Food Integrity by Non-invasive/Non-destructive Methods

Chapter 14: Non-destructive / non-invasive method development, evaluation, and transfer

Dolores Pérez-Marín[1*] and Tom Fearn[2]
1. Agro-Livestock Production Systems Engineering Research Group (ISAG). Non-Destructive Spectral Sensors Unit. Faculty of Agriculture and Forestry Engineering (ETSIAM), University of Cordoba, Campus de Rabanales, 14071 Cordoba, Spain
2. Department of Statistical Science, University College London, 1-19 Torrington Place, WC1E 6BT London, UK

Corresponding Author dcperez@uco.es

Abstract

This chapter attempts to give an overview of the main aspects to be taken into account for spectral data processing using chemometrics, considering the development and validation of calibrations and qualitative predictive models, considerations on signal pretreatments, possibilities for instrument standardization and model transfer, together with the final performance of predictive models for routine use. The main focus will be on NIR spectroscopy. However, the general principles discussed in the chapter are transferable to other non-invasive methodologies that produce multivariate raw measurements. The authors wish to emphasize those aspects that, in their experience, are critical in the development of applications using data from nondestructive spectral sensors.

Keywords: multivariate analysis, chemometrics for spectral data, calibrations, qualitative models, signal pretretaments, derivatives, scatter correction, , instrument cloning

1. Preliminaries

The first step in developing an application is to select the instrumentation that will be used. This means deciding on a measurement method and on a particular instrument. The first question to be answered is which of the many available methods are likely to work successfully for this application. The subsequent chapters of this book give some guidance, and there is an extensive literature on many of the methods that should be consulted. Factors to be considered in choosing a measurement principle and an instrument include the following:

- Is there some solid basis for believing that the candidate method is able to measure the parameter(s) of interest at this level in this type of sample?

- Where will the measurement be made? On a laboratory bench? On-line above a moving conveyor belt? Through a window into a pipe carrying a liquid? In a field? Very different instruments may be indicated for these situations (Figure 1).

- What level of accuracy is needed? Remember here that one of the advantages of most non-invasive methods is the ability to take measurements very rapidly. Taking advantage of this speed to make multiple measurements can dramatically reduce sampling uncertainly, thus compensating for a larger analytical uncertainty.

- Will the instrument need to be calibrated for your application, and if so do you have the resources to do this?

Figure 1. Examples of in situ analysis using portable NIR devices.

Particular considerations in the case of spectroscopy include whether the wavelength range of the instrument is appropriate for the application and, in the case of samples such as intact fruits, whether the penetration depth will be adequate to achieve a representative measurement.

Having chosen an instrument, it is important to optimise carefully the conditions of measurement. The details will vary with the instrumentation, but in general there will be choices that affect what samples the instrument sees, how they are presented to the instrument, and what time is employed for the measurement. Unless the application is a very standard one, this will require some experimentation in a laboratory. In the case of spectroscopy, simply looking at raw measured spectra under various conditions can be very informative about noise levels and about whether the numbers of photons reaching the detector lie in a sensible range. At a slightly more sophisticated level, measures of spectral repeatability, for example the variance between replicate spectra, averaged over wavelengths, can be very useful. For applications where the samples are systematically non-homogeneous, intact fruits for example, it is not enough to have good-looking and highly repeatable spectra, the light needs to have penetrated far enough into the sample to carry the desired information. This is harder to investigate, but can be done in various ways, for example by measuring samples of a range of thicknesses against a background with a spectral signature (Figure 2).

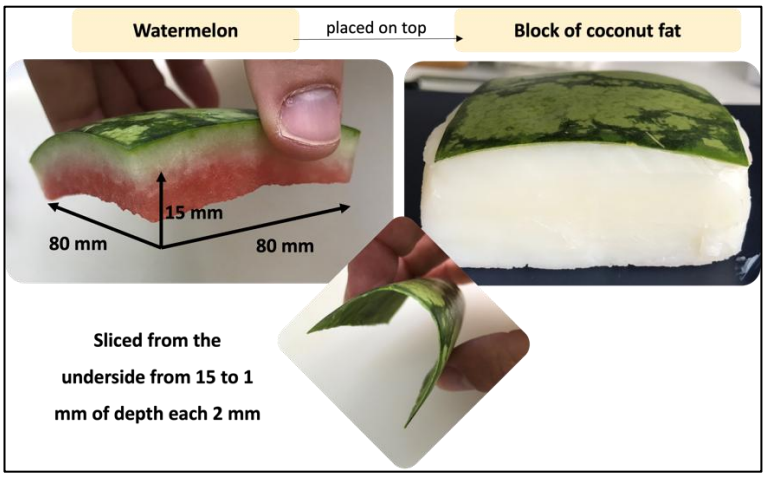

Figure 2. Example of an experiment to measure the deep of light penetration in watermelon (Vega-Castellote et al. 2023).

Effort spent on this part of the method development will usually be more important for success than time spent fine-tuning sophisticated calibration algorithms. No amount of mathematical wizardry will produce a good calibration from bad spectra.

Much of the remainder of this chapter is about the treatment of the multivariate raw data that is measured by many non-destructive methodologies, an area that is often referred to as chemometrics. The focus will be on NIR spectroscopy, whose development as a quantitative method both stimulated and benefited from the development of chemometrics. The general principles discussed below, as well as many though not all of the details, are transferrable to other non-invasive methodologies that produce multivariate raw measurements.

2. Calibration

Most non-destructive methods are indirect and require calibration. Many, especially spectroscopic ones, require so-called multivariate calibration (Martens and Næs 1989) in which the data from multiple measured variables is combined to predict the parameter or parameters of interest. For example, NIR instruments typically measure spectra at tens or hundreds of wavelengths. These spectra are rich in information about both the physical and chemical properties of the sample. The problem is that this information is contained in overlapping peaks in a spectrum which is often strongly affected by scattering effects. The result is that the data from a single wavelength will not be enough to provide a calibration, but methods that use data from multiple wavelengths, perhaps even using the entire spectrum, have proved successful in many applications. Some of these methods will be described briefly in Sections 6 and 7, but first we consider the general approach to making such a calibration.

The complexity of the algorithm used to develop the prediction rule can vary from something as simple as multiple linear regression (MLR) through to very sophisticated methods from machine learning. However, the underlying methodology is generally the same in all cases. A set of samples, called either a calibration set or a training set, is assembled, and for each sample we obtain both the multivariate non-destructive measurement and the laboratory reference measurement that it is desired to predict. The

algorithm learns to predict the reference measurement for these samples, and we then hope that the trained algorithm will be able to accurately predict the reference measurement for other samples not included in the training set. For this scheme to work successfully, it is essential that the training samples adequately represent the population of samples for which it is desired to make predictions. Most obviously, the full range of the parameter or parameters of interest should be covered. Less obviously, but equally important, the training samples need to represent the natural variation in all the other properties that might interfere with the measurement. Thus, for example, one should not calibrate on samples from a single batch or a single supplier if the calibration is to be used on samples from other batches or other suppliers. Variability in the conditions under which the samples are measured also needs to be represented. For example, collecting spectra from the training samples after they have equilibrated to a uniform moisture content and temperature in the laboratory is not a good strategy if the calibration will subsequently be used on samples in which these properties vary.

There is no simple answer to the obvious question "How many training samples do I need". So far as representing population variability is concerned, the number needed will vary considerably with the application. When it comes to calibration algorithms, the more sophisticated the algorithm, the more samples it will need. For a relatively simple algorithm like partial least squares regression (PLSR), 100 samples might well be fine if this allows an adequate representation of all the relevant interfering factors. Sometimes one might get away with fewer, but 10 or 20 is never going to be enough. If you plan to use a sophisticated machine learning algorithm, which will typically have much more flexibility to adapt itself to the training set, expect to need many more training samples (hundreds at least, possibly thousands).

