Special Issue Article

Journal of CHEMOMETRICS

Received: 24 January 2010, Revised: 21 February 2010, Accepted: 23 February 2010, Published online in Wiley InterScience: 2010

(www.interscience.wiley.com) DOI: 10.1002/cem.1310

Principles of Proper Validation: use and abuse of re-sampling for validation

Kim H. Esbensen[a]* and Paul Geladi[b]

Validation in chemometrics is presented using the *exemplar* context of multivariate calibration/prediction. A phenomenological analysis of common validation practices in data analysis and chemometrics leads to formulation of a set of generic Principles of Proper Validation (PPV), which is based on a set of characterizing distinctions: (i) Validation cannot be understood by focusing on the methods of validation only; validation must be based on full knowledge of the underlying definitions, objectives, methods, effects and consequences—which are all outlined and discussed here. (ii) Analysis of proper validation objectives implies that there is one valid paradigm only: test set validation. (iii) Contrary to much contemporary chemometric practices (and validation *myths*), cross-validation is shown to be unjustified in the form of monolithic application of a one-for-all procedure (segmented cross-validation) on all data sets. Within its own design and scope, cross-validation is in reality a sub-optimal *simulation* of test set validation, crippled by a critical sampling variance omission, as it manifestly is based on one data set only (training data set). Other re-sampling validation methods are shown to suffer from the same deficiencies. The PPV are universal and can be applied to all situations in which the assessment of performance is desired: prediction-, classification-, time series forecasting-, modeling validation. The key element of PPV is the Theory of Sampling (TOS), which allow insight into all variance generating factors, especially the so-called incorrect sampling errors, which, if not properly eliminated, are responsible for a fatal inconstant sampling bias, for which no statistical correction is possible. In the light of TOS it is shown how a second data set (test set, validation set) is critically necessary for the inclusion of the sampling errors incurred in all 'future' situations in which the validated model must perform. Logically, therefore, all one data set re-sampling approaches for validation, especially cross-validation and leverage-corrected validation, should be terminated, or at the very least used only with full scientific understanding and disclosure of their detrimental variance omissions and consequences. Regarding PLS-regression, an emphatic call is made for stringent commitment to test set validation based on graphical inspection of pertinent *t–u* plots for optimal understanding of the *X–Y* interrelationships and for validation guidance. QSAR/QSAP forms a partial exemption from the present test set imperative with no generalization potential. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: Principles of Proper Validation (PPV); future performance assessment; test set validation; cross-validation; re-sampling; predictive regression; Theory of Sampling (TOS)

1. INTRODUCTION

Irregularly, but with certainty, discussions break out within chemometrics as to validation—as to *proper* validation. There are few other topics within chemometrics which always lead to personal *opinions*. At times such discussions have been serious, broad-ranging and informative for both seasoned and new participants, while at other times closer to borderline emotional. However, the matter at hand is exclusively a scientific one and the discussion format should be restricted to the proper form and traditions in this arena.

In chemometrics, validation is probably most well known in the context *prediction validation* of which there are (at least) four types: test set validation, cross-validation, 'correction validation' (leverage correction is the prime example), and re-sampling methods (bootstrap, jackknife, Monte Carlo simulation, permutation testing). This tutorial illustrates the central tenets of the Principles of Proper Validation (PPV) by a detailed analysis of prediction validation in the specific *multivariate calibration context*. PPV is a set of general principles which can also be applied e.g. with respect to validation of other data analytical or statistical methods which needs performance testing, e.g. modeling by neural networks, classification evaluation or times series forecasting.

This tutorial does not present the reader with a complete catalog of all the many validation *practices* offered in the statistics and data analysis literature; instead a minimum set of the necessary general phenomenological validation characteristics will suffice. The overlying PPV are concerned with the question of how to establish a *proper* validation that is not derailed by the very many, very different, specific data structures met within data analytical modeling. A major result of the present analysis is that there is only one optimal

* *Correspondence to: K. H. Esbensen, ACABS Research Group, University of Aalborg, Campus Esbjerg, DK-6700 Esbjerg, Denmark.*
E-mail: kes@aaue.dk

a *K. H. Esbensen*
ACABS Research Group, University of Aalborg, Campus Esbjerg, DK-6700 Esbjerg, Denmark

b *P. Geladi*
Unit of Biomass Technology and Chemistry, Swedish University of Agricultural Sciences, SE 90183, Umeå, Sweden

Copyright © 2010 John Wiley & Sons, Ltd.

approach that fulfills all the demands of proper validation, test set validation.

proper *adj.* - adapted or appropriate to the purpose or circumstance
valid *adj.* - sound; just; well-founded; - producing the desired result
validate *v.t.* - to make valid; substantiate; confirm

One reason for much of the often deeply felt differences-of-opinions regarding what constitutes proper validation relates to the fact that validation involves both statistical issues and chemical, physical, data analytical, and physical sampling error issues. A significant proportion of the historical debate simply reflects a too restricted point of view—for example that validation is exclusively a statistical issue, i.e. 'sampling' is simply a matter of drawing from a population of independently sampled i.i.d. measurements (*objects* in the data analysis parlance), which can be called statistical sampling—it most emphatically is not, as argued below. In the present paper a broader, holistic understanding of the compound sampling, analysis and validation issue is advocated, while taking care not to fall into the opposite, equally simplistic position, *viz* that all data matrices result from sampling from heterogeneous material (an unfortunate misinterpretation of several of the papers on the Theory of Sampling (TOS) presented within chemometrics in recent years). However, the gamut of data analysis and data modeling indeed do occupy a realm in which one must always assume the presence of significant sampling errors, which if neglected, will cause grave prediction and validation problems. The rarer cases in which pure statistical sampling complies are simply to be treated in an identical fashion allowing for unity in all validation endeavors.

1.1. Data quality–data representativity–sample representativity

The PPV are introduced with a few discussion points related to the concept of *data quality* and *data representativity* leading to the fundamental issue of *sample representativity*.

Data quality is a broad, but often only loosely defined term; any definition that does not include the specific aspect of data representativity is suboptimal however. This statement is based on an extensive body of experience and literature regarding the 'TOS' [1–23], which is used below in the argumentations *pro* proper validation and *con* cross-validation, most similar re-sampling methods as well as leverage corrected validation. Appendix A gives a *brief* of the principles in the TOS as presented by the selected literature.

The term 'data' is often equaled with 'information'. It is obvious, however, that this can only be in a latent, potential form. It takes data analysis with appropriate, problem-context interpretation to reveal the 'information' residing in e.g. data matrices. In chemometrics the prime interest is very often on data analysis, while issues pertaining to the prehistory of a data table usually receive but scant attention: 'Chemometricians analyse the data . . .'.

One exception is Martens and Martens [19] which addresses: 'multivariate analysis of quality', where the focus is stated to relate to the 'quality of information', which is defined as '. . . dependent on reliability and relevance'. However, *reliability* and *relevance* are open-ended, very general adjectives which, if to be used unambiguously, must be given a specific meaning from the problem context at hand. In a series of recent contributions inducting the TOS into chemometrics, it was argued that a far more relevant characteristicon is *representativity* [1,6,12,14,15,20], because a precise definition is at hand (qualitative as well as quantitative), but mainly because the specific definition in TOS allows for comprehensive understanding of the underlying phenomenon of *heterogeneity*.

Against this backdrop, all data analysis contexts include (at least) a sampling issue, the analytical issue and the data analysis issue. From our analysis of the use and abuse of cross-validation below, it will become clear that 'reliable analytical information' can only be based on *representative samples*. Any valid definition of data quality, therefore, in principle must include some reflections on both representativity and sampling in addition to chemico-physical analysis/measurement before the data analytical issues. Below, we shall introduce a critical distinction between statistical sampling (in the conventional statistical context) and the kind of physical sampling addressed by the TOS. It is mandatory to be competent with respect to both these aspects of 'sampling'.

It is necessary to contemplate the specific origin of any data set, before concentrating on the interesting data structure modeling and interpretation issues, including validation. There may be large, significant, or only small sampling issues involved, the point being that this issue is unknown, and therefore cannot be dismissed *a priori*. In chemometrics, the type of errors colloquially known as 'measurement errors' typically relate to the *X*-data, for example in the form 'instrumental signal errors', but of course also refers to all analytical errors pertaining to the 'reference measurements' (*Y*-data in calibration). These effects are incorporated into the Total Analytical Error (TAE). There exists an extensive experience and literature on representative sampling which clearly shows that the physical sampling errors often dominate compared with the strict analytical and data modeling error effects. An often quoted comparison states that the Total Sampling Error (TSE) typically ranges 10–50–100 × *larger* than the TAE [1–4,7–13,15]. In the discussions pursued in the present paper it is imperative to break the current habit of blatantly disregarding TSE, because of this quantitative dominance. By dealing universally with these issues as if sampling issues *were* always significant, all cases can be treated identically in a rational and efficient manner, covering all combinations of large and/or small statistical errors as well as large/small TOS-sampling errors.

The opposite position (very often met with) is that of assuming that all sampling errors at all times are insignificant. This constitutes an illegitimate generalization however, which is untenable; no general proof of this widespread 'assumption' has ever been presented. This attitude is but a longstanding misunderstanding within chemometrics, statistics and data analysis, as demonstrated by a compelling body of evidence in current validation practice and in the extensive literature [25–72]. We deal with these issues in full detail below. From this perspective, chemometric data analysis without sufficient attention to the full context of relevant pre-data table issues and considerations (physical sampling, statistical sampling) cannot be considered comprehensive; indeed it is incomplete.

The issues regarding validation are not about opinions (personal, institutional), nor about following one or other established schools-of-thought or traditions (thereby trying to dodge a personal responsibility for method selection). All validation issues are fully tractable and lend themselves to

Copyright © 2010 John Wiley & Sons, Ltd.

analysis, rational discussion and sound, objective and impartial conclusions.

2. VALIDATION OBJECTIVES

Validation, in the *exemplar* multivariate calibration context,[†]. means assessing or substantiating that the prediction performance is *valid*, i.e. the objective of validation is to confirm that a particular prediction model will work *according to its purpose*.

This objective does not only refer to the contemporary calibration/modeling situation, but also to the circumstances surrounding the future performance for new 'similar data'. Both the training and the validation data must be 'similar'—emphatically not to one another, as is a prevalent current misunderstanding, but to those new data sets pertaining to the 'future' working situation of the model (delineation of the necessary criteria for 'similarity' is fully elaborated below).

Thus already when designing and selecting a training data set for modeling (calibration) it is imperative also to pay attention to how the model is to be validated. Preferentially one should always be in a position to be able to choose at least one second data set for validation, here generically called a *test set*, with which to represent the future working situation of the particular data model. All prediction models must be validated w.r.t. *realistic* future circumstances. In data analysis, statistics and chemometrics some 10 years ago there was a somewhat rude awakening to the fact that far too little prediction validation was on the agenda. In Höskuldsson's (1997) [17] reassessment of the entire realm of 'Prediction Methods in Science and Technology', it was described how modeling fit assessment dominated as compared to the necessary complementary prediction validation, for which the H-principle of balanced assessment/validation was promulgated. Today there is a much more widespread awareness that modeling fit optimization is a necessary, but not a sufficient, criterion for prediction performance.

3. TEST SET VALIDATION—A NECESSARY AND SUFFICIENT PARADIGM

The central theme of the present foray can be stated in unambiguous terms: All other validation methods are but *simulations* of test set validation.

> **simulate** *v.t.* - to *assume* or have the *appearance* of characteristics of . . .. . .

There exists one universally applicable validation method, which apparently in principle and practice can always be carried out under all prevailing physical, economical, resource allocation constraints—namely the popular cross-validation method, no doubt in large measures because one only needs one data set, the training data set, X_{train}. It will be shown below, however, that cross-validation is always sub-optimal: cross-validation is structurally, by its own design purpose, never able to achieve all the necessary objectives of validation.

On the other hand, the objectives of test set validation are always structurally correct and complete (fully discussed below). If a proper test set *were* always obtainable, no other validation procedure need ever have been introduced; test set validation would then be the only validation method in existence. A full substantiation of the above summary positions follows.

3.1. Validation in data analysis and chemometrics

Internal validation can be used for many different purposes. Below we discuss both legitimate and illegitimate approaches to validation. This tutorial is focused on multivariate calibration for prediction purposes. Cross-validation as used on one data matrix (X) only, e.g. PCA, MCR, PARAFAC is not covered *per se*, but most aspects of the analysis, discussion and conclusions from the prediction scenario can be carried over to these application contexts as well without loss of generality.

4. HISTORY OF CROSS-VALIDATION/ RE-SAMPLING

Already in 1931, Larson [25] observed that the data used in building a model are not good for testing its quality. The earliest specific mention of predictive cross-validation is Lachenbruch 1965 [26], a PhD thesis from which one article was published [27]. Part of the history and use of cross-validation in sociology is documented in a review paper [28]. In the historical perspective the seminal papers on cross-validation in regression were those by Stone and Geisser, independently published in 1974–1975 [29–32]. The paper by Stone [29] includes a long discussion part in which many prominent statisticians of the time give praise and criticism regarding the cross-validation *concept*. One of the experts commenting on the Stone paper of 1974 compares the presented material with Uri Geller's exploits on television, but the paper is nevertheless a statistical masterpiece in its own right.

Stone, Geisser and their contemporaries were well aware of the limitations of their proposed methods and they warned their readers against frivolous use of them. In 1977, the Stone [33] paper on asymptotic properties appeared in which an attempt was made to find out where cross-validation would work and where not. Examples of applied cross-validation activity in the statistics literature are Wold [34], Bowman [35], Picard and Cook [36], Li [37], and Burman [38]. Cross-validation has been used e.g. for density estimation, model comparison, time series analysis, latent variable selection and many more applications in sociology, psychology, medicine etc. The uses in chemometrics form merely a subset of the much broader statistical and data analytical scene. For parameter and density estimation, also bootstrapping and the jackknife were introduced [39,40]. It is important to note that calibration models made by Artificial Neural Networks (ANN) have made extensive use of cross-validation [41] as well.

