

NIR spectroscopy for cacao bean quality measurements

Claudia Beleites^{1,2}, Michael Glitschka¹,
Christoph Böttcher¹, and Andrea Krähmer¹

¹ Julius Kühn-Institut
Königin-Luise-Str. 19, Berlin, Germany
² Chemometric Consulting and Chemometrix GmbH
both Södeler Weg 19, Wölfersheim, Germany

Abstract We present experimental design strategies for developing predictive chemometric models based on NIR spectra of plant materials (here: *Cacao* beans) with a particular focus on two issues: Identifying important confounding factors and choosing a relevant subset of samples for reference analysis.

Keywords: NIR spectroscopy, design of experiments, nested design, hierarchical/clustered data, calibration, regression.

1 Introduction

Project CocoaChain studies cacao/cocoa quality along the processing chain. Setting up an analytical method based on NIR spectroscopy and chemometric data analysis for measuring various aspects of Cacao bean quality and using these NIR spectra to actually derive conclusions about the cacao beans on the first glance have similar sample requirements: Both need a suitable set of samples (specimen) of which NIR spectra are then measured. For the NIRS modeling, however additional reference information is needed, while for quality assessment the obtained NIRS method predicts these characteristics.

DOI: 10.58895/ksp/1000087509-2 erschienen in:

**OCM 2019 - 4th International Conference on Optical Characterization of Materials,
March 13th – 14th, 2019, Karlsruhe, Germany**

DOI: 10.58895/ksp/1000087509 | <https://www.ksp.kit.edu/site/books/m/10.58895/ksp/1000087509/>

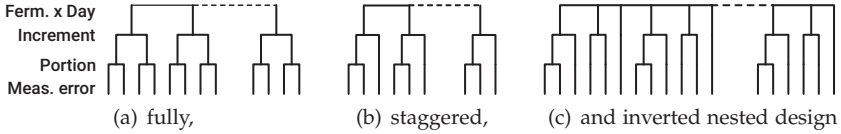


Figure 2.1: Designs of Experiments (DoE) for nested data structures.

Analytical procedures as well as biological systems and production processes often have deeply nested structures of influencing or confounding factors. Experimental effort is most efficiently spent if concentrated on the large and important confounders. At the same time, reference analyses are often expensive and/or time-consuming and in fact a bottleneck for the NIR calibration, emphasizing the need to spend the available experimental effort well.

In his seminal paper [1], Bainbridge presented three different designs of experiments (see fig. 2.1) that allow estimating contributions of nested sources of random error. The fully nested design is easiest to analyze, but it comes at the cost of the number of measurements/samples at the lowest level of the data structure exponentially growing with the number of factors (hierarchy levels) considered. In addition, the degrees of freedom are concentrated in this lowermost level of the data hierarchy, which is also the variance component that is easiest to estimate. In other words, this design is inefficient in its use of measurements for the higher variance components. In contrast, the staggered nested design leads to almost equally distributed degrees of freedom across the data hierarchy. Still, variance component further up are more difficult to estimate, so a design such as the inverted nested design that concentrates degrees of freedom rather on the top levels would be even better. However, at the time Bainbridge wrote his paper (1965) inverted nested design data could not be analyzed. With the nowadays readily available statistical software tools such as mixed models, it is now easily possible to employ such even further thinned out designs.

In this presentation, we explore the application of these designs for analytical method development, namely NIRS calibration and the corresponding reference analyses.

Table 2.1: Overview of the fermentations sampled in Peru 2018. Samples P1 – P4 are commercially imported cacao from the same cooperative where sampling took place in Quillabamba.

Fermentation ID	Region	Variety	Days sampled
KZug	Tingo Maria	CCN51	7
1	Quillabamba	Chuncho	7
2	Quillabamba	Chuncho	6
7	Quillabamba	Chuncho	1 (end)
8	Quillabamba	Chuncho	1 (end)
3	Tarapoto	Forastero CCN51	8
4	Tarapoto	Forastero CCN51	8
9	Tarapoto	Criollo ICS 95 + other	1 (end)
10	Tarapoto	Criollo ICS 95 + other	1 (end)
13	Piura	Cacao blanco + Cacao violeta	7
P1	Quillabamba	Chuncho	n
P2	Quillabamba	Chuncho	n
P3	Quillabamba	Chuncho	n
P4	Quillabamba	Chuncho	n

2 Methods and Material

2.1 Sampling

Samples of *Cacao* beans (seeds of *Theobroma Cacao*) were taken during fermentation: In 2017, a preliminary experiment was run in Ivochote/Peru where specimen were obtained before start of the fermentation (day 0) and then daily until the fermentation was stopped at day 4. In 2018, 10 fermentations in four different regions were sampled. These samples were augmented by an additional 4 samples from commercially imported cacao from four different bags (by Peru Puro, Frankfurt, Germany; two roasted and two unroasted). For an overview, see table 2.1. In order to study sampling error, 2 – 4 increments from prescribed positions in the fermentation box (reactor) were obtained according to a sampling plan. Increments kept apart for the following analyses in order to allow a rough guesstimate of the sampling uncertainty.