3. Looking at the data

Before starting on the actual calibration, it is advisable to investigate the data measured on the training samples. If there are obvious problems of some sort, it is better to deal with them before spending time fitting calibration models. In the case of the reference measurements, this would usually involve no more than checking for unusual or unexpected values. In the case of multivariate data from the non-destructive sensor the task is more complex. If the measurement is spectroscopic, the spectra should be examined. Do the spectra have the expected overall shape? Are a few of them very different from the others? Are there noisy regions that could usefully be discarded, particularly towards the ends of the spectrum? Are there wavelength regions where the absorbance is 100% for some or all of the spectra? Most of these problems should not occur if the sample presentation and measurement conditions have been carefully optimised, but there is always the risk of a few odd spectra whenever many are measured (Figure 3).

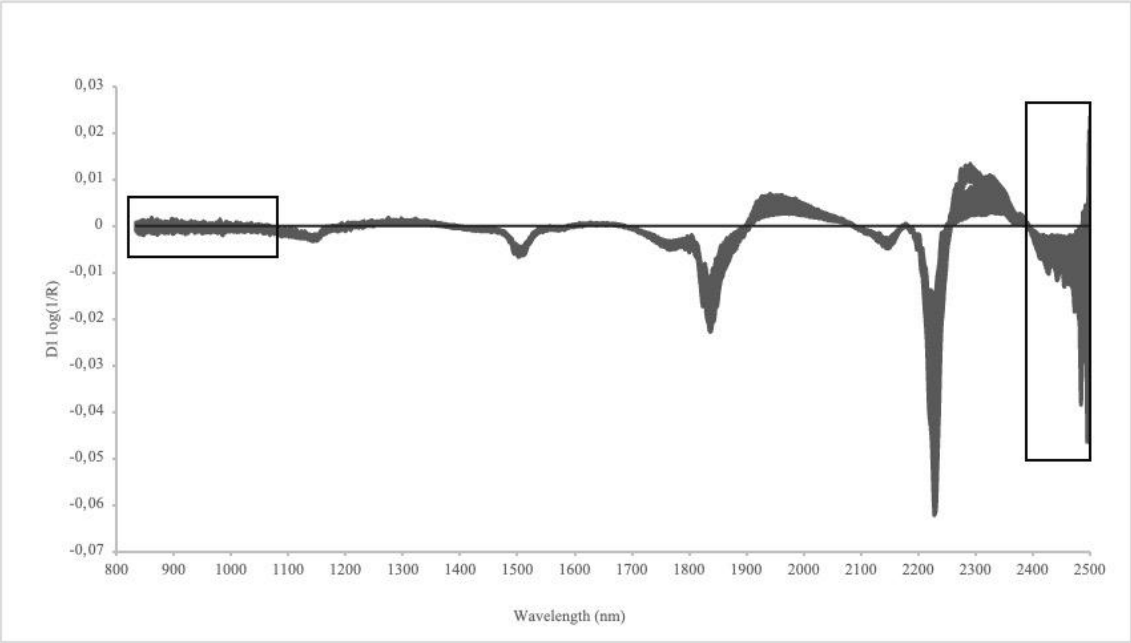

Figure 3. Spectra pretretated with first derivative and 1 data point gap, where the noisy areas in the sides can be appreciated.

There is a limit to what the eye can detect in raw multivariate data. One technique from multivariate statistics that can reveal patterns in multivariate data that would otherwise be difficult to see is principal component analysis (PCA), (Næs et al. 2002). One way to plot multivariate data would be to take one axis for each variable and plot the samples as points in a multidimensional space. With 2 or even 3 variables one could then inspect this plot (Figure 4). With more than 3 variables this is not possible. What PCA does is to centre each variable so that the origin of the axes is in the middle of the cloud of data points and then rotate the axes so that the first new axis, the first principal component (PC), captures as much as possible of the total variation between the samples, the second captures as much as possible of what remains, and so on. Then a plot of any pair of PCs gives us a two-dimensional window into the multivariate space, and it is often the case that looking at a few such plots for the PCs early in the sequence will reveal interesting features of the data. For example, one might see outliers, i.e. samples far from the main cloud of points, or some structure, with the samples grouping into two or more clusters. Outliers should be investigated. Maybe there was a problem with the measurement, or perhaps the outlier is an unusual type of sample. A split into clusters may be good news if the aim of the calibration is to classify samples into groups, assuming the clusters correspond to the groups of interest and not to some other factor. It may be less welcome if the aim is a calibration for some measured property and the split corresponds to an interfering factor, because the clustering shows that the multivariate measurement is strongly affected by this factor. It is also possible that you might see something totally unexpected. The value of looking at plots like this is that you can sometimes be surprised by what you see.

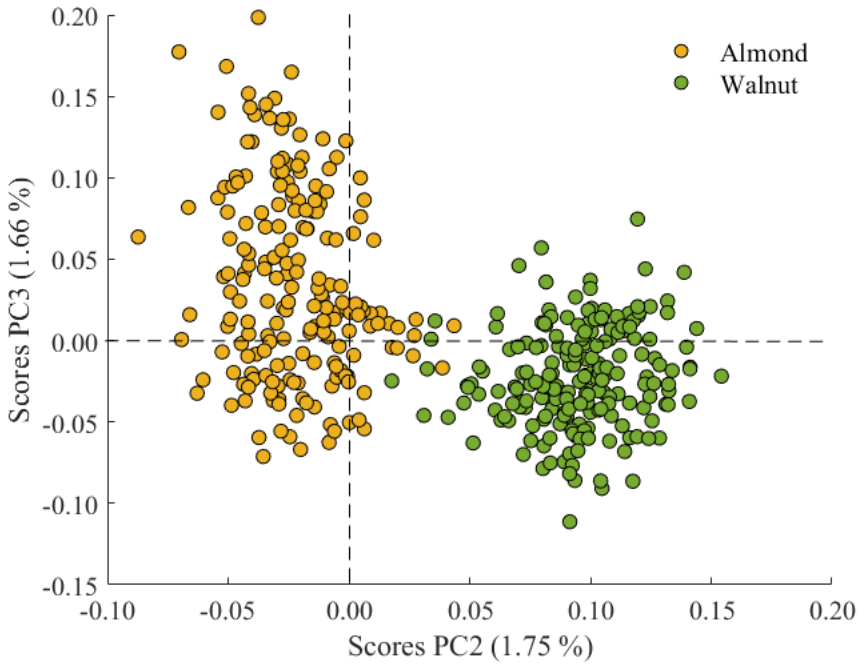

Figure 4. Example of PCA plot of a set of nuts.

4. Data pre-treatment

It is quite common to apply some mathematical pre-treatment to the raw data before inputting it to the calibration algorithm (Figures 5 and 6). The types of pre-treatments used tend to vary with the instrumentation chosen, but are typically aimed at removing high-frequency noise by smoothing what is essentially a continuous spectrum, at removing scattering effects that would otherwise dominate spectral variability between samples, or at standardising input variables by centring and scaling in some way.

Pre-treatment is particularly important for NIR reflectance spectroscopy, where the major part of the variability between the raw absorbance spectra of different samples is often due to scattering. Removing at least some of this can make it much easier for the calibration algorithm to identify and use the relevant signal. The use of derivative pre-treatments will remove additive baseline shifts, with a first derivative removing horizontal linear baselines and a second derivative removing sloping linear baselines. Multiplicative pre-treatments such as Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC) are aimed at the removal of multiplicative scatter effects (Næs et al. 2002). These may be used instead of or in combination with derivatives. None of these methods is completely successful in removing the unwanted variability, because the effects are neither additive nor multiplicative and are wavelength dependent, but they can help a lot.

