Model development using NeoSpectra Instrument

**Code:** PM – 001 / 1

**Version:** 2023. v1

**Author:** Marcal Plans

# Related documents

1. PM – 002/1 Model Implementation in the LabStore
2. PM – 003/1 User Manual and Guidelines for the Generalizer Module
3. PM – 004/1 Quick Guide: Spectral/Calibration Transfer
4. PM – 005/1 Model Transfer – Workflow
5. CL - 001 -1 Bias Calculation and correction - Template

**Contents**

[**Related documents 1**](#_heading=h.1ksv4uv)

[**Introduction 2**](#_heading=h.2jxsxqh)

[**Glossary 2**](#_heading=h.z337ya)

[**Sample preparation 2**](#_heading=h.3j2qqm3)

[**Instrument settings 3**](#_heading=h.1y810tw)

[**Samples used in calibration 3**](#_heading=h.4i7ojhp)

[**Background settings 4**](#_heading=h.2xcytpi)

[**Scanner use on the calibration step 5**](#_heading=h.1ci93xb)

[**Spectral transfer / Model Transfer 5**](#_heading=h.3whwml4)

[**Generalization - Development kit 5**](#_heading=h.2bn6wsx)

[**Validation samples 5**](#_heading=h.qsh70q)

[**Validation criteria 6**](#_heading=h.3as4poj)

[**Model implementation in the LabStore 6**](#_heading=h.1pxezwc)

[**Model maintenance and model update 6**](#_heading=h.49x2ik5)

# Introduction

This document captures all the information items to develop a model using the NeoSpectra Scanner. Since each technology has its own particularity, the present document attempts to include the specific elements for NeoSpectra in a single workflow.

For consistency and convenience, this document is linked to several documents that will be needed for the final implementation of the models in the LabStore or Portal.

# Glossary

*Labstore:* Website where the models can be commercialized like a marketplace.

*Portal:* Cloud based website where samples, models, user etc,.. are managed by Si-Ware’s accounts. Portal gives access to the Labstore and all data and information generated from the NeoSpectra instruments.

*Instrument:* It is the spectrometer; it can also be referenced as the unit or scanner. It is the hardware used to collect NIR spectra.

*Model:* It is the mathematical algorithm used to correlate some data to a parameter or to spectra.

*Samples:* These are the elements of the observations that are used to develop models. Samples are related to the matrix/product/material that is analyzed by the Scanner and it is intended to obtain information by using models.

*Wet Chemistry:* Reference analysis performed on the samples. Wet chemistry is related to the parameters of interest that will be correlated to the spectra by using models.

*Calibration:* it is the part of the process where the model is developed and trained. It is related to calibration samples, model calibration,…

*Validation:* it is the part of the process related to the performance of the model in real field/process conditions. It is related to validation samples, cross-validation,… It is often related to the prediction capability of the models associated to the error of prediction.

*Working range*: range where the model can be used during routing analysis.

# Sample preparation

This part is unique to the application to be developed. The instrument is capable of obtaining the spectra of the samples in several different ways. The developer will decide the best way to prepare the samples based on the following criteria:

1. *Reproducibility of the sample preparation*. To obtain a good model the sample preparation should allow for measurement of the sample spectra in a reproducible way.
2. *Sample processing and preparation time*. Although sometimes the best way to obtain the spectra of the sample is by extensive sample preparation, it may not be economically logical to do so . For example, the best models for protein in corn are obtained by drying the sample, grinding at constant particle size on a benchtop mill with a sieve screen of 1 mm and measuring the spectra of the flour, but in some cases, this sample preparation process may not be acceptable.
3. *Reducing the impact on the parameter of interest.* The sample preparation may alter the parameter of interest. Sample preparation should avoid having this impact. As an example, heavily grinding the sample to a low particle size may heat the sample, reducing the moisture content before it is analyzed by the spectrometer, this may have a detrimental impact if moisture content is the parameter of interest.

