Special Issue

JOURNAL OF CHEMOMETRICS

Received: 5 June 2007, Accepted: 18 October 2007, Published online in Wiley InterScience: 9 July 2008

(www.interscience.wiley.com) DOI: 10.1002/cem.1103

Subset selection strategy

Oxana Y. Rodionova[a]* and Alexey L. Pomerantsev[a]

A new technique for representative subset selection is presented. The advocated method selects unambiguously the most important objects among the calibration set and uses this subset for the model development without significant deterioration in the predictive ability. The method is called boundary subset selection and it is an inherent part of the Simple Interval Calculation (SIC) approach. SIC is a method for linear modeling, which is based on the assumption of error boundedness. The primary SIC consequence is an object status classification (OSClas) that reveals the most influential objects and also designates the most stable and reliable ones. The OSClas is used as the main tool for representative subset selection. The presented results are compared with widely used Kennard–Stone algorithm and D-optimal design procedure employing three real-world examples. Copyright © 2008 John Wiley & Sons, Ltd.

Keywords: representative subset selection; projection methods; SIC method; object status classification

1. INTRODUCTION

Often, in calibration transfer [1], in applying multivariate analysis to the industrial data [2,3], and also for other data analytical objectives, it is necessary to select a special subset from a large(r) calibration set that shall bear the burden of *representing* the entire relevant data model. In general, such a subset must satisfy two opposing requirements: (1) it should be of maximal representativity with respect to the entire set, but (2) it should simultaneously be noticeably smaller than the total set. There have been presented several solutions to this dilemma [2–5], to which we append that of the Simple Interval Calculation (SIC) method. SIC is a method of linear modeling that gives the result of prediction directly in the interval form [6,7] and also provides wide possibilities for the leverage-type OSClas. In subsection 2.1 a brief description of this method is presented. The results of the SIC design for representative subset selection are also compared with two well-known techniques: the Kennard–Stone [8] design and the D-optimal design [9]. All these methods may be used in a situation when no standard experimental design can be applied and all objects are candidates for the representative subset.

It is known that representativity is a vague term, which can be interpreted in different ways. The goal of the paper is to present a technique that selects unambiguously the most important objects among the calibration set and to use this subset for model development without significantly deterioration in the predictive ability of the model. Most likely this subset should be called not a representative, but an influential subset.

An important issue inevitably arises here. How can the predictive ability of a model be evaluated? There are numerous techniques on how to control this, but generally accepted approach still does not exist [10]. The application of the Root Mean Square Error of Prediction (RMSEP) [11,12] calculated with an independent test set should be applied with care as it greatly depends on the 'quality' of this test set. For example if all objects of the test set are situated closely to the centroid of the training set, the predictive ability of the model may be overoptimistic. The inherent flaw of RMSEP is that such a characteristic of predictive ability evaluates the model only on average. Therefore RMSEP is a necessary, but not a sufficient characteristic of model quality. So, there is a need for a more careful comparison of different models' predictive ability using 'traditional' influence plots and/or their novel analogs called the object status plots (OSP) originated form the SIC modeling.

2. THEORY

2.1. SIC basics principles

Let us consider a linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{a} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y}$ is the n-dimensional response vector, $\boldsymbol{a}$ is the p-dimensional vector of unknown parameters, $\mathbf{X}$ is the $(n \times p)$ predictor matrix, $\boldsymbol{\varepsilon}$ is an unknown error vector; ordinarily rank of matrix $\mathbf{X}$ is less than p.

The SIC approach is based on a single assumption that all errors, $\boldsymbol{\varepsilon}$, involved in calibration problem (1) are *limited* (measurement errors in $\mathbf{X}$ and $\boldsymbol{y}$, modeling errors, *etc.*) [7]. The error finiteness means that there exists a maximum error deviation (MED) of error ε, which equals β, that is

$$\exists\beta > 0 \ \mathrm{Prob}\{|\varepsilon| > \beta\} = 0, \text{ and for any } 0 < b < \beta \ Prob\{|\varepsilon| > b\} > 0 \tag{2}$$

where Prob{•} denotes probability that an event occurs. Relying on assumption in Equation (2), and employing given calibration data set ($\mathbf{X}$, $\boldsymbol{y}$) with n samples, it is possible to build the entire

* *Institute of Chemical Physics, Kosygin Str. 4, Moscow 119991, Russia. E-mail: rcs@chph.ras.ru*

a *O. Y. Rodionova, A. L. Pomerantsev*
Institute of Chemical Physics, Kosygin Str. 4, Moscow 119991, Russia

system of inequalities regarding the unknown regression parameters $\boldsymbol{a}$,

$$A = \{\boldsymbol{a} \in R^p : \boldsymbol{y}^- < \boldsymbol{X}\boldsymbol{a} < \boldsymbol{y}^+\}, \text{ where } y_i^- = y_i - \beta,\ y_i^+ = y_i + \beta \quad (3)$$

A is a closed convex set in the parameters' space; it is called the *Region of Possible* (parameter) *Values* (RPV). This is a volumetric analog of the conventional parameter point estimates vector $\hat{\boldsymbol{a}}$, which is calculated by some traditional regression method, for example PLS.

Using the obtained RPV it is possible to solve a prediction problem for any given predictor vector $\boldsymbol{x}$ (e.g. a new spectrum or similar). If parameter $\boldsymbol{a}$ varies over A, it is clear that the predicted value $\boldsymbol{y} = \boldsymbol{x}^t\boldsymbol{a}$ belongs to the interval

$$V = [v^-, v^+], \text{where } v^- = v^- = \min_{\boldsymbol{a}\in\boldsymbol{A}}(\boldsymbol{x}^t\boldsymbol{a}),\ \boldsymbol{v}^+ = \max_{\boldsymbol{a}\in\boldsymbol{A}}(\boldsymbol{x}^t\boldsymbol{a}) \quad (4)$$

The interval V is the result of SIC prediction. In order to find this interval it is not necessary to build RPV explicitly, as the solutions of Equation (4) may be obtained by linear programming methods [13], which are commonly used to find the optima of a linear function on a convex set. However, the limited solutions of a linear programming problem can be found if, and only if the set A is bounded, that is $\mathbf{X}$ is a full-rank matrix. In the opposite case it is necessary to apply a regularization procedure, for example the PLS projection, and further on use a score matrix $\mathbf{T}$ instead of $\mathbf{X}$ in the SIC method [7].

Usually the MED value is unknown and some estimate b is used instead of β. In the present work two β estimates are used. Estimator $b_{\min}$ is defined as follows

$$b_{\min} = \min\{b,\ A(b) \neq \varnothing\} \quad (5)$$

This is a consistent but biased ($b_{\min} \leq \beta$) estimate and $b_{\min}$ is the lower limit of all possible β values. To estimate the upper limit of β we apply a traditional statistical approach [14] to the regression residuals $\boldsymbol{e} = \hat{\boldsymbol{y}} - \boldsymbol{y}$. Therefore it is possible to find an estimator b_{SIC} such that $\text{Prob}\{b_{SIC} > \beta\} > 0.90$ and b_{SIC} is as close to β as possible. This enhanced estimator, b_{SIC}, can be calculated by formula [7]

$$b_{SIC} = b_{\max} C(n, s^2) \quad (6)$$

Here $b_{\max} = \max(|e_1|, \ldots, |e_n|)$ and empirical function $C(n, s^2)$ depends on n that stands for the number of objects in the calibration set, and on the residual variance s^2.

