**OPEN ACCESS DOCUMENT**

Information of the Journal in which the present paper is published:

* Wiley, Journal of Chemometrics, 2015,
* DOI: dx.doi.org/10.1002/cem.2736

Comparison of the Variable Importance in Prediction (VIP) and of the Selectivity Ratio (SR) variable selection methods in the analysis of three different data sets

Mireia Farrésa, Stefan Y. Platikanova, Stefan L. Tsakovskib and Romà Taulera\*

*a Department of Environmental Chemistry (IDAEA-CSIC), Jordi Girona 18, 08034, Barcleona.*

*b Department of Analytical Chemistry, Faculty of Chemistry, Sofia University, James Bourchier Blvd, 1164.*

\*Corresponding author. *E mail address:* roma.tauler@idaea.csic.es (R. Tauler)

ABSTRACT

This study compares the application of two variable selection methods in Partial Least Squares Regression (PLSR), the Variable Importance in Prediction (VIP) method and the Selectivity Ratio (SR) method. For this purpose, three different datasets were analysed: a) physiochemical water quality parameters related to sensorial data, b) GC-MS chemical (organic compound) profiles from fossil sea sediment samples related to sea surface temperature (SST) changes, and c) exposed genes of *Daphnia magna* females related to their total offspring production. Correlation coefficients (R), levels of significance (p-value) and interpretation of the underlying experimental phenomena allowed the discussion about the best approach for variable selection in each case. The comparison of the two variable selection methods in the first water quality data set showed that the SR method is more accurate for sensorial prediction. For the climate data set, when raw TIC GC-MS chromatograms were considered, variables selected using the VIP method were easier to interpret compared to those selected by the SR method. However, when only selected chromatographic peak areas were considered, the SR method detected more efficiently the most relevant variables for SST changes. For the transcriptomic data set, the SR method was found again to be more reliable for variable selection purposes

Keywords: Selectivity Ratio, Variable Importance in Projection, variable selection, partial least squares, (3 to 5 words)

1. INTRODUCTION

The analysis of multi- and megavariate data has become increasingly important in diverse scientific fields like in gene expression microarray data analysis, gas and liquid chromatography-mass spectrometry (GC- and LC-MS), FT-IR, Raman, NMR and MS spectroscopies and hyperspectral imaging, among other. In all these cases, variable selection techniques are a critical step to obtain a good prediction performance and to explain the underlying phenomena. For predicting one or several parameters from a multivariate data set, multivariate linear calibration models based on latent variables like Principal Component Regression ([1](#_ENREF_1)) and Partial Least Squares Regression ([2-4](#_ENREF_2)) methods are used. These methods can process very large data sets even when the number of variables is much larger than the number of samples ([5](#_ENREF_5)). In many cases, most of these variables are little relevant to the investigated problem as they represent variation not related to the response to be modelled and their number can be drastically reduced with minor loss of information. Variable selection methods help selecting a small set of very relevant predictor variables which are correlated to a particular response variable. Variable selection can improve the estimation accuracy by effectively identifying the subset of important predictors and can enhance the model interpretability with parsimonious representation.

There are many approaches that have been proposed as variable selection methods, a large number of them have been extensively described in previous works ([5](#_ENREF_5),[6](#_ENREF_6)). One of the most popular variable selection methods at present is the variable importance in partial least squares (PLS) projection method, which was proposed in 1993 by Wold et al. ([7](#_ENREF_7)) as ‘Variable influence on projection’ (VIP) which is also known as ‘variable importance in projection scores’ or VIP scores by ([8](#_ENREF_8)) . VIP scores are useful in understanding **X** space predictor variables that best explain **y** variance. VIP scores give an estimate of the contribution of a given predictor to a PLS regression model. Target Projection (TP) with Selectivity Ratio (SR) is another popular method which was also proposed as a tool for variable selection in multivariate data analysis ([9](#_ENREF_9)). Variable selectivity ratios are obtained by calculating the ratio of explained to residual variance of the **X** variables on the y target-projected component.