It is recommended for forages *as is*:

* Apply similar mechanical processes before the measurement on the spectrometer. Define the chopper and the mechanical settings.
* It is recommended that the sample particle size not be larger than **5 cm**. Keep the sample in a sealed bag and at a consistent temperature. **Do not freeze samples**.
* When using the **rotator and dish** accessory:
  + The sample will be placed on the dish covering at least **50% of the height of the dish**.
  + For forage, dry hay, and fresh grass type of samples some **pressure must be applied using the screw cap** on the dish to ensure that enough sample is in contact with the glass.
  + For grains and cereals, the screw cap on the dish is not needed.
* When using the **point and shoot method**:

Point and shoot is a type of measurement that does not use any accessory, the sample is located underneath the scanner and the spectra is collected with direct contact.

* + Ensure that the Scanner is perfectly flat to the surface of the sample.
  + Change the location of the scanner on the sample every measurement.

# Instrument settings

There are a lot of parameters that can be modified on a FT-NIR, some of them are related to the measurement itself (scan time and number of measurements per sample) and some of them are related to the algorithm used to transform the signal by using the Fourier Transform.

For most of the applications, the **default FT parameters are the best** parameters to use.

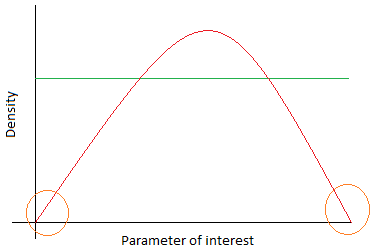
For point and shoot applications (using the saucer or no accessory), the best performance is obtained by **5 measurements per sample and 5 seconds with an interval of 10 seconds** (interval times can vary, but 5 seconds allows sufficient time to move the position of the Scanner in the sample).

Using the **rotator device with the dish**, the recommended settings are 12 **measurements of 5 seconds scanning time and 15 seconds interval. This allows the instrument to collect 6 times the entire area of the dish (180s total).**

# Samples used in calibration

There is no set number of samples that can be given to obtain a robust model. It is known that the number of samples is proportional to the number of sources of variation. NIR is highly dependent on the matrix effect. This is the effect coming from non-related compounds or constituents of the material that co-exist with the parameter/s of interest and they are crucial to obtain a robust model.

The distribution of the parameter of interest is important to balance the models and obtain robust models. In regression models, the **random/uniform distribution (green)** is the best distribution to balance the effect of the extreme samples. It is common to get a **Gaussian/normal distribution (red)** of the parameter of interest. However, this would have some negative effects on the regression model, where extreme samples will have a large effect on the model, and the average of the population would have a large weight on the model, making the models trend to predict the average of the population.



In general, duplicate samples are not good for the NIR models. Duplicate samples are samples that spectroscopically look similar and have similar reference values. The interest is to have samples with **similar levels of the parameter of the interest with different spectra features** (maximizes the matrix effect).

In summary:

* Random distribution of the parameter of interest is the preference for regression models.
* Samples with similar spectra should be removed.
* Sample size is proportional to the sources of the spectral variability. In general, at least 30-40 samples for sources of variability are needed. To simplify, we can extrapolate to **40 samples × Number\_of\_PLS factors** in the model (For example, a moisture model in corn needs 4 PLS Factors meaning at least 160 samples have to be included in the calibration data base ). There are several publications on common parameters in animal feed to use as a starting point for the sample size.

# Background settings

The lamp and sensor reach quickly to the operations temperature and the system has to be reference with a white tile every **15-20 minutes** in steady stage. At the begging of the working day, the recommendation is to take **backgrounds before each sample for at least 6 samples**. This will help to reach the steady stage, after that a background is required every **15 – 20 minutes**.

# Scanner use on the calibration step

It is desirable to use several scanners to collect different samples during the calibration step. Multiple scanners will provide variability between instrument to instrument during the spectra collection that will be integrated on the model using the regression models.

There is no need to collect all the samples on all the instruments. **The samples can be split and measured in a similar proportion.** It is recommended to use **4-5 instruments and measure 20% - 25% of the samples on each one randomly selected**. There is no issue collecting the same samples on different instruments or measuring more samples on one instrument than others.