After the introduction of Partial Least Squares (PLS) regression in the late 1970s [42], chemometricians quickly learned to use PLS regression for making calibration models and using these models for prediction [43–46]. There was a distinct need for a technique for selecting the optimal number of PLS components (optimal model complexity) and for quantitative determination of the predictive performance. Martens *et al.* [47] give the advice of using cross-validation or jackknife to avoiding over-fitting. Ståhle and Wold [48], Haaland and Thomas [49], and Osten [50] are

[†] Exemplar—here used in the sense of Kuhn (1969): 'The Structure of Scientific Revolutions' [25].

among the first to take up a systematic study of this possibility by cross-validation. Later, the textbook [51] introduced cross-validation as a *de facto* standard in multivariate calibration. From then on, the use of cross-validation in all its forms and variants proliferated in chemometrics, mainly because software companies were soon to offer it as a default mode of validation. It is clear that one reason for the widespread popularity of cross-validation is related to its ease of implementation and programming.

5. CROSS-VALIDATION IN CHEMOMETRICS

Cross-validation can be used for ordinary least squares modeling, but then there is no need for estimating how many components should be used. The main use of cross-validation in chemometrics is and has been for estimating the number of components to be used for prediction in latent variable models such as PCR and PLS. As a bonus, an average estimate of the prediction error would *appear* to be at hand as well.

A regression model by PLS (similarly PCR) can be written as follows:

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{\mathrm{r}} + \mathbf{f}_{\mathrm{r}} \tag{1}$$

where $\mathbf{y}$ is the mean-centered vector $(I \times 1)$ of responses; $\mathbf{X}$ is the mean-centered matrix $(I \times K)$ of predictor variables (training data $\mathbf{X}_{\mathrm{train}}$); $\mathbf{b}_{\mathrm{r}}$ is the vector $(K \times 1)$ of regression coefficients; $\mathbf{f}_{\mathrm{r}}$ is the residual vector; $r = 1, \ldots, R$ is a counter for model rank (pseudorank, model dimensionality), and R is the true mathematical rank of $\mathbf{X}$.

One may define The Sum of Squares of the residual $\mathrm{SS}_{\mathrm{res,r}}$ as follows:

$$\mathrm{SS}_{\mathrm{res.r}} = \mathbf{f}_{\mathrm{r}}'\mathbf{f}_{\mathrm{r}} \tag{2}$$

Historically, the primary role of (cross) validation was to find the optimal value for r, r_{opt}. Under-fitting (r too low) results in a high modeling bias, while over-fitting (r too high) leads to an inflated prediction variance. The optimal number of components is the one which gives the best *combination* of low bias and small variance [51,52]. It is clear that there is not a universal solution to such a balancing act—there is a very strong relationship to the specific data structure particularities. The model built in Equation (1) is used for the prediction of new responses, $\mathbf{X}_{\mathrm{new}}$:

$$\mathbf{y}_{\mathrm{hat,r}} = \mathbf{X}_{\mathrm{new}}\mathbf{b}_{\mathrm{r}} \tag{3}$$

where $\mathbf{y}_{\mathrm{hat,r}}$ is the vector of predicted response values, mean-centered with the mean of $\mathbf{y}$ (Equation 1) and $\mathbf{X}_{\mathrm{new}}$ is the matrix $(J \times K)$ of new X-data, different and independent from $\mathbf{X}$ and mean-centered with the same mean values as $\mathbf{X}$ (Equation 1).

$\mathbf{X}_{\mathrm{new}}$ may also be sometimes called $\mathbf{X}_{\mathrm{test}}$. In the conventional statistical context, $\mathbf{X}_{\mathrm{new}}$ is traditionally viewed as coming from the same *population* as the training set. New objects, new measurements are viewed as i.i.d. measurements. This view is distinctly different from that based on a comprehensive TOS-based understanding of 'new measurements' as originated by a compound sampling-and-analysis process (refer to Appendix A).

In order to test the quality of the prediction, *some* true y values corresponding to the objects in $\mathbf{X}_{\mathrm{new}}$ have to be known; these are often called reference measurements in the validation context. This allows the calculation of a prediction residual Sum-of-Squares:

$$\mathrm{SS}_{\mathrm{pre,r}} = (\mathbf{y}_{\mathrm{true}} - \mathbf{y}_{\mathrm{hat,r}})'(\mathbf{y}_{\mathrm{true}} - \mathbf{y}_{\mathrm{hat,r}}) \tag{4}$$

where $\mathbf{y}_{\mathrm{true}}$ is the reference value of the response, sometimes called the 'true value' and $\mathrm{SS}_{\mathrm{pre,r}}$ is the Sum-of-Squares prediction residual for an r-component model.

The most often used prediction performance statistic is

$$\mathrm{RMSEP}_r = [\mathrm{SS}_{\mathrm{pre,r}} J^{-1}]^{1/2} \tag{5}$$

where $\mathrm{RMSEP}_{\mathrm{r}}$ is the Root Mean Square Error of Prediction for an r-component model.

Dividing by J makes comparison between test sets of different sizes possible and taking the square root gives the RMSEP the same units of measurement as the responses, which is often convenient. RMSEP can advantageously be expressed also in a relative manner [%], for example with respect to the average $\mathbf{y}_{\mathrm{true}}$ level.

Plotting $\mathrm{SS}_{\mathrm{pre,r}}$ or $\mathrm{RMSEP}_{\mathrm{r}}$ as a function of r often gives an indication of which value of r gives the 'best model' i.e. the model complexity that results in the lowest prediction error. This is a kind of scree plot, but one with a clear minimum or a low plateau. In either event, it is easy to use this graphical illustration for visual inspection, arriving at r_{opt} with a degree of inter-personal robustness apparently acceptable to everybody. One should realize that r_{opt} often is a fantasy number. There is usually a range of values that for all practical matters are equally good within existing uncertainty bounds.

The average bias can be calculated as

$$\mathrm{Bias}_{,\mathrm{r}} = \mathbf{1}'(\mathbf{y}_{\mathrm{true}} - \mathbf{y}_{\mathrm{hat,r}})J^{-1} \tag{6}$$

where $\mathbf{1}$ represents a vector of ones size $(J \times 1)$.

Assuming a data set, $\mathbf{X}$ consisting of I objects, the different types of (cross) validation proposed in the statistical literature are as follows:

(1) Hold-out: splitting the data in a calibration (C objects) and a test set (T objects): $I = C + T$.
(2) Leave-one-out: each object is left out once; this requires I models to be made [36]. This approach is well known by the name 'Leave-One-Object' out (LOO). This has also been given the unfortunate name 'Full Cross-validation'. This variant is most related to Jackknifing, but can also be found within chemometrics, often then misinterpreted as a particularly strong validation—it most emphatically is not (it is the most slack validation method in existence, see further below).
(3) Leave M out: where all possible subsets of M objects are left out once. This requires $I![(I-M)!]^{-1}[M!]^{\mathbf{-1}}$ models to be made [53]. This variant is mostly related to Bootstrapping, where the same object can appear more than once in the left-out part.
(4) V-fold: here the data set is partitioned into I/V parts and I/V models are made [31]. Segmented validation is a useful cover name for this and related approaches in which one leaves out a fraction of I (a segment).
(5) Monte Carlo based cross-validation; a huge number of subsets of different sizes for holding out is created, accepting any resulting number of repeat objects included in the modeling data basis. In this approach it is the sheer number of (many) subsets, that is supposed to derive the desired information.

The V-fold and LOO methods have been very popular in chemometrics, precisely because they are fast in calculation and easy to implement. Equations (1)–(4) describe a hold-out procedure, but can also be modified to give a V-fold cross-validation procedure. For example, by putting all available data (also $\mathbf{X}_{\mathrm{new}}$) in $\mathbf{X}$, using LOO and V-fold are very easily implemented. In these cases, Equation (3) is used repeatedly for the left-out objects, Equation (1) is used repeatedly on the left-in objects and Equation (4) is used for accumulating results obtained for the left-out objects. Similar to Equation (5), a new definition can be made:

$$\mathrm{RMSECV_r} = \left[\mathrm{SS_{pre,r}} I^{-1}\right]^{1/2} \tag{7}$$

where $\mathrm{RMSCV_r}$ is the Root Mean Square Error of Cross Validation for the r-component model and $\mathrm{SS_{pre,r}}$ is the accumulated sum of squares for the left-out parts.

In summary, cross-validation is typically used to estimate at least two different parameters of the model: (1) r_{opt} and (2) $\mathrm{RMSECV_r}$ obtained for r_{opt}. In the context of the present discussion, the first objective is related to what is often referred to as *internal validation*, while the second is supposed to be related to the *external validation*. A point in the present analysis is that the latter is structurally impossible however.

5.1. Sequential component assessment

Some chemometrics approaches are based on only testing the new, added part when model dimensionality is increased. This is based on the technique from a Wold [34] paper where this was used for PCA models. One basically tests the ratio [54]:

$$R = \mathrm{SS_{pre,r+1}}/\mathrm{SS_{res,r}} \tag{8}$$

where $\mathrm{SS_{pre,r+1}}$ as in Equation (7) is used.

This ratio is demanded to be less than 1 (or 0.95) if the added component can be said to improve the predictive capability of the model. There are confusingly many stopping rules for deciding for which value of components this actually happens. The use of the sequential testing is, however, based on outdated algorithms which calculated principal or PLS components one after another.

5.2. Leverage corrected residuals

Leverage corrected residuals were introduced a long time ago when computers were too slow for cross-validation on *large(r)* data sets. In Equation (1) it is always possible to make the residual part smaller by making r larger. In order to avoid abuse of this property, one may multiply the residuals by a penalty that increases as r goes up. Leverage can reach 1.0 [at the limit] as r increases, so dividing by $[1\text{-leverage}]^{\alpha}$ blows up the residuals if r becomes too large [51]. The theoretical background is a t-test for residuals [55]. Leverage-corrected validation no longer has any serious function.

5.3. Permutation tests

Permutation tests were introduced in chemometrics by Sergio Clementi [56].

A specific approach which can be deployed towards object selection and chance correlations in prediction has seen wide recent application in chemometrics; this approach applies permutation(s) to the response variable $\mathbf{y}$, i.e. randomization of $\mathbf{y}$ [57]. In this approach the ordering of the response vector objects is randomized while the descriptor matrix, $\mathbf{X}$ stays stable with its original ordering. Multivariate calibration models should now be statistically insignificant; t- or F-tests can be employed to discriminate chance fluctuation from real correlation. It is clear that this approach constitutes a very useful check against chance correlations, but is not applicable to assess anything regarding the *future* prediction performance.

5.4. Cross-validation in recent chemometrics history

The questions about how to assess how many components should be used in an optimal PLS model, r_{opt}, and how good the model then becomes are among the most discussed issues in chemometrics.

During the 1990s, cross-validation was established in chemometrics mainly by the proliferation of software packages. Some interesting papers specifically dealing with cross-validation appeared in the literature. Cruciani *et al.* [58] introduced the Standard Deviation of Errors in Prediction (SDEP) parameter, which is RMSECV for a hybrid between cross-validation and bootstrapping. They give an illuminating discussion on the good and bad qualities of the leave-one-out method and show graphically that different subset sizes in the V-fold method have an effect on SDEP. They show this for a number of food chemistry and QSAR examples. The follow-up paper by Baroni *et al.* [59] is about the use of SDEP for variable selection.

Wakeling and Morris [60] introduce a significance test for scree plots used for selecting the number of PLS components. The authors use Monte Carlo simulation to estimate distributional properties of a coefficient of determination, found by cross-validation. Forina *et al.* [61] test the use of validation in near infrared calibration. They compare single evaluation set (SES), cross-validation (CV) and repeated evaluation set (RES). The conclusion is that SES gives unstable results, CV is better and RES is best, but they also mention that bad choices of training and test set and unfamiliarity with the data can be the real source of the differences. Eriksson *et al.* [62] present an excellent paper on validation in QSAR modeling. They mention external validation, cross-validation, permutation testing and graphical methods for checking model and residual. For hold-out (external) validation the authors mention that a proper selection of training and test sets has to be done and that this may be difficult in QSAR situations. Therefore, a combination of permutation tests and cross-validation is the only possible way to go but they also stress checking other model properties than just RMSECV. They show some results using example data sets.

Denham [63] tries to estimate prediction intervals in PLS modeling by bootstrapping and cross-validation in cases where the number of PLS components need not be estimated. Wehrens and van der Linden [64] try to explain bootstrapping as applied to Principal Component Regression (PCR). Among the many subjects handled is variable selection and the comparison of models using all variables with those using only selected ones. Martens and Dardenne [65] use Monte Carlo simulation in a large database (>900 near infrared spectra of maize and protein concentrations). The latter paper is written in a confusing fashion and the conclusions are unfortunately vague. Denham [66] is a follow-up of Denham [63]. This second paper is on the distribution of errors in PLS models. The author compares an analytical approach to re-sampling. The examples show that full cross-validation and bootstrapping are equally good as the analytical approach.

Copyright © 2010 John Wiley & Sons, Ltd.

Other recent papers give good overviews of the validation literature [57,67,70]. Wiklund *et al.* [57] give a comprehensive introduction to the re-sampling situation in chemometrics. They also describe two different ways of doing cross-validation: one model at a time and one component at a time. They show results for a number of data sets and introduce a permutation test. The paper by Filzmoser *et al*. [67] gives a lucid introduction to the cross-validation and test set calibration and validation principles used in their 'repeated double cross validation' approach, but the proposed method still tries to extract too much information from too little data. This approach is discussed in more detail below.

A conclusion from the papers referred to above is that re-sampling methods can be used for very many objectives: (1) selection of a number of PLS components, (2) estimation of an RMSECV for the PLS model, (3) estimation of a distribution and confidence intervals around the RMSECV, (4) comparison of all-variable and variable-reduced models [68] (also called cross-model validation). This easily leads to confusion and certainly a huge overfit as re-sampling is used for all the mentioned purposes based on a *small(er)* data set, as is often the case. Bootstrapping (and Monte Carlo) seems to be just a cover-up creating the *illusion* of getting more out of the data than what is in fact there. Another observation is that many authors seem to be happy about writing validation algorithms than offering comprehensive analysis of the data structures and their impacts on the otherwise impressive algorithms.

Kohonen [69], in a recent thorough overview, remarks that full cross-validation (often termed LOOCV) is generally seen as the universal standard and refers to Gómez-Carracedo *et al*. [70] as the progenitor for this sweeping claim. Kohonen discusses the many historical uses of PRESS (Predicted Residual Sum-of-Squares).