VIP and SR are two of the most frequently used methods in chemometrics for variable selection. VIP scores selection method has been extensively used in different fields and thus for a variety of data types ([10-16](#_ENREF_10)). Recently, Galindo-Prieto and coauthors ([17](#_ENREF_17)) proposed a new VIP approach for orthogonal projections to PLS latent structures (OPLS) to enhance model interpretability. Although the number of applications in scientific works of the SR method is lower than the VIP scores method, SR method is significantly increasing its use at present ([9](#_ENREF_9),[18-21](#_ENREF_18)). VIP and SR methods have been compared in some scientific works together with other variable selection methods ([9](#_ENREF_9),[22](#_ENREF_22),[23](#_ENREF_23)). Rajalahti and coauthors ([9](#_ENREF_9)), compared SR and VIP on mass spectral profiles and stated that VIP approach was not working for biomarker selection, because it proposed too many false biomarkers. Likewise, Tran and coauthors ([23](#_ENREF_23)) compared SR and VIP methods on NIR spectra datasets, and it was found that SR was more reliable for datasets with noisy variables when compared to the VIP, despite that SR was too conservative. More recently, ([24](#_ENREF_24)) investigated PLS discriminant analysis (PLS-DA) combined with VIP and SR on gas chromatography coupled with flame ionization detection (GC-FID) data. In this case, the effect of variable selection was monitored using a bootstrap procedure. It was concluded that SR presented best predictive abilities than VIP, and that the VIP method was biased to select those variables that were present at the highest concentrations and presented large absolute sizes (they have large variances in the PLS-DA model). In this work, VIP and SR variable selection methods are compared for the analysis of three different data sets.

In this work three data sets were analysed with these two methods. In the first data set (water quality data set) analysed in this paper, physiochemical water quality parameters were related to sensorial data of the same water samples ([25](#_ENREF_25)). In this case, the application of PLS to physicochemical and sensory data allowed discrimination among groups of panellist evaluators according to their preferences for different water types.

In the second data set(climate data set), GC-MS chemical (organic compound) profiles from fossil sea sediment samples were correlated with sea surface temperature changes. The application of PLSR allowed the identification of organic compounds whose concentrations in sea sediment stratified samples were changing more with sea surface temperature (SST) ([26](#_ENREF_26)).

And in the third data set (transcriptomic data set), genes from *Daphnia magna* females exposed to sublethal doses of the Selective Serotonin Reuptake Inhibitors (SSRIs) fluoxetine and fluvoxamine were correlated to their corresponding total offspring production ([27](#_ENREF_27)). PLSR was used to investigate the correlations between transcriptome data and total offspring production. And genes contributing more to the prediction of the reproduction responses were selected and compared using the proposed variable selection methods ([27](#_ENREF_27)).

 In this work, a discussion about the reasons why the two variable selection methods (VIP and SR) differ when applied to the same three data sets is presented. Also, the possible advantages and disadvantages of applying these two methods are examined.

3. THEORY

3.1 Partial Least Squares Regression (PLSR)

In this work PLSR ([3](#_ENREF_3),[4](#_ENREF_4),[28](#_ENREF_28)) analysis was performed on each data set: water quality data, climate data and transcriptome data; to investigate the more influent variables in the corresponding model. PLSR is a regression method that combines features from PCA and multiple regression to optimise separation between different groups of samples. PLSR provide information about the correlation structures of the variables and structural similarities or dissimilarities among the variables.

3.2 Variable importance in projection (VIP)

The Variable importance in projection (VIP) selection method was first published by Wold and coauthors ([7](#_ENREF_7)). VIP scores summarize the influence of individual **X**-variables on the PLS model. VIP scores are calculated as the weighted sum of squares of the PLS weights, **w\***, which take into account the amount of explained **y**-variance in each extracted latent variable (dimension). VIP scores give a measure useful to select what are the variables which contribute the most to the y variance explanation. For a given model and data set there will always be only one VIP scores-vector, summarizing all components and **y** -variables.

The VIP score for the jth variable is given as:

$VIP\_{j}= \sqrt{\frac{\sum\_{f=1}^{F}w\_{jf}^{2} ∙ SSY\_{f }∙ J}{SSY\_{total} ∙ F}} $ (1)

Where $w\_{jf}$ is the weight value for j variable and f component, $ SSY\_{f }$ is the sum of squares of explained variance for the fth component and J number of **X** variables. $SSY\_{total} $is the total sum of squares explained of the dependent variable, and $F$ is the total number of components. $VIP\_{j}$is a measure of the contribution of each variable according to the variance explained by each PLS component were $w\_{jf}^{2}$ represents the importance of the jth variable.

In case of one-dimensional **Y**-space, **y**, holds:

$SSY\_{f}=b\_{f}^{2}t\_{f}^{'}t\_{f} SSY\_{total}=b^{2}T^{'}T$ (2)

where **T** is the **X** scores matrix and **b** the PLS inner relation vector of coefficients.

