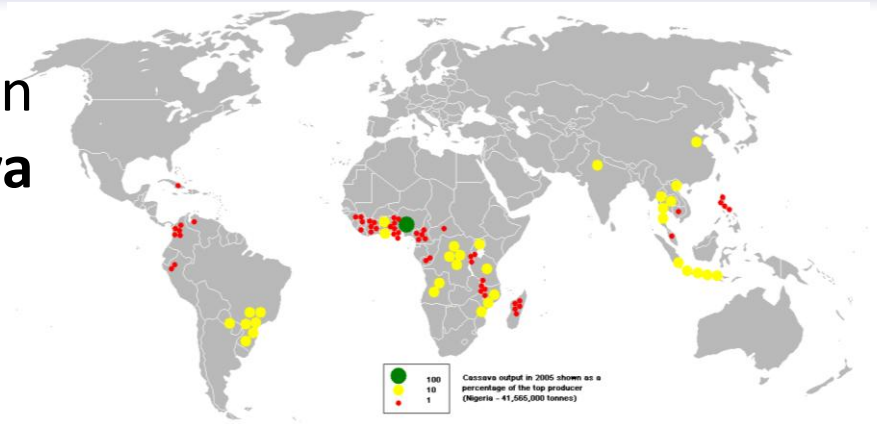
Joint selection of wavelength regions for mid-IR and Raman spectra and variables in PLS regression using Genetic Algorithms

Lidwine Grosmaire[1], Christelle Reynès[2] & Robert Sabatier[2]

[1] *Laboratoire de Physique Moléculaire et Structurale - UMR Qualisud*
Université Montpellier 1 - France
[2] *Laboratoire de Physique Industrielle et Traitement de l'Information - EA 2415*
Université Montpellier 1 - France

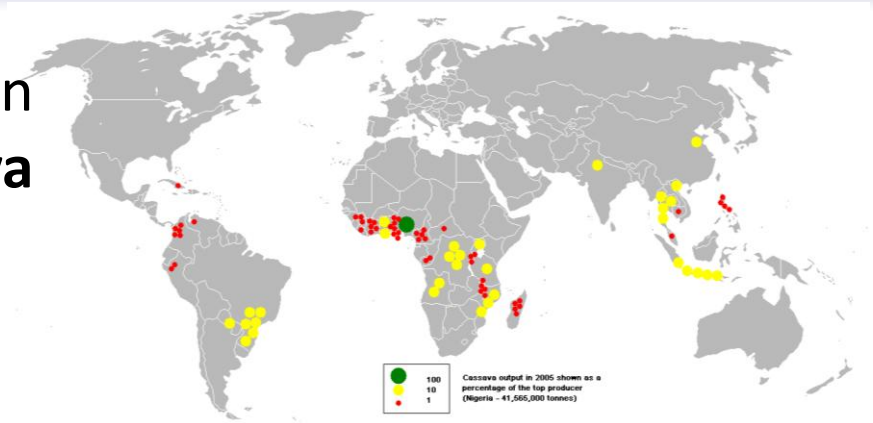
The Context

Increased production and consumption of **cassava**



The Context

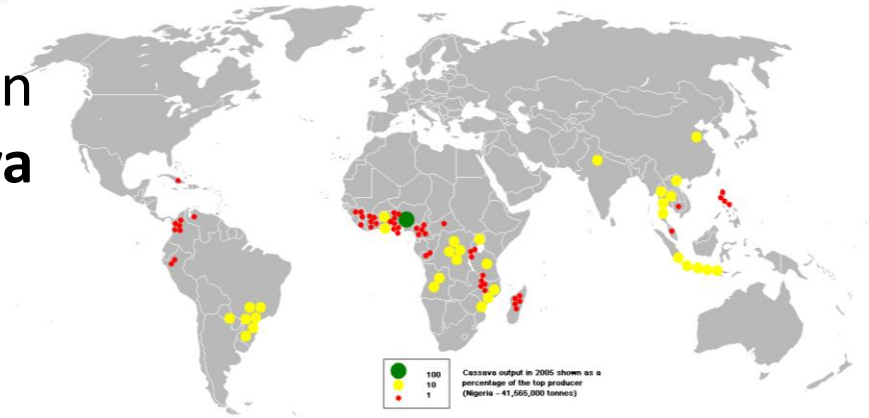
Increased production and consumption of **cassava**