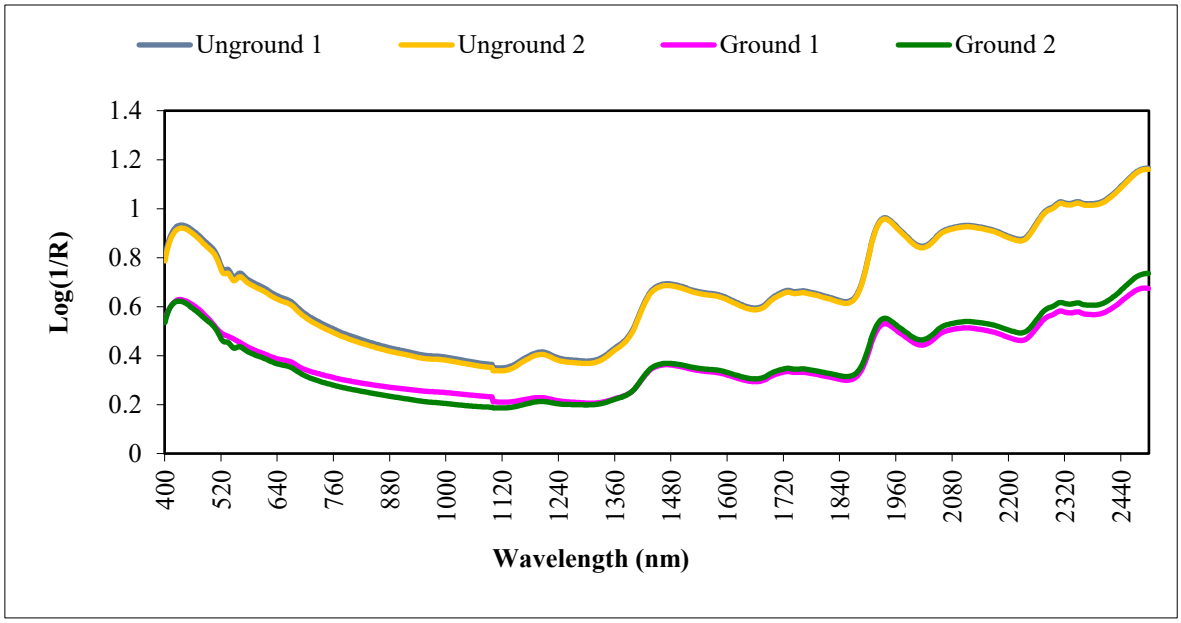

Figure 5. Raw spectra of feed samples analysed ground and unground.

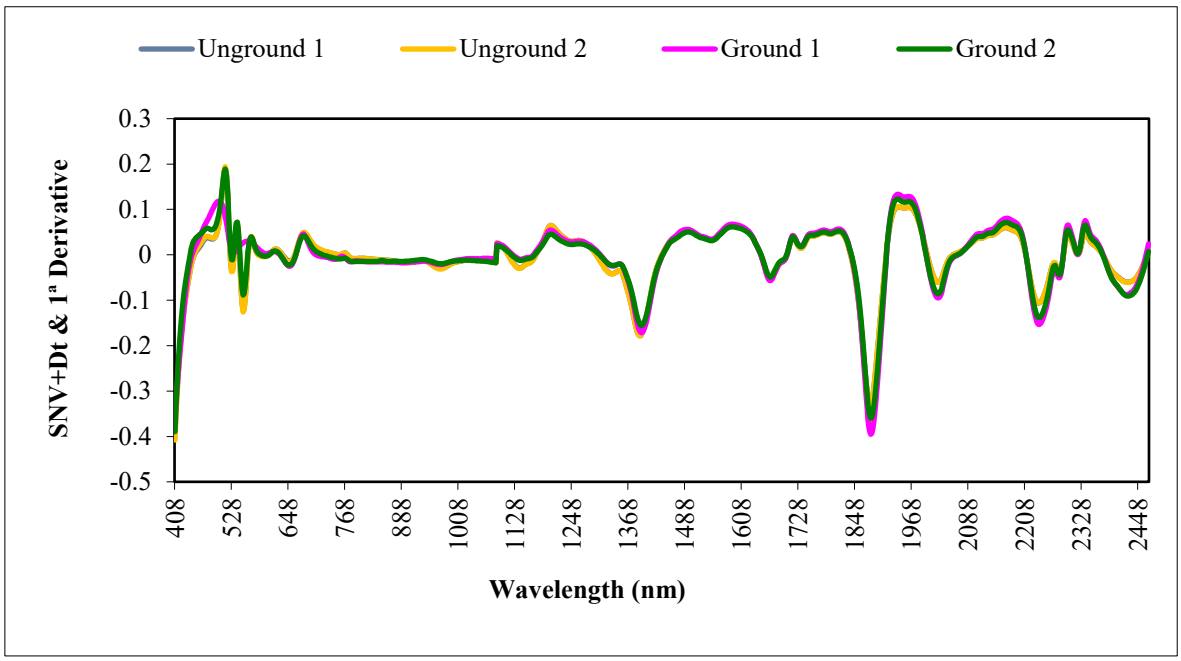

Figure 6. Pretreated spectra of feed samples analysed ground and unground.

Raman spectra are often superimposed on a smooth baseline that arises from fluorescence and may be variable between samples, and for these spectra may be useful to employ a specialised algorithm that tries to remove this baseline (Bocklitz et al. 2011).

Most software packages that focus on multivariate calibration of spectroscopic data will offer a range of suitable pre-treatments. The algorithms employed vary a little, but probably not enough to matter. For example there is not a lot to choose, in most applications, between SNV and MSC. Which pre-treatment, or pre-treatment combination, works best will depend on the instrument and the application, and finding it is part of the tuning process described in the following section.

In any case, it is always important to consider that signal pretreatment remove information from the data, for, although the effects are usually positive, sometimes it could remove interesting data for specific applications. For example, the scatter effect is closely related to the particle size and physical characteristics of the samples analyzed. If we apply any scatter correction method, alone or combined with derivarives, to quantify the chemical composition of a product, the results are usually improved, but if the objective of the application is, for example, the prediction of textural properties or the discrimination between categories of a product, which could be related to some of its physical features, the use of scatter correction may not be recommended. So the choice to use a signal pretreatment should be a thoughtful decision.

5. Tuning and validation

When the calibration exercise involves learning how to predict the parameter of interest from an indirect method such as spectroscopy that provides a large number of measured variables for each sample, a problem known as overfitting may be encountered. When there are more measured variables than samples in the training set, it will usually be possible to find a linear prediction rule that works close to perfectly on the training samples, rather in the same way that we can put a straight line through a pair of points on a two-dimensional plot, or a plane through three points in three dimensions. Even with more samples than variables, some of the noise in the training data will be explained by chance correlations with the many variables available and lead to a trained algorithm that has over-adapted to this particular training set. In the face of this problem most of the algorithms used for multivariate calibration attempt to control the overfitting in some way. For example, Principal Component Regression (PCR) reduces the high-dimensional measured data to a modest number of principal components and uses only this much lower-dimensional data in its prediction rule, which is a multiple linear regression (Martens and Næs 1989, Næs et al. 2002). The challenge is to choose the best number of components. As the number of components increases, the fit to the training data will always get better, so statistics from this fit cannot be used to make the selection. However, if we predict samples that are not part of the training set, the predictions will get worse when we start to overfit to the training set and we can use performance statistics from these predictions to choose the number of factors. Most, if not all, multivariate calibration algorithms require the optimisation of one or more tuning parameters to be carried out in this way.