Since the average of the squared VIP scores equals 1, ‘greater than one rule’ is generally used as a criterion for variable selection ([29](#_ENREF_29)). This is not a statistically justified limit and it can be shown that it is very sensitive to the presence of non-relevant information pertaining to **X** ([23](#_ENREF_23)).

3.3 Selectivity Ratio (SR)

The Selectivity Ratio (SR) ([9](#_ENREF_9),[21](#_ENREF_21)) method is a visualization tool for searching what are the important variables of a multivariate data set in the prediction of a particular property. The ratio between the explained and the residual (unexplained) variance for each variable in the target projection vector (TP) defines the SR for the variable in question. This target projection utilizes both the predictive ability (regression vector) and the explanatory ability (spectral variance/covariance matrix) for the calculation of the selectivity ratio. Given the PLS regression vector, $b\_{PLS}$, Target Projection is performed via the projection of the rows of **X** onto the normalized regression coefficients vector $b\_{PLS}$in Equation 3. In this Equation $t\_{TP}$is proportional to the predicted values, $ŷ=Xb\_{PLS}$. The loadings, $p\_{TP}$**,** are obtained by projecting the columns of **X** onto the score vectors, **tTP**, which again is proportional to $ŷ=Xb\_{PLS}$ in Equation (3) and (4).

$t\_{TP}= Xb\_{PLS}/\left‖b\_{PLS}\right‖$ (3)

$p\_{TP}=Xt\_{TP}/(t\_{TP}^{'} t\_{TP})$ (4)

The ratio of the explained variance ($SS\_{explained,i}$) and of the residual variance for each variable ($SS\_{residual,i}$) in the sum of squares in Equation (5) and (6), respectively, is used then to determine the variable importance, SR, in Equation (7)

$SS\_{i, explained}=\left‖t\_{TPi}p\_{TPi}'\right‖^{2}$ (5)

$SS\_{i, residual}=\left‖e\_{TPi}\right‖^{2}$ (6)

$SR\_{i}=SS\_{explained,i}/SS\_{residual,i}$ (7)

Rajalahti and coauthors ([21](#_ENREF_21)) proposed a F-test to define a boundary between variable regions with high discriminating ability and less interesting regions. In order to determinate which variable has a high discriminatory ability and to reject the null hypothesis (explained and residual variances are the same), the calculated F value (Fcalc) which is equal to SRi from Eq. 7, has to exceed the critical value for the F distribution, Fcrit.

Fcalc = SRi > Fcrit = F (α, N-2, N-3) (8)

Where N is the sample size and α the significance level. In this work, the F*-*test (95%) criterion has been chosen to select the marker candidate.

4. DATA SETS DESCRIPTION

4.1 Water quality data set

This dataset (see Table 1 supplementary material) consisted of 13 physicochemical parameters (**X** predictor variables) and of one overall score for the water taste and flavour evaluation (**y** predicted variable) of 25 bottled mineral and tap water samples covering a wide range of mineralization and chemical composition from different sources ([25](#_ENREF_25)) . Water samples were analyzed using standard methods. Water blends and dilutions were allowed for 48 h of equilibration before analysis. Sodium, potassium, calcium and magnesium and silica concentration levels were analysed by inductively coupled plasma optical emission spectrometry (ICP-OES). Conductivity at 20 ºC, pH and bicarbonate levels were determined by robotic titrosampler. Chloride, nitrate and sulphate concentrations were analysed by ionic chromatography. TDS (dry residue at 180 ºC) levels were measured by gravimetry. Free residual chlorine was analysed by DPD colorimetric method. Sensory analysis were carried out by a panel of selected people ([30](#_ENREF_30)), which were trained according to a flavour profile analysis (FPA) method previously developed in ([31](#_ENREF_31)). Results of these analysis are given in a Table (supplementary material) which summarizes the collected mineral and tap water samples, their mineral composition described as physicochemical parameters and their overall flavour scores. Further information about samples and analysis procedure is given in ([25](#_ENREF_25)).

