Independence rules (or Rules for independence)

A.M.C. Davies
Norwich Near Infrared Consultancy, 75 Intwood Road, Cringleford, Norwich NR4 6AA, UK

When we have developed a multivariate calibration it is essential that the performance of the calibration is tested with a set of samples which are typical and independent. This set has several names but I prefer "Validation set". If you are satisfied with the predictions on these samples then you will claim that you have "validated" the calibration, which suggests that you expect it to continue to give useful results in the future.

We all know what we mean by "independent"; don't we? Of course we do! Wars of Independence have been and are being fought for it. But do we? Have a look at a few examples taken from papers submitted to the *Journal of Near Infrared Spectroscopy*. (I have changed the commodities and analytes to avoid identification of any authors, none of the examples involved olive oil).

Example 1: The object was to test the hypothesis that NIR data could distinguish olive oils from four different regions. One bottle of oil was obtained from each area. 45 spectra of each sample were recorded. 25 spectra from each sample were used for calibration and 20 spectra from each sample were used for validation. It was confirmed that all samples came from the same four bottles!

Example 2: The object was to obtain a calibration for "percent authentic olive" from samples which had been adulterated with a mixture of oils from other sources. The mixed adulterants were added in percentage steps to create a calibration set of 22 samples and a test set of 21 samples.

Example 3: The object was to calibrate for a particular fatty acid in olive oil samples. 180 samples were available. A "match" program was used to select 36 samples for the validation set and the rest were put into the calibration set. [The match program finds the most variable samples in the data set. When used in this way it makes a link between the two sets; see Reference 1 for the suggested safe use of these programs].

So what is meant by "independent" when applied in chemometrics? The *Concise Oxford Dictionary* gives several meanings, the more relevant being:

- (often followed by of) not depending on authority or control
- self-governing.
- not depending on another person for one's opinion or livelihood
- (of income or resources) making it unnecessary to earn one's living.
- unwilling to be under an obligation to others.
- not depending on something else for its validity, efficiency, value, etc. (independent proof).

Although these give the general idea it is perhaps helpful to give a specific definition:

An independent validation set is a set of samples which could not have influenced a calibration.

Notice that this is rather different from saying samples that were not used in the calibration. If the original spectral database contained duplicate spectra it should be obvious (well I hope it is) that you should not have one replicate in the calibration set and one in the validation set. However, if the person collecting the samples actually collected *n* samples and split each sample into half and provided the experimenter with two sets of *n* samples these would not be independent of each other. The experimenter would not know this but the collector would. What about the situation when there are several collectors? They might collect samples from the same sources so that the sample population would contain some samples that only differed by the sampling error but no one would realise that it would be possible to select sets of samples (calibration and validation) which were not independent.

How should the calibrator proceed? Suppose you lived on an island, where olives grew in abundance, that was well separated from any other land and you wanted to develop an NIR method of analysis for olive oil. You could easily assemble a set of samples for calibration but what about the validation set? It has long been the tradition in NIR analysis that you should obtain a large set of samples and divide them into two sets [three if you are using partial least squares (PLS)[2]; using one as the calibration set and the other as the validation set, but is this a true *independent* set? I think the answer is "no", but because it has been used so frequently many people fail to understand the requirement for independence and then make the much worse errors in my examples above.

So if you are the inhabitant of "Olive Island" how do you obtain an independent sample set? You could divide the Island into half and use samples from one half to make the calibration and the other the validation, but then the calibration would be validated for only half of the island. The obvious answer is to make friends with people on the mainland and obtain a set of samples from them. [There is of course confusion with the language because **you** have become dependent on your neighbours but the set can still be independent!]. If this set gives a satisfactory validation then this is an excellent result and you can sell your calibration to your neighbours! However, it is more than likely that the results will be disappointing. Why? Because in addi-

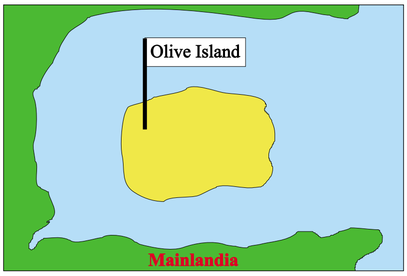

Figure 1. Olive Island.

