Scattered information: philosophy and practice of near infrared spectroscopy

I. Murray
SAC, Aberdeen, AB21 9YA, UK

Introduction

Most of what we know about the fabric of matter and the cosmos comes from its interaction with electromagnetic radiation. Only photons are small enough and energetic enough to probe atomic and molecular structure and so provide instant analysis of composition and quality attributes of materials. Only recently with the proliferation of fast computers and chemometrics has it become possible to test everyday foods and agricultural products routinely for quality.

In this presentation I explain the philosophy and practice of NIR spectroscopy from an applications viewpoint. Applications bound way ahead of the theory but knowledge of the theory helps our understanding of what may and may not be possible now and in future. Spectra can be correlated to composition or quality to provide instant technical information of social and economic relevance. Applications of correlation transform spectroscopy have relevance in agriculture, food, environment, medicine, pharmaceuticals, petrochemicals, polymers, paper and textiles as testified in the presentations here at Cordoba NIR2003.

E–M spectrum and origin of spectra

EM radiation consists of streams of energy packets called photons that behave like a sine wave having orthogonal electric and magnetic vectors propagated at 3×10^8m s^{-1}. The EM spectrum (Figure 1) is a continuum of photon energies E or frequencies ν with the photon energy being proportional to frequency:

$$E = h\nu \quad \text{Bohr–Einstein law} \tag{1}$$

where h is Planck's constant (6.6×10^{-34} J s).

The frequency ν is related to wavelength λ and velocity c by

$$\lambda\nu = c \tag{2}$$

The photon energies of X-rays knock out inner shell electrons causing ionisation while UV-visible photons cause transitions in outermost valence electrons. Infrared photons have much weaker energies that correspond to covalent bond stretching and bending vibrations in molecules while even weaker photon energies of microwaves cause molecules to rotate. The NIR corresponds to vibrational overtones and combinations; echoes of the bond stretching and bending vibrations occurring in the mid-IR.

Near Infrared Spectroscopy: Proceedings of the 11th International Conference.

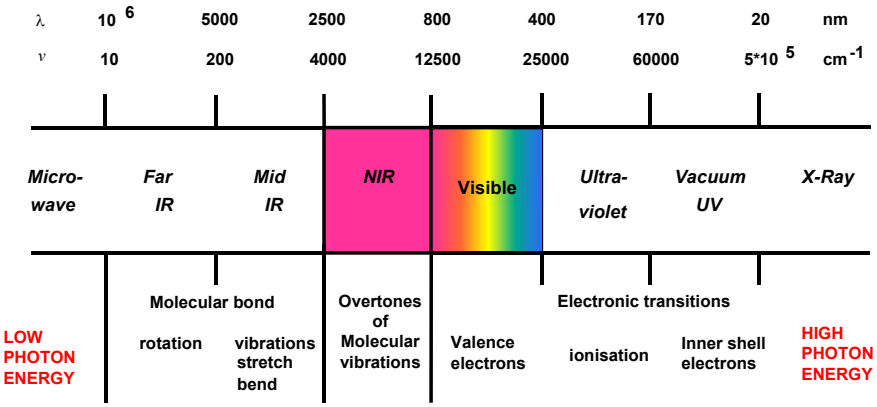

Figure 1. The electromagnetic spectrum

Wavenumber and wavelength

Conventional mid IR uses a frequency related property called wavenumber $\overline{\nu}$ (cm^{-1}) where $\overline{\nu}$ = ν/100c. This is the number of cycles in a one centimetre wave train. Conventional UV-vis and NIR spectroscopy use wavelength λ, usually in nanometres. These units can be converted by

