Discriminant Analysis of Vegetable Oils by Near-Infrared Reflectance Spectroscopy

Karen M. Bewig[1], **Andrew D. Clarke**[a,*], **Craig Roberts**[b] and **Nan Unklesbay**[b]
[a]Department of Food Science and Human Nutrition and [b]Department of Agronomy, University of Missouri, Columbia, Missouri, 65211

Discriminant analysis of four vegetable oil types (cottonseed, peanut, soybean and canola) was performed by near-infrared reflectance spectroscopy. The objective of this study was to provide an alternate method to differentiate vegetable oil types and to classify unknown oil samples. Second derivative spectra of the vegetable oils were subjected to discriminate analysis with Mahalanobis distances principles. A four-wavelength (1704, 1802, 1816 and 2110 nm) equation was derived, which produced a sum of inverse squared distance of 0.0548. Although all four groups were successfully separated with a chi square of 18.9, the soybean oil group is more dispersed in space than the other three groups. Iodine values of the soybean oil samples suggest that this group may have a wide range of hydrogenation states. Discriminant analysis can be successfully used to differentiate vegetable oil types and possibly could also be used to differentiate degree of hydrogenation and oxidative states of oils.

KEY WORDS: Discriminant analysis, NIR, near-infrared spectroscopy, vegetable oil.

Near-infrared reflectance (NIR) is used worldwide for the rapid quantitative determination of moisture, lipid, protein, carbohydrates and fiber in cereals, grains, feeds, meats and dairy products (1). However, it has rarely been used for quantitative or qualitative oil analysis. Using lipid components and hydrogenated oils, Holman *et al.* (2) measured absorbance at 2150 nm and related this to the iodine value of fat. Wetzel (3) investigated the relationship between oil structure, *i.e.*, solid-fat index, degree of unsaturation and carbon number in triglycerides with NIR responses. He correlated carbon number to three major response bands at 1680, 2139 and 2208 nm and established a relationship between lipid structure and NIR spectroscopy. Sato *et al.* (4) developed a foundation for the rapid determination of fatty acid composition in fats and oils by NIR spectroscopy. This was successfully accomplished with pure-triglyceride spectra to reconstruct spectra of several fats and oils and then comparing the calculated spectra to the original spectra.

If the NIR spectra of substances are measured with sufficient accuracy, then any wavelength where absorbance differences exist can serve to classify them (5). The key to achieving qualitative analysis by NIR spectroscopy is the application of the multivariate algorithms, a process often referred to as supervised learning (6). The supervised training method used in NIR spectroscopy is discriminant analysis (7). Discriminant analysis and NIR are being used in the pharmaceutical industry to verify materials packed in appropriate vials and containers (8). A qualitative method, such as discriminant analysis, would be a useful tool in quality control at an oil refinery or end-user production facility. This study was conducted to further develop a qualitative method of analysis to classify oils.

[1]Present address: Continental Baking Co., Checkerboard Square, 2CR, St. Louis, MO 63164.
*To whom correspondence should be addressed at 21 Agriculture Bldg., University of Missouri, Columbia, MO 65211.

The current methods used to classify an oil involve separating and identifying fatty acids or triglycerides by either gas chromatography or high-pressure liquid chromatography. Both methods are costly, time-consuming and potentially hazardous. The iodine value test, which estimates degree of unsaturation, can result in a probable oil classification but is less accurate than chromatographic procedures.

The objective of this study was to determine if NIR could be used to discriminate vegetable oils and to classify unknown vegetable oil samples. If successful, this technique could provide a rapid low-cost nondestructive qualitative identification method for the food industry.