To quantify the quality of SIC prediction two measures are used [7]. The *SIC residual* is the difference between the center of the prediction interval (4) and the reference value y, (scaled by β), so this is a characteristic of bias:

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{1}}{\beta}\left(\boldsymbol{y} - \frac{\boldsymbol{v}^+(\boldsymbol{x}) + \boldsymbol{v}^-(\boldsymbol{x})}{\boldsymbol{2}}\right) \quad (7)$$

The *SIC leverage* is calculated as the width of the prediction interval, divided by the MED β, so it has the character of β-normalized precision:

$$h(\boldsymbol{x}) = \frac{\boldsymbol{1}}{\beta}\left(\frac{\boldsymbol{v}^+(\boldsymbol{x}) - \boldsymbol{v}^-(\boldsymbol{x})}{\boldsymbol{2}}\right) \quad (8)$$

In paper [6] a new OSClas concept was proposed. It was shown that this classification could easily be performed without an explicit construction of the complex RPV in the parameter space. It is instead based on the following statements.

Statement 1. An object ($\boldsymbol{x}$,y) is an insider, iff $|r(\boldsymbol{x},y)| \leq 1 - h(\boldsymbol{x})$;
Statement 2. Calibration object ($\boldsymbol{x}$,y) is a boundary object, iff $|r(\boldsymbol{x},y)| = 1 - h(\boldsymbol{x})$;
Statement 3. An object ($\boldsymbol{x}$,y) is an outlier, iff $|r(\boldsymbol{x},y)| > 1 + h(\boldsymbol{x})$;
Statement 4. An object ($\boldsymbol{x}$,y) is an absolute outsider (explained below) for any y, iff $h(\boldsymbol{x}) > 1$.

Using these statements one can construct an OSP [6], the archetype of which is shown in Figure 1. This OSP has the same appearance for *any* dimensionality of the initial data ($\mathbf{X}$, $\boldsymbol{y}$) and for *any* number of model parameters, which makes it a very powerful tool. Statements 1–4 divide the SIC residual (r) versus SIC leverage (h) plane into three areas, each corresponding to one of the three object categories: insiders (area i in Figure 1), outsiders (area ii) and outliers (area iii). A sample, for which the SIC leverage is greater than one ($h > 1$, area iia in Figure 1, sample 3), cannot be classified as an insider (area i) for any response value. Such samples form a special class of objects, which are called absolute outsiders. According to the SIC approach, *all calibration objects are insiders*. Moreover there are objects that play a special role in calibration. They are boundary samples, or boundary objects (Statement 2), and these objects are of critical significance for model construction. This is because the RPV is not formed by all objects from the calibration set, but only by these boundary samples. Therefore, if one excludes all objects from the calibration set except boundary samples, the RPV will not change (sample 5 in Figure 1 represents a boundary object.). This means that *boundary objects are the most influential ones*, at least from the SIC point of view.

2.2. Kennard–Stone design

The Kennard–Stone [8] design selects a set of objects, which are 'uniformly' distributed over the space defined by the candidates. This is a classic method to extract a representative set of objects from a given data set. In this method the objects are chosen

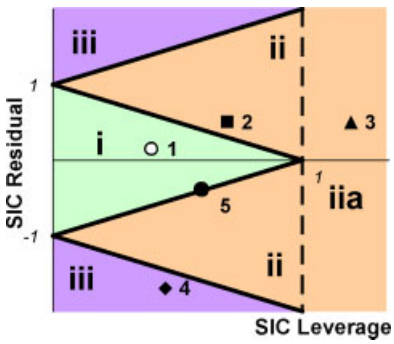

Figure 1. OSP. i: insiders (○, 1); ii: outsiders (■, 2); iia: abs. outsiders (▲, 3); iii: outliers (◆, 4). This figure is available in color online at www.interscience.wiley.com/journal/cem

sequentially. The first two samples are selected by choosing the two farthest apart from each other. The third object selected is the one farthest from the first two samples, etc. Supposing that k objects have already been selected ($k < n$), the $(k+1)$th object in the training set is chosen using the criterion

$$\max_{k<r\leq n}\left(\min(d_{1r}, d_{2r}, \ldots d_{kr})\right) \tag{9}$$

Here d_{jr}, $j = 1,\ldots k$, are the squared Euclidean distances from a candidate object r, not yet included in the representative set, to the k objects already included in the representative set. The procedure is very simple and intensively used [4,5,15–17]. One more advantage of the Kennard–Stone algorithm is that it may be applied to any matrix of predictors; there are no restrictions regarding the matrix multicollinearity.

2.3. D-optimal design

The D-optimal design is a frequently used method [4,18–20]. The principle of this method is to select the objects in order to maximize determinant of the information matrix $|\mathbf{X}^t\mathbf{X}|$ for the linear regression model. Fedorov's [9] exchange algorithm is used for this purpose. Procedure starts with an initial design of the requested size. In each iteration, each point in the design is compared with each point in the candidate list, and the exchange is made for the pair that optimizes the design. The iterative algorithm is terminated when there are no further improvements in the optimality criterion. Samples selected with this criterion are located at the border of the calibration domain. So, D-optimal design, in contrast to the Kennard–Stone procedure, aims to select the most influential, peripheral objects. On the other hand, when the number of variables is larger than the number of objects, D-optimal design cannot be applied directly because of singularity of information matrix.

3. CASE STUDY OVERVIEW

Different strategies of subset selection were investigated for three real world examples. The first data set represents the wheat calibration [6]. To demonstrate various aspects of a subset selection this data set is analyzed in details. The second data set contains the acoustics spectra used for quantitative determination of trace oil concentrations in water [21]. This example evaluates subset selection in application to a rather small data set and a simple two-dimensional PLS model. In the third example, the multistage technological process is under consideration [22]. This data set demonstrates the influence of the representative subset size on the predictive ability of a model.

