Analytica Chimica Acta 1075 (2019) 57–70

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Comprehensive new approaches for variable selection using ordered predictors selection

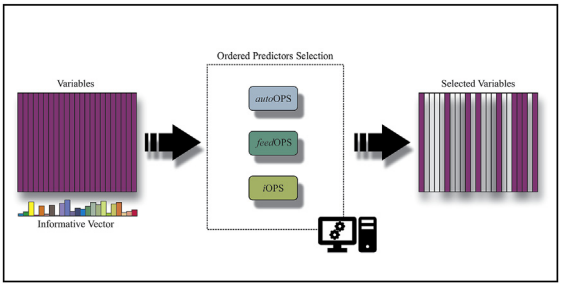
Jussara V. Roque [a], Wilson Cardoso [a], Luiz A. Peternelli [b], Reinaldo F. Teófilo [a, *]

[a] Multivariate Chemical Data Analysis Laboratory, Department of Chemistry, Universidade Federal de Viçosa, 36570-900, Viçosa, MG, Brazil
[b] Department of Statistics, Universidade Federal de Viçosa, 36570-900, Viçosa, MG, Brazil

HIGHLIGHTS

- New OPS algorithms presented high performance for different data structures.
- OPS algorithms were compared to other chemometric feature selection methods.
- OPS presented to be more predictable, reproducible, interpretative, and universal.
- In general, OPS approaches outperformed studied feature selection methods.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 28 January 2019
Received in revised form
15 May 2019
Accepted 16 May 2019
Available online 20 May 2019

Keywords:
Multivariate regression
Feature selection
Chemometrics
Informative vector
Prediction power

ABSTRACT

New strategies of ordered predictors selection (OPS) were developed in this work, making this method more versatile and expanding its worldwide use and applicability. OPS is a recognized method to select variables in multivariate regression and is used by analytical chemists and chemometrists. It shows high ability to improve the prediction of models after the selection of a few and important variables. At the core of OPS is sorting variables from informative vectors and systematically investigating the regression models to identify the most relevant set of variables by comparing the cross-validation parameters of the models. Nevertheless, the first version of the OPS method performs variable selection using only one informative vector at a time and is limited to just one variable selection run. Then, three new strategies were proposed. First, an automatic method was developed to perform variable selection using several informative vectors and their combinations. Second, the feedback OPS is presented, in this new strategy the pre-selected variables would return to a new selection. Last, a method to apply OPS in full array subdivisions called OPS intervals was established. Initially, the new strategies were applied in the six datasets used in the original OPS paper to compare the prediction performance with the new OPS algorithms. After that, twelve new datasets were used to test and compare the new OPS approaches with other variable selection methods, genetic algorithm (GA), the interval successive projections algorithm for PLS (*i*SPA), and recursive weighted partial least squares (rPLS). The new OPS approaches outperformed the first OPS version and the other variable selection methods. Results showed that in addition to greater predictive capacity, the accuracy in the selection of expected variables is highly superior with the new OPS approaches. Overall, the new OPS provided the best set of selected variables to build more predictive and interpretative regression models, proving to be efficient for variable selection in different types of datasets.

* Corresponding author. Department of Chemistry, Universidade Federal de Viçosa, 36570-900, Viçosa, MG, Brazil.
E-mail addresses: rteofilo@gmail.com, rteofilo@ufv.br (R.F. Teófilo).

https://doi.org/10.1016/j.aca.2019.05.039
0003-2670/

Abbreviations

*auto*OPS	automatic ordered predictors selection
Cal	calibration
COR	correlation between each column of matrix **X** with **y**
COV	covariance procedures
URXY	univariate regression between each column of matrix **X** with **y**
*feed*OPS	feedback ordered predictors selection
GA	genetic algorithm
GC	gas chromatography
*h*Mod	number of latent variables of the model
*h*OPS	number of latent variables for ordered predictors selection
hOPS	number of latent variables to generate the best informative vector in OPS method
*i*OPS	ordered predictors selection by intervals
*i*SPA	intervals successive projections algorithm for PLS
MLR	multiple linear regression
MMP-2	matrix metalloproteinases type 2
MS	mass spectrometry
MSC	multiplicative scatter correction
NAS	net analyte signal
NIPALS	nonlinear iterative partial least squares
NIR	near-infrared spectroscopy
NMR	nuclear magnetic resonance spectroscopy
OPS	ordered predictors selection
OPS*v*1	first version of ordered predictors selection
PLS	partial least squares
Pred	prediction
PRODALL	the product of all single vectors simultaneously
QSAR	quantitative structure-activity relationship;
R	correlation coefficient (R_c) of calibration, (R_{cv}) of cross-validation, and (R_p) of prediction
REG	regression coefficients
RMSE	root mean square error (*RMSEC*) of calibration, (*RMSECV*) of cross-validation and (*RMSEP*) of prediction
rPLS	recursive weighted partial least squares
SNV	standard normal variate
SQR	residual information of the reconstructed matrix with *h* latent variables
SVMR	support vector machine regression
UV	ultraviolet spectroscopy
VIP	variable importance on projection
Vis/NIR	visible and near infrared spectroscopy
WGHT	weights
XRF	X-ray fluorescence spectrometry

1. Introduction

Multivariate regression models describe the relationship between the dependent and independent variables when more than one measurement is acquired for each sample [1,2]. There are several multivariate regression methods, such as multiple linear regression (MLR), support vector machine regression (SVMR) and partial least squares regression (PLS). However, the latter is the most commonly used for building first order inverse regression models from data of chemical origin [3].

PLS regression can deal with a large number of highly correlated independent variables, band overlaps, and experimental noise [4,5]. Also, it enables simultaneous modeling and the prediction of more than one analyte [2,3]. However, in some situations, the models cannot provide a satisfactory prediction. This may occur because not all variables are equally essential and variations in the concentrations of the chemical components of interest present in the sample do not cause the same change in all variables [6]. Also, there are non-linear and low signal-to-noise ratio regions that might affect the model. This evidence indicates that the proper selection of variables can significantly improve the efficiency of the multivariate regression model, in addition to making it simpler for interpretations [1,6].

Variable selection is a significant step in multivariate regression, and it has become a fundamental tool in many different research areas. This is because of increased database dimensions, meaning that some variables may be redundant, irrelevant or represent noise [2,6,7]. Models built after the removal of non-informative variables will produce better predictions, a better interpretation with the selection of markers or biomarkers, lower measurement costs through the use of portable instruments, and a decrease in data processing time [1,6].

Several feature-selection methods [8–11] have been applied in several works in the literature [12–16]. Nevertheless, most selection methods are not generalized or efficient for all types of datasets. Different analytical techniques provide distinctive predictor types, and each technique has some well-defined peculiarities [17]. In 2009, Teófilo et al. presented first version of the ordered predictors selection (OPS) algorithm [17], with the advantage of being simple, fast and efficient for the selection of any variable type. Since the publication in 2009, the OPS method has had its potential recognized in the literature [7,18,19], showing a high ability to improve the prediction of multivariate regression models with few variables [15,20,21]. However, the original OPS algorithm performs the variable selection using only one informative vector at a time and is limited to just one variable selection run, which turn out to be limitations of the original version when searching for the best set of variables.

Therefore, the aim of this work was to make a more versatile OPS, expanding its worldwide use and applicability. The development of three new OPS approaches was proposed: *i*) automatic OPS (*auto*OPS), which automatically performs all calculations using either or both informative vectors and its combinations, so the best one is chosen; *ii*) feedback OPS (*feed*OPS), wherein the pre-selected variables would go through a new selection run; and *iii*) interval OPS (*i*OPS), the option to apply OPS in subdivisions of the full array. The new OPS algorithms were developed to be easily understood and executed. Besides that, the algorithms were applied to several types of dataset (sparse or non-sparse), with the proposal to select interpretive variables with greater predictability, and high reproducibility, *i.e.*, selecting the same set of variables when executing the selection more than once.

2. Theory background

2.1. Model evaluation

Regression models were evaluated using statistical parameters

such as the root mean square error (*RMSE*) and correlation coefficient (*R*), according to equation (1) and equation (2), respectively.

$$RMSE = \sqrt{\sum_{i}^{I_m} (y_i - \hat{y}_i)^2} \Big/ I_m \tag{1}$$

$$R = \sum_{i=1}^{I_m} (\hat{y}_i - \hat{\bar{y}})(y_i - \bar{y}) \Big/ \sqrt{\sum_{i=1}^{I_m} (\hat{y}_i - \hat{\bar{y}})^2} \sqrt{\sum_{i=1}^{I_m} (y_i - \bar{y})^2} \tag{2}$$

where y_i and $\hat{y}_i$ are measured and predicted values for each i sample, respectively. The $\bar{y}$ is the **y** values average. When calibration is used, I_m represents the number of samples in the calibration set, and the error and correlation coefficient are the root mean square error of calibration (*RMSEC*) and the correlation coefficient of calibration (R_c), respectively. When internal cross-validation (CV) is used, I_m represents the number of samples in the cross-validation set, and the error and correlation coefficient are the root mean square error of cross-validation (*RMSECV*) and the correlation coefficient of cross-validation (R_{CV}), respectively. When external validation is used, I_m represents the number of predicted samples (*P*) and, in this instance, the error and correlation coefficient are the root mean square error of prediction (*RMSEP*) and the correlation coefficient of prediction (R_P), respectively.

2.2. OPS theory

The OPS method is based on obtaining an informative vector containing information about the best independent variables for prediction. This informative vector can be obtained of several ways from calculations performed with **X** matrix columns (independent variables) and dependent variables (**y**), and its length is equal to the number of independent variables. The original OPS method calculates informative vectors using the regression coefficients (REG), the correlation between each column of matrix **X** with **y** (COR), residual information of the reconstructed matrix with *h* latent variables (SQR), variable importance on projection (VIP), net analyte signal (NAS), and covariance procedures (COV). Details about these vector calculations can be found elsewhere [17].

After obtaining an informative vector, the independent variables of **X** matrix are differentiated according to their corresponding absolute values in the informative vector. The highest absolute value corresponds to the more important independent variable, and the differentiated variables are sorted in descending order of absolute values magnitude. Multivariate regression models are built and evaluated using cross-validation approach. An initial subset of variables (window) is defined to build and evaluate the first model. After, this initial subset is extended by the addition of a fixed number of variables (increment) over the window, and a new model is built and evaluated. Further increments are added until all or a percentage of variables are taken into account, and cross-validation parameters are calculated for each model. Lastly, the variable subsets are compared using the quality parameters calculated during validations, and the best variable subset is defined [17]. Fig. 1 shows a general scheme for the original OPS algorithm.