# Spectral transfer / Model Transfer

Two specific documents are available to ensure the correct process of the spectral transfer also known as calibration transfer (*PM – 004/1 Quick Guide: Spectral/Calibration Transfer, PM – 005/1 Model Transfer – Workflow*). These two documents will guide you through the methodology to prepare the data to be added to the calibration model phase.

# Generalization - Development kit

To be able to generalize the model to all the production instruments, a generalizer algorithm will be applied using 12 - 15 scanners (different from the calibration scanners) on **10-15 samples that cover the range of the parameter of interest**.

This step can be done on site (sending the 15 scanners and running the 10-15 samples at ones) **or** sending the 10-15 samples to the development center in the country/region to measure them on the 15 scanners. Since the sample will be measured several times on several different instruments, it is mandatory to have a **large amount of sample to split the it into 10-15 representative parts** using the appropriate quartering technique, that way we will avoid overuse of the same part of the sample during the measurement of the spectra.

This data will be used to apply the Generalizer, the Generalization can be done locally using the software provided by Si-Ware or can be done by Si-Ware and generalized data will be returned to the developer for further modeling. Generalizer also can be applied on the developer site (*PM – 003/1 - User Manual and Guidelines for the Generalizer Module).* The previous document also goes over the **model calibration step using generalized data**.

# Validation samples

It is common to use a minimum of 10-15 samples to validate the models. These samples have to be completely independent of the calibration set of samples and should cover the working range of the parameter. A **random distribution of the parameter of interest** is the best distribution to be used for this purpose. Samples should be measured on the calibration instruments (4-5 scanners) in the same way that would be used in a **routine analysis** (it should be the same way that the spectra were measured for the calibration set of samples).

As in the Generalizer step, the same sample will be measured several times on several different instruments. Therefore, it is mandatory to have a **large amount of sample to split the sample in 4-5 representative parts** (depending on the number of scanners) using the appropriate quartering technique to avoid the overuse of the same part of the sample during the measurements, which may lead to a change on some constituents of the samples (the most common is moisture loss).

# Validation criteria

Goodness of fit of the model and the figures of merit are commonly used in NIR. For our purpose, it is required to analyze and provide the **R2CV, RMSECV, RMSEP, R2P, the analysis of the proportional bias and systematic bias on the Validation set of samples**. These statistics are obtained from the PLS-ToolBox or any chemometrics software. Use CL - 001 -1 Bias Calculation and correction – Template to validate the models.

# Model implementation in the LabStore

This is the last step in the process. It is also known as publishing the model. Models can be published in the LabStore for external use on a subscription basis or for internal publication in the Portal to be used for corporations. Both processes are similar, although for the LabStore additional steps are required which will need the Si-Ware support team to assist.

The steps to publish the models are explained in a specific document (*PM – 002/1 Model Implementation in the LabStore*).

# Model maintenance and model update

It is up to the developer to decide when it is time to update and maintain the model, but it is a good practice to establish regular checkups of the model performance. Depending on the material/product, a recommended frequency may be every 6 months whereas seasonal products may only be verified once a year. There are some materials/products that have a large environmental component, and it is a good practice to verify performance at the beginning, in the middle, and at the end of the session.

**Checkup of the model** should be done similar to the validation step by using a representative set of samples (**10 – 15 samples**) that **covers the working range** of the model.

The model maintenance **triggered by a verification fail** is usually due to two factors:

1. Predictions of a new samples set compared to the wet chemistry are out of the RMSEP. These samples can be used to enhance the model, and re-validate with a new set of samples.
2. Predictions of the new sample set compared to the wet chemistry provide unseen bias.

The maintenance process is exactly the same as the calibration process, without the generalization and the spectral transfer steps that are already included in the previous version of the model.

**An important point on this note**: models that have **transferred spectra** from a benchtop instrument have to have **a plan** for a periodic update of the model and **replacement of the transferred spectra** for native spectra. It is good practice to establish a strategy, usually 3-5 years, to fully replace the transferred spectra for native NeoSpectra spectra.