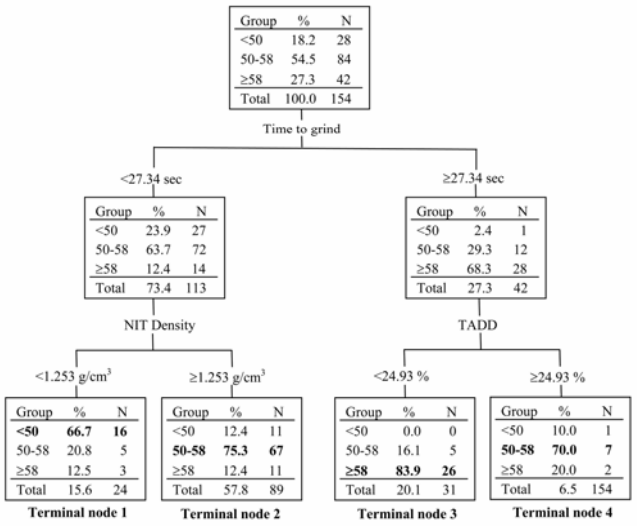

Fig. 4. Decision tree diagram for classification of 100% maize samples into three dry-milled grit yield groups. Each node includes the predefined grit yield groups and their percentages and the numbers at different tree levels. Variable used at a split point is located below in the middle of each node. Threshold values above each node are the point at which the split of the data occurs. Samples are classified into either left or right child node from the mother node according to the threshold values of the variable.

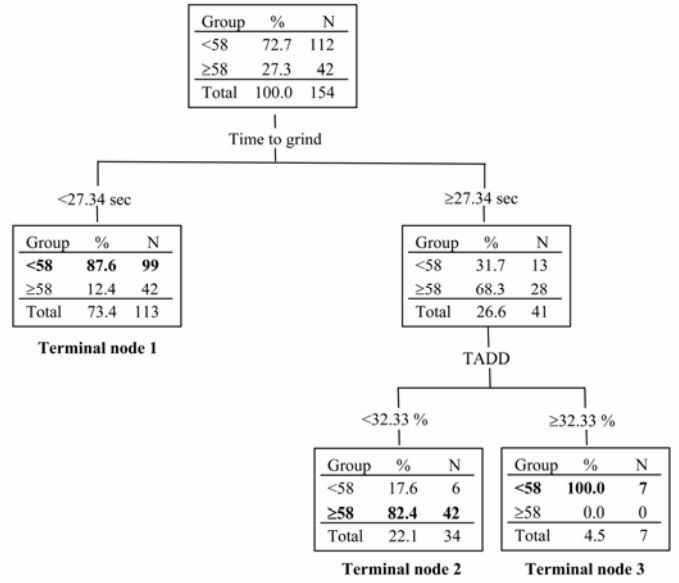

Fig. 5. Decision tree diagram for classification of 100% maize samples into two dry-milled grit yield groups. Diagram as described in Fig. 4.

TABLE IX
Correct Classification Rates of Two Dry-Milled Grit Yield Groups Estimated Using Decision Tree Algorithm with Time to Grind and TADD Variables

	Predicted Group (all samples)			Predicted Group (training set)			Predicted Group (test set)		
Actual Group	**<58**	**≥58**	**Total (% correct)**	**<58**	**≥58**	**Total (% correct)**	**<58**	**≥58**	**Total (% correct)**
<58	106	6	112 (94.6)	74	3	77 (96.1)	32	3	35 (91.4)
≥58	14	28	42 (66.7)	8	21	29 (72.4)	6	7	13 (53.9)
Total	120	34	154 (87.0)	82	24	106 (89.6)	38	10	48 (81.2)

was identical to the decision tree shown in Fig. 5. In fact, the resulting model was only different in the absence of NIT density criterion in the decision tree built by the training and the test data sets for three grit yield groups. The reduction in dry-milled grit yield groups increased the overall correct classification rates (Table IX). With models developed using all the samples or a training data set, 94.6 and 96.1% of the <58% yield group samples were correctly classified, resulting in 87.0 and 89.6% overall correct classification rates, respectively.

Total correct classification rates were similar between discriminant analysis and decision tree models built using all our samples. The decision tree model, however, showed a slightly higher correct classification rate in the training data set, while the discriminant analysis model demonstrated slightly better accuracy in the test data set. Despite lack of sufficient sample numbers in certain dry-milled grit yield groups, the decision tree appeared to be successful in classifying and predicting dry-milled grit yield groups from both an accuracy and a utility view point.

CONCLUSIONS

Low and moderate correlations between maize physical properties and total grit yields in a regression analysis suggested a need to explore alternative rules and procedures that would enable breeders, producers, and processors to more confidently predict maize dry-milling quality. For maize hybrid samples with genetic and environmental diversity measured using only a few physical properties, the suggested classification and predictive models from discriminant analysis and decision tree algorithm were relatively successful in identifying and predicting predetermined dry-milled grit yield groups. These findings suggest a need to further study application of such classification techniques for dry-milling characteristics and possibly other maize production processes, including wet-milling and alkaline processing. While decision trees are often difficult to explain from an academic perspective, their computer-based development is not difficult. In particular, a decision tree algorithm formulated useful classification rules, although classification accuracy was not high enough to use on a commercial scale. From the present study, the decision tree algorithm can be regarded as a superior method compared with discriminant analysis in terms of improved accuracy, the development of clearer classification rules, the use of fewer relevant variables, and a more straightforward interpretation of the result while relying solely on moderate statistical assumptions. Either discriminant analyses or a decision tree algorithm would be sufficient to complement other classification or segregation methods and to screen maize hybrids suitable for dry-milling. To improve correct classification rates and provide more specific classification rules, larger sample sizes and further investigation to identify or develop better measurement techniques/variables are recommended.

