Effects of Preprocessing of Ultraviolet-Induced Fluorescence Spectra in Plant Fingerprinting Applications

BERNARD PANNETON,* JEAN-MICHEL ROGER, SERGE GUILLAUME, and LOUIS LONGCHAMPS

Horticultural R&D Centre, Agriculture and Agri-Food Canada, St-Jean-sur-Richelieu, QC, Canada, J3B 3E6 (B.P.); Cemagref, Montpellier, France (J.-M.R., S.G.); and Laval University, Quebec, Canada (L.L.)

Preprocessing is an important step in data analysis. Dealing with spectral data, normalization is mandatory in order to compare items collected under various conditions. This paper addresses normalization of front-face fluorescence spectroscopy data where spectra are affected by an unknown multiplicative effect. The usual methods for reducing multiplicative problems are reviewed and a more detailed analysis of the normalization by closure is provided based on data on the fluorescence of plants as a means for plant species fingerprinting. As normalization is essentially the reduction of information, some methods of carrying it out are likely to remove either meaningful or discriminant pieces of information. As a result, it is demonstrated that normalization by closure should be performed using spectral data in a range where the spectra contain no information relevant to the problem at hand. This applies provided that in this range the signal-to-noise ratio is high enough. When the noise level is too high, a compromise should be found between preserving useful information and limiting the amount of noise introduced by the normalization procedure. Even if this study were carried out using fluorescence spectra, the overall process is likely to be applied to other spectral data.

Index Headings: **Fluorescence spectroscopy; Preprocessing; Discrimination; Spectra normalization; Plant fluorescence.**

INTRODUCTION

Fluorescence spectroscopy is being investigated as a tool for real-time weed–crop discrimination. Under ultraviolet (UV) induction, plant fluorescence spectra can be segmented as blue-green fluorescence (wide peak around 450 nm) and red or chlorophyll fluorescence with two peaks (685 and 735 nm). Both regions are separated by a significant spectral gap (≈600–650 nm). The exact shape of these spectra depends on many factors, some of them being plant-species specific.[1] If this specificity is strong enough with regard to the effect of other factors, then UV-induced fluorescence spectra can be used as a fingerprint for plant species identification. This approach has been investigated for five plant groups (herbaceous monocots, herbaceous dicots, conifers, hardwoods, and algae) using excitation at 337 nm and measuring the fluorescence at 440, 525, 685, and 740 nm. For herbaceous species, monocots can be discriminated from dicots using the ratio of the square of the intensity of the 440 nm peak to the non-squared intensity of the 685 nm peak.[2]

The red fluorescence is emitted by chlorophyll a.[8] The blue and green fluorescence (BF and GF or BGF in short) results from emission of several fluorescing organic substances in plant tissues and is modulated by reabsorption processes due to photosynthetic pigments. Accordingly, BGF is affected by both leaf tissue structure and content in organic substances. Many of these substances can contribute to BGF (e.g., phenolics in the vacuole and cell wall) and the major contribution comes from the epidermal layer.[5] In sugar beet, the epidermis was also shown to be the main contributor of the BGF of the leaves. In the epidermis indirect evidence from the BGF kinetics indicated that ferulic acid may account for about 70% of the total BGF, leaving 30% unexplained. Evidence was also provided for the inner-filter effect (UV screening) of the flavonoids.[8] More direct evidence on the role of ferulic acid as the main emitter for the blue-green fluorescence was provided by measurement of fluorescence excitation spectra of isolated dried cell walls before and after alkaline hydrolysis of ferulic acid.[6] Compared to dicotyledonous plants, it was also shown that members of the *Poaceae* and other monocotyledonous plants have a much higher blue-green to red fluorescence ratio and that cell walls of these species contained a much higher amount of ferulic acid.[6] Looking at a larger selection of plants, it was concluded that ferulic acid was not the only significant emitter of blue-green fluorescence. Other compounds may be involved and their content can vary among species but appears to be similar on adaxial and abaxial leaf surfaces.[4] On top of the species-to-species differences, other factors such as nutrient stress or temperature can contribute to variations in fluorescence spectra.[3,9] These short highlights of some of the results in the field of UV-induced fluorescence of plants clearly show that there is scope for using UV fluorescence spectra as fingerprints for plant species identification.