From the crop to the starch

The Context

Increased production and consumption of **cassava**



From the crop to the starch

Empirical and small-scale processing

Interest

Good breadmaking ability

The Aims

Physicochemical parameters

↓

Varietal and process impacts on breadmaking ability

The Aims

Physicochemical parameters

↓

Varietal and process impacts on breadmaking ability

Aim

✓Standardize and scale-up the process

✓Improve product quality

✓Industrial development of new gluten-free bread products

The Data

13 varieties x 4 treatments = 52 samples

The Data

13 varieties x 4 treatments = 52 samples

1. **Physicochemical parameters**

The Data

13 varieties x 4 treatments = 52 samples

1. Physicochemical parameters

Breadmaking ability

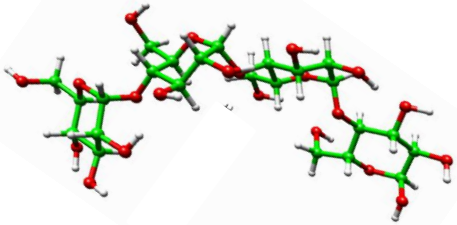
The Data

13 varieties x 4 treatments = 52 samples

1. Physicochemical parameters

Breadmaking ability

Amylose content



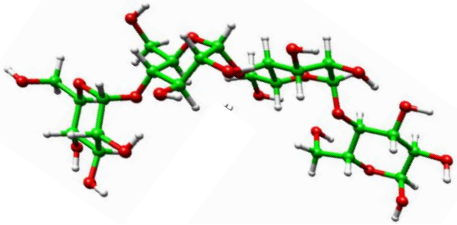
The Data

13 varieties x 4 treatments = 52 samples

1. Physicochemical parameters

Breadmaking ability

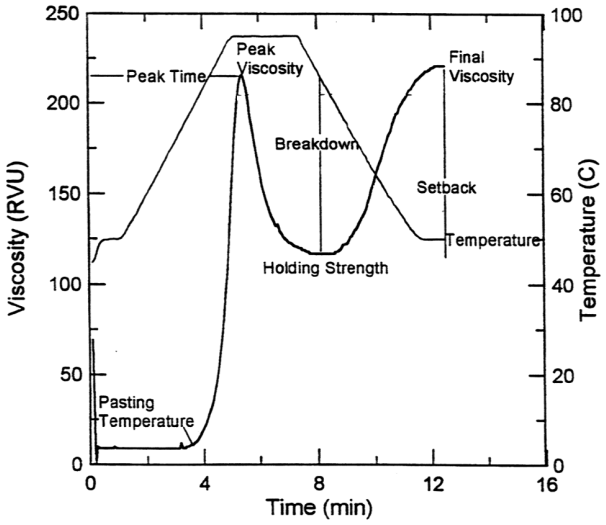
Amylose content



Rapid Visco Analyzer

Pasting behavior

12 variables

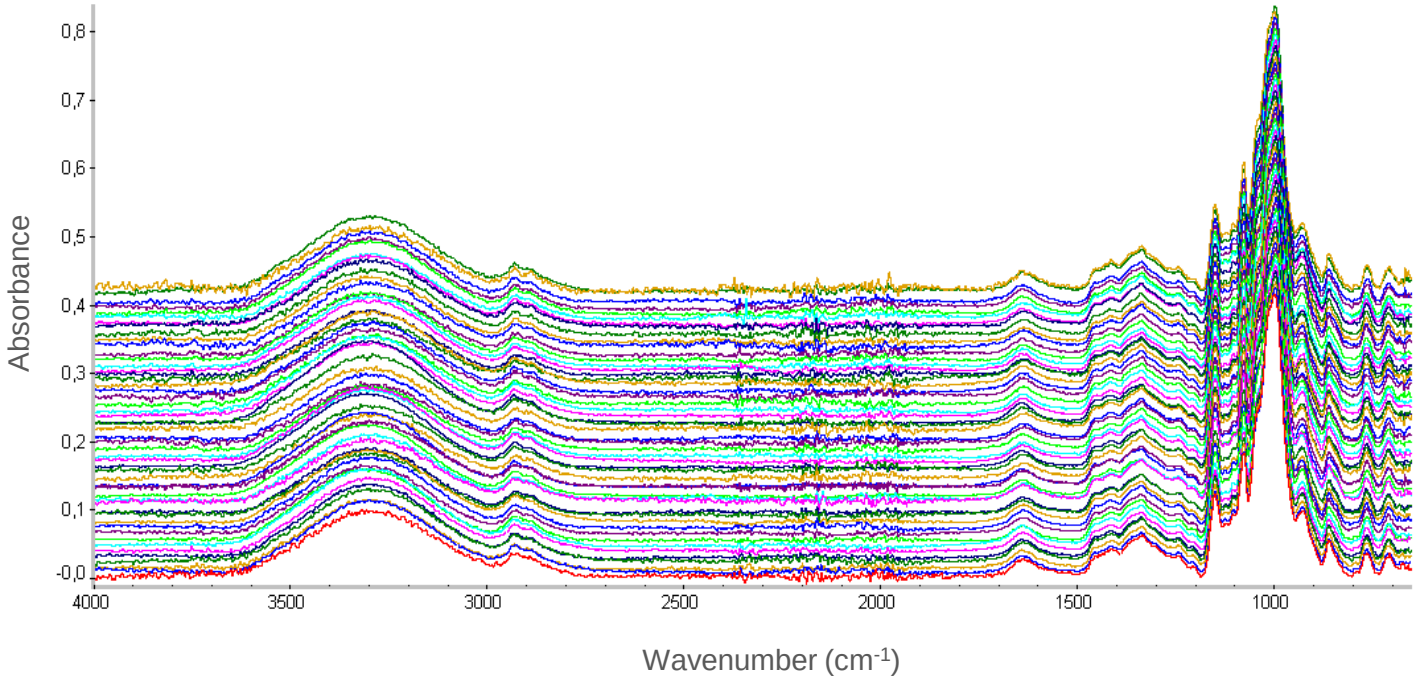


The Data

2. Spectral data

2. Spectral data

Mid-infrared spectra

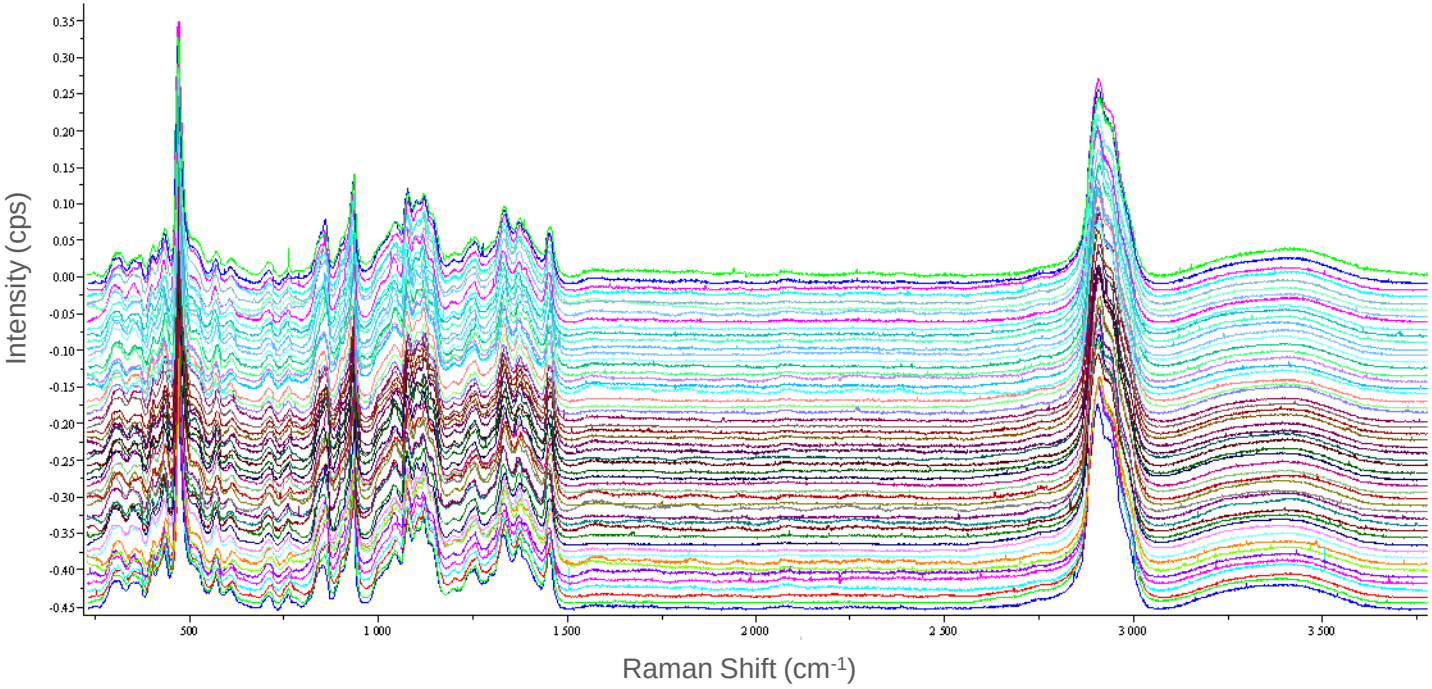


650 – 4000 cm^{-1} ➡ 3351 variables

The Data

2. Spectral data

Raman spectra



