J Wood Sci (2003) 49:275–280

NOTE

Satoru Tsuchikawa · Kaori Yamato · Kinuyo Inoue

Discriminant analysis of wood-based materials using near-infrared spectroscopy

Received: March 13, 2002 / Accepted: July 19, 2002

Abstract This study deals with the suitable discriminant techniques of wood-based materials by means of near-infrared spectroscopy (NIRS) and several chemometric analyses. The concept of Mahalanobis' generalized distance, K nearest neighbors (KNN), and soft independent modeling of class analogy (SIMCA) were evaluated to determine the best analytical procedure. The difference in the accuracy of classification with the spectrophotometer, the wavelength range as the explanatory variables, and the light-exposure condition of the sample were examined in detail. It was difficult to apply Mahalanobis' generalized distances to the classification of wood-based materials where NIR spectra varied widely within the sample category. The performance of KNN in the NIR region (800–2500 nm), for which the device used in the laboratory was employed, exhibited a high rate of correct answers of validation (>98%) independent of the light-exposure conditions of the sample. When employing the device used in the field, both KNN and SIMCA revealed correct answers of validation (>88%) at wavelengths of 550–1010 nm. These results suggest the applicability of NIRS to a reasonable classification of used wood at the factory and at job sites.

Key words Near-infrared spectroscopy · Wood-based material · Mahalanobis' generalized distance · KNN · SIMCA

Introduction

When we utilize wood as architectural or industrial materials, it normally is subjected to several chemical or mechanical processes (e.g., impregnation of antiseptics or fire retardants, application of adhesives, lamination with a polyvinyl chloride film). Therefore, when discarding or recycling such wood-based materials, we must consider the best procedure for distinguishing them clearly into inflammable and incombustible groups in which chemical components should be evaluated accurately and rapidly. Currently, there are few feasible discriminting systems at the factory or at job sites, where the inspectors roughly classify them. Therefore, we need to improve the present situation as quickly as possible from the viewpoint of preserving the environment.

Near-infrared spectroscopy (NIRS) is a nondestructive analytical method for determining the composition of materials.[1,2] Diffuse reflectance or an absorption spectrum of 800–2500 nm allows clear discrimination of various organic compounds. The application of NIRS to such wood-based materials or engineered wood has been reported by some researchers,[3,4] who especially noted the usefulness of quantitative analysis. However, the severe and critical requirement to the wood industry described above may also be satisfied by the use of NIRS.

We[5] have reported that the discrimination of wood species could be performed by means of combining NIRS and Mahalanobis' generalized distance. Its accuracy and reasonability were examined for samples with various moisture contents ranging from oven-dried to the fully saturated free water state. Each wood group was well recognized by the discriminant analysis using second derivative spectra, resulting in a high percentage of correct answers, with good validation. Brunner et al.[6] have also demonstrated the usefulness of Fourier transform (FT)-NIR for classifying wood species. They focused on the principal component analysis, where the original NIR spectra of various sawn cut, or microtomed samples were employed. These previous studies pointed out the significance of the NIR range, which contributed to the discrimination of wood samples with statistical satisfaction, although at first glance the spectra between samples were similar.

In this study, the use of NIRS to classify wood-based materials was examined in regard to social and industrial backgrounds. As a first step of this project, we tried an approximate classification to categorize diverse wood products, so each category included several types of wood-based

S. Tsuchikawa (✉) · K. Yamato · K. Inoue
Graduate School of Bioagricultural Sciences, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
Tel. +81-52-789-4158; Fax +81-52-789-4155
e-mail: st3842@agr.nagoya-u.ac.jp

material. For example, we did not especially distinguish between plywood and laminated veneer lumber (LVL) but classified solid wood and such laminated wood in one group. Several chemometric techniques – Mahalanobis' generalized distance,[7] K nearest neighbors (KNN),[8] soft independent modeling of class analogy (SIMCA)[9,10] – were examined for their accuracy and reasonability. The NIR spectra of wood-based materials were measured by a typical spectrophotometer used in the laboratory and one used in the field. The difference in the accuracy of classification with the spectrophotometer, the wavelength range as the explanatory variable, and the light-exposure condition of the sample were examined in detail.