Core algorithm

The core algorithm of the OPS method consists of the following steps.

calculate an informative vector;
sort in descending order the absolute values of the informative vector and store the sorting index;
sort **X** variables considering the previous sorting index;
define an initial number of variables (window) to be investigated in the ordered **X** array;
define a fixed number of variables to be added over the window (increment);
build multivariate regression models starting with the initial window and then its extension by addition of increments;
store the index of the variables investigated in each subset (window plus increment);
store cross-validation parameters of each model built by each subset;
choose the best set of variables based on cross-validation parameters.

2.2.1. Automatic OPS (autoOPS)

The choice of informative vector is a critical step in the OPS method. Then, the automatic OPS was proposed to perform all steps of variable selection using several informative vectors and provided the best results. In the new approach *auto*OPS, beyond the six informative vectors calculated in the original OPS, two new vectors were used, being the univariate regression between each column of matrix **X** with **y** (URXY), and the vector of weights (WGHT) obtained by the NIPALS [22] (nonlinear iterative partial least squares) algorithm. The URXY vector contains the information about the regression coefficient obtained by each *j* column of **X** ($\mathbf{x_j}$) with **y** response and the difference between the measured and predicted **y**. This vector can be calculated as follows:

$$\begin{aligned} \hat{\mathbf{b}}_j &= (\mathbf{x}_j^t \mathbf{x}_j)^{-1} \mathbf{x}_j^t \mathbf{y} \\ \hat{\mathbf{y}}_j &= \mathbf{x}_j \hat{\mathbf{b}}_j \\ \mathbf{ey}_j &= \mathbf{y} - \hat{\mathbf{y}}_j \\ \mathbf{urxy}_j &= \hat{b}_j(2) \Big/ (\mathbf{ey}_j^t \mathbf{ey}_j) \end{aligned} \tag{3}$$

A univariate regression ($\mathbf{y} = \mathbf{xb}$) is performed using each predictor $\mathbf{x}_j$ and **y**. The regression coefficients **b** (a vector with two elements, being the intercept and slope, respectively) are obtained by classical least squares. The residue ($\mathbf{ey}_j$) between the predicted $\hat{\mathbf{y}}$ and **y** is calculated, and the URXY value of each variable ($\mathbf{urxy}_j$) is subsequently obtained by ration between the second element of **b** (slope) and the residual sum of squares. A high amount of URXY indicates that the corresponding variable should contain valuable information for the model as the residue for this variable presents a small value.

Besides the individual vectors (eight options, six from the original version of OPS and two news), their binary combinations (twenty-eight combinations) and the product of all individual vectors simultaneously (PRODALL) can also be used to search the best set of variables for prediction. All vectors are transformed into absolute values, and the infinite norm was applied. The norm was not applied on COR informative vector. The summary of all informative vectors is shown in Fig. 2.

The new approach performs the selection using one of the following vector input options:

1. Single vector, *i.e.*, only one of the thirty-seven vectors (eight individual, twenty-eight binary combinations and the PRODALL shown in Fig. 2 as the colored boxes).
2. Main vectors, *i.e.*, all eight individual vectors simultaneously (vectors placed on the dashed rectangle in the left side of Fig. 2).
3. Interaction vectors, *i.e.*, all twenty-eight binary combinations of two individual vectors (colored ones above the main diagonal line) plus PRODALL (the top row of Fig. 2) simultaneously.

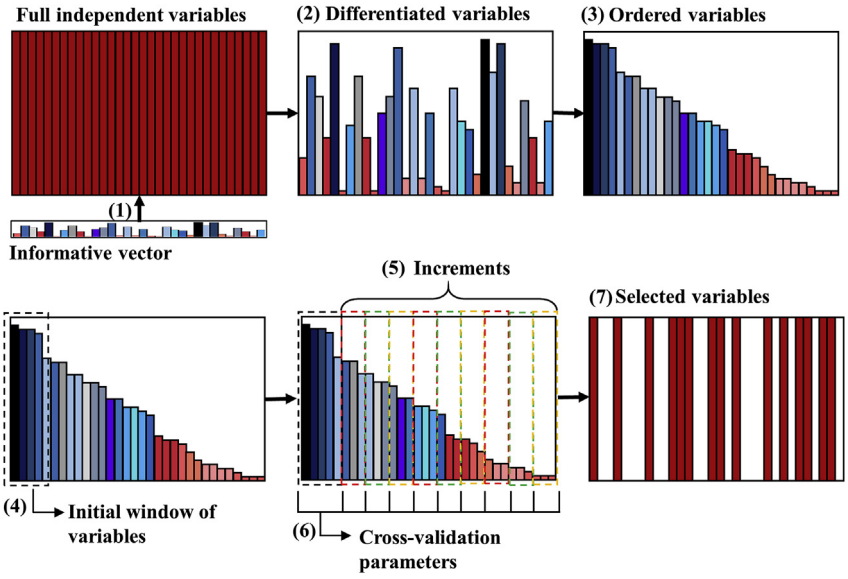


Fig. 1. General scheme of variable selection steps using OPS algorithm. (1) Informative vectors are obtained; (2) the data matrix is differentiated in according to the corresponding absolute values of the informative vector elements; (3) the differentiated variables are sorting in descending order; (4) an initial subset of variables (window) is defined to build and evaluate the first model; (5) this initial subset is extended by the addition of a fix number of variables (increment) over the window, and further increments are added until all, or some percentage of variables are taken into account; (6) cross-validation parameters are calculated for each model; (7) the variable subsets are compared using the quality parameters calculated during validations, and the best variables subset is defined. The different colors represent the weight of the variables and were randomly assigned. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

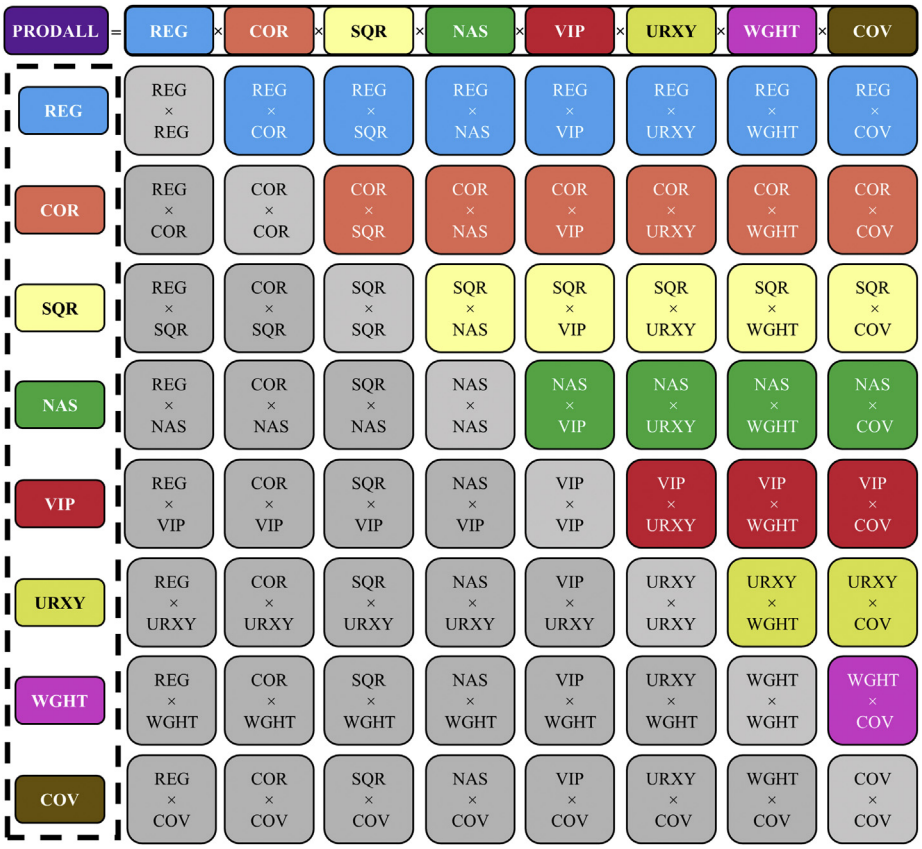


Fig. 2. Summary of all informative vector options to be used in new OPS algorithms for variable selection. REG: regression coefficients; COR: correlation between each column of matrix **X** with **y**; SQR: residual information of the reconstructed matrix with h latent variables; NAS: net analyte signal; VIP: variable importance on projection; URXY: univariate regression between each column of matrix **X** with **y**; WGHT: weights; COV: covariance procedures; PRODALL: the product of all single vectors simultaneously.

4. All vectors, *i.e.*, the main and interaction vectors options simultaneously.

In these options, when more than one vector is initialized in the OPS algorithm (options 2, 3, and 4), the correspondent vectors are used. Based on cross-validation parameters, the best vector is chosen to perform variable selection. The aim is to find the best variable subset to build predictive multivariate regression models. Thus, a selection criterion (*sc*) was defined to choose the vector, which is shown in equation (4).

$$sc_i = RMSECV_i / R_{CV\,i} \quad (4)$$

where *RMSECV* to R_{CV} ratio of each vector *i*, it is obtained and the vector that shows the lowest *sc* is chosen. Therefore, the set of variables selected is used to build the final model.

Particular attention should be given to the number of latent variables used in OPS algorithms. Firstly, the number of latent variables (*h*Mod) used to build the model was determined based on *RMSECV* values obtained by cross-validation procedure. In addition, some informative vectors, as REG, SQR, VIP, NAS, and WGHT, require a number of latent variables to be obtained. However, sometimes, *h*Mod is not able to generate an informative vector with quality for variable selection. Therefore, to find the best number of latent variables for OPS (*h*OPS), a study using the full dataset is performed by increasing the number of latent variables of the model, starting from the pre-determined *h*Mod, and carrying out the variable selection up to a given number of latent variables. Then, varying the *h*Mod value, the *h*OPS is chosen based on the smallest *RMSECV* value. Hence, two types of latent variables are employed in the OPS algorithm, one representing the number of latent variables for model building (*h*Mod) and the other employed to generate the REG, SQR, VIP, NAS, and WGHT informative vector in OPS method (*h*OPS).

