

## Joint selection of wavenumber regions for MidIR and RAMAN spectra and variables in PLS regression using Genetic Algorithms

**Titre:** Sélection conjointe de régions de spectres MidIR et RAMAN et de variables en régression PLS à l'aide d'Algorithmes Génétiques

Grosmaire Lidwine<sup>1</sup>, Reynès Christelle<sup>2</sup> and Sabatier Robert<sup>2</sup>

**Abstract:** Many methods exist for feature selection in PLS regression when there are too many variables. Less methods are available for selecting wavenumber regions for MidIR or RAMAN spectra. In this work, PLS has been coupled with genetic algorithms to allow for the selection of intervals in spectra. This work was motivated by a regression issue about transformation of cassava. Those data consist of three tables: RAMAN spectra, MidIR spectra and physico-chemical variables. The purpose is to adapt to this regression context a strategy previously designed to select intervals in NIR spectra in classification. A new algorithm is proposed to fit such multiblock data in PLS1 regression context. Illustrations on simulated data are performed before application to the real dataset.

**Résumé :** De nombreuses méthodes adaptées pour la régression PLS, s'intéressent aux choix de variables explicatives, quand celles-ci sont en nombre trop important. Quand il s'agit de sélectionner des intervalles pour des spectres, la panoplie des techniques est plus réduite. Dans ce travail, PLS a été associée aux algorithmes génétiques pour permettre la sélection d'intervalles dans des spectres. L'origine de ce travail est une problématique de régression pour des données sur la transformation de manioc. Ces données sont constituées de trois tableaux : des spectres RAMAN, MidIR et des variables physico-chimiques. Il s'agit d'adapter au contexte de régression une stratégie précédemment mise au point pour la sélection d'intervalles uniquement pour des spectres NIR en discrimination. Nous avons développé un algorithme génétique spécialement adapté à ce type de données (multitableau), pour le cas de la régression PLS1. Des illustrations sur des données simulées sont proposées avant l'application au jeu de données réel.

**Keywords:** PLS Regression, Genetic Algorithm, MidIR and RAMAN spectra, Variable Selection, Selection of wavenumber regions

**Mots-clés :** Méthode PLS, Algorithme Génétique, Spectres MidIR et RAMAN, Choix de variables, Sélection d'intervalles

**AMS 2000 subject classifications:** 62H30, 62H25, 62-07, 90-08

### 1. Introduction

Selecting inputs is a very common issue in chemometrics, especially in Partial Least Squares (PLS) regression. Even if PLS is known to work for matrices with more columns than rows, this method improves if an appropriate feature selection has been performed. Several approaches have been proposed (Du et al., 2004; Høskuldsson, 2001; Zou et al., 2007) but few exist dealing

<sup>1</sup> Laboratoire de Physique Moléculaire et Structurale - UMR Qualisud - Université Montpellier 1 - France.

E-mail: [lidwine.grosmaire@univ-montp1.fr](mailto:lidwine.grosmaire@univ-montp1.fr)

<sup>2</sup> Laboratoire de Physique Industrielle et Traitement de l'Information - EA 2415 - Université Montpellier 1 - France

E-mail: [christelle.reynes@univ-montp1.fr](mailto:christelle.reynes@univ-montp1.fr) and E-mail: [robert.sabatier@univ-montp1.fr](mailto:robert.sabatier@univ-montp1.fr)

with wavenumber interval selections, which generally involve more statistically complex methods. Indeed, selecting individual wavenumbers is a chemical nonsense and the strong autocorrelation of variables has to be taken into account in order to produce adequate intervals. Genetic algorithms (GAs) in particular have been used jointly with PLS to select intervals in mid-infrared (midIR) spectra (Leardi, 2001; Leardi and Nørgaard, 2004). A few other specific approaches have been developed such as iPLS (Nørgaard et al., 2000). In this context, a quite general review in interval selection can be found in Høskuldsson (2001).

This work fits into the context of the production and the transformation of cassava. The consumption of this product increases regularly in the world, especially in the tropical regions (Tonukari, 2004). After harvest, cassava is naturally fermented and sun-dried to give sour starch which is mainly used in typical bread production (Bertolini et al., 2001; Demiate et al., 2000; Dufour et al., 1995; Marcon et al., 2009; Mestres et al., 2000; Mestres and Rouau, 1997). In addition, this product is gluten-free and shows good breadmaking ability. However, cassava processing is currently carried out on small and uneconomic farms. Therefore, studies on this starch, especially for industrial purposes, appear to date necessary. The aim of this work is to attempt to explain the breadmaking ability from the different parameters studied.

In this article, we use and adapt a GA previously developed for discriminant analysis to select wavelength interval as inputs in Near IR spectra (Reynes et al., 2006). Here, the objective is not to perform discrimination but to predict a continuous variable (using PLS). Moreover, instead of selecting a set of intervals within a single table, selection is performed through several blocks using intervals in the *spectral* blocks and individual variables in other blocks. Hence, this method allows to combine information of different types in order to predict a continuous variable. Selecting an unknown number of intervals of unknown length combined with an unknown number of individual variables (potentially no interval or no individual variable can be selected) leads to a combinatorial issue which can only be approached through a heuristic method. Among heuristic methods no method outperforms other ones and a choice has to be made. Our expertise in the field of genetic algorithms led us to propose a method based on them. Furthermore, several reviews (see Padgett and Saad 2009 for example) allowed to demonstrate the usefulness of genetic algorithms in this field of research. In the real application, the dataset is a multiblock table with three blocks (measured on the same  $n = 52$  observations) and individual variables have to be selected for physicochemical variables whereas wavenumber intervals should be chosen for two spectroscopic blocks. Hence, the proposed statistical method is not a method using a multiway criterion, however it allows to deal with data structured in multi-blocks as it is able to take into account blocks with different types of data. Before application on the real dataset, a first running on a simulated dataset will show the quality of the new algorithm.