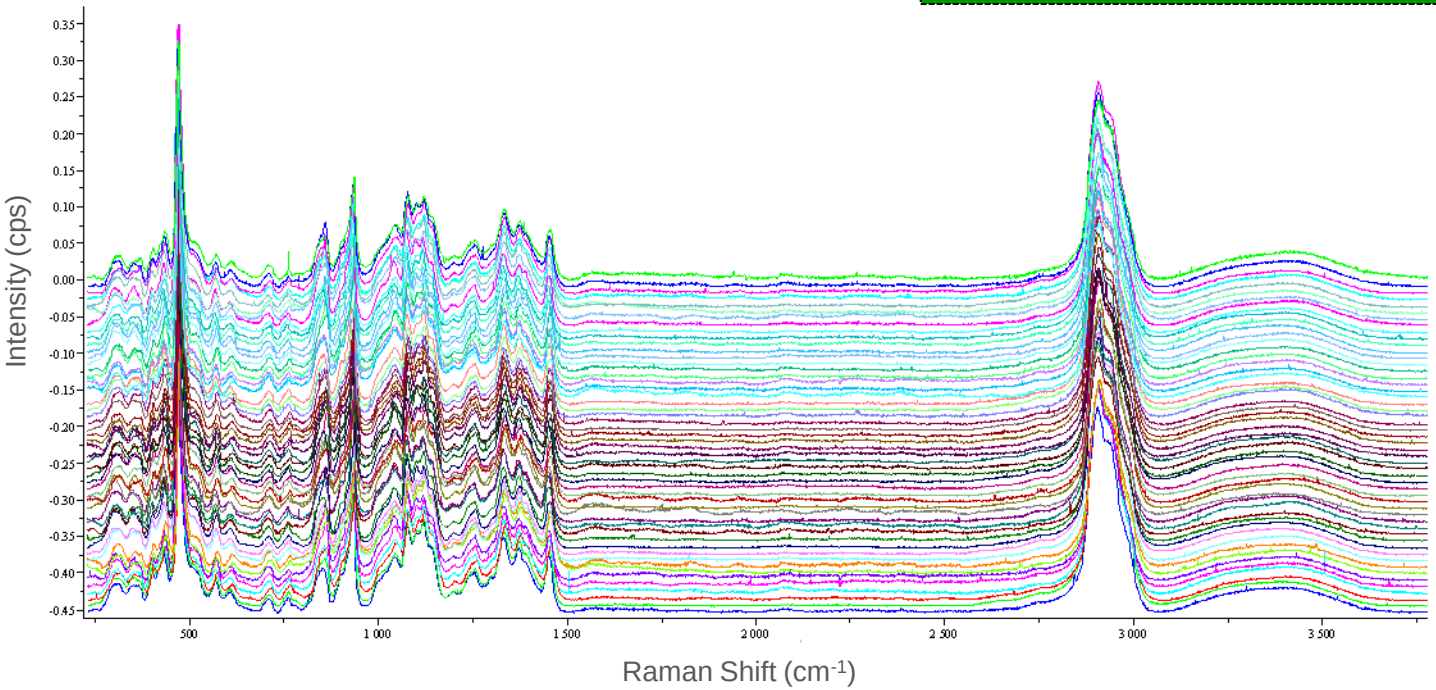
230 – 3800 cm^{-1} ➡ 4562 variables

The Data

2. Spectral data

Raman spectra

Raman + IR
Pre-treatment
Baseline + SNV



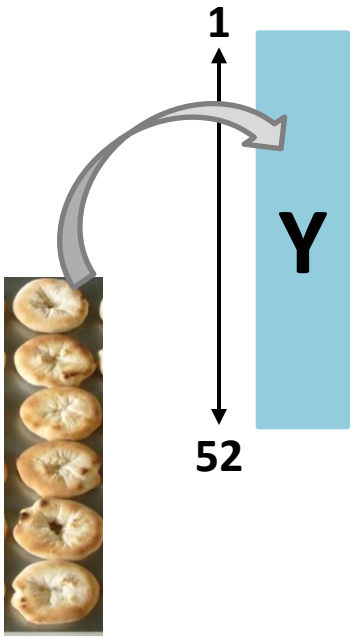
230 – 3800 cm^{-1} ➡ 4562 variables

The Problematic

Data → **Predicting the expansion ability**

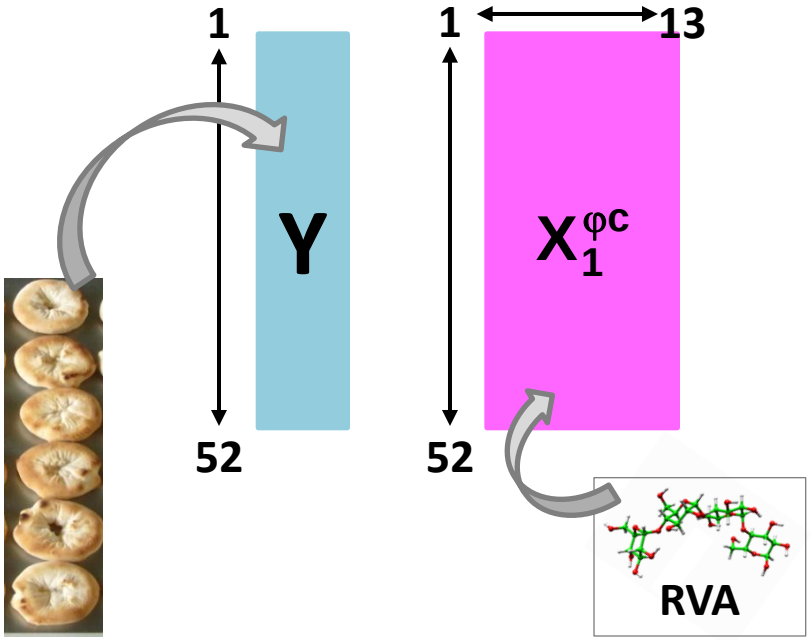
The Problematic

Data → **Predicting the expansion ability**



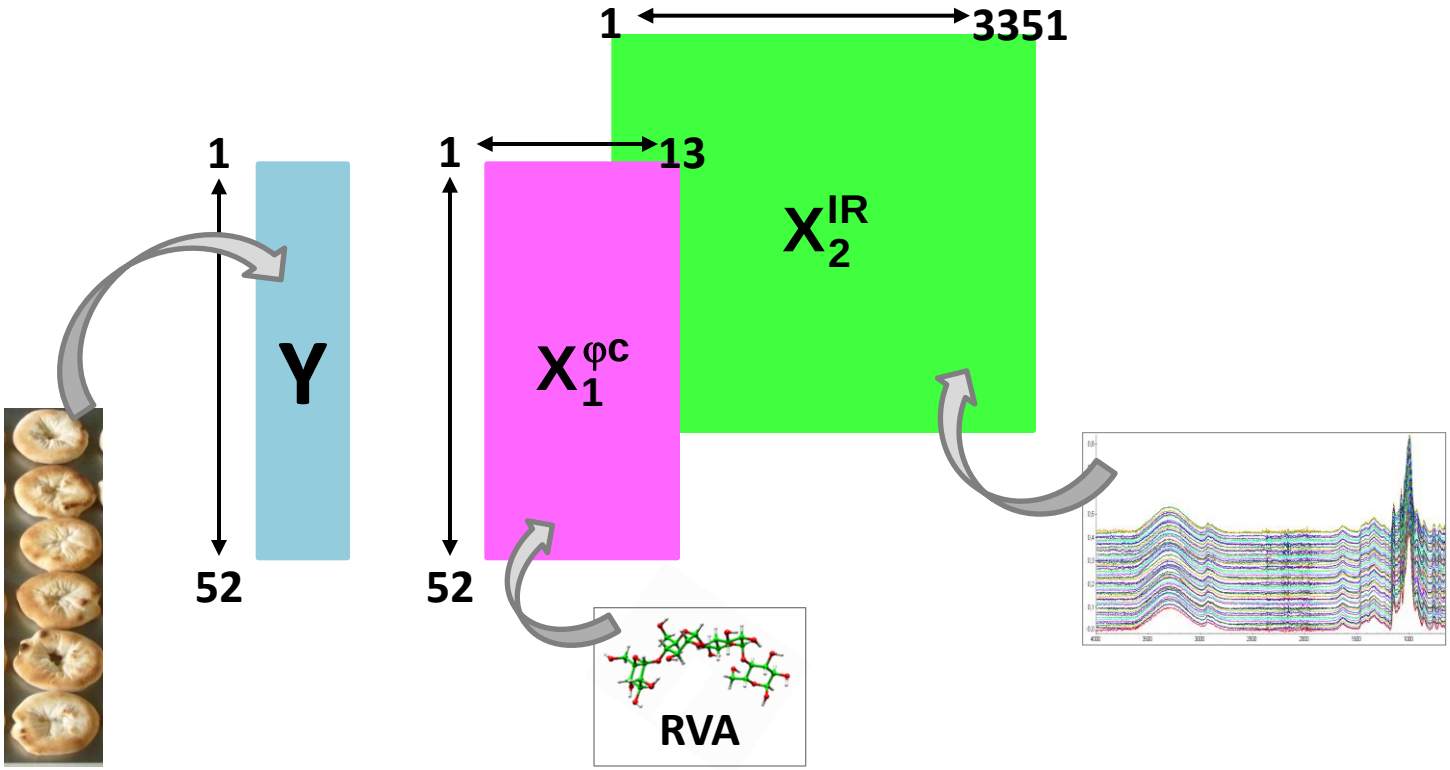
The Problematic

Data → **Predicting the expansion ability**



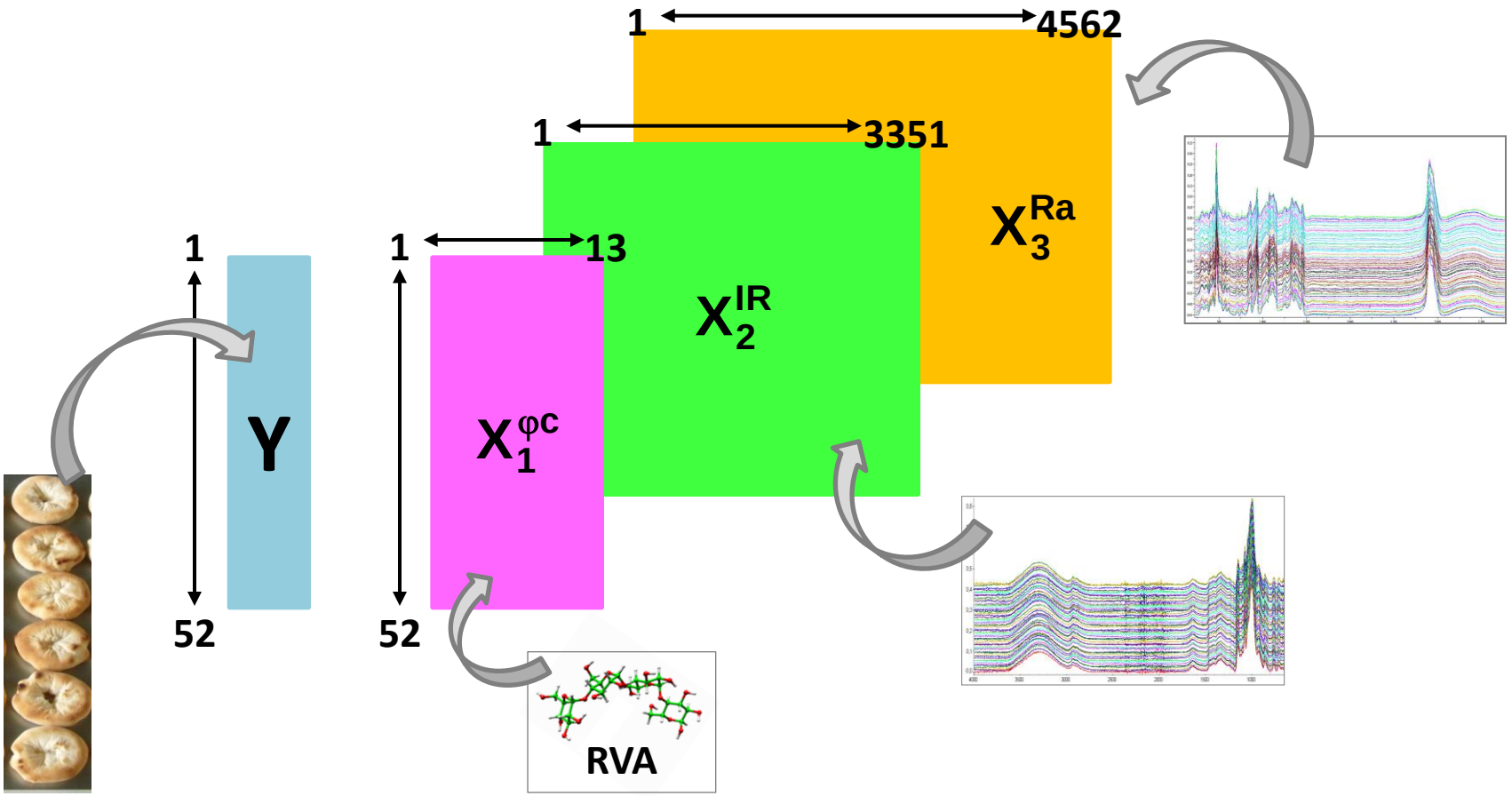
The Problematic

Data ➡ **Predicting the expansion ability**



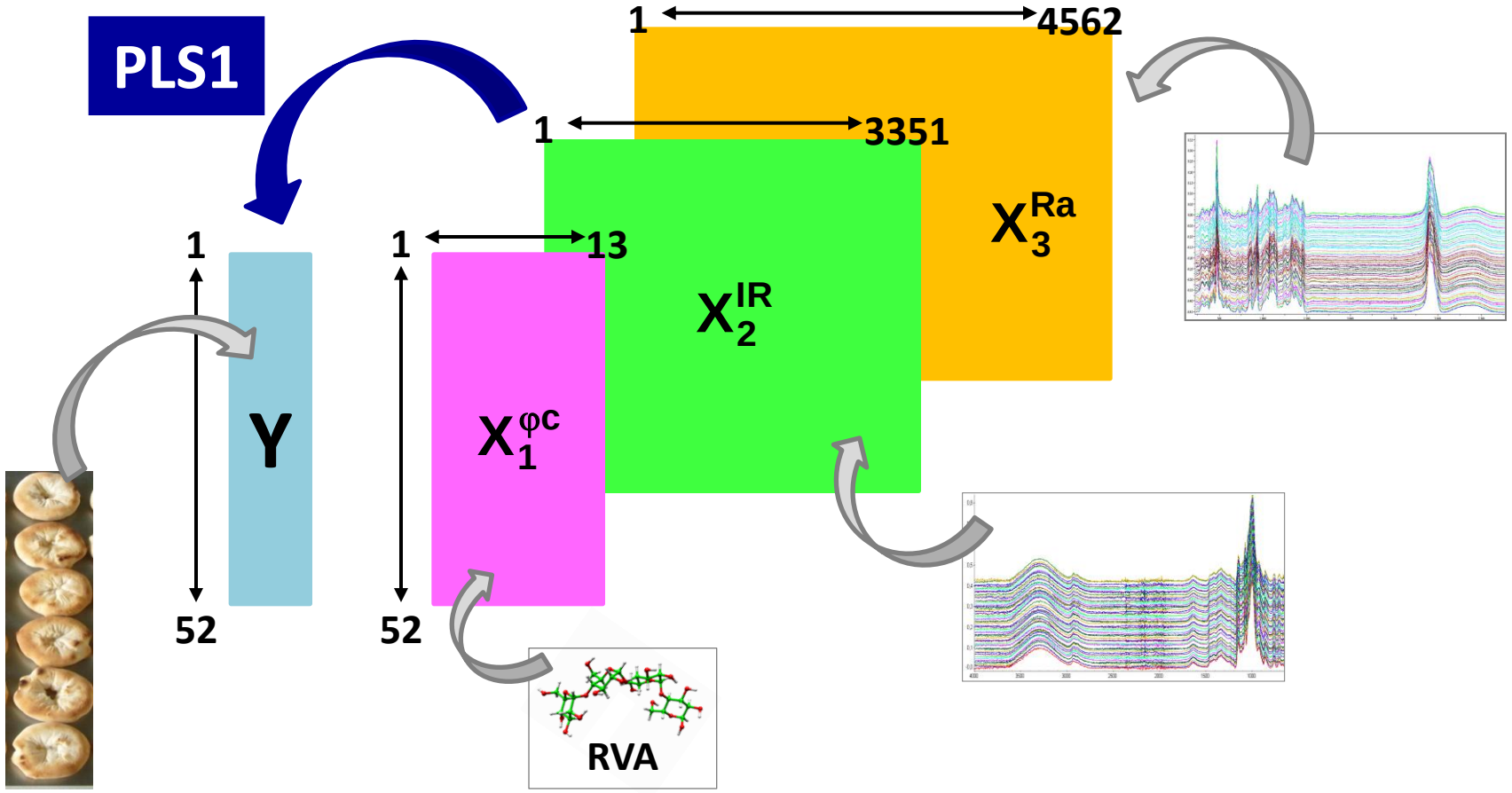
The Problematic

Data → **Predicting the expansion ability**



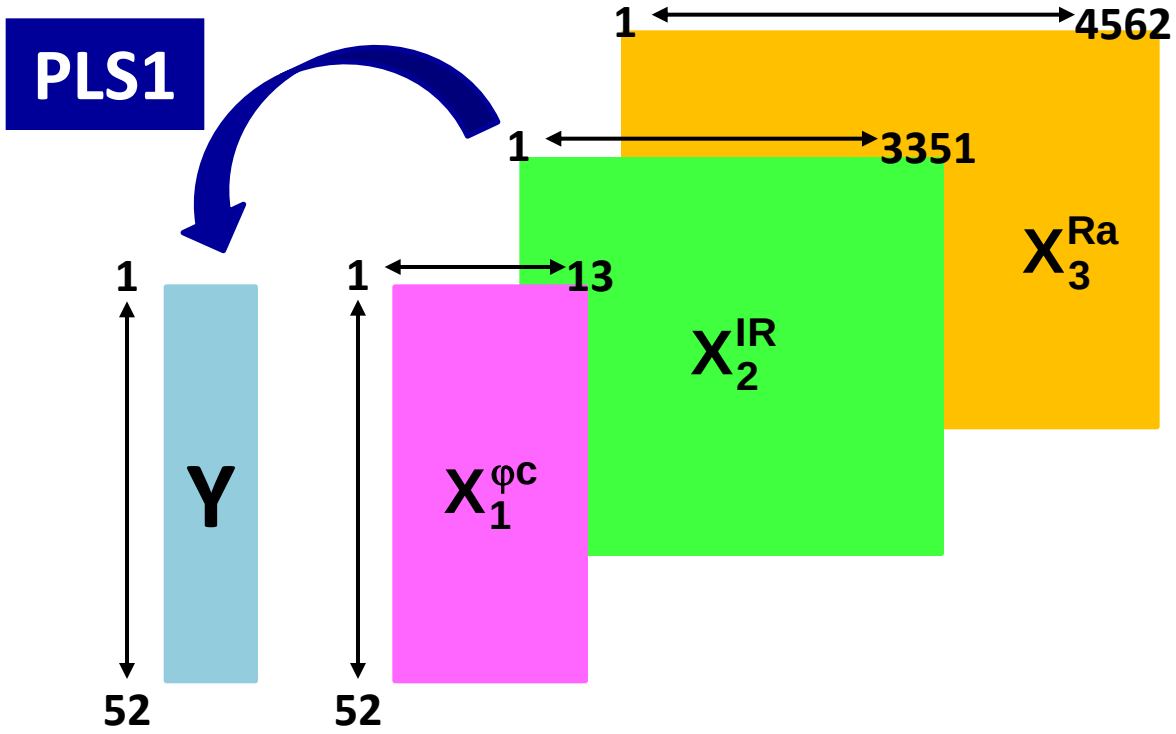
The Problematic

Data → **Predicting the expansion ability**



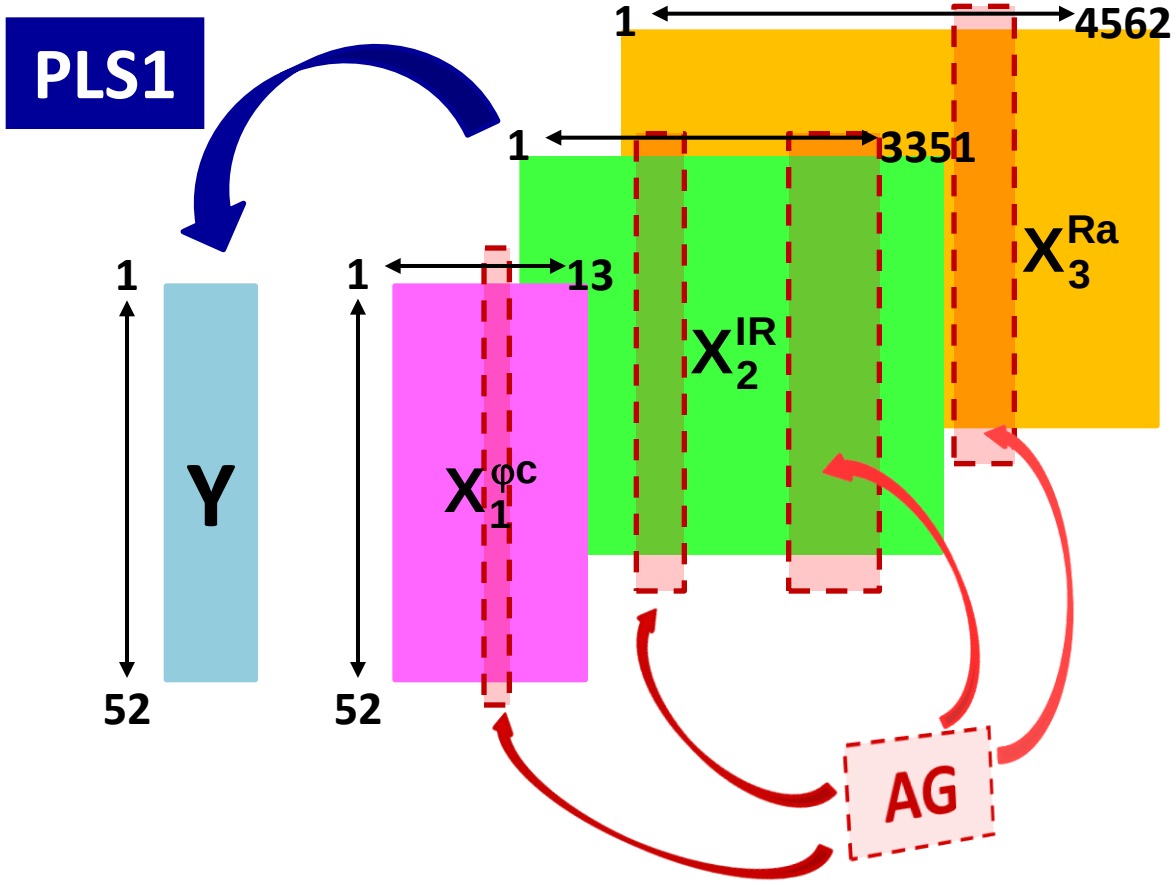
The Problematic

Relevant parameters → **Predicting the expansion ability**



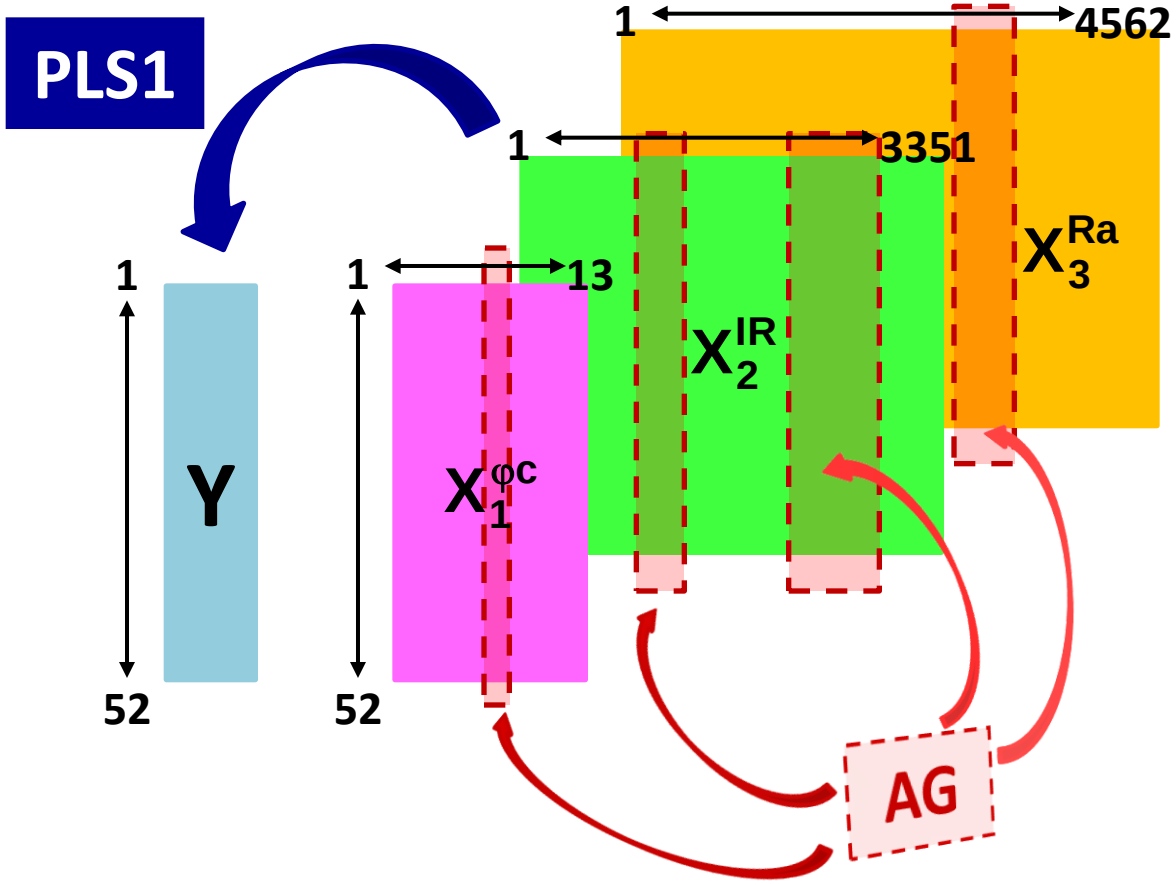
The Problematic

Relevant parameters → **Predicting the expansion ability**



The Problematic

Relevant parameters ➡ **Predicting the expansion ability**



➡AGvPLSm

Genetic Algorithm

➡ optimization method based on the evolution of a « **population** » of potential solutions thanks to genetic mechanisms

Genetic Algorithm

➡ optimization method based on the evolution of a « **population** » of potential solutions thanks to genetic mechanisms