A related approach is based on PRESS for the component $a + 1$, which is compared with the Residual Sum-of-Squares for the total of the preceding *a* components, $RSS_{a.}$ The minimizing criterion is the so-called $\mathbf{R_R}$, defined in the following fashion:

$$\mathbf{R_R} = PRESS_{a+1}/RSS_a$$

Kohonen [71] traces the chemometrics validation history in significant detail; in the discussion below use is made of this very useful overview.

5.5. Models and Monte Carlo simulation

An extreme version of cross-validation is Monte Carlo simulation where very many (thousands or even millions) re-samplings are carried out. Based on the particular scenario investigated and the particular model(s) studied, this approach can be warranted or not. Some examples will explain this. A similar view on models for calibration is also given in Varmuza and Filzmoser [71].

5.5.1. Fundamental models

The laws of physics and the properties of atoms are universal; one often speaks of 'hard models'. Model testing is not needed in this regimen. As an example one can produce a mixture or known composition, particle size, thickness and density and simulate irradiating it with X-ray photons from a well-defined energy distribution. This then allows calculation of very many resulting X-ray fluorescence spectra by Monte Carlo simulation and statistical conclusions about the obtained spectral population can be made. See an example in Czyzycki *et al*. [72].

5.5.2. Unique models with lots of underlying data

As an example, in Process Analytical Technology (PAT) one monitors, studies and controls industrial processes, either *in toto* or with a focus on particularly interesting processing units. An example of using Monte Carlo for this purpose is Sin *et al*. [73] PAT systems are in principle *unique*; no two industrial processes are sufficiently so identical (even if they perform the same task) that they can be treated as case 1 above. Industrial processes are especially also varying over time because of drift, upsets, outlying (unique/no longer representative) samples/measurements or, more fundamentally, because parts erode or wear out, causing drifts/upsets etc.—and eventually have to be replaced. Because of these facts, well known in all PAT sectors, a unique model has to be made for each individual process, and very nearly always have to be *updated* at regular intervals. It should be mandatory also to perform a suitable validation, for every new model updating . . ..

Re-sampling and Monte Carlo approaches can be used for the more narrow model building purpose, usually because a large population of past process measurements exists. If the population of available data is large enough, the hold-out approach is a natural idea. This case cannot be generalized to the future prediction performance validation case however.

5.5.3. Unique models with limited and/or unreliable data

In Quantitative Structure Activity Relationships (QSAR) a rather limited group of molecules are tested, say for toxicity (e.g. by measuring LD50 on some bacteria strain or similar), see Venkatapathy *et al*. [74] for an example. Such LD50 measurements are invariably imprecise. Whatever is done, the relationship between the molecular descriptor indices and properties and LD50 values is often weak. Performing Monte Carlo simulations in order to test prediction performance in such systems is obviously close to meaningless.

Bootstrapping may be compared to Monte Carlo simulation but is less computer-intensive. A simple example can explain this. Assume a certain number of playing cards are available, say: 1 (ace) to 10 of diamonds. This is a finite subpopulation of 10 cards. With these 10 cards all subsets of 5 (poker hands), as an example, can easily be simulated. There are 252 unique such subsets. The means or medians of all these subsets form an impressive histogram, but still there are only 10 playing cards whose sub-distribution is 'analyzed'. None of the other 42 playing cards ever enters this particular bootstrapping application. The key question to be answered here is: which population is studied: the original population of 52 cards?—or the subpopulation of 10 selected cards?—or the population of 252 poker hands made from the subpopulation? Is using 10 cards a legitimate option for making inferences for the entire deck? This would obviously be nonsensical—but this is exactly what re-sampling is doing in most software implementations. If 'ten cards' were replaced by 'ten measured samples', every scientist in his or her right mind would agree that there are only ten samples and that using these as representing the entire lot can only be done based on full understanding of all pertinent issues (lot heterogeneity issues, sampling process errors etc.). Another question that may be asked is what happens to the degrees of freedom when very many models are made. Are there degrees of freedom left for judging a residual standard deviation?

5.6. Discussion: re-sampling for proper validation

Kohonen (2009) [65] presented to the communities concerned about proper validation a most useful comparison of pretty well

Copyright © 2010 John Wiley & Sons, Ltd.

all of the most used validation approaches specified above; in particular he compares RMSECV(LOO), 10-segmented RMSECV, Wold's R (0.95 cut-off limit), R_R and MCCV with RMSEP (test set validation). MCCV is a much used statistic pertaining to these scenarios (Monte Carlo Cross-validation).

Regarding one-and-the-same well selected training data set used for comparison, including a specific test set also produced during the investigations, r_{opt} for these six central validation approaches comes out as 10, 10, 9, 5, 10, and 4 components respectively. In this example cross-validation is not robust, cross-validation approaches by comparison tend to over-fit (test set validation points to four components).[‡] While based on only one data set, there are important implications or more than local interest, as this example echoes very many similar experiences accumulated over decades both within chemometrics and beyond. The key issue is that each data set is unique in the sense of its inherent more-or-less heterogeneous distribution of data objects. 'Competing' validation approaches often end up with similar significantly different results (regarding r_{opt} as well as RMSE estimates). After his extensive comparisons, Kohonen remarks (2009: p. 45) [65]: 'No (validation) method . . . compares to [the] usage of an independent data set', which is in full agreement with the tenor of the present treatment. The comparison above is of course strictly speaking only valid for the specific data set involved. It will always be possible to disagree with the generality of conclusions based on only one data set. No strong claims can be made as to 'all', 'most', or 'many' data sets based on one particular data set only. The validation literature is ripe with the desire for generalizations based on one, or two specific data sets, examples are *legio*, but always remain illegitimate generalizations.

The universal point here is, however, that particulars cannot stand in for general relationships, except in one sense: Only the specific type of data structure present in any given data set may serve as a basis for more general conclusions—if proper caution is executed.

In the present work we aim to derive conclusions, that can be generalized, regarding re-sampling from the point view of primary validation principles only.

5.7. Data structure display via T–U plots

A classical formulation [51] of the PLS algorithm based on Equation (1) is as follows:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 \tag{9}$$

$$\mathbf{u}_1 = \mathbf{y}q_1 \tag{10}$$

where $\mathbf{t}_1$ is the first PLS X-score; $\mathbf{w}_1$ is the first PLS X-weight; $\mathbf{u}_1$ is the first PLS y-score; q_1 is the first PLS y-loading, and $\mathbf{X}$ and $\mathbf{y}$ are mean-centered as defined in Equation (1).

Equations (10) and (11) are given here for only the first component, but $\mathbf{t}_2$ and $\mathbf{u}_2$, $\mathbf{t}_3$, and $\mathbf{u}_3$ etc. can be easily calculated as well. Plotting $\mathbf{u}_a$ against $\mathbf{t}_{a,}$ $a = 1,\ldots, A$ is a reflection of the so-called 'inner relationship'. This 't–u plot' is a very useful vehicle for visual check of whether a possible next component is meaningful or not, as evidenced from the 'inner' partial regressions. Although the PLS algorithm *can* be written without using Equations (9) and (10), it is always informative to assess the data structure by t–u plots.

In order to be able to take proper action with respect to the actual data structure present in training-, test-, or 'future' data sets, a typology of the principal types of data structures associated with multivariate calibration is presented in Figure 1. There are three underlying *'parameters'* characterizing the particular manifestations of any multivariate data structure: (i) the number of objects N, (ii) the degree of linear (or nonlinear) correlation present, and (iii) data clustering, grouping (data

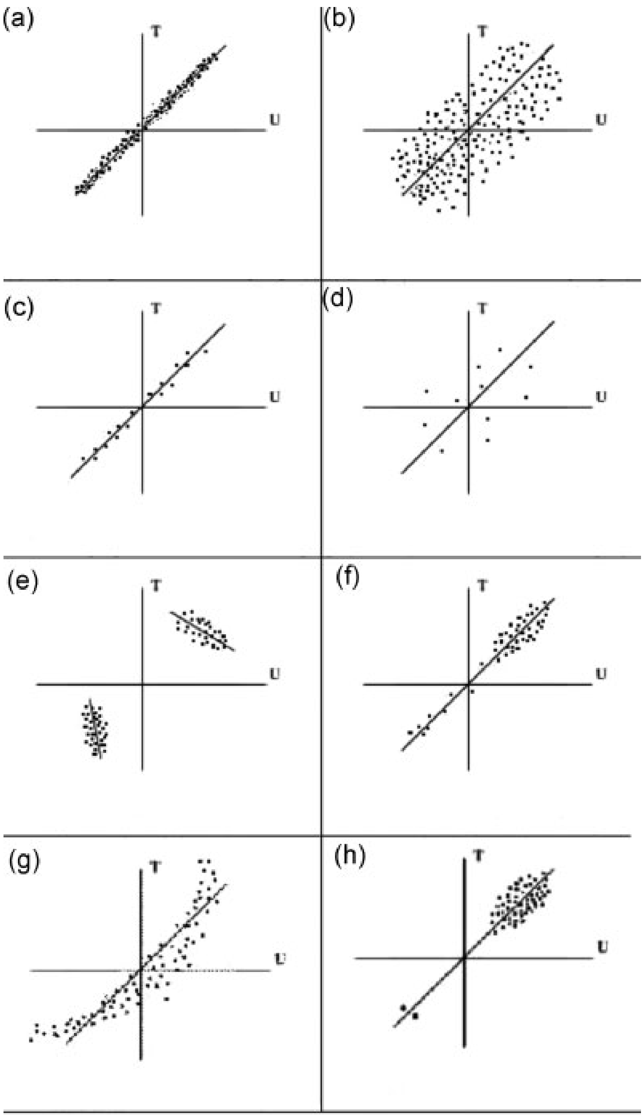

Figure 1. Eight principal covariance (correlation) data structure situations as depicted by their manifestation in PLS t–u plots. There are three influencing parameters which determine the appearance of all T–U data structures: (i) the number of objects N, (ii) the degree of linear (or nonlinear) correlation present, and (iii) data clustering, or grouping (data clumpiness). (a) A-type: strong data structure with many samples; (b) B-type: weak data structure with many samples; (c) C-type: strong data structure with few samples; (d) D-type: weak data structure with few samples; (e) E-type: clumpy data structure; (f) F-type: degenerated data structure; (g) G-type: nonlinear data structure; (h) H-type: the extreme outlier case.

[‡] For completeness, note that Kohonen [66] dismantles use of the R^2 statistics as well; there is no need to go into any particulars here, as it is well known that this statistic is severely sensitive to all data set structure irregularities, *ibid*; see also Høskuldsson [17].

clumpiness). The four first cases shown constitute a systematic series of strong/weak correlation versus small/large N, outlining the gamut of typical data sets for which legitimate PLS models are relevant. Three of the four latter cases represent deviating covariance data structures for which PLS modeling should never even be contemplated.

Based upon the schematics in Figure 1, it is easy to appreciate that an empirical match between cross-validation and test set validation ('similar' r_{opt} and RMSEP) is but a mere reflection of a particular strong correlation between the X- and Y-spaces in a given data set.

T–U plots must be inspected for every regression model being validated. There are traditions within chemometrics which does not include this imperative, instead prescribing 'blind' adherence to one selected version of segmented cross-validation, e.g. full cross-validation, or a fixed number of segments. Even a cursory overview of the principal covariance relationships between X- and Y-spaces delineated in Figure 1 leads to the inescapable conclusion that a fixed, universal number of segments will work in highly different ways depending on the *specific* data structure present. This goes a long way to explain why repeated cross-validation, identical but for alternative starting segment definitions, often may lead to significantly different validation results—this is always a strict consequence of a particular data structure regularity/irregularity. A fixed number of segments can never be said to pay the necessary problem-dependent attention to the many different data structures met with. There is thus ample justified reservation as to the often-claimed 'robustness' of cross-validation. Based on Figure 1, all types of varying validation results are comprehensible and need not lead to confusion.

Upon reflection, this issue is but the reverse side of the also often claimed optimistic, but not fully thought through, cross-validation credo: Validation is on safe ground as long/if several variants of validation, including several different segmented cross-validation result in *similar* validation results (identical number of components, 'similar' RMSECV). All is indeed well if/when this hopeful situation occurs—but the only thing which has been demonstrated is a situation of strongly correlated *X–Y* spaces, as depicted in Figure 1a, c, f, or h. Alas, nothing has been revealed as to the *future* prediction potential, unless it has been independently proved beyond reasonable doubt that this strong *X–Y* correlation remains the defining feature also of other data sets, indeed all other 'future' data sets. Such a demonstration is precisely the objective of test set validation, at least as far as one new data set goes, while all re-sampling approaches only deal with the one-and-only $\mathbf{X}_{\text{train}}$ data set.

The conclusion is clear—cross-validation is not a validation which incorporates information as to the future use of the particular data model. Cross-validation is rather an internal sub-setting stability assessment vehicle; cross-validation speaks only about the robustness of a particular model, as gauged by internal sub-setting of a training data set.

Still more insight can be had—for one-and-the-same data set of the type like cases (a)–(d) in Figure 1, there will always be a strong systematic regularity w.r.t. alternative segmented cross-validations with the number of potential segments increasing [$s = 2,3,4,\ldots N$]. Figure 2 depicts the systematics of 'RMSEP versus # PLS-components' plots, corresponding to the progression of all $(N - 1)$ alternative segmented validations.