4.2 Climate data set

This climate dataset (see Figure 1 supplementary material) consisted of 10 Total Ion Current (TIC) GC-MS chromatographic profiles (2940 retention times) of the neutral fractions extracted from 10 fossil stratified sediment marine samples (**X** predictor variables) and of their corresponding alkenone-based reconstructed sea surface temperature (SST) (y predicted variable). Sediment marine records were taken from site IODP-U1318C, in the continental margin southwest of UK islands during the Integrated Ocean Drilling Program (IODP) Expedition 307 at Challenger Mound in Porcupine Seabight, 420.9 meters below seafloor ([32](#_ENREF_32)). The procedures and equipment for extracting, isolating and quantifying the fossil compounds have been described by ([33](#_ENREF_33),[34](#_ENREF_34)). Purified extract samples were diluted in toluene and analysed using GC-MS. Mass spectra were acquired in electron ionization mode scanning from m/z 42 to 700. Previous to their analysis, the chromatographic profiles of the 10 samples were baseline corrected using asymmetric least squares (AsLS) method ([35](#_ENREF_35),[36](#_ENREF_36)), peak aligned using Correlation Optimized Warping (COW) algorithm ([37](#_ENREF_37),[38](#_ENREF_38)) and mean-centered. Further details of experimental methodology and of data pre-treatment are given in ([26](#_ENREF_26)). Annual mean SST values were reconstructed using a global core-top calibration method and the alkenone unsaturation index UK’37 ([39](#_ENREF_39)) and mean-centered prior to PLSR analysis.

Additionally, a second **X** matrix of predictor variables (see Table 2 supplementary material) was obtained using 77 fossil compounds recognized in the original chromatogram and identified by comparison of their corresponding measured spectra to library data and synthetic standards ([26](#_ENREF_26)). Their chromatographic areas were then integrated to estimate the relative concentration of each of the identified organic compounds. This new **X** matrix was then correlated to the **y** vector having the SST values associated to each of the 10 analysed samples.

4.3 Transcriptome data set

Transcriptome dataset consists of the microarray analysis transcriptional stress responses (1207 deregulated gene fragments) of 12 *D. magna* female samples (**X** matrix of predictor variables) exposed to sublethal doses of the Selective Serotonin Reuptake Inhibitors (SSRIs): fluoxetine and fluvoxamine and their corresponding total offspring production (**y** predicted variables). Tests were performed to determine the effects of the chemicals of interest on transcriptomic responses and reproduction rates on adult stages. Microarrays analyses were performed on isolated RNA from the samples. Microarray results were validated with real-time quantitative polymerase chain reaction (qPCR). Further information about the experiments is found in ([27](#_ENREF_27)).

4. RESULTS AND DISCUSSION

4.1 Water quality data set results

PLSR was applied first to the water quality data set which include sensorial data values (as **y**- variable). **X** (water physicochemical parameters) and **y** (water taste scores) variables were autoscaled. Using leave-one-out cross-validation, PLS modelling resulted in a two-latent variables model which accounted for 68.43% of the **X** data variance and around 89% of the **y** variance.

Table 1 shows the pair-wise correlation coefficients (R) between water quality physicochemical parameters (column variables of **X-**matrix) and water taste scores (**y**-vector), with their level of significance (p-value). The selection obtained using VIP and SR approaches is provided in Figure 1 together with the list of variables correlated to the predicted response **y**.

Figure 1 shows the comparison of the previously shown correlation coefficients, R (Table 1), with the variables selected using VIP and SR methods. The variables along the x-axis are arranged according to their correlation coefficients as it is shown in the map legend at the right side (whiter color means higher correlation).. VIP and SR selected variables are represented in white and non-selected variables are in black

Figure 1 shows that almost the same number of variables were selected by the VIP scores and SR methods, considering a threshold greater than one for VIP scores and 2.04 for SR (F-test, 95%). When the correlation coefficient of variables selected by VIP and SR methods were compared, variables with absolute value of R higher than 0.76 (Table 1) were the same in both variable selection methods (Figure 1). Variable number 10 (Table 1), potassium, with an R value of -0.69 was selected by the VIP method but not by the SR method. Originally, potassium was in such low concentrations in all water samples that was no perceivable in taste by panellists, only sodium (Table 1) was in such a concentration that had effects on taste ([25](#_ENREF_25)). In the VIP calculation, due to the **X**-variable autoscaling, the weighted sum of squares of the PLS weights resulted to be almost equal, and VIP scores resulted to be the same. The selection of potassium variable by the VIP method should be considered doubtful because of their low correlation coefficient value with **y** compared to other selected variables, but potassium had spurious correlation with sodium, this is the reason why it is considered in the PLS model. For this example, the SR method resulted to be more accurate since it did not select potassium as an important variable.