$$\text{wavenumber cm}^{-1} = \frac{10{,}000{,}000}{\text{wavelength nm}}$$

When light is incident on matter it slows down and the wavelength changes but the frequency remains the same. Fermat's Principle states that light takes the path between two points that takes the least time. This leads to the law of regular reflection and Snell's law of refraction that defines refractive index. Incident light (I_0) may be reflected off (I_r), transmitted through (I_t), or absorbed (I_a) by matter. Usually all three processes coexist depending on whether the material is a reflective surface (for example,. mirror), transparent gas, liquid or crystal (for example, quartz) or an opaque substance (for example, carbon black), just to mention the extremes. In diffuse reflectance a perfect matt diffuse reflector will obey the Lambert Cosine law such that it appears equally luminous in all directions. The reference tile in reflectance instruments is an example. Reflection from packed powders always has a small amount of specular or mirror-like reflection superimposed on more predominant diffuse reflection. It is the diffuse component that is analytically useful. The scattering of light from powders arises mostly from sharp angle refraction at boundaries where the pore space is air pockets having a large difference in refractive index from the solid phase. Filling pore spaces with liquids like water and oil enhance penetration, causing absorption for all the components of the matrix as well as those due to the added liquid. For this reason wet soil looks darker than dry soil and an oil spot can make paper transparent.

Scattering

In powder reflectance finer grinding leads to lowered absorption (as log1/*R*) due to less deep penetration and is more marked in regions of greater absorption. This shift is multiplicative rather than parallel causing a ramped baseline offset unique to each specimen. The hallmark of diffuse reflectance of the same material ground to different degrees of fineness is that the standard deviation of their spectra have the same absorption pattern as that of their mean. Ground coloured substances in the visible appear paler as they are ground finer. Several mathematical attempts to remove or normalise this effect have been devised, notably Multiplicative scatter correction (MSC)[1] and standard normal variate/detrend (SNV/DT).[2] MSC centres spectra about the mean of a sample set while SNV/DT acts independently on each sample in turn. While both do a good job reducing spectral variance by "squashing" spectra together, both cause undesirable shifts away from the wavelength regions that intuitively correlate to analytes. The second derivative of log1/*R* is unique among treatments that retain faithful correspondence to the original wavelength space as log reciprocal reflectance. Since it is the *shape* of the spectral signature that conveys information, derivatives with optimised segment and gap are most appropriate in centring spectra around a zero line although they do not eliminate the offset rather they centre it.

Sample presentation modes

Numerous optical effects prevail but the important point is that all the incident photons should be accounted for as being either *reflected, transmitted* or *absorbed*:

$$I_0 = I_r + I_t + I_a \quad (4)$$

In practice the absorbed fraction is assumed to be the fraction not accounted for (dissipated as heat) while either the transmitted or reflected fraction is measured relative to a control to measure the incident radiation I_0. 'Lost photons' are assumed to be absorbed photons. Herein lies many dangers. Sampling, sample preparation and presentation are important because no amount of smart maths can recuperate information lost due to spectra collected under unsatisfactory measurement conditions.

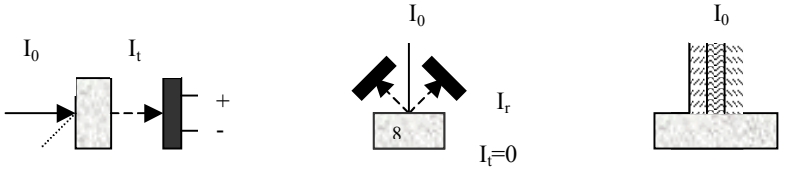

Figure 2. Sample presentation modes

Transmission mode

In transmission the detector is placed *behind* the specimen. In transmission through transparent isotropic specimens (fluids) of fixed path length *l* (usually 1 cm) the Beer–Lambert law applies to monochromatic light:

$$\text{Absorbance}_\lambda = \log 1/T = \log (I_0/I_t) = \log I_0 - \log I_t = \varepsilon c l \quad (5)$$

where ε is a constant of proportion called the molar absorptivity and c is the concentration in moles per litre

Absorbance, being a log ratio of two intensities, has no units so we invent "absorbance units" noting that A has a log scale such that $A = 1$ means 10% of the photons are detected while $A = 2$ means 1% are detected etc. relative to the standard. Sixteen bit resolution ($2^{16} = 65{,}536$) corresponds to an upper limit of 4.8 absorbance units. Provided the absorbance is not too large, good signal to noise ratios are observed.