There will always be a *lowest* RMSECV when the number of segments is at its maximum, N (corresponding to LOOCV). Conversely, when $s = 2$, RMSECV will be at its *maximum*. These

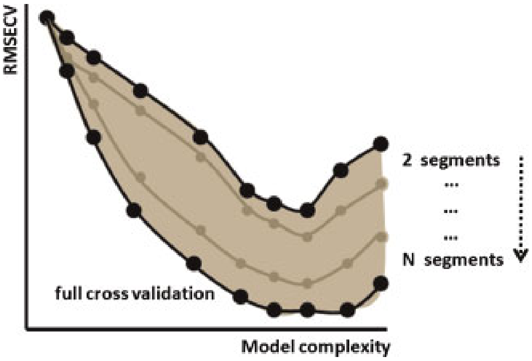

Figure 2. Systematic behavior of RMSECV as a function of the number of PLS components in the model. Prediction Y-error variance estimations decrease as a function of increasing number of cross-validation segments [$s = 2,3,4,\ldots, N$]. Irregular data structures will cause minor deviations from from the general relationship depicted.

relationships hold for all non-extreme covariance data structures when cross-validation is performed on one-and-the-same data set. Exceptions occur but are simply due to a larger degree of data structure irregularity, which of course at times will result in minor deviations from these principal systematic relationships. This generic illustration was first presented in the seminal textbook on multivariate calibration [51], but the significance for the validity of cross-validation was apparently overlooked. Based alone on a gradual reduction in the number of segments, RMSECV will increase and *vice versa*. Illegitimate conclusions may easily result if the data analyst succumbs to the temptation to select the particular segmentation level, s, which happens to correspond to the lowest RMSECV—such a *voluntary* this approach is wholly unscientific however. The objective of validation is most emphatically not to select the lowest possible RMSECV among a set of alternative segmented cross-validations—the objective is to estimate the most *realistic* prediction MSE as applicable to the situation pertaining to all future data sets. As shall be shown, this actually precludes cross-validation altogether.

Careful inspection of the pertinent *t–u* plot of any multivariate calibration model is the only remedy possible in order fully to understand and interpret results stemming from 'blind' cross-validations. This is possibly a reason why some traditions actively avoid inspection of *t–u* plots; these give an inconvenient insight into the real-world *X–Y* data structure which may indeed be very different from the universally applied assumption of a reasonably strong correlation between X and Y [Figure 1 (a–c)].

Such potential validation information does not reach the data analyst if systematic inspection of *t–u* plots is not on the agenda.

5.8. Remark on multiple validation approaches

When using segmented cross-validation several times over, or when using a multitude of different validation approaches, there is always a tacit wish that all (most) approaches will lead to (practically) the same optimal number of components. When this happens it is *claimed* that this situation is significant of a successful validation. From the above it follows that this is but a hollow truism—upon reflection, it is clear that all that is proven again is that a particular data structure is characterized by a strong (X, Y) correlation, and again nothing regarding the universal application of this or that variant of re-sampling approaches on future data sets was proven. There are a much higher number of data structures for which this does not hold.

5.9. Remark on non-sequential components

When data structure modeling allows for the possibility that not all sequential components necessarily are of interest, e.g. multi-block modeling, variable selection a.o. improvement of modeling fit alone (in combination with *none*, or *some* prediction performance issues), there are a suite of well-known statistics and optimization criteria available, see overviews in e.g. Kohonen [65], Høskuldsson [17], Gómez-Carrasco [66]. In these situations, as well as concerning nonlinear modeling/prediction it is necessary to consider all possible exponent- and cross-terms. The number of terms grows exponentially and it is no surprise that no strict, formalized way to handle this situation exists; but more-or-less voluntary applications of one or more re-sample techniques certainly have no superior merit on its own. A particularly inspired approach concerns the use of PDF (Pseudo Degrees-of-Freedom), van der Voet [75], which directly compares prediction errors from both test set validation and cross-validation.

5.10. Verdict on data set splitting

'Why is sequential application of an identical sampling protocol in order to produce two distinct data sets, X_{train} (now) and X_{test} (reflecting the future application situation as possible) different from splitting a doubly large X_{train} simply sampled in one operation?' This is undoubtedly the most often heard question/remark in discussions on re-sampling for validation. This issue reflects a strong desire for a simple, universal method. Alas, the complexities of proper validation (realistic assessment of the future MSE) do not comply with such a shotgun approach as this is most unfortunately misnamed 'test set splitting' suggestion.

Taking care of the number of measurement (samples, objects), N, is the easiest obligation of the experimentalist/sampler/analyst/data analyst. As is well known it is far more important to be in relevant control of the variance influencing factors when trying to secure a sufficiently representative ensemble of N objects (samples). There are usually few very useful guidelines in this game, except the universal stipulation that the training set must *span* the range of Y-values in a sufficient fashion, which is a problem-dependent issue; often this is the only consideration given to the issue of 'representativity'. This is a much too shallow understanding however.

The present work emphasizes the critical issue of understanding the phenomenon of heterogeneity as well as being in full command regarding identification and elimination of all 'incorrect sampling errors' (ISE), lest an uncontrollable sampling bias dominate the measurement uncertainty budget. The sampling process, dominantly governed by influences from ISE if these are not properly eliminated, does not give rise to analytical data which follows any known statistical distribution, Appendix A and literature [1–23]. However, the data structure of any data set, as depicted in a t–u plot, is a fair reflection of the sum total of all influencing factors on the measurement uncertainty. Sampling using a specific protocol acknowledging the principles of TOS ensures that all circumstantial conditions influence the sampling process in a comparable manner regarding both X_{train} and X_{new}, i.e. they are given the same opportunity to play out their role irrespective of who is doing the sampling, and who is doing the sample preparation and analysis, etc. This is the role of an objective sampling_and_analysis protocol.

Circumstantial conditions are capricious; however, they are time-varying—and in general they defy systematization. But any second sampling from a lot (we very deliberately refrain from using the too simplistic terminology 'from the population') will always reflect the sampling/analysis objectively—precisely at the time, or at the place (in the future setup) pertaining to the second sampling session. The crux of the matter is that the sampler has no control over which, and to which degree, these conditions may have changed *between* sampling the training data set, X_{train} and the 'future data set', X_{new}. This simple 'second sampling at a future time/place' is the crucial information carrier 'from the future application situation' that must be included in the validation procedure.

This difference can be of any magnitude: insignificant, intermediate or gross. But this ambiguity is immaterial since the changing conditions cannot be described or quantified; they are specifically those factors and circumstances which are not actively involved; indeed cannot actively be involved in the specification of the sampling situation. All that matters is that the second data set will unquestionably display the data structure reliably as related to the second sampling session (or to a third etc.). To the degree that this has changed, there is now a trustworthy representation hereof involved in the validation, as illustrated by Figure 3.

What is implied here is that to the degree such differences are found to be present between the training data set sampling and the future application situations (time, place), these must be involved in the validation procedures, or the model estimations ($\mathbf{r}_{\text{opt}}$, RMSEP, other . . .) will not be fully relevant for the purpose it was intended to serve. It is immaterial that it often is argued that in 'some cases' this difference is not of sufficient magnitude to be of influence. The winning argumentative point is that it is not possible to identify such cases *a priori*. All situations are unknown at the time of decision of which type of validation to perform. As a matter of fact, the only way this dilemma could ever be circumvented would be by carrying out both a test set validation and one or more cross-validation alternatives. If this is the case,

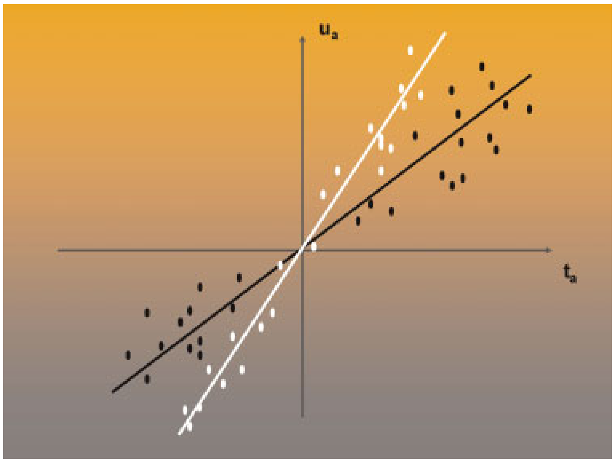

Figure 3. Synoptic display of X_{train} and X_{new} (X_{test}) furthering an objective basis for evaluation of the empirical data structure differences and their qualitative and quantitative expressions pertaining to the two *independent* sampling/analysis sessions. The two data sets shown display significantly different loading weights, w, and the training set (white) displays a distinctly smaller variance than the test set (black). Principal sketch; increasing data structure irregularity will blur these principal relationships visually, but the basic principle displayed remain the same.

Copyright © 2010 John Wiley & Sons, Ltd.

one would never accept the structurally inferior cross-validation, however, relative to the more realistic test set alternative.

The *logical* conclusion is to decree a *test set validation imperative*. Nothing adverse will ever result from *always* applying test set validation, and one is supplied with superior information in one operation, since test set validation as a matter of fact delivers estimates of both r_{opt} and RMSEP—while everything is uncontrollably risky (possibly downright wrong) by basing a re-sampling validation approach an in principle un-testable assumption of global training set representativity.

This line of argumentation also sheds light on the issues surrounding partial use of cross-validation. It is often claimed that cross-validation is a good help to determine the 'correct' number of factors, i.e. cross-validation is often accepted for *internal* validation purposes—while it is often claimed that for the most reliable estimation of RMSEP, a 'completely independent' i.e. an *external* validation data set is required. The present treatment of course has no quarrels with the latter stipulation—but is in total disagreement regarding any use of cross-validation for determination of $\mathbf{r_{opt}}$.

It is not in the best interest to invoke a two-step validation approach. By using a one-step procedure, test set validation, one is directly presented the most reliable estimate of RMSEP based on the objectively correct number of components, all in one go, since test set validation manifestly always will include all sampling and measurement uncertainty effects originating from whatever changed circumstantial conditions.

A traditional argument is often heard: 'IF there is a significant difference between the data structures of the training vs. the test set—is there any reason to suppose that an overarching model will fit the data? Without prior knowledge as to this difference . . . why should the model based on the training data set also fit the "future" data sets?' – This is a very pertinent question, as it goes directly to the heart-of-the-matter of validation: If there is a significant data structure difference, it is imperative that the training data set model is evaluated only by the test set approach which manifestly incorporates the second (future) data structure. The resulting RMSEP will *always* result in a higher RMESP estimate than any cross-validation alternative, and will *ipse facto* be the more *realistic* estimate, see Figure 4.

There is always the alternative to augment the training data set with the test set (or portions hereof) if it can be established that the data structure difference is beyond acceptance. Such a situation is a strong indication that the data structure at the training data set calibration is inherently unstable however, and nobody would accept just a local cross-validation in such a case. In some situations, typically in process contexts, there is a relaxed acceptance of the inevitable lapse into such a unacceptably changed data structure with 'time' (or 'location'); the order of the day is of course simply to accept this and to instigate some form of 'model updating'. Within many process technology communities, updating is a standard issue. There may be several alternative ways to go about this, all strongly dependent on the particular contextual setting, but the main issue is that cross-validation is correctly viewed as quite unable to deliver the necessary realistic MSE estimate of future performance.

The above arguments are the principal reasons for not splitting a training data set, however large. With splitting there is no information pertaining to the future application situation at all, the number of objects notwithstanding. A massive redundancy in numbers is mistaken for a suitable basis for future performance validation. Test set validation is the best possible valid attempt to remedy this predicament—by securing (at least) one new data set from as far in the 'future' as is logistically possible. By accepting that the circumstantial conditions may/will change (it is only a matter of time), a test set is the best one can do within reason.

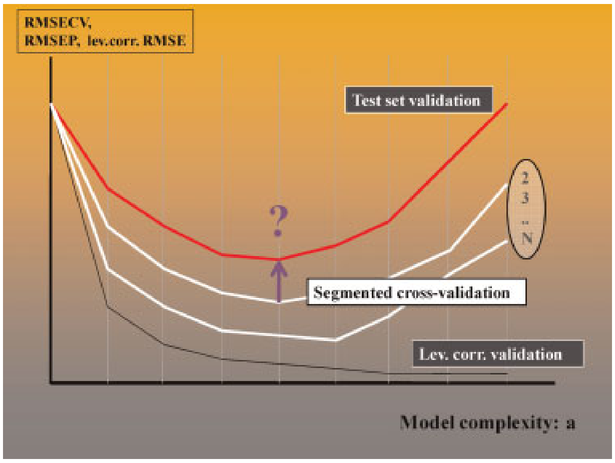

Figure 4. Relationships between RMSE estimates as a function of model complexity (a). Leverage-corrected RMSE estimates are universally lower than those pertaining to either re-sampling or test set validation. Segmented cross-validation estimates are structurally lower than those stemming from test set validation. For one-and-the-same data set, X_{train}, the systematic relationships between the different segment variants are only indicated here, they were laid out in completion in Figure 2. Stronger-and-stronger (X,Y) correlation will result in more-and-more *similar* curves in this type of plot, see text for full details.

From this discussion it also transpires that a regimen of systematic, regular test set validation model checking is a wise approach within the arena of quality monitoring and quality control. The above discussion appears particularly easy to understand in the process technology, process monitoring and control settings. Proper process sampling in this context is treated specifically elsewhere [20,23].

5.11. Cross-validation does have a role—model comparisons

All is not lost for cross-validation however. In the arena of model comparison (models of structurally identical nature, but of optional alternative parameter settings: e.g. different pre-processing alternatives, different X-variable selection alternatives, etc.) cross-validation would appear to be a particularly relevant approach. For this specific purpose, cross-validation furnishes precisely what is needed, a general identical model performance framework within which alternative parameter settings/values can be objectively compared. In this context it is formerly a necessity to use the same number of segments for all alternative validations. In this area of applied validation, there is very good use for cross-validation, although it is interesting to contemplate how one is to deal with the possibility of a different number of components r_{opt} for alternative optimized models. Even for this legitimate use of cross-validation there is always a demand for responsible disclosure of the structural RMSE underestimation deficiencies a.o.

The special case of QSAR/QSAP is treated more specifically below.

6. DISCUSSION

6.1. Systematics of cross-validation

It is advantageous to treat all cross-validation variants under a common systematic heading, here termed *segmented cross-validation*. This allows significant simplification in discussing historically disparate variants: Leave-one-object-out (LOO), the plethora of differently segmented cross-validations and the so-called 'test set split' option (which is a particularly obfuscating terminology for an otherwise straight 2-segmented cross-validation approach. Indeed this name could not have been chosen in a worse fashion; 'test set split' is but a *myth*).

Depending on the fraction of training set samples (totaling N) held out for validation, an optional range of $(N-1)$ potential cross-validation *segments* will be available for the data analyst, the number of segments falling in the interval [2, 3, 4,. . ., $(N-1)$, N]. Various 'schools-of-thought' of cross-validation have developed over history, within chemometrics and elsewhere, some favoring 'full cross-validation' (one object per segment; N segments in total—LOO), some defining 10 segments as the canonical number and others favoring similar *schemes* each with its own preference (e.g. 3, 4, or 5 segments)—whereas a small, but steadily growing minority see more complexity in this issue than a more-or-less voluntary selection from the space of $(N-1)$ optional variants of segmented cross-validation.