If a large number of samples are available for the training, one possibility would be to split them into 2 sets, one for training, one for tuning. More usually samples are at a premium, and then cross-validation (Næs et al. 2002) may be used for the tuning. In this procedure, the training samples are divided into blocks of approximately equal size. Then each block in turn is removed from the training set and predicted by a rule trained on all the remaining ones. At the end of this loop, all samples have been predicted once using a rule trained on other samples, and the resulting performance statistics can be used to optimise tuning parameters. The block sizes can be as small as one sample, but both theoretical considerations and practical experience suggest that splitting into around 4-10 blocks is a better option. Whether one uses a single split or cross-validation, it is important to ensure that the predicted samples are genuinely different to the training samples. For example, if the training set comprises 2 subsamples of each of 100 bulk samples, the duplicates should be kept together during any split, so that data from one is not used to predict the other. An even more extreme case would be to have replicate spectra on the

same physical sample split across the training and tuning sets. Averaging the replicate spectra before splitting is probably the best way of being sure to avoid this.

Once the prediction algorithm is tuned, its performance needs to be assessed on a further set of independent samples. Nomenclature varies, but here this will be called validation, and the set of samples a validation set. It might be thought that the performance statistics from the tuning exercise would provide an adequate assessment of likely future performance, but the problem with this is that the optimisation involved in the tuning means that these performance statistics are almost certainly over-optimistic. If the algorithm is a simple one, such as PCR with its single tuning parameter, this over-optimism may not be serious. If the algorithm is more sophisticated, for example a neural network with many tuning parameters, there is much more scope for over-optimism. One consequence of this is that comparing algorithms of different complexity via their performance on the samples used to tune them is not valid.

Validation sets may be split off from the training set and kept apart throughout the training exercise or, better, they may be a completely separate set of samples collected and measured after the training is complete. The more genuinely independent the samples in the validation set, the more convincing will be the validation. For example, if you have data on samples of fruit from three harvest years, the best validation statistics will be achieved by ensuring that all three years are represented in both training and validation sets. The most convincing validation though, would be to take the samples from year three as the validation set. After all, you presumably plan to use the calibration in future years and a realistic assessment of its likely performance when used in this way is what is needed.

6. Quantitative calibration algorithms

When the number of predictor variables measured by the sensor is small, NIR reflectance measurements at four wavelengths for example, multiple linear regression (MLR) may be an adequate tool for deriving a multivariate calibration. MLR is a standard statistical tool used across a wide range of fields for finding the best fitting linear equation relating one dependent variable to multiple predictor variables (Freedman 2009). Its advantages are that it is extremely simple and that there are no tuning parameters to be optimised. Its disadvantages are that it needs a training set with many more observations than variables and that the fitting becomes unstable when the predictor variables are strongly correlated amongst themselves. This latter problem is common with NIR spectral data, though spectral pre-treatments, especially multiplicative ones, can reduce its severity. With measurements at just a handful of wavelengths, the simple scheme of choosing one reference wavelength and, separately for each sample, expressing the data as ratios to the measurement at this wavelength, can work quite well. Its disadvantages mean that MLR does not work well when the number of variables is more than a few. For example, early NIR instruments with six interference filters were routinely calibrated with MLR, but when later versions had wheels with 19 filters, MLR was too unstable to use with all 19 measurements. To fit a linear prediction equation when there are large numbers of predictors, for example an NIR spectrum measured at 100 or even 1000 wavelengths, something more sophisticated than MLR is needed.

The most commonly used approaches, when the number of predictors is large, involve some sort of dimension reduction applied to the predictor variables, allowing MLR to be

used with a smaller number of variables. One option is to use a small subset of the original predictors. The issue with this is the computational cost of the search. There are approximately 10^8 subsets of size 4 from 100 for example, so an exhaustive search would be time consuming. There are stepwise search procedures and stochastic search methods of various sorts that are designed to avoid an exhaustive search, and these have been used successfully in applications like this one (Andersen and Bro 2010). Because there are so many possible subsets, there is a high risk of overfitting, and care needs to be taken to tune and validate properly as described in Section 5. In any case, the final selection needs to make some sort of sense if it is to be used with confidence. In an NIR application, for example, there should be at least one wavelength in the subset that can be associated with what is being predicted. If you know from the chemistry that one particular variable should be included, include it from the start, don't waste time and effort letting the computer find it or, worse, not find it.

Variable selection methods can and do work, and may be the preferred approach in cases where measurement time can be reduced by not measuring all the variables. However, the alternative approach of reducing the dimension of the problem by constructing a small number of new variables or factors that aim to capture most of the information in the full set is both simpler and quicker. One method for doing this, PCA, has been described in Section 3, and its combination with MLR to make PCR was described in Section 5. The tuning parameter for PCR is the number of factors to be included in the regression step. This can be optimised using cross-validation or a separate tuning set as described in Section 5. An alternative approach that is sometimes suggested is to choose the number of factors that explain some arbitrarily chosen proportion of the total variance in the original variables, 99% say. However, this number of factors will not necessarily be the best choice for prediction. Choosing by optimising cross-validated predictions is to be preferred. The input variables should always be centred, i.e. have their mean subtracted, before input to the PCA step. Whether they should also be scaled so that each has the same standard deviation is less clear cut. If the input variables are measured in different units or have variability that ranges over orders of magnitude, scaling seems essential to give each of them a chance to contribute. On the other hand, the different amount of variability at different NIR wavelengths results at least in part from the fact that there are stronger signals in some places, and it may be preferable not to scale spectral data so that this information is preserved. If in doubt, it may be worth trying both options.

PCR suffers from the problem that at the dimension reduction stage PCA is constructing new variables without being told what they are to be used for. Capturing as much as possible of the variability in the set of original variables is not necessarily the same as constructing variables that will be good for the prediction task. The information most useful for prediction may not appear in the early principal components, with the result that quite a few factors are needed in the MLR to get good predictions. PLSR is an alternative approach that aims to construct more relevant factors. These factors are, like those from PCA, linear combinations of the original variables. The difference is that they are constructed, sequentially, to optimise a criterion that rewards both explaining variance in the original variables and having a high correlation with the y to be predicted: the covariance between the constructed variable and y. PLSR works very like PCR but will almost always produce solutions that use fewer factors. For this reason, and because there is plenty of good software available, it has become the standard approach in NIR spectroscopy and is widely used in other applications. As with PCR, the number of factors needs to be optimised using cross-validation or a separate tuning set. The usual

performance statistic for this optimisation in quantitative prediction problems is root mean squared error (RMSE), which is the square root of the average of the squared prediction errors on the samples predicted. Other statistics exist, but this is the simplest and most obviously relevant one, as it directly measures the size of the prediction errors.

For the vast majority of applications of multivariate calibration, PLSR works very well and nothing more sophisticated is required. Its main limitation is that it produces a linear predictor because the factor construction and the final prediction equation, both involve linear combinations of variables. In some situations nonlinear predictors will work better. One example is when the range of the y to be predicted is very large, moisture content ranging from 5% to 50% say. Then it is common to see curvature in the observed versus predicted plot. Another is when the sample sets involved are large and heterogeneous, when PLSR may not be flexible enough to adapt to this heterogeneity. In such cases it may be necessary to turn to a more flexible algorithm. There are many to choose from, with local regression methods, support vector machines (SVM), and artificial neural networks (ANN) being some of the common options (Pérez-Marín et al. 2007). At the time of writing this, convolutional neural networks (CNN) are the current hot topic (Walsh et al. 2023). Describing all of these is beyond the scope of this chapter. A warning that applies to all of them is that the more flexible the algorithm the more care needs to be taken to avoid overfitting. The latest deep learning approach might look more exciting, but it is wise to try PLS first and, if it works, not to look any further.