## 2. The real dataset

In this work, native and modified cassava starches from Colombia have been investigated in order to provide information on expansion ability. 13 varieties have been selected (10 from highland and 3 from lowland) after 4 controlled processes: non fermented oven-dried (NFO), fermented oven-dried (FO), non fermented sun-dried (NFS) and fermented sun-dried (FSR). The bread-making ability, which is represented by the values of loaf volume, is strongly impacted by the varietal effect and also by the treatment effect. The problematic is to try to explain one variable of

interest (breadmaking ability) from different physicochemical parameters (amylose content and RVA parameters), MidIR and Raman spectra.

Starches are composed basically of two different glucose polymers: amylose (linear polymer) and amylopectin (branched polymer). The amylose content is determined by Differential Scanning Calorimetry (DSC) (Bertolini et al., 2001). When starch is dispersed in water, the viscosity of the medium changes as a function of the temperature (Thomas et al., 1999). These changes in viscosity are recorded with a Rapid Visco Analyser (RVA) during controlled mixing, heating and cooling programs. Dry starch, when heated in water, starts swelling as the water is absorbed by the granules. Past a certain state, the phenomenon is irreversible and is called gelatinisation which is a temperature range in which the starch granules lose their forms and cristallinity. Once the gelatinised temperature is reached, the viscosity increases rapidly as the swelling reaches a maximum (peak viscosity). During this process, the amylose molecules leach out of the granules creating a medium around the particles. The granules keep swelling until they have reached equilibrium. After maximum viscosity is reached, if the temperature continues to increase, the granules will burst and disperse completely. Solubilised starch polymers and remaining insoluble granular fragments have a tendency to reassociate after heating. This reassociation is referred to as retrogradation and results in an ordered structure and an increase in viscosity. Retrogradation is especially evident when starches are cooled and results in the formation of crystalline aggregates and a gelled texture. RVA profiles obtained lead to 12 parameters of time, temperature and viscosity during the gelatinisation and retrogradation processes.

Mid-infrared and Raman spectra were recorded in the  $650\text{--}4000\text{ cm}^{-1}$  (3351 variables) and  $230\text{--}3800\text{ cm}^{-1}$  (4562 variables) spectral ranges, respectively. Raw spectra were pre-processed using a baseline correction and a vectorial normalisation of the total spectral region with Lab-Spec5 software. The baseline correction mentioned is a linear one which is realised with a linear regression through the chosen points on spectrum and removes the linear baseline thus found.

Finally, 52 samples (13 varieties x 4 treatments) were analysed to determine chemical and physicochemical characteristics as (i) the breadmaking ability, which represent the dough expansion, (ii) the amylose content, (iii) the RVA parameters and (iv) the MidIR and Raman spectra.

### 3. Reminder about Genetic Algorithms

Genetic Algorithms (GA), introduced by Goldberg (1989); Holland (1975), are powerful tools in optimisation issues. They are inspired by nature imitation and especially by natural selection. In our context, a GA will be used to determine the selected intervals by optimising a regression-based criterion. Before detailing the developed method, let us begin with an introduction about the general unfolding and main steps of GAs.

GAs are population based heuristics, indeed, the algorithm begins with a population of  $T_{pop}$  potential solutions to the optimisation problem. Those solutions are encoded by a numerical vector allowing a complete description of each solution. Then, this population evolves according to three operators: cross-over, mutation and selection. Selection is a crucial step as it favours the survival of good individuals with regards to the optimised criterion. This fitness value quantifies the quality of each solution with regards to the problem. Cross-over and mutation are independent from the problem and are therefore randomly performed without consideration to the fitness value. Then, any GA can be described as in Table 1.

TABLE 1. *General unfolding of GAs.*


---

<b>Main steps of a GA:</b>
1. construction of the first generation,
2. evaluation,
<b>while</b> stopping criteria not met <b>do:</b>
3. selection step,
4. crossover,
5. mutation,
6. evaluation,
<b>end.</b>

---

Concerning the stopping criterion, the chosen approach is to set a maximum number of generations,  $N_{gene}$ . Global convergence is ensured by the fulfillment of the theoretical required conditions (see Section 4.7) hence there are two usual ways of stopping a GA: defining a maximum number of generations or defining a minimum fitness improvement. We consider that they are equivalent as they both require prior steps of trials. Indeed, the fitness variation control might seem smarter but as the global behaviour of the fitness function is unknown, it is impossible to define what is a negligible improvement of its value without performing several trials aiming at studying the global evolution of the fitness value. Actually, performing such trials leads to being able to define both the negligible fitness improvement and the maximum number of generations required to reach it. We think that using a maximum number of generations is more honest as it does not lead to believe that the criterion is based on a fake knowledge of the fitness function.

#### 4. A new proposal for multiblock datasets

The proposed method consists in simultaneously using PLS1 for a multiblock table  $X$  and a GA whose characteristics fit each block specificities. In our application,  $X = [X_1, X_2, X_3]$  with  $X_1$  the matrix of physicochemical variables,  $X_2$  gathers MidIR spectra and  $X_3$  Raman spectra. The output variable,  $y$  as well as  $X$  are measured on the same  $n = 52$  observations.

Seeking for presentation simplicity, the proposed method is described for  $K = 3$  blocks and PLS1 but it can easily be generalised to  $K > 3$  and/or PLS2. From now on, the developed approach will be denoted selGAmPLS (selection Genetic Algorithm in multiblock PLS).

