GENETIC ALGORITHMS
in Variable Selection

Riccardo Leardi

Department of Pharmaceutical and Food Chemistry and Technology
University of Genoa - ITALY

People at Dow Chemical were reading the literature ...

ELSEVIER

Chemometrics and Intelligent Laboratory Systems 41 (1998) 195–207

Chemometrics and intelligent laboratory systems

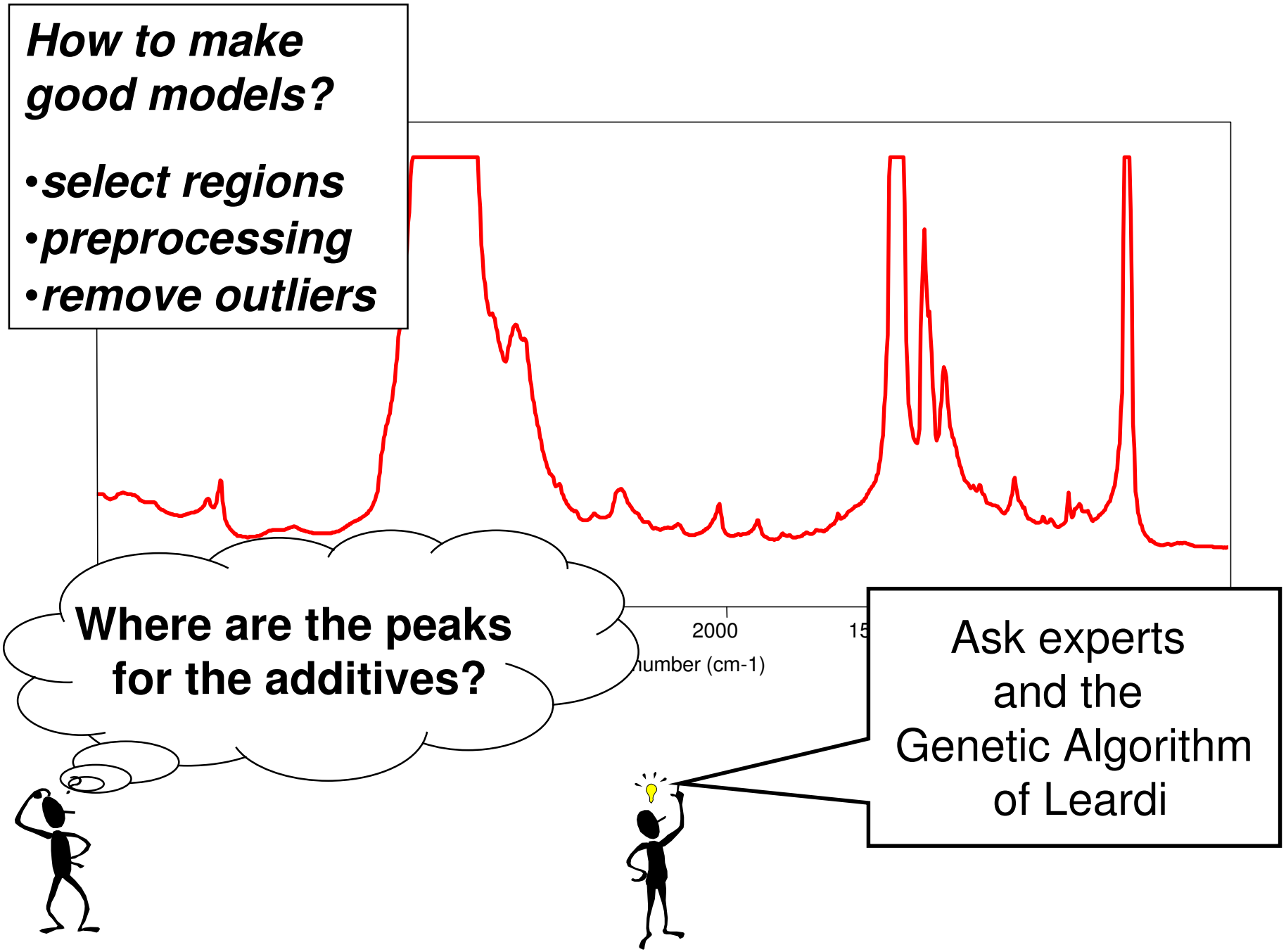
Genetic algorithms applied to feature selection in PLS regression: how and when to use them