EXPERIMENTAL PROCEDURES

Oil samples. Four varieties of vegetable oil commonly used in the food industry were used in the discriminant analysis equation: cottonseed, peanut, soybean and canola. Samples were randomly selected lots of virgin oil from various oil suppliers: Plains Cooperative Oil Mill (Lubbock, TX); Riceland Foods (Stuttgart, AR); Oilseeds International (Fresno, CA); C & T Refinery (Charlotte, NC); Archer Daniels Midland (Decatur, IL); and Cargill (Gainesville, GA). Three additional oil types, corn, sunflower and olive, were purchased from a local retail market for equation validation purposes. Corn oil was produced by Wesson, Inc. (Fullerton, CA) sunflower oil by Schnuck's Markets (Bridgeton, MO) and olive oil by Pompeian (Baltimore, MD). All samples were stored in the dark at 25°C and analyzed by NIR spectroscopy within 30 d.

Chemical procedures. Iodine values were determined on each oil within 90 d by the American Oil Chemists' Society Official Method Cd 1-25 (9). Fatty acid composition was determined on all soybean oil samples by the gas chromatographic methods 969.33 and 963.22 approved by AOAC (10). Fatty acid analyses were performed by personnel at the University of Missouri Agricultural Experiment Station Chemical Laboratory (Columbia, MO).

NIR measurements. Forty-six oil samples were analyzed in triplicate by NIR. Log (1/reflectance) spectra were recorded at 4-nm intervals from 1100 to 2500 nm with an InfraAlyzer 500 (Bran + Luebbe, Buffalo Grove, IL) NIR analyzer. A stainless-steel cup with a 1-mm deep chamber, suitable for viscous fluids (PN 189-B419-01; Bran + Luebbe), and a quartz cover glass was used for sample presentation. All spectra were recorded at 25°C. An IBM PC equipped with IDAS software (Bran + Luebbe) was used to collect, store and manipulate data. Second-derivative transformation of all spectra was performed by using the following criteria: 4 nm between output points, 4 nm in moving average, 12 nm per derivative segment and 12 nm between derivative segments. Second-derivative manipulation creates a baseline at zero and resolves many overlapping peaks not easily seen in the original spectra. Derivative transformation created a new wavelength search range of 1116 to 2474 nm. The data file was randomly separated into two files: (i) A data file for the discriminant analysis equation consisting of nine

Copyright © 1994 by AOCS Press

cottonseed, eight peanut, eight soybean and five canola oil samples and (ii) a validation data file consisting of three cottonseed, four peanut, four soybean and two canola oil samples. A third file was created with the sunflower, corn and olive oil samples. The discriminant data file was subjected to a wavelength search by means of the IDAS discriminant analysis program. Three-, four- and five-wavelength searches were tried.

Mahalanobis distances. The multidimensional distance measures were computed with the matrix Equation 1 (11):

$$D_{ij}^2 = (X_{ij} - \overline{X}_j)/M(X_{ij} - \overline{X}_j) \quad [1]$$

where D_{ij} is the Mahalanobis distance from the *i*th sample to the location of the *j*th material in the multidimensional space, X_{ij} is the vector of absorbance data from the unknown sample, M is the inverse pooled covariance matrix from all the materials in the training set, and X_j is the location in multidimensional space of the *j*th material. The unit distances measured are called Mahalanobis distances (11).

Sum of inverse squared distances (SISD). This statistic was calculated with Equation 2 (11):

$$\Sigma \left(\frac{1}{D_{ij}}\right)^2 \quad [2]$$

where D_{ij} is the Mahalanobis distance between all pairs of group *i* and *j*.

Normalized distances. Mahalanobis distances are normalized by the root mean squared (RMS) group size (Equation 3):

$$RMS_j = [\Sigma D_{ij}^2/(n_j - 1)]^{1/2} \quad [3]$$

where RMS_j is the size of the data from the *j*th material, D_{ij} is as defined in Equation 2, and n_j is the number of training samples of the *j*th material (11).

Wavelength selection. The optimum set of wavelengths was selected by computing the distances D_{ij} between all pairs of groups *i* and *j*, then forming the SISD. The groups that were closest together contributed most heavily to this sum. This method selected those wavelengths that resulted in the smallest sum for separation of the closest groups (5).

