analytical chemistry

Article

pubs.acs.org/ac

Simple and Effective Way for Data Preprocessing Selection Based on Design of Experiments

Jan Gerretzen,[†,‡] Ewa Szymańska,[†,‡] Jeroen J. Jansen,[†] Jacob Bart,[§] Henk-Jan van Manen,[§] Edwin R. van den Heuvel,[∥] and Lutgarde M. C. Buydens*[,†]

[†]Radboud University, Institute for Molecules and Materials, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands

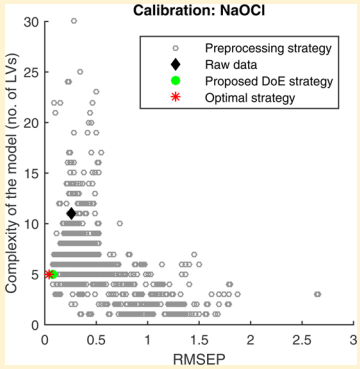
[‡]TI-COAST, Science Park 904, 1098 XH Amsterdam, The Netherlands

[§]AkzoNobel, Supply Chain, Research & Development, Zutphenseweg 10, 7418 AJ Deventer, The Netherlands

[∥]Eindhoven University of Technology, Den Dolech 2, 5600 MB Eindhoven, The Netherlands

Supporting Information

ABSTRACT: The selection of optimal preprocessing is among the main bottlenecks in chemometric data analysis. Preprocessing currently is a burden, since a multitude of different preprocessing methods is available for, e.g., baseline correction, smoothing, and alignment, but it is not clear beforehand which method(s) should be used for which data set. The process of preprocessing selection is often limited to trial-and-error and is therefore considered somewhat subjective. In this paper, we present a novel, simple, and effective approach for preprocessing selection. The defining feature of this approach is a design of experiments. On the basis of the design, model performance of a few well-chosen preprocessing methods, and combinations thereof (called *strategies*) is evaluated. Interpretation of the main effects and interactions subsequently enables the selection of an optimal preprocessing strategy. The presented approach is applied to eight different spectroscopic data sets, covering both calibration and classification challenges. We show that the approach is able to select a preprocessing strategy which improves model performance by at least 50% compared to the raw data; in most cases, it leads to a strategy very close to the true optimum. Our approach makes preprocessing selection fast, insightful, and objective.



Data preprocessing involves the conversion of the original, raw data to *cleaned* data, in which unwanted variation has been removed. Sources of such unwanted variation include, e.g., baseline drifts in spectroscopic measurements or misalignment in chromatographic elution profiles. These sources are unrelated to the goal for which data was collected, such as predicting the concentration of a compound from spectroscopic data.

A multitude of different preprocessing methods has been developed to remove a variety of artifacts in data from different analytical platforms.[1−6] In a complete data analysis procedure, often more than one preprocessing method is applied. The consecutive application of different preprocessing methods is called a preprocessing strategy.[7] Each preprocessing strategy consists of a number of different steps (e.g., baseline correction, smoothing), and in each step, a specific preprocessing method is applied.

Preprocessing can make or break data analysis.[7] This implies that a wrong choice of preprocessing is detrimental to the predictive power of a chemometric model (see, e.g., Figure 4). Moreover, existing preprocessing selection procedures are limited and not suited for large data sets.[7] Nowadays, many researchers consider preprocessing a burden. Therefore, it is of the utmost importance that an approach is developed to obtain an optimal preprocessing strategy within reasonable time.

The existing selection procedures include visual inspection of data after preprocessing and the evaluation of the preprocessed data with quality parameters such as correlation. The most common selection procedure is a fit-for-use approach: simply trying a few preprocessing methods or strategies and selecting the one with the best model performance. It is, however, very unlikely that the few preprocessing strategies selected *a priori* lead to an optimal result. A straightforward way to overcome this is the evaluation of many or all possible strategies and simply selecting the optimal one. However, this is a very time-consuming procedure and therefore not feasible for large data sets and the multitude of available preprocessing methods.

Without a preprocessing selection procedure, the preprocessing strategy is often based on experience: knowledge about the instrument, sample, and different preprocessing methods. Although such a preprocessing strategy will probably have an acceptable model performance, it may not be the optimal one.

This paper presents a novel, fast, simple, and effective approach for preprocessing selection. The key idea of the approach is design of experiments (DoE), to evaluate a well-selected number of different preprocessing strategies. Inter-

Received: July 27, 2015
Accepted: November 19, 2015
Published: November 19, 2015

DOI: 10.1021/acs.analchem.5b02832
Anal. Chem. 2015, 87, 12096−12103

pretation of main effects and interactions shows which preprocessing steps are relevant to the data and which are not. For each relevant step, the optimal preprocessing method is subsequently found using a simple search algorithm. In this work, the selection criterion is solely based on optimal model performance. At a later stage, model interpretability will be included as well (see the Discussion section).

We focus on preprocessing for calibration and classification challenges in spectroscopic data. Four preprocessing steps are considered: baseline correction, scatter correction, smoothing, and scaling. These are the commonly used and important preprocessing steps.[7] Our selection of preprocessing steps and methods is not exhaustive and may be altered, e.g., based on sample or instrument knowledge or experience with certain preprocessing methods, which will be discussed later as well.