Riccardo Leardi [a,*], Amparo Lupiáñez González [b]

[a] *Dipartimento di Chimica e Tecnologie Farmaceutiche e Alimentari, via Brigata Salerno (ponte), University of Genoa, 16147 Genova, Italy*

[b] *Departamento de Química Analítica, Facultad de Ciencias, University of Granada, Granada, Spain*

Received 7 November 1997; accepted 10 April 1998


How to make good models?
•select regions
•preprocessing
•remove outliers
2000
umber (cm-1)
Where are the peaks for the additives?
Ask experts and the Genetic Algorithm of Leardi

WHAT I GOT

- **FTIR data of polymer films (1873 wavelengths)**

- **Concentrations of 2 additives (no names)**
  - **Additive B (42 + 28 samples)**
  - **Additive C (109 + 65 samples)**

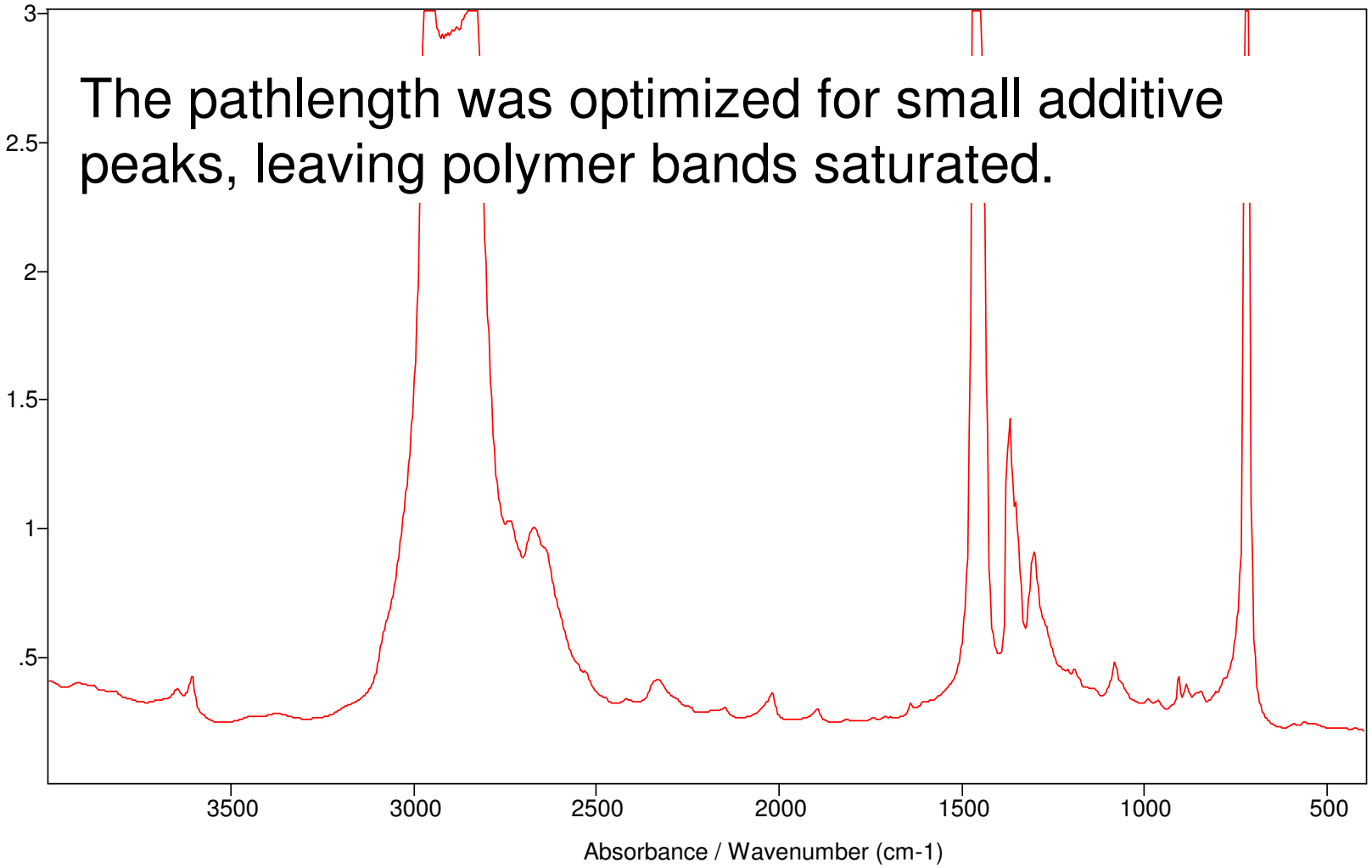
- **NO information about suggested regions**