##### 4.1. Overview of selGAmPLS

The method consists of two nested steps: on one hand, the selection of relevant wavenumber intervals (namely,  $\tilde{X}_2$  for MidIR data and  $\tilde{X}_3$  for RAMAN data) and individual variables ( $\tilde{X}_1$  a subset of the physicochemical variables) gathered in  $\tilde{X} = [\tilde{X}_1, \tilde{X}_2, \tilde{X}_3]$  and on the other hand, the modelling of  $y$  thanks to PLS. Finally, number of PLS components  $A$  will be automatically chosen thanks to cross-validation inside the algorithm. In the following paragraphs, the different parts of the GA are described.

#### 4.2. The fitness for selection of wavenumber regions and variables

The optimisation issue has to make a balance between two objectives: achieving a precise modelling of  $y$  while keeping a simple model using as few variables as possible. This last constraint allows to obtain a more interpretable model as well as a model which is less prone to overfitting. Hence, the fitness function can be defined as follows:

$$fitness = cor^2(y, \hat{y}) + c \times (\alpha N_{selvar} + \beta) \quad (1)$$

where  $cor$  is the usual correlation coefficient between  $y$  and its modelling  $\hat{y}$  thanks to cross-validated PLS1 applied on  $\tilde{X}$ . The second term of the fitness value deals with the total number of selected variables (summing both individual variables and wavenumbers selected by intervals), denoted  $N_{varsel}$ . The  $c$  coefficient aims at balancing those two parts.  $\alpha$  and  $\beta$  allow to scale  $N_{selvar}$  into  $[0, 1]$ .

In order to choose meaningful wavenumber intervals, during GA generations, a minimal interval width denoted  $w_{min}$  and a minimal distance between successive intervals,  $d_{min}$ . If two successive intervals are less than  $d_{min}$ , the two intervals are merged.

#### 4.3. Solution encoding

The first step in GAs is to define the encoding allowing to describe any potential solution as a numerical vector. In our application, a solution is a subset of variables (either individual ones or intervals). In order to encode solutions the three blocks are combined into a single matrix so that variables are numbered from 1 to 7926 (13+3351+4562). Selected intervals are encoded by both the number of the first variable of the interval and the number of the last one. Individual variables are encoded as their number repeated twice.

#### 4.4. Crossover operator

The objective of this step is to make combinations of the characteristics from the previously selected solutions in order to potentially gather interesting selections. Two parents (solutions) are crossed to obtain two children (to keep the same population size). A one point crossover is applied and results in exchanging the first part of one parent characteristics with the same part of the other parent characteristics. To perform the crossover, the population is randomly permuted in order to form  $T_{pop}/2$  new pairs. Each of these pairs has probability  $\pi_c \in [0, 1]$  (the crossover rate) to undergo crossover. Then,  $\pi_c$  represents the proportion of the population which is crossed. It is important to notice that this operator is completely independent from the issue. Indeed, such a crossover can generate children which are indiscriminately better or worse than the parents according to their fitness value.

#### 4.5. Mutation operator

This step brings the necessary hazard to efficiently explore the solution space. It has to ensure that any point of this space can be reached. Moreover, if a local optimum is obtained, mutations will

avoid a too quick convergence to it. Thus, efficient convergence of the GA is highly dependent upon this step.

In practice, mutation consists in adding, deleting or modifying an interval or individual variable. The mutation probability,  $\pi_m$ , is defined to control the mutation rate in the population: each solution has probability  $\pi_m$  to undergo mutation.

#### 4.6. Selection operator

The goal of selection step is to select, at each generation, the best individuals according to their fitness values. However, keeping only the very best individuals would probably lead to quick convergence. In such a case, the selected solution would be likely to be suboptimal. The selection scheme described in [Reeves and Rowe \(2003\)](#) will be used here. In this method, a selection probability is computed for each individual of the current population. In this way, a good solution (with regards to the fitness) will have a higher probability to be selected but even a bad solution will have a positive probability to be selected. This allows to maintain heterogeneity in the population. It has to be taken into account that a bad solution can possibly evolve to a solution which is better than the best one in the current population (especially at the beginning of the algorithm).

For each individual, the fitness is computed. Then, the solutions are ranked according to their fitness value (the best one having the highest rank). Thus, to each solution  $i$ , a rank  $r(i)$  is affected. These ranks allow to compute a selection probability as follows:

$$p(i) = a \times r(i) + b. \quad (2)$$

$a$  and  $b$  parameters are computed so that:

- the sum of selection probabilities equals one:  $\sum_{i=1}^{T_{pop}} (a \times r(i) + b) = 1$ ,
- the best individual has a selection probability twice as high as the individual having median rank:  $p(T_{pop}) = 2p(T_{pop}/2)$ .

#### 4.7. Convergence necessary and sufficient conditions

[Bhandari et al. \(1996\)](#) defined two necessary and sufficient conditions for such an algorithm to converge:

- elitism step: the best solution in the present population has a fitness value no less than the fitness value of the optimal individuals from the previous populations,
- each solution has a positive probability of going to an optimal solution within any given iteration.

The first condition is quite easy to implement. The best solution at each generation has to be automatically introduced in the next generation. For the second condition, it depends on the mutation step. Indeed, the mutation step has to be defined so that any solution can evolve into any other one in any given number of iterations. This is allowed by the previously defined mutation operator.

#### 4.8. Implementation of selGAmPLS methodology

In the use of selGAmPLS, some parameters have to be chosen.  $T_{pop}$ ,  $N_{gene}$ ,  $\pi_c$  and  $\pi_m$  are the GA dependent tuning parameters and have been given default values thanks to preliminary analyses (results not shown). The proposed default parameters are:  $T_{pop} = N_{gene} = 200$ ,  $\pi_m = 0.9$  and  $\pi_c = 0.5$ . It should be noted that the values of those parameters only influence convergence speed and not convergence itself (see previous section for more about convergence).