An ideal instrument would only measure the exact property of interest with perfect accuracy. In the simplest cases, the observed quantity is linearly related to the recorded signal, but in other cases, the relationship is more complex. For example, light scattering creates a nonlinear multiplicative problem in reflectance spectroscopy.[7] A similar process occurs with front-face fluorescence spectroscopy of vegetation.[14] The measuring instrument also picks extra signals such as noise and interferences and may be unstable in time. Preprocessing is performed after data acquisition to modify the data before further analysis. Most of the time, its purposes are to remove bias, to linearize the response of the variables and to remove variations which are not of interest in the analysis (interfering variance). Linearization is often desired because linear responses are easier to model than nonlinear ones. Unwanted variance will require more detailed models (more degrees of freedom) to isolate interfering variance from the variance of interest.[13] The selection of proper preprocessing steps can be based on the knowledge of the physics of the phenomenon used to probe a sample and on the expected result to derive from the measurements.

Considering a practical application in the field, it must be recognized that it will be difficult, if even possible, to obtain the absolute intensity of the fluorescence spectra. On one hand, the

Received 4 September 2008; accepted 4 April 2008.
* Author to whom correspondence should be sent. E-mail: pannetonb@agr.gc.ca.

0003-7028/08/6207-0747$2.00/0

geometry of the target–instrument system cannot be controlled. The probe to target distance will change constantly and the orientation of the target area (leaf blades) is random. This geometrical instability will result in a multiplicative effect on the signal. Plant fluorescence can be modeled by a modified Kubelka–Munk theory,[11] and it follows that the signal varies with leaf thickness, resulting in a further multiplicative effect.[7] In this context, proper preprocessing of the data to normalize spectra with respect to intensity is required (homothetic transformation). This should be performed in such a way that the potential for weed–crop discrimination is maintained or enhanced.

In general, a pure multiplicative effect can be modeled as

$$x_{i,j} = \zeta_{i,j} \cdot \beta_i + \varepsilon_{i,j} \tag{1}$$

where $x_{i,j}$ is the measured quantity for sample i in the waveband j, $\zeta_{i,j}$ is the true signal, β_i is the multiplicative factor for sample i, and $\varepsilon_{i,j}$ is the measurement noise. If β_i can be estimated, one can apply the following correction:

$$\zeta_{i,j} + \varepsilon_{i,j}/\hat{\beta}_i = x_{i,j}/\hat{\beta}_i \tag{2}$$

where $\hat{\beta}_i$ is the estimate of β_i. Equation 2 shows that when the multiplicative correction is applied, the apparent noise level will change from sample to sample and this may lower the performance of subsequent data processing.

Defining a method for performing a multiplicative correction amounts to defining an appropriate method for calculating $\hat{\beta}_i$. The most direct method is the use of an internal standard (see Ref. 7 for details and practical examples). The signal from the standard must be distinct and uncorrelated to the signal from the sample being analyzed. In fluorescence spectroscopy, this can be achieved by introducing a known amount of a fluorophore that is excited by the wavelength of the excitation source but that emits in a wavelength range where the samples do not emit. When this can be achieved, the multiplicative correction factor can be estimated as

$$\hat{\beta}_i = \frac{\sum_j (w_j \cdot x_{i,j})}{\sum_j (w_j \cdot x_{\mathrm{ref},j})} \tag{3}$$

where w_j are weights equal to 1 in the wavelength range corresponding to the unique emission range of the internal standard and 0 otherwise.

When no internal standards are available, MSC (multiplicative signal correction) can be used (see Refs. 7 and 13 for details and practical examples). Here, it is assumed that the true spectrum is the sum of a reference spectrum $X_{\mathrm{ref},j}$ and a sample-specific variation about this reference $\tilde{\zeta}_{i,j}$. Formally:

$$x_{i,j} = \zeta_{i,j} \cdot \beta_i + \varepsilon_{i,j} = (X_{\mathrm{ref},j} + \tilde{\zeta}_{i,j}) \cdot \beta_i + \varepsilon_{i,j} \tag{4}$$

Ideally, the multiplicative correction $\hat{\beta}_i$ should be estimated by ordinary least squares regression of $x_{i,j}$ on $(X_{\mathrm{ref},j} + \tilde{\zeta}_{i,j})$, that is:

$$\hat{\beta}_i = \frac{\sum_j x_{i,j} \cdot (X_{\mathrm{ref},j} + \tilde{\zeta}_{i,j})}{\sum_j x_{i,j}^2} = \frac{\sum_j x_{i,j} \cdot X_{\mathrm{ref},j} + \sum_j x_{i,j} \cdot \tilde{\zeta}_{i,j}}{\sum_j x_{i,j}^2} \tag{5}$$

From the right-hand side of Eq. 5, one can infer that if the sample-wise correlation between $x_{i,j}$ and $X_{\mathrm{ref},j}$ is much stronger than the correlation between $x_{i,j}$ and $\tilde{\zeta}_{i,j}$, then $\hat{\beta}_i$ can be estimated by a regression of $x_{i,j}$ on $X_{\mathrm{ref},j}$. This is the basis for standard MSC. In other words, MSC is appropriate when most of the signal comes from a steady source and the signal of interest produces fluctuations of relatively small amplitudes and with little correlation with the one from the steady source.