In each example the various sets and subsets of samples are considered. To avoid any confusion and in order to make the statement more clear, the following systematic notation is used. The G set is the entire set of objects (samples) under investigation. This is further divided into the calibration C set and the test T set, which together give the G set, that is C set + T set = G set. The numbers of objects in these sets are noted by N_G, N_C and N_T, correspondingly. The PLS model calibrated over C set and validated over T set is named Model_C. The number of PCs in this model, L, is fixed to be used in further PLS modeling. After that, some subset selection algorithm is applied to the C set in the PLS score domain calculated for Model_C. This procedure designates objects to be included in a representative subset. Such a subset is called B set or K set or D set (SIC, Kennard–Stone or D-design) with respect to the method applied. Notations N_B, N_K and N_D are used for the number of objects in such subsets. The rest of C set samples that are out of the representative subset are named redundant objects. They form a set, which is called RB or RK or RD set, in dependence of the algorithm. It is clear that any representative subset, for example B set, jointed with the corresponding redundant set, for example RB set, gives the C set, that is B set + RB set = C set. The new PLS models which are recalibrated over B, K and D sets and validated over T set, are named Model_B, Model_K and Model_D, correspondingly. All of them employ the same number of PCs as Model_C, but span the different PLS spaces. The values of the root mean squared error of calibration (RMSEC) calculated with respect to the corresponding model are named by RMSEC_C, RMSEC_B, RMSEC_K and RMSEC_D. They are calculated over their own training sets, for example

$$\mathrm{RMSEC\_C} = \sqrt{\frac{1}{N_C - L}\sum_{n=1}^{N_C}(y_n - \hat{y}_n)^2}$$

where y_n are the C set reference values and $\hat{y}_n$are the corresponding calibrated values. Similarly, the values of the RMSEP are marked by RMSEP_C, RMSEP_B, RMSEP_K and RMSEP_D. They are calculated over the same T set using the pertinent model

$$\mathrm{RMSEP\_C} = \sqrt{\frac{1}{N_T}\sum_{n=1}^{N_T}\left(y_n - \hat{y}_n^C\right)^2}$$

where y_n are the T set reference values and $\hat{y}_n^C$ are the corresponding values, predicted with Model_C. Sketches of all data sets and models under investigation are presented in the flow-chart (Figure 2).

For each example several SIC models are constructed using the scores matrices obtained by the corresponding models: Model_C, Model_B, etc. However, for each SIC model, the same b_{SIC} estimate is used. This value is calculated for Model_C and it is not revised further as well as the number of PCs obtained from Model_C.

To show that a subset X (X may be B, K or D) is indeed representative the following procedure is undertaken:

1. Build a PLS model, Model_X, based on the X set with the given number of PCs, and construct the corresponding SIC model with the given bSIC value.
2. Validate Model_X using the test T set
3. Employ Model_X for prediction of samples from the redundant RX set.
4. Compare the results of calibration and prediction with results obtained from Model_C. (10)

A number of models or subsets characteristics are used in the examples. Among them there are the SIC and PLS residuals that are calculated with respect to y values using Equation (7) for SIC and formula $r_{PLS} = y - \hat{y}$ for PLS. The SIC and PLS leverages are calculated by Equation (8) for SIC and standard formula for PLS [11]. In each case it will be clear which model (C, B, K or D) is applied for calculation of a specific characteristic.

In graphical presentations a special marking will be used over the paper in all figures (Figure 3–Figure 7). All objects from the

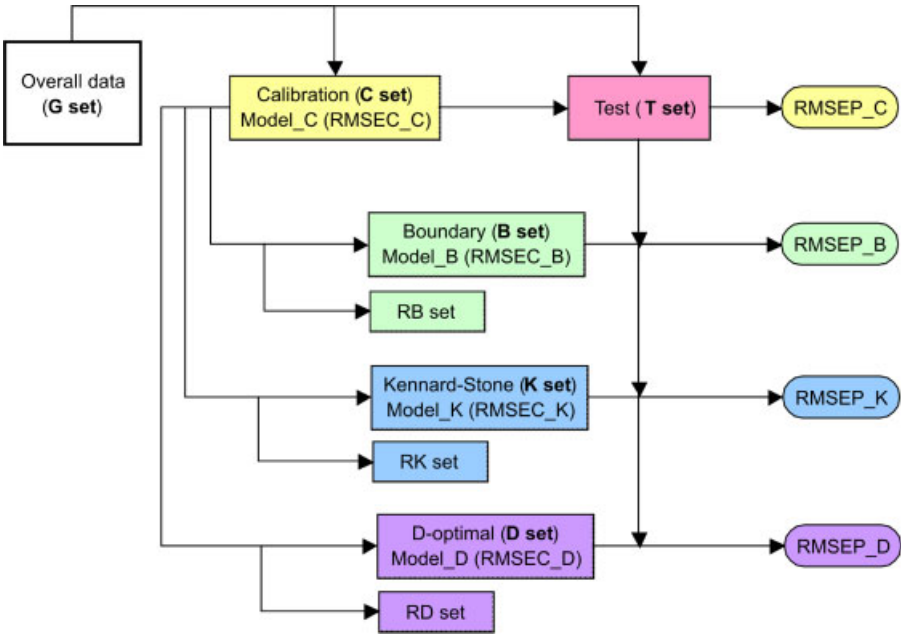

Figure 2. Data sets and models under consideration. This figure is available in color online at www.interscience.wiley.com/journal/cem

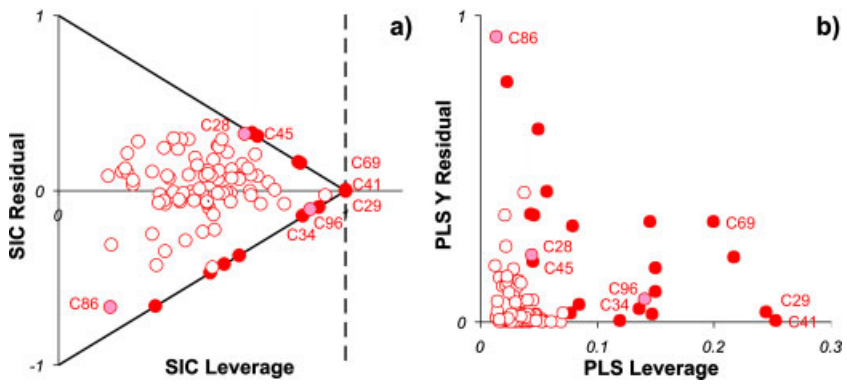

Figure 3. Wheat data: Model_C with four PLS components. Calibration set: ○, insiders; ●, boundary samples. (a) SIC OSP; (b) influence plot. This figure is available in color online at www.interscience.wiley.com/journal/cem

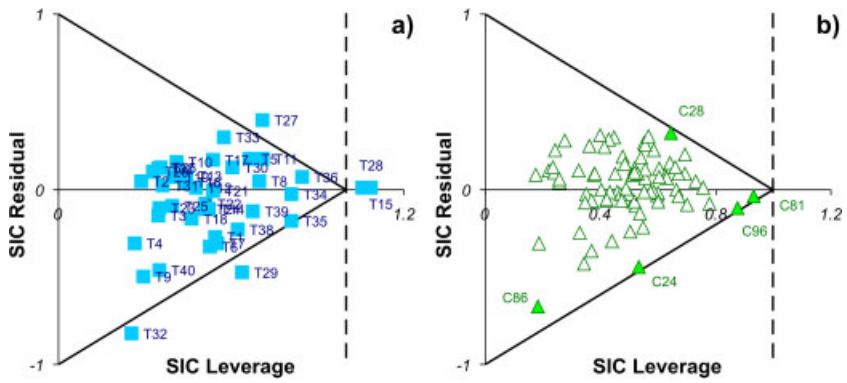

Figure 4. Wheat data: SIC OSP for Model_B. Prediction of the T set and the RB set (a) ■, test objects; (b) △-RB objects. This figure is available in color online at www.interscience.wiley.com/journal/cem

