Analytica Chimica Acta 893 (2015) 14–24

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Tutorial

Validation of chemometric models – A tutorial

Frank Westad [a, *], Federico Marini [b]

[a] CAMO Software AS, Nedre Vollgate 8, N-0158 Oslo, Norway
[b] Dept. of Chemistry, University of Rome "La Sapienza", I-00185 Rome, Italy

HIGHLIGHTS

- The different approaches to validation are presented and discussed.
- Data-driven vs hypothesis-oriented.
- Illustration of the effects of adopting different strategies.

GRAPHICAL ABSTRACT

ARTICLE INFO

Article history:
Received 2 February 2015
Received in revised form
29 May 2015
Accepted 29 June 2015
Available online 10 August 2015

Keywords:
Validation
Chemometrics
Resampling
Test set
Cross-validation

ABSTRACT

In this tutorial, we focus on validation both from a numerical and conceptual point of view. The often applied reported procedure in the literature of (repeatedly) dividing a dataset randomly into a calibration and test set must be applied with care. It can only be justified when there is no systematic stratification of the objects that will affect the validated estimates or figures of merits such as RMSE or R^2. The various levels of validation may, typically, be repeatability, reproducibility, and instrument and raw material variation. Examples of how one data set can be validated across this background information illustrate that it will affect the figures of merits as well as the dimensionality of the models. Even more important is the robustness of the models for predicting future samples. Another aspect that is brought to attention is validation in terms of the overall conclusions when observing a specific system. One example is to apply several methods for finding the significant variables and see if there is a consensus subset that also matches what is reported in the literature or based on the underlying chemistry.

Contents

1. Introduction 15
2. Theory 15
 2.1. Data-driven vs. hypothesis-driven validation 15
  2.1.1. Data driven validation 15
 2.2. Hypothesis driven validation 17
  2.2.1. Confirmation of theory/application specific knowledge 17
  2.2.2. Scientific significance, induction vs. deduction 18

* Corresponding author.
E-mail address: fw@camo.com (F. Westad).

http://dx.doi.org/10.1016/j.aca.2015.06.056
0003-2670/

3. Data 18
3.1. Oat flour 18
3.2. Tablets 18
3.3. Beer 18
3.4. QTL genetic marker data 18
4. Results 18
4.1. Example of sample selection 19
4.2. Example of validating model performance across replicates etc. 19
4.3. Confirmation of findings across data-analytical methods 21
4.4. Confirmation when the "truth" is known 22
4.5. Validation of two methods for variable selection and subsequent prediction 22
5. Conclusions 23
References 23

1. Introduction

It is fair to say that validation, as a concept, is one of the most important aspects in science. In this tutorial, we will – unlike many other papers on validation in quantitative sciences, like analytical chemistry – not solely put emphasis on the numerical aspects. We feel that validation must be presented at a more conceptual level, and be driven by an underlying hypothesis given the actual application [1]. The tradition we see in other sciences, such as medicine, is to more formally set up a research hypothesis, which may be confirmed or rejected. From the authors' experience, this is not so prominent in analytical chemistry. One reason may be that the analytical chemists' role in research project is as a "source" of analytical results in a larger picture. However, as the scientist closest to the aspects of sampling, measurement procedure, instrument suitability etc., the analytical chemist must convey the importance of validated findings as the basis for the conclusions of the particular study, being it medicine, biology, forensics and so on.

When this is said, there exists in analytical chemistry and chemometrics a strong awareness of the importance of validation, and the necessity of validating models with unknown samples is often highlighted.

This tutorial aims at presenting validation in a wide context, with examples taken from analytical chemistry to illustrate how the level of validation and the choice of methods for analyzing data may impact the conclusions and chemical insight gained.

2. Theory

2.1. Data-driven vs. hypothesis-driven validation

As stated above a large portion of the scientific publications on validation concerns aspects of numerical nature, such as repeatability of measurements and prediction error in quantitative models. This may be looked upon as *internal*, or *data-driven*, validation (induction, empirical), where the analytical result is discussed within the scope of the project. One may also look at validation in an *external* or *hypothesis-driven* context, where the results are confirmed by theory (deduction, first principles) or existing knowledge. The distinction between empirical and first principle models may not be so obvious, as in the scientific-philosophical context of deduction and induction. In fact, most formulas in physics and chemistry are based on experiments. Textbook formulas are often approximations, although often conveyed to the reader in the basic courses of the curriculum as the true relation. One example is the ideal gas law $PV = nRT$, which for gases at high pressure is extended to $PV = zmRT/M_w$. z is here the compressibility factor and M_w the molecular weight. Numerous other examples exist in chemical engineering in fields like flow theory, drying and distillation. Another aspect of this is if the sensors we use measure "das Ding an sich", as the German philosopher Kant expressed it. Temperature has been measured for centuries by thermometers that are all based on indirect correlations (quicksilver, ethanol, thermocouple). Thus, one may look at a spectrometer applied for multivariate calibration as just as fundamental as many other (univariate) sensors applied in chemistry. Also, it must be stressed that the principle for all these sensors is based on inverse causality; it is the concentration of a chemical compound that gives an absorbance value. In general, it is not advised to extrapolate empirical models, but it is not given that first-principle models are universally valid. One example is Beer–Lambert's law, which is often extrapolated beyond the suggested range, where linearity holds. This more generic discussion is not pursued further in this tutorial but it is the authors' view that this topic deserves more attention in the scientific community.

2.1.1. Data driven validation

2.1.1.1. Analysis of variance (ANOVA). One way to validate a system/process based on hypotheses is to setup an experimental design, where the sources of variations are systematically varied to generate a structured data table and partition the overall variance into the various sources of error by use of ANOVA [2–4]. Thus, such designs may help in understanding the causality of our system since one can for the basic factorial designs estimate the effect of Factor A independently of Factor B. This is a valid approach for investigating known variables that may influence the results, which can be set to specific quantitative values or categorical levels. ANOVA may also be used for analyzing empirical data in general but once the variables are not orthogonal there is no "truth" as such regarding which way to estimate the sum of squares [5–9]. Without going into details, this is one reason why latent variables methods are widely used in chemistry because many variables may have the same information content in a specific application. Applying ANOVA in these situations leaves us with two choices: a) remove some of the model variables to avoid collinearity (indeterminacy) b) keep them but knowing that the order in which they are listed in the data table may affect the p-values and size and signs of the effects estimated.

2.1.1.2. Test-set versus cross-validation. When the objective is to establish a calibration model for predicting quantities such as concentration, the most conservative validation is to test the model on a representative independent test-set of sufficient size. This has been discussed in length in Ref. [10]. Then, it may be debated, given the specific application, what is meant by such a test-set; should it allow for extrapolation of the calibration space?; is the assumption that the model shall be robust towards change in sample matrix, raw materials, chemical reagents, etc.? For general comments

about extrapolation of empirical models see Section 2.1 above. These sources of variation, that are in principle unknown for future objects, can be to some extent quantified by several approaches. This is where an application-specific level of validation, in terms of cross-validation (CV) of the calibration objects, may be applied. In Ref. [11] the authors make a general comment that, for sample sets >50, test set validation is preferred, whereas cross-validation is best for small to medium datasets. In situations with small data sets, aggregation may improve model stability [12]. Even though CV is generally regarded as the second best validation method, in this context it is useful for several purposes:

1. The number of objects is limited: In situations where one can not readily "ask" for more objects but must rely on the first set of objects, because the underlying conditions in the objects' universe have changed. Typical examples can be found in biology, environmental research, aquaculture etc. In this situation, one will need all objects to build the best model for interpretation, to include the underlying phenomena in the model, and to ensure high stability; one can thus not afford to put aside 30–40% of the objects as a test set. A test-set of insufficient size or sample variation may give a worse estimate of future prediction error than cross-validation. One rule of thumb is to apply cross validation if the number of samples is smaller than 40.
2. The main purpose of establishing a model may not in itself be for predicting or classifying new objects, but to understand the inherent structure in the system under observation. In chemometrics, this relates to so-called latent variables, that may convey the basic chemical or biological phenomena. The interpretation of such models is highly dependent on the number of latent variables, and, therefore, it is vital to assess the correct dimensionality of the model, i.e., in more mathematical terms, the model *rank*. In this context, it is of uttermost importance to distinguish between numerical rank, statistical rank and the domain-specific rank. In particular, numerical rank indicates the dimensionality of the largest subspace spanned by the samples, i.e., the number of components needed to exactly reconstruct the data matrix(-es) with zero error. On the other hand, the term statistical (chemical) rank is used to denote the number of latent variables needed to approximate only the systematic (informative) variation present in the data, assuming that the remaining components (not included in the model) account for irrelevant or spurious variation and, in general, error contributions. This rank may be estimated by one of many statistical procedures. Lastly, the term domain-specific (or application-specific) rank is sometimes used to indicate that number of components which reflects particular background hypotheses or background knowledge. Note that, even though a representative test-set is present, it is nevertheless important to find the correct model rank for predicting the test set; if the model rank is incorrect, the figure of merit (R^2, RMSEP, false classification rate …) may not be a good estimate of future objects (for which there will not be any reference values).
3. The objects in a data table can be stratified into groups based on background information about the origin of the objects. Such groups are a consequence of the experimental set-up of the study. Typical stratifications are:
   - Across instrumental replicates (repeatability)
   - Reproducibility (analyst, instrument, reagent…)
   - Sampling site and time
   - Across treatment/origin (year, raw material, batch…)

Cross-validation performed at the various grouping level will give important information about the stability of the model and which sources of variation that need special attention. Thus, even if a test set has been defined as the proper way of validating the model (or process or system in a wider context), the calibration set must be validated with CV at the appropriate level. If not, the model dimensionality may not be conservative enough and the test set is predicted with a suboptimal number of variables or components. It is also important that the test set consists of samples from other levels of the underlying data structure. Assume data for determination of total organic carbon (TOC) in soils have been sampled at 10 geographical sites, for analysis with a chemical reference method and spectroscopic measurements. Without thinking of the underlying structure, one may be tempted to divide the samples randomly into a calibration set (70%) and test set (30%) as the validation scheme. Then, if the test set prediction error is close to the calibration error one concludes proof of concept. However, for a model to be applied in practice, which involves predicting TOC at *other* sampling sites, a suggested validation scheme is: 1. Model seven of the ten samples sets and cross-validate across the sampling sites to give a conservative number of latent variables or subset of variables. 2. Predict the three other sampling sites with this model. This validation scheme reflects the prediction error at new unknown sampling sites, although no guarantee can be given that this first set of ten sample origins spans all other sites. One may also repeat this procedure.

In Ref. [13] it is pointed out that "cross-validation demonstrates prediction, but is an unlikely scenario in industrial applications, where concomitant data acquisition for model development and test materials would be unwieldy". In this context, the same applies to random splits into calibration and test sets. Hawkins, Douglas and Kraker [14] comments: "A further technical issue is a common misapplication of cross-validation, in which it is applied only partially, leading to incorrect results. Statistical theory and empirical investigation verify the efficacy of cross-validation when it is applied correctly". Quantitative Structure-Activity Relationships (QSAR) is a field where a number of variable selection procedures have been used (and sometimes misused). In these applications, there is no apparent grouping of the molecules that would serve as a basis for systematic cross-validation except to group into various classes of compounds which of course is a very conservative approach. Golbraikh and Tropsha [15] concluded that the validated R^2 from leave-one-out CV is necessary but not a sufficient condition for the model to have a high predictive power. Baumann, Albert and Von Koorff [16] conclude that, given no stratification of the samples, leave-multiple-out CV is preferred. In the field of forensics, O'Connell, Ryder, Leger and Howley [17] report the use of a "robust segmented cross-validation". A variant of CV, the so-called Monte-Carlo CV (MCCV) has been presented [18,19], but these studies are not discussing the concept of CV in a broader scientific context.

It must be mentioned that cross-validation cannot serve as a criterion to decide on the best model out of many when variable selection is performed. The more conservative Cross Model Validation (CMV) is a useful alternative in such situations [20–23]. The expected estimate of future prediction error, however, is the same for CMV as for CV, if the samples are homogeneously distributed. Another way to overcome the problem of using the same criterion to select a subset of variables and the error is to divide the objects into a calibration, a validation and a verification set, where the verification set is the "proof of the pudding".

The aspects above will be exemplified in the Results section.

2.1.1.3. Resampling methods: bootstrap, jack-knifing and cross-validation. Resampling methods are widely used to estimate parameters and/or their uncertainty in a model [24,25]. The simplest case is the estimation of the mean of a population. In a multivariate context, resampling methods are applied to estimate the parameters and their uncertainty with two objectives: a) To estimate the

dimensionality of the model in terms of latent variables; b) To estimate the uncertainty of individual variables, in order to find the relevant ones (out of many).

The main difference between Jack-knifing and bootstrapping is that bootstrapping is resampling with replacement; thus, in, e.g., one bootstrap sample-subset of size 100, one particular sample may appear more than once. There is also a distinction between conditional and unconditional bootstrapping. In the unconditional approach, bootstrap is carried out on the original data, so that – at the *m*th iteration – matrix(–es) $\mathbf{X}_m$ (and $\mathbf{y}_m$) are built by sampling with replacement from the complete set of objects available and, each time, the model is calculated using the selected subset of samples. On the other hand, conditional approach operates by sampling with replacement from the residuals obtained after fitting the model on the complete (non-stochastic) data set. For instance, if one assumes the model to be

$$\mathbf{y} = \mathbf{Xb} + \mathbf{f} = \hat{\mathbf{y}} + \mathbf{f} \quad (1)$$

i.e., a multivariate calibration situation, where **b** is the vector of regression coefficients and **f** that of residuals, while $\hat{\mathbf{y}}$ denotes the vector of model estimates, then, at the *m*th iteration, the bootstrapped version of the vector **y** ($\mathbf{y}_m$) is built according to:

$$\mathbf{y}_m = \hat{\mathbf{y}} + \mathbf{f}^*_m \quad (2)$$

where $\mathbf{f}^*_m$ is a vector obtained by sampling with replacement from the residuals of the full model. Accordingly, the model estimates are obtained by fitting the vector $\mathbf{y}_m$ to the original **X** matrix. Here it must be stressed that, since in the multivariate (bilinear modeling) context it may be difficult to decide on the dimensionality of the data a priori, bootstrap on the original data (unconditional) is the approach more frequently used.

The original Jack-knifing procedure used the mean of the sub-models as the reference in estimating the variance, whereas in cross-validation the model on all object is used as the reference. The difference between the two approaches is, according to Efron [26], of order 1/(number of objects). Intuitively, it is more relevant to use the model on all objects as the reference in our opinion. Cross-validation gives, in general, slightly higher uncertainty estimates for the parameters than bootstrapping, and the estimates reflect how the validation was done: across, e.g., replicate, sample, raw materials, year or production site as mentioned in the previous section. The data analyst must therefore have access to all qualitative information about the samples to perform the model validation given the underlying stratification of the objects.