There *always* exists this range of *(N-1)* potential cross-validation variants for any given data set with N samples, but no set of principles for objective determination of the optimal number of segments has ever been offered. Esbensen [23] offered a first foray only; the issue is complex.

6.2. Phenomenology of cross-validation versus test set validation

Usually there is more focus on strict adherence to one or other fixed cross-validation *procedure*, complete with preferred number of segments (*scheme*), than openness with respect to what exactly are the precise assumptions and prerequisites behind cross-validation. This is troublesome, as no amount of *pro et con* discussion of specific numbers of segment will reveal the underlying structural problems associated with all types of cross-validation.

Against this backdrop Esbensen [21] analyzed the operative aspects of cross-validation versus test set validation, further corroborated and more powerfully illustrated by Esbensen and Lied [22]. The general conclusion arrived at here, in parallel with similar insights found sparingly across the validation literature, is that cross-validation is aimed at performing as a particularly effective *simulation* of test set validation. However, while for a superficial comparison cross-validation performs in a *similar fashion* to test set validation, there is one critical dissimilarity: there is only one data set involved in cross-validation, X_{train}.

This is aggravated by the fact that any modification always concerns a reduction of its size, i.e. *some* local modeling being based on a subset of N objects only. Any of the local models used for estimation of the number of components to be included in the regression model and/or to estimate RMSEP will manifestly be based on a voluntarily reduced data set of an undisputedly smaller number of objects than N. This constitutes a massive contradiction however. As soon as the training set has been defined by the data analyst, this means that all the objects herein are needed for its purported representativeness. None of the objects in a properly defined training data set are therefore available for the kind of voluntary exclusion demanded in cross-validation. It is only fair that once the data analyst (or the informed data supplier) has made the utmost efforts involved in securing an absolutely optimal training data set, this particular set configuration must remain unmodified (re. the number of objects, their spanning ranges, their 'representativity' . . .)—or else all credibility to the training set definition is lost. After the training data set has been codified into the necessary protocol, there can be no change.

6.3. Two worlds of sampling understanding

It is necessary to distinguish clearly between the process of statistically drawing from a population of i.i.d. objects (this process is here termed *sampling*$_{STAT}$) and physical sampling, i.e. materializing N individual 'samples' (objects in data analysis parlance) from a heterogeneous system (a lot) (this process is here termed *sampling*$_{TOS}$) in order to eliminate any possibility of confusing one sampling process with another. The critically important distinction is that while *sampling*$_{STAT}$ assumes selection of N objects from a population of otherwise *similar* objects (all objects are *similar* in all aspects than w.r.t. the analyte), *sampling*$_{TOS}$ is addressing a significantly heterogeneous target by a selection process. The first situation corresponds with the occurrence of the type of sampling bias, which conventionally can be subjected to the statistical bias correction. In the latter, fully TOS-error influenced situation, the sampling bias is always varying (inconstant), and very nearly never small enough to neglect, rather to the contrary. A complementary distinction: a (statistical) sample is a subset of a population: *sample*$_{STAT}$, while a *sample*$_{TOS}$ refers to one individual physical subpart of lot material, with a specific mass (*mass*$_{SAMPLE}$), which is the result of a specific sampling process (*sampling*$_{TOS}$). TOS focuses on how to obtain a 'representative sample' (*sample*$_{TOS}$), as opposed to non-representative 'specimens'. All essential definitions and relationships in TOS are given in the published literature [1–23]; Appendix A summarizes a minimum background to all claims regarding TOS in the present context.

6.4. Proper validation—the role of the sampling bias (present vs. future)

The above discussion illustrates the fundamental distinction between the statistical and the TOS data quality contexts. Without full and unambiguous understanding of these differences, unnecessary confusion and heated discussions without any possibility of ever reaching a common ground will only continue. Lot and material heterogeneity and the way it is conceptualized and quantified constitutes the singular key to understanding the differences between *sampling*$_{STAT}$ and *sampling*$_{TOS}$. A recent overview in summary form of TOS in relation to sampling, chemometric data analysis and validation can be found in Reference [23]; process sampling in the PAT regimen was analyzed in Reference [20]. The full argumentation for claiming, as we do in this work, that one cannot understand validation in the conventional i.i.d. statistical context only is presented here. No matter how offensive this may appear at first sight, what is meant is only that a sufficient TOS heterogeneity understanding is *also* needed.

Copyright © 2010 John Wiley & Sons, Ltd.

Validation using only one data set must be appreciated based on the key issues pertaining to the TSE: any *N*-object data set constitutes a specific realization of an *N*-tuple of individual TSE materializations. Cross-validation precludes any other possibility than this singular manifestation on the ensemble of the *N* objects in the training set. By voluntarily having access to one data set only (one set of objects drawn from the population in the statistical parlance) cross-validation simply never includes the possibility of incorporating empirical information as to the future prediction situation and the specific data structure reigning here (which *may* or *may not* be different, one will never know). Instead, various declarations of faith in the *assumption* that all training data sets are *always* representative of the future application situations have been offered both in many different versions in the literature and in other contemporary discussion fora. These claims are, however, invalidated by their own limited scope, as they are invariably dressed up in the framework of a statistical population only.

The key argumentation in the present work revolves around sufficient understanding of the complexities incurred by TOS to this complacent view. Standing on the statistical context only, the data analyst is conveniently relieved of all responsibility with respect to the representativity of any-and-all future data sets. It is this claimed universality which strikes a completely wrong tenor—it is simply untrue in the light of TOS and its extensive experience.

The central issue is then that cross-validation sub-models only reduces the calibration or validation basis of X_{train}. It is off-handedly assumed that any-and-all of these sub-models used for estimating both $\mathbf{r_{opt}}$ and RMSEP can further reliable information of how the model will also perform on future data sets (test set, application data sets ...). This tacit assumption stands in stark contrast to the necessity to relate validation to the *real* future prediction situation. For cross-validation there rests an enormous burden of proof in this context. It is fair to say that within chemometrics no cross-validation school has presented anything akin to proof of how this voluntary reduction of the singular training data set may relate to prediction performance in the future. This procedural tradition simply has no rational foundation.

The future prediction situation will have to be characterized by at least one new data set X_{new}. The central issue here is that all new data set will be associated with a new realization of the ensemble TSE manifestations, which is never identical with that for X_{train}; all new data sets will *per force* encompass a new set of TOS-sampled objects, each with a new individual TSE manifestation.

The main lesson from TOS' more than 50 years of practical experience is that there is no such thing as a constant sampling bias—the sampling bias changes with every new sampling from heterogeneous materials. For heterogeneous lots, there exists no possibility for the kind of bias-correction offered in statistics, as this is based on a faulty over-simplified bias constancy assumption. Real world heterogeneity is much more complex than the statistical model of a population of i.i.d. objects, with a very few non-consequential exceptions (infinitely diluted solutions, and similar systems, which do not allow any generalization). This key insight is furnished by even a rudimentary understanding of the phenomenon of heterogeneity, DH_L, distributional heterogeneity and CH_L, compositional heterogeneity. No manner of statistical modeling of instrument/signal error propagation will even begin to model the material heterogeneity which leads to TSE in the range of 10–50–100 × the TAE, Appendix A. These are facts which are fundamental and irreducible characteristics behind all (*X*,*Y*) data spaces. The data analyst viewpoint is not broad enough to deal with the complete heterogeneity, sampling, analysis, data analysis context.

If each individual samples$_{\text{TOS}}$ contribute in a varying, uncontrollable degree to the overall ensemble bias, it follows that the resulting data structure will be different for each new ensemble of *N* objects—specifically this means that each-and-any future *sample$_{STAT}$*, a test set, by necessity must be different from the training set; it is only a matter of to which degree different. It follows that there can never be any guarantee that the specific training set realization will necessarily also be representative of all future data sets—even if each ensemble repeats the same number of *N* individual TSE materializations. This is the crucial distinction between physical TOS sampling and statistical sub-setting sampling upon which the present argumentation rests.

It also follows that bias is a function of the specific heterogeneity expressed by the lot material and form, as well as of the specific sampling process employed. In fact TOS specifies no less than three types of bias-generating errors stemming from the sampling process itself, which manifest themselves if not specifically eliminated; this fact constitutes a most serious reason to be informed and reasonably conversant regarding TOS in practice. In addition, the ensemble bias will change its nature and magnitude also as a function of which and how many objects be included in the training and the test set(s).

Therefore, it is mandatory to test all implicit or explicit *assumptions* regarding sampling bias constancy. As it happens this is not difficult, nor is it associated with prohibitive outlays in terms of work, resources or economy—in fact this can actually be done 'free of charge' in practice every time a prediction model is to be used for its designated purpose—by the simple procedure of test set validation option. All predictions are here always based on a new data set, that is, a new ensemble of TOS-sampled objects (*N* objects). Therefore, sampling will *per force* have been involved at the very least a second time (or more) in exactly the same fashion as for the training data set. However, all X_{NEW} data sets constitute a completely new, independent *sample$_{STAT}$*, containing a valid representation of the relevant 'heterogeneity information' pertaining to the future use.

With the help of the relationships shown in Figure 4 it is possible to delineate the universal deficiency displayed by all types of segmented cross-validation (compared with leverage-corrected validation as well): Test set validation will always result in the *highest* estimate for RMSEP than any of the segmented cross-validation alternatives (and often very much higher than the leverage-corrected RMSE estimate)—indeed the *most realistic* estimate.

Figure 4 summarizes an extensive accumulated experience with validation of many hundreds of data sets, representing all types of data structure depicted in Figure 1, especially all those of more regular appearance, types (a)–(d). This systematic understanding is communicated by many professional data analysts as well. In our own decades of chemometric experiences (teaching, professional, consulting), innumerable data analyses have also led to similar results based on very many, diverse data structure types. Obviously at times there may also exist partly deviating curves to the ones depicted, but these are invariably always related to just more irregular data structures.

The 'gap' illustrated with a vertical arrow represents the missing second TSE-component which can only be quantified by comparing the test set and the cross-validation results (in whatever guise). This can be said to constitute the 'missing link' w.r.t. the MSE estimate in re-sampling/cross-validation. This is the quantitative measure of the missing TOS-error component which can only be incorporated by incorporating a second data set. From this insight, one can conclude that most types of variants of either re-sampling types in general, or the internal cross-validation type in particular, by necessity are inferior with respect to their own purported objectives, precisely because the dominating TSE contributions from the virtual set of all possible future data sets are never involved. It follows that cross-validation should logically and scientifically be discontinued. Only test set validation can stand up to the logical demands of all the characteristics of proper validation. One should henceforward observe a test set imperative.

Based on the above discussion it is possible to comprehend the following analysis of an often presented 'combined approach' which is characterized by *external splitting* of a test set (from the master training set) combined with *internal cross-validation*. A recent paper by Filzmoser *et al*. [67] presents the hitherto most evolved version of this approach in a comprehensive systematic framework of intensive repetitions of both internal and external loops (termed 'repeated double cross validation'). The common feature of this and various closely similar proposals is that the master training data set is subjected to a barrage of repeated internal and external splitting and re-samplings from which, it is claimed, it is possible to obtain superior validation results (sometimes based upon evaluation by conventional statistical methods, e.g. a_{opt} frequency distributions, residual histograms). There are minor variations as to the exact degree to which new sub-model fitting process takes place at all or selected levels and repetitions, but this is all unfortunately a moot point: No manner of advanced repetition of model optimizations (wheels-within-wheels) based on cross-validation, the fundamentals of the present critique stands. These approaches represent a powerful school-of-thought within current chemometrics, a tradition which is adamantly insisting on a two-step framework: internal estimation of model complexity first (universally this is carried out by cross-validation), followed by external 'test set' or 'validation set' estimation of the resulting RMSEP. While this latter is associated with a correct insistence that external test set validation must be fully independent from model optimization (there is a very clear demand that no objects must have been involved in both calibration and validation), sadly this is often just lip service to the test set imperative, since the point of origin is splitting from the master training set—indeed all objects ultimately originate from the one-and-only training data set. While much original thought has gone into these procedures and principles, as long as they are not based upon full understanding of the salient reasons why -, and the principles behind the test set imperative (PPV), these approaches fall from the exact same critique and reasons as their simpler counterparts which were all shown to be inadequate above. In fairness, some proponents within this two-step consensus actually insist of proper test sets in the present form—alas also insisting that internal cross-validation is the proper procedure for model optimization, for which reason these are equally at fault in the final evaluation. The crucial issue, exceedingly difficult to abandon, is the inability to comprehend the inherent inferior specification of model optimality if based on one data set only. Interestingly phrases like: '***within** the population of the data used*', '*although calibration set and test set have been selected randomly, the resulting SEP$_{TEST}$ values could be (**just by chance**) too optimistic or too pessimistic, depending on **how representative** this separation was*' and '*assuming that **all** new samples are from the same data population as the samples used for model creation* . . .' [67] [**emphasis** by present authors] clearly reveal that use of terms and issues like chance, population, representativity cannot be used at large but must be based on appropriate understanding of the TOS and its intimate impact on all central validation issues, delineated in full in Appendix A.

6.5. Disclaimer on universalities

Above it was argued as if there always, without exception, will be significant TOS errors present, leading to the demonstrated 'extra-statistical sampling variances'. The present authors fully acknowledge, however, that there be situations in which the sampling$_{\mathrm{TOS}}$-errors can be demonstrated to be of only insignificant magnitude(s). To the degree that some cases of this nature do exist, if/when/where appropriately demonstrated, the present TOS-augmentation can safely be disregarded—but the burden of proof-of-existence must lie with the too glib contrary general representativity assumptions. It is comforting that all test set validations can be directly compared with any particular re-sampling preference as well (e.g. it is always possible also to perform a cross-validation on X_{train}). The opposite case is universally impossible, and the shortcomings stemming from a re-sampling procedure only are dramatically well illustrated by Figure 4.