The last method for estimating the multiplicative correction is called normalization by closure.[7] In this case, the correction factor is estimated by

$$\hat{\beta}_i = \sum_j w_j \cdot x_{i,j} = \sum_j w_j \cdot (\zeta_{i,j} \cdot \beta_i + \varepsilon_{i,j}) \tag{6}$$

where w_j are the weights and where it is implied that $\Sigma_{j=1}^{K} w_j = 1$. Two limit cases appear:

(1) $w_j = 1/K$; $\hat{\beta}_i = \bar{x}_i$; that is, the standard normalization by closure, and
(2) $w_{j0} = 1$; $w_j = 0$, $\forall j \neq j_0$; $\hat{\beta}_I = x_{i,j0}$; that is, the normalization by a single wavelength.

When applying the standard normalization by closure, all resulting spectra have equal mean value. This puts emphasis on the shape of the spectra.

In our application, it is not possible to work with an internal standard because plants are to be probed as they appear in the field using front-face fluorescence. *A priori*, MSC is not appropriate as variations in fluorescence spectra from one plant species to the other are large to the point where the idea of a mean or reference spectra would be dubious. Therefore, normalization by closure appears to be the best option, and this paper will concentrate on this approach and especially on setting up w_j.

Some "artificial" intercorrelations can be generated when applying $\hat{\beta}_i$ to $x_{i,j}$ if one or some of the wavebands j dominate the sum in Eq. 6.[7] In addition to introducing unwanted intercorrelation, normalization by closure based on a spectral region carrying significant information (high amplitude) is not desirable as information is partially destroyed inside this spectral region. In the limiting case where the spectral region is limited to a single waveband, the information at that waveband is lost for further processing. In the spectrometry domain, where the variables are highly correlated, a single wavelength normalization at λ_i decreases the amount of information in the neighborhood of λ_i. To avoid introducing undesirable autocorrelations, the weighing vector w_j can be set to 0 in regions of the spectra where the signal is strong and carries useful information. In this case, the noise term in Eq. 6 may dominate and result in a noisy $\hat{\beta}_i$. This would amplify unwanted variance. In our classification application, there might be regions of the spectra where the signal is strong but carries very little discriminating information. Such a region would be an ideal candidate for normalization purposes. In the end, the optimum choice of a weighing vector w_j depends on the signal-to-noise ratio as a function of wavebands, on the autocorrelation structure of the spectra, and on the covariance between the classes and the wavebands. To investigate the effect of the shape of the weighing vector, one of the choices is to use a single waveband for normalizing the signal (all $w_j = 0$ except for a single j) at a time and to span the whole spectral region. This is the approach that was retained for this work. The objective was

TABLE I. List of weed species and corn hybrids.

Plant group	Species or corn hybrids
Corn hybrids	Elite 60T05
	Monsanto DKC 26-78
	Pioneer 39Y85 (RR)
	Syngenta N2555 (*Bt*,LL)
Monocotyledonous weeds	*Digitaria ischaemum* (Schreb.) I
	Echinochloa crus-galli (L.) Beauv.
	Panicum capillare (L.)
	Setaria glauca (L.) Beauv.
Dicotyledonous weeds	*Ambrosia artemisiifolia* (L.)
	Amaranthus retroflexus (L.)
	Chenopodium album (L.)
	Capsella bursa-pastoris (L.) Med.

to find the best waveband for computing $\hat{\beta}_i$ in the context of plant species fingerprinting using front-face fluorescence.

MATERIALS AND METHODS

Experiments. The experiment was conducted on seedlings of four corn hybrids and eight plant species. These 12 plant species/hybrids were in three groups (Table I): four *Zea Mays* (L.) hybrids, four annual monocotyledonous grasses, and four annual dicotyledonous weeds. Eight specimens of each plant species/hybrid were cultivated in a growth chamber in 1.07 L pots (12.7 cm diameter) of soil-less mix (Promix BX, Premier Horticulture, Quebec, Canada). Nutrients (20–12–20 at 95 g/L) were mixed with water and this mixture was applied as needed. The growth chamber temperature was 20 °C during the day and 12 °C at night, with a plateau of one hour at 16 °C between each change of temperature. The photoperiod was 16 hours of light and 8 hours of darkness. The relative humidity was maintained around 55%. Light was provided by 1000 W metal halide lamps at a distance of approximately one meter from the pots, giving about 600 W/m^2. Plants were laid out randomly on the table and a rotation of the pots was implemented to avoid adverse effects from any light or temperature gradients within the growth chamber.