In the laboratory, the field sample increments were further divided into 6 portions à 20 beans each for chemical reference analysis plus 100 beans for a cut test per fermentation-day (i. e., 25 – 50 beans per increment, depending on the number of increments avail-

able for the fermentation-day in question). Approximately 25 beans per fermentation-day were kept for possible further experiments. The large remainder of the material undergoes further roasting experiments and sensory analysis.

Laboratory sample splitting was done in a 2-stage process. First, an in-house constructed sample divider was used to obtain a pre-set fraction of the sample containing at least the number of beans required for chemical analysis and cut test. From these, the required number of beans for each portion was *randomly* selected.

2.2 Spectroscopic Measurements and Reference Analyses

Portions for chemical reference analyses were peeled and ground. NIR spectra between 3600 and 12500 cm^{-1} of $2 \times$ approximately 1 ml of the ground material were measured with a Bruker MPA (Bruker, Ettlingen/Germany) NIR spectrometer at 8 cm^{-1} spectral resolution.

Afterwards, the 2017 material material was de-greased and extracted with methanol. In order to keep track of analytical error, this was done in duplicate as was the actual chromatography run.

2.3 Design of Experiments and Statistical Analysis

Simulation In order to see the effects of the different experimental designs proposed by Bainbridge [1], we set up a simulation comparing these designs and their analysis by mixed models for two common scenarios comprising 3 levels of data (e. g. primary/field samples, portions/aliquots, and measurements):

1. the number of possible measurements at the lowest level is limited — a situation frequently encountered with time-consuming wet-chemical reference analyses
Here, we simulate 48 measurements in total, and in consequence according to the chosen design 12, 16, and 24 primary samples with 1 or 2 aliquots/portions each.
2. the number of available samples at the topmost level is limited — a situation rather typical for studies of biological systems.
The simulation has 12 primary samples and in consequence 48, 36 and 24 measurements in total (again 1 or 2 aliquots).

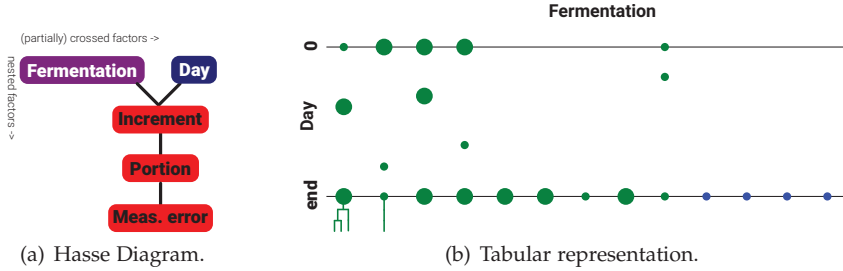


Figure 2.2: Thinned out design of experiments for reference analyses (2018 material): **(a)** The Relationship among the various factors. Blue: Fixed factor fermentation *Day*, red: Random factors (confounders: The field sampling variance in *Increment*, biological bean-to-bean variance covered in *Portion* as well as the analytical random error in the lab procedure *Meas.error*), violet: Factor *Fermentation* comprises a mix of fixed and random factors such as random batch-to-batch variation as well as possible variation due to region and variety (fixed factors) which cannot be separated due to lack of samples. **(b)** Thinned-out experimental plan showing the partially crossed design for factors *Fermentation* and *Day* chosen for reference analysis. Large dots mark *Fermentation* \times *Days* where two increments and for one of them two portions were selected. For the other increment as well as for the *Fermentation* \times *Days* marked with small dots, a single portion was randomly chosen (depicted exemplarily for the two leftmost *Day n* samples). Blue circles mark the samples taken from commercially imported cacao.

The data is univariate random with mean zero and variance 1 for each of the data hierarchy levels (nested random factors). We obtain point estimates as well as bootstrapped 95 % confidence intervals for the 3 variance components via mixed models [2]. The simulation comprises 1000 runs per scenario.