7. Qualitative calibration

For some specific applications of nondestructive spectral sensors in the assessment of food integrity, a qualitative approach could be more suitable. Some examples of these applications are those related to:

- the absence or presence of an ingredient — such as the presence of beef in a minced meat,
- belonging to a specific class or commercial category — such as crop variety or the geographical origin of rice or cheeses, or
- authentication issues or fraud detection — such as the authentication of extra virgin olive oil, the detection of adulteration of sweet almonds with bitter almonds or the detection of melamine in milk powder.

Qualitative analysis aims to classify samples according to their spectral features, grouping together samples with similar spectra and separating them from those with different spectra (Downey 1996). Thus, many qualitative methods are based on computing the similarity between the spectra of different samples, using tools such as spectral correlation or spectral distances. Among the different options of distances that exist, the most common is the Euclidean distance or variations of it, although, since it does not use data normalized to the size and shape of the groups, its use poses the problem that it can give rise to very different solutions by simply changing the scales of measurement of the variables. A very popular alternative is the Mahalanobis distance, which incorporates a data standardization procedure. However, this could give rise to problems if there is high collinearity in the data (Vandeginste et al. 1998; Heise and Winzen 2004).

In qualitative analysis, two main types of classification methods can be identified depending on whether or not the class assignment of the training set is *a priori* known (Naes et al. 2002):

1.- **Unsupervised methods**. This do not require any prior knowledge of the objects to perform the grouping or classification but the methods define the classes. These are considered as exploratory methods, for which the main purpose is to identify structures within the population. The most common unsupervised methods are PCA, described in Section 3, and cluster analysis. The latter is a multivariate classification technique that defines the natural patterns of grouping within collectives, so that the profiles of samples within the same group are very similar to each other (internal group cohesion) and those of samples from different groups are different (external group isolation). Other unsupervised pattern recognition methods, based on ANN, include Kohonen neural networks (Kohonen et al. 2001).

2.- **Supervised methods**. Supervised pattern recognition methods are taught using training sets in which the classes and the class membership of the samples are already defined (Vandeginste et al. 1998). There is a wide range of chemometric procedures for supervised qualitative analysis. Among these methods, can be distinguished two main types: discriminant analysis and class modeling methods. The main difference between these two types of approaches lies in the fact that discriminant analysis methods divide the space into as many regions as classes defined in the learning collective and classify any unknown sample into one of these classes according to which of the regions it falls in. On the other hand, class modeling techniques establish a boundary separately for each class, with samples inside the boundary judged to be possible members of the class. As a result samples may be assigned to one class, to several (when the spaces are overlapped) or none of the classes (Massart et al. 1988; Vandeginste et al. 1998). Class modelling is the only possible choice when the focus is on a single class (Olivieri 2017).

The most commonly applied discriminant methods in the field of spectroscopy are linear discriminant analysis (LDA) or quadratic discriminant analysis (QDA), usually with prior dimension reduction by PCA, K nearest neighbours (KNN) and discriminant analysis by regression, usually PLS Discriminant analysis (PLS-DA). The most commonly used modeling methods are SIMCA (Soft Independent Modelling of Class Analogy) and artificial neural networks (Vigneau et al., 2000; Naes et al., 2002).

Discriminant analysis by regression, usually uses PLS (usually known in this application as PLS1) or the algorithm PLS-DA (usually known as PLS2), being this last

With just two groups PLSR can be used for classification by defining a dummy 0/1 variable representing class membership uses a training set in the usual way for the development of the predictive model that can be subsequently applied to the classification of new unknown samples (Naes et al. 2002; Heise and Winzen 2004; Kramer et al. 2004). According to Naes et al. (2002), the choice of these values for the dummy variables (0/1) is arbitrary, and other pairs of values can also be considered.

With more than two classes, the multivariate (that is to say multivariate y) version of PLS, PLS2, is needed. The resulting algorithm, usually referred to as PLS-DA, is one of the most widely used qualitative approaches to manage spectral data, and in particular is very popular in NIRS food applications. Although it can be used to discriminate between 2 or more classes, it is often recommended to discriminate at one time between just two classes. In the case of discriminating between more than three the idea is to perform the discrimination in cascade between pairs of classes subsequently. When PLS-DA is

applied, discrete variables of group membership or non-membership are used, and the independent variable becomes a vector of classes, so that each sample has a value of 1 in the class to which it belongs and a value of 0 for the rest of the classes (Naes et al. 2002; Heise and Winzen 2004). For the prediction of unknown samples, a value of the discriminant variable is obtained for each of the established classes; thus, a value of 1 in one of the categories would represent a perfect assignment to it, while a value of 0 would indicate non-membership of that class. In practice the predictions are not 0 or 1 but lie on a continuous scale, so that a threshold, of 0.5 for example, may be used for the assignment of class membership (Figure 7).

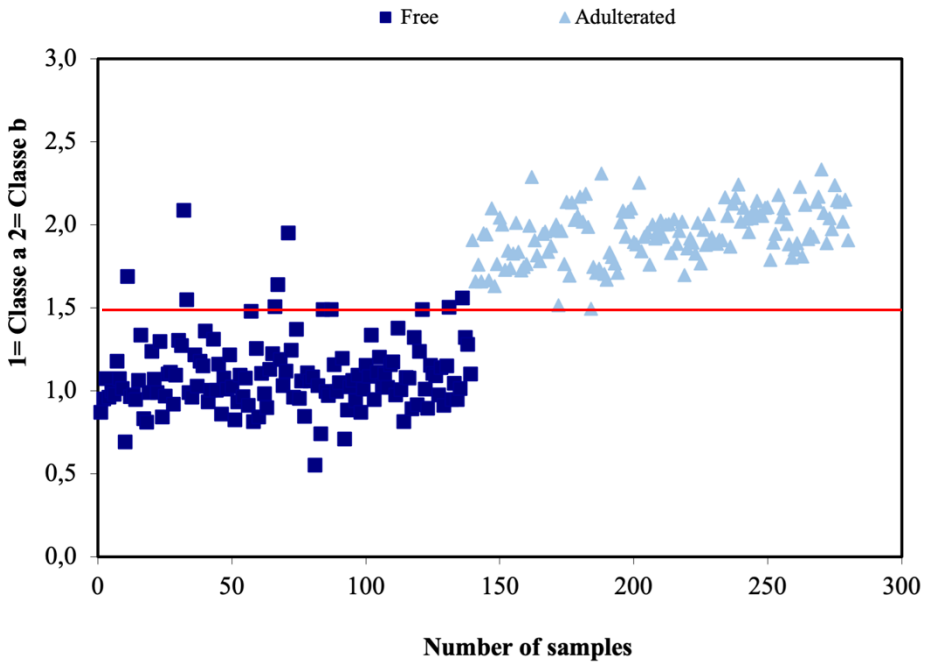

Figure 7. Plot representing the discrimination of the model between adulterated and non adulterated feeds.

The criteria considered for the classification of unknown samples could be diverse. Naes et al. (2002) and Kramer et al. (2004) indicated that the unknown samples should be assigned to the group in which they present a higher predicted value of the discriminating variable, while other authors point out that the mean value of the discriminating variables can be used as the discriminatory limit (McElhinney et al. 1999). Downey (2000) explains that the use of the mean value would be applicable if the classes were balanced in terms of number of samples, but recommends calculating the discriminatory limit between classes by taking into account the standard deviation of the predicted values for each class. Another similar approach, described by Murray et al. (2001), would be to consider an uncertainty factor, that is to consider the value of the minimum difference for classifying samples in a given category as the product of the model fitting error by an uncertainty factor. Thus, a sample is classified uncertainly in a class when the difference between the value given to that sample in its class and the value given to it in the other classes is lower than the value of the minimum difference. Another option is to use ROC (receiver operating characteristic) curves to establish the optimal threshold value. Thus, in a two-axis plot is represented the true positive rate (y-axis) and the false positive rate

(x-axis), the experimental outcomes are connected by a line that constitutes the ROC curve (Olivieri 2017).