For each pot, the date of emergence was recorded and data were gathered at 10, 20, and 30 days after emergence. Fluorescence was measured on the uppermost fully developed leaf. At 20 and 30 days, two measurements were performed on each leaf. For the dicotyledons, one datum was read from the main leaf vein and another was read from the leaf blade. For the monocotyledons, one measurement was performed on a point at the base of the leaf (lower 25% of the leaf blade) and another one at its apex (top 25% of the leaf blade). At 10 days after emergence, the leaves were so small that only a single measurement point, randomly located on the leaf, was used. Measurements were performed in a greenhouse under natural daylight to avoid the 60 Hz interference from artificial lighting. Plants were placed in the greenhouse one hour before measurements were carried out. The whole experiment was repeated in time on three separate occasions (three blocks of data). Therefore, the data set comprised a total of 1440 spectra (three blocks, three plant groups, four hybrids/species per group, eight replicates per hybrids/species, and five readings for each). Some data were rejected for various reasons (growth problems, instrumentation problems, etc.), yielding a validated data set made of 1392 spectra.

Plant fluorescence was induced by a xenon flash lamp (Spectra-physics Series Q Housing 60000 with a 5 J Xe pulsed arc lamp) controlled by an Oriel 68826 power supply (9 μs pulse width). The flash output was coupled to a fiber-optic bundle (Oriel 77578) using a condensing lens assembly (Oriel 60076) and a bandpass filter centered at 327 nm (20 nm full width at half-maximum (FWHM)). The induced fluorescence was collected by another fiber-optic bundle (Oriel 77532) and transported to the spectrograph (Oriel MS125 1/8 m). Using a length gauge, both fiber optics were positioned 5 mm above the leaf and pointed to the same spot on the leaf, 2.3 mm in diameter. The leaf blade was positioned perpendicular to the probe as judged by the operator. The spectrograph was modified by the insertion of a high-pass filter at the input port (400 nm: 5% at 388 nm and 80% at 405 nm) to cut off second-order effects. An internal charge-coupled device (ICCD) detector (Andor, DH 712–18F/03, 5 ns, Phosphore P43) was coupled to the spectrograph to record the spectrum. This camera has 512×512 active pixels. The wavelength axis of the spectra was along the horizontal axis with a resolution of 0.95 nm per pixel, and the central portion covering the range from 400 to 760 nm (378 pixels or wavebands) was retained. Vertically, the signal from 170 pixels in the center of the array was summed to yield the required one-dimensional (1D) spectra.

Fluorescence signals were acquired under ambient light using the same technique used by Norikane and Kuruta.[10] Under ambient light and without the UV excitation, 11 spectra were acquired at 10 Hz and averaged. Then, 11 spectra were acquired under UV flash excitation (10 Hz) and averaged. The difference between the two resulting spectra is the pure induced fluorescence signal. The fluorescence spectra were smoothed using an 11-point moving average filter.

The data acquisition procedure described above yielded a data set where spectra can be directly compared one to the other because the main source of variance was the sample-to-sample differences and other sources of variance (instrument noise and positioning errors) were small.

Data Processing and Analysis. Before applying preprocessing, individual spectra were further smoothed using the Savitzky–Golay technique[13] with a third-order polynomial and a 21-channel-wide window.

In order to evaluate the pertinence of preprocessing, a simple discriminating model based on computing the Mahalanobis distance was carried out. The Mahalanobis distance was defined as[12]

$$d_c^2 = (\mathbf{X} - \mu_c)\Sigma_{cr}^{-1}(\mathbf{X} - \mu_c)' \tag{7}$$

where d_c is the vector of the Mahalanobis distance of all samples with respect to class c, $\mathbf{X}$ is the matrix of the $x_{i,j}/\hat{\beta}_i$, μ_c is the center of mass of the class c, and Σ_{cr} is the matrix of variance-covariance for class c reduced to remove the effect of noise. This reduction was achieved as follows. First, by singular value decomposition[13] of $\mathbf{X}_c$ (the matrix containing the spectra of class c):

$$\mathbf{X}_c = \mathbf{USV}' \tag{8}$$

The k largest values of the $\mathbf{S}$ diagonal have been inverted and the others set to 0, yielding the matrix $\mathbf{S}^-$. Then, Σ_{cr}^{-1} was calculated as:

$$\Sigma_{cr}^{-1} = \mathbf{V}(\mathbf{S}^-)^2\mathbf{V}' \tag{9}$$

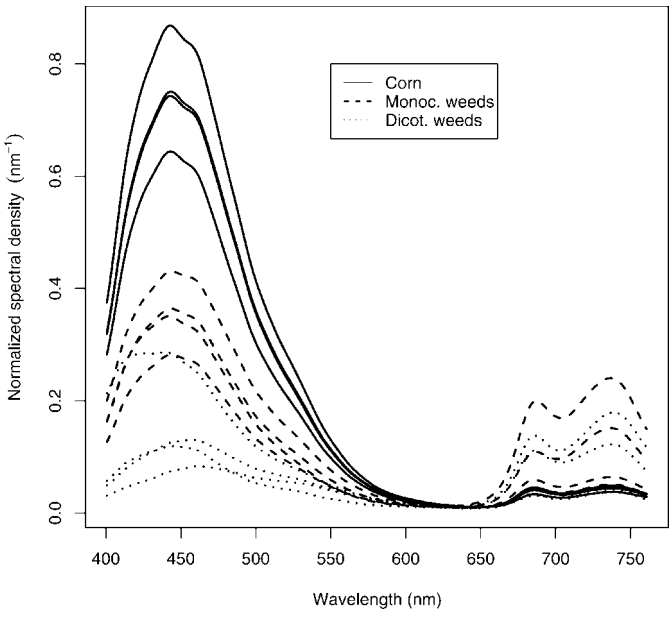

FIG. 1. Mean raw spectra. Spectra averaged per plant species/hybrids.

The number k has been estimated using principal component analysis (PCA) on $(\mathbf{X}_c - \mu_c)$. The appropriate number of principal components that should be retained was determined using the Scree-test of Cattell.[12] For each class, a vector of Mahalanobis distances was computed and a sample was assigned to the class corresponding to its smallest Mahalanobis distance.

RESULTS AND DISCUSSION

Raw plant fluorescence spectra were averaged per plant species/hybrids (Fig. 1). The spectra all displayed the characteristics of plant fluorescence. The blue-green fluorescence (BGF) peak at about 450 nm had a large amplitude. The amplitude of this peak depends strongly on the excitation wavelength. Excitation at 327 nm as performed here provided a much larger BGF signal than excitation at 360 nm or higher. In the blue-green region, spectra formed well-defined groups. Corn was distinct from all weeds and there was little confusion between monocotyledonous (monocots) and dicotyledonous (dicots) weeds. In the chlorophyll fluorescence (ChlF) part of the spectrum, there was more confusion between the three groups.

A principal component analysis was performed on the whole data set. Cross-validation was used to select the appropriate number of principal components (PCs).[13] After five PCs, the root mean square error of cross-validation was minimum. Loadings of the first five principal components were plotted against wavelength (Fig. 2). Each of the first five PCs explained variance in the BGF or ChlF regions with no or little variance explained in the other region. For example, the first PC (PC1) explained variance in the BGF region but none in the ChlF region. Furthermore, PC1, PC2, and PC3 displayed an alternating pattern (BGF, ChlF, BGF). From these observations, it is concluded that the emission of BGF is a process that is independent of ChlF emission. This is in agreement with available knowledge about the mechanisms of fluorescence that clearly shows that BGF and ChlF have