DATA

Eight data sets with different goals have been analyzed with the proposed approach. Most details and results can be found in the Supporting Information. Here, we discuss the results for two data sets: a NIR data set on compound calibration and a NIR data set on classifying Rochefort beers.[7,8] In those references, however, a mid infrared data set was used for exactly the same purpose instead of NIR. Therefore, the section on the classification data only highlights relevant experimental details for the NIR data set.

Calibration Data: Background. The NIR calibration data set relates to waste treatment of a chlorine gas (Cl_2) production facility. The gaseous waste effluent of this facility contains chlorine, which needs to be removed for environmental reasons. For this purpose, a so-called caustic scrubber is used.

In this scrubber, waste gases are led through a solution containing NaOH. NaOH reacts with chlorine in the waste gases to produce NaOCl and NaCl:

$$Cl_2 + 2NaOH \rightarrow NaOCl + NaCl + H_2O$$

NaOH, however, also reacts with CO_2 in air, leading to Na_2CO_3:

$$2NaOH + CO_2 \rightarrow Na_2CO_3 + H_2O$$

These reactions show that the concentrations of NaOH, NaOCl, and Na_2CO_3 play an important role in the scrubber. Therefore, continuous monitoring of the concentrations of these compounds is required. NIR spectroscopy is already proven useful for online monitoring of caustic scrubbers.[9−11]

Calibration Data: Experimental Setup. Different samples (n = 13) containing various amounts of NaOH, NaOCl, and Na_2CO_3 have been prepared, in which the concentrations of all compounds represented values likely to occur in the scrubber. A NIR spectrum of each sample was obtained at five different temperatures (15 °C, 21.3 °C, 27.5 °C, 33.8 °C, and 40 °C, the range in which the scrubber operates), leading to 65 NIR spectra in total. The compositions of the different samples are given in Table S-1. In the remainder of this paper, these samples will be referred to as compositions (i.e., a composition indicates a sample with a specific amount of NaOH, NaOCl, and Na_2CO_3). Six NIR spectra have additionally been measured and will be used as a validation set. These spectra were recorded by measuring three different compositions (independent of the training set) at two temperatures per composition (see Table S-2).

The NIR spectra were recorded on a Multi Purpose Analyzer (Bruker) with a transflection probe with adjustable path length (Hellma Mini Immersion Probe Saphir). Each spectrum was recorded as the average of 32 scans with a resolution of 16 cm^{-1}. In total, each spectrum contains 1102 data points. The raw spectra are displayed in the top panel of Figure 1.

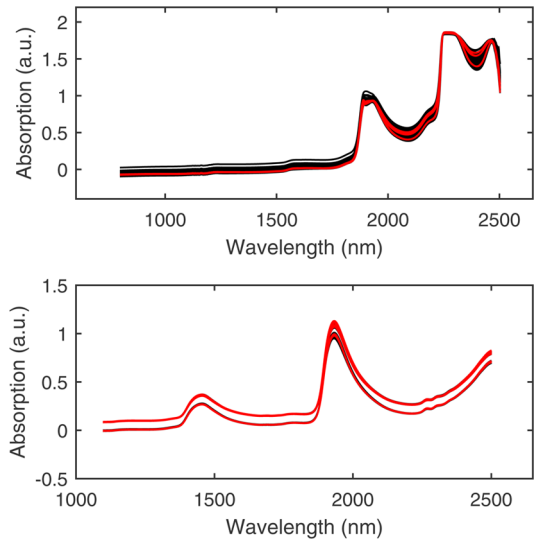


Figure 1. Top panel: plot of the raw calibration data. Spectra from the training set (n = 65) are shown in black; validation spectra (n = 6) are shown in red. Bottom panel: raw classification data. Black represents Rochefort 8°, red Rochefort 6° and 10°.

Classification Data: Experimental Details. To discriminate Rochefort 8° from Rochefort 6° and 10° beers, NIR spectra of beers from both classes were recorded using a scanning spectrophotometer (NIRSystems 6500). Spectra were recorded in duplicate between 400 and 2498 nm, from which the average spectrum in the 1100−2498 nm wavelength range was used for data analysis; each spectrum has 700 data points.

Separate training and validation sets were measured. The training set consisted of 44 spectra, of which 28 are Rochefort 8° beers (Figure 1). The validation set consisted of 30 spectra, of which 20 are Rochefort 8° beers. As already described,[7] the spectra have been measured in two different batches, leading to the offset visible in Figure 1. This offset is unrelated to the beer class.

METHODS

General Approach. The key idea of our approach is that only design of experiments (DoE) is used to find an optimal preprocessing strategy (see Figure 2). In the DoE, different preprocessing steps are evaluated as factors; these are baseline correction (B), scatter correction (St), noise removal by smoothing (Sm) and scaling (Sg), always applied in this order, and it is assessed whether or not each factor influences model performance (The order of the steps is discussed further in the Discussion section.). For this, we have chosen the simplest DoE: the full-factorial design. Second, for each preprocessing step deemed relevant by the DoE, the optimal preprocessing method for that step is obtained from a broader set of methods. All preprocessing methods used in this work are based on extensive studies.[7] Table S-3 presents an overview of all methods used.