***auto*OPS algorithm**

The algorithm *auto*OPS consists of the following steps.

choose *k* informative vectors to be studied;
for *i* = 1 to *k*
  calculate the *i* informative vector;
  run the **core of OPS algorithm** using the *i* informative vector;
  store the *i* set of selected variables;
  apply equation 4;
  store sc_i;
end
the lowest *sc* indicates the informative vector that selects the best set of variables.

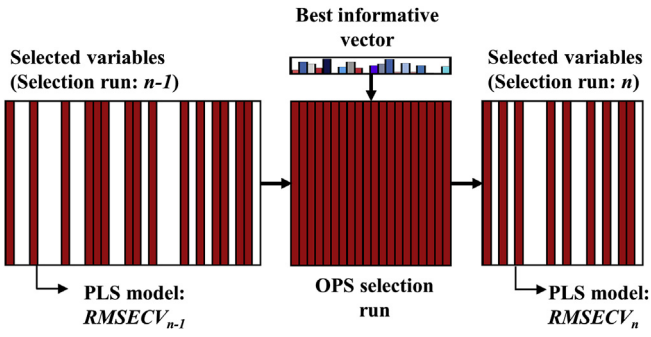


Fig. 3. Scheme of the *feed*OPS algorithm where the pre-selected variables return to a new selection run until specific criteria are achieved. *RMSECV*: root mean square error of cross-validation, *n*: loop counter.

2.2.2. Feedback OPS (*feed*OPS)

The *feed*OPS has the purpose of applying feedback in the OPS algorithm, wherein pre-selected variables can return to a new selection run until specific criteria are achieved (Fig. 3). This new strategy works in a loop constrained by one or more rule; when one of them is attained the loop stops.

In *feed*OPS, *RMSECV* was taken as a reference parameter. Convergence is achieved when in two consecutive selection runs, the relative difference (*rd*) (equation (5)) is less than an *rd* value defined or when the *RMSECV* value increases instead of decrease. Thus, the previous loop is taken as the best selection. Besides that, the maximum number of selection runs can be defined. Therefore, the last loop is considered the best selection. In each selection run, the *auto*OPS is applied to provide the best informative vector.

$$rd = RMSECV_n - RMSECV_{n-1} / RMSECV_n \quad (5)$$

where *n* is the number of selection runs.

***feed*OPS algorithm**

The algorithm *feed*OPS consists of the following steps.

define convergence criteria (*rd* and *l* = max number of loops);
n = 1;
while $rd_{calc} > rd$ *or* $n < l$
  run ***auto*OPS** algorithm;
  store $RMSECV_n$ of the selected informative vector;
  % The *RMSECV* of full **X** matrix was used to compare with *RMSECV* in the first loop.
  calculate rd_{calc} (Equation 5);
    if $RMSECV_n - RMSECV_{n-1} > 0$
      break.
    End
  $n = n + 1$;
end
obtaining the final set of selected variables.

2.2.3. OPS intervals (*i*OPS)

In this new approach, the OPS is applied in subdivisions of the full array. This search strategy is called OPS intervals (*i*OPS) and is showed in Fig. 4. Firstly, the number of variables in each interval is defined. The entire array is subdivided, and the variable selection is performed. In each interval, the *auto*OPS or the *feed*OPS is applied. This step is mainly used to reduce the number of variables in each interval. The intervals comprising only the selected variables will be merged into a new matrix and a new variable selection will take place considering predetermined window and increment. The *auto*OPS (or *feed*OPS) is applied in the new array, and the best set of variables is reached.

***i*OPS algorithm**

The algorithm *i*OPS consists of the following steps.

define the number of variables in each interval (at least 50);
apply ***auto*OPS** or ***feed*OPS** algorithm in each interval;
find the best set of variables for each interval;
create a new array containing the selected variables in each interval;
apply ***auto*OPS** or ***feed*OPS** algorithm in the new array;
obtaining the final set of selected variables.
% We nickname this algorithm of *crème de la crème*.

Once all new approaches have been explained, Fig. 5 shows a chart that represents the options to apply the new OPS algorithms and summarizes the names of different options of OPS algorithm application.

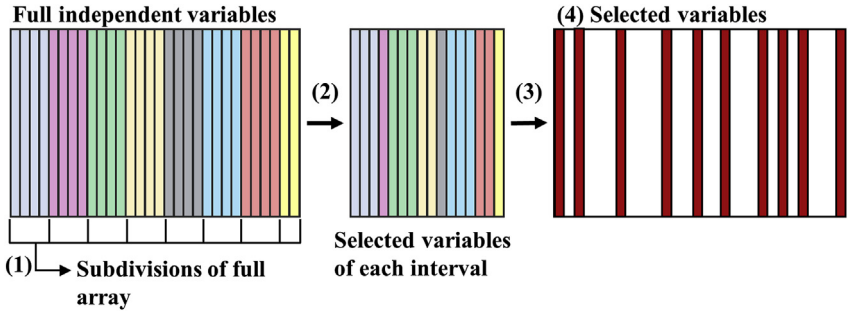


Fig. 4. Scheme of *i*OPS algorithm where the variable selection is performed in each interval; subsequently, these selected variables are ordered, and a new selection is performed to find the best set of variables. (1) The array is subdivided; (2) *auto*OPS or *feed*OPS is applied in each interval; and a new matrix is created with the selected variables in each interval (3) *auto*OPS or *feed*OPS is applied in the new matrix; (4) the best set of variables is reached. *RMSECV*: root mean square error of cross-validation.

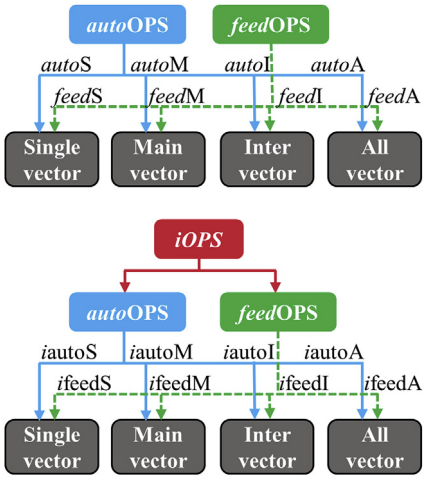


Fig. 5. Scheme of the new OPS algorithms (*auto*OPS, *feed*OPS, and *i*OPS) combined with the four vector options. Names of OPS options derived by the combination of new OPS approaches with all vector options are arranged in the lines for the arrows.

Each new approach combined with each vector option generates sixteen ways to apply the OPS. For example, using *auto*OPS with a single vector, the option is called autoS. The *feed*OPS with the main vector option is named *feed*M. The application of *i*OPS with *auto*OPS and all vectors option is called *i*autoA.

3. Experimental

The OPS algorithms were written in function *.m* to MATLAB 2019a (Math Works, Natick, USA) and are fully authorial (available at www.deq.ufv.br/chemometrics). This section is about the validation of the new approaches of OPS (*auto*OPS, *feed*OPS, and *i*OPS). Initially, the new approaches were applied in the six datasets used in the original OPS paper [17] to compare the prediction performance with the new OPS algorithms. Then, twelve new datasets, comprising different types of data, were used to compare the performance of the new OPS approaches with other variable selection algorithms used by chemometrists, *i.e.*, genetic algorithm [10] (GA), intervals successive projections algorithm for PLS [8] (*i*SPA), and recursive weighted partial least squares [11] (rPLS). All calculations were performed in MATLAB environment.

3.1. Modeling

The six datasets studied in the original OPS paper were used to build PLS models using variables selected by the new OPS approaches (*auto*OPS, *feed*OPS, and *i*OPS). They were tested using all vectors, the most complete OPS option. The calibration and prediction sets were the same used in the original paper [17]. In addition, the method leave-*N*-out cross-validation was applied, where *N* was set as 10% of the total sample number in the training set. The range calibrated, the size of calibration and prediction sets, and the pre-treatment performed in each dataset can be found elsewhere [17]. The new OPS algorithms were applied using different windows and increments of variables for each dataset, and 100% of variables were tested. In *feed*OPS, a minimum difference of 2% between two consecutive *RMSECVs* and a maximum number of ten loops were set as convergence criteria. In *i*OPS, when the *feed*OPS option was used to run the selection, the convergence criteria were the same as used in *feed*OPS. The full matrix was divided into intervals of 10% of the its size, limited in at least fifty variables. Additionally, *h*OPS was calculated for each interval in *i*OPS.

Twelve new datasets were used to build PLS models using all variables of the **X** matrix (full models) and selected variables using the new OPS, GA, *i*SPA, and rPLS algorithms. All three new approaches of OPS were tested for each dataset using all vectors, the most complete OPS option. A total of thirty-seven informative vectors were evaluated. In order to compare the OPS with the other three variable selection algorithms, the new strategy that provided the best OPS model for each dataset was chosen between the three new strategies available using the all vectors option. This choice was made based on the parameters described in the section *"2.1 Model evaluation"*.

The full datasets were split into calibration and prediction sets using Kennard-Stone algorithm [23]. Table 1 shows information about the property and range calibrated, the size of calibration and prediction sets, and the pre-treatment performed in each dataset. The **y** variable was mean centered for all the properties and different pre-treatments and their combinations were studied for each full matrix **X**. The pre-treatments tested were mean center, autoscale, smoothing, first and second derivative, multiplicative scatter correction (MSC), normalize, baseline, and standard normal variate (SNV). The combinations of two, three and four pre-treatments were also studied. The pre-treatments that presented a model with the lowest *RMSECV* value was chosen for each dataset. All algorithms performed the variable selection with datasets equally pre-treated.

The new OPS algorithms were applied using the window of 10 and increments of 5 variables, 100% of variables were tested, random cross-validation was applied, where splits were set at 10% of **X** matrix rows. Only QSAR models were built using leave-one-out cross-validation. In *feed*OPS, as convergence criteria, 2% as the minimum difference between two consecutive *RMSECVs* and ten as the maximum number of loops. In *i*OPS, when the option to run the

Table 1
Information about datasets used to perform validation of new strategies of OPS algorithm.