RESULTS AND DISCUSSION

Oil spectra. Most differences in NIR spectra patterns of oils are evident from 1600 to 1800 nm and from 2100 to 2200 nm (4). For this reason, only the region between 1580 and 2220 nm will be presented in the spectral figures. Figure 1, which presents a representative spectrum from each oil sample group, reveals visible differences around 2110 nm. Goddu (12) reported that the region near 2100 nm could be used for determination of terminal saturation. The region around 1700 nm shows absorbance due to the first overtone of the C-H stretching vibration (13), but little difference was observed for these spectra. Because the discriminate equation was derived from second-derivative spectra, the remaining discussion will address the transformed spectra. The specific regions around 1704, 1802, 1816 and 2110 nm, the wavelengths suggested by the discriminant analysis search, are shown in Figures 2–4. Figure 2 presents the portion of spectra from 1680 to 1728 nm for all oil types and shows the difference in absorbance at wavelength 1704 nm, the first wavelength chosen in the discriminant equation. It is obvious that peanut and soybean oil have different absorbances than cottonseed and canola. This may be related to oleic acid content, a constituent that is much higher in peanut oil as compared to cottonseed oil (14).

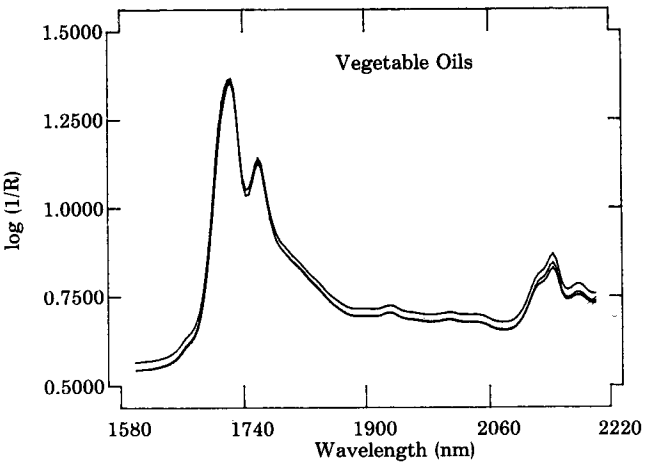

FIG.1. Log (1/R) spectra of all four vegetable oil types. Abbreviation: R = reflectance.

Figure 3 presents the portion of spectra from 1788 to 1844 nm and includes the next two wavelengths in the equation, 1802 and 1816 nm. Canola oil seems to be separated from the other oils at both wavelengths, and canola is the most unsaturated of these four oils (14). Findings of Sato *et al.* (4) report that the region near 1800 nm contains information about the degree of saturation of fatty acid moieties. Canola oils also have higher erucic and linolenic acids than other oils (14).

Figure 4 presents the spectral region from 2100 to 2128 nm and includes the fourth wavelength, 2110 nm, used in the equation. At this wavelength, soybean oil and cottonseed oil have different absorbances than peanut and canola oil. Total percentage of polyunsaturation can be slightly higher in soybean and cottonseed due to higher linoleic acid content (14).

Discriminant analysis equation. The equation that gave the best separation of the four oil groups was a wavelength selection of 1704, 1802, 1816 and 2110 nm, as determined by the SISD. By summing the inverse of the squared distances, the groups that are close together are given more weight than the groups that are far apart. Thus, the SISD statistic reflects mainly the contribution of the closest groups. The farther apart the groups are, the smaller the SISD will be (15).

The Mahalanobis distances between groups are as follows: 17.13 (from cottonseed to peanut); 17.49 (from cottonseed to soybean); 31.13 (from cottonseed to canola); 4.97 (from peanut to soybean); 17.87 (from peanut to canola); and 17.08 (from soybean to canola). Three wavelengths, 1704, 1802 and 2110 nm, produced an SISD of 0.0670. The four-wavelength equation added 1816 nm with a resulting SISD of 0.0548. A five-wavelength search added 2292 nm and produced an SISD of 0.0525. Adding one more wavelength only decreased the SISD by 0.0023.