The size of the collectives that make up each class is a key aspect that affects the discrimination results obtained. Downey (2000) points out that PLS-DA analysis works best when the established categories present similar structures, dispersion and covariances; otherwise, the author notes that it is better to use other qualitative approaches for classification. Naes et al. (2002) indicate that when the existing classes are unbalanced in terms of number of samples, the separation limit between classes will be displaced, so that the model will classify more samples in the best represented group.

Qualitive models, in general, must be validated using an external independent set to evaluate the real performance of the model. In this case, this step is even more important than in quantitative approaches. For the evaluation of the performance of qualitative models, the most used statistics are the percentage of correctly-classified samples (ratio of correctly-classified samples to total samples), the percentage of correctly-classified samples in each class, the sensitivity or ratio of samples classified positive to total number of positive samples, the specificity or ratio of samples classified negative to total number of negative samples, the number of false positives (total number of samples not belonging to a class classified as in this class by the model) and number of false negatives (total number of samples belonging to a class that are not classified in the class by the model) (Pérez-Marín et al. 2006).

8. Why does it work?

When a multivariate calibration has been developed, it is important to make some effort to try to understand why it works. If the calibration algorithm produces a prediction equation linear in the variables measured by the instrument, as does PLS for example, then examining the coefficients in this equation can give some idea of which of these variables make large contributions to the prediction (Martens and Næs 1989). In the case of NIR spectroscopy, the coefficients can be plotted against wavelength to give a curve resembling a spectrum. Given the complexity of NIR spectra, this is very unlikely to resemble closely the spectrum of the analyte being predicted, but hopefully it will show some features that can be related to this analyte. Tables of wavelength assignments that can help with this interpretation are available in the literature (Shenk et al. 2001; Williams and Norris. 2001). This is not an exact science, partly because the same absorption band may have multiple possible sources in the samples being measured, but at least a tentative interpretation is usually possible. If you cannot see any evidence of a chemical basis for the success of the calibration, it is a cause for concern.

One situation in which interpretation may prove difficult is when the instrument is measuring the analyte of interest by using the fact that it is correlated with some other property of the samples. For example, NIR calibrations for mycotoxins in cereals have been developed by exploiting the fact that NIR, which cannot measure anything that is present in such tiny concentrations, can measure the damage caused to the sample by the mould that produces the toxin. This may correlate well with the amount of mycotoxin present, thus allowing a calibration to be developed. The use of so-called secondary correlations like this one is dangerous. If the correlation is stable in the population of samples of interest, the calibration may be expected to work. If it is specific to the training and validation sets, perhaps because of the way they have been produced, it will probably

not work. In any case, this is one step further from a direct measurement, and this extra step may break down in the future. Investigating whether the calibration relies on a secondary correlation is one part of an interpretation.

When the calibration algorithm is nonlinear, and especially if, like an artificial neural network for example, it is essentially a black box, interpretation becomes much more difficult. It is possible to produce what might be called a pseudo coefficient vector by applying perturbations to the inputs to the algorithm, see for example Cui and Fearn (2018), and this may be a useful approach.

9. Calibration transfer

It is well known that NIRS instruments, even those of the same model and commercial brand, often present differences or deviations in their response, despite the fact that in recent years manufacturers have been making an increasingly important effort to standardize the response and to reduce these differences between different instrument units. Therefore, one of the most important aspects for the development and implementation of NIRS technology, at industry and laboratory level, is the possibility of using the models or databases obtained in a given NIRS instrument in a different NIRS instrument, since it is not possible to reuse them directly. In addition, with the development of a new generation of portable and on-line instrumentation, instrument cloning has become more interesting, since it is necessary for cost and efficiency reasons not to have to start from scratch every time when a new instrument is acquired, but rather to take advantage of reusing the information already existing in the previous equipment.

Standardization of NIRS instruments is therefore a critical process. Since the objective of cloning is to obtain information regarding the difference between instruments, all other sources of variation must be avoided or minimized to achieve a successful process (Fearn, 2001).

Differences between NIRS instruments usually result in two types of effects. One refers to shifts of the absorption peaks of the spectra of the same samples along the axis that represents the measurement wavelengths. The second one corresponds to differences in the absorbance levels, i.e., there is a shift on the y-axis of the spectrum collected in one device with respect to the one collected in another. Moreover, this second effect can be additive, multiplicative, a combination of both or even have a more complex form (Fearn, 2001).

There are several methods for cloning NIR instruments. Some of these standardization methods are designed to correct for only one of these two problems, while others can correct for both effects. Improvements in instrumentation have in many cases eliminated the problem of wavelength scale variations: in some instruments the use of internal references allows wavelength alignment correction (Brimmer and DeThomas, 1992). However, the variation in absorbance levels persists between instruments, and it is more difficult for instrument manufacturers to standardize their instruments to eliminate this problem (Fearn, 2001).

The objective behind cloning NIRS instruments is to share predictive models or databases developed in one instrument with others, obtaining the same or a very similar answer in different instruments. The various chemometric approaches available for the transfer of

prediction models between different instruments can be classified into three main groups (Bouveresse and Massart, 1996):

- Development of robust calibrations that can be transferred without the need for standardization. This is not realistic most of the time.
- Adjustment of the existing calibration to work properly on the new equipment. This usually refers to a slope and/or bias adjustment, or adjustment of the coefficients and wavelengths used in the calibration.
- Adjustment of the spectra collected on an instrument, usually called a satellite instrument, to make them as similar as possible to those of the instrument, often called the master, on which the calibrations or database to be transferred were developed. Initially, this strategy was used to make the satellite instrument equal to the master, making necessary to modify the spectra recorded in the new instrument always before applying the calibrations, and enlarging only the models in the master instrument. But later on, it was predominant to transfer the database of the old instrument to the new one, and then develop new calibrations in the new instrument, enabling the enlarging of calibrations with new samples analysed in the new instrument. For the standardization of NIRS instruments by spectral adjustment, there are three main methods in the related literature: the Shenk and Westerhaus (1989) algorithm, DS ("direct standardisation") and PSD ("piecewise direct standardisation").

These spectral adjustment methods, usually referred to as direct standardization, generally use a set of samples analyzed in the two instruments to be cloned to adjust the response of one instrument to make it spectrally similar to another. One of the advantages of the application of these procedures is that it is not necessary to have the reference data of the cloning samples, unlike with other standardization methods. However, in order to achieve satisfactory results, it is crucial that the same samples are analyzed under identical conditions in each of the instruments to be cloned, so that the differences found between spectra can be attributed to instrumental differences and not to changes in the cloning samples. This is relatively easy to achieve when the instruments are located in the same laboratory and sealed samples or inert reference materials are used, but less so when cloning is performed remotely and using unsealed samples (Pérez-Marín et al, 2006b).

Shenk and Westerhaus (1989, 1991) proposed the first NIRS instrument standardization algorithm through spectral transfer. This algorithm performs a univariate correction of the full spectrum in two stages. It performs a wavelength correction followed by a correction of the spectral intensity or photometric response correction of the satellite instrument. Both adjustments are stored in a standardization file, which is subsequently used for transferring spectra and models between different instruments. A description of the mathematical process on which this algorithm is based can be found in Bouveresse et al. (1994).