Dataset	Voltammetry	Fluorescence	MS	Raman	NIR	UV
Property	Ascorbic Acid	Catechol	Ethanol	Iodine Value	Lignin	Phorbol esters
Range	21.5–100.0	1.0–8.0	12.79 - 15.09	52.0–69.0	18.04–28.36	0.72 - 3.58
Unit	μmol L^{-1}	μmol L^{-1}	vol (%)	g I_2 100 g fat $^{-1}$	% (w/w)	mg g^{-1}
Size	22 × 525 (Cal)	28 × 221 (Cal)	34 × 200 (Cal)	75 × 5667 (Cal)	216 × 1038 (Cal)	106 × 656 (Cal)
	10 × 525 (Pred)	10 × 221 (Pred)	10 × 200 (Pred)	30 × 5667 (Pred)	40 × 1038 (Pred)	32 × 656 (Pred)
Column-wise pre-treatment	Mean Center	Mean Center	Mean Center	Auto	Mean Center	Mean Center
Row-wise pre-treatment	Baseline	None	None	SNV	Baseline + 2nd Deriv (3) + MSC	1st Deriv (15)

Dataset	NMR	GC	Vis/NIR	XRF	QSAR	MS
Property	Pentanol	Overall Quality	Xylene	Ni	MMP-2 pIC_{50}	Peroxide Value
Range	0–100	1.13 - 4.50	4.00–15.01	0.062 - 15.890	6.25 - 8.28	1.05–33.62
Unit	%	–	weight percent (%)	%	–	meq Kg^{-1}
Size	181 × 2334 (Cal)	119 × 2294 (Cal)	25 × 316 (Cal)	12 × 261 (Cal)	25 × 439 (Cal)	20 × 106 (Cal)
	50 × 2334 (Pred)	40 × 2294 (Pred)	5 × 316 (Pred)	3 × 261 (Pred)	6 × 439 (Pred)	5 × 106 (Pred)
Column-wise pre-treatment	Mean Center	Auto	Mean Center	Mean Center	Auto	Mean Center
Row-wise pre-treatment	None	MSC + 1st Deriv (19) + Norm	1st Deriv (3)	None	None	None

Cal: calibration set; Pred: prediction set; Auto: autoscaling; SNV: standard normal variate; 1st or 2nd Deriv: type of derivative with the window between parenthesis; MSC: multiplicative scatter correction; Norm: normalize to unit area; MMP-2 pIC_{50}: Half-maximal inhibitory concentration negative logarithm of type 2 matrix metalloproteinases.

selection using *feed*OPS was used, the convergence criteria were the same as those used in *feed*OPS. The full matrix was divided into intervals of 10% of the its size, limited in at least fifty variables. Additionally, *h*OPS was calculated for each interval in *i*OPS. Random cross-validation was applied to build full models, where splits were set at 10% of **X** matrix rows. Although the OPS algorithms require several parameters to optimize the variable selection, only **X** and **y** are mandatory inputs for all new OPS algorithms. It is possible to perform the variable selection using some default parameters and automatic choices, but the result may not be the optimal one.

The GA is a method used for solving optimization problems based on natural selection processes and genetics that mimic biological evolution [10]. Initially, a starting population is constituted for a series of different individuals and the number of individuals is defined as the population with a defined size. These individuals are analyzed and crossed according to the parameters introduced into the algorithm. At the end of the process, a vector consisting of zeros and ones indicates whether variables should be included (1) or not (0). The parameters used to perform the variable selection were optimized (results not shown) and the following conditions were used: a population of 52, a maximum generation of 300, a mutation rate of 0.008, a window width of 1, convergence of 80, 50 terms included at initiation, cross-over rule of 2, random cross-validation was applied with the number of subsets to divide data into for cross-validation of 10, cross-validation iteration of 1 and 3 replicate runs. The GA variable selection was performed using the *.m* functions from and PLS-Toolbox 8.2 (Eigenvector Research, Inc. Wenatchee, USA).

The *i*SPA combines the noise-reduction properties of PLS with the possibility of discarding non-informative variables in SPA-MLR algorithm [24]. In this method, it is assumed that the variables have been divided into non-overlapping intervals, usually, with the same length. First, the columns of **X** are partitioned according to the intervals of variables previously defined and subjected to a sequence of projection operations that result in the creation of chains. Second, PLS models are built using leave-one-out cross-validation for each combination of intervals, and the best combination of intervals is then chosen by the smallest *RMSECV* [8]. The *i*SPA algorithm was applied dividing the spectra into 20 intervals and the maximum number of intervals was selected as 20. This algorithm for MATLAB was provided by the authors.

The rPLS method uses the dependent variable **y** to guide the variable weighting recursively. This method iteratively reweights the variables using the regression vector calculated by PLS [11]. Random cross-validation with splits of 10 and the number of iterations infinite were the default settings used in variable selection. This algorithm for MATLAB was retrieved from www.models.life.ku.dk/algorithms.

The variable selection was carried out in triplicate on the calibration set. Thus, three models were built for OPS, GA, *i*SPA, and rPLS. Each model was applied on the prediction set, and the calculated *RMSEP* values were pairwise compared by Tukey test. Differences with $p < 0.05$ were considered significant.

3.2. Datasets

The six datasets used in the original OPS paper were near-infrared spectroscopy (NIR), fluorescence spectroscopy, a simulated dataset, Raman spectroscopy, gas chromatography (GC), and quantitative structure-activity relationship (QSAR) data.

For NIR spectra of diesel samples [25], the following physical properties were modeled: boiling point at 50% recovery, cetane number, density, freezing temperature of the fuel, total aromatics, and viscosity. Mixtures of standard solutions of six different analytes were used in fluorescence [26]: catechol, hydroquinone, indole, resorcinol, L-tryptophane, and DL-tyrosine. The simulated dataset consisted of twenty mixtures simulated by using ultraviolet-type spectra from four analytes and their respective concentrations randomly generated. For Raman spectra of *Escitalopram*® tablets [27], the dependent variable referred to the amount of active substance in the tablets. GC dataset of fuel samples (Pirouette® software – Infometrix, Inc) is formed by thirty-five independent variables consisting of the chromatograms peak areas and three dependent variables comprising the following physical properties: flash point, freeze point, and specific gravity. The QSAR dataset [28] consists of fourteen molecular descriptors for forty-eight HIV-1 protease inhibitors and the dependent variable was the *in vitro* inhibition activity.

For NIR dataset, the new OPS algorithms were applied using window of five and increments of two variables (window 5, increment 2), Raman dataset (window 10, increment 5), fluorescence dataset (window 5, increment 2), GC dataset (window 2, increment 1), QSAR dataset (window 2, increment 1), and simulated dataset (window 5, increment 1).

The twelve new datasets used to compare full models and selected variables using the new OPS, GA, *i*SPA, and rPLS algorithms are described below.

Voltammetry: This dataset was presented by Teófilo [29] and consists of a mixture of standard solutions of three biological interest compounds (dopamine, uric acid, and ascorbic acid). Square wave voltammetry (SWV) with boron-doped diamond electrode was employed to quantify these analytes. The potential range investigated was from 0.01 to 1.5 V.

Fluorescence Spectroscopy: This dataset was presented by Teófilo [29] and consists of a mixture of six standard phenol solutions (hydroquinone, guaiacol, p-cresol, m-cresol, catechol, and phenol). The excitation wavelength was fixed at 275 nm, and the emission wavelength was scanned from 270 to 380 nm.

Mass Spectrometry (MS): (1) The first mass dataset was obtained from http://www.models.ku.dk/datasets and was presented by Skov et al. [30]. A mass spectrum was obtained for each sample of wine in the range of 5–204 *m*/*z* using the electron ionization mode at 70 eV. Fourteen properties were evaluated in the original work [30], but only ethanol was calibrated in this work. (2) The second contains mass spectra collected from the headspace of butter samples that had been artificially aged to produce various levels of spoilage. The reference values for rancidity are peroxide values. The data were supplied by Leatherhead, UK, and obtained from Pirouette® (Infometrix, Inc) software. The mass spectrum was obtained in the range from 44 to 149 *m*/*z*.

Raman Spectroscopy: This dataset was presented by Lyndgaard et al. [31], and it is available at http://www.models.ku.dk/datasets. The samples were 16 pork carcasses taken from the daily production stock of a slaughterhouse for determining the fatty acid composition of pork backfat as a function of the iodine depth profile. The Raman spectra were acquired using a total of 16 accumulations of 1 s exposure and were stored as Raman shifts in the range 1800–200 cm^{-1}.

Near Infrared Spectroscopy (NIR): This dataset is composed of sugarcane leaf NIR spectra that was used to predict the lignin content of sugarcane stalks, and it was provided by Assis et al. [16]. The NIR spectra were acquired from 10000 to 4000 cm^{-1}.

Ultraviolet Spectroscopy (UV): This dataset was obtained from Roque et al. [32], and consists of UV spectra of *Jatropha curcas* extracts in the range of 210–350 nm. Phorbol esters content was the modeled property.

Nuclear Magnetic Resonance Spectroscopy (NMR): This dataset is available at http://www.models.ku.dk/datasets, and was presented by Winning et al. [33]. It consists of NMR spectra of a designed set of 231 simple alcohol mixture (propanol, butanol, and pentanol), and each spectrum was acquired in the range of 0.64–3.84 ppm.

Gas Chromatography (GC): Chromatographic profiles of volatile roasted coffee compounds provided by Ribeiro et al. [34] to predict scores of coffee beverage overall quality. The chromatograms dataset used in this work comprises of the range from 1.25 to 21.83 min of retention time.

Visible and Near Infrared Spectroscopy (Vis/NIR): The data were obtained from Pirouette® (Infometrix, Inc) software. This dataset contains spectra of hydrocarbon mixtures (heptane, isooctane, toluene, xylene, and decane) from two different diode array spectrometers. Absorbances from 470 to 1100 nm were collected.

X-Ray Fluorescence Spectrometry (XRF): X-ray fluorescence spectra were obtained from Pirouette® (Infometrix, Inc) software. Wang et al. [35] presented this dataset. Spectra of nickel alloys plus elemental concentrations in the alloys were measured from $2\theta = 44°–70°$. The nickel concentrations were determined by wet chemistry.

Quantitative structure-activity relationship (QSAR): De Melo [36] presented this dataset of matrix metalloproteinases type 2 (MMP-2) with 31 cinnamoyl pyrrolidine derivatives, where 439 molecular descriptors were obtained.

4. Results and discussion

4.1. Algorithms

The finest OPS model of each dataset is presented in Table 2; regarding this, no tendency is observed. In order to present the detailed results obtained using each one of the new approaches (*auto*OPS, *feed*OPS, and *i*OPS), it was selected a dataset that shows the best result using each of one them. For *auto*OPS was chosen XRF; for *feed*OPS, voltammetry; and for *i*OPS, GC dataset.