THE CHALLENGE

To verify if Genetic Algorithms could find a model characterized by:

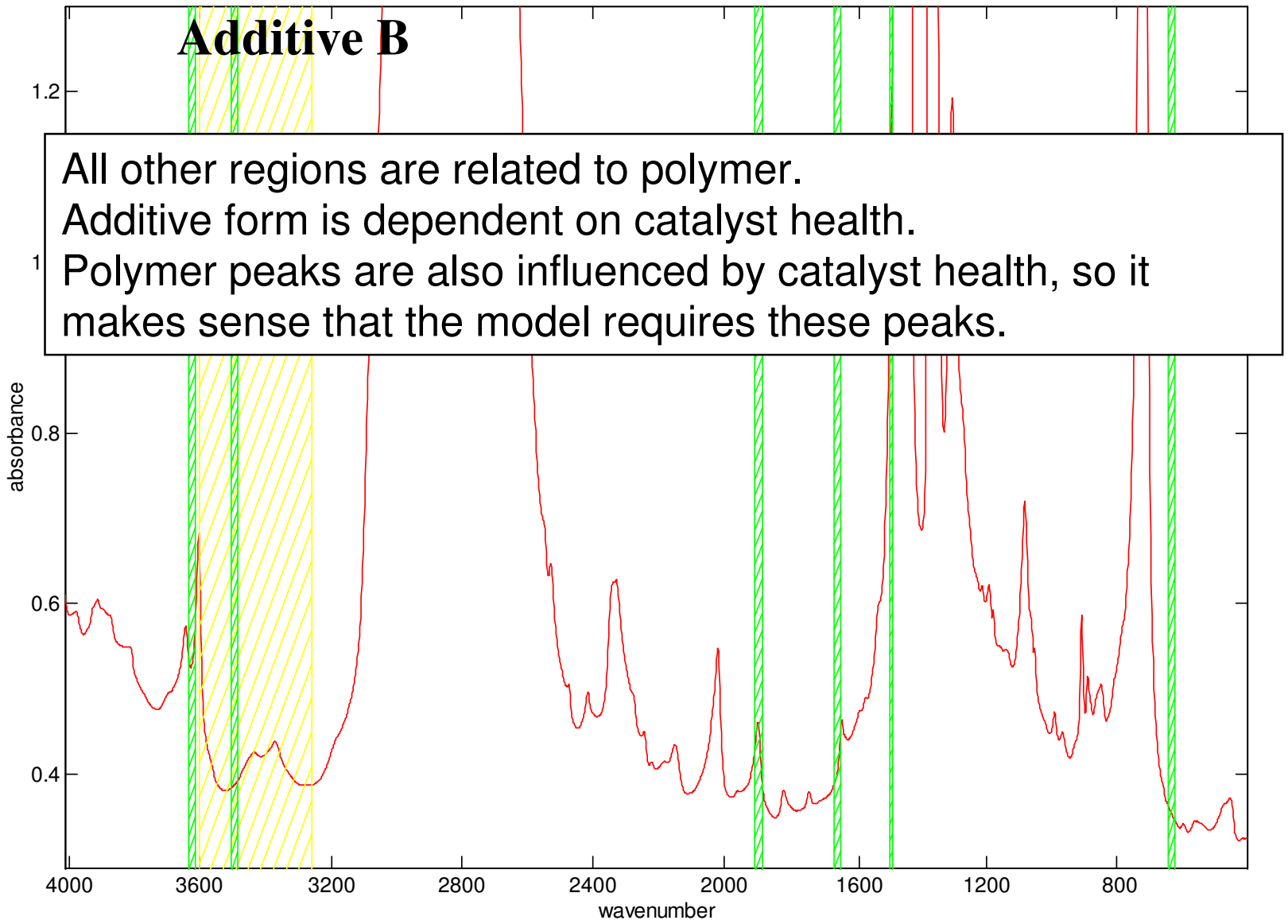
- **good predictive ability**
- **“logical” regions**