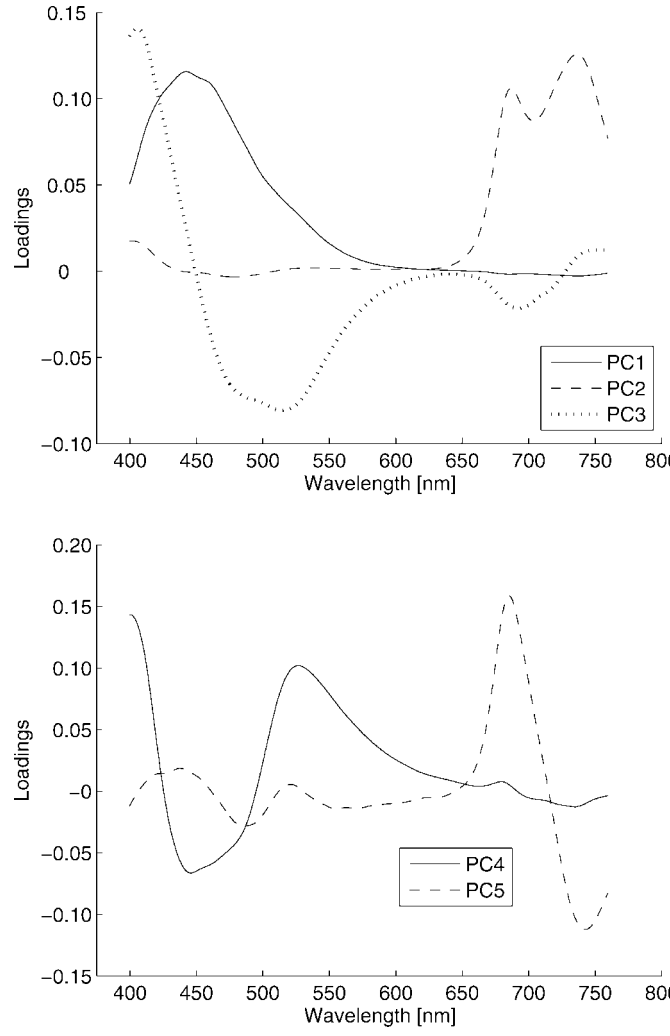

FIG. 2. Loading plots. First five principal components (PC1 to PC5) from raw data. Principal component analysis performed over the whole mean-centered data set.

distinct origins and change independently in response to physiological and environmental factors.[1]

For monocots, BGF was very significant (Fig. 1) and as discussed previously, normalization by closure can introduce "artificial" intercorrelation. This means that normalization by the sum over the BGF region of the spectra was not desirable. The "artificial" intercorrelation is clearly visible when looking at the loadings of the first three principal components (PCs). With raw data (Fig. 2), the first and third PC captured variance in the BGF region while the second PC captured variance in the ChlF region. When normalization by closure was applied, "artificial" intercorrelation was introduced between BGF and ChlF as all of the first three PCs captured variance in both regions (Fig. 3).

The discriminating model was applied to the smoothed raw data (no multiplicative correction). Classification success rate was calculated for four classification schemes (Table II, first line): per species/hybrids (12 classes), per plant group (corn/monocots/dicots), monocots/dicots (corn being a monocot), and corn/weeds. Very good classification results were obtained with the corn/weed scheme, followed by the corn/monocots/dicots scheme. This was in agreement with the clear

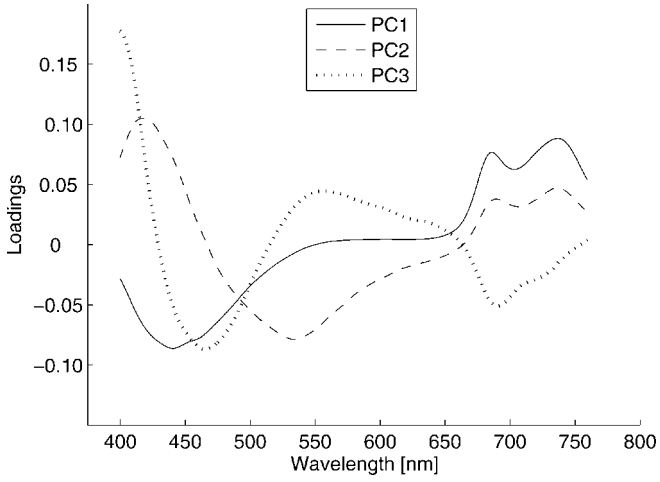

FIG. 3. Loading plots. First three principal components (PC1 to PC3) from data normalized by closure (all wavebands used). Principal component analysis performed over the whole mean-centered data set.

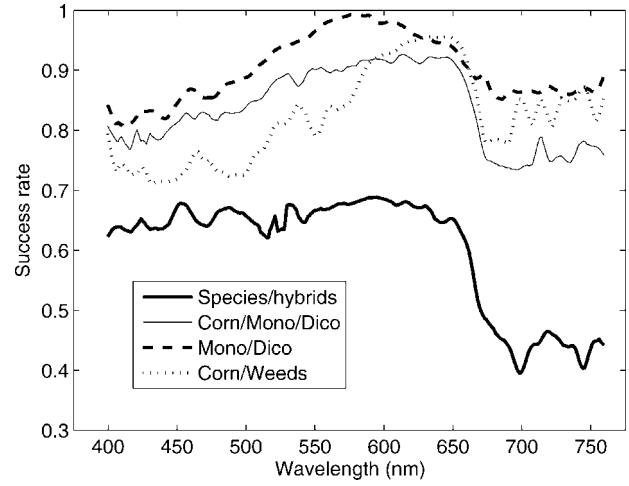

FIG. 4. Success rate for single waveband preprocessing.

segmentation of spectra according to these classes (Fig. 1). Classification of raw spectra by species/hybrids gave poor results.