Here absorbance is linearly related to concentration with slope ε so that the analyte concentration can be calculated or interpolated from a calibration plot. However if the specimen is not perfectly transparent but is cloudy, milky or even opalescent then light scattering will be confounded with absorption and non-linearity will occur especially at shorter wavelength. As scattering increases (and depending on the nature of the scatter) a point is reached when the back scattered radiation flux, directed back towards the source, becomes far greater than that emerging by diffuse transmission through the specimen. The specimen gets more opaque with longer path length so that better signal to noise may be obtained in R mode. Certainly more light will be reflected than transmitted in such cases so that absorbances are encouragingly lower in R mode. However a word of warning, the re-emitted light may emerge from only a very thin surface film that may not represent the bulk specimen. This is particularly true about wet specimens having physiological water concentrations around 70%.

Reflectance mode

In the limit, when a specimen is "infinitely" thick, it is opaque and no photons emerge in transmission ($I_t = 0$) so that all the incident photons must either be reflected or absorbed. An intermediate mode, ambiguously called "transflectance", places a diffuse reflector behind the specimen to ensure photon return and prevent structured contribution from the backing plate. Most everyday substances we wish to analyse are substantially opaque. A representative solid angle of the light reflected from the specimen is measured relative to a matt ceramic reference tile as control to measure I_0. In R mode, absorbance is defined as:

$$\text{Absorbance}_\lambda = \text{Log } 1/R = \log (I_0/I_r) \propto \text{number of chromophores encountered in light path} \quad (6)$$

In reflectance the mean path length is indeterminate and varies with each specimen and with the wavelength - penetrating less deep in regions of higher absorption. So the reflectance spectrum of just one specimen at all wavelengths does not arise from exactly the same depth profile of the specimen. The Beer–Lambert relation still applies but path length, concentration and the proportionality constant are combined. One solution lies in measurements made at several wavelengths allowing a matrix solution. In the simplest case of measuring one analyte, a sensor wavelength and a non-absorbing reference wavelength measurement could be solved as a pair of simultaneous equations for two unknowns: path length and concentration. In practice multiple linear regression (MLR) was applied to the spectra of a calibration set of typical specimens to generate an equation of the form:

$$\%\ \text{Analyte} = B_0 + B_1 w_1 + B_2 w_2 + B_3 w_3 + \ldots . B_n w_n \quad (7)$$

where B_0 is the regression constant and B_x are partial regression coefficients, w_x are absorbances at specified wavelengths 1, 2, 3…..n (± derivative pre-treatment) added stepwise in order of their relevance as explained variance.

Collinearity

While MLR provides a solution there is a snag - collinearity**.** Regression expects all explanatory wavelengths to be independent and not inter-correlated otherwise the matrix becomes ill-conditioned and unstable solutions arise. Adjacent wavelengths are sure to be highly correlated; so too are all wavelengths in R mode that suffer baseline offset. Also features that arise from the same chromophore such as its first and second overtones will inter-correlate. As wavelength data are often highly inter-correlated, it is better to transform these into a new set of non-correlated ("mutually orthogonal") variables such as principal components or partial least squares vectors before regression. However this necessary step comes at the cost of losing track of an easy interpretation of how the model relates back to spectra. More recent progress in NIR is attributed to these new chemometric calibration methods.

Vibrational spectroscopy

For a diatomic molecule to absorb infrared radiation it must vibrate such that it changes its *electric dipole moment*. A non-linear molecule with n atoms will have 3n-6 normal vibrational modes. These are of two kinds: stretching and bending. Bond stretching along the inter-nuclear axis requires more energy than bending the angle between bonds. So stretching occurs at higher frequencies than bending. If a photon of *exactly* the correct energy impacts on a molecule the photon is absorbed and the molecule is raised from the ground vibrational state (V=0, where almost all molecules reside at room temperature) to the first exited vibrational state (V=1).