Classification methods

When we have qualitative or quantitative information based on spectra, we can focus on multivariate spectra to classify or evaluate substances. As several overtone or combined-tone bands of organic compounds overlap in the NIR region, it is important to find the key information by applying an effective chemometric technique(s). The classification modeling by KNN and SIMCA are briefly described here. Because we introduced the concept of Mahalanobis' generalized distance in the last report,[5] its description is omitted here.

K nearest neighbors

The KNN procedure attempts to categorize an unknown sample based on its proximity to samples already categorized, similar to the Mahalanobis' generalized distance technique. Specifically, the predicted class of an unknown depends on the class of its k nearest neighbors, which accounts for the name of the technique. In a fashion analogous to polling, each of the k closest training set samples votes once for its class; the unknown sample is then assigned to the class with the most votes. An important part of the process is to determine an appropriate value for k (the number of neighbors voting).

The general expression for the Euclidean distance d_{ab} between the known and unknown sample is

$$d_{ab} = \left[\sum_{j=1}^{m} \left(\boldsymbol{a}_j - \boldsymbol{b}_j \right)^2 \right]^{1/2} \quad (1)$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the data vectors for the known and unknown sample, respectively. A data vector contains m explanatory variables as the absorbances of a restricted wavelength range.

If d_{ab} has a low value, it means there is high similarity between the two data. This way, d_{ab} between the unknown sample and several known samples, which have already been classified, are calculated. Finally, the cluster including the unknown sample is defined. We call a concrete procedure 1NN, 3NN, 5NN, and so on. In the case of 1NN, the cluster that includes the known sample nearest to the unknown sample is selected. In the case of 3NN, the three known samples nearest to the unknown sample are examined successively. If the cluster of two or three samples coincides with each other, the unknown sample may be included in it. When the similarity of the cluster does not satisfy such conditions, we cannot determine that the cluster includes the unknown sample.

Soft independent modeling of class analogy

The SIMCA method was first introduced by Wold.[9] In contrast to KNN, which is based on distances between pairs of samples, SIMCA develops principal component models for each training category. An attractive feature of SIMCA is its realistic prediction options compared to KNN.

Also with SIMCA, all known samples are classified into several clusters. Principal component analysis (PCA) is performed for each cluster, for which a restricted pth dimensional space is constructed. The area of cluster including the known s samples containing n measurement values (i.e., the absorbances of the restricted wavelength range) is defined as RSD and is calculated as follows.

$$RSD = \left[\sum_{i=1}^{s} \sum_{k=1}^{n} \frac{e_{ik}}{(s-p-1)(n-p)} \right]^{1/2} \quad (2)$$

where e_{ik} is the residual of ith known sample. After performing PCA for each cluster, a so-called SIMCA-box is presented. The classification is based on the distance D between the unknown sample and each SIMCA-box. D is compared with RSD. If D is much smaller than RSD, the unknown sample may be assigned to the cluster.

Materials and methods

Samples

Wood-based materials of various types are commonly used under diverse conditions; however, it is not advisable to examine the detailed classification at the first step of this project. We should examine comprehensively the reasonability of using NIRS to classify wood-based materials. Therefore, the samples were classified in five typical categories, as described in Table 1: solid wood, laminated wood, particle- or fiberboard, impregnated wood, and overlaid wood. These categories were approved by the Wood Technological Association of Japan. The dimensions of the samples were 50 × 30 × 10 mm (sample surface 50 × 30 mm). Each sample was measured in the air-dried condition. In this study, we also controlled the light-exposure condition of the sample, which might be regarded as a simulation of used wood. The discriminant analysis was performed on two sample groups as follows.

1. The samples did not suffer forced exposure. The sample volume for each category consisted of 16 specimens: 12

Table 1. Tested species

Solid wood
Japanese cedar (*Cryptomeria japonica*)
Japanese cypress (*Chamaecyparis obtusa*)
Douglas fir (*Pseudotsuga menziesii*)
Western hemlock (*Thuja heterophylla*)
Japanese zelkova (*Zelkova serrata*)
Oak (*Quercus crispula*)
Japanese ash (*Fraxinus mandshurica*)
Japanese beech (*Fagus crenata*)
Laminated wood
Plywood
Laminated veneer lumber
Particleboard or fiberboard
Particleboard
Medium-density fiberboard
Insulation fiberboard
Impregnated wood
Hard fiberboard
Preservative-treated wood
Overlaid wood
Fire retardant-treated wood
Plastic film-overlaid plywood
Printed paper sheet-overlaid plywood
Plastic film-overlaid particleboard
Wrapping of plastic profiles with thermoplastic foils
Thick fancy veneer-overlaid flooring
Thin fancy veneer-overlaid flooring

each were employed for the dataset, and 4 each were employed for the validation set. There were a total of 80 samples.