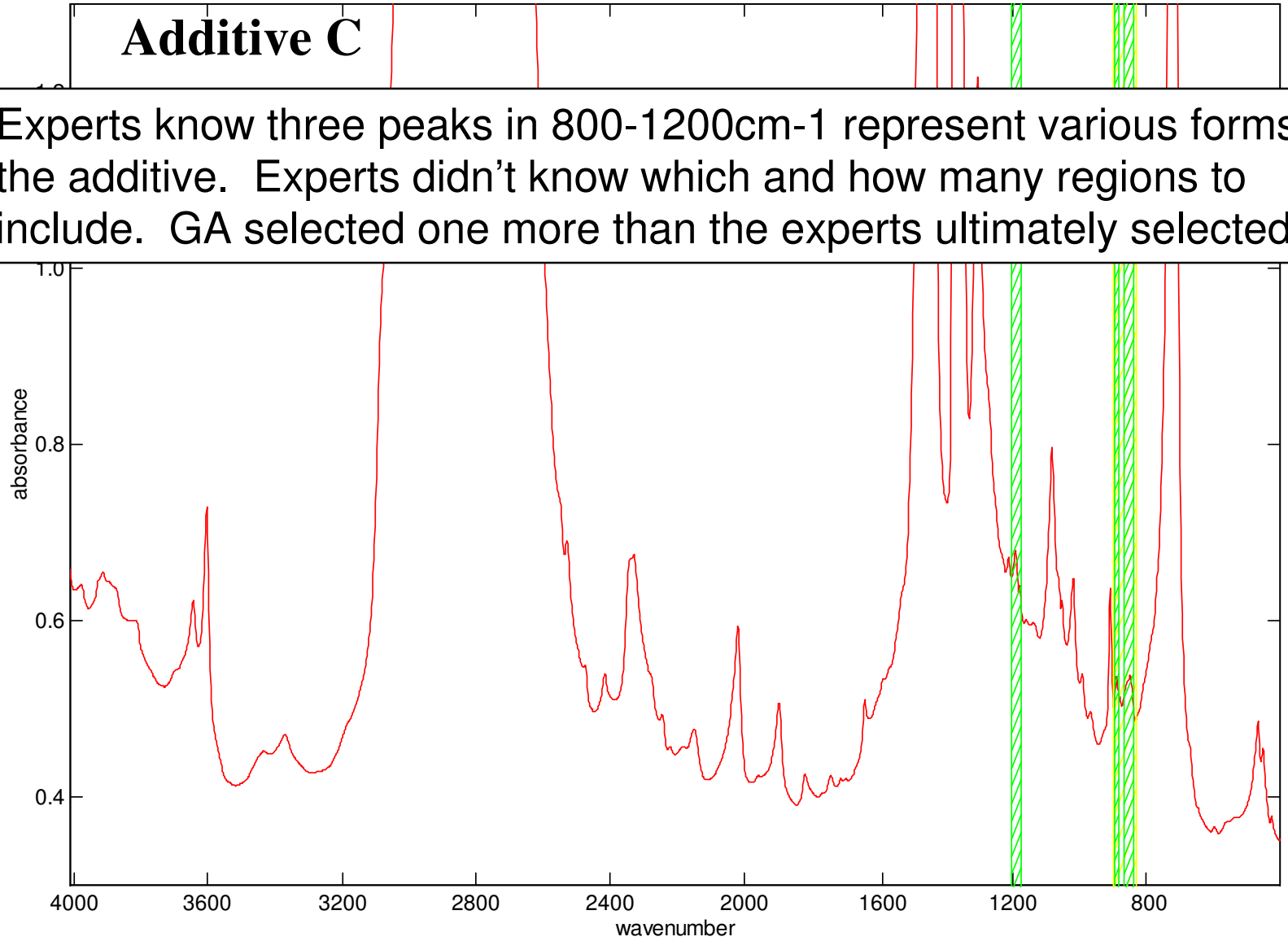
These spectra are not pretty



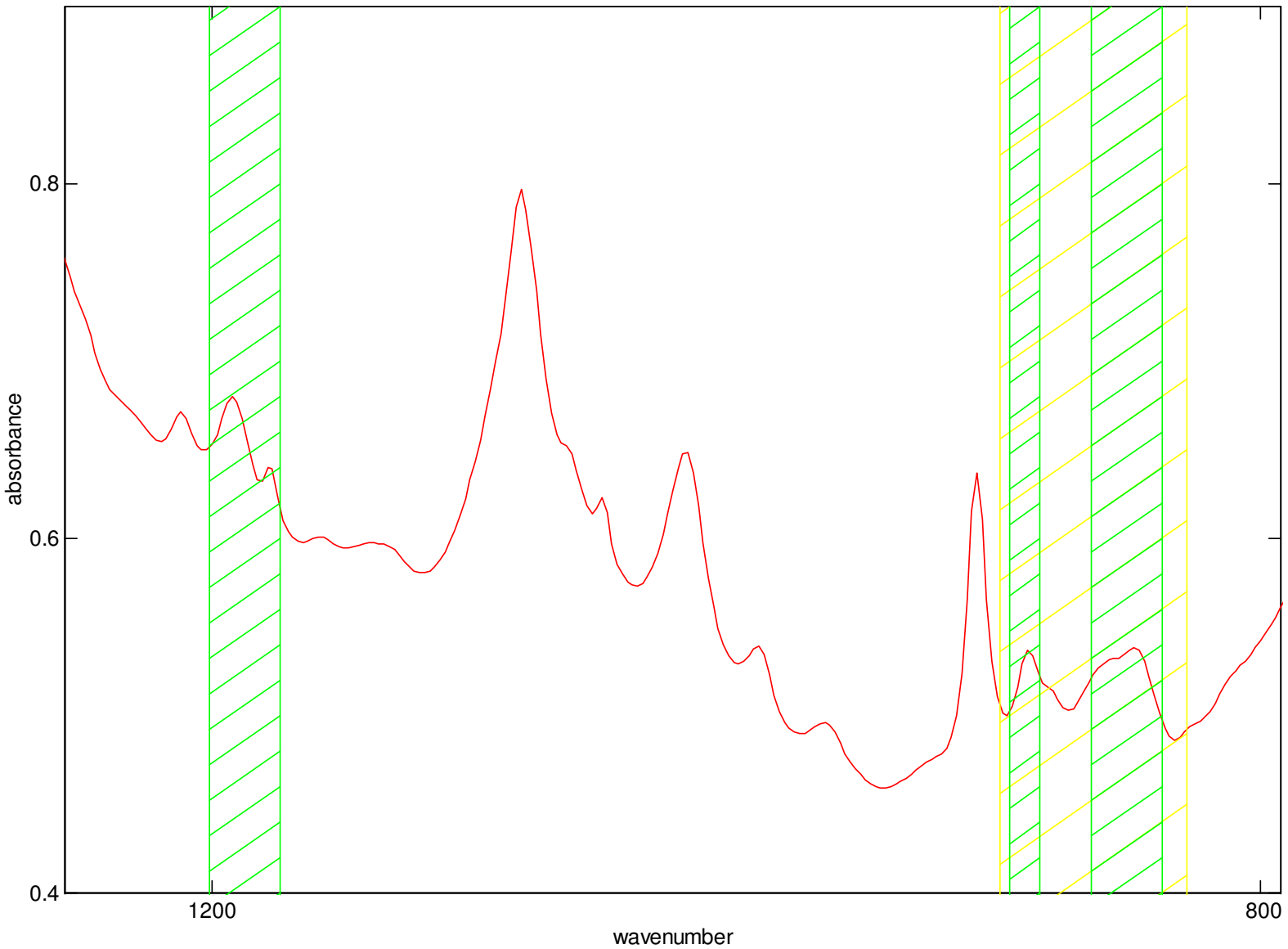
THE RESULTS

	Additive B	Additive C
RMSEP (GA)	**48**	**47**
RMSEP (expert)	**54**	**48**
regions (GA), cm-1	***3634-3616*** ***3506-3485*** **1906-1884** **1662-1645** **1493-1487** **644-623**	**1200-1175** ***895-885*** ***864-839***
regions (expert)	**3600-3260**	**899-829**


Additive B
All other regions are related to polymer.
Additive form is dependent on catalyst health.
Polymer peaks are also influenced by catalyst health, so it makes sense that the model requires these peaks.
absorbance
1.2
1
0.8
0.6
0.4
4000
3600
3200
2800
2400
2000
1600
1200
800
wavenumber



Experts know three peaks in 800-1200cm-1 represent various forms of the additive. Experts didn't know which and how many regions to include. GA selected one more than the experts ultimately selected.