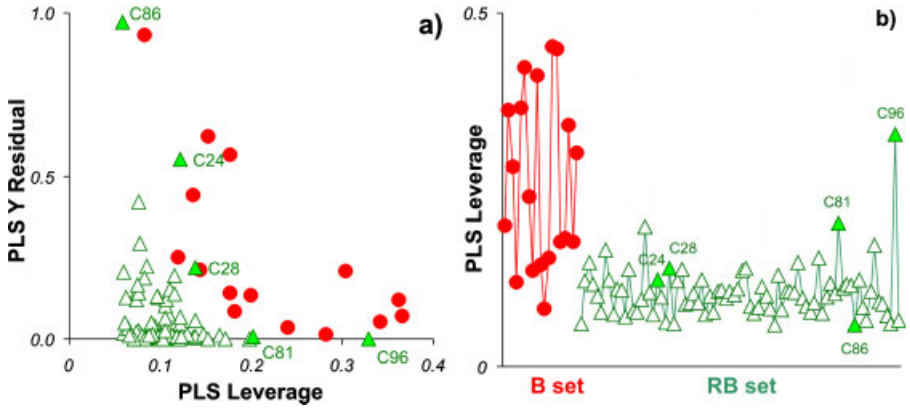

Figure 5. Wheat data: Four PLS component Model_B. ●, B set; △, RB set. (a) PLS *y* residuals versus PLS leverages; (b) PLS leverages for the B and RB sets. This figure is available in color online at www.interscience.wiley.com/journal/cem

calibration set, in spite of their further role in different models, are marked by dots. These objects are further annotated with their SIC designations: filled dots represent *boundary samples*, while open dots present insiders. All objects from the test set are denoted by squares and objects from the redundant sets are marked by triangles. The shaded markers emphasize the objects, which may be of special interest, and therefore mentioned in the text.

4. CASE STUDY 1: DETERMINATION OF WATER IN WHOLE WHEAT FROM NIR SPECTRA

4.1. Exploratory analysis

The **X** matrix consists of NIR spectra in the range of 908–1120 nm, recorded at 118 wavelengths; the reference ***y*** vector includes

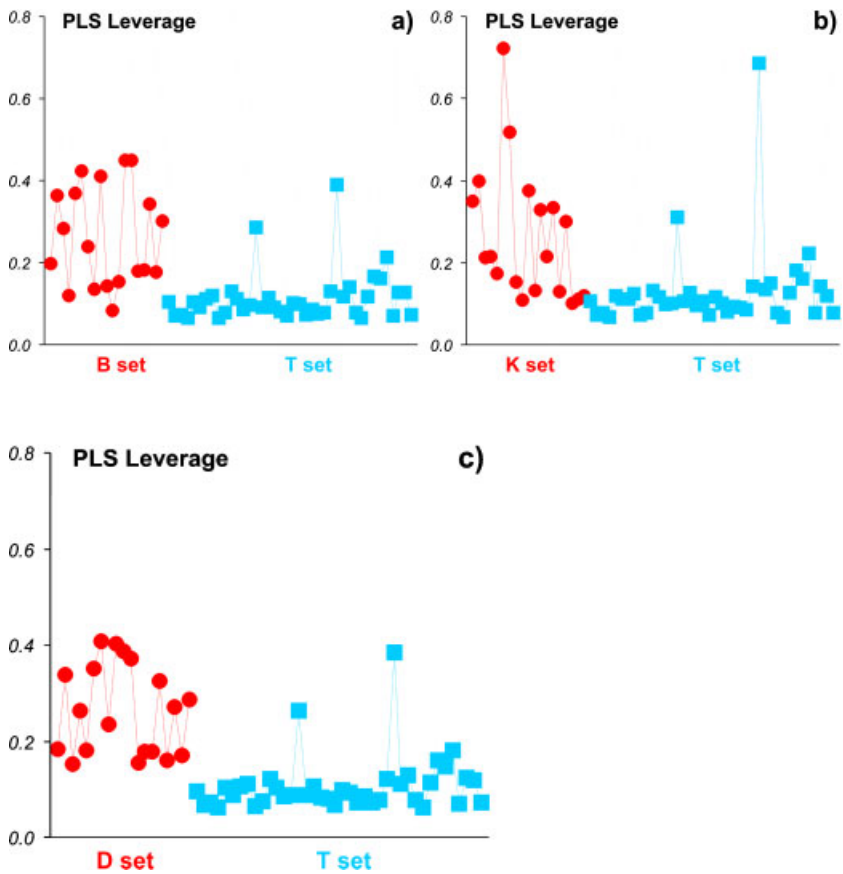

Figure 6. Wheat data: PLS leverages for various models. ■, test T set; ●, training subsets (B, K and D sets). (a) Model_B; (b) Model_K (c); Model_D. This figure is available in color online at www.interscience.wiley.com/journal/cem

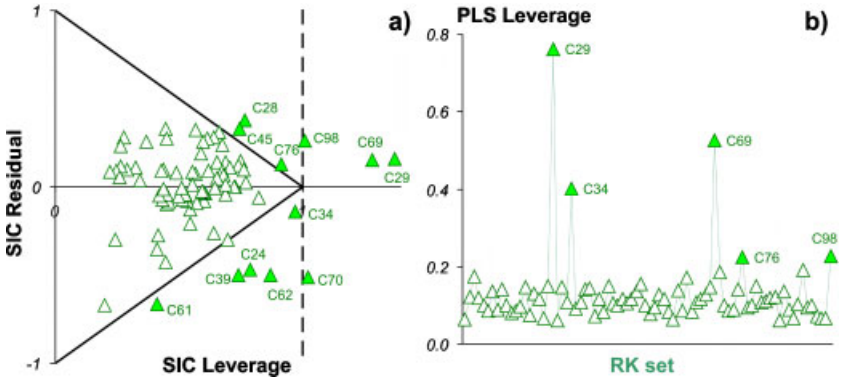

Figure 7. Wheat data: Model_K with four PLS components. △, RK objects; ▲, marked RK objects. (a) OSP for RK set; (b) leverage versus object number. This figure is available in color online at www.interscience.wiley.com/journal/cem

moisture content of 139 samples as quantified in the laboratory by a standard analytical method (evaporation loss-of-weight). Outliers are already removed. The data are centered and scaled to unit variance. For the beginning, the entire data set (139 objects) is used for modeling. Employing 10%-out cross-validation, the PLS model with four components is built. Summary results of PLS regression are presented in Table I.

The first column of Table I reports dimension. The next column shows cumulatively how much of **X** has been used. For example the first score vector accounts for 31.84% of **X**-variance, two score vectors account for 86.86%, etc. Similarly column 3 shows how much of the variation of ***y*** is explained. Last column shows the correlation coefficient r_{tu} between the score vector ***t*** and the associated ***u*** vector. There is a noticeable relationship between the first four sets of vectors.

This PLS subspace is used further for statistical comparison of the training and test sets. It is known [2] that the results of comparison depend not on the data itself, but also on the working variable-subspace, which in its turn depends on the calibration model, that is the number of PLS components.