PLS-1 and PLS-DA were chosen for the analysis of the calibration and classification data sets, respectively. PLS and PLS-DA have already proven suitable methods for similar cases

DOI: 10.1021/acs.analchem.5b02832
Anal. Chem. 2015, 87, 12096−12103

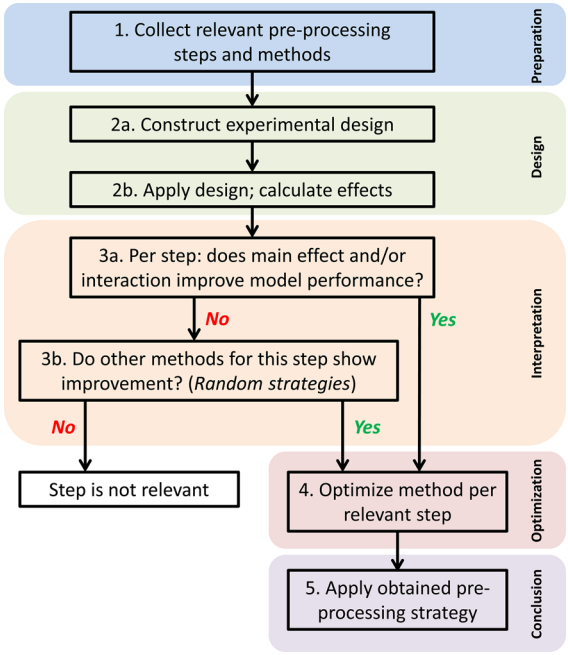


Figure 2. Flowchart of the DoE approach.

many times.[12−15] The presented approach can be used with other data analysis methods as well. All programming was performed using MATLAB (version 8.4.0 (R2014b), The MathWorks Inc., Natick, MA).

PLS and Cross-Validation. PLS-1 is a commonly applied regression method[16,17] that aims to predict a response vector $\boldsymbol{y}$ given a data matrix $\boldsymbol{X}$, by estimating the regression coefficients $\boldsymbol{b}$:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{E} \quad (1)$$

$\boldsymbol{E}$ is the residual matrix, which should be minimized to obtain an optimal model. For PLS-DA, $\boldsymbol{y}$ is binary, indicating class membership.

When an optimal $\boldsymbol{b}$ is found, the performance of the model should be tested by applying the model to new and unseen data (i.e., using $\boldsymbol{b}$ and a new $\boldsymbol{X}$ to generate predictions $\hat{\mathbf{y}}_{new}$). The root-mean-square error of prediction (RMSEP) is a common measure of PLS model performance: the lower the RMSEP, the better predictions the model generates for new data. For classification, the percentage of misclassification is often used as a model performance measure.

Cross-validation was used to optimize the number of latent variables (LVs) for each PLS model.[18,19] For the calibration data, cross-validation was performed by leaving one composition (i.e., five spectra) out at a time, leading to a 13-fold cross-validation procedure. The optimal number of LVs, between 1 and 35, in the cross-validation procedure was obtained by using a selection algorithm.[20] In short, it compares the model performance of a PLS model with a + 1 LVs with that of a model with a LVs by using a randomization t-test; the model with a LVs is preferred if no significant difference in model performance exists (more details can be found in the Supporting Information). The more conventional way of selecting the number of LVs with the lowest root mean square error of cross-validation (RMSECV) often indicated an extremely high number of LVs (at or around 35) and was therefore not used. Using the breakpoint (elbow) in the curve of RMSECV vs number of LVs was also not used, because it is not straightforward to automate.

For the classification data, PLS-DA with a leave-10%-out cross-validation procedure was performed to determine the optimal number of LVs (again between 1 and 35). Here, the number of LVs leading to the lowest percentage of misclassifications was chosen. Classes were coded as −1 and +1, and we accounted for unequal class size.[21]

The DoE Approach. DoE, also referred to as Experimental Design (XPD), provides a way to reveal which factors influence the response of an experiment.[22] Its goal is to evaluate the influence of each factor on the response. Here, we propose a *full factorial* design (i.e., the simplest design) to evaluate the influence of each preprocessing step on model performance. Each factor is varied at two levels: the low level ("−") and the high level ("+"). This leads to a design matrix with four factors (i.e., the four preprocessing steps) and 2^4 = 16 experiments (Table 1).

Table 1. Design Matrix As Used in Our Approach

experiment	baseline	scatter	smoothing	scaling
1	+	+	+	+
2	+	+	+	−
3	+	+	−	+
4	+	+	−	−
5	+	−	+	+
6	+	−	+	−
7	+	−	−	+
8	+	−	−	−
9	−	+	+	+
10	−	+	+	−
11	−	+	−	+
12	−	+	−	−
13	−	−	+	+
14	−	−	+	−
15	−	−	−	+
16	−	−	−	−

The aim is to assess for each preprocessing step whether it influences model performance or not. In practice, that implies that an evaluation should be made as to whether the model performance significantly increases when performing a specific step compared to not performing it. This already determines the value for the low level for all steps: the low level indicates that the step should not be performed (For scaling, the low level indicates "mean centering", since this is customary for virtually all PLS models.).