This set of default parameters is quite pessimistic and allows to obtain good results for data up 8000 variables (the maximum number of variables used in our applications).  $\pi_m$  and  $\pi_c$  values do not need to be modified. The  $T_{pop}$  value provides a satisfactory exploration of solution space, it may be tuned by the user if he wants to intensify search but the simplest solution is to keep it at default value and to tune  $N_{gene}$ . Indeed, there is a very simple tool to visually study the required number of generations: a plot is automatically generated after each generation allowing to show if average parameters of the population stabilise or not. Hence, if the parameters are stable from 50 to 100 generations, the user may want to reduce  $N_{gene}$ , on the contrary, if parameters are still evolving after 200 generations, the user should try a higher  $N_{gene}$ . It is then possible to use the algorithm without any parameter tuning except if a graphical convergence is not observed after 200 generations (which should not be the case with less than 8000 variables). So tuning parameters is most of time not necessary.

The other parameters  $N_{selvar}$ ,  $w_{min}$  and  $d_{min}$  are more connected to the application and must be chosen according to physicochemical criteria.

The algorithm has been implemented into R (R Development Core Team, 2011) and can be obtained on request to the corresponding author.

### 5. Applications

In this part, before working with the dataset described in Section 2, we used a simulated dataset in order to assess the method performances.

#### 5.1. The simulated dataset

The simulated dataset is constructed so that it contains information close to the target dataset. The variable of interest (whose value is to be predicted),  $y$ , is randomly generated according to a uniform distribution for  $n = 60$  observations. The explanatory variables are organised into  $K = 4$  blocks: three blocks simulating spectra type data and one block of individual variables (similar to physico-chemical variables in the real dataset).

The spectra type blocks are produced according to the method proposed in Shariati-Rad and Hasani (2010). They are generated by combinations of Gaussian peaks, with different variances. Each spectrum contains  $p_1 = p_2 = p_3 = 500$  wavenumbers. An individual gaussian noise (with standard deviation of 0.03% of the maximum value in each block) and a global background noise (analogous to baseline) is added to each block of spectra. Spectra profiles contain three peaks for the second block and four peaks for the first and third blocks. In the first two blocks, the second peak is correlated to  $y$  with correlation coefficients of 0.7265 and 0.7596 respectively. No peak of the third block is related to  $y$ . Uncorrelated peaks are randomly generated with average absolute

TABLE 2. Results obtained by the three methods applied on the simulated dataset (var nb: number of variables used in the model, int nb: number of intervals chosen by the algorithm, A: PLS model dimension chosen thanks to PRESS,  $R_{CV_{train}}^2$ : average squared correlation coefficient between  $y$  and  $\hat{y}$  obtained with the training set using 4-fold cross validation,  $R_{test}^2$  squared correlation coefficient between  $y$  and  $\hat{y}$  obtained with the test set). For selGAmPLS, the given results are averages of the ten runs.

Method	var nb	int nb	A	$R_{CV_{train}}^2$	$R_{test}^2$
selGAmPLS	60.4	3.5	3.8	0.7986	0.6877
PLS + VIP (>1)	226	9	2	0.7636	0.6204
PLS	1510	-	2	0.6715	0.6383

values of correlation coefficients with  $y$  of 0.1468 for the first block, 0.0707 for the second one and 0.1308 for the last one. The average spectrum for each block is depicted in the right panel of Fig. 1.

The fourth block, like in the real dataset, is a collection of  $p_4 = 10$  individual variables, generated by ten gaussian laws. The first two variables are generated so that they have a linear correlation coefficient with  $y$  of 0.7135 and 0.5912 respectively. The eight other ones are randomly generated (leading to an absolute value of linear correlation coefficient with  $y$  ranging from 0.0564 to 0.2425).

The obtained data are split into two subsets: the learning set contains forty randomly chosen observations and the test set contains the twenty remaining ones. selGAmPLS has been applied to the simulated dataset with the following parameter values:  $T_{pop} = 200$ ,  $N_{gene} = 200$ ,  $\pi_m = 0.9$ ,  $\pi_c = 0.5$  (one run takes about 22 minutes on a desktop computer with processor Intel Core i7 2.93 GHz). The results obtained after ten runs of selGAmPLS are given in Fig. 1 and Tab. 2. According to Fig. 1, it appears that in the final populations most solutions contain selected intervals from the target ones (shaded in gray) for the first two blocks, most solutions contain no selected intervals in the third block and at least one of the two correlated individual variables of the fourth block. Moreover, if we focus at the best solution of each run, no one selected intervals in the third block. Furthermore, seven out of ten solutions contain intervals overlapping the targets ones and at least one of the two correlated individual variables of the fourth block.

As far as we know there is no other method allowing to perform comparable variable selection (allowing both interval and individual selection through multiblock tables). If we wanted to compare to other interval selection algorithms, our blocks would have to be concatenated. Yet, selecting intervals on the physicochemical parameters is not relevant and there is a risk of selecting intervals which overlap two blocks. That is why we chose to compare to the standard method to perform individual variable selection in PLS context: VIP methodology (Wold et al., 1993). The results of selGAmPLS were compared with PLS used on the whole dataset and with the variables selected through VIP with a threshold value of one. By looking at Tab. 2, it appears that selGAmPLS achieves the most stringent choice of variables and the best  $R^2$  value for the test set.