A recent comparison of re-sampling techniques to determine the optimal number of components in PLS regression was reported in Ref. [27].

2.1.1.4. How to select subsets of samples for calibration and validation. When a single training/test split is performed on the data set, an intelligent choice of the samples to be put in each set is needed in order to be able to produce reliable considerations based on the obtained results. Different criteria have been proposed in the literature to operate an intelligent splitting of the available samples among the sets and they all share the same concept, i.e., to try to span the sample space as uniformly as possible.

Historically, the first algorithm to select a representative subset of the available samples so that they span as uniformly as possible the design space was proposed by Kennard and Stone [28]. In detail, given a set of candidate samples, Kennard and Stone algorithm aims at selecting the most diverse among them to be included in the training set, according to a *maximin* criterion. Indeed, at first the distances among all pairs of samples are computed and the two most distant samples are selected to be included in the training set. Successively, for each of the remaining candidate samples, the minimum distance to all the already selected samples is computed, so that the one showing the maximum value of this minimum distance is in turn selected to be included in the training set. The whole procedure is then repeated until the desired number of training samples is selected.

As the Kennard-Stone approach tries to concentrate as much of the diversity in the original data set in the training sample, depending on the data configurations it could lead to overoptimistic results. Based on these considerations, a modification of the algorithm aimed at maintaining a comparable diversity between the two sets was proposed by Kennard himself (even though it was left unpublished until it was discussed by Snee [29]). The corresponding algorithm, named Duplex, starts as the original Kennard-Stone, by computing all the distances among samples and selecting the two most distant samples to be included in the training set; however, it continues by putting the second two most distant samples in the test set. Successively, the most diverse samples according to the already mentioned *maximin* criterion (maximum minimum distance) are in turn added to one or another set, until the desired splitting ratio and the requested number of samples are obtained.

Another way of achieving an intelligent splitting of the data set so that the training samples span as uniformly as possible the design space is to select the samples according to a D-optimality criterion [30]. The principle of D-optimal designs is to select a subset among the candidate samples so to maximize the determinant of the information matrix ($\mathbf{X}'\mathbf{X}$): this determinant is maximized when the selected samples span as much as possible of the space of the whole data. With respect to Kennard-Stone algorithm, in selecting the samples D-optimal approach privileges more high leverage and peripheral points.

Lastly, another possibility of selecting training samples among a set of candidate object is to use clustering techniques, like k-means or Kohonen mapping. In particular, the latter technique has proved very effective in several occasions in producing a representative data splitting [31,32]. A Kohonen neural network operates by mapping samples from an N-dimensional space onto a discrete 2-dimensional grid of neurons, so that objects that have similar properties in the original space will map to the same or to neighboring nodes. Accordingly, by selecting a proper dimensionality of the 2D neural network, one can make so that more objects map to the same neuron. Then, for each position of the 2D grid a certain fraction (in general, from 2/3 to ¾, depending on the numerosity of the data set and the density of the mapping) of the objects is selected to be included in the training set, and the remaining are used for validation.

2.2. Hypothesis driven validation

This section presents some aspects regarding validation in terms of confirming hypotheses, theoretical or first-principle models and the “true model”. Some points to consider in this context are.

- Confirm existing knowledge, e.g., from literature and other sources
- The true underlying model is found, maybe with small adaptations, if the system under observation is not exactly identical (other chemicals within the same group of compounds)
- Recognize the underlying profiles or inherent latent variables in chemistry or biology

2.2.1. Confirmation of theory/application specific knowledge

A level of validation that is of special interest to the analytical chemist is the method's ability to find the true signal of the

chemical compounds in a system. Assuming that the total signal acquired with a suitable instrument is free from unknown interferences, then the observed signal should be the sum of the true signal of individual compounds times the concentration:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} \tag{3}$$

Multivariate methods that may be suited to find the true signals and the corresponding concentrations are e.g., Multivariate Curve Resolution (MCR) [33], Independent Component Analysis (ICA) [34], SIMPLISMA [35]. It should also be mentioned that the true signals of unknown interferences can also be estimated using the appropriate method for the given data. Constraints may be imposed to improve the chance of success: non-negativity, unimodality, closure and some type of equality constraints based on local rank and on selectivity from previously known information are those, which are most commonly adopted. In the context of mixture analysis, the use of constraints represents one of the simplest yet most effective tools to deal with the problem of rotational ambiguity, i.e., the fact that – given the model in Eq. (3) – one could find a set of transformation matrices **T**, so that the decomposition:

$$\mathbf{X} = (\mathbf{C}\mathbf{T})\left(\mathbf{T}^{-1}\mathbf{S}^{\mathrm{T}}\right) \tag{4}$$

would give the same fit of the data. In this context, constraints are needed to get narrower and narrower band solution, because of the lack of unique solution for bilinear data in general. However, as pointed out by Rajko [36], particular care in the use of the proper constraints and conditions and in the interpretation of the results as, despite the reduction in the rotational ambiguity, the obtained solution(s) may not only not be equal to the true one, but even not lie in the feasible region.

These methods will generally give the relative concentrations of the objects, thus the true concentration for one sample may be needed for the estimation of the actual concentration for the compound of interest. Parallel Factor Analysis (PARAFAC) [37] can be used for 3 and higher dimensional data, and a unique solution can exist when some mild conditions are fulfilled, e.g., according to Kruskal ranks. Examples of such data are hyphenated analytical techniques: Excitation-emission fluorescence, LC–UV, LC–MS, GC–MS. PARAFAC may also apply to NMR (COSY, NOSY), LC-MS-MS and other combination of techniques, but the basic assumption is that the data have a tri-linear structure. In this context, multichannel imagery does not fall into the same category, as there is usually no linearity assumption on the information in the 2-dimensional image space that will fit into the PARAFAC framework, and unfolding images for analyzing the data pixel-wise followed by re-mapping information to the image domain is the most common approach.

2.2.2. Scientific significance, induction vs. deduction

Although this is not the main focus of this tutorial, a brief paragraph on the more deeply scientific view on the difference between induction and deduction is included.

Munck et al. [38] present a holistic view on the scientific process and how multivariate methods in food science can play the role as a basis for hypothesis generation and confirm theory. In science, deduction is the scientific method of starting with theory, generating hypotheses and performing experiments to verify or falsify the theory, i.e., going from the general to the specific. For many chemical and biological systems, it is difficult to use basic physical formulas for modeling the system to the required level in terms of explained variance or prediction ability. When this is said, ab initio methods play an important role in fields such as QSAR and spectroscopy. Induction, on the other hand, starts with analysis of experimental data (empiric) and, from there, one may generate hypotheses that can lead to general theory (first principle models) and new insight. Many of the early physicists and chemists started with experimental work that led to basic theory. There should be no conflict of interest towards either one approach; it is when our empirical findings confirm our theory and background knowledge that the causal effects and true underlying relationships in a system are validated.

3. Data

3.1. Oat flour

The first data set used to show the outcomes of different single splitting techniques, and to show the effects of validating across different factors, is made of the NIR spectra within the range 800–2498 nm recorded in reflectance mode on 166 naked oat flour samples; each sample was analyzed in replicate. The samples come from 12 different varieties and from 3 different harvesting years (2006, 2007, and 2008). A more detailed description of the data set can be found in Ref. [39].

3.2. Tablets

The second data set used to show the outcomes of different splitting techniques was the basis of the ShootOut at the 2002 Chambersburg meeting [40]. It includes the NIR spectra of 654 pharmaceutical tablets recorded in transmission mode in the interval 600–1898 nm with two different spectrometers.