2. The samples were exposed to simulated sunlight using a WEL-SUN-D (Suga Test Instruments) for 37.5, 75, and 150 h, respectively. These terms corresponded to the natural outdoor exposure times of 2.5, 5, and 10 months, respectively. Of course, it would be preferable to expose the wood for a much longer time to examine the application of this technique to the used wood. The sample volume for each category was same as in item 1.
3. The members of the dataset and validation set were changed four times in each category to check the accuracy of the validation.

Measuring apparatus

We measured each sample using two types of NIR spectrophotometer. Analyses for laboratory use and for field use were considered.

The InfraAlyzer 500 from Bran & Luebbe Co. was employed as the typical instrument for laboratory use; it was labeled the L-type. It includes a diffraction grating and an integrating sphere for obtaining spectral data. The optical fiber probe was used for direct attachment between the sample and the detector. In this system, NIR spectra with high wavelength resolution (about 0.1–1.0 nm) can be measured continuously using diffraction gratings. However, it takes a significant length of time (30 seconds to several minutes) to obtain a repeatable, stable spectrum. The wavelength of incident light varied from 800 to 2500 nm at a stepwidth of 4 nm.

Model fruit selector K-BA100 (Kubota Co.) was employed as the instrument for field use; it was labeled the F-type. It includes a diffraction grating and a multichannel linear-array detector for obtaining spectral data. This device, operating in the interactance mode, was designed to ascertain the quality of fruits or vegetables growing in the field. The attachment optical fiber, employing the interactance method, is useful for this original purpose; however, it was not available for wood samples because of the extreme light propagation along the longitudinal direction of wood fiber. Consequently, the measurement was performed by keeping a distance of 10 mm between the sample surface and the fiber probe. The diffusely reflected light can be detected under this condition. The measurement time for one spectrum takes only 5 s, but the linear image sensor restricted the measurable range at short wavelengths ranging from 550 to 1010 nm.

Outline of experiment

The procedures for discriminant analysis are as follows. The NIR spectra for provided samples were measured by L-type and F-type spectrophotometers, respectively. They were divided into datasets, constructing each category and the validation set as unknown data. The discriminant analysis on the basis of Mahalanobis' generalized distance, KNN, and SIMCA were then examined. The members of the dataset and validation set were changed four times in each wood sample category.

Mahalanobis' generalized distance was applied to NIR spectra measured by the L-type instrument. The two or three wavelengths for the best separations between five categories of wood-based materials were determined from the overall measurable wavelengths (800–2500 nm).

KNN and SIMCA were also applied to NIR spectra from L-type and F-type instruments. For each device, several wavelength ranges were established taking into consideration the spectroscopic characteristics of the electromagnetic wave. The three ranges A_F, B_F, and C_F for the F-type instrument corresponded to the visible range (550–800 nm), the measurable NIR range for this device (800–1010 nm), and the overall measurable wavelength range (550–1010 nm), respectively. The three ranges A_L, B_L, and C_L were specified for the L-type spectrophotometer. They corresponded to the short wavelength range in NIR that was mainly assigned to the second overtone (800–1400 nm); the long wavelength range in NIR, which was mainly due to the first overtone and the combination band (1400–2500 nm); and overall NIR range (800–2500 nm), respectively. The specifications for each device and the established wavelength ranges are summarized in Table 2.