$$\Delta E = E_2 - E_1 = h\nu \qquad 8)$$

This requires a photon from the mid IR where the absorption is very strong and the transition is called the *fundamental*. The fundamental has a *classical* frequency ν that derives from Hooke's law as:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \qquad (9)$$

where $\mu = m_1m_2/(m_1+m_2)$ and m_1 and m_2 are the atomic masses and k is the force constant (stiffness) of the bond. The important point here is that atomic masses and the bond stiffness dictate the frequency or wavelength of absorption. Every chemical group in a molecule having different component atoms with different mass cause peaks at different frequencies characteristic of that group for example:

	$\overline{\nu}$ cm^{-1}	λ nm
CH stretch	2975–2840	3360–3520
OH stretch	3670–3230	2730–3100
NH stretch	3540–3300	2830–3030

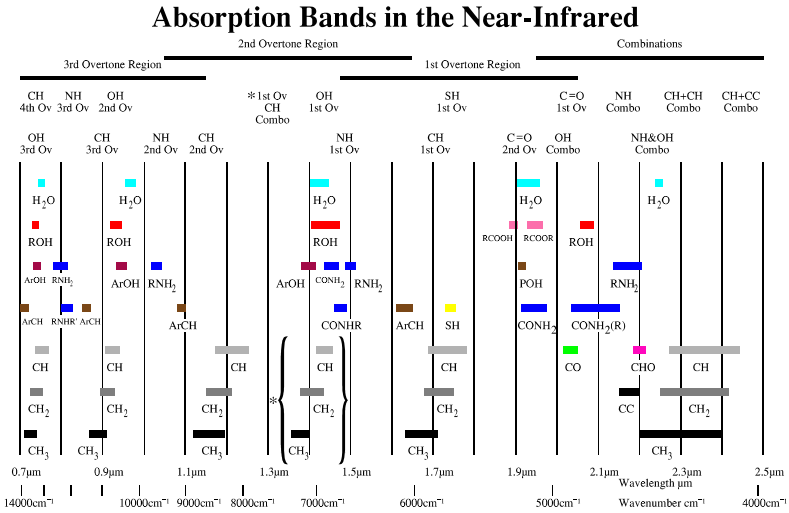

Figure 3. Absorption bands in the near infrared

Weak forces like hydrogen bonding shift bands to longer wavelength (lower frequency). A correlation map built up from experience of scanning pure chemicals can show where characteristic bands arise from particular functional groups[3]. (Fig.3)

In classical mechanics only *one* vibrational frequency is allowed although all amplitudes are possible. Equation (9) represents a ball and spring model for the molecule and its covalent bond that together act as a *simple harmonic oscillator*.

The classical model is a good approximation but if it were strictly true there would be no bands in the NIR at all! The ball and spring model obeys Hooke's Law:

$$F = -kq \qquad (10)$$

where the restoring force (F) is proportional to the displacement (q) from equilibrium with k as the force constant. In fact bonds are not quite like springs in that they are more easily stretched than compressed so that they behave as anharmonic oscillators.

Quantum mechanics leads to only discrete vibrational energy levels being allowed. These turn out to be nearly (but a bit less than) whole number multiples of the fundamental frequency. They are overtones or harmonics that occur with much lower probability than the fundamental and so they are an order of magnitude weaker, involve larger energy jumps that correspond to resonance conditions with more energetic photons from the NIR. Each successive overtone gets progressively weaker, more anharmonic and departs more from whole number multiples of the fundamental frequency. So the energy levels converge. The NIR consists of such overtones and combinations of mid IR fundamental vibrations especially those that are most anharmonic such as those involving the low mass hydrogen atom for example, –CH, –OH and –NH. This turns out to be most appropriate for gross composition analysis of food since its major components: water, lipid, carbohydrate and protein are well characterised by HOH, $–CH_2–$, –CHOH– and –CONH– features respectively.

Quantum mechanics applied to the harmonic oscillator introduces the vibrational quantum number V = 0, 1, 2, 3….so that the vibrational energy is given by:

$$E_{vib} = (V+½)h\nu \tag{11}$$

where ν is the classical frequency.