3.3. Beer

The data were taken from Ref. [41]: 60 beer samples were measured with a dispersive near infrared spectra NIRSystems Inc. (model 6500) spectrophotometer at 20 °C in the VIS/NIR region (400–2500 nm). Transmission spectra were recorded using a 10 mm quartz cell directly on the undiluted fresh beer, and spectral data collected at 2 nm intervals in the range from 400 nm to 2250 nm were converted to absorbance units, giving a total of 926 variables. The dependent variable was Extract, which was analyzed by Carlsberg A/S, and the range of extract concentrations was 4.23–18.76 mg/L.

3.4. QTL genetic marker data

249 samples of the plant species *Phytolacca dodecandra* from different locations in Ethiopia were subject to RAPD genetic marker analysis. The independent variables (X) were 70 binary variables (RAPD markers) and the dependent variable (y) was the altitude at where the plants had grown [42].

4. Results

The sections below present various aspects of validation:

- Optimal selection of samples
- Correct and wrong validation across test set evaluation given stratification of the objects
- Two methods for variable selection giving the same subset of variables
- Methods giving the same estimate of the significance for a Design of Experiment application where ANOVA is the benchmark
- Comparison of uncertainty estimates for variable selection and the impact on prediction

4.1. Example of sample selection

As already discussed in Section 2.1.1.4, different algorithms can be used for the intelligent splitting of the available data into a single training/test pair. In this paragraph, the results of the application of Kennard-Stone, Duplex, D-optimal and Kohonen-based sample selection schemes on the two data set described in Section 3.1 and 3.2 will be presented. In particular, the results obtained in the case of the oat data set, where a 2:1 splitting ratio was adopted, are reported in Fig. 1. It is apparent from the figure that, accordingly to what could be expected based on the theoretical considerations reported in Section 2.1.1.4, both Kennard-Stone and D-optimal based approaches tend to capture as much as possible of the sample diversity in the training set, so that practically none of the samples which are relatively far from the bulk of objects in the plot is included in the test set. On the other hand, Kohonen, and to a greater extent, Duplex, provide a more representative selection, maintaining the same diversity among the sets. This situation is even more evident in Fig. 2, where the outcomes for the tablets data set (a training/test splitting ratio of 3:2 was used) are reported. Indeed, in this case when Kennard-Stone or D-optimal algorithms are applied, none of the samples in the second cluster is included in the validation set.

In general, from the results reported above it is possible to conclude that the use of an intelligent splitting criterion allows governing with a reasonable confidence, to what extent the diversity originally present in the data set will be preserved in the training and test subsets and, as a consequence, to direct the choice about which strategy to use, depending on the specific modeling to be carried out. For instance, if there is the suspect of possible outliers or extreme points in the data set, one could choose to use a robust calibration approach to build the models and, and so one would like to have all the most diverse samples in the training set. With such an approach, strategies like Kennard-Stone or D-Optimal would be recommended. On the other hand, in almost all the other situations, where one aims at capturing the same diversity in both sets, duplex (or Kohonen-based selection) should be preferred.

In general, the possibility of controlling or tuning the desired outcome by selecting a specific strategy makes the use of "intelligent" splitting algorithms to be preferred over random selection. Indeed, especially for small sample sets, random splitting leads to a high variance of the model estimates, if the selection is not repeated a sufficient number of times (in our experience, for most data sets, at least 30 iterations are needed to have an acceptable stability the solutions). Moreover, in most of the cases, the mean outcomes over the different random training/test splits are comparable or even worse to those obtained by duplex or Kohonen (and, depending on the data, also Kennard-Stone or D-optimal).

Here it must be stressed that, although in this section attention was focused mainly on discussing how the way of selecting the training samples influences the final model, also choosing the proper number of samples plays a key role. Indeed, the choice of the splitting ratio to be adopted must reflect two concurrent issues to be compromised: the number of training samples must be enough to build a stable and reliable model and the number of test set samples should allow a representative generalization of the obtained results. With moderately numerous data set (50–100 samples) training/test splitting ratios of 3:1 to 2:1 normally work well and the fraction of test set objects may be increased even further in the case of larger data set, where more samples are available. Moreover, it is worth noticing that, with medium-small data sets, the choice of the selection strategy becomes even more important and duplex outperforms random selection, which gives comparable predictions only when a high number of repetitions is adopted.

4.2. Example of validating model performance across replicates etc.

As mentioned above, the level of which the validation is performed is important for a liable estimation of future prediction error (RMSEP). As an illustration, we will use the data set made up of the NIR spectra of oat flour samples described above. The 332 objects are divided into various subsets pertaining to the level of

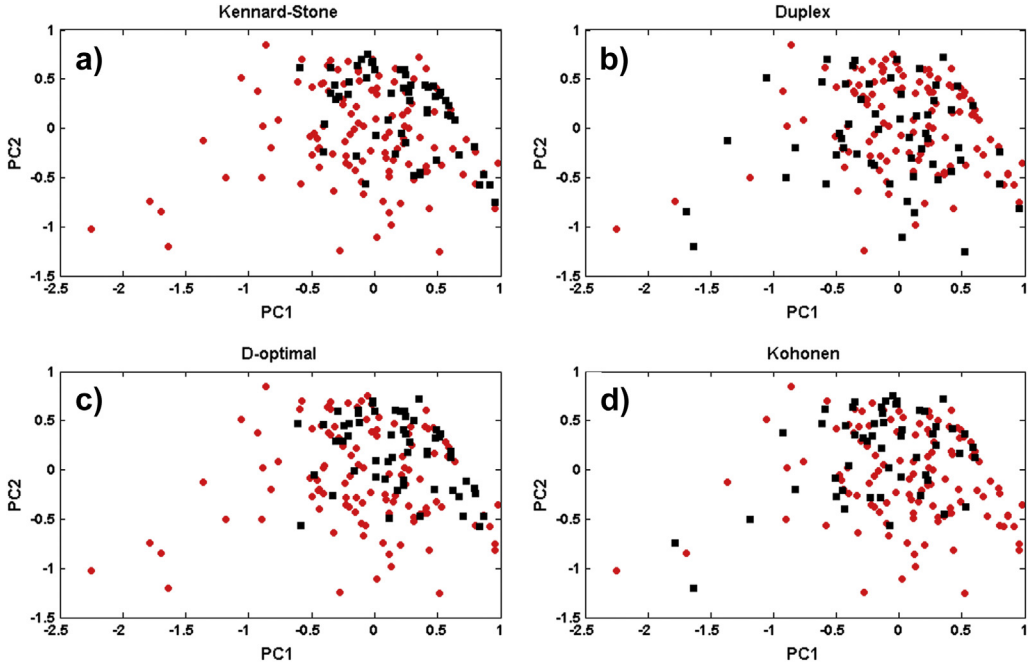

Fig. 1. Effect of the algorithm used for training/test splitting (with a 2:1 ratio) on the distribution of samples as evaluated on a data set made of NIR measurements on 166 naked oat flour samples. (a) Kennard-Stone; (b) duplex; (c) D-optimal; (d) Kohonen. Legend: training set = red circles; test set = black squares. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