General unfolding:

❶ Building of an initial population whose characteristics are as diverse as possible

Genetic Algorithm

optimization method based on the evolution of a « **population** » of potential solutions thanks to genetic mechanisms

General unfolding:

❶ Building of an initial population whose characteristics are as diverse as possible

❷ Evolution of this population through three operators:

- mutation
- crossover

} Independent of the optimization problem

- selection

} Allows to quantify the solution quality

Genetic Algorithm

optimization method based on the evolution of a « **population** » of potential solutions thanks to genetic mechanisms

General unfolding:

❶ Building of an initial population whose characteristics are as diverse as possible

❷ Evolution of this population through three operators:

- mutation
- crossover

} Independent of the optimization problem

- selection

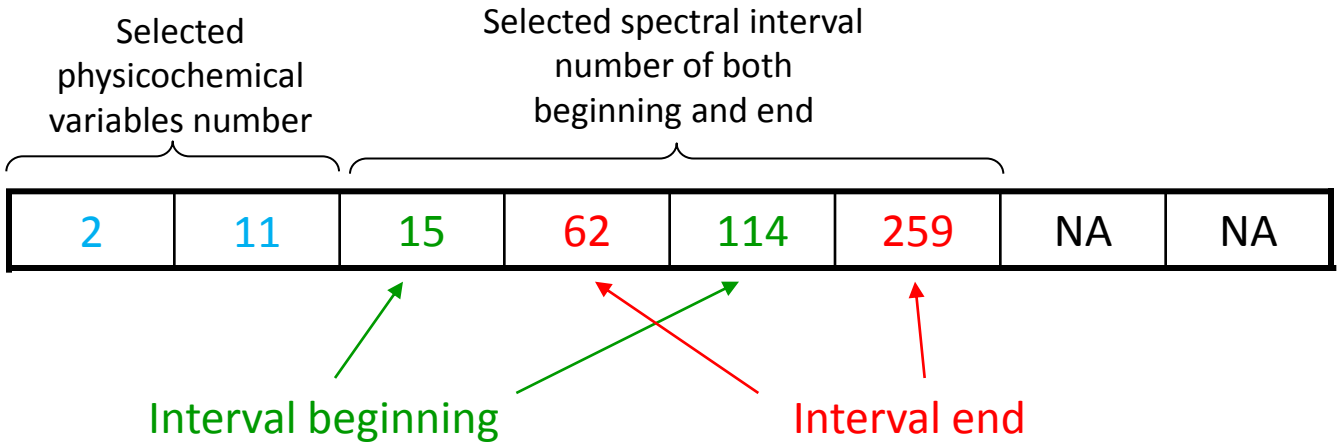
} Allows to quantify the solution quality

❸ Final population obtained after convergence criterion met

Genetic algorithm application to feature selection

One solution = a subset of variables (individual physicochemical variables + spectral intervals)

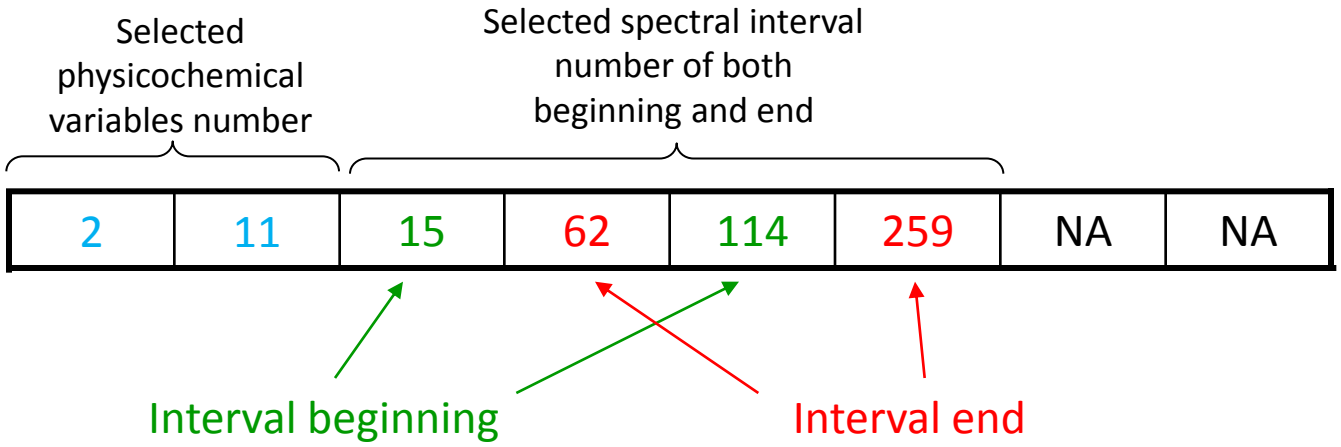
- **Individual encoding:**



Genetic algorithm application to feature selection

One solution = a subset of variables (individual physicochemical variables + spectral intervals)