A consequence of Heisenberg's uncertainty principle arises here in that even in the vibrational ground state ($V = 0$) there is still some vibrational energy—so-called zero point energy. However, this would still impose equally spaced vibrational energy levels that lead again to just one absorption band. In contrast anharmonicity leads to unequally spaced vibrational energy levels that converge up to the dissociation energy of the bond when the bond is broken. The potential energy well for a diatomic molecule plots the potential energy (y axis) versus the inter nuclear distance (x axis) which has an equilibrium value, the bond length. This PE well takes the form of a parabola for Hooke's law for the harmonic oscillator whereas the parabola becomes lopsided and distorted for the anharmonic oscillator when the bond is stretched to breaking point. This shape is called a Morse curve (Figure 4).

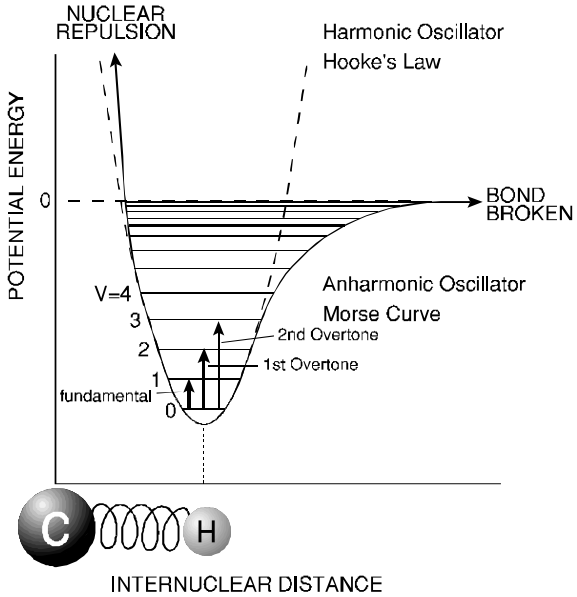

Figure 4. Potential energy well for a covalent bond

When the Schrödinger equation is solved using a cubic potential it gives converging energy

$$E_{vib} = \bar{\omega}_e\left(V + \tfrac{1}{2}\right) - \bar{\omega}_e\bar{x}_e\left(V + \tfrac{1}{2}\right)^2 \quad \text{cm}^{-1} \tag{12}$$

levels that fit observation:

where $\overline{\omega}_e$ is the vibrational wavenumber which a classical oscillator would have for an infinitesimal displacement from equilibrium and $\overline{\omega}_e\overline{x}_e$ is called the anharmonicity constant. A snag is that $\overline{\omega}_e$ cannot be measured directly unlike the case of the harmonic oscillator. Values for $\overline{\omega}_e$ and $\overline{\omega}_e\overline{x}_e$ have to be estimated from a least squares fit from the V ← 0 observed overtone line spacing using the equation:

$$\overline{E}_v - \overline{E}_0 = \overline{\omega}_e V - \overline{\omega}_e\overline{x}_e V(V+1) \quad (13)$$

Here the theory needs modified by the observed line spacing. The important point is that *theory* has to appeal to *practical* measurement to arrive at consensus.

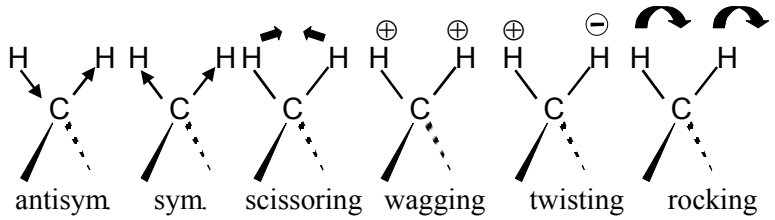

Figure 5. Normal mode vibrations of the methylene group –CH_2-

The theory of vibration–rotation spectra was developed from simple diatomic molecules usually isolated in the gas phase. Only simple cases allow solution of equations. In polyatomic liquids or solids, mixtures, or complex tissues, empirical approaches have to be used. Effects arising from neighbouring groups in the molecule, associative effects of solvents, hydrogen bonding and Fermi resonance conspire to complicate any further progress connecting theory and practice. The methylene group –CH_2– is a good example of the complexity of vibrational spectra that occurs in most organic compounds (Figure 5). It has six normal modes; two stretching and four bending, of which five are IR active. These give rise to a series of overtone and combination tones recurring across the NIR. The spectrum of polyethylene is an example.