Using the established PLS model, a pertinent SIC model is built and β estimates are calculated, namely $b_{min} = 1.03$, $b_{SIC} = 1.5$. The results of this modeling, that is the number of PLS components and β estimates are used for further analysis.

4.2. Data analysis with calibration set and test set: Model_C

Applying random selection, the initial data set is divided into the calibration set comprising 99 objects and the test set containing 40 objects. As it is mentioned in the Section 'Introduction', the quality of the test set with respect to the calibration set and constructed model is very important for reliable evaluation of model prediction ability. The generalization of Bartlett's test is applied for comparison of variance–covariance matrices of two data sets. This test lets us compare the orientation of the clouds of points in space as well as compare the dispersion of the data around their respective means. Additionally, the Hotelling T^2 test checks that two centroids are situated at the same place in space. The implementation of these two statistical tests is described in detail in Reference [2]. Statistical comparison of the calibration and test sets gives the following results, $C = 12.3$ ($C_{crit} = 18.4$) for Bartlett's test and $F = 0.99$ ($F_{crit} = 2.44$) for the Hotelling T^2 test.

The PLS modeling of the calibration set produces a four-component model. The summary results of PLS regression are presented in Table II and they are similar to those reported in Table I. External validation is done by means of the test set. In accordance with the general rules of notation given in Section 3, this model is named as Model_C.

Using Model_C, the SIC model and the corresponding OSP [Figure 3(a)] can be established. Following the OSClas rules, 19 calibration objects that are located on the border of the triangle (Statement 2) are boundary samples. Further, all the objects from the calibration C set are named C1, C2,. . ., C99. All the objects from the test T set are denoted T1, T2,. . ., T40.

Comparing the SIC OSP, [Figure 3(a)] and the traditional influence plot [11,12] [Figure 3(b)] one can see that all the most influential samples revealed from the influence plot are at the same time the boundary samples determined by OSClas. For example objects C29, C41 and C69 are boundary samples and

Table I. Wheat data [PLS modeling for initial data set (G set, 139 objects)]. Bold digits indicate the optimal model complexity and characteristics corresponding to chosen PCs

PCs	**X**%-explained	***y***%-explained	RMSEC	RMSEP (CV)	r_{tu}
1	31.84	51.42	0.622	0.638	0.72
2	86.86	61.93	0.551	0.565	0.47
3	95.58	81.51	0.384	0.400	0.72
4	**99.50**	**89.65**	**0.288**	**0.299**	**0.66**
5	99.70	89.81	0.285	0.301	0.13
6	99.77	90.11	0.281	0.300	0.17

Table II. Wheat data (general characteristics for PLS Model_C). Bold digits indicate the optimal model complexity and characteristics corresponding to chosen PCs

PCs	**X**%-explained	***y***%-explained	RMSEC_C	RMSEP_C	r_{tu}
1	30.15	51.80	0.639	0.504	0.72
2	86.76	62.83	0.561	0.496	0.48
3	95.44	82.98	0.379	0.354	0.74
4	**99.45**	**90.74**	**0.280**	**0.313**	**0.63**
5	99.70	90.94	0.277	0.311	0.13
6	99.74	91.49	0.268	0.313	0.20

they may be treated as important judging by both plots, OSP and influence plot. If there are similar objects, only one of such objects is included in the boundary subset. Influential object C28 is an insider, but it is located very close to the border and to the boundary object C45. The same situation can be seen with C96 (insider) and C34 (boundary). In the influence plot these samples are also located very close to each other. So, the SIC OSClas helps to reveal all important samples in the calibration dataset (Statements 1–2).

The comparison of OSP and influence plots (Figure 3) demonstrates that the concept of boundary samples makes sense not only inside the SIC approach, but it also characterizes the data set structure optimally and may be useful for a representative subset selection. At the same time OSP provides a strict numerical distinction between the boundary and non-boundary objects.

4.3. Boundary subset: Model_B

According to the OSClas concept, boundary samples detected by the SIC model are then the most important objects for modeling and all the other samples may be treated as redundant in this sense. Further, the boundary set (B set, $N_B = 19$) and 'redundant' set (RB set, $N_{RB} = 80$) will be distinguished. The RB set is formed by the calibration objects (C set) that are not included in the B set. It can be expected that (1) the B set may be used instead of the calibration C set for model building; (2) the objects from the redundant RB set may be reliably predicted with the model calibrated over the B set. From the regression point of view, there should be a rather small RMSEP for the RB set. From the SIC point of view, all these objects should be insiders.

To show that the B set is representative the procedure presented in Equation (10) was applied. As it may be seen from Table III (rows 1, 2), the calibration error (RMSEC_B) for Model_B is significantly worse than that for Model_C. This may be easily explained as for Model_B only the most peripheral, influential objects in calibration set were used. All objects that are located closer to the center of the model were eliminated. However, just these objects have small residuals and small leverages simultaneously.

On the other hand, there is no significant deterioration in predictive ability of Model_B in comparison with Model_C.

But this is the comparison of two models on average. Now let us analyze the uncertainty in prediction for individual objects from the test T set. The SIC status of samples from the T set determined by both models is absolutely the same, therefore only one OSP plot [Figure 4(a)] is presented here. The SIC leverage for each object from the T set calculated by Model_C coincides with the corresponding SIC leverage calculated by Model_B. This implies that the width of the SIC prediction interval calculated by Model_C and Model_B is similar for each object from the test set.

Several outsiders, T15, T27, T28, T29, T32 confirm that some objects from the T set differ from the B set objects. The presence of the mild outsiders among the test samples is important for the trustworthy validation of calibration model stability.

Now let us consider how Model_B predicts insiders, that is the objects collected in the RB set. If the boundary approach is correct, then the RB set consists of the most stable samples, with small leverage and small residual values. First, let us consider the results of the 'traditional' PLS prediction. In Figure 5(a), the PLS *y* residuals are plotted against the PLS leverages. The boundary samples (closed dots) are the most influence samples. On the other hand, PLS leverages for all boundary samples do not vary seriously [Figure 5(b)]. It is worth mentioning that there are no training objects (B set) with very low or very high leverages. This means that all objects play practically the same role in model building. As for the test objects (from the RB set), all of them have rather low leverages and from moderate to low residuals. Only

Table III. Wheat data: summary table (prediction ability of different PLS models with four components)

Model	Training set	RMSEC	Test set	RMSEP
Model_C	C set	0.280	T set	0.313
Model_B	B set	0.426	T set	0.330
			RB set	0.246
Model_K	K set	0.212	T set	0.339
			RK set	0.311
Model_D	D set	0.266	T set	0.333
			RD set	0.305

object C86 has a high PLS residual, but its PLS leverage is the lowest [Figure 5(a)].

Thus one can conclude that all RB samples are close to the center of the model and reliable in prediction. Prediction reliability is confirmed by Table III (row 3).

At the same time, the SIC OSP [Figure 4(b)] shows that all samples from the RB set are indeed *insiders*. It may be interesting to analyze the location of some objects in the influence plot [Figure 5(a)] in comparison with OSP [Figure 4(b)]. Object C86 has a high PLS residual [Figure 5(a)] and it also has a high SIC residual [Figure 4(b)], though it is located in the insider area as its SIC leverage, and at the same time PLS leverage is rather low. Object C96 has the highest PLS leverage among the RB samples [Figure 5(b)] and at the same time it is located very close to the insider border and even to the border of absolute outsiders.