- **Individual encoding:**

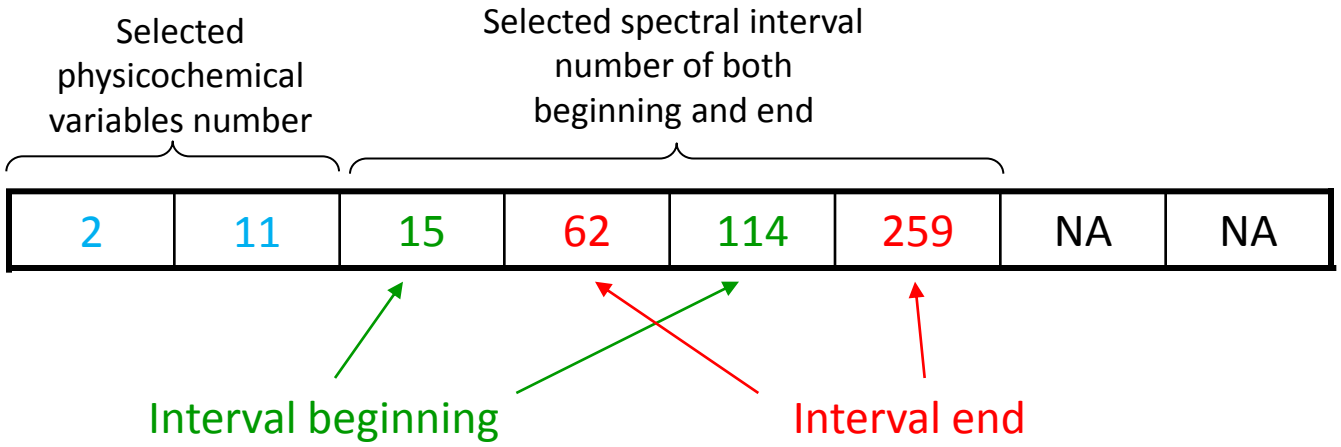


- **Initial population:** random generation of selected variables
  - Choice of the number of selected variables and/or interval between 1 and *Nmax*
  - For each selected variable:
    - Choice of the table in which the next variable is to be chosen

Genetic algorithm application to feature selection

One solution = a subset of variables (individual physicochemical variables + spectral intervals)

- **Individual encoding:**



- **Initial population:** random generation of selected variables

  - Choice of the number of selected variables and/or interval between 1 and *Nmax*

  - For each selected variable:

    - Choice of the table in which the next variable is to be chosen

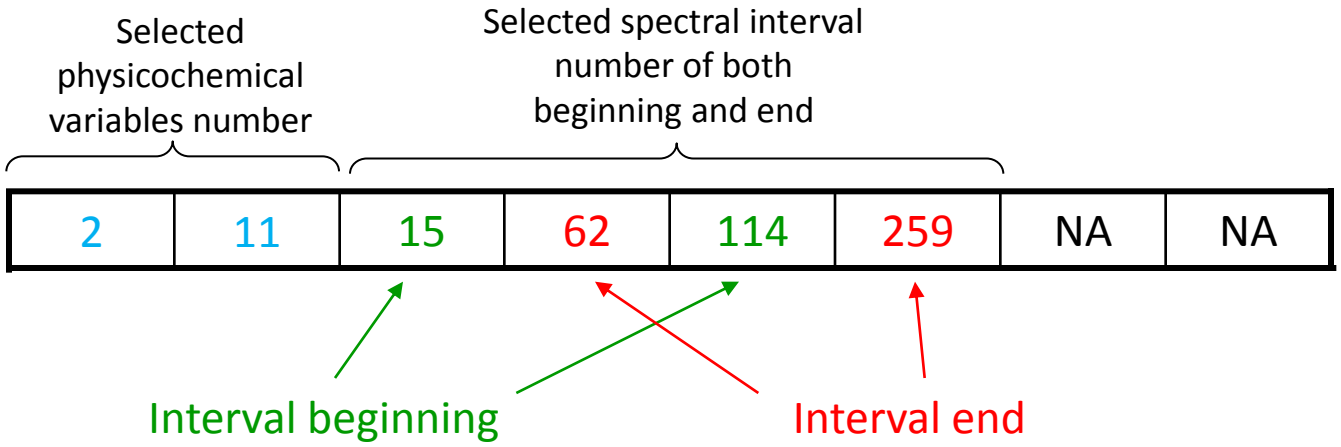
physicochemical table → one variable

spectral tables → one interval

Genetic algorithm application to feature selection

One solution = a subset of variables (individual physicochemical variables + spectral intervals)

- **Individual encoding:**



- **Initial population:** random generation of selected variables

  - Choice of the number of selected variables and/or interval between 1 and *Nmax*

  - For each selected variable:

    - Choice of the table in which the next variable is to be chosen

Genetic algorithm application to feature selection

- **Mutation operator:** allows to explore the solution space
  - Add
  - Delete
  - Relocate

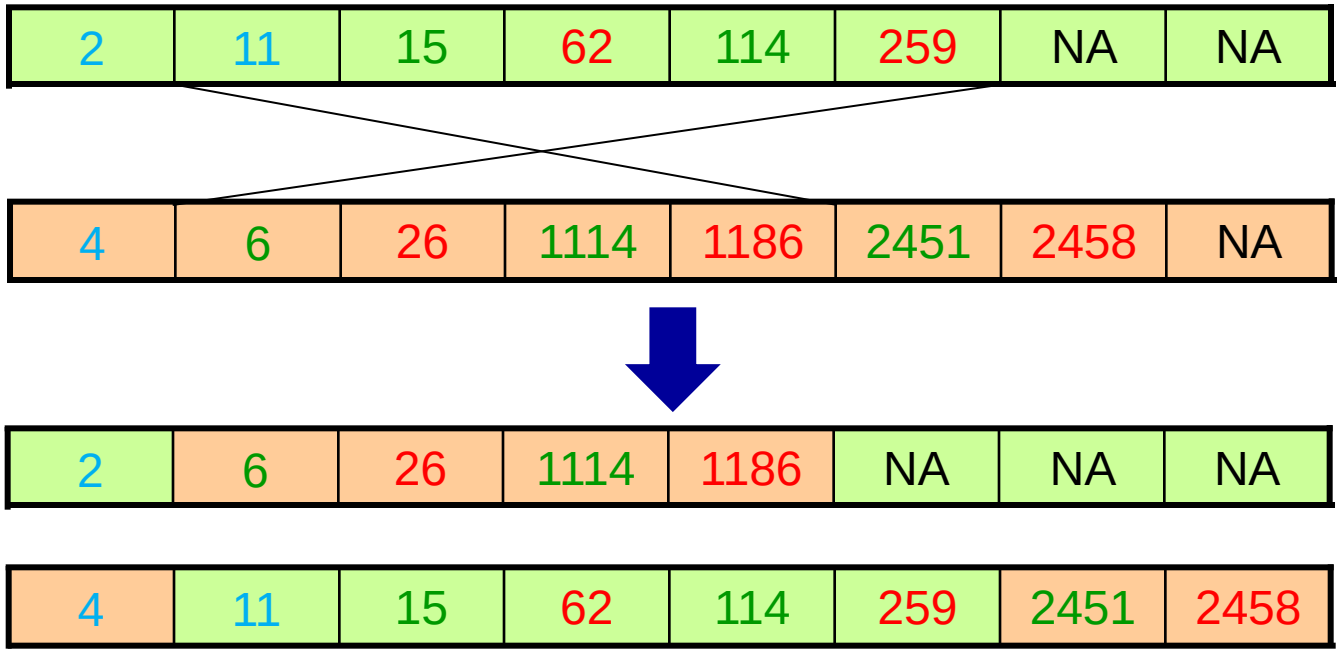
  } an interval and/or variable

Genetic algorithm application to feature selection

- **Mutation operator:** allows to explore the solution space
  - Add
  - Delete
  - Relocate

  } an interval and/or variable

- **Crossover**: allows to combine characteristics of previously selected individuals



Genetic algorithm application to feature selection

- **Fitness function:** quantification of solution quality

$$fitness = \underbrace{R^2_{CV}\left(y, \hat{y}\right)}_{\text{Model precision}} + \underbrace{c \times \left(\alpha\left(N_{\text{varsel}} + \beta\right)\right)}_{\text{Model parsimony}}$$

$R^2_{CV}\left(y, \hat{y}\right)$: linear correlation coefficient between the real and predicted capacity of panification

N_{varsel} : number of selected variables and/or interval

c, α, β : normalisation parameters

Genetic algorithm application to feature selection

- **Fitness function:** quantification of solution quality

$$fitness = \underbrace{R^2_{CV}\left(y, \hat{y}\right)}_{\text{Model precision}} + \underbrace{c \times \left(\alpha\left(N_{\text{varsel}} + \beta\right)\right)}_{\text{Model parsimony}}$$

$R^2_{CV}\left(y, \hat{y}\right)$: linear correlation coefficient between the real and predicted capacity of panification

N_{varsel} : number of selected variables and/or interval

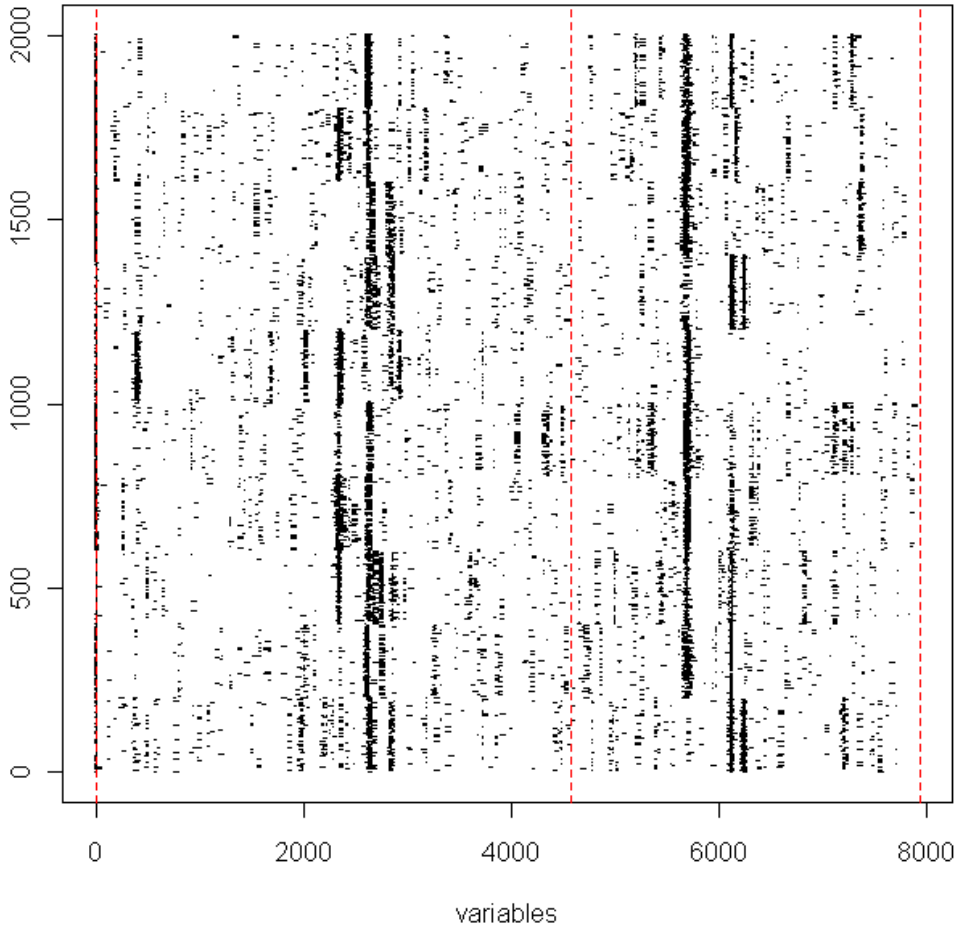
c, α, β : normalisation parameters

- **Selection operator:**
  - Ranking of solutions according to the fitness (the best one having the highest rank)
  - Compute selection probability for each solution : $p(\#k) = \gamma\, k + \nu$