Even the simplest amino acid, glycine: H_2N–CH_2–COOH, with just 10 atoms will give rise to 24 normal modes of vibration each of which will have many overtones and combinations that overlap and cluster in the NIR. Although the situation may seem hopelessly complex, a great averaging takes place such that the spectra of macromolecules and even the architecture of tissues give rise to apparently simple, smooth rolling spectra in the NIR that are still analytically useful. Progress involves an empirical process of examining spectra of pure compounds, proportional mixture models, food matrices before and after extraction of components like water and lipid. Mixture models do help identify useful wavelengths and devise the best *S/N* presentation of specimens to an instrument but such models are at best qualitative and are never adequate for quantitative analysis of 'real' specimens. Even experience with 'closed' sample sets with internal cross validation or independent validation sets are not proven until tested on 'open' sets stretching into the future. In this sense NIR uses inductive rather than deductive reasoning so that any calibration model is tentative, empirical and open to revision in the light of new information or experience.

Whereas traditional chemical analysis isolates analytes from the sample matrix by separation techniques like solvent extraction or chromatography, NIR allows the intact sample matrix to participate in *predicting* an analyte using a multivariate model developed on a 'real' specimen population. In many cases this involves *aliasing* the analyte to major components of the matrix.

Water

Water holds a special interest in NIR spectroscopy not least because attempts to accurately measure moisture in grain led K. H. Norris to the rebirth of this long neglected region. A love-hate relationship exists between NIR workers and the strongly absorbing physiological water in plant and animal tissue. Water is an extremely strong absorber with broad bands due to hydrogen bonding. These obscure and dilute the matrix, limiting penetration and causing high absorbance, deformed peaks and detector saturation. Wet samples limit the use of longer wavelengths.

An isolated water molecule has a boomerang shape with 104º bond angle and 0.96 Å bond length with two lone pairs of electrons ("bunny's ears") on oxygen making for near tetrahedral co-ordination in liquid and ice. This results in formation of hydrogen bonds between neighbouring water molecules causing unexpectedly high melting and boiling point - water would be expected to exist as a gas. Having three atoms (n) and being non-linear leads to 3n–6 = 3 normal modes for water: symmetric stretch v_1, asymmetric stretch v_3 and bend v_2 (Figure 6)

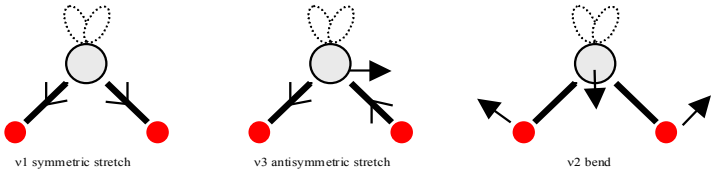

Figure 6. Normal mode vibrations of water

In the NIR the $v_3 + v_2$ stretch–bend combination absorbs about 1940 nm and the $v_1 + v_2$ first overtone occurs from 1412 to 1490 nm. Both bands moving to longer wavelength with increasing hydrogen bonding. A band at 970 nm arises from the $v_3 + v_2$ combination overtone and at 739 nm from the $v_1 + v_2$ combo third overtone. Temperature, pH and solvation shells around salt ions affect water band position and shape causing challenging problems as well as providing opportunities to measure analytes like sodium chloride that otherwise have no absorption bands at all.

Water has a formula weight of 18 so 18 grams or 18 mL is one mole and water itself is 55 molar in water! Table 1 lists NIR bands for liquid water[4]. The absorption coefficient (ε) is 114 Litre mole^{-1}cm^{-1} at 1940 nm and 26 at 1450 nm. Thus a 1 µm path length would have an absorbance of 0.63 at 1940 nm and 0.14 at 1450 nm. Water *in vivo* in plant and animal tissues is therefore a seriously strong absorber causing noisy distorted band heads and obliterating lesser chromophores. In fresh tissue containing 70% water, wavelengths beyond 1300 nm may not be useful or at best provide information from just a surface film. In such excessively wet tissues information is more likely to be got at shorter wavelength in the Herschel infrared and in "windows" (like 1700 nm) between major water peaks where S/N is more favourable. An unusual situation where troughs are more useful than peaks.