Results and discussion

Discriminant analysis based on Mahalanobis' generalized distance

Mahalanobis' generalized distances between the five categories were calculated for the dataset, where the best two

Table 2. Specifications for each device and established wavelength range for statistical procedures

Parameter	L-type (device for laboratory use)	F-type (device for field use)
Wavelength range	800–2500 nm	550–1010 nm
Detector	Integrating sphere with PbS detector	Multichannel linear array detector
Measurement time	5 s	30 s to several minutes
Applied chemometric technique	Mahalanobis' generalized distance KNN SIMCA	KNN SIMCA
Selected wavelength for analysis	Mahalanobis' generalized distance: 800–2500 nm (total NIR range) KNN and SIMCA A_L: 800–1400 nm (NIR range due to the second overtone band) B_L:1400–2500 nm (NIR range due to the first overtone and combination band) C_L: 800–2500 nm (total NIR range)	KNN and SIMCA A_F: 550–800 nm (visual range) B_F: 800–1010 nm (NIR range available for F-type) C_F: 550–1010 nm (visual + NIR range)

KNN, K nearest neighbors; SIMCA, soft independent modeling of class analogy; NIR, near-infrared

Table 3. Results of discriminant analysis of wood-based materials based on Mahalanobis' generalized distance

Light-exposure condition of the sample and selected wavelengths (nm)	Correct result
No exposure	
1945,[a] 1985	61%
945, 1945,[a] 1985	75%
Exposure for 37.5, 75, and 150 h	
865, 985[a]	48%
985,[a] 1465,[a] 1985	71%

The wavelengths were selected from second derivative spectra measured by the spectrophotometer for laboratory use (L-type). The NIR range for selecting the wavelength was 800–2500 nm

[a] Wavelength derived from the absorption of wood components[1]

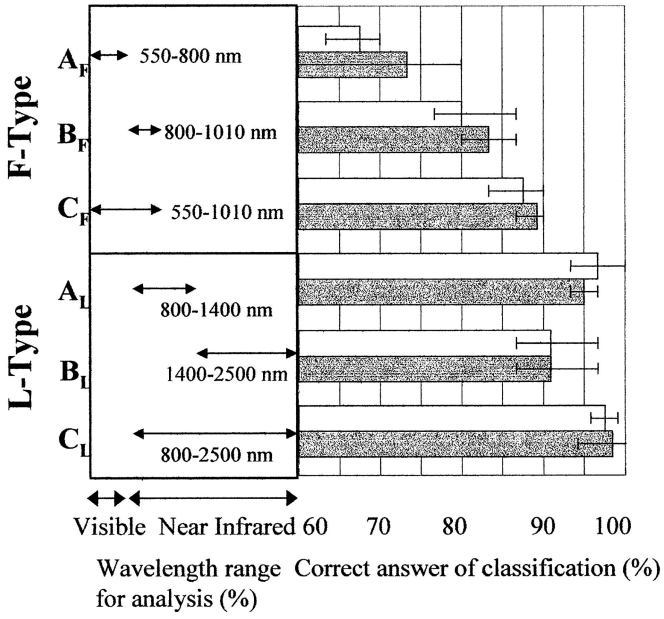

Fig. 1. Results of discriminant analysis of wood-based materials by K nearest neighbors (KNN). *Open bars*, nonexposed sample; *gray bars*, light-exposed sample; *L-Type*, instrument for laboratory use; *F-Type*, instrument for field use

or three wavelengths were chosen. For this procedure, the restricted wavelength in the spectrum is focused to classify or identify the sample by matching the location and strength of absorbance peaks to those of known substances.

Table 3 shows the results of the discriminant analysis based on Mahalanobis' generalized distances. According to our recent report[5] in which we examined the discrimination of wood species using this technique, eight wood species having various moisture contents could be easily classified with the correct answer more than 98% of the time. The variation of spectra with wood-based materials should be larger than that with wood species. Therefore, we first estimated that these categories could be well separated by this classification method. However, the maximum correct answers were limited to 75% for the nonexposure group. Furthermore, the spectroscopic reasonability of the selected wavelength was unelear. Some wavelengths were selected from water absorption bands, whereas the moisture content was not suitable as the explanatory variable in this case. This means that the wavelength selection has no significance.

Mahalanobis' generalized distances cannot be applied to the classification where the spectra vary widely from that of the sample. The selection of wavelengths, which could be explained by the specified position in the restricted *n*th dimensional space, may be difficult. Although we may find a correct answer by increasing the number of selected wavelengths as explanatory variables, it will have little effect or dramatically improve them.

Discriminant analysis based on KNN

Figure 1 shows the results of the discriminant analysis based on KNN. The white and gray bars indicate the correct answers for the nonexposed and light-exposed samples, respectively. For the classification at a wavelength range of 800–2500 nm using the L-type (C_L) instrument, we found a high rate of correct answers (>98%) independent of the light-exposure conditions. On the other hand, the correct classification results evaluated by A_L and B_L were slightly less.