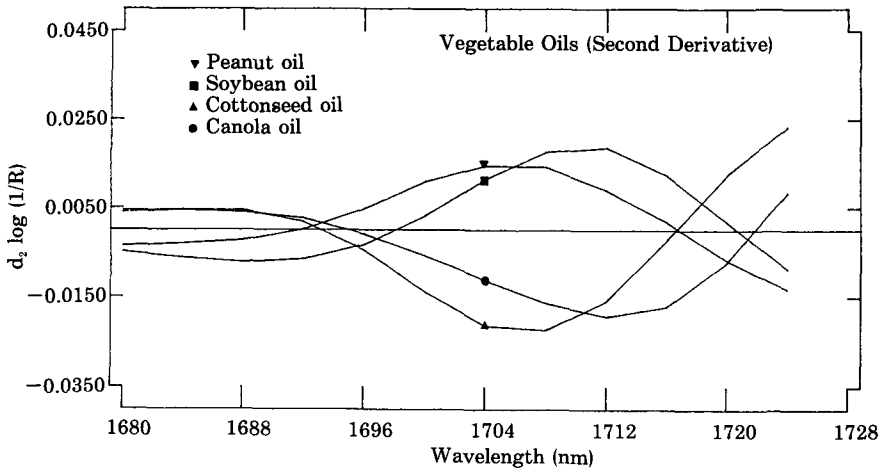

FIG. 2. Wavelength region from 1680 to 1728 nm for second-derivative spectra of four vegetable oils.

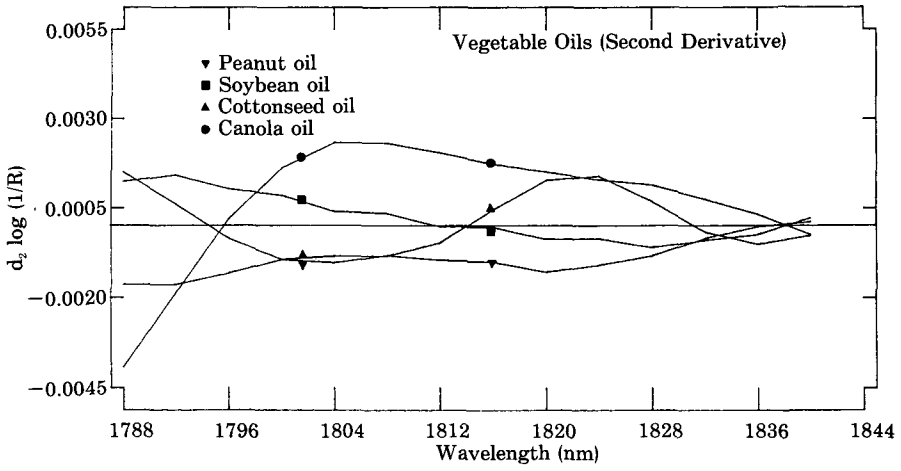

FIG. 3. Wavelength region from 1788 to 1844 nm for second-derivative spectra of four vegetable oils. See Figure 1 for abbreviation.

This suggested that four wavelengths were sufficient to discriminate among oil groups. The Mahalanobis distances revealed adequate separation between all groups except the peanut and soybean oil groups. Upon further inspection of sample locations on a Log-Log plot (Fig. 5), it was evident that one soybean oil sample (S8), had absorbed in the region of the peanut oil group. This explains the low Mahalanobis distance of 4.97 between the two groups. Group pairs that are closer than six times the Mahalanobis distance may likely overlap and can lead to misclassification of unknown members of either group (15). If this sample were deleted, the Mahalanobis distances between the peanut and soybean groups would increase above 6.00, and the SISD would decrease to 0.0150. However, this sample was retained in the data file to represent the random selection of oils intended in this study. The total inverse squared distance for each wavelength represents the weight or influence of each wavelength to the equation for purposes of wavelength deletion. All wavelengths were almost equally weighted. Wavelength 2110 nm contributed the most with a total inverse squared distance of 0.0943, and 1816 nm contributed the least with a total inverse squared distance of 0.0630. Deleting wavelength 1816 nm from the equation increased the SISD instead of decreasing it, which indicated that this wavelength was contributing significantly to the equation.