Reference Analysis DoE The 2017 material was analysed with a fully nested design employing 3 portions (“Aliquot”) \times double extraction \times double chromatography.

As the capacity for wet chemical processing and reference analysis is limited, we employ a thinned-out design for selecting samples for a first round of reference analyses of the 2018 material: Of the 44

available *Fermentation* \times *Days* 30 are selected so that all fermentation are covered always with samples of the fermented cacao beans. Where samples were taken throughout the fermentation, a sample from before the start as well as an additional sample from one randomly selected day in between was chosen in addition. This reduces the number of portions to be analyzed from 264 to 45, i.e. about 17 % while still retaining the ability to check the approximate contribution of variance of the various confounders. This procedure takes into account the relations of the various factors, namely that *Fermentation* and *Day* are partially crossed while *Increments* are nested within *Fermentation* \times *Days*, *Portions* within *Increments* and further random errors characterizing the analytical method are nested within *Portions*, see the so-called Hasse-Diagram (see [3] for a discussion).

Based on these reference analyses, a preliminary calibration will be performed which then allows to select a further portions for reference analysis in order to achieve a good and roughly uniform coverage of calibration samples in concentration space.

All statistical analyses were performed in R [4], in particular using packages *hyperSpec* [5] for handling of the spectra and [2] for mixed models.

3 Results

Simulation In a preliminary simulation experiment, we compared the three experimental designs described by Bainbridge [1]. We observe (fig. 2.3) across all designs and for both scenarios that the variance at the uppermost level is somewhat overestimated while the lower two variance components (factors) are on average well estimated (though the lowermost is slightly underestimated with the inverted nested design). As expected, the fully nested design allows precise estimation of the variance of the lowermost factor, but the uppermost variance estimation is highly imprecise: The median confidence interval width for the estimated standard deviation of the topmost factor is about twice the actual standard deviation. In comparison, the staggered nested design achieves a slight improvement on the confidence interval width for the uppermost factor which comes at the cost of slightly worse precision in the variance estimates of the middle and lowermost factors.

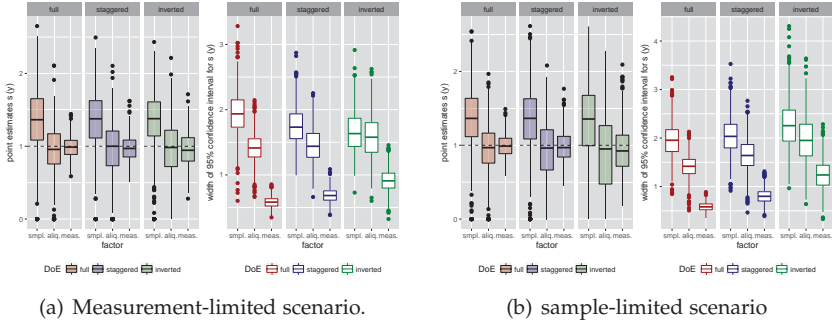


Figure 2.3: Results of the mixed-model variance component estimation for fully nested, staggered nested and inverted nested designs on simulated data. Filled boxes: Distribution of point estimates and empty boxed distribution of the estimated confidence interval widths for the 1000 simulation runs.

The inverted design goes further in this direction and achieves similar precision for the middle and uppermost factors. The lowermost factor is still estimated with higher precision. These trends are similar for both scenarios. However, in the measurement-limited scenario the confidence interval width for the topmost factor improves from fully nested over staggered nested to inverted nested design, whereas for the sample-limited scenario all variance estimates get increasingly imprecise in the same order. This is plausible considering that here for few samples, the already limited number of measurements is progressively decreased, while for the measurement limited scenario the number of measurements stays constant and is distributed across a larger number of primary samples.

We conclude that the inverted nested design can be recommended in situations where the total number of measurements, i. e. the number of samples at the lowermost factor, is limited. If, instead the number of primary samples (of the topmost factor) is the limitation, a fully nested design yields the maximum amount of information that can be gotten from the limited number of samples.