For the high level, one of the many available methods should be chosen. Such a method should lead to improved model performance if the respective data artifact indeed requires correction. On the basis of our experience with spectroscopic data, both from literature and our own experience,[7] we have preselected a method for each factor. For baseline, the selected method is AsLS (Asymmetric Least Squares[23,24]) and for scatter SNV (Standard Normal Variate[25]). For smoothing, the Savitzky−Golay algorithm was chosen, with parameters for window width and order exactly in the middle of the evaluated options (i.e., window width 9 points and polynomial order 3[7,26]). Finally, Pareto scaling was chosen for the scaling step, because this is one of the more commonly applied scaling methods in infrared data.[27]

DOI: 10.1021/acs.analchem.5b02832
Anal. Chem. 2015, 87, 12096−12103

Table 2. Design Matrix Including Response Variables[a]

experimental design					response	
experiment	baseline	scatter	smoothing[b]	scaling	$RMSEP_{NaOCl}$	%-$misclass_{Rochefort}$
1	AsLS	SNV	yes	Pareto	0.466	7.42
2	AsLS	SNV	yes	MC	0.568	7.42
3	AsLS	SNV	none	Pareto	0.577	7.42
4	AsLS	SNV	none	MC	0.563	7.42
5	AsLS	none	yes	Pareto	0.083	19.82
6	AsLS	none	yes	MC	0.194	11.36
7	AsLS	none	none	Pareto	0.081	19.82
8	AsLS	none	none	MC	0.199	11.36
9	none	SNV	yes	Pareto	0.826	7.42
10	none	SNV	yes	MC	0.568	11.36
11	none	SNV	none	Pareto	0.825	7.42
12	none	SNV	none	MC	0.570	11.36
13	none	none	yes	Pareto	0.289	19.82
14	none	none	yes	MC	0.270	19.82
15	none	none	none	Pareto	0.288	19.82
16	none	none	none	MC	0.263	19.82

[a]Abbreviations: asymmetric least squares (AsLS); standard normal variate (SNV); mean centering (MC). The two rightmost columns show the RMSEP (root mean square error of prediction) values of prediction of NaOCl and the percentage of misclassification for the Rochefort data respectively. The number of LVs for each experiment and response is optimized using cross-validation (no. of LVs not shown). [b]"Yes" indicates Savitzky–Golay smoothing, with window size of 9 px and a 3rd order polynomial.

The response variable for the calibration data is RMSEP and for the classification data the percentage of misclassification in the validation data, such that for both data sets a lower response indicates improved model performance. The effect (i.e., influence) of each preprocessing step can be calculated by taking the average of the response variable of this particular step at the high level ($\bar{y}_+$) minus the average response variable at the low level ($\bar{y}_-$):

$$\text{effect} = \bar{y}_+ - \bar{y}_- \tag{2}$$

A negative effect indicates that $\bar{y}_+$ is smaller than $\bar{y}_-$, i.e., performing a specific step has led to a decrease in the response variable and thus improved model performance. The complete design matrix is shown in the first five columns of Table 2.

Significance of Effects. In general, the significance of an effect is determined by a pooled variance, obtained from response values of repeating all rows in the DoE. For this purpose, we bootstrap the original data to artificially create new data on which the same DoE (Table 2) is applied. To keep comparability with the results from the original, non-bootstrapped data set, the number of LVs is not optimized but set to the values as found for the nonbootstrapped data. The number of bootstrap samples to use is that number where the variability (s_{effect}) in response values (i.e., RMSEP or percentage of misclassifications) has become constant.

From the response values obtained during bootstrapping, the variance in response is calculated for each row in the design and these variances are subsequently *pooled* over the 16 DoE strategies to obtain one, averaged variance:

$$s^2_{pooled} = \frac{\nu_1 s_1^2 + \nu_2 s_2^2 + \ldots + \nu_{16} s_{16}^2}{\nu_1 + \nu_2 + \ldots + \nu_{16}} \tag{3}$$

Here, s_n^2 is the variance in responses of row n in the DoE; ν_n is the amount of bootstrap samples for row n (denoted r_n) − 1, $\nu_n = r_n - 1$. Finally, the pooled variance is converted into a standard deviation for each effect s_{effect} using the following equation:

$$s_{effect} = \sqrt{s^2_{pooled} \times \left(\frac{1}{N_+} + \frac{1}{N_-}\right)} \tag{4}$$

In this equation, N_+ and N_- are the number of values used in the calculation of $\bar{y}_+$ and $\bar{y}_-$, respectively. An effect is often deemed significant if its value is at least 2−3 times larger than s_{effect}.

This approach implicitly assumes equal variances and normally distributed responses over the response values for all 16 experiments. Since this does not completely hold (lower response values are also associated with a lower variance), the logarithm of the response values is used instead for calibration; for classification, the arcsin of the percentages of misclassification is used.

Interaction Effects and Their Interpretation. Using DoE, interaction effects can be calculated as well, due to the full-factorial way of setting up the experiment. In our case, the design consists of four different factors, from which six two-factor interactions can be constructed (B × St, B × Sm, B × Sg, St × Sm, St × Sg, and Sm × Sg) as well as four three-factor interactions (B × St × Sm, B × St × Sg, B × Sm × Sg, and St × Sm × Sg) and one four-factor interaction (B × St × Sm × Sg).

The interpretation of a two-factor interaction effect in our approach is as follows:[28] if the effect of the interaction has a *negative* value, a decrease in response (i.e., improved model performance) is expected when the two factors involved are *simultaneously* changed from the low to the high level. Vice versa, an increase in response is expected when the effect of the interaction has a positive value. The net effect Eff_{net} on model performance when including two preprocessing steps A and B can therefore be calculated via

$$Eff_{net} = Eff_A + Eff_B + Eff_{A\times B} \tag{5}$$

This only holds in the case of two factors and their interaction. The interpretation of three- and four-factor interactions is less straightforward,[28] and they are therefore further neglected in this work (see the Discussion section).