4.2. Climate data set results

In this case, PLSR was initially applied to TIC GC-MS chromatograms of the neutral fractions extracted from 10 fossil sediment marine samples (**X**-variables, with dimensions of 10x2940) and to the reconstructed sea surface temperature variables, SST values (**y**-variable, with dimensions 10x1). Using leave-one-out cross-validation, first PLS latent variable (LV1) accounted for 64.59% of **X** data variance and as much as 93.01% of the dependent variable**y** (SST) variance. PLSR results confirmed therefore, the presence of a strong correlation pattern between the 10 TIC chromatographic elution profiles of the neutral fraction and the SST changes in sediment samples.

From the application of PLSR to the pre-processed TIC chromatograms, the VIP method selected 50 variables with VIP scores greater than one, and the SR method selected 58 variables with SR scores greater than 3.73 (F-test, 95%). In Figure 1, the selection obtained using VIP and a SR method is provided, together with the list of variables that should have a relation to the predicted response **y**. All SR variables had absolute R values higher than 0.75 whereas there were a subset of the VIP variables which had absolute R values lower than 0.6. The SR method gives a higher sensitivity and the VIP method a lower specificity. However, there are a subset of variables selected by the SR method (Figure 2) with high correlation coefficients with **y** but not selected by the SR method.

When the SR variable selection method was applied to raw chromatographic data, SR scores resulted higher when the corresponding correlation coefficients with the **y**-variable (Table supplementary material) were also higher, close to one, regardless the shape of the chromatographic peak. In some cases, significant SR values were obtained at regions of the chromatographic profile where no peak was present. In Figure 3 two examples are given, where TIC chromatogram regions at retention times 11.6 and 12.1 minutes were chemically meaningless, but SR values were rather high. In contrast, in Figure 3, all VIP scores are placed at retention times where, indeed, there are chromatographic peaks.

PLSR analysis was also applied to the autoscaled peak areas (concentrations) of the identified organic compounds in the GC-MS analysis (**X**-matrix of predictor variables, of dimensions of 10x77) of sea sediment samples. Using a leave-one-out cross-validation strategy, PLSR LV1 accounted for 42.70% of the **X** data variance and for almost 93% of the **y** data variance. In this case, from the 77 **X** variables tested, there were 42 with VIP scores greater than one. And 16 variables were significant according to the F-test (95%) for the SR method (see Figure 4). All these 16 SR variables were in the same group as the 42 VIP variables.

Correlation coefficients of the 16 variables selected by both, VIP and SR, methods were greater than 0.8 (p < 0.005). There were three variables with absolute values of R near to one (p < 0.005), which were rejected by VIP and not by SR (variables 4, 15 and 73 of concentration **X** matrix, in Table 5 of supplementary material). All other variables selected by VIP and not by SR methods presented an absolute R value between 0.62 and 0.72 (p < 0.005).

When variables selected using VIP method from the GC-MS TIC chromatograms (50 in total) were compared with those selected from the concentrations of the 77 identified organic compounds (42 in total), 24 common variables were encountered to be relevant in both analysis. All 24 variables had an absolute R value higher than 0.75 (p < 0.05) for the TIC chromatograms, and higher than 0.62 (p < 0.05) for the peak areas. When comparing the variables selected by SR scores from TIC chromatograms (58 in total) with those selected from peak areas (16 in total), 11 variables proved to be common in both analyses with absoluter R values higher than 0.90 (p < 0.0005)

To summarize, 11 variables were selected to be important by VIP and SR methods using both the TIC chromatograms data set and the peak areas data set (see Table 2). Table 2 shows that short chain *n*-alkanes (eicosane, heneicosane, docosane, tricosane) and long chain *n*-alkanes (from pentacosane to dotriacontane) were in high abundances at lowest SST values.

Short chain *n*-alkanes originate from a variety of sources, such as marine algae and bacteria ([40](#_ENREF_40),[41](#_ENREF_41)). Long chain *n*-alkanes have a terrestrial source because they are major lipid molecules in epicuticular waxes of vascular plants, and they are common components of eolian winds dust. As discussed in ([26](#_ENREF_26)), climatic processes at lower temperatures transported higher fluxes of terrigenous material (terrestrial compounds) at IODP U1318C site.

The analysed TIC chromatograms contained many variables (retention times) and many of them had no chemical meaning (not related with chemical compounds). Therefore when they were analysed by PLSR, clear differences between SR and VIP variable selection methods were encountered. Since the SR method is based on a variance ratio evaluation (see method section) and each retention time is an individual variable, many of the selected variables by SR were not chemically interpretable (Figure 3). Although correlation coefficients of selected variables by SR were also high, they could not be considered to be reliable markers of the sought changes in SST values.