QSAR/QSAP contexts may often constitute a fundamentally different situation than what has been delineated above. The objects in question here are typically molecules etc.—in this context there is often no FSE or GSE (nor ISE), i.e. there are no 'measurement errors' as regards the definition of the objects in the *X*-space. Often also, the data sets involved are by nature representatives of the 'small sample case' (i.e. a small number of objects). As regards *X*-variables, however, there may, or may not, still be measurement errors involved depending upon whether quantum chemistry calculations or direct measurements are involved in descriptor quantifications, while *Y*-values (activities, functional properties) clearly are not exempt from the present context. In order not to create confusion and futile debate, the present authors have no desire to declare blanket inclusion of QSAR/QSAP in the present call for a test set imperative; here would indeed appear to be good reasons to claim special circumstances which merit application of careful and reflected use of cross-validation [62,68,74].

6.6. Remark on several test sets

With the above very few, quite specific exceptions, the test set concept is universally the most *realistic* prediction validation possible. This is so because all relevant errors components are guaranteed to be included: all *X*-errors are incorporated in X_{NEW} in fully realistic fashion, as are all *Y*-errors, sampling$_{\mathrm{TOS}}$ as well as sampling$_{\mathrm{STAT}}$. We may trust this to be the optimal validation approach because all future use of the prediction model necessarily will involve identical conditions for similar new sampling and analysis.

Arguments can easily be raised for invoking a postulated need for several test sets: Of course more than one test set will always allow for more valid assessment, since more test set realizations correspond to more examples of the future in-work prediction scenario for the prediction model etc. However, here we see no special need to make matters more difficult by going to this extreme. One properly materialized test set will suffice to incorporate the principal information from the future situation as best possible given the dominating objections and postulated budget or effort constraints, which is often claimed not even to allow for one test set. In a rational context, it is evident that a decision regarding the real need for several test sets will be based much more on specific problem-dependent characteristics, always related to the specific data background at hand.

7. CONCLUSIONS

Re-sampling and cross-validation approaches to validation work on one data set only, X_{train}. This scenario was analyzed in detail. The tradition of cross-validation is particularly strong; its current use is mainly based on unsubstantiated *assumptions* of the training set always being fully representative of the population, indeed for all types of data sets, also those pertaining to 'future' use—in splendid disregard of their extremely varying origins and varying data structures. This widespread tacit assumption was shown to be untenable in the light of the significant bias-generating sampling errors described in the TOS. It is critically necessary to be able to competently identify and eliminate all the so-called Incorrect Sampling Errors (ISE).

Instead of an almost endless series of partial exemplifications (based on particular date set structures) presented in the literature, and from which no valid generalization can ever be made, we have alternatively established *first principles* regarding validation. The PPV were outlined based on a set of key distinctions:

(i) Validation cannot be understood by focusing on the methods (*schemes*) of validation only; proper validation must be based on full knowledge of the underlying definitions, objectives, methods, effects and consequences.
(ii) Analysis of common validation objectives implies that there is only one valid paradigm, formulated as the test set validation imperative.
(iii) Contrary to much contemporary chemometrics validation practice and *myths*, cross-validation is shown to be unjustified in its current form of monolithic application of one principal type of procedure (segmented cross-validation) on any-and-all data sets. Within its scope and design, cross-validation is shown as but a sub-optimal *simulation* of test set validation only, crippled by a critical sampling variance omission, because it is based on only one data set, the training data set.

Many re-sampling validation methods were shown to suffer from the same deficiencies.

The PPV are simple and universal and can be applied to all situations in which assessment of performance is desired—be this prediction-, classification-, time series forecasting-, modeling validation a.o. The new element in PPV is the TOS, which is needed in order to be able to identify and eliminate all bias-generating sampling errors (incorrect sampling errors) which are responsible for unnecessary, inflated heterogeneity-induced measurement variances, and for which there are no statistical corrections possible. Invoking the complete body of theoretical and practical experiences from over 50 years' of application the TOS, it was shown to be untenable to continue with bland, unjustified *assumptions* regarding representativity. A salient brief, Appendix A, supported by a full set of references, argued how the TOS is able to describe, correct and reduce to *a priori* acceptable levels for all kinds of material or lot heterogeneity errors as well as eliminate those errors originating from the sampling process itself. Sampling variance, in the form of both $\mathrm{sampling}_{\mathrm{TOS}}$ and sampling $_{\mathrm{STAT}}$ is the result of a much more complex interaction between a specific sampling process and the material heterogeneity in question, than what is contained in the traditional statistical population concept alone. On this basis it was concluded that re-sampling and cross-validation approaches miss out critically with respect to the crucial $\mathrm{sampling}_{\mathrm{TOS}}$ variance, which can only be accommodated by a test set (a second independent sampling—more than one if so desired locally, but this is not a universal demand), without which re-sampling validation will universally underestimate the realistic prediction error. There is no theoretical way to derive any approach that can estimate the magnitude of this missing part. For this reason, re-sampling and cross-validation should logically be terminated or only used in practice with full disclosure of the critical deficiencies outlined. QSAR/QSAP constitutes a special case, in which informed use of cross-validation may be well merited, especially in the 'small sample case', although in the final evaluation, a test set will still always reign superior.

Regarding the main chemometrics method PLS-regression, a call was made for commitment to test set validation based on graphical inspection of *T–U* plots for optimal understanding of the operative *X–Y* interrelationships. Simple visual inspection will also allow a reliable premonition of the outcome of any particular validation approach, especially if based on the complete sampling variance understanding ($\mathrm{sampling}_{\mathrm{stat}}$ and $\mathrm{sampling}_{\mathrm{TOS}}$). There is no justification continuing to reject the work effort involved in securing a test set for validation purposes, acknowledging that this is the only approach which eliminates the deficiencies outlined. The comparatively rare occasions when a test set is manifestly not an option (historical data a.o.) have absolutely no generalization power—and the comprehensive understanding delineated here will stand the data analyst in good stead also when forced to perform some form of re-sampling or other. Full disclosure of the structural MSE underestimation deficiency is mandatory in all cases.

Many reasons are given in scores of traditional arguments for continued use of cross-validation and re-sampling for validation. Our critique against can be summarized:

- Complacency: one approach/method for all data sets, disregarding vastly different data correlation structures
- Focus on algorithms, implementation and software, without critical thinking
- Unwillingness to investigate consequences of traditional statistical population assumptions
- Resistance against the TOS for complementary understanding re. heterogeneity and sampling process issues
- Confusion regarding fundamental (hard) versus soft data models
- Admiration of mathematics and no interest in how 'data' originate (TOS)
- Blind adherence to traditions or schools of thought: 'This is the way chemometrics has been doing validation for close to 40 years . . .'
- Two-step approaches involving cross-validation cannot be justified however personably cloaked in *a priori* or *a posteori* statistical re-sampling procedures.

REFERENCES

1. Petersen L, Minkkinen P, Esbensen KH. Representative sampling for reliable data analysis: theory of sampling. *Chemom. Intell. Lab. Syst.* 2005; **77**(1–2): 261–277.
2. Gy P. *Sampling for Analytical Purposes*. Wiley: Chichester, UK, 1998.
3. Pitard FF. *Pierre Gy's Sampling Theory and Sampling Practice* (2nd edn). CRC Press LLC: Boca Raton, USA, 1993.
4. Smith PL. *A Primer for Sampling Solids, Liquids and Gases – Based on the Seven Sampling Errors of Pierre Gy*. ASA SIAM: Philadelphia, PA, USA, 2001.
5. Gy PM. The analytical and economic importance of correctness in sampling. *Anal. Chim. Acta* 1986; **190**: 13–23.
6. Petersen L, Dahl CK, Esbensen KH. Representative mass reduction: a critical survey of techniques and hardware. *Chemom. Intell. Lab. Syst.* 2004; **74**: 95–114.
7. Minkkinen P. Evaluation of the fundamental sampling error in the sampling of particulate solids. *Anal. Chim. Acta* 1987; **196**: 237–245.
8. Heikka R, Minkkinen P. Comparison of some methods to estimate the limiting value of the variogram, vh(j), for the sampling interval j ¼ 0 in sampling error estimation. *Anal. Chim. Acta* 1997; **346**: 277–283.
9. Gy P. Sampling of discrete materials – III: quantitative approach – sampling of one dimensional objects. *Chemom. Intell. Lab. Syst.* 2004; **74**: 39–47.
10. Gy P, Marin L. Unbiased sampling from a falling stream of particulate material. *Int. J. Miner. Process.* 1978; **5**: 297–315.
11. Gy P. Does your mechanical sampler do what it's supposed to? *Coal Min. Process* 1981; **34**: 71–74.
12. Petersen L, Esbensen KH. Representative process sampling for reliable data analysis – a tutorial. *J. Chem.* 2006; **19**(11–12): 625–647.
13. Halstensen M. Experimental multivariate sensor technology and development of system prototypes for industrial multi-phase characterisation: selected forays. *Doctoral Thesis*, HIT, 2001.
14. Esbensen KH, Friis-Petersen HH, Petersen L, Holm-Nielsen JB, Mortensen PP. Representative process sampling – in practise: variographic analysis and estimation of total sampling errors (TSE). *Proceedings 5th Winter Symposium of Chemometrics (WSC-5), Samara 2006*. Chemom. Intell. Lab. Syst. 2007; 88(1): 41–49.
15. Esbensen KH, Minkkinen P (Eds). Special Issue: 50 Years of Pierre Gy's Theory of Sampling. Proceedings of First World Conference on Sampling and Blending (WCSB1). Tutorials on Sampling: Theory and Practise. *Chemom. Intell. Lab. Syst.* 2004; **74**(1): 236.
16. Minkkinen P. Weighting error – is it significant in process analysis? In *Proceedings Third World Conference on Sampling and Blending (WCSB3)*, Porto Alegre, 23–25 October 2007; Costa J.F.C.L., Koppe, J.C., (Eds). 2007; pp. 59–69. ISBN 978-85-61155-00-1.
17. Høskuldsson A. 1996; *Prediction Methods in the Sciences*. Thor Publishing: Denmark. ISBN 87-985941-0-9
18. Miserque O, Pirard E. Segregation of the bulk blend fertilizers. *Chemom. Intell. Lab. Syst.* 2004; **74**: 215–224.
19. Martens M, Martens H. *Multivariate Analysis of Quality. An Introduction*. Wiley: Chichester, 2001; p.p 445. ISBN0-471-97428-5.
20. Esbensen KH, Paasch-Mortensen P. Process sampling (theory of sampling) – the missing link in process analytical technologies (PAT). In *Process Analytical Technologies*, (2nd edn). Bakeev K (ed.). Blackwell: Oxford, 2009; (in print). Chapter 3.
21. Esbensen KH. *Multivariate Data Analysis – in Practise. An Introduction to Multivariate Data Analysis and Experimental Design*, (5th edn). CAMO AS Publ: Oslo, 2001; p. 598. ISBN 82-993330-2-4.
22. Esbensen KH, Lied TT. Principles of Image Cross-validation (ICV): representative segmentation of image data structures. In *Techniques and Applications of Hyperspectral Image Analysis*, HF, Grahn P Geladi (eds). Wiley: Chichester, 2007; Chapter 7, pp. 155–180. ISBN 978-0-470-01086-0.
23. Esbensen KH, Julius LP. 2009; Representative sampling, data quality, validation – a necessary trinity in chemometrics. In *Comprehensive Chemometrics, Volume*, **4**, Brown S, Tauler R, Walczak R (eds). Elsevier: Oxford, 1–20.
24. Kuhn TH. *The Structure of Scientific Revolutions*, (2nd edn), (enlarged). University of Chicago Press: Chicago, 1970.
25. Larson S. The shrinkage of the coefficient of multiple correlation. *J. Edu. Psychol.* 1931; **22**: 45–55.
26. Lachenbruch P. Estimation of error rates in discriminant analysis. *PhD Thesis*, University of California, Los Angeles, 1965
27. Lachenbruch P, Mickey M. Estimation of error rates in discriminant analysis. *Technometrics* 1968; **10**: 1–11.
28. Camstra A, Boomsma A. Cross-validation in regression and covariance analysis. *Socio. Method. Res.* 1992; **21**: 89–113.
29. Stone M. Cross-validatory choice and assessment of statistical predictions. *J. Royal Stat. Soc.* 1974; **36**: 111–147.
30. Geisser S. A predictive approach to the random effect model. *Biometrika* 1974; **61**: 101–107.
31. Geisser S. A predictive sample reuse method with applications. *J. Am. Stat. Assoc.* 1975; **70**: 320–328.
32. Stone M. Cross-validation and multinomial prediction. *Biometrika* 1974; **61**: 509–515.
33. Stone M. Asymptotics for and against cross validation. *Biometrika* 1977; **64**: 29–35.
34. Wold S. Cross-validation estimation of the number of components in factor and principal component analysis. *Technometrics* 1978; **20**: 397–405.
35. Bowman A. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* 1984; **71**: 353–360.
36. Picard R, Cook R. Cross-validation of regression models. *J. Am. Stat. Assoc.* 1984; **79**: 575–583.
37. Li K. From Stein's unbiased risk estimates to the method of generalized cross-validation. *Anna. Stat.* 1985; **13**: 1362–1377.
38. Burman P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika* 1989; **76**: 314–503.
39. Efron B. Bootstrap methods: another look at the jackknife. *Anna. Stat.* 1979; **7**: 1–26.
40. Wehrens R, Putter H, Buydens L. The bootstrap: a tutorial. *Chemom. Intell. Lab. Syst.* 2000; **54**: 35–52.
41. Leite E, de Souza C. Artificial neural networks applied to mineral potential mapping for copper-gold mineralizations in the Carajas Mineral Province, Brazil. *Geophys. Prospect.* 2009; **57**: 1049–1065.
42. Wold H. Open path models with latent variables. The NIPALS nonlinear iterative partial least squares approach. In *Quantitative Wirtschaftsforschung: Wilhelm Krelle zum 60 Geburtstag*. Albach H, Helmstedter E, Henn R (eds). Mohr: Tübingen, Germany, 1977; 729–754.
43. Martens H, Jensen S-A. Partial least squares regression: A new two-stage NIR calibration method. In *Progress in Cereal Chemistry and Technology. Proceedings of 7th World Cereal and Bread Congress Prague*. J, Holas J Kratochvil (eds). Elsevier: Amsterdam, the Netherlands, 1983; 607.
44. Frank I, Kalivas J, Kowalski B. Partial least squares solutions for multicomponent analysis. *Anal. Chem.* 1983; **55**: 1800–1804.
45. Wold S, Martens H, Wold H. The multivariate calibration problem in chemistry solved by the PLS method. In Lecture Notes in Mathematics, A, Ruhe S Kågström (eds). *Proceedings Conference Matrix Pencils, March, 1982*; Springer: Heidelberg, Germany, 1983; 286–293.
46. Wold S, Ruhe A, Wold H, Dunn W. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalize dinverses. *SIAM J. Sci. Stat. Comput.* 1984; **5**: 735–743.
47. Martens H, Wold S, Martens M. A layman's guide to multivariate data analysis. In *Food Research and Data Analysis*. H, Martens H Russwurm (eds). Applied Science Publishers: London, 1983; 473–492.
48. Ståhle L, Wold S. Partial least squares with cross-validation for the two-class problem. A Monte Carlo study. *J. Chemom.* 1897; **1**: 185–196.
49. Haaland D, Thomas E. Partial least squares methods for spectral analysis: I Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.* 1988; **60**: 1020–1193.
50. Osten D. Selection of optimal regression via cross validation. *J. Chemom.* 1988; **2**: 39–48.
51. Martens H, Naes T. *Multivariate Calibration*. Wiley: Chichester, UK, 1989.
52. Faber N. A closer look at the bias-variance tradeoff in multivariate calibration. *J. Chemom.* 1999; **13**: 185–192.
53. Shao J. Linear model selection by cross validation. *J. Am. Stat. Assoc.* 1983; **88**: 313–486.
54. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C, Wold S. *Multi- and Megavariate Data Analysis. Part I. Basic Principles and Applications*. Umetrics Academy: Umeå, Sweden, 2006.
55. Hamilton L. *Regression with Graphics. A Second Course in Applied Statistics*. Duxbury Press: Belmont CA, 1991.
56. Clementi S. Personal Communication.
57. Wiklund S, Nilsson D, Eriksson L, Sjöström M, Wold S, Faber K. A randomization test for PLS component selection. *J. Chemom.* 2007; **21**: 427–439.