Thus it can be stated that the goal has been reached: (1) the model constructed with the help of the selected subset can predict the test samples with the accuracy that is not worse than the prediction error evaluated on the whole data set; (2) the high accuracy of prediction for the RB objects conform the 'redundancy' of these samples for model construction; (3) the boundary set is indeed significantly smaller than the training set, 19 samples out of 99.

It is worthy of mentioning that RMSEP calculated for the RB set is rather small (see Table III). This may be easily explained by the concept of the boundary samples. After the selection of the most important (boundary) objects from the calibration set, the samples that are left are the most 'average' ones and are situated closer to the center of the model and have low *y*-residual values too. That is why prediction of these samples shows much lower value of RMSEP. This should be taken into consideration when the calibration and test sets are selected. If test set consists only of 'average' samples, the RMSEP will be small, but such a model will be overoptimistic. Though in practice there are some situations when an investigator uses calibration set on a wider range, adding extreme samples, but validates the model on the test set that represents a narrower region. We consider that such a tactic may be effective in some practical cases, but it may not be used in general case for the methods of model comparison.

4.4. Investigation of representative properties of different subsets

In this section the boundary approach will be compared with the Kennard–Stone algorithm (see Subsection 2.2) and the D-optimal design (see Subsection 2.3). As these methods do not specify the subset size, it was fixed equal to $N_K = N_D = 19$ that is the size of boundary subset, $N_B = 19$. Employing the calibration C set and Model_C, the Kennard–Stone set (K set) is formed. Objects that are left in the C set comprise a 'redundant' set (RK set). Then, the procedure given by Equation (10) is applied to the K set. This establishes Model_K with four PLS components and the corresponding SIC model with the related OSP. The same procedure is also repeated with the D-optimal algorithm and 19 samples are selected from the C set in accordance with procedure described in Subsection 2.3. These samples form the D set. The objects left in the calibration set are collected in a 'redundant' set (RD set). This results in a Model_D with four PLS components, SIC model and corresponding OSP. Once again it should be emphasized that for validation of all models the same test T set is used.

Comparing RMSEP values for the various models (Table III, rows 1, 2, 4, 6) one can state that on average the prediction abilities of these models are slightly different. For Model_B and Model_D the OSClas of the test objects is absolutely similar. As to Model_K, one more absolute outsider appears there. Nevertheless the SIC leverages for Model_K are sometimes equal to, but mostly greater, than the SIC leverages for Model_B. As the SIC leverage (8) shows the size of the SIC prediction interval, the prediction with Model_K will be less precise than with Model_B for the individual test objects. As to Model_D, the SIC leverages for the individual test objects coincide with those for Model_B. The same situation may be observed for the PLS leverages shown in Figure 6. One can ascertain that the maximum PLS leverage values in training objects in Model_B [Figure 6(a)] and Model_D [Figure 6(c)] are essentially less than in Model_K [Figure 6(b)]. The leverages for Model_B and Model_D are very similar both for the test and training objects.

Therefore it may be concluded that the D and B sets have similar properties and the models obtained using these subsets have a comparable predictive ability. The model established on the base of the K set is worse in this sense. This may be easily explained by analyzing the strategy for each subset selection. Boundary approach and the D-optimal design try to select the most important, peripheral objects, whereas the Kennard–Stone algorithm selects objects uniformly.

Now let us analyze how Model_K predicts redundant samples (RK set). Studying the corresponding OSP [Figure 7(a)], one can see that there are many objects classified as outsiders. From the SIC point of view, the K set cannot be considered as a representative one, as while predicting the redundant samples, not all of them are insiders. There are totally 12 outsiders and 4 absolute outsiders among them. As to the PLS diagnostics, RMSEP for the RK set is nearly the same as for the T set, and there are objects with high leverages that may even be treated as outliers [i.e. C29 in Figure 7(b)].

Model_D better predicts its redundant objects (RD set). Due to OSClas, only seven objects within the RD set are outsiders. All of them are located very close to the model and there are no absolute outsiders at all. Again models built with the help of the B and D sets demonstrate similar prediction properties.

4.5. Different training sets

To confirm the boundary approach and to show that such a successful subset selection is not fortuitous, the following procedure was repeated 10 times:1. Divide initial data set (G set, $N_G = 139$) randomly on the training set (C set, $N_C = 99$) and the test set (T set, $N_T = 40$).2. For each such pair of the C and T sets establish the appropriate PLS model with four components and SIC model, with predefined $b_{SIC} = 1.5$ (Model_C).3. For each C set and the corresponding Model_C determine the appropriate B, K and D sets and apply to them the procedure given by Equation (10).

It is natural that the number of boundary samples and the values of RMSE vary slightly, but on the whole the results in Table IV confirm the effectiveness of the SIC approach.

The last row of Table IV demonstrates the average results (after 10 runs) for the four types of models.

So, the presented example reveals that boundary samples, which form an influential subset, are not only the significant objects for SIC models, but also constitute the relevant objects for bilinear projection models.

Table IV. Wheat data: PLS models with four components (boundary subset selection and models' evaluation for 10 calibration/test sets). Bold digits are the main results that were obtained in the course of statistical simulations

Run #	N_B	Model_C		Model_B		Model_K		Model_D	
		RMSEC	RMSEP	RMSEC	RMSEP	RMSEC	RMSEP	RMSEC	RMSEP
1	18	0.258	0.359	0.328	0.372	0.209	0.362	0.155	0.362
2	19	0.309	0.227	0.456	0.249	0.304	0.281	0.289	0.267
3	19	0.280	0.312	0.426	0.330	0.212	0.339	0.266	0.335
4	21	0.292	0.281	0.471	0.305	0.253	0.304	0.295	0.325
5	24	0.289	0.287	0.449	0.278	0.305	0.293	0.245	0.311
6	21	0.292	0.281	0.471	0.305	0.253	0.304	0.295	0.325
7	18	0.290	0.292	0.469	0.278	0.264	0.283	0.258	0.289
8	21	0.284	0.304	0.423	0.317	0.202	0.328	0.244	0.319
9	22	0.277	0.315	0.477	0.329	0.274	0.334	0.224	0.348
10	21	0.295	0.276	0.453	0.318	0.206	0.315	0.234	0.342
Mean values	0.287	**0.293**	0.442	**0.308**	0.248	**0.314**	0.251	**0.322**	

5. CASE STUDY 2: DETERMINATION OF TRACE OIL CONCENTRATIONS IN WATER

The **X** matrix consists of 1024 acoustic frequency variables (after FFT). The response vector ***y*** represents reference concentrations of oil in the samples that were specially prepared in the test laboratory. The calibration C set consists of 40 observations (objects) and the test T set also consists of 40 observations.