DOI: 10.1021/acs.analchem.5b02832
Anal. Chem. 2015, 87, 12096−12103

Optimization of Relevant Steps. After having determined which preprocessing steps are relevant using DoE, the optimal method for each step should be chosen, ultimately leading to an optimal preprocessing strategy. For this purpose, we use a sequential optimization approach, in which the optimal preprocessing method for each step is sought in a sequential way. First, the method for the first relevant step is optimized; all subsequent steps are not performed. Next, the method for the second relevant step is optimized, where the first step is performed with the already optimized method. This continues until all relevant steps have been optimized.

RESULTS

Here, we will focus on the results of predicting the concentration of NaOCl and classifying the beer samples. The results for the other two compounds in the calibration data, NaOH and Na_2CO_3, can be found in the Supporting Information as well as results from all other investigated data sets. An additional discussion follows at the end of this section.

Prediction of NaOCl. Table 2 shows the RMSEP values for the 16 different preprocessing strategies as evaluated by the DoE. Effects, both main effects and interactions, are calculated for each of the four preprocessing steps (Figure 3). The height

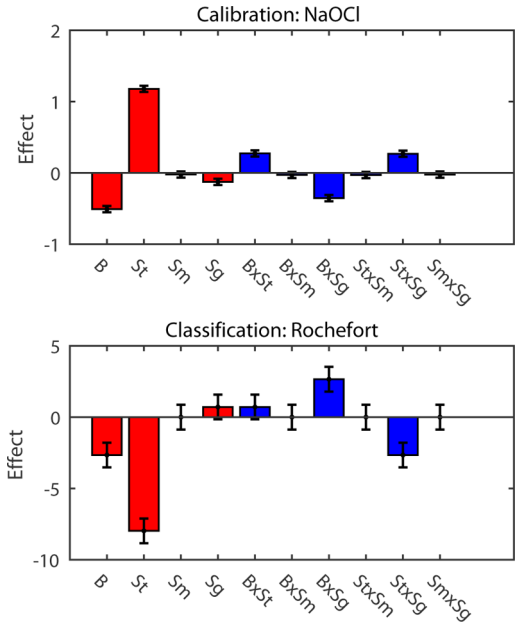


Figure 3. Effect values for all main effects (red) and two-factor interactions (blue), based on predicting the concentration NaOCl (upper panel) and classifying Rochefort beers (bottom panel). Errorbars indicate $\pm 2s_{effect}$. Abbreviations: Baseline (B); Scatter (St); Smoothing (Sm); Scaling (Sg). Interaction effects are shown with a "x", e.g., "B × St" indicates the two-factor interaction between Baseline and Scatter. Effect values are based on the logarithm (calibration) or arcsin (classification) of the response values from Table 2.

of a bar indicates the effect value; the errorbars indicate the size of $\pm 2s_{effect}$. Bootstrap samples are constructed to estimate the value of s_{effect}. A total of 150 bootstrap samples were chosen, because the value of s_{effect} has stabilized at that number of bootstrap samples (Figure S-1).

From Figure 3, it follows that scatter correction (St) has a very large positive effect, i.e., it leads to an increase in RMSEP. Furthermore, all effects that include smoothing are insignificant, an effect value of 0 is within the interval given by the effect value $\pm 2s_{effect}$, and thus smoothing does not influence RMSEP at all. Figure 1 shows that the data under study are indeed not very noisy.

The only two effects that obviously lead to a decrease in RMSEP are baseline correction (B) and the baseline correction-scaling interaction (B × Sg). Since Sg only has a very small positive effect, if significant at all, there will be a net decrease in RMSEP if doing both B and Sg compared to only B. Therefore, we conclude that baseline correction and scaling are the relevant preprocessing steps.

It may be that St and Sm are excluded due to an unfortunate choice for the high level of these steps (i.e., St or Sm in fact do lead to a reduction in RMSEP, but the selected method at the high level does show this; see step 3b in Figure 2). To protect our approach against such failure of the selected methods, model performance is evaluated of six additional preprocessing strategies. In these additional strategies, only the methods selected for the irrelevant preprocessing steps are varied, while the methods for the relevant steps are fixed to their high level setting. If these strategies cannot improve model performance any further, it is very likely that a particular preprocessing step is rejected because it indeed has no effect on model performance and not because of failure of the selected high level method. This approach is further referred to as *random strategies*.

In this case, RMSEP values of six additional strategies are calculated that include different methods for St and Sm, while fixing the method for B and Sg to AsLS and Pareto, respectively. These values are subsequently compared to row 7 in Table 2: in that way, it can be assessed whether different methods for Sm and St can further lower the RMSEP of application of only B and Sg (0.081). For NaOCl, it appeared that all six additional strategies had a much higher RMSEP compared to 0.081 (range, 0.327−1.238), so this supports the conclusion that only B and Sg need further optimization.

The standard deviation in RMSEP values from the 150 bootstrap samples for experiment 7 is among the lowest of all 16 experiments (Figure S-2). Performing B and Sg thus not only leads to an increase in model performance, it also leads to a more robust model (compared to strategy 16, the raw data): small changes in data lead to only small changes in model performance.