Copyright © 2010 John Wiley & Sons, Ltd.

58. Cruciani G, Baroni M, Clementi S, Costantino G, Riganelli D, Skagerberg B. Predictive ability of regression models. Part 1. Standard deviation of prediction errors (SDEP). *J. Chemom.* 1992; **6**: 335–346.
59. Baroni M, Cruciani G, Clementi S, Costantino G, Riganelli D, Oberrauch E. Predictive ability of regression models. Part 1. Selection of the best predictive PLS model. *J. Chemom.* 1992; **6**: 347–356.
60. Wakeling I, Morris J. A test of significance for partial least squares regression. *J. Chemom.* 1993; **7**: 291–304.
61. Forina M, Drava G, Boggia R, Lanteri S, Conti P. Validation procedures in near-infrared spectrometry. *Anal. Chim. Acta* 1994; **295**: 109–118.
62. Eriksson L, Johansson E, Wold S. QSAR Model Validation, In F Chen, G Scüürmann, (eds) *Quantitative Structure-Activity Relationships in Environmental Sciences – VII. Proceedings of the 7th International Workshop on QSAR in Environmental Sciences, June 24-28, 1996, Elsinore, Denmark* SETAC Press: Pensacola, Florida, 1997; pp. 381-397.
63. Denham M. Prediction intervals in partial least squares. *J. Chemom.* 1997; **11**: 39–52.
64. Wehrens R, van der Linden W. Bootstrapping principal component models. *J. Chemom.* 1997; **11**: 157–171.
65. Martens H, Dardenne P. Validation and verification of regression in small data sets. *Chemom. Intell. Lab. Syst.* 1998; **44**: 99–121.
66. Denham M. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *J. Chemom.* 2000; **14**: 351–361.
67. Filzmoser P, Liebmann B, Varmuza K. Repeated double cross validation. *J. Chemom.* 2009; **23**: 160–171.
68. Baumann K, Stiefl N. Validation tools for variable subset regression. *J. Comput. Aided Mol. Design* 2004; **18**: 549–562.
69. Kohonen J. Advanced chemometric methods: applicability on industrial data. *Dr of Science (Technology) Thesis*, Lappeenranta University of Technology. Acta Universitatis Lappeenrantaensis 353, ISBN 978–952-214- 814-8.
70. Gómez-Carracedo MP, Andrade JM, Rutledge DN, Faber K. Selecting the optimum number of partial least squares components for the calibration of attenuated total reflectance mid-infrared spectra of undersigned kerosene samples. *Anal. Chim. Acta* 2007; **585**: 253–265.
71. Varmuza K, Filzmoser P. *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press: Boca Raton FL, 2009; 103.
72. Czyzycki M, Bielewski M, Lankosz M. Quantitative elemental analysis of individual particles with the use of micro-beam X-ray fluorescence method and Monte Carlo simulation. *X ray Spectrom.* 2009; **38**: 487–491.
73. Sin G, Gernaey K, Lantz A. Good modeling practice for PAT applications: propagation of input uncertainty and sensitivity analysis. *Biotechnol. Prog.* 2009; **25**: 1043–1053.
74. Venkatapathy R, Wang C, Bruce R, Moudgal C. Development of quantitative structure-activity relationship (QSAR) models to predict the carcinogenic potency of chemicals I. Alternative toxicity measures as an estimator of carcinogenic potency. *Toxicol. Appl. Pharmacol.* 2009; **234**: 209–221.
75. van der Voet H. Pseudo-degrees of freedom for complex predictive models: the example of least-squares. *J. Chemom.* 1999; **13**: 195–208.
76. Esbensen KH, Paoletti C, Minkkinen P. Pitard F. (2009) Developing meaningful international sampling standards – where do we stand today? The world's first horizontal (matrix-independent) sampling standard. in Duggan, S. (Ed.) *Proceedings of 4th World Conference on Sampling and Blending (WCSB4)*, Cape Town Oct. 21-23, 2009. p.163. The South African Institute of Mining and Metallurgy, Symposium Series S59. ISBN 978-1-920211-29-5.
77. Pitard FF. 2009; *Pierre Gy's Theory of Sampling and C.O.Ingamell's Poisson Process Approach. Pathways to representative sampling and appropriate industrial standards. Dr of Technology Thesis.* Aalborg University, campus Esbjerg. Denmark. ISBN: 978-87-7606-032-9.

APPENDIX A

Summary of the Theory of Sampling (TOS) [1–23]

A.1. Introduction

Naturally occurring materials in science, technology and industry (including materials being processed in the analytical laboratory) are *heterogeneous* at all operative scales related to sampling. Therefore sampling cannot be satisfactorily carried out in practice, without a working understanding of the phenomenon of *heterogeneity* and how heterogeneity can be *counteracted* in the sampling process. All sampling process *interacts* with the heterogeneous material making up a lot. Because of this, sampling is far from trivial as all sampling procedures unavoidably will be affected by the heterogeneity of the lot material at all scales larger than the operative sampling tool. In addition, the sampling process creates sampling errors of its own, due to non-compliance with the practical, mechanical, maintenance and operative procedural tenets of TOS. For stationary lots this generates five principal types of sampling errors (in this paper process sampling is not covered, full details are found elsewhere [20,22]). The five stationary sampling errors suffice for full understanding of the principles of TOS.

The objective of representative sampling is directed at analyzing the conditions under which it is guaranteed that a reliable sample with an analyte concentration, a_S, sufficiently close to the true average lot concentration, a_L can be obtained. TOS shows that all such conditions rests with the sampling process; it is not possible to ascertain whether a specific 'sample' is representative or not from inspection of the sample itself.

A minimum understanding of TOS includes: heterogeneity, five sampling errors, the Fundamental Sampling Principle, lot dimensionality, proper methods for mass reduction, sampling correctness, seven Sampling Unit Operations (SUO) and the replication experiment.

A.2. Heterogeneity

Heterogeneity of stationary lots and materials has two fundamental aspects: Constitutional Heterogeneity (CH) and Distributional Heterogeneity (DH).

The constitutional heterogeneity represents the heterogeneity dependent on the physical or chemical differences *between* individual lot units, which TOS terms 'fragments'; *'grains*' is a useful imaginary metaphor for 'fragments' e.g. mineral grains, seed grains, kernels, biological cells. Any given target to be sampled (characterized by lot geometry, material type and - state, grain-size distribution etc.) exhibits a CH which is an inherent property of the lot. Thus CH plays out its role at the inter-grain scale of any lot. CH can only be reduced by *altering the physical state of the material*.

The distributional heterogeneity complements this characterization by describing all aspects of heterogeneity dependent upon the *spatial distribution* in the lot, as gauged by the operative sample size (volume/mass) used. This sampling unit can conveniently be imagined as the proverbial *sampling scoop*. The physical manifestations of DH are stratification, segregation and/or local groups-of-fragments concentrations with a significant higher, or lower, analyte concentration than the average lot concentration, a_L. DH can actively be reduced by using a suite of 'correct' sampling methods to be delineated further below. DH can never be larger than CH (in a sense DH is a complicated fraction of CH) and CH can *never* be strictly zero. Dependent on the purpose and scale of sampling (scoop size), CH *may* be close to negligible, but it is never nil. Homogeneity is defined as the (*theoretical*) limiting case of zero heterogeneity. If a homogeneous material did actually exist, sampling would not be needed—as all sampling errors would be zero, i.e. all 'samples' would be identical.

A.3. Constitutional heterogeneity (CH)

TOS defines a *heterogeneity contribution* to the total lot heterogeneity by firstly focusing on the individual fragments. TOS characterizes all fragments according to the component of interest (the analyte, *A*), described by the proportion (or grade), a_{i}, and the fragment mass, M_{i}. If a lot consists of N_{F} individual fragments with individual masses, M_{i}, with an average fragment mass, M_{i}^{-}, with lot grade a_{L} and a lot mass M_{L}, the heterogeneity contribution from each individual fragment, h_{i}, can be calculated as:

$$h_{\mathrm{i}} = \frac{(a_{\mathrm{i}} - a_{\mathrm{L}})}{a_{\mathrm{L}}} \cdot \frac{M_{\mathrm{i}}}{M_{\mathrm{i}}^{-}} = N_{\mathrm{F}} \frac{(a_{\mathrm{i}} - a_{\mathrm{L}})}{a_{\mathrm{L}}} \cdot \frac{M_{\mathrm{i}}}{M_{\mathrm{L}}}$$

Heterogeneity contributions are *dimensionless* intensive units. h_{i} delineates both the compositional deviations of each fragment, while also compensating for variation in the fragment masses; larger fragments result in a larger influence on the total heterogeneity than smaller ones. This viewpoint constitutes a major distinction from 'classical statistics' where all population units contribute equally (with equal statistical mass). h_{i} constitutes an appropriate compound measure of mass-weighed heterogeneity as contributed by each fragment in the lot.

The total constitutional heterogeneity of the lot, $\mathrm{CH_L}$, can further easily be defined as the variance of the distribution of the heterogeneity contributions of all fragments:

$$\mathrm{CH_L} = s^2(h_{\mathrm{i}}) = \frac{1}{N_{\mathrm{F}}} \sum_{\mathrm{i}} h_{\mathrm{i}}^2 = N_{\mathrm{F}} \sum_{\mathrm{i}} \frac{(a_{\mathrm{i}} - a_{\mathrm{L}})^2}{a_{\mathrm{L}}^2} \cdot \frac{M_{\mathrm{i}}^2}{M_{\mathrm{L}}^2}$$

A.4. Distributional heterogeneity (DH)

By ascending one scale level, from the scale of fragments to the operative level of one sampling unit (*sampling scoop*), one is able to cover the complementary realm of lot distributional heterogeneity, $\mathrm{DH_L}$. No longer concerned with the lot consisting of the totality of N_{F} fragments, any lot can alternatively be considered as being made up of a number of potential sampling volumes, N_{G}, commensurate with the operative volume of the sampling tool. Other than this hierarchical operative scale difference, the focus is identical, *viz* quantitative description of the differences in composition (concentration) of the analyte, *A*, *between* these sampling volumes (index *n*), a_{n}. $\mathrm{DH_L}$ can be calculated via a strict analog to the first definition of heterogeneity carried by a single fragment. A group-of-fragments, *group* for short (index *n*), G_{n}, similarly carries an amount, a contribution of the total lot heterogeneity, h_{n}, which can be calculated from the grade of the group in question, a_{n}, the group mass, M_{n}, the average group mass, M_{n}^{-}, and the average grade over all groups,a_{n}^{-}:

$$h_{\mathrm{n}} = \frac{(a_{\mathrm{n}} - a_{\mathrm{L}})}{a_{\mathrm{L}}} \cdot \frac{M_{\mathrm{n}}}{M_{\mathrm{n}}^{-}} = N_{\mathrm{G}} \frac{(a_{\mathrm{n}} - a_{\mathrm{L}})}{a_{\mathrm{L}}} \cdot \frac{M_{\mathrm{n}}}{M_{\mathrm{L}}}$$

The distributional heterogeneity for the entire lot can likewise be calculated as the variance of all group heterogeneity contributions:

$$\mathrm{DH_L} = s^2(h_{\mathrm{n}}) = \frac{1}{N_{\mathrm{G}}} \sum_{n} h_{\mathrm{n}}^2 = N_{\mathrm{G}} \sum_{n} \frac{(a_{\mathrm{n}} - a_{\mathrm{L}})^2}{a_{\mathrm{L}}^2} \cdot \frac{M_{\mathrm{n}}^2}{M_{\mathrm{L}}^2}$$

Due to the fact that the aggregate sum of all (virtual) groups constitutes the physical lot in its geometric entirety, it is clear that $\mathrm{DH_L}$, in fact, is a measure of the *spatial heterogeneity* exhibited by the lot.

This two-scale understanding of the heterogeneity of any lot (system, material)—fragments versus group-of-fragments—constitutes a most effective theoretical concept in TOS with which one is able to understand and deduce several important key issues of representative sampling of heterogeneous materials. $\mathrm{DH_L}$ accounts for the material heterogeneity in a specifically relevant form, namely that corresponding the specific sampling size (mass/volume) used, M_{S}. It is equally possible to ascertain the quantitative effect of the lot heterogeneity interacting with *alternative* sampling processes, for example using alternative sampling volumes.