Preprocessing was applied to the data set using all available wavebands as candidate normalizing wavebands (0.95 nm wide). For each of the four classification schemes, curves of the success rate as a function of normalizing wavelength were generated (Fig. 4). The four curves displayed some high-frequency noise. It was traced to shifts in selecting k on a per class basis. For example, as the normalizing waveband changed, the sawtooth pattern in the curve for the corn/weeds classification around 700 nm was associated with k oscillating between 3 and 4 for the corn class and between 3 and 6 for the weeds class. The overall larger scale variations are the ones of interest.

For wavebands in the BGF region, classification as monocots/dicots resulted in the highest success rate, approaching 100% near 580 nm. Classification per species/hybrids yielded the lowest success rate over the whole wavelength range. This result was quite normal and should be attenuated by the implicit difficulty due to the large number of classes. Corn/monocots/dicots classification had a higher success rate in the BGF region than the corn/weeds classification, showing that splitting weeds as monocots and dicots improved the performances. In the ChlF region, classification as monocots/dicots still had a higher success rate followed by corn/weeds, corn/monocots/dicots, and the species/hybrids classification. The fact that classification as monocots/dicots yielded better success rates made sense as it is the only classification scheme that follows botanical classification. Interestingly, this was also the scheme in which proper preprocessing produced the larger increase in classification success rate (Table II). Nevertheless, there seems to be scope for the corn/monocots/dicots or corn/weeds classifications, as success rates larger than 90% were obtained. The three-class scheme would be useful in the context of weed control in agriculture as specific herbicides can be used for monocot and dicot weeds. Finally, using data in the ChlF region ($>$685 nm) cannot be recommended as in all cases this resulted in a lower success rate when compared to performance obtained with raw smoothed data.

In general, using a waveband in the 550 to 650 nm region resulted in better classification performances except for the species/hybrids scheme. This is the region of the spectral gap between the BGF and the ChlF. For the monocots/dicots classification, the optimum waveband (579 nm) was located slightly below the minimum in the fluorescence spectra (630–635 nm) and for the corn/weeds classification, the optimum waveband was located slightly above (646 nm). These small shifts from the minimum of the spectra illustrated the compromise between avoiding spectral regions carrying most of the information (BGF and ChlF) and avoiding regions of lower signal-to-noise ratio (SNR), i.e., near the minimum of the spectra. An estimate of the SNR was computed as the residual of a five-component PCA. The estimated SNR varied from 20 to 95 dB (Fig. 5). On average at 20dB, the signal is ten times stronger than the noise. For the corn/monocots/dicots classification, there appeared to be no such shift as the curve was virtually flat from $\approx$575 nm to $\approx$645 nm, but this was the range where both monocots/dicots and corn/weeds classification performed better. This last observation indicated that the decrease in SNR induced by selecting a waveband near the minimum in the spectra was not significant enough to impair the classification performances for the corn/monocots/dicots classification.

TABLE II. Classification success rates for various multiplicative corrections.

	Classification scheme			
	Spec./ hyb.	Mono/ Dico	Corn/ Mono/Dico	Corn/ Weeds
Raw data	0.64	0.86	0.91	0.95
Standard normalization by closure	0.70	0.93	0.91	0.85
Single waveband (nm)				
593	0.69			
579		0.99		
614			0.93	
646				0.96
Wavelength range (nm)				
618 ± 25	0.70			
587 ± 25		0.99		
596 ± 25			0.94	
634 ± 25				0.96

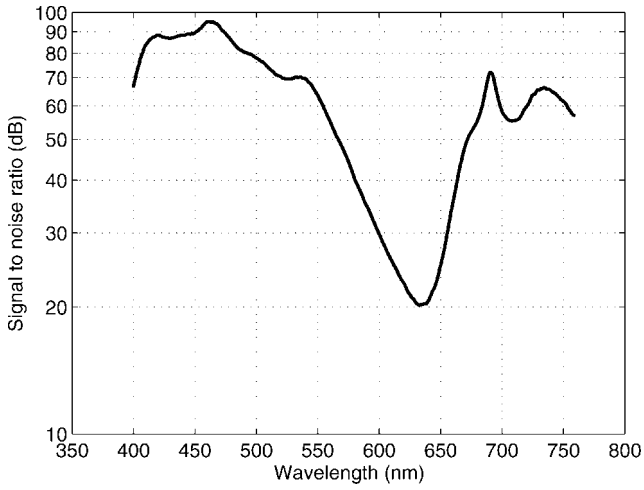

FIG. 5. Estimated signal-to-noise ratio.