Next, the optimal method for each significant step has to be determined by a simple sequential optimization scheme. Sequential optimization of B and Sg leads to the evaluation of 14 different preprocessing strategies: 7 to select the best baseline method and subsequently 7 to select the best scaling method. The preprocessing strategy obtained in this way is fourth order detrending plus Pareto scaling, with an RMSEP value of 0.0757.

A preprocessing strategy has been obtained that improves much in terms of RMSEP over the raw (i.e., meancentered) data (0.263, row 16 in Table 2) and also in robustness (Figure S-2). We have also compared the obtained strategy with the true optimal strategy, the strategy from among all 4 900 possible strategies leading to the lowest RMSEP.[7] Therefore, we applied all 4 900 different preprocessing strategies, calculated 4 900 different PLS models with cross-validation, and assessed the RMSEP for each PLS model (Figure 4).

The strategy obtained using the DoE approach is very close to the true optimum. Moreover, the model complexity has lowered compared to a model built on the raw data. The DoE

DOI: 10.1021/acs.analchem.5b02832
Anal. Chem. 2015, 87, 12096−12103

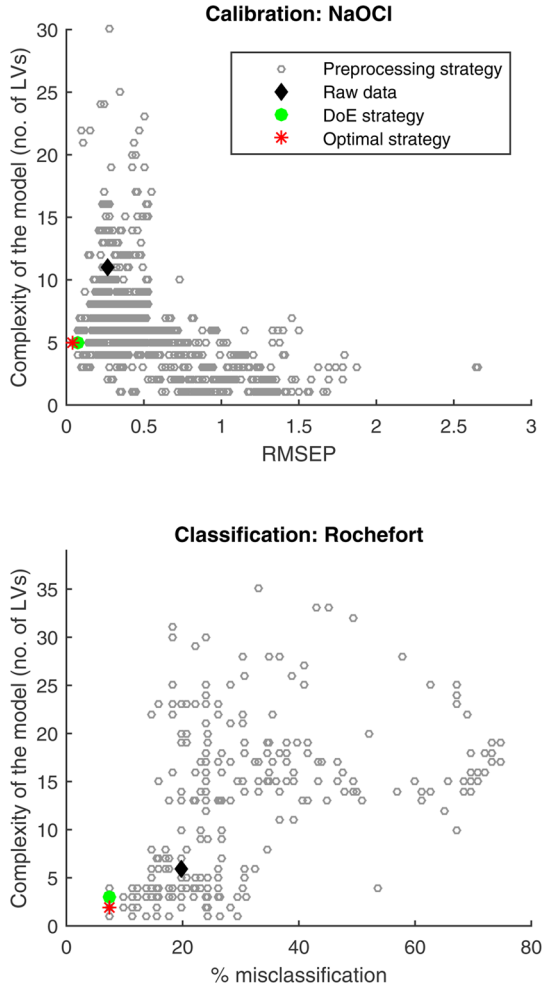


Figure 4. Model performance of all 4 900 different preprocessing strategies versus the complexity of the model (i.e., the number of LVs). Upper panel, calibration data (NaOCl); lower panel, Rochefort classification. The black diamond indicates the result of the raw (i.e., mean centered) data. The green dot shows the outcome of the final strategy found using our DoE approach and the red star the strategy with the lowest model performance that would have been found by examining all 4 900 strategies.

approach is thus able to provide an optimal preprocessing strategy, with evaluation of only a small fraction of all 4 900 possible preprocessing strategies.

Classifying Rochefort Beers. Percentage of misclassification values for the 16 strategies are shown in Table 2 for the classification data. The bottom panel of Figure 3 shows the main effects and second-order interactions for these data where the error bars are again based on 150 bootstrap samples.

Two main effects (B and St) have a negative effect. The B × St interaction does not seem significant, since 0 is within the confidence limits. Furthermore, all effects with smoothing are 0, again indicating that smoothing does not have an effect on model performance. Also for this data, there does not seem to be any noise (Figure 1). B and St are therefore deemed relevant steps.

Scaling (Sg) has one relevant interaction: St × Sg. However, since baseline and scatter correction are already included, there will be no additional increase in model performance when including scaling: main effect Sg is not significant, while B × Sg and St × Sg cancel out. Therefore, scaling is not considered as a relevant step here.

The conclusion to only include B and St is validated by additional strategies, where the settings for B and St are fixed. It appeared that all six additional strategies have exactly the same percentage of misclassification as for only B and St (experiment 4 in Table 2; percentage of misclassification 7.42%) and therefore the conclusion to only include B and St and to disregard Sg is justified.

The strategy obtained with the DoE is baseline correction via AsLS and scatter correction via either max scaling, L2 norm scaling or SNV (they all lead to the same response). The location of these strategies in the full *preprocessing landscape* is shown in Figure 4. Apparently, the most optimal preprocessing strategy in terms of model performance has been found. Only the model complexity could have been 1 or 2 LVs lower with another preprocessing strategy. Again, we can conclude that an optimal preprocessing strategy is obtained using the DoE approach.

DISCUSSION

The presented approach performs well for all eight investigated data sets (Table S-4). The approach does not only lead to a more efficient preprocessing selection, it also provides more insight in preprocessing, especially in cases where preprocessing steps are not deemed relevant. The general framework presented in Figure 2 is valid for other data types (step 1), different types of designs (step 2) and other models and performance measures (step 3) as well.