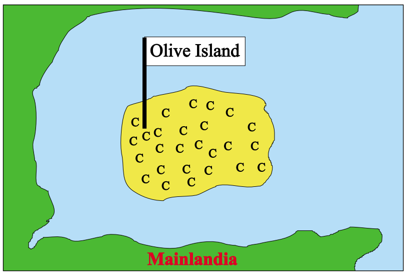

Figure 2. Calibration set for oil content of olives from Olive Island.

Figure 3. "The calibration is OK but where can I get a validation set?"

tion to being independent we also require samples to be "typical" and it is unlikely that olives grown in an ocean environment would be the same as those grown in a continental climate. [Something very much like this actually happened at the dawn of the current phase of NIR analysis. Wheat calibrations were developed in the USA but the instrument manufacturers were disappointed when they tried to sell instruments with these calibrations for wheat grown in Europe].

So where do independent calibrations come from? The answer is "next year". These samples are independent and hopefully they will be typical and with each passing year you can increase the size of the calibration set and test calibrations with a new independent set. Current wheat calibrations are based on thousands of samples!

Of course you cannot really wait for next year so the best you can do is to collect a large set of samples from your island, say 200, and split it into two. This is usually done by sorting the samples according to the value of the analyte then placing alternate samples into one or other sets. This gives you two sets that will have very similar characteristics. The results from the validation will be over optimistic but, except in the case of extremely variable commodities, not too seriously in error. Then if the results are promising, next year you will have a calibration based on 200 samples (from last year) and you can validate it with samples from this year.

This imaginary island does not cover all possibilities for sets which may not be of agricultural materials but I hope it can help people to see what they need to achieve. Tom Fearn's recommendation is to try to think of all sources of variation in the samples and make sure that they are **all** included in the validation set. His other words of wisdom are to realise that the ***ideal*** validation set is a random selection of all the samples you will **ever** want your calibration to predict! This is impossible but it helps you to see the disadvantage of splitting one data set into two.

I mentioned, in passing, that when you are doing PLS you need three sets of data. These are a calibration set, a number of factors selection set and the validation set. These requirements were explained in an earlier article[2] and in more detail in our book.[3] I should also mention that sometime ago I wrote an article about the dangers of using sample selection programs and the error of using the non-selected samples as the validation set.[1] This is a frequent error which would be avoided if calibration developers spent a little more time thinking about Olive Island and the "Fearn recommendations". I leave you to use these tests to see what was wrong with the "validation" samples in my three examples. There is a chapter on validation in our book.[3]

Acknowledgements

If anyone does recognise (or thinks that they recognise) their work in my examples, I ask for forgiveness in the interests of education. I am grateful to Professor Tom Fearn for his thoughts, but the responsibility for any errors is entirely mine.

References

1. A.M.C. Davies, *Spectrosc. Europe* **8(4),** 27 (1996).
2. www.spectroscopyeurope.com/TD_10_2.pdf
3. T. Næs, T. Isaksson, T. Fearn, and T. Davies, *A User-Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, Chichester (2002).

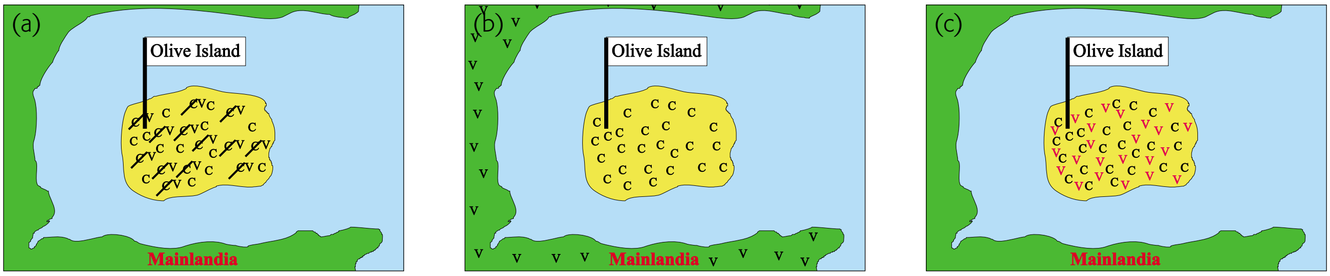

Figure 4. Possible validation sets: (a) use half the calibration set; (b) get them from Mainlandia; (c) wait until next year. C, calibration sample from year one; V, validation sample from year two.