    γ and ν are fixed so that the best solution selection probability is twice higher than the median solution one and the sum of all selection probabilities is one
  - Build a new population according to these probabilities with constant size

Results

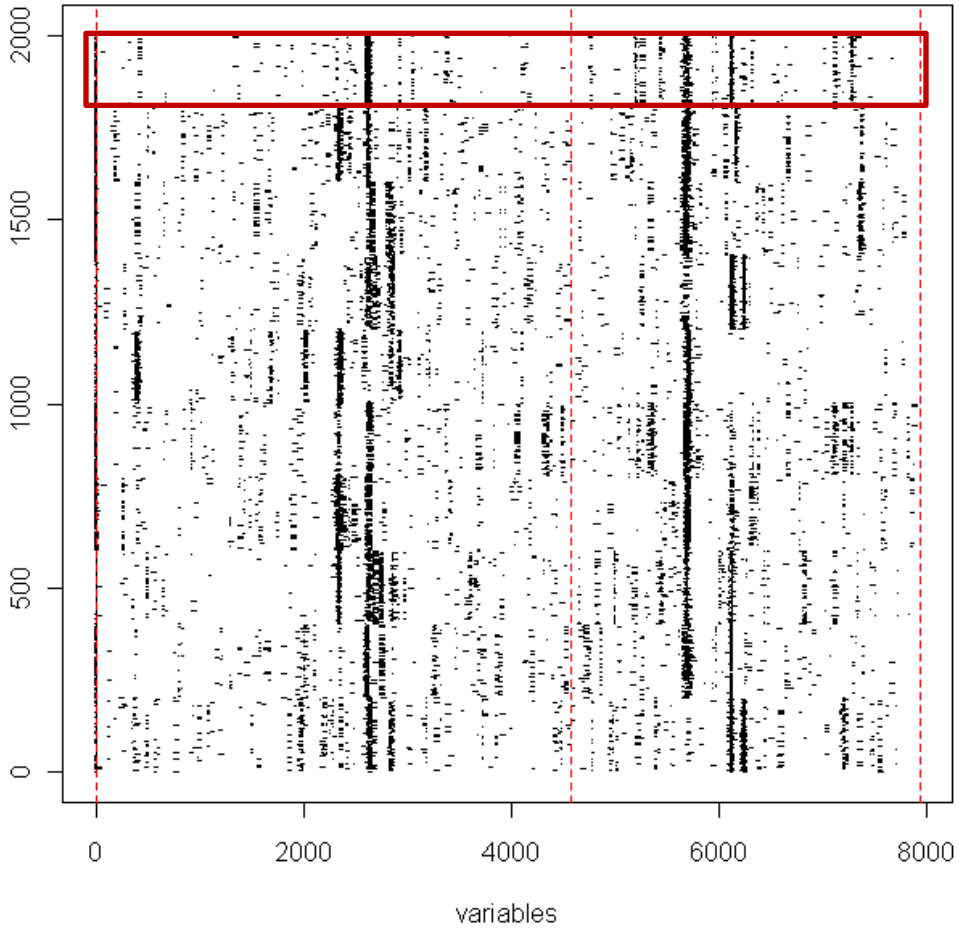
GA results for 10 runs with population of size 200 and 100 generations:



Final populations characteristics:
selected variables are indicated by black points

Results

GA results for 10 runs with population of size 200 and 100 generations:

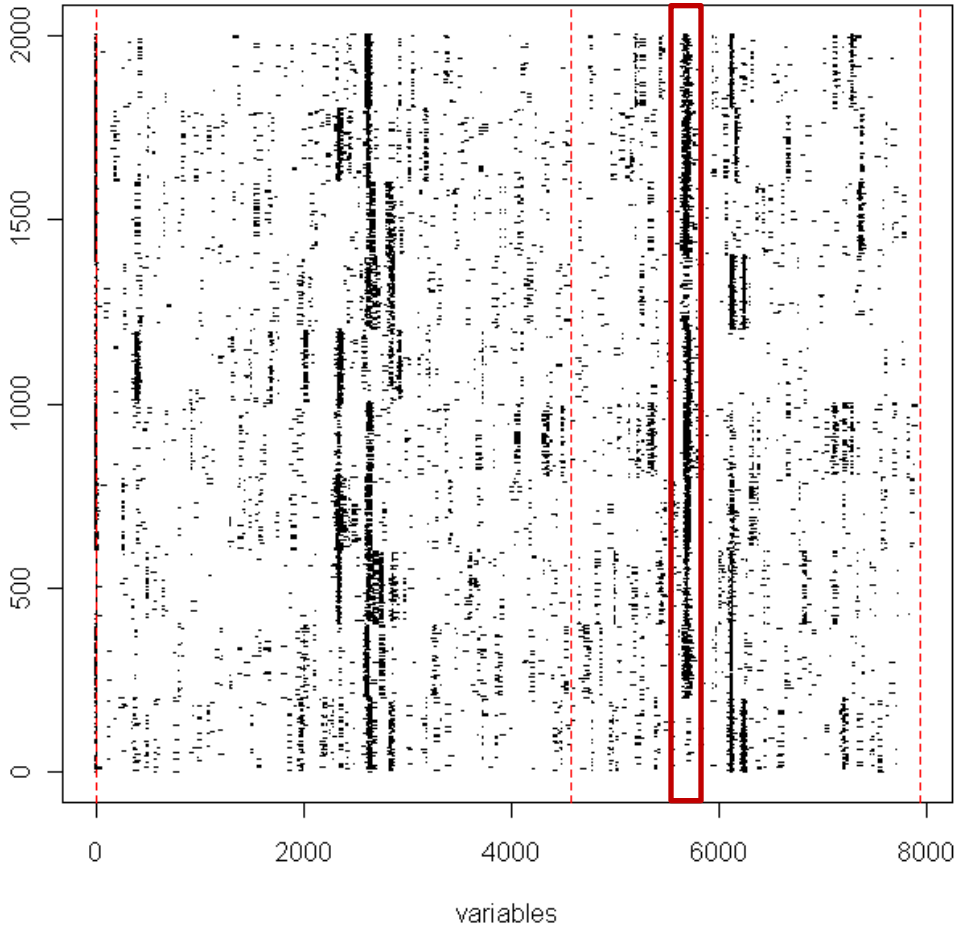


Final populations characteristics:
selected variables are indicated by black points

Individual populations seem to have converged

Results

GA results for 10 runs with population of size 200 and 100 generations:



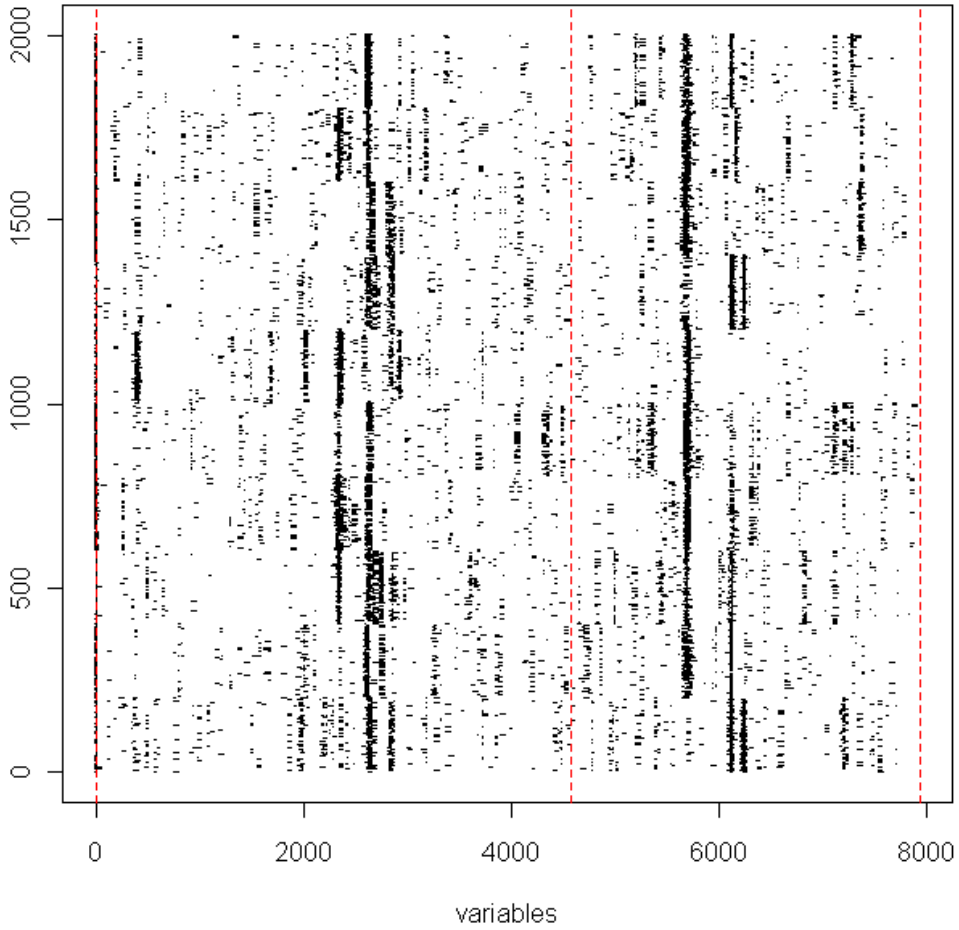
Final populations characteristics:
selected variables are indicated by black points

Individual populations seem to have converged

The 10 final populations are quite close indicating a global convergence of the GA

Results

GA results for 10 runs with population of size 200 and 100 generations:



Final populations characteristics:
selected variables are indicated by black points

Individual populations seem to have converged

The 10 final populations are quite close indicating a global convergence of the GA

Finally, are retained:

4 physicochemical variables
4 RAMAN intervals
2 Mid-IR intervals

$R^2=0.9936$ $R^2_{CV}=0.8273$

Results

Comparison with other methods

Necessity of feature selection ➡ comparison with PLS on all variables

Relevance of feature selection method ➡ comparison with PLS + VIP

Method	# var	# comp	R^2	R^2_{CV}
PLS	7926	7	0.7836	0.6605
PLS + VIP	4	3	0.7210	0.6650
AGvPLSm	311	12	0.9936	0.8273

Comparison of different method results (number of selected variables, number of retained components, R^2 and cross-validation R^2).

PLS — noisy model + quite bad performances

PLS + VIP — only 4 variables (physicochemical ones)

Results

Comparison with other methods

Necessity of feature selection ➡ comparison with PLS on all variables

Relevance of feature selection method ➡ comparison with PLS + VIP

Method	# var	# comp	R^2	R^2_{CV}
PLS	7926	7	0.7836	0.6605
PLS + VIP	4	3	0.7210	0.6650
AGvPLSm	311	12	0.9936	0.8273

Comparison of different method results (number of selected variables, number of retained components, R^2 and cross-validation R^2).