The RMS distances for each oil group are 1.87 (cottonseed oil); 1.36 (peanut oil); 2.86 (soybean oil); and 1.24 (canola oil). The RMS indicates the size of the group. Figure 5 also shows that the soybean oil group is more dispersed in space than the other three groups. The chi square (X^2) value was calculated as 18.9. The critical value of X^2 for 3 df at a confidence level of 0.995 is 12.84. So, the differences between the RMS group sizes did not result from random variation.

The discriminant analysis equation was verified by a validation set containing the same oil types but not actual samples used in the equation set. A second validation set was also used, which contained oil types that are different than those used in the equation set ("other" oil types).

First validation. The first validation was tested with the

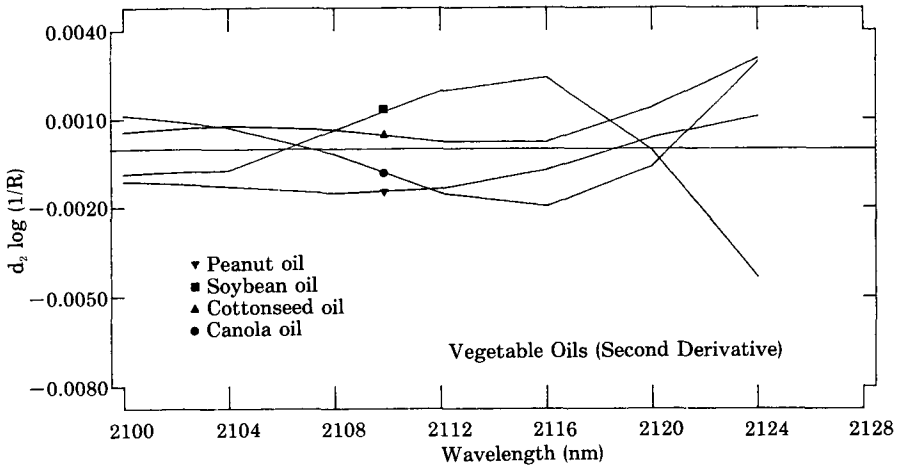

FIG. 4. Wavelength region from 2100 to 2128 nm for second-derivative spectra of four vegetable oils. See Figure 1 for abbreviation.

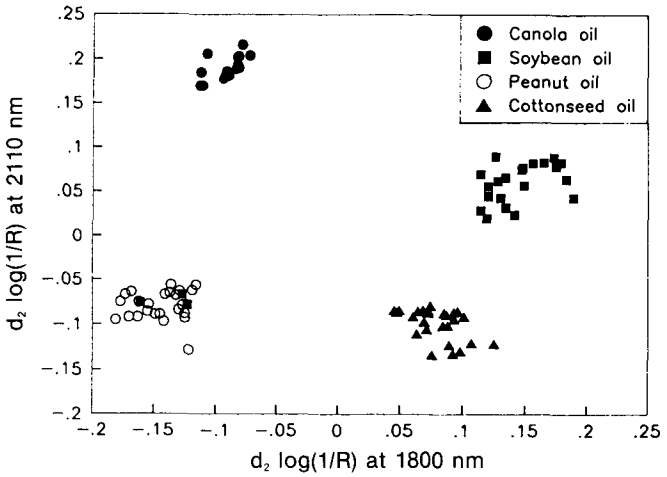

FIG. 5. Second-derivative spectral data points of all oil samples plotted at wavelengths 2110 nm *vs.* 1800 nm. Note that one soybean oil sample (in triplicate) is located in the peanut oil grouping. See Figure 1 for abbreviation.