0.8
0.6
0.4
absorbance
1200
800
wavenumber

This is exciting!!!

- **Variable Selection is a very important step for developing a good multivariate model.**
- **This tool provides an automated approach when expertise is not available or the variables are not known (ex. Octane number).**

VARIABLE SELECTION METHODS:

"UNIVARIATE": select those variables that have the greatest correlation with the response

"SEQUENTIAL": select the best variable and then the best pair formed by the first and second and so on in a forward or backward progression. A more sophisticated approach applies a look back from the progression to reassess previous selections

"MULTIVARIATE (PLS-ORIENTED)": Interactive Variable Selection, Uninformative Variable Elimination, Iterative Predictor Weighting PLS, Interval PLS, ...

GENETIC ALGORITHMS

GENETIC ALGORITHMS

Genetic Algorithms (GA) mimick the evolution of a species according to the Darwinian theory.

Each experimental condition, coded by a sequence of 0's and 1's, is treated as the genome of an individual, whose "performance" is considered as its "fitness"

Operators of a classical GA:

Select-copy: simulates the fights for mating, in which the best individuals have the highest probability of success, and therefore of spreading their genome

Cross-over: simulates the mating between two individuals, producing two offsprings, whose genetic material is derived from that of the two parents

Mutation: as in nature, rarely occurring random phenomena, producing random changes in the genetic material

AN EXAMPLE OF GA APPLIED TO FEATURE SELECTION

(for sake of simplicity, assume 10 variables)

chromosome 1: 0010011001 (model made by variables 3, 6, 7, 10)

chromosome 2: 1000110011 (model made by variables 1, 5, 6, 9, 10)

Cross-over: genes 1, 4, 6, 8 are swopped

offspring 1: 1010011001

offspring 2: 0000110011

Mutation: gene 2 of offspring 2 is mutated

offspring 1: 1010011001 (variables 1, 3, 6, 7, 10)

offspring 2: 0100110011 (variables 2, 5, 6, 9, 10)

The main problems of ”Classical GA”

- overfitting

- lack of reproducibility

When applied to spectral data sets (as any other selection method)

- non ”spectroscopically logical” selections (”dispersed” wavelengths rather than regions)

Modifications have been made to the standard GA in order to:

- **make it more suitable to the feature selection problem**
- **reduce the risk of overfitting**

Further modifications have been made to make it especially suitable for spectral data sets

Detailed description of the algorithm goes well beyond the scope and the time of this talk

DRAWBACK OF GA-PLS

- **Huge computation time (owing to the increased computing power this limitation is becoming less and less relevant)**

Data set **APPLE JUICES** (Research Institute of Geisenheim (Germany), Department of Wine Analysis and Beverage Research)

638 German apple juices from five different years (1999, 2000, 2001, 2002, 2003)

FT-IR spectra (1054 wavelengths) by Wine Scan FT120 (Foss Electric A/S) (only wavelengths 1-550 are taken into account)

5 responses (Brix, density, Folin C, TEAC, total acidity)

Variable selection performed on 229 samples (1999, 2000)