The best OPS model obtained for XRF dataset was using the *auto*OPS algorithm. In Fig. 6A1 the *sc* value obtained for each of the thirty-eight vectors is shown. It is noticeable that the vector SQR × URXY showed the lowest value (dashed red line). Additionally, not all vectors were able to improve prediction quality regarding the full model (solid red line). Some of them presented consistent improvement of model prediction. These results indicated that vector combinations performed better than individual vectors. This conclusion cannot be generalized for all datasets, but it is enough to show the importance of combinations. Fig. 6A2 shows the OPS plot for recognition of the best set of variables. It is a detail of the best result shown in Fig. 6A1 for SQR × URXY vector. This plot displays an increase of the variables number, related to the addition of new variables to the first variables window, and a decrease in *RMSECV* and increase in R_{CV}. This type of plot is very useful to visualize the variable selection and help in choosing the best set. The ideal subset of variables is found when the ration of *RMSECV* and R_{CV} is minimum (equation (4)). Sometimes, an increase of the variables number is accomplished by a small variation in *RMSECV* value. In this way, the subset of selected variables may not be ideal, because instead the *RMSECV* increase with the addition of new noninformative variables, this parameter remained almost constant. Then, an overestimated subset of variables can be erroneously selected.

In Fig. 6A2, with few variables, the OPS model performance is worse than the full model (solid red line), and as the number of variables increases, the *RMSECV* decreases and R_{CV} increases until the optimal number of variables is reached (dashed red line). After that, the *RMSECV* value increases and R_{CV} decreases with the inclusion of noninformative variables.

Voltammetry dataset was used to present the result provided by *feed*OPS (Fig. 6B). For this dataset, eight selection runs were

Table 2
Best OPS model, informative vector and the number of latent variables used to perform the selection (*h*OPS) obtained for each dataset.

Dataset	Property	OPS[a] model	Vector
Voltammetry	Ascorbic Acid	*feed*	NAS (19)
Fluorescence	Catechol	*i*auto	SQR (5) × COV
MS	Ethanol	*i*auto	REG (18) × COR
Raman	Iodine Value	*i*auto	URXY
NIR	Lignin	*feed*	REG (19)
UV	Phorbol esters	*auto*	REG (9) × COR
NMR	Pentanol	*auto*	SQR (7)
GC	Overall Quality	*i*feed	REG (16)
Vis/NIR	Xylene	*i*auto	REG (10)
XRF	Ni	*auto*	SQR (11) × URXY
QSAR	MMP-2 pIC_{50}	*i*auto	REG (19)
MS	Peroxide value	*feed*	REG (9) × NAS (9)

MMP-2 pIC_{50}: Half-maximal inhibitory concentration negative logarithm of type 2 matrix metalloproteinases.
[a] Best result. *h*OPS values are presented between parentheses.

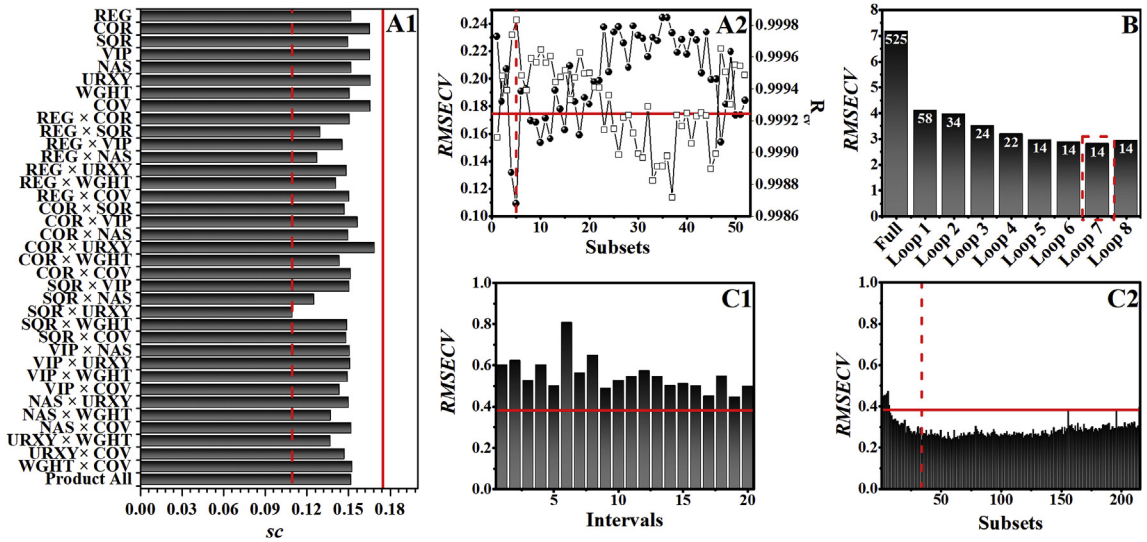


Fig. 6. (A1) Minimum *sc* values obtained for all informative vectors used in the *auto*OPS algorithm for the XRF dataset. The selected vector was detached as dashed red line. (A2) Detailed results of variable subsets obtained by the *auto*OPS algorithm using the SQR × URXY vector and your respective *RMSECV* and R_{CV} values for the XRF dataset. The dashed red line means the selected subset. Black spheres are *RMSECV* values showed in the left axis, and black squares are R_{CV} values showed in the right axis. (B) *Feed*OPS typical plot (number of selection runs and *RMSECV* values) for the voltammetry dataset. The dashed red line indicates the loop where the convergence was attained, and the set of selected variables was obtained. (C1) *i*OPS plot for the GC dataset with *RMSECV* values obtained by PLS models built using the selected variables in each interval of the full array (first variable selection step). (C2) *i*OPS plot for the GC dataset with *RMSECV* values obtained by PLS models built using the subsets of the merged matrix built from selected variables in each interval. A dashed red line shows the subset were the best set of variables were selected. A solid red line represents the *RMSECV* of full model in each subplot of *i*OPS. *sc*: selection criterion; *RMSECV*: root mean square error of cross-validation. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

performed. The number of informative variables selected and the *RMSECV* value decreases in each loop until one of the criteria is achieved. The maximum of ten loops and the relative difference between two consecutive *RMSECV* values lower than 2% were set as constraints. Convergence was reached with seven loops, where the difference between *RMSECV* of the seventh and eighth loop was smaller than 2%. Therefore, the set of variables was selected by the NAS vector in loop 7 (dashed red line). As a result, a meaningful reduction of the number of variables (from 525 to 14) and improved model performance parameters was reached.

Regarding *i*OPS, the GC dataset results were chosen to be presented. Fig. 6C1 shows the *RMSECV* obtained by each interval of the full array using variable selection. In this example, the selection of variables in each subdivision was not able to enhance the predictive capacity of PLS models. So, in the last step of *i*OPS, the selected variables in each interval were merged into a new matrix. A new variable selection took place considering window and increment of variables, and the best set of variables was obtained. In Fig. 6C2, the result of this last step is shown. *RMSECV* value decreases with the addition of the increments of variables until finding the finest subset of variables (dashed red line) and increases with the inclusion of noninformative variables in the model.

4.2. Modeling

The three new OPS approaches provide four different ways to apply OPS since the *i*OPS can be applied using *auto*OPS or *feed*OPS. The PLS models built using the selected variables for the six datasets used in the original OPS paper are shown in Fig. 7.

The number of selected variables is placed in the left-side axis and is represented by bars. The informative vectors are placed inside the bars. The *RMSEP* values are located in the right-side axis and are represented by the black spheres. The *h*Mod for each model is shown below the spheres. In general, the prediction performance of the new approaches outperformed the first version of OPS (OPS*v*1) algorithm in all datasets.

The *h*Mod used in the original paper was taken as reference to run the new OPS algorithms, and for most of the cases, this number was kept and for the others a smaller *h*Mod was chosen automatically. The informative vector used for all datasets in the original paper was REG or some combination with it. With the new approaches, different informative vectors were chosen, including the PRODALL vector for three properties (Fig. 7B, H, and 7J), proving the importance of new vectors and their combinations to select more predictive variables. For some datasets, the number of selected variables were slightly higher than the ones selected in the original paper, but the prediction was better.

In Fig. 7R–U, it was not possible to apply the *i*OPS algorithm, since the GC (Fig. 7R–T) and QSAR (Fig. 7U) datasets have thirty-five and fourteen variables, respectively. For GC (specific gravity – Fig. 7T), the *RMSEP* value was the same for OPS*v*1, *auto,* and *feed* approaches, but the number of selected variables and the informative vector were not the same. For QSAR (Fig. 7U), the *RMSEP* value, the number of selected variables and the informative vector were the same for OPS*v*1, *auto* and *feed* approaches. Both datasets have a few number of variables, and it is possible that this fact influenced these results.

The PLS models obtained using the twelve datasets previously presented are described. The mean results of the performance parameters of PLS models are shown in Table 3. Concerning the *RMSECV* values, the OPS models were smaller than full models for all datasets. Regarding the cross-validation results, OPS was the most predictive for voltammetry and NIR; GA was the most predictive for fluorescence, MS (ethanol), NMR, Vis/NIR, XRF, QSAR, and MS (peroxide value); rPLS was the most predictive for GC and OPS and rPLS presented similar results for Raman and UV.