The other key aspect of the cloning process, together with the choice of the method, is the selection of the sample set for cloning, for which two fundamental requirements must be considered: their stability and representativeness. This means that the cloning samples must be physically and chemically stable, since otherwise the differences found between the spectra collected in the equipment to be cloned would not be exclusively instrumental in nature, but also due to changes in these samples. In addition, the cloning samples must be representative of the calibration samples, being in the same range of optical densities.

As pointed out by Dardenne and Biston (1990), cloning can be considered as a calibration that estimates optical densities as a function of other optical densities.

Although with the use of spectral adjustment algorithms, which do not require the reference data of the cloning samples, it is possible to use artificial samples or generic references for instrumental standardization, their application for the transfer of calibrations of agrifood products is usually not satisfactory. This is probably due to the different ranges of optical densities covered in such a case by the calibration and cloning collectives, and therefore to the lack of representativeness (Fearn, 2001).

Another option is the use of sealed capsules of real samples. However, the use of sealed capsules is not always possible for reasons of availability and cost, and sealing does not always ensure the stability of the sample, since depending on the nature of the product, physical and chemical changes may occur to a greater or lesser degree. The alternative would be to use real unsealed samples, analogous to the calibration samples, as a cloning set, but it is important to design a protocol that ensures that the piece of sample seen by both instruments is equivalent. This problem is exacerbated when we need to standardise instruments that are not located in the same laboratory. Pérez-Marín et al. (2006b) developed for first time a methodology for cloning NIR instruments remotely and transfer calibrations of intact feeds (pellets, extruded,etc.) with excellent results (Figure 8). In addition an example of the application of this methodology for cloning a benchtop and a portable instrument for the analysis of fresh forages can be found in Pérez-Marín et al. (2022).

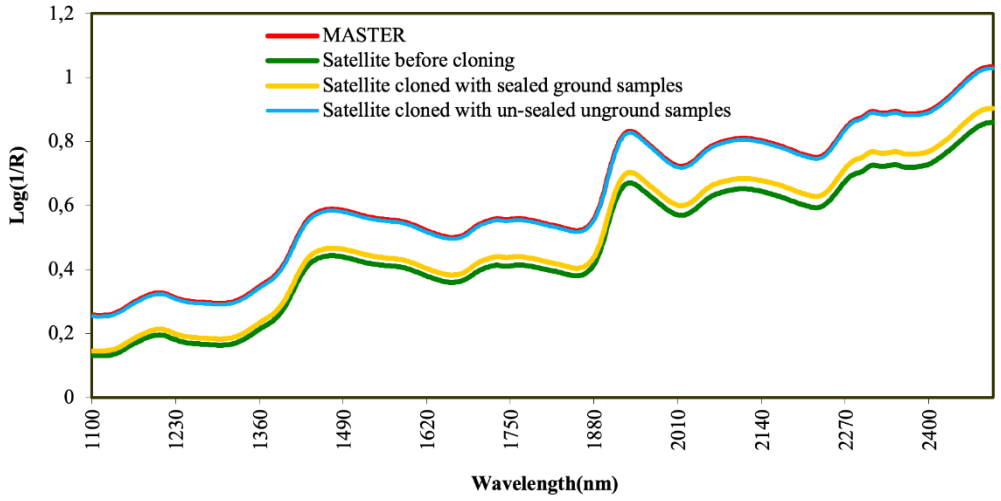

Figure 8. Effect of cloning in two benchtop instruments using sealed and un-sealed samples.

Another important aspect to consider in the standardization process is the number of cloning samples selected. It must be taken into account that the cloning samples must provide sufficient information regarding instrumental differences. Thus, the number of cloning samples to be used will be mainly conditioned by two factors: the complexity of the instrumental differences and the type of standardization method used (Bouveresse and Massart 1996). Ideally, it would be desirable to use only one cloning sample

representative of the calibration set (for example the sample closest to the center of the population). However, this is only possible for standardization of NIRS instruments that do not require wavelength adjustment, i.e., correction on the abscissa axis. In this case, only a photometric correction is performed between the instruments. When cloning requires wavelength correction in addition to the absorbance scale adjustment, several samples (usually a minimum of 3, although 10 or more are recommended) are required for the standardisation.

In general, the results of standardisation can be evaluated by means of two tests. First, to test the spectral repeatability and adjustment between the cloned instrument, the RMS(c) statistic can be used. This statistic is the averaged root mean square of differences corrected for the bias between spectra obtained in two instruments or between subsamples scanned on the same instrument. The second test is used to evaluate equation or database transfer between the two instruments, using a validation set analysed in both instruments. The procedure uses the statistic SED (standard error of differences) to compare the predicted values for the validation set between the master and satellite instruments, before and after the standardization procedure (Pérez-Marín et al. 2006b).

10. Monitoring and maintenance of calibrations

When a calibration or a qualitative model is ready for routine use, we can obtain predictions of unknown samples. Together with the predictions, usually we have the values of some statistics that allow us to know if the prediction is or not reliable. These statistic can be of two types: those that measure the spectral distances between the predicted samples and the training ones, and those statistics that evaluate the predictions themselves. For the first type, the most popular are the Mahalanobis distances (H) - both the global Mahalanobis distance (GH) and the neighbor Mahalanobis distance (NH) - or the Hotelling T^2. These statistics allow us to know if the predicted sample is close to the training samples, thereby avoiding extrapolation (Shenk et al. 2001). For the second type, the Student T of the chemical residuals is used. In this case, as the reference value is not available, the prediction may be compared to the mean value of the parameter in the calibration, to evaluate if there prediction is out of the calibration range.

If the samples predicted are considered outliers, for example those samples with GH values higher than 3 or T values higher than 2.5, the spectral analysis should be repeated if possible, ideally with a reference measurement as well. If the distances and prediction are confirmed, the sample can be considered as not represented in the calibration, so this type of samples are the candidates to enlarge the calibrations. Thus, one reason why the performance of a calibration may become worse is that it is now seeing samples of a type that were not represented in the calibration exercise, perhaps because a supplier has changed or because all the training samples came from one year's harvest and the instrument is now seeing samples from a new harvest. In such a case, the calibration may need more than a simple correction in order to learn to cope with this extra variability. Examples of the previously unseen type of sample will need to be added to the training set, and the calibration reworked and revalidated. Usually, one recalibration per year is recommended.

The performance of calibrations can drift, e.g. because something in the instrument or the measurement setup changes slowly over time, or undergo a sudden change, e.g. because

the instrument has a component changed or because the raw materials being measured come from a new supplier and are different to anything in the calibration set.

To protect against either of these problems it is desirable to monitor the instrument by comparing its predictions against measurements using the reference method that it was calibrated against. As a routine monitoring procedure, the predictions on a small number of samples could be compared with reference measurements on a regular basis, say once a week. A good way to treat the resulting data would be to plot the average difference between instrument and reference on a control chart with suitable warning and action limits (Montgomery 2005). Such a chart provides a clear visualisation of the instrument's performance, and an indication of when some corrective action may be needed. Something to avoid at all costs is the making of regular small adjustments to the calibration when the differences between instrumental and reference measurements are within the normal variation expected for the system, and the use of a control chart will prevent this. If the control chart does flag a problem, the first step is to check whether there are any obvious changes to the measurement system, for example a viewing window that has become contaminated, as well as running any diagnostics available on the instrument. Remember to include the reference measurement procedure in these checks: maybe the discrepancy comes from a change in the way this was conducted. If no cause can be identified and corrected, it may be necessary to make an adjustment to the calibration. This will normally require a larger number of samples to be measured by both instrumental and reference methods. These samples should cover the full range of the calibration. The number needed will be dependent on the instrumentation involved, but for NIR spectroscopy around 10-20 should be sufficient. Often a scatter plot of reference *versus* predicted will reveal that there is still a good correlation between the two sets of results but that either the intercept or slope of the line has changed. Then either a bias or skew and bias correction (Fearn 2001) may fix the problem. In the case of a known disturbance to the system, such as a lamp change in the instrument, it would be wise to proceed straight to this more extensive check without waiting for the control chart to detect any change.