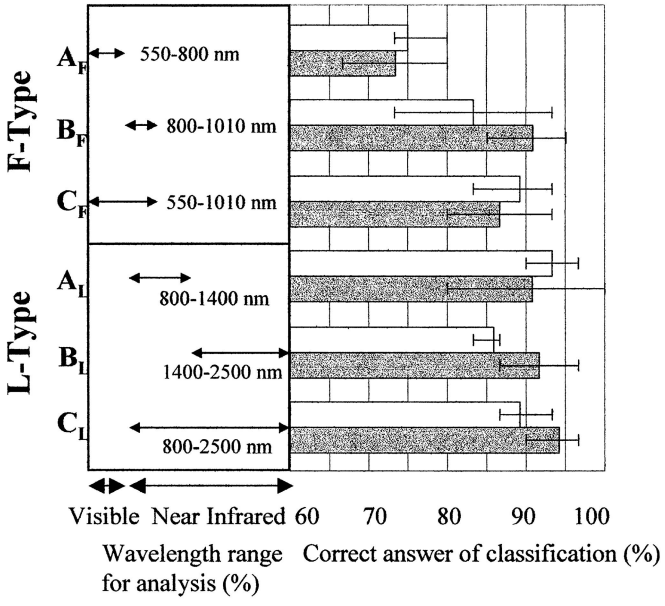

Fig. 2. Results of discriminant analysis of wood-based materials by soft independent modeling of class analogy (SIMCA). *Open bars*, nonexposed sample; *gray bars*, light-exposed sample

In the case of the F-type instrument, we found the correct answer about 88% of the time at the wavelength range of 550–1010 nm (C_F). Whereas the color condition varied with the sample categories, the number of correct classification answers using the visible range (A_F) was low. The correct answer evaluated by B_F is higher than that by A_F in all cases. This suggests that the NIR range includes effective information for the classification of wood-based materials, even though it is a limited NIR range (800–1010 nm).

Discriminant analysis based on SIMCA

Figure 2 shows the results of the discriminant analysis based on SIMCA. The white and gray bars indicate the correct answer for the nonexposed and light-exposed samples, respectively. In the case of L-type, the number of correct answers decreased almost to that of KNN. The correct classification answers obtained by the F-type instrument were slightly more numerous than even obtained with KNN independent of the selected wavelength or the light-exposure state of the sample; however, the variation in results increased.

KNN and SIMCA are based on the assumption that the closer samples lie in the measured space, the more likely it is that they belong to the same category. This idea of proximity implies the concept of distance. KNN and SIMCA are similar techniques that differ in their definition of distance. SIMCA is statistically more realistic than KNN, so we must consider the results in regard to our demands for the classification of wood-based materials.

Reasonability of classification analysis of wood-based materials

We examined several classification procedures for wood-based materials under diverse conditions. The results are summarized to estimate and conclude which method is suitable for our purpose. The analytical method of Mahalanobis' generalized distances may not satisfy us in terms of accuracy. Therefore, we should examine the statistical comparison of KNN and SIMCA for each device.

As shown in Fig. 1, the KNN method using the overall NIR range (C_L) with the L-type instrument gave nearly all correct answers (100%) independent of the light-exposure state of the sample. Needless to say, such a procedure should be recommended as the best analytical classification method. It is obvious that correct classification by A_L occurred at a higher rate than by B_L for both KNN and SIMCA. Therefore, we can conclude, interestingly, that the relatively short NIR range of 800–1400 nm includes more effective information than the longer NIR range of 1400–2500 nm, whereas the sensitivity of the NIR region to chemical features of the materials commonly increases with the increase in wavelength. This result provides useful suggestions for when a new instrument is designed to classify used wood.

On the other hand, for the F-type instrument the classification analysis based on both KNN and SIMCA provided correct answers for validation more than 88% of the time for the overall measurable range (C_F) independent of the light-exposure condition of the sample. The spectroscopic information about visible range plus only a short NIR range (800–1010 nm) may eventually be found suitable for the classification analysis. As described above, comparative merits between KNN and SIMCA depend on our demand for real classification of the wood-based materials. Although the number of correct answers did not reach 90% in this case, we may presently accept this degree of analytical accuracy when considering the limited measurable wavelength. We did not find that light exposure has a significant effect on the classification analysis, perhaps because of the relatively short exposure time for the sample.