The *RMSEP* values are followed by superscript letters indicating the Tukey test results. Different letters mean that the models are significantly different. For MS (ethanol) and UV, all five PLS models (full and four variable selection models) differ significantly between then, being the OPS, the most predictive model. Only for fluorescence dataset, OPS and GA models were statistically

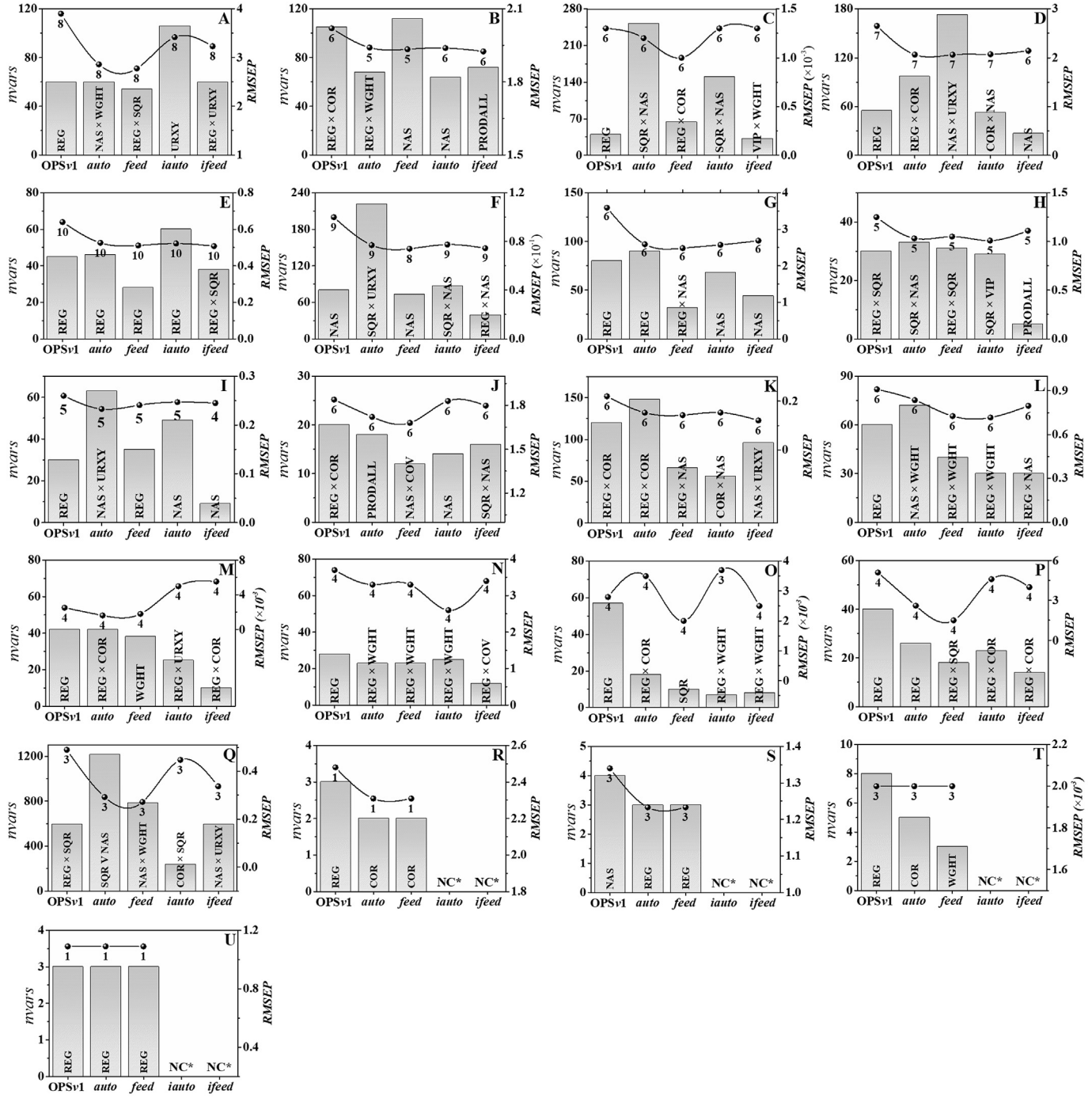


Fig. 7. Number of selected variables placed in the left-side axis (light gray bar) and *RMSEP* values placed in the right-side axis (black spheres). The informative vector and the *h*Mod are placed in bars and black spheres, respectively. The values outside the bar are the percent of decrease (negative values) or increase (positive values) of *RMSEP* values or the number of variables regarding full model represented by the horizontal line on zero. (A) NIR – boiling point at 50% recovery, (B) NIR – cetane number, (C) NIR – density, (D) NIR – freezing temperature of the fuel, (E) NIR – total aromatics, (F) UV–viscosity, (G) Fluorescence – catechol, (H) Fluorescence – hydroquinone, (I) Fluorescence – indole, (J) Fluorescence – resorcinol, (K) Fluorescence – L-tryptophane, (L) Fluorescence – DL-tyrosine, (M) Simulated – analyte 01, (N) Simulated – analyte 02, (O) Simulated 03 – analyte 03, (P) Simulated – analyte 04, (Q) Raman – active substance in *Escitalopram* ® tablets, (R) GC – flash point, (S) GC – freeze point, (T) GC – specific gravity, and (U) QSAR – *in vitro* inhibition activity. *RMSEP*: root mean square error of prediction, *h*Mod: number of latent variables of the model. *NC: not calculated.

Table 3
Performance parameters of PLS models obtained using all variables and those selected by OPS, GA, *i*SPA, and rPLS algorithms.

	Voltammetry - Ascorbic Acid					Fluorescence - Catechol					MS - Ethanol				
	Full	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**
nvars	525	14	74–80	52–79	7	221	30	32–35	33–166	28	112†	10	13–22	79	7
*h*Mod			6					3					6		
RMSECV	8.2	3.4	4.8	10.0	3.7	0.6	0.4	0.3	0.6	0.5	0.46	0.36	0.29	0.48	0.35
R_{CV}	0.943	0.991	0.982	0.920	0.989	0.989	0.994	0.997	0.989	0.993	0.160	0.666	0.686	0.100	0.517
RMSEP	7.6^{c}	4.0^{a}	6.1^{b}	9.2^{d}	8.4^{c}	0.8^{b}	0.7^{a}	0.6^{a}	0.9^{b}	0.8^{b}	0.68^{e}	0.36^{a}	0.42^{b}	0.48^{c}	0.56^{d}
R_P	0.969	0.995	0.983	0.962	0.963	0.989	0.994	0.994	0.988	0.990	0.220	0.865	0.802	0.783	0.551
	Raman - Iodine Value					**NIR - Lignin**					**UV - Phorbol Esters**				
	Full	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**
nvars	5667	25	2132–2224	284–3116	84	1038	135	293–358	518–882	149	656	50	148–164	66	51
*h*Mod			2					10					8		
RMSECV	2.1	2.0	2.1	2.5	2.0	1.42	0.59	0.81	1.53	0.88	0.29	0.23	0.30	0.25	0.23
R_{CV}	0.763	0.801	0.785	0.644	0.806	0.733	0.959	0.921	0.698	0.906	0.855	0.912	0.849	0.890	0.909
RMSEP	2.0^{a}	1.9^{a}	2.0^{a}	3.2^{b}	2.4^{ab}	0.89^{c}	0.67^{a}	0.75^{b}	1.02^{d}	0.80^{b}	0.27^{c}	0.24^{a}	0.28^{d}	0.26^{b}	0.32^{e}
R_P	0.880	0.908	0.892	0.720	0.848	0.928	0.960	0.948	0.902	0.941	0.947	0.958	0.943	0.952	0.924
	NMR - Pentanol					**GC - Overall Quality**					**Vis/NIR - Xylene**				
	Full	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**
nvars	2334	198	985–1167	2101	199	2294	185	762–772	115–1492	736	316	45	59–63	175–285	27
*h*Mod			4					5					5		
RMSECV	0.68	0.65	0.61	0.68	0.69	0.37	0.28	0.20	0.49	0.19	0.21	0.14	0.06	0.19	0.15
R_{CV}	0.999	0.999	0.999	0.999	0.999	0.918	0.953	0.977	0.847	0.979	0.999	0.999	0.999	0.999	0.999
RMSEP	2.45^{b}	2.32^{a}	2.45^{b}	2.45^{b}	2.44^{b}	0.38^{b}	0.36^{a}	0.49^{d}	0.48^{d}	0.40^{c}	0.16^{c}	0.12^{a}	0.13^{b}	0.18^{d}	0.14^{b}
R_P	0.995	0.996	0.996	0.995	0.996	0.925	0.937	0.887	0.877	0.920	0.999	0.999	0.999	0.999	0.999
	XRF - Ni					**QSAR - MMP-2 pIC_{50}**					**MS - Peroxide Value**				
	Full	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**	**Full**	**OPS**	**GA**	***i*SPA**	**rPLS**
nvars	261	30	79–86	79–105	248	439	195	48–117	373–439	26	106	40	26–53	69–91	104
*h*Mod			4					3					6		
RMSECV	0.173	0.149	0.117	0.246	0.172	0.45	0.40	0.31	0.45	0.33	0.44	0.39	0.28	0.48	0.49
R_{CV}	0.999	0.999	0.999	0.999	0.999	0.647	0.731	0.834	0.669	0.829	0.999	0.999	0.999	0.999	0.999
RMSEP	0.185^{b}	0.139^{a}	0.204^{c}	0.208^{c}	0.185^{b}	0.51^{c}	0.37^{a}	0.51^{c}	0.51^{c}	0.48^{b}	0.39^{b}	0.32^{a}	0.52^{c}	0.39^{b}	0.39^{b}
R_P	0.999	0.999	0.999	0.999	0.999	0.204	0.564	0.255	0.206	0.263	0.999	0.999	0.999	0.999	0.999

nvars: number of variables; *h*Mod: number of latent variables of the model; *RMSECV*: root mean square error of cross-validation; R_{CV}: correlation coefficient of cross-validation; *RMSEP*: root mean square error of prediction, R_P: correlation coefficient of prediction. Units: Voltammetry – ascorbic acid (μmol L^{-1}), Fluorescence – catechol (μmol L^{-1}), MS – ethanol (vol - %), Raman – iodine value (g I_2 100 g fat^{-1}), NIR – lignin (% w/w), UV – phorbol esters (mg g^{-1}), NMR – pentanol (%), GC – overall quality (not applied), Vis/NIR – xylene (weight percent - %), XRF – Ni (%), (K) QSAR – MMP-2 pIC_{50} (not applied), and MS – peroxide value (meq Kg^{-1}). MMP-2 pIC_{50}: Half-maximal inhibitory concentration negative logarithm of type 2 matrix metalloproteinases. †Full dataset after removal of variables that not present variance.

equivalent. Furthermore, *RMSEP* values of full, OPS, GA, and rPLS do not differ significantly in Raman dataset.

In some cases, the least predictive model was not the full. For example, in voltammetry, NIR, GC, Vis/NIR, and XRF, *i*SPA models have the highest *RMSEP* value. Also, the *RMSEP* was higher than full in rPLS for UV and GC datasets, and in GA for UV, GC, XRF, and MS (peroxide value).

In QSAR models, it is recommended to verify the possibility of chance correlation using the y-randomization test [36], where the y response was randomized five hundred times. Then, OPS and rPLS models do not show chance correlation. However, GA, *i*SPA, and full models presented chance correlation (results not shown).

