TECHNICAL NOTE TN2022_4 – WHAT TO CONSIDER WHEN BUILDING AND IMPLEMENTING A REGRESSION MODEL

Introduction

Hyperspectral imaging provides multivariate data, which in turn can be used to build quantitative models. Since version 1.1.0.34 of Insight, a regression algorithm is available (PLS). The aim of this TN is to highlight what to take into account when building such models.

PLS = PARTIAL LEAST SQUARE

Article

When building a quantitative model, sample preparation is crucial. The following parameters need to be thoroughly thought when preparing the training samples:

- The quantitative range of interest
- The quantity of training samples, and their distribution
- The accuracy with which the variable of interest is known

Once the model is built, its implementation and usage also require other parameters to be considered than just software related. Those parameters will also be discussed in this TN, including:

- Camera SNR
- Illumination stability
- Relevant pre-processing

We also encourage the reader to go through the TN about camera transferability. We believe it would provide a wider scope to the user.

- **The quantitative range and the distribution of the training samples**

  Building a regression model requires a rather large amount of training samples. For these samples, the variable of interest (e.g. moisture, fat, etc...) needs to encompass the range of interest (interpolation rather than extrapolation). If for example one wants to measure the moisture content of wooden chips, and the moisture of the samples in general ranges between 15 and 30%, it is important that the training samples used to build the model have a moisture level between 15 and 30%, at least. On the opposite, if the moisture level of all the training samples is above 50%, a model will be built, but with a very low predictive performance between 15 and 30%. A Regression model has much better interpolative performance than extrapolative one.

  Also, the distribution of the variable related to the training sample is important. We recommend to have it evenly distributed over the full range of interest, and above (as a margin for outliers). If we go back to the previous example about the moisture of wooden chips, the panel of the training samples should represent the full expected range and beyond, with moisture level ranging between ca. 10 and 40%, with constant steps. For instance, 5 samples at 10%, 5 at 12% 5 at 14%, ... until 5 samples at 40%. If the distribution is not even, the accuracy of the resulting model may also be uneven.

Over-calibration is another important aspect in modelling and it relates to the choice of the training samples. If in the above example most of the samples (let say 20) are for instance with a moisture level of 18%, and very little for the other values (only 1 or 2 samples), the model will tend to be biased toward 18%.

- **The accuracy of the training samples**

  The accuracy of the regression model also directly depends on the accuracy of the training samples, meaning the accuracy at which the variable of interest is known. Going back to the wooden chips application, if the moisture of the training sample is known with an accuracy of ±5%, the resulting model will not be as accurate as if the accuracy of the training samples would be known at ±1%. Building a model on accurate samples will provide a model with better performance.

As we previously highlighted, the choice and the quality of the training samples is crucial when building a quantitative model. However, once the model is built, more factors still need to be considered for a successful implementation.

- **A good SNR**

  A PLS model scrutinizes tiny differences within spectra to make quantitative estimates. A camera providing spectra with the highest SNR will ease the extraction of the information, making the model more accurate. For more information about SNR, we encourage the reader to go thorough the TN “SNR”.

- **A stable illumination**

  The illumination is also an important parameter to take into account. Besides its spectral emission, which should be as even as possible over the spectral range of interest, its temporal stability is crucial. As mentioned previously, quantitative models rely on small spectral differences, and if the illumination fluctuates, these induced variations could be mis-interpreted by the algorithm. We recommend the reader to go though the TN “Illumination”.

- **A relevant pre-processing**

  Some pre-processors highlight features in spectra for more efficient information extraction. In a case of quantitative analysis, those spectral feature are typically bound with absorption or reflection peaks. Those are characterized by two assets: i) their position and ii) depth. Both properties are underscored by derivative and SNV pre-processors. Those should be considered when building a PLS model.

  We recommend the reader to go through the TN “Pre-processors”.

Finally, we would like to highlight that some quantitative applications, do not necessarily required regression models, even though they sound like. They could be based first on a hard sorting (with classes), and then on a pixel count. Let’s take an example related to meat, especially looking at a T-bone steak. Such a steak typically includes a large part of red meat, a bone and some white fat around on some edges. One could first make a qualitative analysis, estimating ca. 50 000 pixels as “red meat”, 15 000 as “bones” and 20 000 as “fat”. A fat level quantification could be 20 000 / (50 000 + 15 000 + 20 000) = 0.235, which means 23.5%. In this approach, no regression model is actually needed. However, if someone else is looking at fat content in minced meat, or as intramuscular one within the red meat of the above mentioned T-bone steak, then a regression model is needed.

Since SPECIM Insight 1.1.0.34, regression model for quantitative analysis can be built. Besides, relevant pre-processors are also available.

Disclaimer

This technical note is prepared by SPECIM, Spectral Imaging Ltd. and for generic guidance only. We keep all the rights to modify the content.

Version history

Version	Date	Author	Comments
1.0	Oct. 4th 2022	MMA	