## 5.2. The cassava starches dataset

The algorithm has been run ten times with the same parameters as for the simulated dataset (one run takes about 46 minutes on a desktop computer with processor Intel Core i7 2.93 GHz). Taking into account the best solution of the ten final populations, an average cross-validated  $cor^2(y, \hat{y})$

TABLE 3. Results obtained by the three methods applied on cassava data (var nb: number of variables used in the model, A: PLS model dimension chosen thanks to PRESS,  $R^2$ : squared correlation coefficient between  $y$  and  $\hat{y}$  obtained in the training set,  $R_{CV}^2$  average squared correlation coefficient between  $y$  and  $\hat{y}$  obtained through 10-FCV).

Method	var nb	A	$R^2$	$R_{CV}^2$
selGAmPLS	311	12	0.9936	0.8273
PLS + VIP	4	3	0.7210	0.6650
PLS	7926	7	0.7836	0.6605

value of 0.885 was achieved with 0.017 standard deviation. On average, 2.7 physicochemical variables, 2.8 MidIR intervals (corresponding to an average of 46.3 wavenumbers) and 1.8 RAMAN (37.6 wavenumbers) intervals have been selected.

Fig. 2 gives a representation of the selected variables across the ten final populations. Selected variables are concentrated in some favoured zones confirming the convergence of the algorithm even if a few equivalent solutions can be found.

Actually, four physicochemical variables, two MidIR intervals and four Raman intervals have been selected. It corresponds to a total of 311 variables covering about 4% of the total number of variables.

In order to evaluate the relevance of this selection, it has been compared to a PLS run on the whole datasets (7926 variables) and to a PLS run on the variables selected through VIP methodology. This approach allowed to selected only four physicochemical variables and no spectroscopic wavenumbers. The corresponding results are given in Tab. 3.

By applying PLS to the whole dataset, the obtained model fails to extract interesting information from those very noisy data (many variables, sometimes strongly correlated). The VIP selection only extracts four physicochemical variables and the obtained model is quite interesting achieving less overfitting than the PLS on all variables (the loss between learning set and cross-validated results is smaller for PLS+VIP). selGAmPLS achieves the best modelling for both learning and cross-validated results.

According to these results, it appears that physicochemical variables play a very important part in modelling (using only four of them leads to performances very close to using the whole dataset). However, taking into account spectroscopic variables allows to account for 16% extra variability.

The physicochemical variables selected are derived from RVA data (cf. Fig. 3): peak viscosity, holding strength, breakdown and relative breakdown. As the temperature is increased, the starch granules swell and increase the viscosity of the paste until the peak viscosity is reached. This parameter is related to the granule size and reflects the molecular degradation of starch (Dias et al., 2011). The holding strength is related to the behaviour of starch during processing and illustrated the stability of the paste during cooking (Dias et al., 2011; Klug Tavares et al., 2010). The breakdown parameter is related with the rigidity of swollen granules and with the amylopectin content which is responsible for water uptake (Juhász and Salgo, 2008). Finally, these three parameters are positively correlated with the water absorptivity (Tamaki et al., 2005) and they are therefore key pasting parameters to describe the breadmaking ability.

Regarding spectroscopic variables, as samples analysed have very similar chemical compositions (15.7 to 21.7 % of amylose), the vibrational spectra recorded do not show significant differences (cf Fig. 4 and 5). However, the spectral regions selected by selGAmPLS, whether high or low frequency do not correspond to specific vibrational bands or combination bands, but contain

relevant information as they significantly improve the model.

To give more sense to the interpretation of selected data, especially on spectroscopic ones, we have eliminated non informative regions in MidIR and Raman spectra and worked in the MidIR range between 800 and 1180  $\text{cm}^{-1}$  and in the raman range between 50 and 1200  $\text{cm}^{-1}$ .

This spectral region was shown to correspond to the conformational and crystalline order of starch (van Soest et al., 1995). The bands in this region essentially result from C-O and C-C vibrational modes that are highly coupled, making the exact band assignment difficult (Goodfellow and Wilson, 1990; Kizil et al., 2002).

After another run of selGAmPLS, two MidIR intervals and four Raman intervals have been selected. The two wavenumber ranges selected in MidIR spectra (960-983  $\text{cm}^{-1}$  and 1017-1045  $\text{cm}^{-1}$ ) are sensitive to the crystalline state of starch (Mutungi et al., 2011). Moreover, the bands at 1044  $\text{cm}^{-1}$  and 1017  $\text{cm}^{-1}$  are associated to the ordered and amorphous structure of starch, respectively (van Soest et al., 1995). The ratio of absorbances 1044/1017  $\text{cm}^{-1}$  can be used to quantify the degree of order in starch samples (Mutungi et al., 2011). In the case of the Raman spectra, three selected intervals (517-523  $\text{cm}^{-1}$ , 628-669  $\text{cm}^{-1}$  and 799-822  $\text{cm}^{-1}$ ) are associated with the skeletal modes of the glucose ring, whereas the range between 996 and 1036  $\text{cm}^{-1}$  is related to starch crystallinity, as in MidIR. Therefore, the spectral regions corresponding to the crystalline order of starch play an important role in the breadmaking ability.

## 6. Conclusion

Feature selection is an important issue in data analysis. It allows to provide more powerful and meaningful models leading to a better understanding of studied phenomena. When dealing with thousands of variables, an exploration of all possible combinations is unworkable and a step by step approach is most of time not appropriate to provide an optimal model. Furthermore, usual methods (including VIP) select individual variables and not intervals that are the only meaningful entities in spectrometry. Using a heuristic method is therefore necessary.

Genetic algorithms provide a very adaptable and efficient solution when dealing with both several kinds of variable selections (individual vs intervals) and multiblock tables. It can be noticed that generalisation to PLS2 is easy, for example by using the Escoufier RV instead of  $R^2$ .