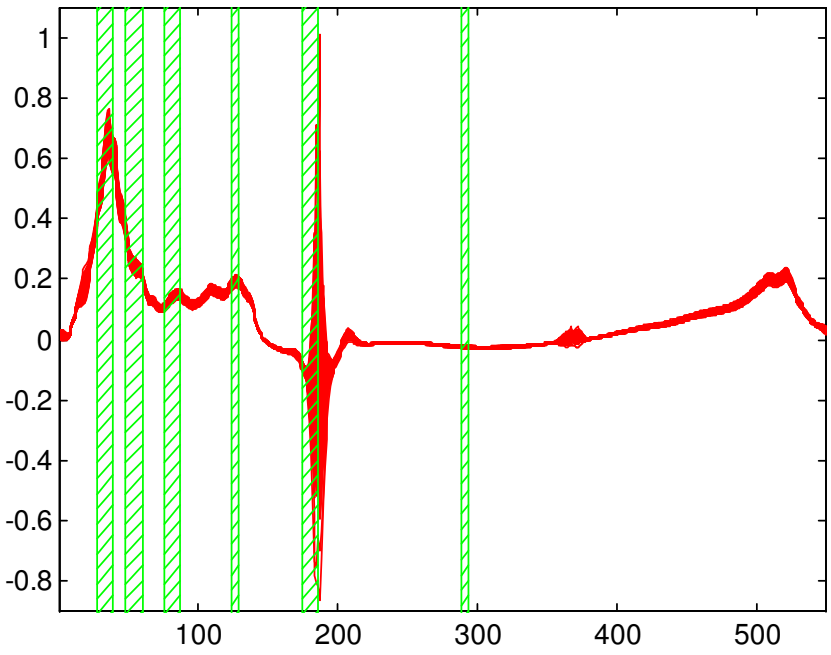
GOALS OF THIS STUDY

Compare GA to a commercial package for variable selection (Foss) in what concerns the distribution and the interpretability of the selected wavelengths (possibility of designing a filter instrument)

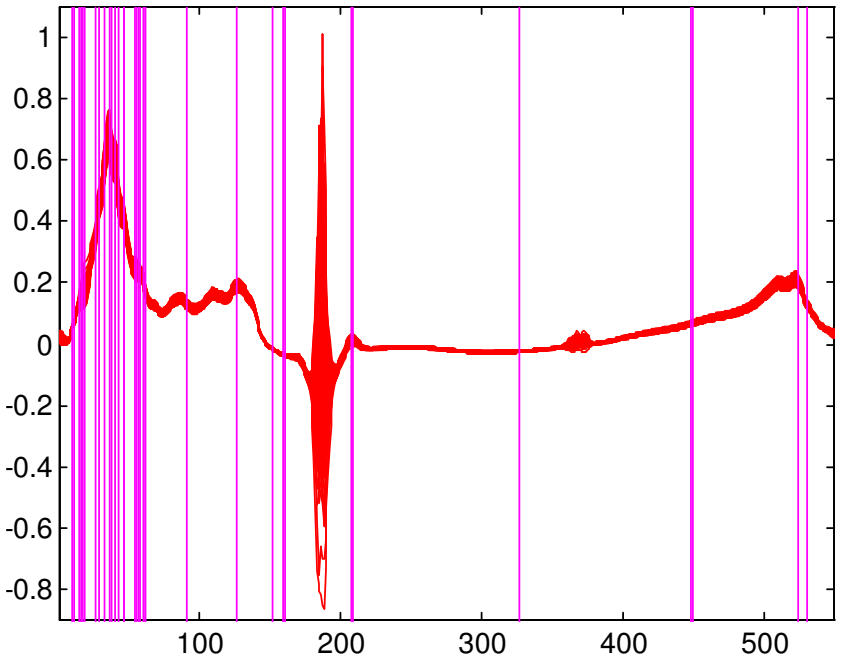
Check if the regions selected on the 1999-2000 samples have a good predictivity also for the following years

Since it is not possible to keep the same PLS model throughout the years, determine the size of the training set required to get an acceptable RMSEP