four-wavelength equation. Table 1 lists the results. Any sample with a Mahalanobis distance of three times or less was classified in the corresponding group. If Mahalanobis distances were greater than 3.0, the sample was classified as a member of the group having the lowest Mahalanobis distance. Two samples (S3 and S6) were more than three times the Mahalanobis distance from any reference group but were correctly classified as soybean oil because they were closest to the soybean oil reference group. Validation was considered successful except that two samples were misclassified. Sample S5 was classified as canola oil, even though it was soybean oil. Sample P4 was classified as soybean oil when it was actually peanut oil. The P4 sample, however, appeared close to the peanut group with a Mahalanobis distance of 5.86. Soybean oils were involved in both incorrectly classified samples, supporting the earlier observation that the soybean oil reference group was dispersed and within six times the Mahalanobis distance from the peanut oil reference group. All the cottonseed and canola oils were classified correctly.

Mahalanobis distances assume that the region of space occupied by one material has the same size, shape and

TABLE 1

Mahalanobis Distances and Classifications for Validation Samples

Sample number	Actual oil type	Mahalanobis distances[a]				Classified as:
		C	P	S	CL	
C4	Cottonseed	1.49	16.54	17.07	31.08	Cottonseed
S3	Soybean	14.58	15.16	13.99	19.76	Soybean[b]
S5	Soybean	32.06	18.39	17.74	1.51	Canola[c]
CL3	Canola	29.77	16.68	15.91	2.00	Canola
S6	Soybean	24.45	12.99	12.14	25.24	Soybean[b]
P4	Peanut	18.75	5.86	2.60	17.96	Soybean[c]
P6	Peanut	17.82	2.99	5.07	17.37	Peanut
C8	Cottonseed	0.59	16.99	17.39	31.14	Cottonseed
P8	Peanut	18.43	1.65	4.89	17.27	Peanut
C10	Cottonseed	1.00	17.99	18.35	31.82	Cottonseed
S7	Soybean	19.08	5.45	2.04	16.72	Soybean
CL6	Canola	30.58	17.40	16.53	1.61	Canola
P10	Peanut	16.10	1.56	5.21	18.43	Peanut

[a]Mean values of triplicate near-infrared reflectance readings. Abbreviations: C = cottonseed oil group; P = peanut oil group; S = soybean oil group; CL = canola oil group.
[b]More than three times Mahalanobis distance, but correctly classified.
[c]Classified incorrectly.

TABLE 2

Normalized Distances and Classification for Validation Samples

Sample number	Actual oil type	Normalized distances[a]				Classified as:
		C	P	S	CL	
C4	Cottonseed	0.79	12.16	5.96	24.99	Cottonseed
S3	Soybean	7.79	11.14	4.88	15.89	Soybean[b]
S5	Soybean	17.14	13.52	6.19	1.21	Canola[c]
CL3	Canola	15.91	12.26	5.56	1.61	Canola
S6	Soybean	13.24	10.45	4.66	20.29	Soybean[b]
P4	Peanut	10.02	4.31	0.92	14.44	Soybean[c]
P6	Peanut	9.53	1.00	1.77	13.97	Peanut[d]
C8	Cottonseed	0.31	12.49	6.07	25.04	Cottonseed
P8	Peanut	9.85	1.21	1.71	13.97	Peanut[d]
C10	Cottonseed	0.53	13.22	6.41	25.59	Cottonseed
S7	Soybean	10.20	4.00	0.71	13.44	Soybean
CL6	Canola	16.34	12.79	5.77	1.29	Canola
P10	Peanut	8.61	1.15	1.82	14.82	Peanut[d]

[a]Mean values of triplicate near-infrared reflectance readings. Abbreviations: See Table 1.
[b]More than three times normalized distance but correctly classified.
[c]Classified incorrectly.
[d]Less than three times normalized distance from two reference groups but classified correctly.

orientation as that occupied by every other material (16). As previously discussed, the group of data points representing soybean oil was larger than the space occupied by the other three oil groups. The inverse pooled covariance matrix adequately described the shape and orientation of each group, but not the size (16). If some groups were more dispersed than others, then the use of normalized distances could minimize the three classification errors: False failure to classify, incorrect classification and falsely classifying a sample when the classification should have been ambiguous (even if the sample was correctly classified) and thus results in a more accurate analytical technique (16).