VIP scores obtained in the PLSR analysis of the TIC chromatograms data set with one single component (LV1) were easier to interpret and gave a profile with chromatographic shape. In contrast to variables selected by the SR method (Figure 2), VIP selected variables had absolute R values lower than 0.6 (Figure 2, Table 3 supplementary material), which means that some of them were less correlated with SST changes. However, the number of relevant variables selected by the SR method was lower for the chromatographic peak areas data set than for the TIC chromatograms data set. This was because in this case, when compound concentrations (integrated peak areas) were used, baseline, background and noisy contributions were discarded and the possibility of false positives was drastically diminished.

4.3 Transcriptomic data set results

PLS analysis of the transcriptomic data set consisted of control and SSRIs treated daphnia samples. Using leave-one-out cross-validation PLSR modelling resulted in a two-latent variables model that accounted for 57.8% of X data variance and explained around 80% of the total y variance.

VIP scores greater than 1 and SR scores higher than 3.14 (F-test, 95%) were selected and compared. Only 6 variables resulted significant by the SR scores method, with 5 of them coincident with those selected by the VIP scores method (271 in total). Owing to the fact that variable selection methods aim at selecting a small set of very relevant variables, and that with a VIP threshold greater than one gave too many variables, VIP threshold value was incremented to 2 and 154 variables resulted important. In this group, the same 5 coincident variables (with VIP and SR method) were again included. Finally, when VIP threshold was incremented up to 3, a total number of 84 variables resulted to be important, but now only 3 of them coincided with those variables selected by SR scores (Figure 3).

Correlation coefficients (R) between the variables selected by SR (F-test 95%) and VIP scores (threshold of 2 and 3) were calculated (see Table 6 supplementary material). 56 variables with VIP scores higher than 2 presented absolutes correlation coefficients of 0.6 or higher (p < 0.05), and 42 variables with VIP scores higher than 3 presented absolutes correlation coefficients of 0.6 or higher (p < 0.05). It has to be pointed out that some of the variables selected by the VIP scores method, as shown in Figure 5, were little correlated with the independent variable. There were some variables with very low correlation coefficients (with R values lower than 0.1) selected by the VIP scores method even with a threshold of 2 and 3 (see table 6 Supplementary material). VIP method finds variables that are important not only because of their possible correlation with the **y** variable, but also because they describe significantly the **X** variance ([5](#_ENREF_5)). In contrast, the six variables selected by the SR scores method presented always a large absolute correlation coefficient value, higher than 0.64 (p < 0.05) (see Table 6 supplementary material). Therefore, in terms of number of false positives, the SR method should be considered a better variable selection method.

In this transcriptomic data example, there were a large number of variables which were selected as relevant using the VIP scores method which were not selected as relevant using the SR method, even though their correlation coefficients with the **y**-variables were also high (supplementary material). As stated previously by other authors ([23](#_ENREF_23)), as a consequence of the reduced number of degrees of freedom in the F-test used in the SR method a very small number of variables are finally selected, which is sometimes a very conservative decision. A large number of false positives are therefore excluded by the SR variable selection method, yet this can imply the cost of excluding also some relevant variables. Therefore, the number of false negatives may be the main disadvantage of the SR method, as it has been shown for this transcriptomic data set and also previously for the climatic data set (see section 4.2), where some highly correlated variables with the **y** vector were discarded by the SR variable selection method.

Two new PLSR analyses were performed using the same data set of this section 4.3.But, in this case, the initial total number of 1207 variables was reduced selecting only the relevant variables from the previous developed PLSR model. In the first analysis, the VIP selected (threshold greater than 3) variables were used (**X**-matrix of predictor variables of dimensions of 1207x84) and in the second analysis, the SR selected variables were used (**X**-matrix of predictor variables of dimensions of 1207x6). PLSR modelling using VIP selected variables resulted in two-latent variables model that accounted for 47.69% of **X** variance and for 93.35% of **y** variance. The second PLSR model using SR selected variables with two latent variables 2 LV, accounted for 89% of **X** variance and for almost 81% of **y** variance. If the aim of the variable selection method is to extract a low number of possible biomarkers of the investigated treatment (total offspring production of females treated with SSRIs), SR method should be considered more accurate. As clearly shown above, half of the variables selected by VIP already explained a 93% of **y** variance. Whereas in the case of the SR method, with only 6 variables selected by SR, already a total of 81% of the **y** variance was explained. In terms of transcriptomic data interpretation, 6 genes may not be enough for testing the hypothesis of the study. In the present study, this hypothesis was whether low concentrations of SSRIs affected offspring production and/or juvenile developmental rates though different mechanisms of action ([27](#_ENREF_27)). The number of genes involved in *D. magna* metabolic pathways is most probably in the hundreds and their identification is a complex task, therefore selection (by SR method) of only 6 genes is insufficient/scarce for a global interpretation of the changes in the metabolic pathways. In this case the SR threshold could be lowered as it was done in previous work ([27](#_ENREF_27)), where those genes with SR values higher than the mean were finally considered to be the most relevant for explaining the offspring.