PLS — noisy model + quite bad performances

PLS + VIP — only 4 variables (physicochemical ones)

AGvPLSm — very good model in description + cross-validation by using a reasonable number of variables split among the three tables

Conclusion

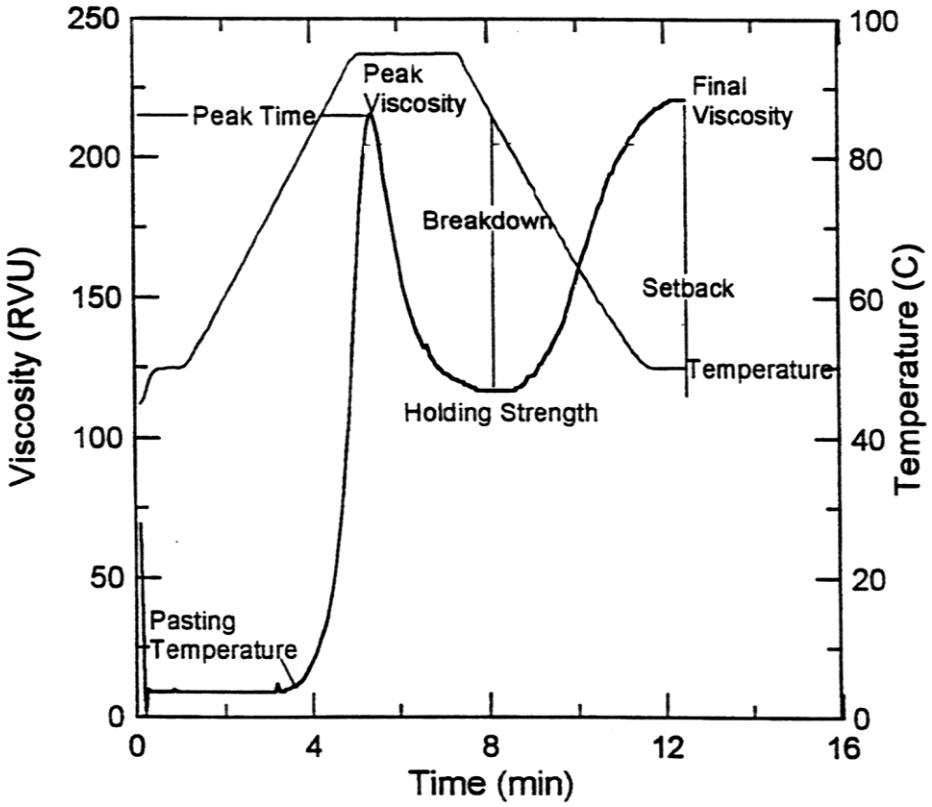
Interpretation of selected variables

1. **Physicochemical parameters**

Conclusion

Interpretation of selected variables

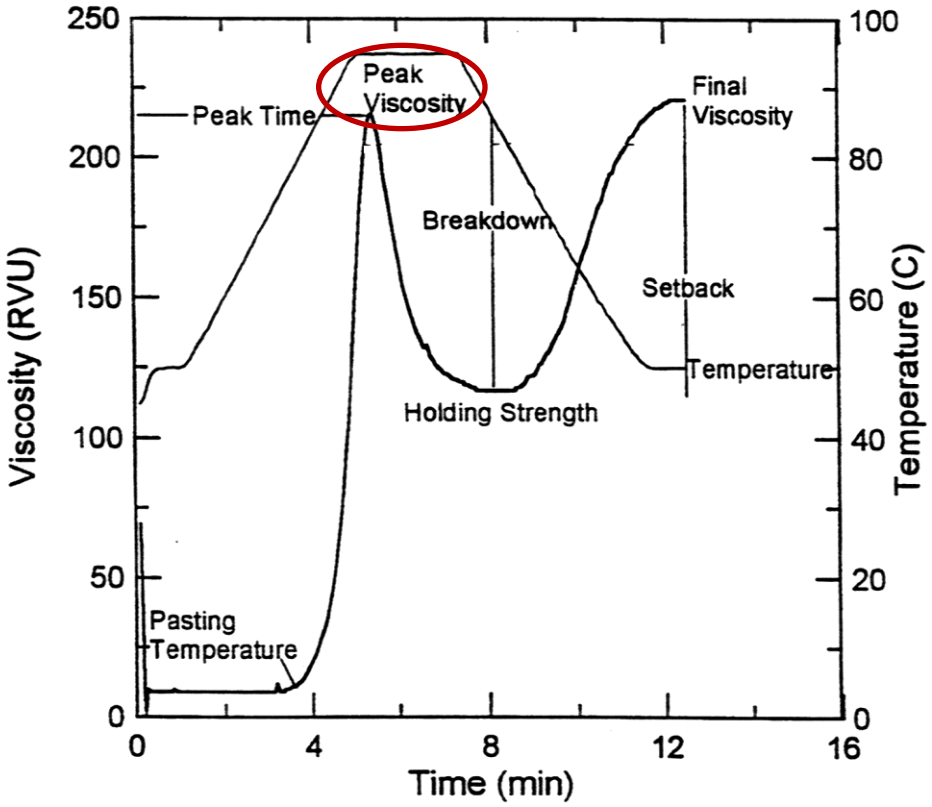
1. **Physicochemical parameters** → RVA



Conclusion

Interpretation of selected variables

1. **Physicochemical parameters** → RVA

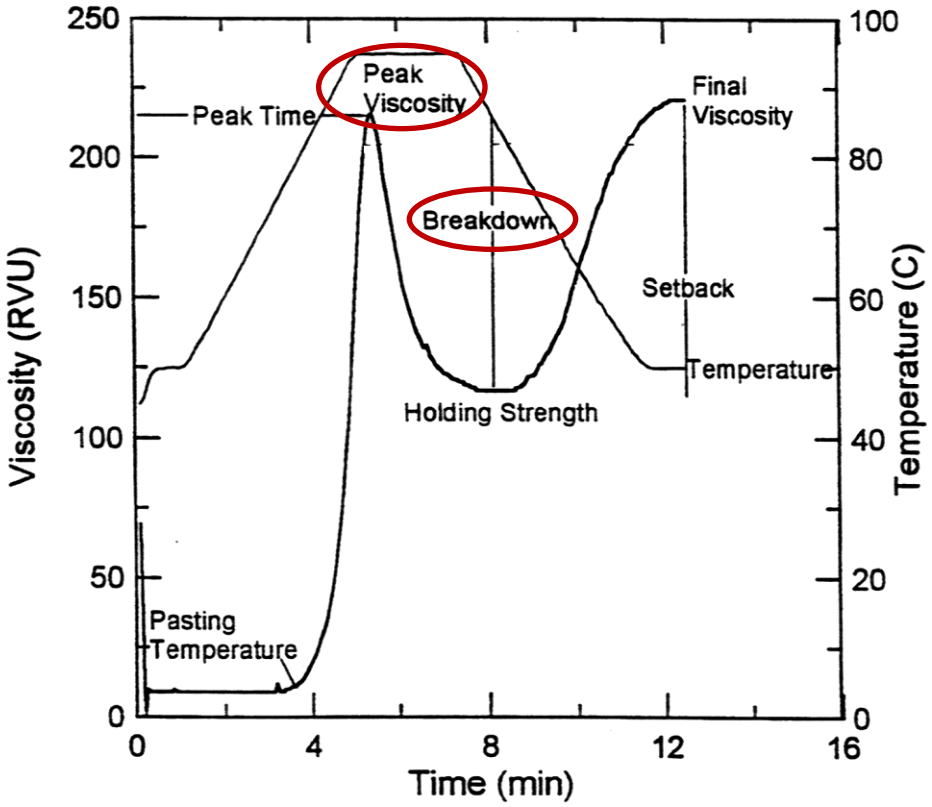


Granule size

Conclusion

Interpretation of selected variables

1. Physicochemical parameters → RVA



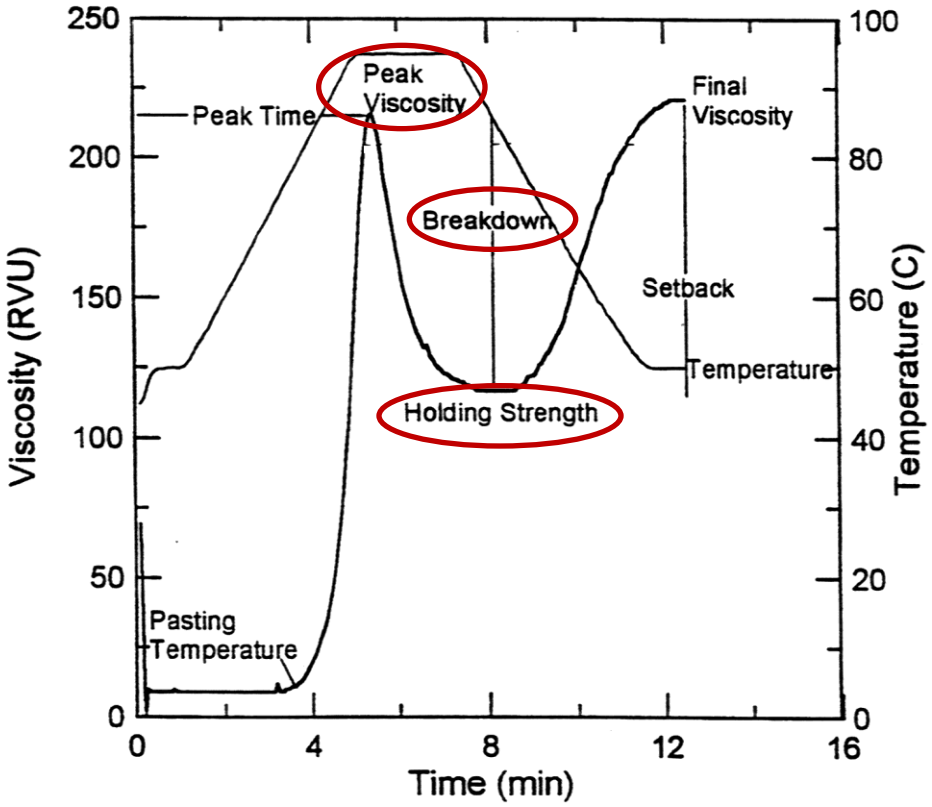
Granule size

Granule weakness

Conclusion

Interpretation of selected variables

1. Physicochemical parameters → RVA



Granule size

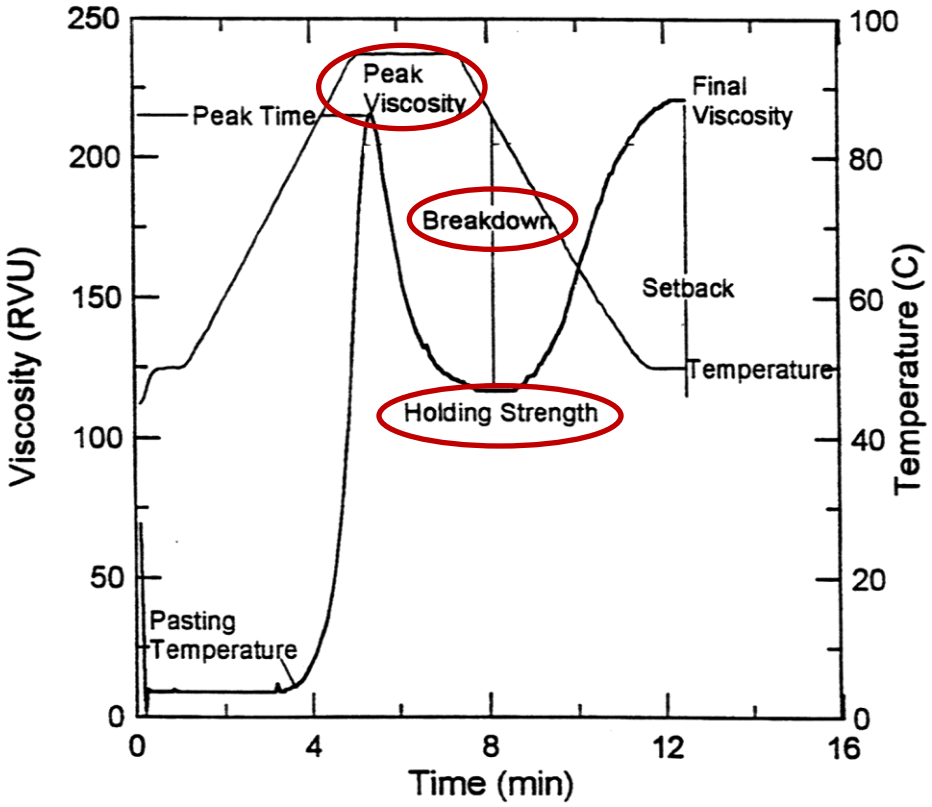
Granule weakness

Lowest paste viscosity

Conclusion

Interpretation of selected variables

1. Physicochemical parameters → RVA



Granule size