Special attention should be given to MS (ethanol) dataset. Results for *i*SPA and rPLS models were not obtained using as the start point of selection the matrix with all two hundred variables because those algorithms were not able to perform the variable selection. In *i*SPA, the algorithm started the selection but paralyzed and did not finish the selection. For rPLS, the selection has not even begun; an error appears indicating that the **X** data, not present variance. From these, it can be presumed that both methods, *i*SPA, and rPLS, show the limitation to select variables in datasets like GC and MS, where the baseline does not have variance, and a peak suddenly arises. Thought, in this work, these algorithms perform the selection in a GC and another MS dataset. This can be explained by the presence of noise in the baseline, which confers variance to the entire set of variables. So, to perform the selection with *i*SPA and rPLS algorithms, the variables that not present variance were eliminated and the full matrix now consists of 112 variables. This new **X** was used to build the full model and to start the variable selection.

In Fig. 8 is shown a bar plot with the *RMSEP* values and the number of selected variables of each variable selection algorithm for each dataset. These results are shown in percentage of increase or decrease regarding the full models. For OPS models, the percentage of decrease in *RMSEP* values ranged from 5 to 47% and the percentage of decrease in number of variables ranged from 56 to 99%. For GA and *i*SPA models, an error bar is shown because these two variable selection methods do not show stability when executing the selection more than once. *i*SPA was reproducible only in MS (ethanol), UV, and NMR datasets. OPS and rPLS do not show any variance of selected variables in all twelve datasets, indicating that they are highly reproducible.

Fig. 9 presents the selected variables for all datasets. Each subplot shows the selected variables for each algorithm (OPS, GA, *i*SPA, and rPLS), and the frequency at which a variable was selected. The variable selection by OPS occurred, in general, where well-defined regions are observed, *i.e.*, peaks with rather clear physical meaning were selected. In some datasets, baseline regions are also selected, but despite this, the models are predictive. Besides that, the OPS selected few variables in all datasets, while the other methods

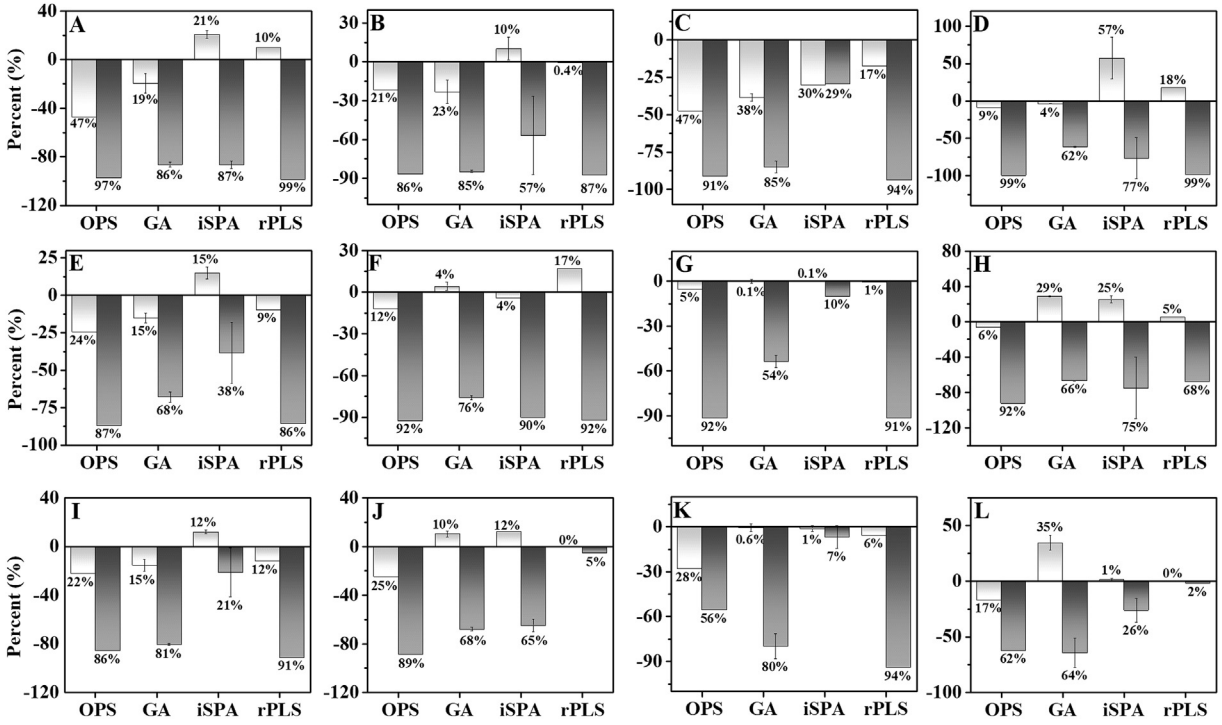


Fig. 8. *RMSEP* values of variable selection models (light gray bar) and the number of selected variables (dark gray bar). The values outside the bar are the percent of decrease (negative values) or increase (positive values) of *RMSEP* values or the number of variables regarding full model represented by the horizontal line on zero. (A) Voltammetry – ascorbic acid (μmol L^{-1}), (B) Fluorescence – catechol (μmol L^{-1}), (C) MS – ethanol (vol - %), (D) Raman – iodine value (g I_2 100 g fat^{-1}), (E) NIR – lignin (% w/w), (F) UV – phorbol esters (mg g^{-1}), (G) NMR – pentanol (%), (H) GC – overall quality, (I) Vis/NIR – xylene (weight percent - %), (J) XRF – Ni (%), (K) QSAR – MMP-2 pIC_{50}, and (L) MS – peroxide value (meq Kg^{-1}). *RMSEP*: root mean square error of prediction.

selected almost all variables in some cases.

The variables selected by the four methods do not match in most datasets. Few common variables are selected in all four methods (with the maximum frequency) as can be seen at the top of each subplot in Fig. 9. Therefore, the selection methods studied in this work do not converge to the same result of selected variables. Even with some very similar selection strategies to OPS (rPLS uses the regression vector, and *i*SPA uses intervals), the other methods are not able to achieve the same result; they are very different. For example, in Fig. 9B, the variables selected by OPS are more informative than those selected by the other methods since these variables are in the region with higher intensity in the pure emission spectra of catechol (305–320 nm). In Fig. 9G, the NMR spectra are composed by four signals. GA and *i*SPA selected variables throughout the spectra, including baseline. The rPLS and the OPS selected four and two signals, respectively. The NMR signal between 3.00 and 3.25 chemical shift with high intensity is pentanol specific, and the OPS algorithm was able to select this specific signal, proving that OPS finds the region with higher selectivity. Thus, the new OPS was able to select more informative and predictive variables.

5. Conclusions

The new comprehensive approaches of OPS, *auto*OPS, *feed*OPS, and *i*OPS algorithms were developed and compared to the OPS*v*1 and the other three methods of variable selection concerning prediction capacity. The prediction performance of the new strategies outperformed the OPS*v*1. The new OPS algorithms were successfully applied to several types of dataset (sparse or non-sparse). They selected interpretive variables with greater predictability and were highly reproducible, selecting the same set of variables when executing the selection more than once. In general, the new strategies were able to reduce the number of variables in all datasets significantly. Besides that, they were better than the other methods in the external prediction of all datasets. Although the new approaches presented good results, there is not a best strategy nor informative vector suitable for all datasets or dataset type. A unique algorithm that automatically executes the three new OPS approaches is easily programmable. Overall, the OPS proved to be a universal and powerful method that significantly improved the prediction ability of the models, making it simpler to interpret, and showing excellent stability by selecting the same set of variables when executing the selection more than once.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Conflict of interest

The authors have conflicts of interest with researchers that defend the algorithms used in the comparison with OPS, who may have a biased evaluation of our work.

CRediT authorship contribution statement

Jussara V. Roque: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Wilson Cardoso:** Data curation, Formal analysis, Validation, Visualization, Methodology, Writing - original draft. **Luiz A. Peternelli:**