Table 2 presents the same validation set as normalized distances. Overall, the normalized distances are shorter (lower) than Mahalanobis distances. The same samples (S3 and S6) were more than three times the normalized distances but were classified correctly. Samples S5 and P4 were misclassified. In addition, three samples (P6, P8 and P10) were within three times the normalized distances of two reference groups (peanut and soybean), but were correctly classified as peanut oil. Again, all cottonseed and canola oil were classified correctly.

Iodine values. Iodine values were obtained for all oil samples. Cottonseed oil had an iodine value range of 102.6–119.3, which compared to the reference typical cottonseed frying oil range of 103–113 (17), representing the entire range sufficiently. Peanut oil had an iodine value range of 93.9–103.9, which compared to the reference typical peanut frying oil range of 84–102 (17), representing the middle to upper range. Canola oil ranged from 108.6–120.8. The calculated mean reference iodine value for canola oil is 120.0 (18). These canola oils represented the lower to middle range of typical canola oils. Soybean oil ranged from 97.0–136.7, the widest iodine value range of all four oils; it had a standard deviation of 12.0, the highest of all oil groups.

The wide range of iodine values would suggest large variation in degree of saturation or hydrogenation within this group. This could account for the wider dispersion of data points in the soybean group as compared to the other three groups and may account for the soybean oil group overlapping with the peanut oil group. This would also support the validation set results, which showed some peanut and soybean oil samples involved in misclassifications. Two reference iodine value ranges were found for soybean oil, one for soybean oil of 125–138 and another for processed soybean oil of 86–103 (17). Swern (14) reports an iodine value range for soybean oil of 120–141. It seems that the soybean oil group used in this study could possibly contain more than one population of soybean oil according to the degree of hydrogenation.

The sunflower, corn and olive oils were representative of typical vegetable oils. Calculated reference mean values were as follows: sunflower = 136.0, corn = 128.0 and olive = 82.0 (18).

Fatty acid profiles of soybean oil samples. Fatty acid profiles were performed on all soybean oil samples because one sample (S5) was misclassified as being closest to the canola oil group based on both the Mahalanobis and normalized distances. The samples had undergone autooxidation during the storage period before the profiles were performed. This was evident by the broader peaks and lower than expected methyl ester results as compared to reference standards. Even though the actual values of the methyl ester compositions were affected by the autooxidation, all soybean oil samples had similar values for C16:0, C18:0, C18:1, C18:2 and C20:1, except one. Not surprisingly, sample S5 had different values, especially for C18:0 and C20:1. The methyl ester weight percent for C18:0 was 0.32 for this sample, whereas all others ranged from 1.34 to 2.98. The C20:1 was 3.46 for sample S5, while all others were less than 0.26. Typical soybean oil is reported to have a C18:0 range of 4.0 to 4.6 and a negligible C20:1 content (19). Canola oils have a C18:0 range of 1.2 to 1.8 (19) and a C20:1 range of 0.8 to 2.3 (19,20). Given the NIR identification of S5 as canola oil and the fatty acid profile that suggests a composition more similar to canola than to soybean oil, one may conclude that the oil used for validation was actually a canola oil. However,

given that no other information is available to identify the sample and that the suppliers were credible, the alternative conclusion was that the S5 sample was a soybean oil that reflects an extreme from the population. If so, the NIR technique does identify unusual samples and can provide an opportunity to further analyze samples prior to acceptance in food processing applications.