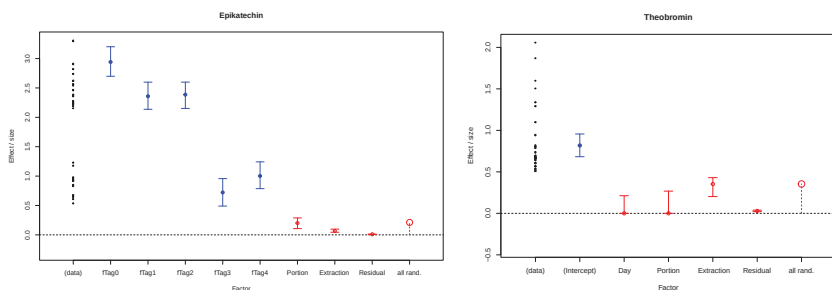


Figure 2.4: Changes in epicatechin content (**left**) during fermentation (blue, fixed factor) and comparison with the order of magnitude of 3 levels of confounding variance (red, random factors) in the laboratory processing chain for the reference analyses: Variance due to the laboratory sample division including the biological bean-to-bean variance, random error of the extraction procedure and random error of the actual chromatographic reference analysis. Theobromin content (**right**) is not expected to vary significantly during the fermentation, so only one fixed factor (blue) for the content is modeled and the day-to-day variation is considered mostly random field sampling error.

Epicatechin content mixed model. Fig. 2.4 shows preliminary results for analyte epicatechin of the 2017 material. Epicatechin is a secondary plant metabolite and antioxidant. Epicatechin content decreases during fermentation with a marked drop after day 2, i.e. when fermentation conditions are changed from anaerobic to aerobic. In addition to the fixed factor describing the temporal development, we extracted variance estimates for three confounding factors nested within our samples. The uppermost factor, *Portion* contains variance due to each portion consisting only of a limited number of cacao beans and is thus related to the biological bean-to-bean variance. We also observe that the random errors contributed by further steps in the laboratory processing (*Extraction*) and actual measurement (*Residual*) are each considerably smaller than the next. This corresponds well with the rule of thumb that the measurement noise contributed by modern analytical equipment is typically much smaller than the random errors introduced in the processing of samples and that sampling is often the step introducing most uncertainty. We conclude that the most efficient use of further

experimental resources here is to examine more and/or larger portions of cacao beans. In other words, upcoming work (2018 material) will benefit from examining more primary samples and in turn switching to an inverted nested design.

In contrast, for theobromine, we found that the extraction procedure was the main contributor of uncertainty, and in consequence focus our attention to that part of the analytical procedure. Note that here we also see the difficulty in estimating variance components further up in the data hierarchy: *Portion* (bean-to-bean variance) and *Day* (field sampling error) cannot reliably be estimated. This is a direct consequence of the large variance at *Extraction* level.

4 Acknowledgements and Contributions

Financial support of the project “CocoaChain” (IGF 169 EN/3) by the AIF (Arbeitskreis industrielle Forschung) and FEI (Forschungskreis der Ernährungsindustrie) is highly acknowledged.

The authors thank H. Balster (2017), M. Nourisson, C. Bahmann (2018) and K. Zug (2017 + 218) for taking the field samples, as well as F. Tietz for preparing the extracts for chromatographic analysis of the 2017 material and M. Harke for helping with preparation of the lab samples, NIRS and some reference measurements.

The field sampling plans were constructed by AK (2017) and CBe (2017, 2018). Lab sample division, randomized measurement plans, selection of samples for reference analysis and statistical/chemometric analyses including the simulation study were performed by CBe. The reference analysis plan and methodology was devised by AK, CBö and CBe. Chromatographic analyses were set up, optimized and performed by MG and CBö.

5 Summary

Inverted nested designs of experiments allows us to construct a highly thinned-out set of samples for reference analysis while for NIR spectroscopy where sample preparation and measurement are less time-consuming and more suitable for automation we employ a fully nested design. Both designs allow estimation of nested variance components,

but the inverted nested design in our case runs time-consuming wet-lab preparation and reference analysis for only $\approx 17\%$ of the available samples in the first round of a two round calibration strategy. In a second round, more samples will be added and their selection will be guided by the preliminary calibration obtained in the first round.

We use mixed models to estimate contributions of several sources of uncertainty, i. e. confounding factors, along the analytical-chemical processing chain of the samples. This in turn allows us to focus further work where it is most efficient: On the respectively largest source of uncertainty in the processing chain.

References

1. T. R. Bainbridge, "Staggered, nested designs for estimating variance components," *Industrial Quality Control*, pp. 12–20, 1965.
2. D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
3. G. W. Oehlert, *A First Course in Design and Analysis of Experiments*. (W. H. Freeman), 2010. [Online]. Available: <http://users.stat.umn.edu/~gary/book/fcdae.pdf>
4. R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: <https://www.R-project.org/>
5. C. Beleites and V. Sergo, *hyperSpec: a package to handle hyperspectral data sets in R*, 2018, r package version 0.99-20181022. [Online]. Available: <http://hyperspec.r-forge.r-project.org>