REFERENCES

Andersen C.M., Bro R. (2010). Variable selection in regression – a tutorial. J. Chemometrics 24: 728-737.

Bocklitz T., Walter A., Hartmann K., Rösch P., Popp J. (2011). How to preprocess Raman spectra for reliable and stable models? Anal. Chim. Acta 704: 47-56.

Bouveresse E., Massart D.L. (1996). Standardisation of near-infrared spectrometric instruments. A review. Vibrational Spectroscopy 11: 3-15.

Bouveresse E., Massart D.L., Dardenne P. (1994). Calibration transfer across near-infrared spectrometric instruments using Shenk's algorithm: effects of different standardisation samples. Anal. Chim. Acta 297: 405-416.

Brimmer P.J., De Thomas F.A. (1992). Stability of NIR instrumentation: Control of critical parameters affecting analytical precision and transferability of NIR calibrations. In: Making Light Work: Advances in Near Infrared Spectroscopy, I. Murray and I.A. Cowe (ed.), VCH, Weinheim, p. 35.

Cui C., Fearn T. (2018). Modern practical convolutional neural networs for multivariate regression: applications to NIR calibration. Chemometrics and Intelligent Laboratory Systems 182: 9-20. http://doi.org/10.1016/j.chemolab.2018.07.008

Dardenne P., Biston R. (1990). Standardization procedure and NIR instrument network. In: Proceedings of the 3rd International Conference on Near Infrared Spectroscopy. Biston R. and Bartiaux-Thill N. (ed.). Agricultural Research Centre, Gembloux, Belgique, p. 655.

Downey G. (1996). Review. Qualitative analysis in near-infrared region. Analyst 119: 2367-2375.

Downey G. (2000). Discriminant PLS—questions and answers from a listserver. NIR News 11(1): 9-12.

Fearn T. (2001). Standardisation and calibration transfer for near infrared instruments: a review. J. Near Infrared Spectrosc. 9: 229-244.

Freedman D.A. (2009) Statistical Models: Theory and Practice. Cambridge University Press, Cambridge, UK.

Heise H.M., Winzen R. (2004). Chemometrics in Near-Infrared Spectroscopy. In: Near-Infrared Spectroscopy. Principles, Instruments, Applications. Siesler H.W., Ozaki Y., Kawata S. and Heise H.M. (ed.), Wiley-VCH.

Kohonen T., Schroeder M. R., Huang T.S. (2001). Self-Organizing Maps, 3rd edn. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Kramer R., Workman J.Jr., Reeves III J.B. (2004). Qualitative analysis. In: Near Infrared Spectroscopy in Agriculture. Roberts C.A., Workman J.Jr and Reeves III J.B. (ed.). Agronomy Serie No. 44. ASA, CSSA and SSSA Publishers, Madison, Wisconsin, USA, pp. 175-206.

Martens H., Næs T. (1989). Multivariate Calibration Wiley, Chichester, UK.

Massart D.L., Vandeginste B.G.M., Deming S.M., Michotte Y., Kaufman L. (1988). Chemometrics: a textbook. Data handling in science and technology 2, Elsevier Science, Amsterdam, The Netherlands.

McElhinney J., Downey G. (1999). Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. J. Near Infrared Spectrosc. 7: 145-154.

Montgomery D.C. (2005). Introduction to Statistical Quality Control, 5th edn. Wiley

Murray I., Aucott L.S., Pike H.I. (2001). Use of discriminant analysis on visible and near infrared reflectance spectra to detect adulteration of fishmeal with meat and bone meal. J. Near Infrared Spectrosc. 9: 297–311.

Naes T., Isaksson T., Fearn T., Davies A. (2002). A user-friendly guide to Multivariate Calibration and Classification. NIR Publications, Chichester, UK.

Olivieri P. (2017). Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues – A tutorial. Analytica Chimica Acta 982: 9-19.

Pérez-Marín D., Garrido-Varo A., Guerrero J.E. (2006). Optimization of Discriminant Partial Least Squares Regression Models for the Detection of Animal By-Product Meals in Compound Feedingstuffs by Near-Infrared Spectroscopy. Applied Spectroscopy. doi: 10.1366/000370206779321427

Pérez-Marín D., Garrido-Varo A., Guerrero J.E. (2006b). Remote NIRS instrument cloning and transfer of calibrations to predict ingredient percentages in unground compound feedingstuffs. J. Near Infrared Spectrosc. 14: 81-91.

Pérez-Marín D., Garrido-Varo A., Guerrero J.E. (2007). Review. Non-linear regression methods in NIRS quantitative analysis, TALANTA 72: 28-42. https://doi.org/10.1016/j.talanta.2006.10.036.

Pérez-Marín D., De la Roza B. Entrenas J.A., Garrido M. Garrido A. (2022.) Transferring a large data library of fresh total mixed rations from a benchtop to 2 portable near-infrared spectrometers for on-farm real-time decisions. J. Dairy Sci. 105, 3, 1-13. https://doi.org/10.3168/jds.2021-21032

Shenk J.S. (1992). Networking and calibration transfer. In: Making Light Work: Advances in Near Infrared Spectroscopy. Murray I. and Cowe I.A. (ed.), VCH. London, UK, p. 223.

Shenk J.S., Westerhaus M.O. (1989). U.S. Pat., 4866644, Sept. 12 1989.

Shenk J.S., Westerhaus M.O. (1991). New standardization and calibration procedure for NIRS analytical systems. Crop Sci. 31:1694-1696.

Shenk J.S., Workman J.J., Westerhaus M.O. (2001). Application of NIR Spectroscopy to Agricultural Products. In: Handbook of Near Infrared Analysis, 2nd edn. Burns D.A. and Ciurczak E.W. (ed.), Practical Spectroscopy Series, Vol. 27. Marcel Dekker, USA.

Vandeginste B.G.M., Massart D.L., Buydens L.M.C., De Jong S., Lewi P.J., Smeyers-Verbeke J. (1998). Handbook of Chemometrics and Qualimetrics: Part B. Elsevier Science B.V. Amsterdam, The Netherlands, p. 649.

Vega-Castellote M., Sánchez M.-T., Wold J.P., Afseth N.K., Pérez-Marín D. (2023). Near infrared light penetration in watermelon related to internal quality evaluation. Postharvest Biology and Technology 204:112477. https://doi.org/10.1016/j.postharvbio.2023.112477

Vigneau E. Qannari El M., Devaux M.F. (2000). Chapitre 12: Méthodes prédictives. In: La spectroscopie infrarouge et ses applications analytiques. Bertrand D. and Dufour E. (ed.), Collection Sciences&Techniques Agroalimentaries, Editions TEC&DOC, París, Francia, p. 295.

Williams P.C., Norris K. (2001). Near-Infrared Technology in the Agricultural and Food Industries, 2nd edn. American Association of Cereal Chemists, Inc. St. Paul, Minnesota, USA.

Walsh J., Neupane A., Koirala A., Li M., Anderson N. (2023). Review: The evolution of chemometrics coupled with near infeared spectroscopy for fruit qualty evaluation. II. The rise of convolutional neural networks. J. Near Infrared Spectrosc. 31: 109-125. doi: 10.1177/09670335231173140.