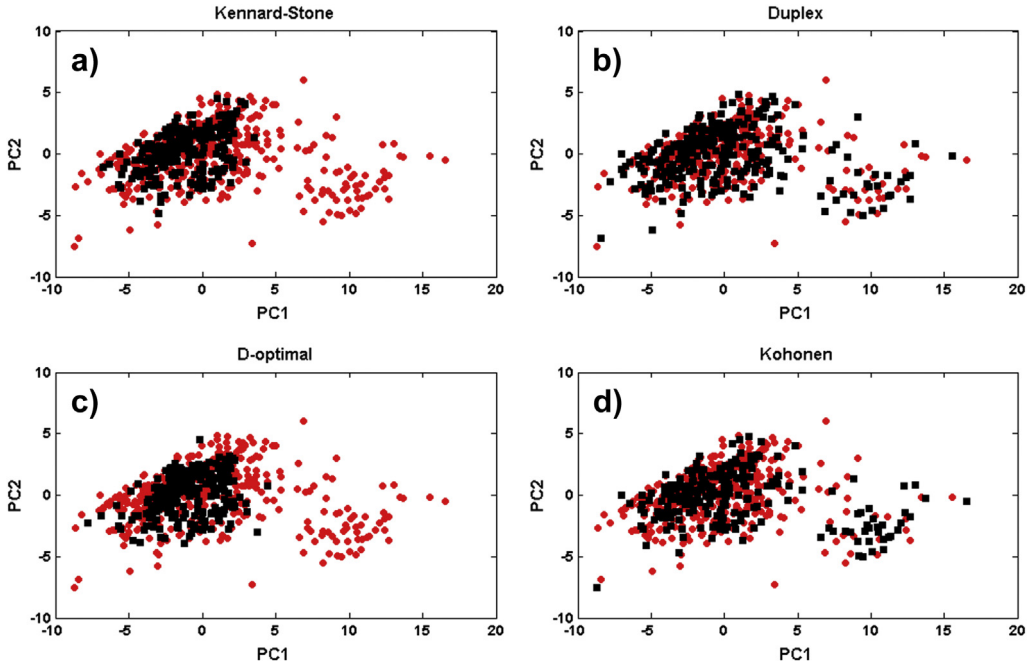

Fig. 2. Effect of the algorithm used for training/test splitting (with a 3:2 ratio) on the distribution of samples as evaluated on a data set made of NIR measurements on 654 pharmaceutical tablets (IDRC shootout 2002). (a) Kennard-Stone; (b) duplex; (c) D-optimal; (d) Kohonen. Legend: training set = red circles; test set = black squares. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

validation: (across replicate, sample, cultivar and year). The data were averaged over the two replicates to give 166 samples with unique reference values as basis for the cases D and E below.

Case A: Random test-set validation.

A common procedure for a given data set is to randomly divide the objects into a calibration and test set. This means that many physical samples will be represented by one replicate in the calibration set and one in the test set, but some samples will have both replicates in either calibration or test set. This scheme does not give specific information with respect to any stratification of the objects due to the level of validation. The RMSEC of this model after the optimal number of factors may serve as a baseline as this error represents the sample and measurement error, to what extent the instrument is suited for modeling the dependent variable and the uncertainty of the reference method.

Case B: Random cross validation.

A common validation scheme is to cross validate randomly. In this case a 20 segments validation was chosen, giving an RMSEC of 0.35 and RMSECV 0.40 of after 10 factors.

Case C: Cross validation across replicates.

This validation schemes takes systematically all replicates for the same physical sample out during cross validation. In this case with two replicates it means that replicate 1 objects are taken out and a model is established for replicate 2 and vice versa. Thus, the validation is a test of how precise one can re-measure the same physical sample.

Case D: Cross validation across physical samples.

The next level in validation if replicates are present is to keep systematically replicates for the same physical samples out during cross-validation. An alternative is to take the average over replicates and predict the individual replicates with the model. The average over replicates was the basis for the following cases.

Cases E & F: Validation across type of cultivar and year.

A more conservative approach than validating with leave-one-out or randomly with 10 segments is to validate across the type of cultivar. From the 12 types of cultivars three of them were assigned to a test set (47) and nine kept as a calibration set (119). Samples for most of the cultivars were measured for all three years. PLS regression with cross-validation over cultivar for the calibration samples gave an RMSEC of 0.39 and an RMSECV of 0.54 after seven factors. As a comparison leave-one-out CV gave an RMSECV of 0.44. Prediction of the three cultivars not included in the calibration set gave an RMSEP of 0.58 indicating that CV across cultivar gives a better estimate of RMSEP of samples for unknown cultivars. When looking at the Hotelling's T^2 statistic, one of the new cultivars was found to be outside of the critical limit which may induce a higher residual for these samples; 0.58 is slightly higher than 0.54. For the model based on years 2006 and 2007 the validation scheme is important for the prediction of the test set from year 2008. The two alternatives were: 1. random 10-segment CV, 2. validate across year, i.e. two segments. The first scheme indicates 7 factors and an RMSEC of 0.35 and an RMSECV of 0.40, the 2008 test samples gave an RMSEP of 0.78. The conservative validation across year indicated 5 factors, an RMSEC of 0.46 and RMSECV of 0.60 and an RMSEP of 0.63 when predicting year 2008. Leave-one-out CV gives an RMSECV of 0.44. As can be visualized in this case for the Hotelling's T^2 statistic all but one sample lie outside the critical limits, thus the model is extrapolated when predicting year 2008. Nevertheless, the RMSEP is not significantly different for the RMSECV when validating across year. Table 1 shows a comparison of the validation and test set schemes presented above for the PLS regression models.

As can be seen in Table 1 the RMSE values are much higher for cases D and E; the cases that are close to a realistic situation for an industrial application. Thus, case A which is a common way to divide samples into calibration and test-set may lead to over-fitting. The reason for this is that, e.g., in the case of replicated measurement for the same physical sample, the replicates might be split into the calibration and test set respectively. The validation is not performed by keeping a physical sample out, which is the operational use of the model for future samples.

Another important aspect is also how the validation scheme affects the estimation of the stability of the model parameters. Figs. 3

Table 1
Comparison of the validation and test set schemes presented above for the PLS regression models for oat flour samples.

Validation scheme	No. of objects	No. of factors	RMSEC	RMSECV	RMSEP
A: Random calibration and test	210/122	8	0.37	–	0.44
B: Random cross validation, 20 segments	332	10	0.35	0.41	–
C: Keeping replicates out	332	8	0.35	0.37	–
D: Keeping sample out	166	8	0.37	0.44	–
E: Model based on 9 cultivars; test set 3 cultivars	118/47	7	0.39	0.44	0.58
F: Model validated randomly year 2006–2007; test 2008	113/53	7	0.35	0.40	0.78
F2: Model validated across year 2006–2007; test 2008	113/53	5	0.46	0.63	0.60

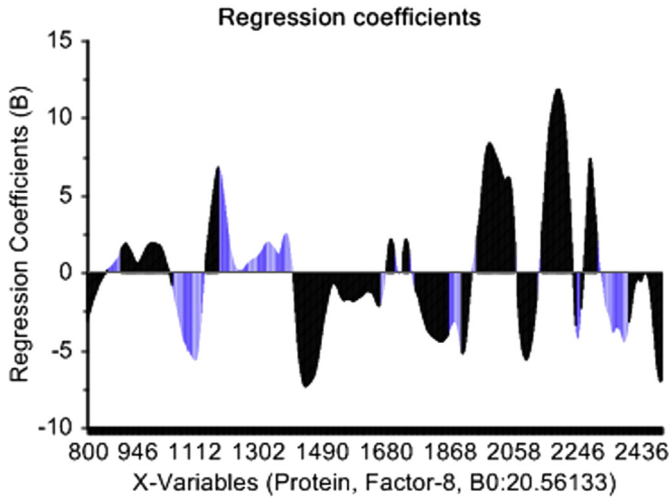

Fig. 3. Oat flour data set: significant variables from uncertainty test when validating across replicates.