Compared to using a single normalizing waveband, normalization by closure using all wavebands resulted in almost equal or lower success rates (Table II). The larger difference was for the corn/weeds classification. This was to be expected because corn was the class where the dominance of the BGF peak over the spectra was the largest and as such, where the introduction of "artificial" intercorrelation was the most detrimental.

Calculations were performed using 50 contiguous wavebands (48 nm band) for computing the normalization with a central wavelength in the range from 424 to 736 nm. The best normalizing band was selected (Table II). For all the classification schemes, the success rates were virtually equal to the rates obtained when normalizing from a single waveband. It was expected that using a wider wavelength range would result in improved classification since for uncorrelated white noise the contribution of the error term to $\hat{\beta}_i$ in Eq. 6 should decrease, increasing the SNR of $x_{i,j}/\hat{\beta}_i$. The noise level in our data was generally low and this explained why using large normalizing wavebands did not improve classification performances. In field applications, the SNR is likely to be reduced and the normalization scheme will have to be adapted to cope with this real-life situation.

CONCLUSION

Data preprocessing is usually achieved in a routine way, even in a blind way, the objective being to prepare the data for future use, e.g., to build a decision system. This work showed that each of the preprocessing steps carries important consequences. Our claim is that preprocessing has to be done according to the physical and/or biological meaning of the available information as well as according to the model selected to extract this information.

Spectrum normalization is required in order to compare data collected under variable conditions. This study showed, and recalled, that a normalization is not at all a neutral operation. It comes to information reduction. The normalization range has to be carefully chosen: areas with discriminant information have to be avoided, as well as areas of low SNR. When non-discriminant wavelengths present a low signal level, a good practice consists of applying normalization according to a narrow range of wavelengths to improve robustness.

Preprocessing was evaluated on the basis of classification success rates using all the available wavebands as candidate normalization data. This was computed with twelve different classes, with each class corresponding to one plant species/hybrids. It is interesting to notice that the classification performances were improved when groups that are meaningful to the user and that follow botanical classification were considered, such as monocots versus dicots or corn versus weeds.

The next step is the building of a discriminant system based on the preprocessed data and to extend results from spectra collected under ideal laboratory conditions to field data.

ACKNOWLEDGMENTS

The excellent technical support of Mr. G. St-Laurent and Mrs. M. Piché is hereby acknowledged. This work was financially supported by Agriculture and Agri-Food Canada and by a Natural Sciences and Engineering Research Council of Canada Discovery Grant.

1. Z. G. Cerovic, G. Samson, F. Morales, N. Tremblay, and I. Moya, Agronomie **19,** 543 (1999).
2. E. W. Chappelle, F. M. Wood, W. W. Newcomb, and J. E. McMurtrey III, Appl. Opt. **24,** 74 (1985).
3. E. W. Chappelle, J. E. McMurtrey III, F. M. Wood, and W. W. Newcomb, Appl. Opt. **23,** 139 (1984).
4. G. A. Johnson, S. V. Mantha, and A. D. Day, J. Plant Physiol. **156,** 242 (2000).
5. M. Lang, F. Stober, and H. K. Lichtenthaler, Radiat. Environ. Biophys. **30,** 333 (1991).
6. H. K. Lichtenthaler and J. Schweiger, J. Plant Physiol. **152,** 272 (1998).
7. H. Martens and T. Naes, *Multivariate Calibration* (John Wiley and Sons, New York, 1996).
8. F. Morales, Z. G. Cerovic, and I. Moya, Biochim. Biophys. Acta **1273,** 251 (1996).
9. F. Morales, Z. G. Cerovic, and I. Moya, Aus. J. Plant Physiol. **25,** 325 (1998).
10. J. H. Norikane and K. Kuruta, Trans. ASAE **44,** 1915 (2001).
11. A. Rosema, W. Verhoef, J. Schroote, and J. F. H. Snel, Remote Sens. Environ. **37,** 117 (1991).
12. G. Saporta, *Probabilités, Analyse des donnéess et statistiques* (Editions Technip, 1990).
13. B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig, and R. S. Koch, *PLS Toolbox 4.0 -Reference Manual for use with Matlab* (Eigenvector Research Inc., Wenatchee, WA, 2006).
14. J. Zarco-Tejada, J. R. Miller, G. H. Mohammed, and T. L. Noland, Remote Sens. Environ. **74,** 582 (2000).