ACKNOWLEDGMENTS

We thank the Andersons Endowment administered through the Ohio Agricultural Research and Development Center of The Ohio State University, AgraMarket Quality Grains, Frito Lay Inc., and seed companies for providing maize samples and financial support in this study. We gratefully acknowledge Darin Joos, Phill Devillex, Kraig Roozeboom, and Weston Johnson for their assistance.

LITERATURE CITED

Baker S., Herrman, T. J., and Loughin, T. 1999. Use of regression and discriminant analyses to develop a quality classification system for hard red winter wheat. Cereal Chem. 76:890-893.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984. Classification and Regression Trees. Wadsworth International Group: Belmont, CA.

Dombrink-Kurtzman, M. A., and Bietz, J. A. 1993. Zein composition in hard and soft endosperm of maize. Cereal Chem. 70:105-108.

Dorsey-Redding, C., Hurburgh, C. H., Johnson, L. A., and Fox, S. R. 1990. Adjustment of maize quality data for moisture content. Cereal Chem. 67:292-295.

Dorsey-Redding, C., Hurburgh, C. H., Johnson, L. A., and Fox, S. R. 1991. Relationship among maize quality factors. Cereal Chem. 68:602-605.

Duensing, W. J., Roskens, A. B., and Alexander, R. J. 2003. Corn dry milling: Process, products, and applications. Pages 409-447 in: Corn Chemistry and Technology. P. J. White and L. A. Johnson, eds. AACC International: St. Paul, MN.

Johnson, D. E. 1998. Discriminant analysis. Pages 217-285 in: Applied Multivariate Methods for Data Analysts. D. E. Johnson, ed. Duxbury Press: Pacific Grove, CA.

Kirleis, A. W., and Stroshine, R. L. 1990. Effects of hardness and drying air temperature on breakage susceptibility and dry-milling characteristics of yellow dent corn. Cereal Chem. 67:523-528.

Lawton, J. W., and Faubion, J. M. 1989. Measuring kernel hardness using Tangential Abrasive Dehulling Device. Cereal Chem. 66:519-524.

Lee, K. M., Herrman, T. J., Lingenfelser, J., and Jackson, D. S. 2005. Classification and prediction of maize hardness-associated properties by using multivariate statistical analyses. J. Cereal Sci. 41:85-93.

Nelson, S. O. 1980. Moisture-dependent kernel- and bulk-density relationships for wheat and corn. Trans. ASAE 23:139-143.

Pan, Z. 1996. Physical properties and dry-milling characteristics of six selected high-oil maize hybrids. Cereal Chem. 73:517-520.

Paulsen, M. R., and Hill, L. D. 1985. Corn quality factors affecting dry milling performance. J. Agric. Eng. Res. 31:255-263.

Paulsen, M. R., Watson, S. A., and Singh, M. 2003. Measurement and maintenance of corn quality. Pages 159-219 in: Corn Chemistry and Technology. P. J. White and L. A. Johnson, eds. AACC International: St. Paul, MN.

Peplinski, A. J., Paulsen, M. R., and Bouzaher, A. 1992. Physical, chemical, and dry-milling properties of corn of varying density and breakage susceptibility. Cereal Chem. 69:397-400.

Pomeranz, Y., Martin, C. R., Traylor, D. D., and Lai, F. S. 1984. Corn hardness determination. Cereal Chem. 61:147-150.

Pomeranz, Y., Czuchajowska, Z., Martin, C. R., and Lai, F. S. 1985. Determination of corn hardness by the Stenvert Hardness Tester. Cereal Chem. 62:108-112.

Pomeranz, Y., Hall, G. E., Czuchajowska, Z., and Lai, F. S. 1986. Test weight, hardness, and breakage susceptibility of yellow dent corn hybrids. Cereal Chem. 63:349-351.

Rao, C. R. 1973. Linear statistical inference and its applications. John Wiley and Sons: New York.

Reddy, P. W. 1996. Corn dry milling using roller mills: Maximizing low fat grit production. PhD dissertation. Kansas State University: Manhattan, KS.

Robutti, J. L. 1995. Maize kernel hardness estimation in breeding by near-infrared transmission analysis. Cereal Chem. 72:632-636.

Shandera, D. L., Jackson, D. S., and Johnson, B. E. 1997. Quality factors impacting processing of maize dent hybrids. Maydica 42:281-289.

USDA. 1990. General information. Pages 15-17 in: Grain Grading Procedures. USDA Federal Grain Inspection Service: Washington, DC.

Watson, S. A. 2003. Description, structure, and composition of the corn kernel. Pages 69-106 in: Corn Chemistry and Technology. P. J. White and L. A. Johnson, eds. AACC International: St. Paul, MN.

Wehling, R. L., Jackson, D. S., and Hamaker, B.R. 1996. Prediction of corn dry-milling quality by near-infrared spectroscopy. Cereal Chem. 73:543-546.

Witten, I. H., and Frank, E. 2000. What's it all about? Pages 1-36 in: Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations. I. H. Witten and E. Frank, eds. Morgan Kaufmann Publishers: San Francisco, CA.

Wu, Y. V., and Bergquist, R. R. 1991. Relation of corn grain density to yields of dry-milling products. Cereal Chem. 68:542-544.

[Received April 24, 2006. Accepted November 29, 2006.]