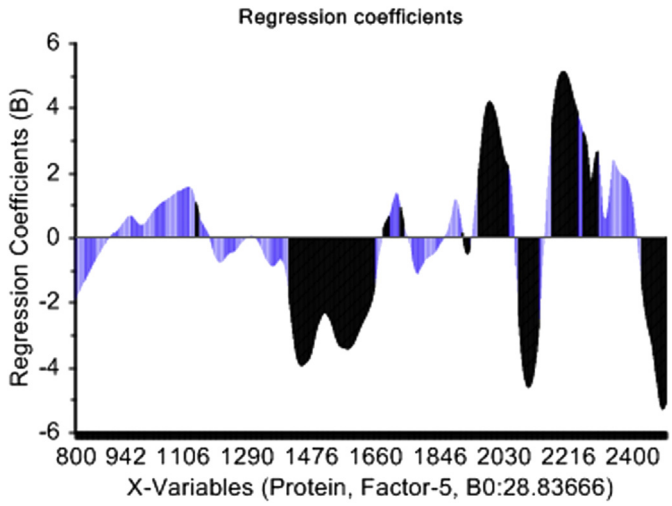

Fig. 4. Oat flour data set: significant variables from uncertainty test when validating across years.

and 4 show the parts of the spectral range that are found to be significant for Cases C and F. As expected the more conservative validation scheme (F) yields fewer significant spectral ranges since the stability of the model is more susceptible to variations over year.

4.3. Confirmation of findings across data-analytical methods

Another important aspect of science, which is often neglected, is the validation related to investigating if several methods give the same result. As scientists tend to focus more or less narrow-mindedly on their own field of interest, already published results related to a manuscript under submission are not known or this information is not proactively pursued. In the chemometric community, it is fair to say that the focus has mostly been on the quantitative aspects, such as showing that one's own method is superior to existing ones. In this context, the grounds of which the various results are compared are also related to the level of validation. In Ref. [43] the authors report that equivalent results were obtained for various multivariate methods for spectroscopic determination of metal ions. Greensill and Walsh [44] compared 10 methods for calibration transfer of models on NIR instruments, and in Ref. [45] the results for classification models for a pharmaceutical product were investigated across three instrument vendors. Several other papers, where various methods have been compared, have been published the past years [46,47]. In Ref. [48] it is highlighted that comparison of methods with extensive search for the "best model" will also lead to optimistic results and that the division into calibration and test set must be considered carefully.

It is common in scientific publications to read how the authors claim that their method is "the best" in terms of model performance, exemplified by prediction error or classification rate. Articles concerning variable selection are no exception in this case. However, in many cases, there is no statistical inference whether one (novel) method is significantly better than another (existing) method. One may say that publishing a paper where e.g. a 5% reduction in prediction error is more of an academic drill than a practical aspect for an on-line method. Furthermore, the evaluation of if models are significantly different is often based on ad-hoc interpretation, e.g., claiming that an error of 0.45 is lower than 0.49, which is hardly the case.

From a scientific viewpoint, it may be of more interest to evaluate if various methods with variable selection as the objective give the same subset of variables that are regarded as being relevant. Also important is if these variables confirm existing knowledge about the system/process, and if they can explain causality and are not just due to indirect correlation that have predictive power for the empirical domain studied. If several methods have the same objective, they *should* give similar results if they are suited for the purpose. One example is given in Fig. 5, which shows results from two methods for variable selection. The data can be found in Ref. [42] and represent genetic analysis (Quantitative Trait Loci) of various samples of an Ethiopic plant. The dataset consisted of 234 samples and 70 genetic markers. One objective in the study was to find the genetic markers for which the concentration changes as the plant adapts to the altitude. In this case both Partial Least Squares Regression (PLSR) with jack-knifing [41,49] and an implementation of genetic algorithms [50] were applied to find the best subset of variables. The figure shows that the important variables presented as regression coefficients from jack-knifing (b) match the most frequently found variables in 100 runs of the genetic algorithm (a), where a subset of five variables was the modeling criterion. For visual convenience the jack-knife based p-values are represented as $-\log(p)$ rather than the p-value itself.

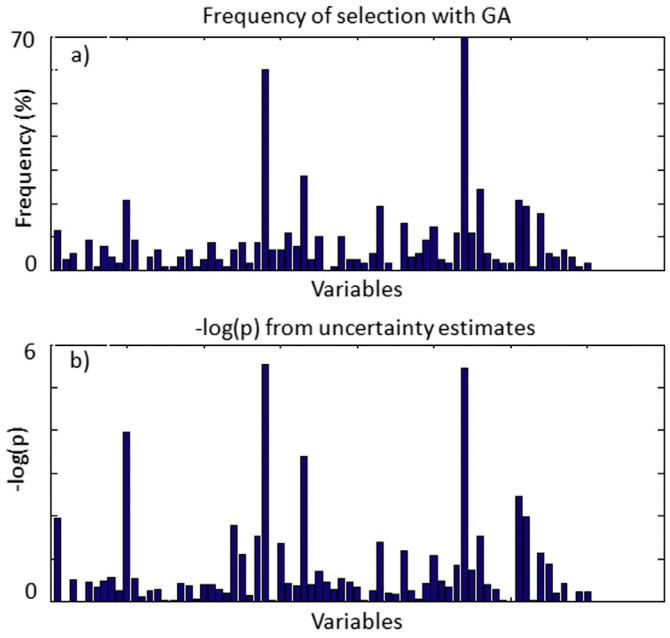

Fig. 5. QTL genetic marker data set: a comparison of two methods for variable selection, genetic algorithm and uncertainty estimates from jack-knifing.

4.4. Confirmation when the "truth" is known

Estimation of model parameters, both their value and their uncertainty, is essential in all empirical models. In this context one should ideally compare to the "true" values if such exist. As multivariate models with many variables do not fulfill, in general, the requirement of ANOVA that the variables should not be correlated, we have chosen a structured data table generated by the design of experiment (DoE) for comparison. The well-known chemist and statistician George Box came up with a nice educational example to show the principles of DoE by letting the participants in a workshop make "paper helicopters" to investigate the impact of a number of variables describing dimensions etc. [51]. The data that were the basis for the results in Table 2 were taken from the second part of the experiment with a subset of the variables as input to a reduced design after the first screening. The two last columns are from a PLS regression model with bootstrapping and jack-knifing respectively [41]. The jack-knife procedure estimates the uncertainty of the model parameters by calculating the difference between the model with all samples and the individual models, i.e. when some samples were kept out. These differences are squared and summed for all the cross-validation segments as the basis for the standard deviation of each model parameter. The bootstrap method is similar except that the mean of all models is used as the reference model in estimating the uncertainties. A t-test is then applied to give a p-value for each parameter, in this case individual variables in the regression coefficient vector. For an orthogonal design and one response variable the PLS regression captures all eigenvalues in X in the first factor.

Table 2
p-values for various estimation methods – helicopter data.