Granule weakness

Lowest paste viscosity

↓

Pasting parameters

↓

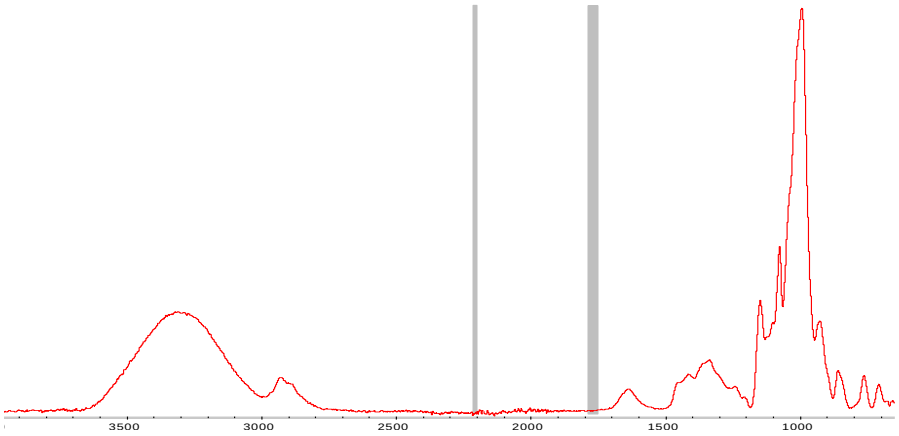
Water absorption

Conclusion

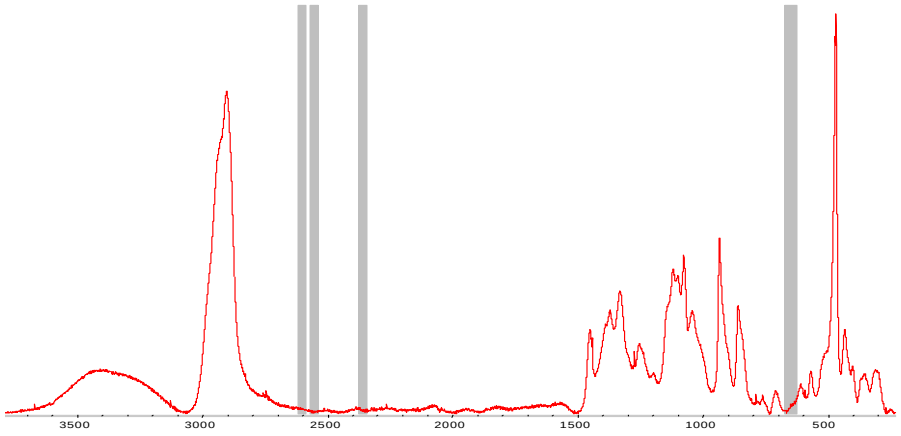
Interpretation of selected variables

1. **Spectral data** Selected variables ⇨ Intervals of wavenumbers

Mid-infrared spectra



Raman spectra

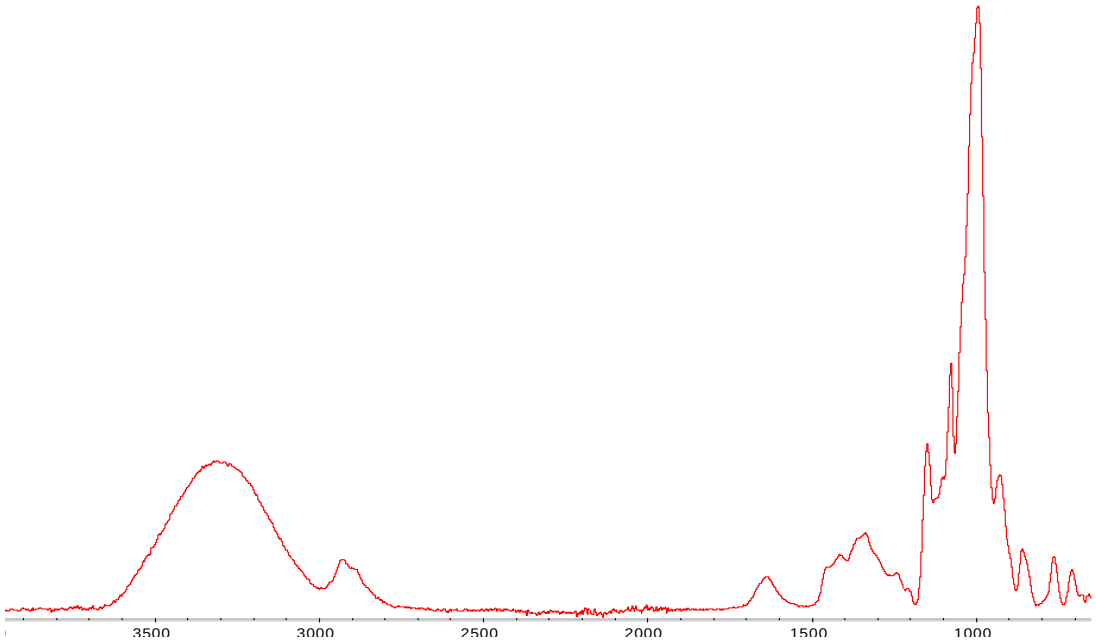


Selected range { No matching with specific vibrational bands
Improve the model

Perspectives

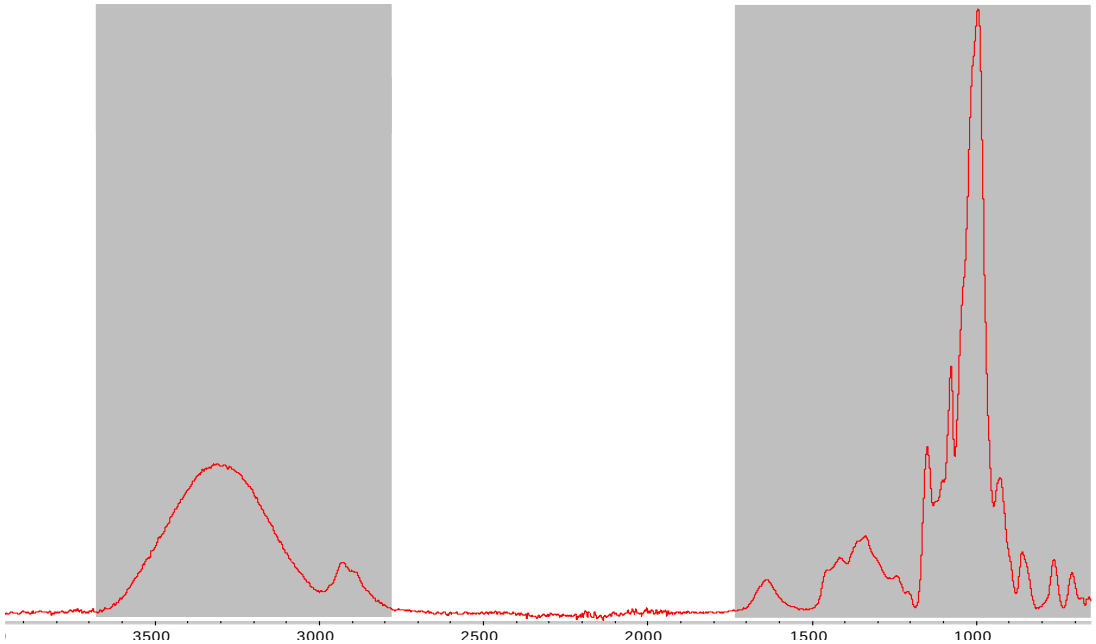
Perspectives

1. Constraint to the most informative areas of spectra



Perspectives

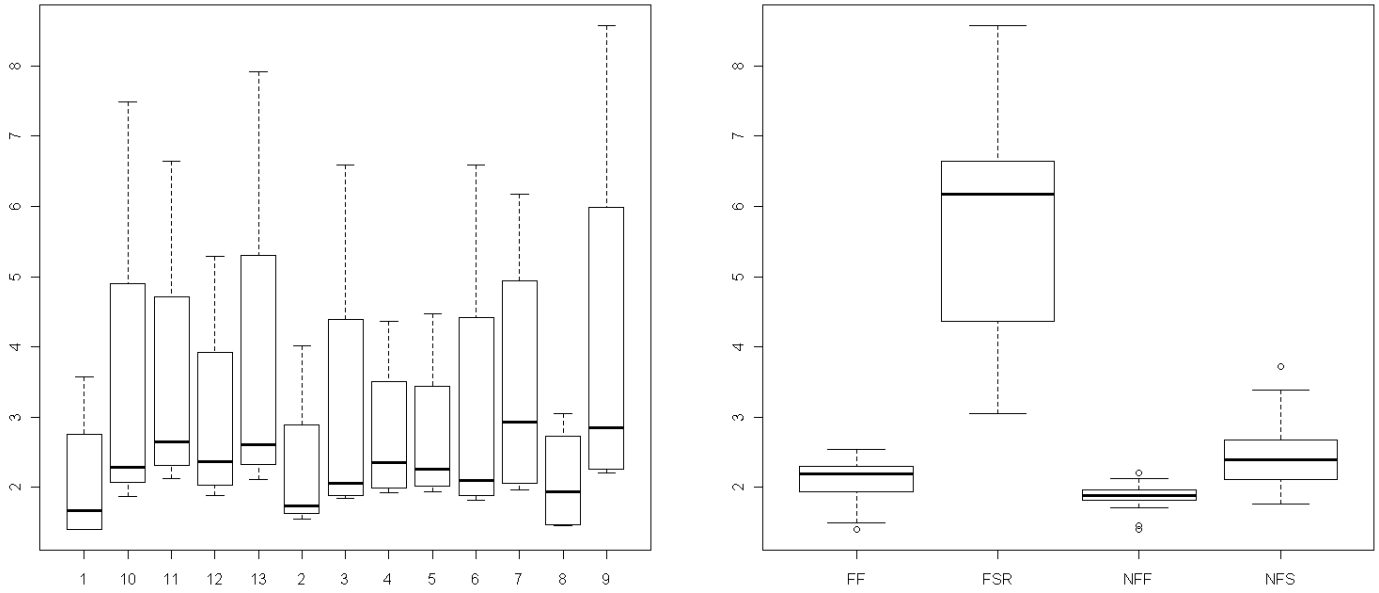
1. Constraint to the most informative areas of spectra



Perspectives

1. Constraint to the most informative areas of spectra

2. Collect more data → get closer to our goal and select a variety



Distribution of the breadmaking ability by variety and treatment

Aknowledgements

Thank you for your attention

Aknowledgements to

Pedro Maldonado, from Qualisud, for physicochemical data
Dominique Dufour, from CIAT in Colombia, for providing cassava starches