From all steps included in the preprocessing strategy, scaling is the only step that does not (directly) relate to the removal of a data artifact. Baseline correction methods, scatter correction methods, and smoothing all aim to remove a certain artifact from the data (baseline, scatter, and noise, respectively) which may hamper the construction of a chemometric model. Scaling, on the other hand, is mainly part of a preprocessing strategy to make the data more suitable for data analysis, such as when autoscaling as to remove the influence of variables measured on different scales.

Scaling should always be the last step in a preprocessing strategy. Indeed, when performing scaling, e.g., prior to baseline correction, the scaling effect will be partly attenuated by the baseline correction method. The other three steps, however, may be changed in order. For this work, we have set the order of steps in the strategy to a reasonable order. Some scatter correction methods are also able to remove a baseline,[29] but baseline correction methods are generally better suited for this task. Therefore, baseline correction is performed prior to scatter correction.

The order of steps should be fixed before evaluating the DoE, because effects and interactions will change if steps are performed in a different order. The design could be extended with another factor, i.e., the order of steps, which is especially useful when a reasonable order is not known on beforehand. This factor may have multiple levels such that multiple orders can be examined. The number of experiments to perform (and also the number of chemometric models to build) is, however, increasing in such an approach.

The preprocessing steps and methods used in this work are the most common ones for preprocessing spectroscopic data. For all investigated data sets, our selection performed well. Therefore, our selection may be used if a user is not experienced in preprocessing and wants to obtain a reasonable

DOI: 10.1021/acs.analchem.5b02832
Anal. Chem. 2015, 87, 12096−12103

preprocessing strategy. Of course, more experienced users can alter the approach based on their experience and knowledge, e.g., by taking different preprocessing steps into account, selecting different preprocessing methods for the DoE, or changing the order in which the steps are applied. The influence of this order can even be examined using the design itself, by adding a fifth factor representing the order of applying the different methods, possibly with multiple levels. Almost all preprocessing methods used in this work do not require any parameter optimization, which simplifies the use of this approach. Therefore, other popular preprocessing methods, such as Extended Multiplicative Signal Correction (EMSC)[30,31] have not been considered, since these methods often require a more thorough parameter optimization or input based on prior knowledge.

In this work, we have solely focused on model performance, without taking interpretability of the model into account. In future work, interpretability will be included by evaluating, e.g., the Variable Importance in Projection (VIP)[32] or the newly proposed Significance Multivariate Correlation (sMC).[33] When both predictive ability and interpretability need to be assessed, multicriterion experimental designs may be used, which allow for the simultaneous evaluation of more than one response variable, for example, by using desirability functions.

In the current setup, the validation set is the same throughout the full approach. Since many steps involved make use of the same validation set (DoE, bootstrap), it may be that the final model is heavily focused toward this set such that the obtained model performance does not fully resemble the expected model performance for new samples. We have investigated this issue by using an additional validation set for the Corn data set, and it appeared that the RMSEP of this additional validation set was only slightly worse compared to that of the original validation set (see the Supporting Information). Also, the same preprocessing steps were deemed relevant. However, splitting the data in three parts is recommended if the data size permits to do so. As a separate study, it may also be investigated whether bootstrapping the validation set as well is advantageous, if the data cannot be split in three parts.

Additional validation may also be used to evaluate whether three- and four-factor interactions could indeed be neglected. These interactions should not be neglected if a model including these interactions is significantly better than a model including only two-factor interactions. Since interpretation of the higher-order effects is not straightforward, it may be best to deem all steps relevant for the subsequent optimization when higher-order interactions appear relevant.

Flåten and Walmsley have also used an DoE approach to optimize calibration model parameters, including a few parameters related to preprocessing.[34] Their approach is however fundamentally different from our work: in their paper, only one method was used for each preprocessing step, e.g., only derivatization for baseline correction. In other words, it was decided on beforehand that baseline correction should be performed using derivatives. Our approach tries to avoid this by first evaluating whether baseline correction should be performed *at all*, before optimizing a specific method. Moreover, their approach also involves optimizing model parameters such as the no. of LVs via DoE, while the current approach optimizes this using cross-validation.

Calculation time for the current approach is approximately 15−30 min, depending on the number of bootstrap samples, the maximum number of LVs in cross-validation, the number of different preprocessing steps and methods to consider, and the size of the data set. Calculation time does not depend on the number of relevant preprocessing steps, since the sequential optimization procedure is very fast. The evaluation of all 4 900 different preprocessing strategies, also including proper cross-validation for each strategy, already takes a day for a relatively small data set such as the caustic scrubber data and is therefore not feasible. For larger data sets and/or more different preprocessing steps and methods, the difference in calculation time between our approach and the evaluation of all possible strategies will increase even further.

CONCLUSION

In this paper, we have presented a novel and much needed approach to obtain an optimal preprocessing strategy within reasonable time. The approach uses design of experiments to systematically evaluate the influence of each preprocessing step on the final strategy. On the basis of the outcome of the design, preprocessing methods should be optimized for each step leading to a decrease in model performance.