Variable	ANOVA	JK PC1	BS PC 1
Block	0.613	0.671	0.583
Wing area	0.961	0.970	0.962
Wing ratio	0.005	0.024	0.004
Body width	0.882	0.913	0.888
Body length	0.001	0.008	0.001

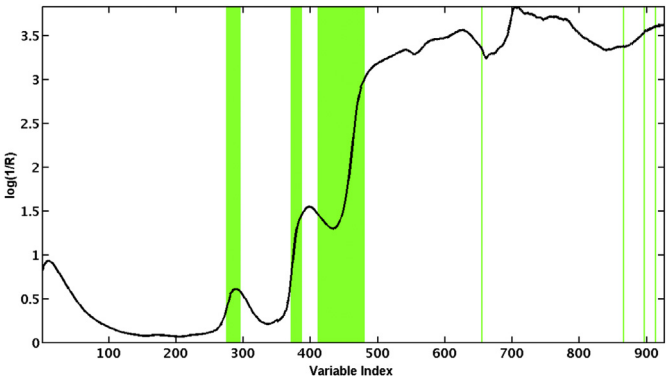

Fig. 6. Beer data set: significant variables from jack-knife uncertainty estimates.

4.5. Validation of two methods for variable selection and subsequent prediction

In the example above, the ANOVA results from the experimental design can be regarded as the reference. However, in empirical data collected from a process, it is not to be expected that the samples are suited for ANOVA as ANOVA only gives an unambiguous results in the case of orthogonal and mixture designs. One may use the data on NIR spectroscopy on beer samples in Ref. [41] for illustration. The data consist of 950 variables in which around 400 of them are hampered by noise because of detector saturation. The 40 calibration samples were cross-validated with 5 segments (venetian blinds) in a PLS regression model. A cross-validation/jack-knife estimation of the uncertainties for each regression coefficient for a 5 factor model was employed (Fig. 6). A similar model was performed by unconditional bootstrap. The bootstrap procedure was repeated 1000 times and 1,00,000 times respectively and yielded many noisy variables to be significant (Fig. 7 shows for the 1000 repeated bootstrap estimates). The significant variables were selected for predicting the 20 test samples. Applying these subsets of variables gave a prediction error for the 20 test samples of 0.22 (jack-knife) and 0.39 (bootstrap, 1000) respectively, This indicates that the bootstrap procedure admits too many noisy variables through the "noise filter", and thus prediction of new samples suffers from higher error. Moreover, no significant improvement, neither in terms of prediction error nor of selected variables, was observed by increasing the number of bootstrap repetitions from 1000 to 1,00,000. It is not only the numerical aspect that is of importance in this example but also the interpretation of the

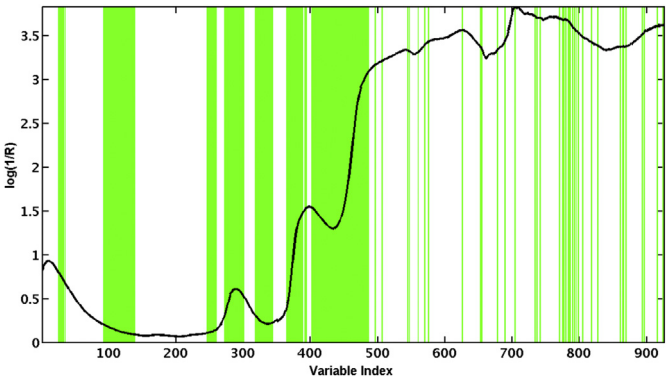

Fig. 7. Beer data set: significant variables from 1000 bootstrap repetitions.

Table 3
Results from modeling and prediction of the beer data.

Model	Variables applied	RMSEC	RMSE[a]	RMSEP
First model	All variables	0.18	1.09	0.73
Jack-knife	Selected variables	0.14	0.27	0.20
Bootstrap	All variables	0.18	1.30	0.73
Bootstrap	Selected variables	0.05	0.42	0.39

[a] RMSECV for cross-validation, RMSEBS for bootstrap.

significance variables found by the two methods. This is in line with the results in Table 3: the jack-knife estimates are more conservative.

The results from cross-validation and test-set validation are given in Table 3.

It should be mentioned that although the examples above are from models where Partial Least Squares Regression was chosen as the method for multivariate regression, the conclusions and principles are valid for any regression method.

5. Conclusions

Validation of chemical systems and processes in general has many facets. The examples shown in this tutorial illustrate that, for numerical validation, the automatic splitting in a calibration and test set can only be justified when there is no stratification of the objects that may influence the model results. This can affect the interpretation with respect to either model dimensionality, or identifying which variables that are important, and what is the true relationships between variables in the system. Proper validation is also imperative to not give unrealistic (i.e., optimistic) estimates of the ability to classify new samples or quantitative prediction of the dependent variable(s) of interest in a regression model. Although the mantra in validation is that a pure independent test set is always required, the validation of the calibration set must nevertheless reflect any subgroups of objects that describe uncontrolled variation which will be unknown for future samples. In all cases, validation is a procedure that aims at providing an answer to a question that has always to be kept in mind: the nature of the question to be answered or of the hypotheses to be verified must always guide the choice of the proper validation strategy to be followed. Accordingly, in this tutorial, we hope to have sketched some of the possible lines along which, depending on the cases, a proper validation can be carried out, both from quantitative and qualitative points of view, and which questions one should always ask oneself in order to design the correct strategies.