The original (raw data) PLS model shows a nonlinearity in the first component ***t*–*u*** plot, signifying a nonlinear relationship between the oil concentration and its influence on the effective surface tension of water. Therefore the raw y values are transformed by $y = \log(1 + y_{raw})$, which is sufficient to linearize the relationship. The final PLS modeling yields two-component model, which explains a total of 60% of **X**-variance, but 99.9% of ***y***-variance.

Eight boundary samples are detected for the given C set by OSClas, with $b_{min} = 0.154$, and $b_{SIC} = 0.29$. Afterwards, three subsets are selected: the B, K and D sets and three corresponding models are established. Models' names have the same meaning as presented in Figure 2. The predictive ability of PLS models are presented in Table V.

Model_B shows good prediction properties and it also treats 'redundant' objects from the RB set as reliable objects (column 3 in Table V). Other models also show satisfactory predictive ability on average. OSClas generated by the SIC models presents more detailed analysis shown in Table VI. Here such a characteristic is the number of objects that are regarded as the outsiders and absolute outsiders in the T set, and various 'redundant' (RB, RK and RD) sets.

Table VI shows that the status of the test samples determined by Model_C slightly differs from results for Model_B. This minor discrepancy is not essential and may be easily explained as only eight samples are used as the training, B set, for Model_B. Nevertheless this example confirms the efficiency of selecting the representative subset using the boundary approach.

6. CASE STUDY 3: PRODUCTION PROCESS MODELING

The **X** matrix consists of 25 process variables. The response vector ***y*** represents the final quality of the end product. The calibration C set comprises 102 observations (objects) and the test T set consisting of 52 objects. The data are centered and scaled as described in Reference [22]. As previously, the T set is used throughout the example for external validation for all established models. The PLS modeling yields a seven-component Model_C, which explains a total of 99.5% of **X**-variance and 99.9% of ***y***-variance.

Forty-six boundary samples are detected for the given C set by OSClas, with $b_{min} = 0.040$ and $b_{SIC} = 0.058$. Afterwards three subsets are selected for evaluation: the B, K and D sets and three corresponding models are built. Models' names have the same meaning as shown in Figure 2. The predictive abilities of the obtained models are shown in Table VII. Similar to the previous examples there is no deterioration of predictive power of Model_B in comparison with Model_C. On the other hand, the predictive performance of Model_K and Model_D is worse. It is interesting to analyze how many common/different objects are

Table V. Trace oil concentrations in water, PLS models with two components

Model_C	Model_B		Model_K		Model_D	
T set	T set	RB set	T set	RK set	T set	RD set
0.092	0.100	0.070	0.122	0.111	0.106	0.08

Values in the table are RMSEP values calculated for the sets indicated in each column.

Table VI. Trace oil concentrations in water: SIC-modeling (OSClas for different sets)

Training set	Test set ($N_T = 40$)		Redundant sets ($N_R = 32$)	
	Outsiders	Abs. outsiders	Outsiders	Abs. outsiders
C set (Model_C)	8	1	—	—
B set (Model_B)	10	0	3	0
K set (Model_K)	17	1	9	0
D set (Model_D)	10	1	5	0

Table VII. Production process modeling, PLS models with seven components

Model_C	Model_B		Model_K		Model_D	
T set	T set	RB set	T set	RK set	T set	RD set
0.018	0.018	0.010	0.020	0.018	0.027	0.028

Values in the table are RMPEP values calculated for sets indicated in each column.

included in each subset. Sketches of the three subsets are presented in Figure 8.

The percentages in intersection areas show the portion of common objects. Only 20% of calibration objects are selected by all strategies. That means that in general each method forms its own subset.

In some applications, a representative subset comprising 45% of calibration objects may be considered to be too large. So, it is important to analyze the influence of the subset size on the prediction quality of the corresponding models. According to OSClas, the minimum number of boundary objects is determined by b_{min} value. In the example, the minimal boundary set consists of eight objects ($N_B \geq 8$). Amplifying b from $b = b_{min}$ to $b = b_{SIC}$, the increasing B sets are formed. In parallel, the Kennard–Stone algorithm and the D-optimal design are used for comparison of the subsets (K and D sets) of the same size. For each selected subset a new PLS model with seven components is established, and RMSEC and RMSEP (over the same T set) are calculated. In such a case, RMSEC and RMSEP may be considered as functions of the subset size (Figure 9) evaluated for the three types of models (Model_B, Model_K and Model_D).

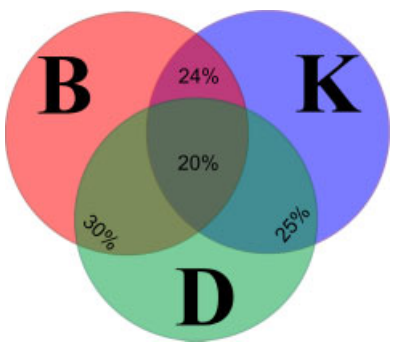

Figure 8. Subsets of 46 objects each. B–B set, K–K set, D–D set. This figure is available in color online at www.interscience.wiley.com/journal/cem

Dashed line (line 4, Figure 9) indicates RMSE values for Model_C. The dimensionality of the PLS space is not changed, but, of course, each subset forms its own PLS vector system. That is why RMSEC and RMSEP are not monotonous functions of the subset size. Explicit instability at the initial interval from 8 till 15 samples conforms that corresponding models are unstable. It is obvious that, in general, the subset of 8 or 10 objects is too small for establishing a model with 7 PLS components. As expected, the larger the subset, the closer the RMSE values to that calculated for Model_C. Figure 9(a) (curve 1) shows that RMSEC_B for each B set is always greater than those characteristics for other kinds of subsets. This confirms the role of boundary objects as the most influential ones that have rather high residual and high leverage. On the other hand RMSEP_B in most cases is the best (the lowest) among other similar values and it converges with RMSEP_C value obtained with overall Model_C. See Figure 9(b) (curves 1 and 4).

Analyzing curve 1 in Figure 9(b) we can state that if the goal is to choose a subset of objects that could be reliably used for the model construction without any compromise in predictive ability, not less than 42 objects should be used. This example shows not only the importance of boundary objects, but also confirms that the number of boundary objects, detected by OSClas is very close to optimal.