In this work, the result obtained is very interesting for a predictive use. In terms of interpretation, the method allowed to highlight the importance of some physico-chemical variables and to select a small number of short intervals to significantly complete the resulting model. These intervals are difficult to interpret for chemists and therefore do not participate in understanding the phenomenon but are essential for the quality of the final model. However, the use of part of the spectra gives more sense to the interpretation of physicochemical interpretation at the expense of the quality of the model. It appears that the wavelength regions which have been selected to explain the breadmaking ability are those associated to the ordered and amorphous structures of starch. Due to the small learning sample size, improving model interpretation implies a loss in prediction. Despite this small size probably implies a part of overfitting, it nevertheless provides interesting insights in the model meaning. Further analyses on new data will allow to confirm these promising results.

## References

- Bertolini, A., Mestres, C., Lourdin, D., Valle, G., and Colonna, P. (2001). Relationship between thermomechanical properties and baking expansion of sour cassava starch (polvilho azedo). *Journal of the Science of Food and Agriculture*, 81(4):429–435.
- Bhandari, D., Murthy, C., and Pal, S. (1996). Genetic algorithm with elitist model and its convergence. *International Journal of Pattern Recognition and Artificial Intelligence*, 10(6):731–747.
- Demiate, I., Dupuy, N., Huvenne, J., Cereda, M., and Wosiacki, G. (2000). Relationship between baking behavior of modified cassava starches and starch chemical structure determined by ftir spectroscopy. *Carbohydrate Polymers*, 42(2):149–158.
- Dias, A., Zavareze, E., Elias, M., Helbig, E., Da Silva, D., and Ciacco, C. (2011). Pasting, expansion and textural properties of fermented cassava starch oxidised with sodium hypochlorite. *Carbohydrate Polymers*, 84(1):268–275.
- Du, Y., Liang, Y., Jiang, J., Berry, R., and Ozaki, Y. (2004). Spectral regions selection to improve prediction ability of pls models by changeable size moving window partial least squares and searching combination moving window partial least squares. *Analytica chimica acta*, 501(2):183–191.
- Dufour, D., Brabet, C., Zakhia, N., Chuzel, G., Egbe, T., Brauman, A., and Treche, S. (1995). Influence de la fermentation et du séchage solaire sur l'acquisition du pouvoir de panification de l'amidon aigre de manioc. *Agbor Egbe T., Braumann A., Griffon D., Trèche S. Transformation Alimentaire du Manioc. Editions ORSTOM: Paris, Francia*, pages 399–417.
- Goldberg, D. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-wesley.
- Goodfellow, B. and Wilson, R. (1990). A fourier transform ir study of the gelation of amylose and amylopectin. *Biopolymers*, 30(13-14):1183–1189.
- Holland, J. (1975). *Adaptation in natural and artificial systems*. Number 53. University of Michigan press.
- Høskuldsson, A. (2001). Variable and subset selection in pls regression. *Chemometrics and intelligent laboratory systems*, 55(1-2):23–38.
- Juhász, R. and Salgo, A. (2008). Pasting behavior of amylose, amylopectin and their mixtures as determined by rva curves and first derivatives. *Starch-Stärke*, 60(2):70–78.
- Kizil, R., Irudayaraj, J., and Seetharaman, K. (2002). Characterization of irradiated starches by using ft-raman and ftir spectroscopy. *Journal of agricultural and food chemistry*, 50(14):3912–3918.
- Klug Tavares, A., Zanatta, E., da Rosa Zavareze, E., Helbig, E., and Guerra Dias, A. (2010). The effects of acid and oxidative modification on the expansion properties of rice flours with varying levels of amylose. *LWT-Food Science and Technology*, 43(8):1213–1219.
- Leardi, R. (2001). Genetic algorithms in chemometrics and chemistry: a review. *Journal of chemometrics*, 15(7):559–569.
- Leardi, R. and Nørgaard, L. (2004). Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *Journal of Chemometrics*, 18(11):486–497.
- Marcon, M., Kurtz, D., Raguzzoni, J., Delgadillo, I., Maraschin, M., Soldi, V., Reginatto, V., and Amante, E. (2009). Expansion properties of sour cassava starch (polvilho azedo): Variables related to its practical application in bakery. *Starch-Stärke*, 61(12):716–726.
- Mestres, C., Boungou, O., Akissioe, N., and Zakhia, N. (2000). Comparison of the expansion ability of fermented maize flour and cassava starch during baking. *Journal of the Science of Food and Agriculture*, 80(6):665–672.
- Mestres, C. and Rouau, X. (1997). Influence of natural fermentation and drying conditions on the physicochemical characteristics of cassava starch. *Journal of the Science of Food and Agriculture*, 74(2):147–155.
- Mutungi, C., Onyango, C., Doert, T., Paasch, S., Thiele, S., Machill, S., Jaros, D., and Rohm, H. (2011). Long- and short-range structural changes of recrystallised cassava starch subjected to in vitro digestion. *Food Hydrocolloids*, 25(3):477–485.
- Norgaard, L., Saudland, A., Wagner, J., Nielsen, J., Munck, L., and Engelsen, S. (2000). Interval partial least-squares regression (ipls): a comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, 54(3):413–419.
- Padgett, C. and Saad, A. (2009). Genetic algorithms in chemistry: Success or failure is in the genes. In *Applications of Soft Computing*, pages 181–189. Springer.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Reeves, C. and Rowe, J. (2003). *Genetic algorithms: principles and perspectives: a guide to GA theory*, volume 20.

Springer.