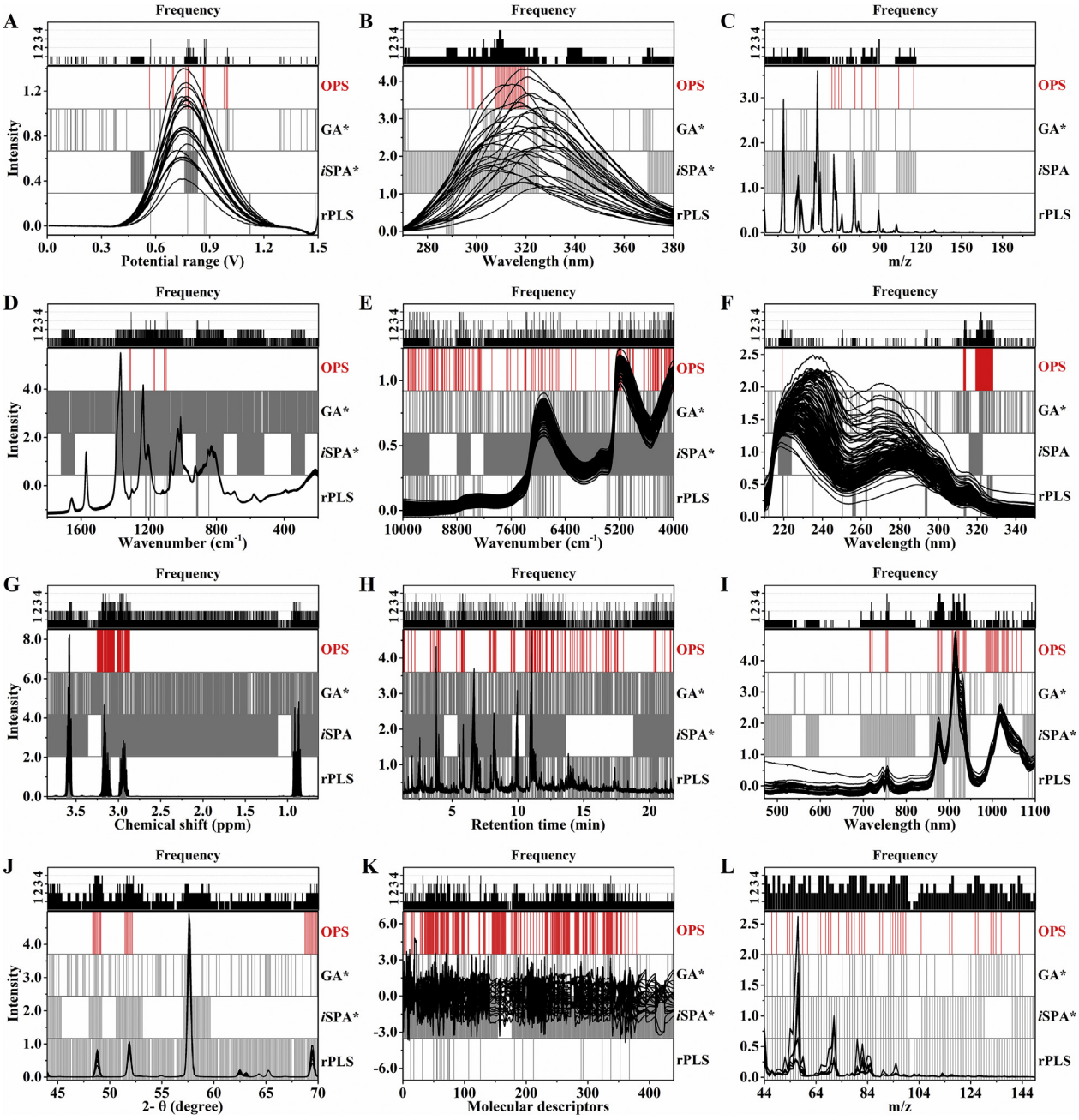


Fig. 9. (A) Voltammetry – ascorbic acid, (B) Fluorescence – catechol, (C) MS – ethanol, (D) Raman – iodine value, (E) NIR – lignin, (F) UV – phorbol esters, (G) NMR – pentanol, (H) GC – overall quality, (I) Vis/NIR – xylene, (J) XRF – Ni, (K) Autoscale QSAR – MMP-2 pIC50, and (L) MS – peroxide value with the variables selected (vertical lines) by OPS, GA, *i*SPA, and rPLS. The top graph in each subplot is the frequency that a variable was selected by the four variable selection methods. *Variables selected by the model with lowest *RMSEP*.

Supervision, Validation, Visualization, Writing - review & editing. **Reinaldo F. Teófilo:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Visualization, Writing - review & editing.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - process 310303/2015-0 and 310503/2015-9. The authors are grateful to Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for scholarship, and also to Juliano S. Ribeiro and Eduardo B. de Melo for providing GC and QSAR datasets, respectively. We are especially grateful for the reviewers' valuable comments and suggestions that contributed to improve the quality of our work.

References

[1] C.M. Andersen, R. Bro, Variable selection in regression-a tutorial, J. Chemom. 24 (2010) 728–737, https://doi.org/10.1002/cem.1360.
[2] B. Nadler, R.R. Coifman, The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration, J. Chemom. 19 (2005) 107–118, https://doi.org/10.1002/cem.915.
[3] R.G. Brereton, Introduction to multivariate calibration in analytical chemistry, Analyst 125 (2000) 2125–2154, https://doi.org/10.1039/b003805i.
[4] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of

chemometrics, Chemometr. Intell. Lab. Syst. 58 (2001) 109–130, https://doi.org/10.1016/S0169-7439(01)00155-1.
[5] S. Wold, J. Trygg, A. Berglund, H. Antti, Some recent developments in PLS modeling, Chemometr. Intell. Lab. Syst. 58 (2001) 131–150, https://doi.org/10.1016/S0169-7439(01)00156-3.
[6] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (2010) 14–32, https://doi.org/10.1016/j.aca.2010.03.048.
[7] M. Goodarzi, Y. Vander Heyden, S. Funar-Timofei, Towards better understanding of feature-selection or reduction techniques for Quantitative Structure–Activity Relationship models, TrAC Trends Anal. Chem. (Reference Ed.) 42 (2013) 49–63, https://doi.org/10.1016/j.trac.2012.09.008.
[8] A. de Araújo Gomes, R.K.H. Galvão, M.C.U. de Araújo, G. Véras, E.C. da Silva, The successive projections algorithm for interval selection in PLS, Microchem. J. 110 (2013) 202–208, https://doi.org/10.1016/j.microc.2013.03.015.
[9] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486–497, https://doi.org/10.1002/cem.893.
[10] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemometr. Intell. Lab. Syst. 41 (1998) 195–207, https://doi.org/10.1016/S0169-7439(98)00051-3.
[11] Å. Rinnan, M. Andersson, C. Ridder, S.B. Engelsen, Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS, J. Chemom. 28 (2014) 439–447, https://doi.org/10.1002/cem.2582.
[12] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Predictive-property-ranked variable reduction with final complexity adapted models in partial least squares modeling for multiple responses, Anal. Chem. 85 (2013) 5444–5453, https://doi.org/10.1021/ac400339e.
[13] A.M.K. Pedro, M.M.C. Ferreira, Nondestructive determination of solids and carotenoids in tomato products by near-infrared spectroscopy and multivariate calibration, Anal. Chem. 77 (2005) 2505–2511, https://doi.org/10.1021/ac048651r.
[14] R.M. Balabin, S.V. Smirnov, Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta 692 (2011) 63–72, https://doi.org/10.1016/j.aca.2011.03.006.
[15] I.R.N. de Oliveira, J.V. Roque, M.P. Maia, P.C. Stringheta, R.F. Teófilo, New strategy for determination of anthocyanins, polyphenols and antioxidant capacity of Brassica oleracea liquid extract using infrared spectroscopies and multivariate regression, Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 194 (2018) 172–180, https://doi.org/10.1016/j.saa.2018.01.006.
[16] C. Assis, R.S. Ramos, L.A. Silva, V. Kist, M.H.P. Barbosa, R.F. Teófilo, Prediction of lignin content in different parts of sugarcane using near-infrared spectroscopy (NIR), ordered predictors selection (OPS), and partial least squares (PLS), Appl. Spectrosc. 71 (2017) 2001–2012, https://doi.org/10.1177/0003702817704147.
[17] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48, https://doi.org/10.1002/cem.1192.
[18] J. Yuan, S. Yu, T. Zhang, X. Yuan, Y. Cao, X. Yu, X. Yang, W. Yao, QSPR models for predicting generator-column-derived octanol/water and octanol/air partition coefficients of polychlorinated biphenyls, Ecotoxicol. Environ. Saf. 128 (2016) 171–180, https://doi.org/10.1016/j.ecoenv.2016.02.022.
[19] C.S.W. Miaw, C. Assis, A.R.C.S. Silva, M.L. Cunha, M.M. Sena, S.V.C. de Souza, Determination of main fruits in adulterated nectars by ATR-FTIR spectroscopy combined with multivariate calibration and variable selection methods, Food Chem. 254 (2018) 272–280, https://doi.org/10.1016/j.foodchem.2018.02.015.
[20] C. Assis, L.S. Oliveira, M.M. Sena, Variable selection applied to the development of a robust method for the quantification of coffee blends using mid infrared spectroscopy, Food Anal. Methods. 11 (2018) 578–588, https://doi.org/10.1007/s12161-017-1027-7.
[21] Í.P. Caliari, M.H.P. Barbosa, S.O. Ferreira, R.F. Teófilo, Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods, Carbohydr. Polym. 158 (2017) 20–28, https://doi.org/10.1016/j.carbpol.2016.12.005.
[22] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17, https://doi.org/10.1016/0003-2670(86)80028-9.
[23] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148, https://doi.org/10.1080/00401706.1969.10490666.
[24] R.K.H. Galvão, M.C.U. Araújo, W.D. Fragoso, E.C. Silva, G.E. José, S.F.C. Soares, H.M. Paiva, A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm, Chemometr. Intell. Lab. Syst. 92 (2008) 83–91, https://doi.org/10.1016/j.chemolab.2007.12.004.
[25] O.O. Soyemi, M.A. Busch, K.W. Busch, Multivariate analysis of near-infrared spectra using the G-programming language, J. Chem. Inf. Comput. Sci. 40 (2000) 1093–1100, https://doi.org/10.1021/ci000447r.
[26] N. Klaas, M. Faber, R. Bro, Standard error of prediction for multiway PLS, Chemometr. Intell. Lab. Syst. 61 (2002) 133–149, https://doi.org/10.1016/S0169-7439(01)00204-0.
[27] M. Dyrby, S.B. Engelsen, L. Nørgaard, M. Bruhn, L. Lundsberg-Nielsen, Chemometric quantitation of the active substance (containing C≡N) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-Raman spectra, Appl. Spectrosc. 56 (2002) 579–585, https://doi.org/10.1366/0003702021955358.
[28] R. Kiralj, M.M.C. Ferreira, A priori molecular descriptors in QSAR: a case of HIV-1 protease inhibitors, J. Mol. Graph. Model. 21 (2003) 435–448, https://doi.org/10.1016/S1093-3263(02)00201-2.
[29] R.F. Teófilo, Métodos quimiométricos em estudos eletroquímicos de fenóis sobre filmes de diamante dopado com boro,, Thesis (Doutorado em Química), Universidade Estadual de Campinas, 2007.
[30] T. Skov, D. Ballabio, R. Bro, Multiblock variance partitioning: a new approach for comparing variation in multiple data blocks, Anal. Chim. Acta 615 (2008) 18–29, https://doi.org/10.1016/j.aca.2008.03.045.
[31] L.B. Lyndgaard, K.M. Sørensen, F. Berg, S.B. Engelsen, Depth profiling of porcine adipose tissue by Raman spectroscopy, J. Raman Spectrosc. 43 (2012) 482–489, https://doi.org/10.1002/jrs.3067.
[32] J. V Roque, L.A.S. Dias, R.F. Teófilo, Multivariate calibration to determine phorbol esters in seeds of Jatropha curcas L. Using near infrared and ultraviolet spectroscopies, J. Braz. Chem. Soc. 28 (2017) 1506–1516, https://doi.org/10.21577/0103-5053.20160332.
[33] H. Winning, F.H. Larsen, R. Bro, S.B. Engelsen, Quantitative analysis of NMR spectra with chemometrics, J. Magn. Reson. 190 (2008) 26–32, https://doi.org/10.1016/j.jmr.2007.10.005.
[34] J.S. Ribeiro, F. Augusto, T.J.G. Salva, M.M.C. Ferreira, Prediction models for Arabica coffee beverage quality based on aroma analyses and chemometrics, Talanta 101 (2012) 253–260, https://doi.org/10.1016/j.talanta.2012.09.022.
[35] Y. Wang, X. Zhao, B.R. Kowalski, X-ray fluorescence calibration with partial least-squares, Appl. Spectrosc. 44 (1990) 998–1002, https://doi.org/10.1366/0003702904086867.
[36] E.B. de Melo, A QSAR study of matrix metalloproteinases type 2 (MMP-2) inhibitors with cinnamoyl pyrrolidine derivatives, Sci. Pharm. 80 (2012) 265–281, https://doi.org/10.3797/scipharm.1112-21.