References

[1] R. Harshmann, How can I know if it's real?“ A catalogue of diagnostics for use with three-mode factor analysis and multidimensional scaling, in: H.G. Law, C.W. Snyder Jr., J. Hattie, R.P. McDonald (Eds.), Research Methods for Multimode Data Analysis, Praeger, New York, 1984, pp. 566–591.
[2] C.M. Andersen, New Aspects of Chemometrics Applied to Spectroscopy, The Royal Veterinary and Agricultural University, 2003, pp. 54–55. Ph.D. Thesis.
[3] C.M. Andersen, R. Bro, P.B. Brockhoff, Quantifying and handling errors in instrumental measurements using the measurement error theory, J. Chemom. 17 (2003) 621–629.
[4] R. Liu, W. Chen, K. Xu, The influence of experimental design on the model precision in the noninvasive glucose sensing by near-infrared spectroscopy, Proc. SPIE 6826 (2008) (Optics in Health Care and Biomedical Optics III), 682626/1–682626/10.
[5] R.R. Hocking, The Analysis of Linear Models, Brooks/Cole, Monterey, CA, 1985.
[6] O.J. Pendleton, M. Von Tress, R. Bremer, Interpretation of the four types of analysis of variance tables in SAS, Comm. Stat. Theor. Meth. 15 (1986) 2785–2808.
[7] S.R. Searle, G. Casella, C.E. McCulloch, Variance Components, John Wiley and Sons, Hoboken, NJ, 1992.
[8] J. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, sixth ed., Pearson Education Limited, Harlow, UK, 2010.
[9] R.G. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, John Wiley and Sons, New York, NY, 2003.
[10] K. Esbensen, P. Geladi, Principles of proper validation: use and abuse of re-sampling for validation, J. Chemom. 24 (2010) 168–187.
[11] G. Kos, H. Lohninger, R. Krska, Validation of chemometric models for the determination of deoxynivalenol on maize by mid-infrared spectroscopy, Mycotoxin Res. 19 (2003) 149–153.
[12] C. Beleites, R. Salzer, Assessing and improving the stability of chemometric models in small sample size situations, Anal. Bioanal. Chem. 390 (2008) 1261–1271.
[13] J.E. Wood, D. Allaway, E. Boult, I.M. Scott, Operationally realistic validation for prediction of cocoa sensory qualities by high-throughput mass spectrometry, Anal. Chem. 82 (2010) 6048–6055.
[14] D.M. Hawkins, J. Kraker, Deterministic fallacies and model validation, J. Chemom. 24 (2010) 188–193.
[15] A. Golbraikh, A. Tropsha, Beware of q^2!, J. Mol. Graph. Model 20 (2002) 269–276.
[16] K. Baumann, H. Albert, M. Von Korff, A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations,, J. Chemom. 16 (2002) 339–350.
[17] M.-L. O'Connell, A.G. Ryder, M.N. Leger, T. Howley, Qualitative analysis using Raman spectroscopy and chemometrics: a comprehensive model system for narcotics analysis,, Appl. Spectrosc. 64 (2010) 1109–1121.
[18] Q.S. Xu, Y.Z. Liang, Monte Carlo cross validation, Chemom. Intell. Lab. 56 (2001) 1–11.
[19] Q.S. Xu, Y.Z. Liang, Y.P. Du, Monte Carlo cross-validation for selecting model. Prediction error in multivariate calibration, J. Chemom. 18 (2004) 112–120.
[20] J.S. Urban Hjort, Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap, Chapman and Hall, London, UK, 1994, pp. 40–56.
[21] L. Nørgaard, R. Bro, PLS regression in the food industry. A study of N-PLS regression and variable selection for improving prediction errors and interpretation, in: M. Tenenhaus, A. Morineau (Eds.), Les Methods PLS. Symposium International PLS 99, CISIA – CERESTA, France, 1999, pp. 187–202.
[22] E. Anderssen, K. Dyrstad, F. Westad, H. Martens, Reducing over-optimism in variable selection by cross-model validation, Chemom. Intell. Lab. Syst. 84 (2006) 69–74.
[23] F. Westad, N.K. Afseth, R. Bro, Finding relevant spectral regions between spectroscopic techniques by use of cross model validation and partial least squares regression, Anal. Chim. Acta 595 (2007) 323–327.
[24] M. Stone, Cross-validatory choice and assessment of statistical prediction, J Roy, Stat. Soc. B 36 (1974) 111–147.
[25] B. Efron, The Jackknife, the Bootstrap, and Other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982.
[26] B. Efron, Bootstrap methods: another look at the jackknife, Ann. Stat. 1 (1979) 1–26.
[27] L. Xu, Q.S. Xu, M. Yang, H.Z. Zhang, C.B. Cai, J.H. Jiang, H.L. Wu, R.Q. Yu, On estimating model complexity and prediction errors in multivariate calibration: generalized resampling by random sample weighting (RSW), J. Chemom. 25 (2011) 51–58.
[28] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[29] R.D. Snee, Validation of regression models: methods and examples, Technometrics 19 (1977) 415–428.
[30] W. Wu, B. Walczak, D.L. Massart, S. Heuerding, F. Erni, I.R. Last, K.A. Prebble, Artificial neural networks in classification of NIR spectral data : design of the training set, Chemom. Intell. Lab. Syst. 33 (1996) 35–46.
[31] K. Rajer-Kanduč, J. Zupan, N. Majcen, Separation of data on the training and test set for modeling: a case study for modeling of five colours properties of a white pigment, Chemom. Intell. Lab. Syst. 65 (2003) 221–229.
[32] F. Marini, A.L. Magrì, R. Bucci, A.D. Magrì, Use of different artificial neural networks to resolve binary blends of monocultivar Italian olive oils, Anal. Chim. Acta 599 (2007) 232–240.
[33] A. De Juan, R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution, Anal. Chim. Acta 500 (2003) 195–210.
[34] F. Westad, M. Kermit, Independent component analysis, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, vol. 2, Elsevier, Oxford, UK, 2009, pp. pp.227–248.
[35] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 65 (1991) 1425–1432.
[36] R. Rajko, Comments on “near-infrared hyperspectral unmixing based on a minimum volume criterion for fast and accurate chemometric characterization of counterfeit tablets”, Anal. Chem. 82 (2010) 8750–8752.
[37] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171.
[38] L. Munck, L. Norgaard, S.B. Engelsen, R. Bro, C.A. Andersson, Chemometrics in food science – a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance, Chemom. Intell. Lab. 44 (1998) 31–60.
[39] S. Bellato, V. Del Frate, R. Redaelli, D. Sgrulletta, R. Bucci, A.D. Magrì, F. Marini, Use of near infrared reflectance and transmittance coupled to robust calibration for the evaluation of nutritional value in naked oats, J. Agric. Food Chem. 59 (2011) 4349–4360.
[40] G.E. Ritchie, Pharmaceutical Analysis/New Technology, Purdue Pharma L.P.,

444 Saw Mill River Road, Ardsley, NY 10502. Data set available at: http://www.idrc-chambersburg.org/ss20022012.html (Last accessed 02.01.2013).
[41] F. Westad, H. Martens, Variable selection in near infrared spectroscopy based on significance testing in partial least squares regression, J. Near Infrared Spectrosc. 8 (2000) 117–124.
[42] A. Bjørnstad, F. Westad, H. Martens, Analysis of genetic marker-phenotype relationships by jack-knifed partial least squares regression (PLSR), Hereditas 141 (2004) 149–165.
[43] Y. Ni, S. Chen, S. Kokot, Spectrophotometric determination of metal ions in electroplating solutions in the presence of EDTA with the aid of multivariate calibration and artificial neural networks, Anal. Chim. Acta 463 (2002) 305–316.
[44] C.V. Greensill, K.B. Walsh, Calibration transfer between miniature photodiode array-based spectrometers in the near infrared assessment of mandarin soluble solids content, J. Near Infrared Spectrosc. 10 (2002) 27–35.
[45] A. Kazeminy, S. Hashemi, R.L. Williams, G.E. Ritchie, R. Rubinovitz, S. Sen, A comparison of near infrared method development approaches using a drug product on different spectrophotometers and chemometric software algorithms, J. Near Infrared Spectrosc. 17 (2009) 233–245.
[46] O. Preisner, J.A. Lopes, J.C. Menezes, Uncertainty assessment in FT-IR spectroscopy based bacteria classification models, Chemom. Intell. Lab. 94 (2008) 33–42.
[47] P. Murtaugh, Performance of several variable-selection methods applied to real ecological data, Ecol. Lett. 12 (2009) 1061–1068.
[48] J. Reunanen, Overfitting in making comparisons between variable selection methods, J. Mach. Learn. Res. 3 (2003) 1371–1382.
[49] H. Martens, M. Martens, Multivariate Analysis of Quality. An Introduction, John Wiley and Sons, New York, NY, 2001.
[50] R. Leardi, Application of genetic algorithm–PLS for feature selection in spectral data sets, J. Chemom. 14 (2000) 643–655.
[51] G.P.E. Box, Teaching engineers experimental design with a paper helicopter, Qual. Eng. 4 (1992) 453–459.

Federico Marini is researcher in Analytical Chemistry at the University of Rome "La Sapienza". His research activity is focused on all aspects of chemometrics, ranging from the application of existing methods to real world problems in different fields (food science, cultural heritage, drug analysis, environment, -omics disciplines and so on) to the design and development of novel algorithms (particularly in the field of classification and multi-way analysis). In 2012 he won the Elsevier Chemometrics and Intelligent Laboratory Systems Award "for his achievements in chemometrics". He is author of more than 70 papers in international peer-reviewed journals, and recently he edited and coauthored for Elsevier the book "Chemometrics in food chemistry".

Frank Westad received his M.Sc in Chemistry and Data analysis in 1988, and completed his Ph.D thesis "Relevance and parsimony in multivariate modelling" in 2000. His working experience includes Senior Research Scientist positions at GE Healthcare and the Norwegian Food Research Institute. He has published 35 papers and has also written chapters for chemometric textbooks. Over the years he has worked in projects involving Design of Experiments and chemometrics for a number of industries. He is now holding the position as Chief Scientific Officer at CAMO Software.