In this study we performed a series of measurements and analyses under known favorable experimental conditions for the sample (e.g., sample surface or moisture content). Needless to say, condition in the field where the measurements and analyses must be performed are for more severe and contaminated. Furthermore, the adequate or achieved correct answers must be further considered. In the future we will examine such analyses using real used wood and clarify several problems that must be overcome to achieve a reasonable performance.

Conclusions

We sought to find a suitable discriminant technique for wood-based materials using NIR spectroscopy and several chemometric techniques. The concepts of Mahalanobis'

generalized distance, K nearest neighbors (KNN), and soft independent modeling of class analogy (SIMCA) were evaluated to examine their accuracy and reasonability for this purpose. The differences in accurate classification with the spectrophotometer, the wavelength range as the explanatory variable, and the light-exposure condition of the samples were examined in detail. NIR spectra were measured by a spectrophotometer typically used in laboratory (L-type) and another used in the field (F-type).

Mahalanobis' generalized distances could not be used to classify wood-based materials when the NIR spectra varied widely within the sample category, as the selection of the wavelengths, which could be explained by the specified position in the restricted *n*th dimensional spaces, became difficult. KNN, using the entire NIR region (800–2500 nm) when the L-type instrument was employed, exhibited a large number of correct answers for a validation rate of more than 98% independent of the light-exposure conditions of the sample. This means that the NIR region includes much useful information for classifying wood-based materials. With the F-type spectrophotomer, there were correct answers about 88% of the time in the measurable wavelength range (550–1010 nm). SIMCA, using the L-type instrument, provided fewer correct answers than KNN. In contrast, the F-type spectrophotometer provided a slighter higher rate of correct classifications than did KNN independent of the selected wavelength or the light-exposure state of the sample. Application of the visual range plus only a short NIR range (800–1010 nm) may eventually be suitable for both KNN and SIMCA. Although the number of correct answers estimated by the F-type spectrophotometer did not reach 90%, we may presently accept such analytical accuracy because of the limited measurable wavelength.

Finally, the analytical methods we recommend are KNN for the L-type spectrophotometer and both KNN and SIMCA for the F-type instrument. It is important to determine the comparative merits of the devices and chemometrics techniques because of our need to classify wood-based materials. These results suggest the applicability of NIR spectroscopy to the classification of used wood in factory and job settings.

Acknowledgments The authors thank Gifu Prefectural Human Life Technology Research Institute and Kubota Co. for their support. We also thank Professor Dr. Shiro Kimura and Dr. Hideyuki Yokochi for their constructive discussions about the research.

References

1. Osborne BG, Fearn T (1988) Near infrared spectroscopy in food analysis. Longman, London
2. Burns DA, Ciurczak EW (1992) Handbook of near-infrared analysis. Marcel Dekker, New York
3. Kniest C (1992) Characteristics of urea resined wood particles by NIR spectroscopy. Holz Roh Werkstoff 50:73–78
4. Niemz P, Korner S, Wienhaus O, Flamme W, Balmer M (1992) Applying NIR spectroscopy for evaluation of the hardwood/softwood ration and resin content in chip mixtures. Holz Roh Werkstoff 50:25–28
5. Tsuchikawa S, Inoue K, Noma J, Hayashi K (2003) Application of near-infrared spectroscopy to wood discrimination. J Wood Sci 49:29–35
6. Brunner M, Eugster R, Trenka E, Bergamin-Strotz L (1996) FT-NIR spectroscopy and wood identification. Holzforschung 50:130–134
7. Mark HL, Tunnell D (1985) Qualitative near-infrared reflectance analysis using Mahalanobis distances. Anal Chem 57:1449–1456
8. Kowalski BR, Bender CF (1972) Pattern recognition: a powerful approach to interpreting chemical data. J Am Chem Soc 94:5632–5642
9. Wold D (1976) Pattern recognition by means of disjoint principal components models. Pattern Recognition 8:127–139
10. Sharaf MA, Illman DL, Kowalski BR (1986) Chemometrics. Wiley, New York