Second validation. The three oils, sunflower, corn and olive, that were not used as reference groups in the discriminant equation were subjected to validation. The resulting Mahalanobis distances were all above 3.0 as expected. However, when we used the normalized distance, corn oil was misclassified as soybean oil with a distance of 2.74. The sunflower oil with a normalized distance of 3.20 was also close to the soybean oil group. This could also be due to the larger soybean oil group dispersion.

It has been shown that discriminant analysis can be used to separate the spectra of these four vegetable oils, and the equation can be used to classify unknown oil samples accurately. It also has been shown that discriminant analysis may be able to separate different degrees of hydrogenation, as in the case of soybean oils.

Because of the differences seen in using normalized distances rather than Mahalanobis distances in classifying these vegetable oil samples, more studies are needed to determine the most accurate statistic for the types of samples being addressed.

Discriminant analysis as an NIR qualitative method could have a viable place in the oil refinery and end-user production facilities.

ACKNOWLEDGMENTS

We thank the staff of the Agricultural Experiment Station Chemistry Laboratory at the University of Missouri-Columbia for assistance in performing the fatty acid profiles. Eagle Snacks, Inc., and especially Henry Davis, provided valuable technical support. Contribution from Missouri Agricultural Experiment Station Journal Series Number 11,951.

REFERENCES

1. Dunmire, D., and R. Williams, *Cereal Foods World 35*:913 (1990).
2. Holman *et al.*, cited by Williams, P. and K. Norris (eds). *Near-Infrared Technology in Agricultural and Food Industries*, American Association of Cereal Chemists, Inc., St. Paul, 1987, p. 204.
3. Wetzel, D.L., *Lipid Structures and NIRA,* Presented at the 4th International Symposium NIRA, Kansas State University, Manhattan, 1983.
4. Sato, T., S. Kawano and M. Iwamoto, *J. Am. Oil Chem. Soc. 68*:827 (1991).
5. Mark, H., and D. Tunnell, *Anal. Chem. 57*:1449 (1985).
6. Sharaf, M.A., D.L. Illman and B.R. Kowalski, *Chemometrics*, Wiley, New York, 1986 p. 228.
7. Martens, M., and H. Martens, *Appl. Spectr. 40*:303 (1986).
8. Williams, P., *Chem. in Australia 57*:410 (1990).
9. *Official Methods and Recommended Practices of the American Oil Chemists' Society,* 4th edn., American Oil Chemists' Society, Champaign, 1990, Method Cd 1-25.
10. *Official Methods of Analysis of the Association of Official Analytical Chemists*, 15th edn. Arlington, 1990, pp. 963–965.
11. Mark, H., *Analytica Chimica Acta 223*:75 (1989).
12. Goddu, R.F., *Anal. Chem. 29*:1790 (1957).
13. Holman, R.T., and P.R. Edmonson, *Ibid. 28*:1533 (1956).
14. *Bailey's Industrial Oil and Fat Products*, edited by D. Swern, Vol 4., 4th edn. John Wiley and Sons, New York, 1979 pp. 352–422.
15. Anon., *IDAS-pc Users Reference Manual*, Addendum 3, Bran + Luebbe, Buffalo Grove, 1989, pp. 1.1–3.32.
16. Mark, H., *Anal. Chem. 58*:379 (1986).
17. Vandaveer, R.L., *Chipper/Snacker 42*:35 (1985).
18. Leo, D.A., *J. Am. Oil Chem Soc. 60*:301 (1983).
19. Ackman, R.G., and W.M.N. Ratnayake, *The Role of Fats in Human Nutrition*, 2nd edn., Academic Press, New York, 1989, pp. 454, 455.
20. Eskin, N.A.M, and C. Frenkel, Section A, Proceedings of 13th World Congress International Society Fat Research 1977 pp. 1–9.

[Received June 24, 1993; accepted November 26, 1993]