TOS terms the fundamental sampling volume: the *increment*. All increments may either be used as for making up a *composite sample*, see further below, or it may be used as a single increment sample, termed a *grab sample*. The most important aspect of any sampling process is the size of the sampling unit M_{S}. From TOS, it is clear that a single-scoop sample is almost never acceptable (results in unacceptably inflated sampling bias, see further below), so M_{S} is nearly always to be understood as the compound mass of a composite sample, unless specifically stated otherwise.

Unlike for $\mathrm{CH_L}$, which is an intrinsic characteristic of the given material, $\mathrm{DH_L}$ can actively be altered (reduced), especially by choosing a smaller sampling tool thereby increasing the number of increments in composite sampling (in process sampling this means increasing the sampling frequency), and/or the lot can be thoroughly mixed, blended etc. In large lots, forced mixing is often impractical or impossible; in such cases increasing the number of increments is the only option for reliable sampling. If there is a significant segregation or grouping (fragment clustering) in the lot, increasing the sample size, M_{S}, only results in a comparatively minor effect and will soon reach an impractical limit. By way of contrast and effectiveness composite sampling is always a good choice of action. TOS has much to say (all negative) regarding the universal futility of grab sampling, which is *never* representative in practice against all realistic heterogeneous lots and materials. Grab sampling is never reliable and should accordingly be abolished.

It follows that sampling from a heterogeneous lot can never result in identical analytical results; there will always be a *sampling variance* (more accurately, a sampling_cum_analysis distribution) as expressed by a set of analytical results. Even a set of identically replicated samples (carried out following an identical *protocol*) will give rise to a distinct, non-zero sampling variance (see section below: *replication experiment*). This is solely due to the fact that no sampling process can eliminate the effect of heterogeneity for any lot—its role is to reduce this effect as much as possible, and to be able to quantify the remaining sampling variance. It *may* happen that particular systems *may* possess extraordinarily small heterogeneities etc. but no generalizations regarding universal relationships re. 'homogeneity' or 'sufficiently homogeneity' etc. can be drawn from such particulars. It is highly advisable always to treat any lot material as if it carried a significant heterogeneity.

A.5. Sampling error versus practical sampling

Analysis of the phenomenon of heterogeneity [1–4,14] outlines three factors which are responsible for the magnitude of the distributional heterogeneity:

- CH_L (constant for a given material)
- *Grouping* (depends on the size (volume/mass) of the extracted increments)

- *Segregation* (depends on the spatial distribution of fragments in the lot)

Both segregation and grouping can be quantified if need be; methods and equations are described in detail in the pertinent literature (and further references herein). More important is how to counteract the effects arising here from in practical sampling. In order to extract samples from heterogeneous materials with sufficiently low sampling variation it is necessary to minimize DH_L. For any *given material state*, the case of reducing the two phenomenological factors *grouping* and *segregation* can principally be achieved in only two ways:

- Decreasing the size of the extracted increments, thereby increasing the number of increments (or increasing the sampling frequency) combined to form a given sample mass, M_S (this approach *counteracts* grouping and segregation on the scale of the sampling tool volume).
- Mixing/'homogenizing' the lot (*reduces* macro-scale lot segregation)

If these measures are insufficient for a given sampling process and total error acceptance level, it will be necessary to reduce the constitutional heterogeneity itself, which necessitates physical reduction of the fragment sizes, *comminution* (grinding or crushing), and/or increasing the total sample mass, M_S. Comminution is by far the most effective of these two options, following:

$$\mathrm{var(FSE)} = C \cdot d^3/M_S$$

in which C is a material constant (constant for a given grain-size distribution state) and d is the top-diameter of the material (termed d_{95}) [1–14].

See further in Section Sampling Unit Operations.

A.6. Total Sampling Error (TSE)—Fundamental Sampling Principle (FSP)

All analytical results are associated with an analytical uncertainty, expressed as the variance of the TAE. Following analysis of the entire sampling process, TOS aggregates all other sources of error from sampling as the TSE. TAE and TSE together form the Global Estimation Error (GEE).

Figure 1 Zero-dimensional sampling errors and their TOS interrelationships.

TAE is often in good control in the laboratory, and is *usually* of only little concern in comparison to sampling, as TAE is always significantly smaller than the sum of all sampling errors, TSE. In fact TSE is very often 20–50–100 × larger than TAE [1–8]. Exceptions would reflect only very uniform materials with a truly exceptional small heterogeneity—such are rare indeed.

TSE has many sources. The objective of representative sampling is to identify, eliminate or reduce all contributing sampling errors. While much of the sampling procedure and sampling equipment issues and efforts are to some extent under control of the sampler, the part from constitutional heterogeneity is dependent on the material properties only. This error is termed the Fundamental Sampling Error (FSE), as it cannot be altered for any given system (lot, geometry, material, state, size distribution); it is FSE which is reduced by crushing/comminution. On the other hand, contributions from the spatial distribution of the material are not fixed and can more easily be altered. This is dependent not only on the material characteristics itself (DH_L) but also on the sampling procedure and whether which counteraction measures are invoked (if for example mixing can be applied before sampling). The variation stemming from distribution heterogeneity is represented by the Grouping and Segregation Error (GSE).

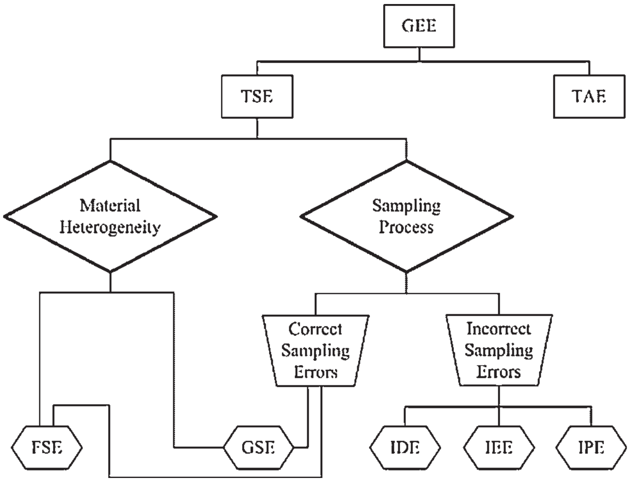

Figure A1. Zero-dimensional sampling errors and their TOS interrelationships.

All possible extractions from the lot (all possible virtual increments) must have the same probability of being selected, of being materialized. This critical stipulation is called the Fundamental Sampling Principle (FSP). FSP must *never* be compromised otherwise all possibilities of documenting accuracy (unbiasedness) of the sampling process are abandoned. FSP implies physical access to all geometrical units of the lot. TOS contains a wealth of practical guidelines of how to achieve compliance with FSP [1–16].

TOS employs a strict terminology, in which all aspects of non-compliant sampling can be specifically named. Thus, TOS specifies as 'correct' only those features that will contribute towards the ultimate goal of being able to demonstrate *representativeness* of the particular sampling process employed. To which further: The sum of FSE and GSE is termed the '*Correct Sampling Errors*' (CSE), as they are not due to erroneous sampling or wrong procedures; in fact CSE always occur to some degree, even when the sampling procedure is 'correct' (meaning accurate), hence their somewhat peculiar name. Errors that are connected to erroneous sampling procedures are contrarily summed as the *Incorrect Sampling Errors* (ISE).

ISE comprise four parts, one stemming from not *delineating* correct increments from the lot, the second from not *extracting* exactly what was delineated and a third form of error is induced after the extraction of the increment (or sample).

The *Increment Delineation Error* (IDE) can be avoided by always selecting (delineating) an increment that completely covers the relevant dimensions of the lot, for instance a complete cross-sectional slice if the lot is a (very) long pile of material, or a 'drill core' to the very bottom of the layer(s) of interest if the lot is a three-dimensional volume or of a similar shape. The *Increment Extraction Error* (IEE) arises when particles inside the delineated increment do *not* end up in the sample, for instance by bouncing off the increment tool edges—or by being blown away as dust, or if particles outside the delineated increment find their way into the sampling tool, contamination. The usual dictum is that only particles or fragments with their center of gravity inside a delineated increment should become part of the increment actually being extracted. TOS contains very detailed descriptions of all ISE and ditto recipes for their elimination.

The third incorrect error arises when the sample is altered *after* extraction, for instance by absorbing moisture, by spillage, cross-contamination or similar. Sample tampering and downright fraud is also a type of 'error' which is likewise collected under the term Incorrect Preparation Error (IPE). There is also another, a fourth ISE, the Incorrect Weighting Error (IWE), which sometimes plays a role too, but mostly concerning process sampling; IWE is not treated here.

All the incorrect errors can be minimized, in fact all can always be completely eliminated. This is the definition of 'correct' sampling: There are many practical, mechanical aspects of this issue, all relatively simple, almost trivial to implement, but only if they are properly recognized and one is willing to invest the work necessary. There are usually no cheap 'fixes' to a non-representative sampling procedure, but modification to a correct counterpart is neither an expensive approach, but mostly has to do with realizing and accepting often minor changes in procedures only. The TOS literature deals a great length with all these issues, with various types of focus on explaining the connection between the individual errors and how they can be minimized completely. The selected literature list in the text is comprehensive [1–23], and further references can be found in abundance here.

A.7. Seven Sampling Unit Operations

A set of simple *Sampling Unit Operations* (SUOs) has been formulated, which constitute a complete set of procedures and general principles regarding *practical* sampling [14].

These unit operations can be grouped according to their use:

Three general principles: normally utilized only once in planning or optimization of new or existing sampling procedures:

- Transformation of lot dimensionality (transforming 'difficult to sample' 2D and 3D lots to 'easy to sample' 1D lots). It is always possible to acquire some form of *specimen* from 3D and 2D lots, but whether this is based on probabilistic, correct, unbiased methods is a much more difficult issue—estimates of the primary sampling errors are difficult, sometimes impossible to come by, as are useful estimates of lot heterogeneity and composition a_L.
- Characterization of 0D sampling variation by a replication experiment
- Characterization of 1D (process) variation by variography [20,23]

Four practical procedures: often used several times over during practical sampling:

- Lot or sample homogenization by mixing or blending
- Composite sampling, using the smallest possible increments
- Particle size reduction (comminution)
- Representative mass reduction [6]

As but two examples of the use of SUO: If the Fundamental Sampling Principle appears difficult to uphold (for example for large stationary lots) Sampling Unit Operation #1 must be invoked [Lot Dimensionality Transformation]. In this fashion all 'impossible-to-sample' lots (includes also 2D, 3D lots) can in fact very often be transformed into a 1D lot configuration, by far the easiest configuration for representative sampling.

All primary sampling must employ composite sampling (SUO # 5), unless it has been specifically proven that acceptable sampling quality can be otherwise achieved based upon single increments e.g. using CV_{rel} (see immediately below).

The theory pertaining to the individual SUOs is explained in full in the TOS literature.

A.8. Replication experiment—quantitative sampling variance

The quantitative effect of DH_L interacting with a particular sampling process (i.e. a sampling process using a specific sample mass in a specific sampling plan (grab sampling, composite sampling, other . . .) can be quantified by extracting and analyzing a number of *replicate samples* 'covering the entire geometry of the lot' and calculating the resulting empirical variance of the analytical results a_S. Often a relatively small number of primary samples will suffice, though never less than 10. This procedure is termed a *replication experiment*.

The replication experiment must be governed by a fixed protocol that specifies how the sampling and analysis methods are to be repeated. It is essential that both primary sampling and all sub-sampling and mass-reduction stages, sample preparation etc. are replicated in an *identical* fashion. It is a critical requirement that all Incorrect Sampling Errors (ICS) have been eliminated, i.e. that only *correct sampling* is employed. This principle is called TOS' preventive paradigm.

It is possible to employ a standard statistic to this type of replication experiment. The relative coefficient of variation, CV_{rel} is a very useful measure of the magnitude of the standard deviation (STD) in relation to the average (X_{avr}), often advantageously expressed as a simple percentage:

$$CV_{rel} = [(STD)/X_{avr}] \times 100$$

This *in toto* sampling variance is specifically influenced by the specific heterogeneity of the material *as expressed by* the current sampling procedure. When it is observed that all sampling errors (TSE), primary sampling error, secondary, tertiary, etc. including all errors incurred by mass reduction [6]—it transpires that CV_{rel} is a particularly apt summary characterization of the GEE. It is very convenient to use CV_{rel} for initial characterization of an existing sampling procedure—as well as to compare the numerical %-age resulting from modified, hopefully improved procedures.

Currently international efforts are aimed at formulating the world's first so-called 'horizontal sampling standard' (matrix-independent). This work has a.o. also focused on developing a rationale for specification of an authoritative threshold level for CV_{rel} as a practical maximum acceptable sampling variance for *significantly heterogeneous materials and systems*. A level corresponding to 35% has been suggested preliminarily based on extensive practical experience [76]. While applying well to all such systems, this quantitative threshold level is not to be viewed as a universal excuse to let go of individual responsibility however. It is

Copyright © 2010 John Wiley & Sons, Ltd.

important to acknowledge that for many systems in which the heterogeneity is substantially less (so-called *uniform systems*), the CV_{rel} threshold should be set as low as 15–20%, which is the level at which the sampling process takes on the characteristics of a Poisson process [77].

The only way to perform documentable and reliable representative sampling is by adhering to TOS' preventive paradigm:

(1) Elimination of all Incorrect Sampling Errors (eliminate IDE, IEE, IDE, IWE)
(2) Reduction of the remaining sampling variance (reduce FSE + GSE)

Failure to comply with stipulation 1 guarantees development of an inconstant sampling $bias_{TOS}$, which is impossible to estimate and therefore also impossible to correct for. History and the literature is ripe with examples of unawareness of TOS, sampling bias and the unavoidable consequences hereof, which range from important-to-substantial economic loss, to worse . . ., to fatal . . ..

Lack of proper attention to stipulation 2 is tantamount to missing out on due diligence. Whether in science, technology and industry it is not enough to address only the TAE. The overwhelming measurement uncertainty issues always lie with ill-informed, improper, non-representative sampling. This has the above direct bearing on the issues of proper validation as well.

Copyright © 2010 John Wiley & Sons, Ltd.