5. CONCLUSIONS

In this paper, VIP and the SR variable selection methods have been compared for three different datasets. In general, the VIP method selected a higher number of variables than the SR method. Variables selected by the VIP method were sometimes false positive candidates, while those selected by the SR method gave false negative candidates. VIP scores variable selection method was more reliable than the SR method for large datasets such as the raw chromatographic data in the climate data set. In contrast, for other types of preprocessed or transformed data sets, like for the physicochemical variables dataset, both methods detected efficiently the most relevant variables. In the transcriptomic data set and in the t when chromatographic peak areas (concentrations) from the climate data set were used, variables selected by the SR method described most of the **y** variance, whereas, only a low number of variables selected by VIPs were contributing to the description of the **y** variance. Final decision about the best approach should be performed according to the optimal description of the experimental data and from the sounder interpretation of results.

REFERENCES

1. Næs T, Martens H. Principal component regression in NIR analysis: Viewpoints, background details and selection of components. Journal of Chemometrics 1988;2(2):155-167.

2. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Analytica Chimica Acta 1986;185(0):1-17.

3. Wold H. Estimation of Principal Components and Related Models by Iterative Least squares. Multivariate Analysis: Academic Press; 1966. p 391-420.

4. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 2001;58(2):109-130.

5. Andersen CM, Bro R. Variable selection in regression—a tutorial. Journal of Chemometrics 2010;24(11-12):728-737.

6. Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems 2012;118(0):62-69.

7. Wold S, Johansson A, Cochi M, editors. PLS-partial least squares projections to latent structures Leiden: ESCOM Science Publishers; 1993. 523-550 p.

8. Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C, Wold S. Multi- and Megavariate Data Analysis, Part 1, Basic Principles and Applications: Umetrics AB; 2006.

9. Rajalahti T, Arneberg R, Berven FS, Myhr K-M, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemometrics and Intelligent Laboratory Systems 2009;95(1):35-48.

10. Pérez-Enciso M, Tenenhaus M. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. Hum Genet 2003;112(5-6):581-592.

11. Modlich O, Prisack H-B, Munnes M, Audretsch W, Bojar H. Predictors of primary breast cancers responsiveness to preoperative Epirubicin/Cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signatures. Journal of Translational Medicine 2005;3(1):32.

12. Farrés M, Piña B, Tauler R. Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC–MS. Metabolomics 2014:1-15.

13. Serranti S, Cesare D, Marini F, Bonifazi G. Classification of oat and groat kernels using NIR hyperspectral imaging. Talanta 2013;103(0):276-284.

14. Eriksson L, Hermens JM, Johansson E, Verhaar HM, Wold S. Multivariate analysis of aquatic toxicity data with PLS. Aquatic Science 1995;57(3):217-241.

15. Yang J, Zhao X, Liu X, Wang C, Gao P, Wang J, Li L, Gu J, Yang S, Xu G. High Performance Liquid Chromatography−Mass Spectrometry for Metabonomics:  Potential Biomarkers for Acute Deterioration of Liver Function in Chronic Hepatitis B. Journal of Proteome Research 2006;5(3):554-561.

16. Eriksson L, Gottfries J, Johansson E, Wold S. Time-resolved QSAR: an approach to PLS modelling of three-way biological data. Chemometrics and Intelligent Laboratory Systems 2004;73(1):73-84.

17. Galindo-Prieto B, Eriksson L, Trygg J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). Journal of Chemometrics 2014:n/a-n/a.

18. Gomez-Carracedo MP, Ferre J, Andrade JM, Fernandez-Varela R, Boque R. Objective chemical fingerprinting of oil spills by partial least-squares discriminant analysis. Analytical and Bioanalytical Chemistry 2012;403(7):2027-2037.