7. DISCUSSION AND CONCLUSIONS

The new method for representative (influential) subset selection has been presented. It is based on a combination of the SIC approach with chemometric bilinear projection methods (PCR, PLS). One of the main issues of the SIC approach is the OSClas, which has been shown to be a powerful and visually simple

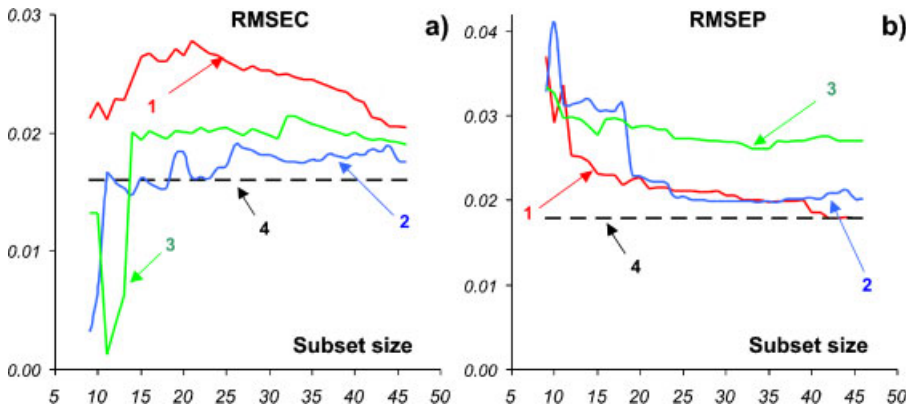

Figure 9. Production process modeling, PLS models with seven components. RMSE in dependence of the size of the representative subset, 1: Model_B, 2: Model_K, 3: Model_D, 4: Model_C. This figure is available in color online at www.interscience.wiley.com/journal/cem

instrument for the detailed analysis of the status of individual objects and therefore useful for the representative subset selection. The three different real-world examples presented here originate from various chemical problems; they have very diverse internal data structures and PLS models of different complexity are established for them. This allowed us to demonstrate the advocated method in a variety of settings. The examples show that the boundary samples, which form the influential subset, are not only significant objects for SIC models, but they constitute the relevant objects for bilinear projection models as well.

It was shown that the strategy of selecting the representative subset using the boundary samples is not user dependent, that is not subjective. It is also important that the SIC approach uses no new or extra parameters, which cannot be evaluated using the data set and have to be set *a priori*.

It is worth mentioning that the term 'redundant sample' in application to insiders in a training set should not be interpreted directly. Of course such kind of samples is useful for the initial data overview. The more samples are in the training set, the more accurate is the determination of model complexity, that is the number of principal components in PCR, or PLS modeling, and the more accurate is the b_{SIC} estimate.

Analyzing all three techniques for the subset selection described in the paper we can state the following. The Kennard–Stone procedure is effective in the case when it is necessary to divide the initial data set into two equivalent sets, for example into the training and test sets. The Kennard–Stone method selects samples uniformly and, therefore, it works less effectively than the other two procedures for the purpose of selecting the most important objects. The D-optimal design and boundary approach demonstrated perfect performance for the purposes of the influential subset selection. Nevertheless we believe that the SIC method has several advantages. It: (1) determines the unambiguous number of influential objects for the data and model under consideration; and (2) it takes into account not only X values, but also *y* values.

At the same time, the D-optimal design is often used for the selection of the training set. Here the researchers should realize that after moving all the influential objects to the training set, the remaining test set (samples left in the initial data set) represents a narrower region and such a test set validation may be misleading.

8. SOFTWARE

All presented SIC calculations were made with software, programmed and implemented as an add-in for Excel, which includes: various NIPALS algorithms [11] for bilinear matrix decompositions, a standard Simplex algorithm [13,23] for optimizations, as well as a necessary suite of special procedures, for example for preprocessing, transformations, etc. This software is developed for internal use, but all algorithms are of well-known types and may be easily implemented using various standard packages. The Kennard–Stone and D-optimal procedures are programmed using VBA and Excel interface.

REFERENCES

1. Bouveresse E, Massart DL. Standardisation of near-infrared spectrometric instruments: a review. *Vib. Spectrosc.* 1996; **11**: 3–15.
2. Jouan-Rimbaud D, Massart DL, Saby CA, Puel C. Characterization of the representativity of selected sets in multivariate calibration and pattern recognition. *Anal. Chim. Acta* 1997; **350**: 149–161.
3. Höskuldsson A. Variable and subset selection in PLS. *Chemom. Intell. Lab. Syst.* 2001; **55**: 23–38.
4. Wu W, Walczak B, Massart DL, Heuerding S, Erni F, Last IR, Prebble KA. Artificial neural networks in classification of NIR spectra data: Design of training set. *Chemom. Intell. Lab. Syst.* 1996; **33**: 35–46.
5. Dantas Filho HA, Harrop Galvao RK, Dantas Filho HA, Harrop Galvao RK, Ugulino Araujo MC, da Silva C, Bezerra Saldaha TC, Jose GE, Pasquini C, Raimundo IM Jr., Rodrigues Rohwedder JJ. A strategy for selecting calibration samples for multivariate modelling. *Chemom. Intell. Lab. Syst.* 2004; **72**: 83–91.
6. Rodionova OYe, Esbensen KH, Pomerantsev AL. Application of SIC (Simple Interval Calculation) for object status classification and outlier detection —comparison with PLS/PCR. *J. Chemometrics* 2004; **18**: 402–413.
7. Rodionova OYe, Pomerantsev AL. Principles of simple interval calculations. In Progress in Chemometrics Research, Pomerantsev AL (ed.). Nova Science Publishers: NY, 2005; 43–64.
8. Kennard RW, Stone LA. Computer aided design of experiment. *Technometrics* 1969; **11**: 137–148.
9. Fedorov VV. Theory of Optimal Experiments. Academic press: New York, 1972. (Moscow University, English translation by Studden WJ, Klimo EM).
10. Faber K. Comparison of two recently proposed expressions for partial least squares regression prediction error. *Chemom. Intell. Lab. Syst.* 2000; **52**: 123–134.
11. Martens H, Naes T. Multivariate Calibration. Wiley: New York, 1998.

12. Næs T, Isaksson T, Fearn T, Davies T. Multivariate Calibration and Classification. NIR Publications: Chichester, UK, 2002.
13. Dantzig G. Linear Programming and Extensions. Princeton University Press: Princeton, NJ, 1963.
14. Gumbel E. Statistics of Extremes. Columbia University Press: NY, 1962.
15. Rajer-Kanduc K, Zupan J, Majcen N. Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. *Chemom. Intell. Lab. Syst*, 2003; **65**: 221–229.
16. Yoon J, Lee B, Han C. Calibration transfer of near-infrared spectra based on compression of wavelet coefficients. *Chemom. Intell. Lab. Syst.* 2002; **64**: 1–14.
17. Park K-S, Ko Y-H, Lee H, Jun C-H, Chung H, Ku M-S. Near-infrared spectral data transfer using independent standardization samples: a case study on the trans-alkylation process. *Chemom. Intell. Lab. Syst.* 2001; **55**: 53–65.
18. Andersson PM, Sjostro M, Wold S, Lundstedt T. Strategies for subset selection of parts of an in-house chemical library. *J. Chemometrics* 2001; **15**: 353–369.
19. Cruciani G, Baroni M, Carosati E, Clementi M, Valigi R, Clementi S. Peptide studies by means of principal properties of amino acids derived from MIF descriptors. *J. Chemometrics* 2004; **18**: 146–155.
20. Wold S, Josefson M, Gottfries J, Linusson A. The utility of multivariate design in PLS modelling. *J. Chemometrics* 2004; **18**: 156–165.
21. Esbensen KH. Multivariate Data Analysis—In Practice (4th edn). CAMO: ASA: Oslo, Norway, 2000.
22. Pomerantsev AL, Rodionova OYe, Höskuldsson A. Process control and optimization with simple interval calculation method. *Chemom. Intell. Lab. Syst.* 2006; **81**: 165–179.
23. Taha H. Operations Research. An Introduction (3rd edn, vol. 1). MacMillan: New York, 1982.