To summarize, a user first selects the relevant preprocessing steps (i.e., factors in the design) for a specific data set. Next, a specific experimental design is chosen (e.g., full factorial) and the preprocessing methods to be used at the high level are defined, the low level always implies *do nothing*. The design is then applied, effect values are calculated for each factor (and possibly interactions) and a conclusion is drawn for each preprocessing step whether it is relevant for the data under study or not. Additional factor levels may be used to strengthen these conclusions (*random strategies*). Finally, the relevant steps, if any, can be further optimized by, e.g., the proposed sequential optimization. In this work, a selection of preprocessing steps and methods is provided that covers the basic preprocessing requirements for spectroscopic data.

All examples investigated in this work show considerable improvement in model performance in reasonable time using our approach. Furthermore, the approach is generic and can be applied to data from many different analytical platforms.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b02832.

Details on selection procedure for number of LVs; additional results for a selection of different data sets; results of using an additional validation set; and experimental description of additional data sets including references (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: chemometrics@science.ru.nl. Phone: +31-24-3653192. Fax: +31-24-3652653.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research received funding from The Netherlands Organization for Scientific Research (NWO) in the framework of Technology Area COAST.

DOI: 10.1021/acs.analchem.5b02832
Anal. Chem. 2015, 87, 12096−12103

REFERENCES

(1) Rinnan, A.; van den Berg, F.; Engelsen, S. B. *TrAC, Trends Anal. Chem.* **2009**, *28*, 1201−1222.
(2) Cen, H. Y.; He, Y. *Trends Food Sci. Technol.* **2007**, *18*, 72−83.
(3) Zeaiter, M.; Roger, J. M.; Bellon-Maurel, V. *TrAC, Trends Anal. Chem.* **2005**, *24*, 437−445.
(4) Daszykowski, M.; Stanimirova, I.; Bodzon-Kulakowska, A.; Silberring, J.; Lubec, G.; Walczak, B. *J. Chromatogr. A* **2007**, *1158*, 306−317.
(5) Smolinska, A.; Blanchet, L.; Buydens, L. M. C.; Wijmenga, S. S. *Anal. Chim. Acta* **2012**, *750*, 82−97.
(6) Lommen, A. *Anal. Chem.* **2009**, *81*, 3079−3086.
(7) Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. C. *TrAC, Trends Anal. Chem.* **2013**, *50*, 96−106.
(8) Engel, J.; Blanchet, L.; Buydens, L. M. C.; Downey, G. *Talanta* **2012**, *99*, 426−432.
(9) Phelan, M. K.; Barlow, C. H.; Kelly, J. J.; Jinguji, T. M.; Callis, J. B. *Anal. Chem.* **1989**, *61*, 1419−1424.
(10) Grant, A.; Davies, A. M. C.; Bilverstone, T. *Analyst* **1989**, *114*, 819−822.
(11) Seasholtz, M. B. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 55−63.
(12) Rambla, F. J.; Garrigues, S.; delaGuardia, M. *Anal. Chim. Acta* **1997**, *344*, 41−53.
(13) Kucheryavskiy, S.; Lomborg, C. J. *Food Chem.* **2015**, *176*, 271−277.
(14) Goodarzi, M.; Sharma, S.; Ramon, H.; Saeys, W. *TrAC, Trends Anal. Chem.* **2015**, *67*, 147−158.
(15) Chen, Q. S.; Zhao, J. W.; Liu, M. H.; Cai, J. R.; Liu, J. H. *J. Pharm. Biomed. Anal.* **2008**, *46*, 568−573.
(16) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(17) Dejong, S. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251−263.
(18) Osten, D. W. *J. Chemom.* **1988**, *2*, 39−48.
(19) Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A. *Metabolomics* **2008**, *4*, 81−89.
(20) Vandervoet, H. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 313−323.
(21) Brereton, R. G.; Lloyd, G. R. *J. Chemom.* **2014**, *28*, 213−225.
(22) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wikström, C.; Wold, S. *Design of Experiments: Principles and Applications*; MKS Umetrics AB: Umea, Sweden, 2008.
(23) Eilers, P. H. C. *Anal. Chem.* **2003**, *75*, 3631−3636.
(24) Eilers, P. H. C. *Anal. Chem.* **2004**, *76*, 404−411.
(25) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. *Appl. Spectrosc.* **1989**, *43*, 772−777.
(26) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627.
(27) Wold, S.; Antti, H.; Lindgren, F.; Ohman, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175−185.
(28) Lazić, Z. i. R. *Design of Experiments in Chemical Engineering: A Practical Guide*; Wiley-VCH: Weinheim, Germany, 2004; p 313.
(29) Rinnan, A. *Anal. Methods* **2014**, *6*, 7124−7129.
(30) Martens, H.; Stark, E. *J. Pharm. Biomed. Anal.* **1991**, *9*, 625−635.
(31) Martens, H.; Nielsen, J. P.; Engelsen, S. B. *Anal. Chem.* **2003**, *75*, 394−404.
(32) Chong, I. G.; Jun, C. H. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103−112.
(33) Tran, T. N.; Afanador, N. L.; Buydens, L. M. C.; Blanchet, L. *Chemom. Intell. Lab. Syst.* **2014**, *138*, 153−160.
(34) Flåten, G. R.; Walmsley, A. D. *Analyst* **2003**, *128*, 935−943.

DOI: 10.1021/acs.analchem.5b02832
Anal. Chem. 2015, 87, 12096−12103