19. Rajalahti T, Kroksveen AC, Arneberg R, Berven FS, Vedeler CA, Myhr K-M, Kvalheim OM. A Multivariate Approach To Reveal Biomarker Signatures for Disease Classification: Application to Mass Spectral Profiles of Cerebrospinal Fluid from Patients with Multiple Sclerosis. Journal of Proteome Research 2010;9(7):3608-3620.

20. Karimi S, Hemmateenejad B. Identification of discriminatory variables in proteomics data analysis by clustering of variables. Analytica Chimica Acta 2013;767(0):35-43.

21. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr K-M, Kvalheim OM. Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles. Analytical Chemistry 2009;81(7):2581-2590.

22. Andries JPM, Heyden YV, Buydens LMC. Predictive-property-ranked variable reduction in partial least squares modelling with final complexity adapted models: Comparison of properties for ranking. Analytica Chimica Acta 2013;760(0):34-45.

23. Tran TN, Afanador NL, Buydens LMC, Blanchet L. Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). Chemometrics and Intelligent Laboratory Systems 2014;138(0):153-160.

24. Krakowska B, Stanimirova I, Orzel J, Daszykowski M, Grabowski I, Zaleszczyk G, Sznajder M. Detection of discoloration in diesel fuel based on gas chromatographic fingerprints. Analytical and Bioanalytical Chemistry 2014:1-12.

25. Platikanov S, Garcia V, Fonseca I, Rullán E, Devesa R, Tauler R. Influence of minerals on the taste of bottled and tap water: A chemometric approach. Water Research 2013;47(2):693-704.

26. Farrés M, Martrat B, Grimalt JO, Tauler R. Chemometrics extraction of climatic signals from fossil organic molecules accumulated in marine sediments up to 8.65 Ma old (IODP-U1318). 2014.

27. Campos B, Garcia-Reyero N, Rivetti C, Escalon L, Habib T, Tauler R, Tsakovski S, Piña B, Barata C. Identification of Metabolic Pathways in Daphnia magna Explaining Hormetic Effects of Selective Serotonin Reuptake Inhibitors and 4-Nonylphenol Using Transcriptomic and Phenotypic Responses. Environmental Science & Technology 2013;47(16):9434-9443.

28. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Analytica Chimica Acta 1986;185:1-17.

29. Chong I-G, Jun C-H. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 2005;78(1–2):103-112.

30. Devesa R, Fabrellas C, Cardeñoso R, Matia L, Ventura F, Salvatella N. The panel of Aigües de Barcelona: 15 years of history. Water Science and Technology. Volume 49; 2004. p 145-151.

31. Devesa R, Cardeñoso R, Matía L. Contribution of the FPA tasting panel to decision making about drinking water treatment facilities. Water Science and Technology. Volume 55; 2007. p 127-135.

32. Ferdelman TG, Kano A, Williams T, Henriet J-P, Scientists E. Expedition 307 Scientists, 2006. Site U1318. Proc IODP: Washington, DC (Integrated Ocean Drilling Program Management International, Inc.); 2006.

33. Villanueva J, Grimalt JO. Pitfalls in the chromatographic determination of the alkenone U37k index for paleotemperature estimation. Journal of Chromatography A 1996;723(2):285-291.

34. Villanueva J, Pelejero C, Grimalt JO. Clean-up procedures for the unbiased estimation of C-37 alkenone sea surface temperatures and terrigenous n-alkane inputs in paleoceanography. Journal of Chromatography A 1997;757(1-2):145-151.

35. Eilers PHC. A Perfect Smoother. Analytical Chemistry 2003;75(14):3631-3636.

36. Eilers P, Boelens H. Baseline Correction with Asymmetric Least Squares Smoothing. 2005.

37. Nielsen N-PV, Carstensen JM, Smedsgaard J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. Journal of Chromatography A 1998;805(1-2):17-35.

38. Tomasi G, Berg Fvd, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. Journal of Chemometrics 2004;18(5):231-241.

39. Müller PJ, Kirst G, Ruhland G, von Storch I, Rosell-Melé A. Calibration of the alkenone paleotemperature index U37K′ based on core-tops from the eastern South Atlantic and the global ocean (60°N-60°S). Geochimica et Cosmochimica Acta 1998;62(10):1757-1772.

40. Blumer M, Guillard RRL, Chase T. Hydrocarbons of marine phytoplankton. Marine Biology 1971;8(3):183-189.

41. Volkman JK, Johns RB, Gillan FT, Perry GJ, Bavor Jr HJ. Microbial lipids of an intertidal sediment—I. Fatty acids and hydrocarbons. Geochimica et Cosmochimica Acta 1980;44(8):1133-1143.