Table 1. NIR bands for liquid water in transmission at 20°C[4].

λ region µm	λ_{max} µm	λ_{max} cm^{-1}	Path length cm	Band pass nm	Absorption coefficient L mol^{-1} cm^{-1}
0.7–0.9	0.76	13200	12.97	6	0.026
0.9–1.15	0.97	10300	1.99	3	0.46
1.15–1.35	1.19	8400	0.99	3	1.05
1.35–1.8	1.45	6900	0.09	3	26.0
1.8-2.5	1.94	5160	0.02	4	114.0

Calibration strategy

Before embarking on calibration the first question to consider is the intended purpose of analysis by NIR. "Quality" means fitness for purpose. Any analysis must be fit for the purpose. As NIR is a secondary method that depends on another primary method for calibration, this reference method must be capable of doing a good job. This is in terms of precision (repeatability) and accuracy (approach to the correct result) in relation to the variance of the analyte in the sample population to be tested. The ratio of SEL to the SD_{pop} gives a guide to the best value for R^2 that may be obtained on NIR calibration:

$$\frac{SEL}{SD_{pop}} = \sqrt{\frac{n(1-R^2)}{n-2}} \cong \sqrt{\left(1-R^2\right)} \text{ for large n} \qquad (14)$$

where n is the number of samples.

By substituting for SEL and SD_{pop} a projected value for R^2 can be found. If SEL was half the value of SD_{pop} then the best R^2 could be is 0.75 Improvement can only come from a smaller SEL or a larger SD_{pop} from a wider range in concentration. The reciprocal of this ratio is called RPD[5] and is often given as a figure of merit for a method. Any standard error of estimate (SEL, SEC, SECV or SEP) can be substituted to find the corresponding R^2. If the projected R^2 is too small and no scope for improvement exists then much effort and disappointment can be avoided by not attempting calibration at all!

Calibration set

Selection of the calibration sample set is the most critical part of calibration. The acronym W.E.P.T. applies here. A Wide range in analyte concentration, Even, rather than normal distribution, Precise reference analyses and "Typical" samples are a counsel of perfection for a calibration set. The most frequently asked question is how many samples are needed to establish a calibration. The answer is the least number that adequately represent the population. As it is easier to scan samples than do reference analyses, the best policy is to scan many and use their spectra to select the least number that adequately represent the whole population. In this way redundant samples are eliminated and the cost of reference analysis can be cut. The ISI software uses the H statistic or Mahalanobis distance to achieve this. Global *H* measures the distance of a sample from the centroid while Neighbourhood *H* (N*H*) measures the average distance to six nearest neighbours. It detects 'lonely' specimens that are genuine outliers. Only one specimen is needed to characterise each neighbourhood. One thing worth noting is that an acceptable calibration model can be obtained from a large amount of relatively imprecise data.[6] Validation is where accuracy matters most. The remarkable thing about statistical sampling is that in a large population with fixed variability, the standard error of the mean depends mainly on the number of samples and only to a small extent on the fraction of the population sampled. The mean of 100 specimens is almost as precise whether the population is 200,000 or 20,000 or 2,000. So sampling can greatly reduce the amount of measurement needed.

The snag with traditional calibration lies in the fact that any normally distributed population of similar samples usually forms a diffuse swarm in multivariate space with the population density decreasing outwards from the mean. The distance G*H* measures this departure from the mean or centre in much the same way as the standard deviation. G*H* values greater than 3.0 are considered outliers. Clearly there will be a trade off between the scope of a calibration model, the range it can cope with, and its ability to discriminate between samples that are close in composition.