- Reynes, C., Souza, S., Sabatier, R., Figueres, G., and Vidal, B. (2006). Selection of discriminant wavelength intervals in nir spectrometry with genetic algorithms. *Journal of Chemometrics*, 20(3-4):136–145.
- Shariati-Rad, M. and Hasani, M. (2010). Selection of individual variables versus intervals of variables in pls. *Journal of Chemometrics*, 24(2):45–56.
- Tamaki, M., Kihara, R., Okuda, M., Aramaki, I., Katsuba, Z., and Tsuchiya, T. (2005). Properties of starch and protein of hattan-type varieties of rice suitable for brewing original hiroshima sake. *Plant production science*, 8(5):586–591.
- Thomas, D., Atwell, W., and of Cereal Chemists, A. A. (1999). *Starches*. Eagan Press Minnesota.
- Tonukari, N. (2004). Cassava and the future of starch. *Electronic journal of biotechnology*, 7(1):5–8.
- van Soest, J., Tournois, H., de Wit, D., and Vliegthart, J. (1995). Short-range structure in (partially) crystalline potato starch determined with attenuated total reflectance fourier-transform ir spectroscopy. *Carbohydrate Research*, 279:201–214.
- Wold, S., Johansson, E., and Cocchi, M. (1993). Pls-partial least squares projections to latent structures. *3D QSAR in drug design*, 1:523–550.
- Zou, X., Zhao, J., and Li, Y. (2007). Selection of the efficient wavelength regions in ft-nir spectroscopy for determination of ssc of fuji apple based on bpls and fipls models. *Vibrational spectroscopy*, 44(2):220–227.

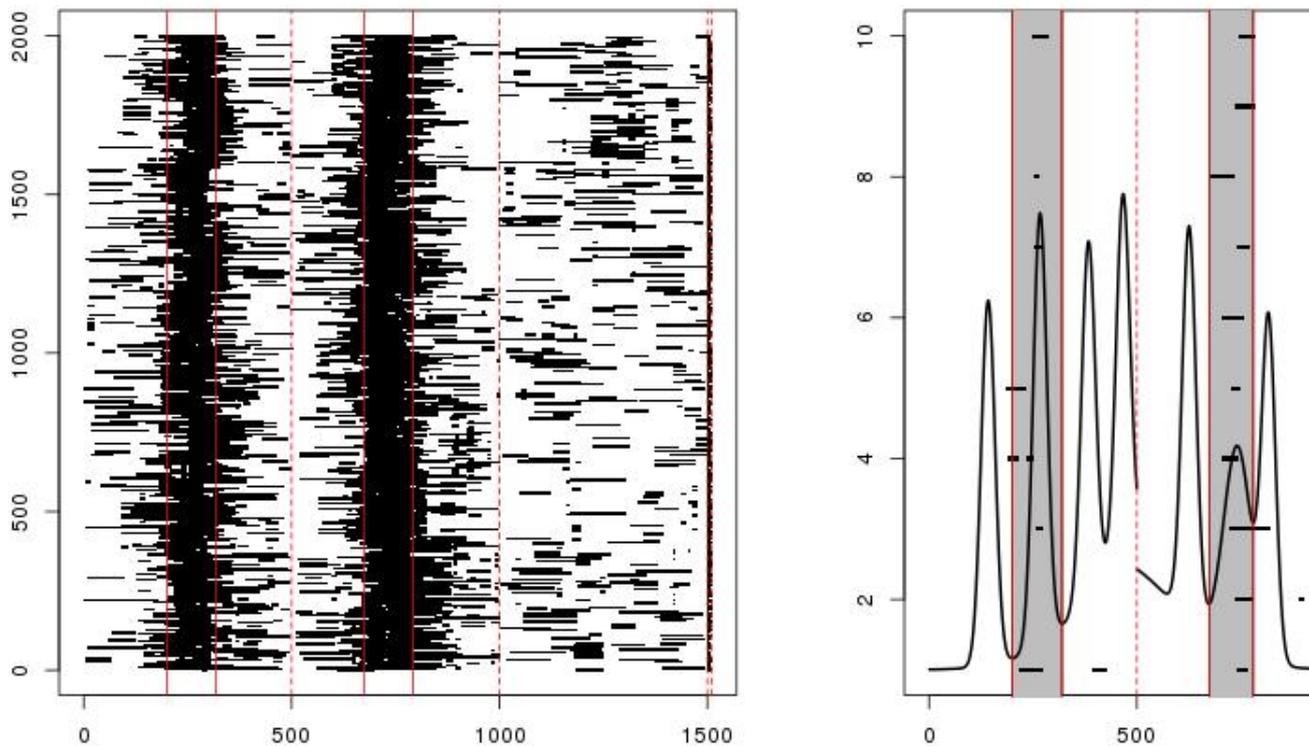


FIGURE 1. Representation of the final solutions of selGamPLS for the simulated dataset. On the left panel, selected variables (depicted in black points) in the ten final populations (that is to say 2000 solutions) are given on the vertical axis. The horizontal axis gives the variable numbers (from 1 to 1510). Blocks are separated by vertical dotted lines. The target intervals are surrounded by vertical solid lines. On the right panel, only the best solution of each run is represented in the same way. The average spectra of the first three blocks are also given.