While apparently excellent calibration models can be performed, lack of robustness on independent validation frequently occurs. The cause is difficult to track. Over fitting with too many terms introduces spurious correlation unique to the calibration set especially if there are few samples. Failure to balance the composition range and H statistic of the calibration and validation sets likewise risks failure. Poor reference method, changed sample preparation or presentation will compound these errors leading to lack of robustness. Persistence usually results in a working model at the end of the project.

Outliers and "typical" samples

Outliers can be of two kinds t and H. A sample assigned the wrong reference value will fail the Students' t test on prediction. Its reference analysis must be repeated and checked. One barley among 100 soy specimens will have an atypical spectrum and be detected as an H statistic outlier. An H outlier is a "lonely" sample with no neighbours within its domain. Spectra are an excellent 'xenoprobe' for detecting such foreign specimens. But if a sufficient number of these are deliberately added they will no longer be detected as outliers.

The snag is what do we define as a "typical" sample? We can define what we do in calibration as an "omnianalyser" that reports any analyte or attribute of a specimen based on inference from a reference set whose membership is validated—but how is membership validated? The inference arises from some form of correlation transform derived from spectra. It is easy to see that models may be based on populations of just one kind of cereal or across several cereal species or all types of forage—fresh or fermented or across mixed feeds manufactured for several species of farmed animals. Just one barley sample among 100 wheat specimens would be a spectral outlier while 50 such among 100 would give a compromise calibration that may accommodate both cereals but at a cost of lost precision. A trade-off must exist between the scope or breadth of a calibration model and its ability to discriminate between two specimens that are nearly but not quite the same. One approach has addressed this problem. The "LOCAL" concept of Shenk and Westerhaus (ISI) uses a large spectral library to select specimens closest to any new specimen and perform an analysis using just these nearest neighbours. Artificial neural networks (ANN) are another calibration method that copes with non-linear modelling of complex sample populations. Calibration strategy remains at the research edge of chemometrics so we can expect future developments.

Conclusions

NIR spectra are weak echoes of their mid IR fundamental vibrations. They owe their existence to anharmonicity. Theory has to appeal to practical measurements of the observed line spacing to concur with observation. Bands are repeated as an overlapping attenuated series across the region. They provide means of scaling response appropriate to concentration; like a built-in dilution series. Absorption band position on the wavelength axis permits qualitative identification of an analyte while the peak amplitude on the absorbance axis permits quantitative measurement in diffuse transmittance or reflectance. Many bands, including those of the matrix participate in the model. Collinearity occurs everywhere in spectra. This needs to be overcome using principal component analysis or partial least squares vectors.

Correlation transform modelling is applied to a set of specimens typical of those to be tested in future along with their reference analyses. The entire sample matrix participates in analysis by aliasing analytes or attributes to major components. The selection of calibration and validation sample sets that are representative and appropriate to the purpose of analysis remains problematic. Robustness of any calibration model and avoiding over-fitting by having many authenticated specimens, remains the main challenge of analysis by NIR spectroscopy. The compelling need for

speed, timeliness in reporting and not least the cost suggest that for low value, high volume commodities like food and feed that testing will either be done by NIR in real time or it may not be done at all.

References

1. H. Martens, S.A. Jensen and P. Geladi, *Proceedings of the Nordic Symposium on Applied Statistics,* Ed by O.H.J .Christie. Stokkland Forlag, Stavanger, Norway, pp. 205–234 (1983).
2. R.J. Barnes, M.S. Dhanoa and S.J. Lister, *Appl. Spectrosc.* **43,** 772 (1989).
3. I. Murray *Near Infrared Diffuse Reflectance/Transmittance Spectroscopy*, Ed by J. Holló, K.J. Kaffka and J.L. Gönczy. Akadémiai Kiadó, Budapest, Hungary, pp. 13–28 (1987).
4. J.A. Curcio and C.C. Petty, *Journal of the Optical Society of America* **41(5),** 302 (1951).
5. P.C.Williams and D. Sobering, in *Near Infrared Spectroscopy: The Future Waves*, Ed by A.M.C. Davies and P.C. Williams. NIR Publications, Chichester, UK, pp. 185–188 (1996).
6. D.B. Coates, *Spectroscopy Europe* **14(4),** 24 (2002).