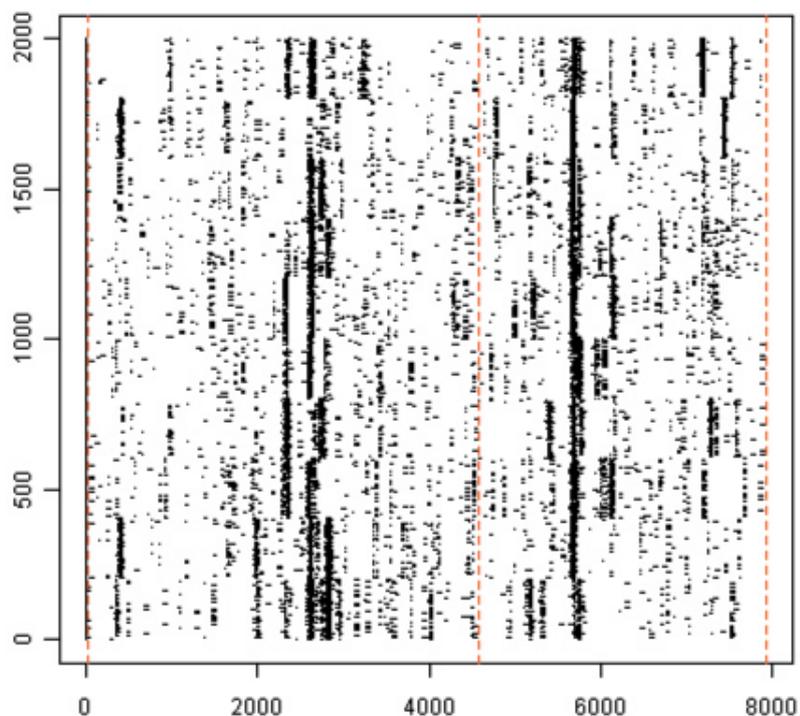


FIGURE 2. Representation of selected variables (depicted in black points) in the ten final populations (that is to say 2000 solutions on the vertical axis). The horizontal axis gives the variable numbers (from 1 to 7926 like in solution encoding). Blocks are separated by vertical dotted lines.

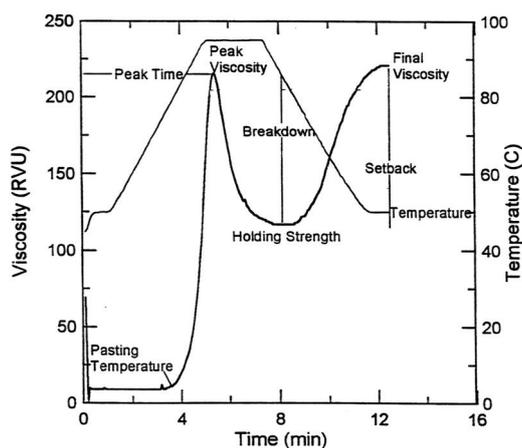


FIGURE 3. General pasting profile of starch using RVA and showing typical pasting parameters.

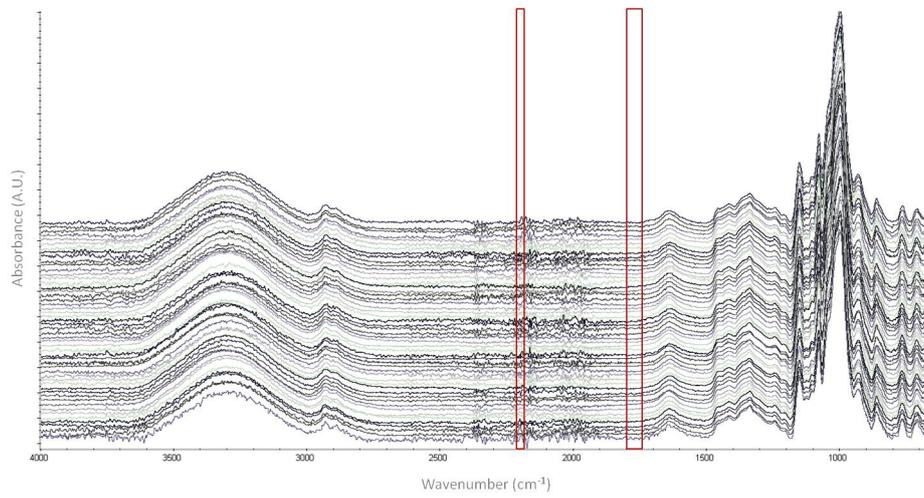


FIGURE 4. *MidIR spectra of cassava starch samples and selected wavenumber regions with selGAmPLS.*

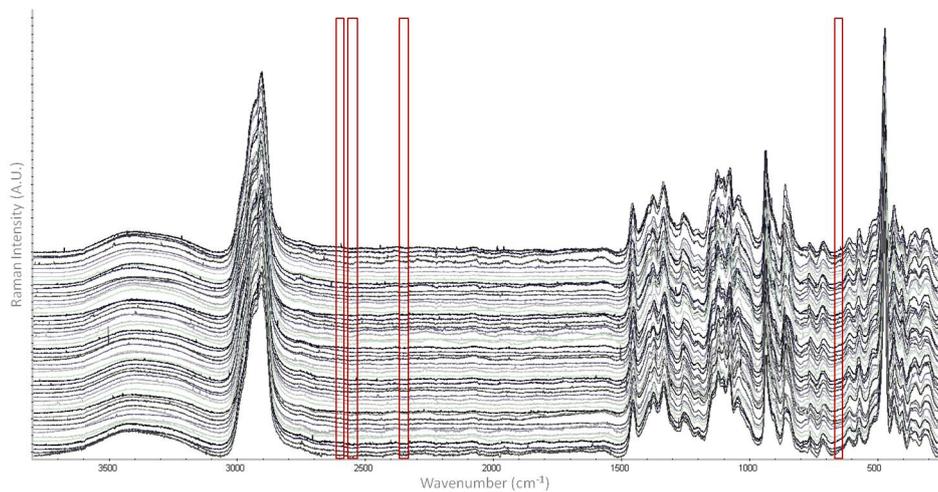


FIGURE 5. *Raman spectra of cassava starch samples and selected wavenumber regions with selGAmPLS.*