BRIX

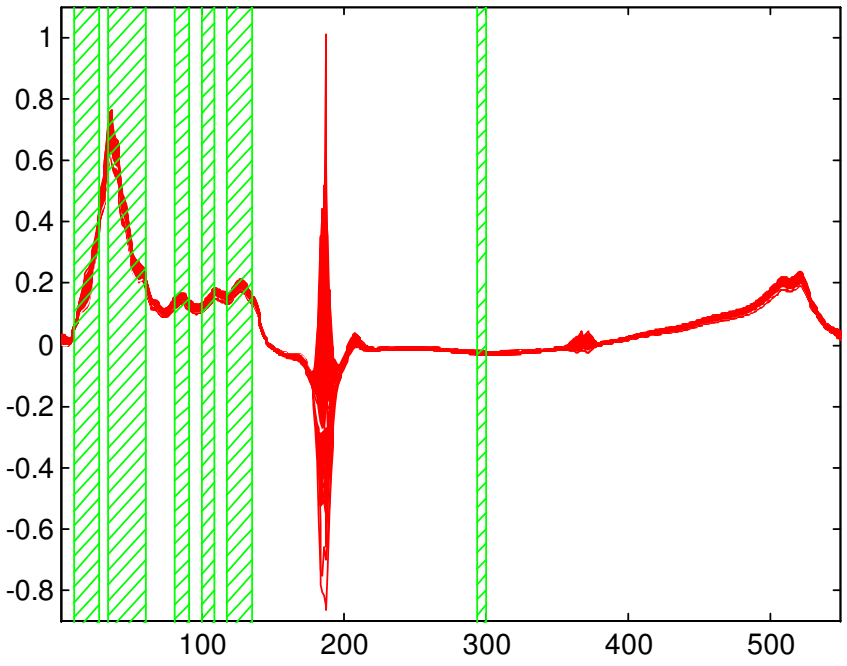


GA

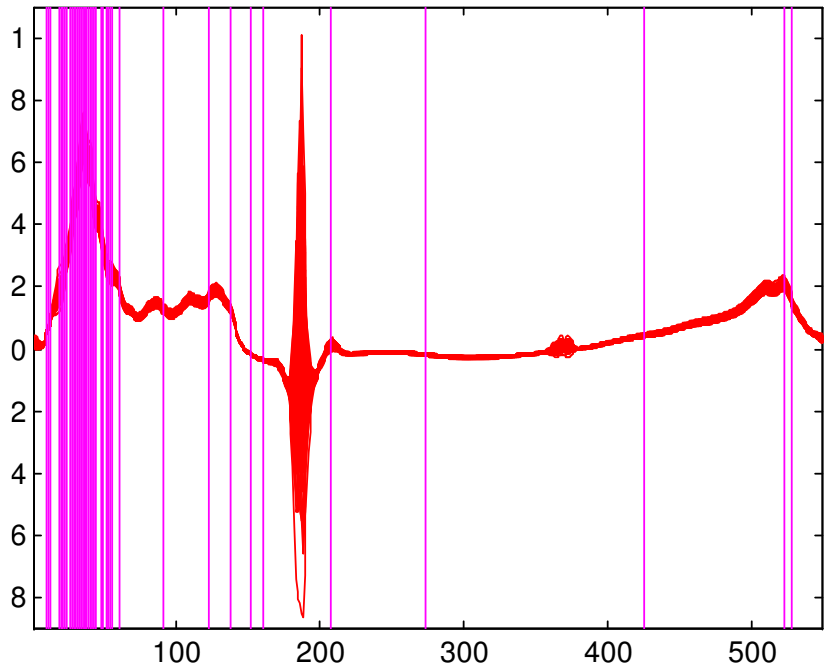


Wine Scan

DENSITY

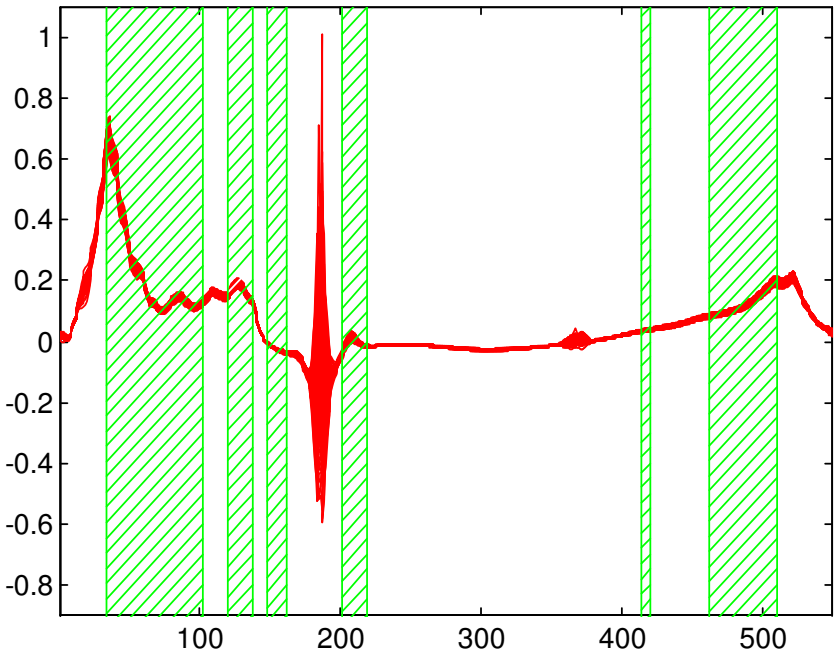


GA

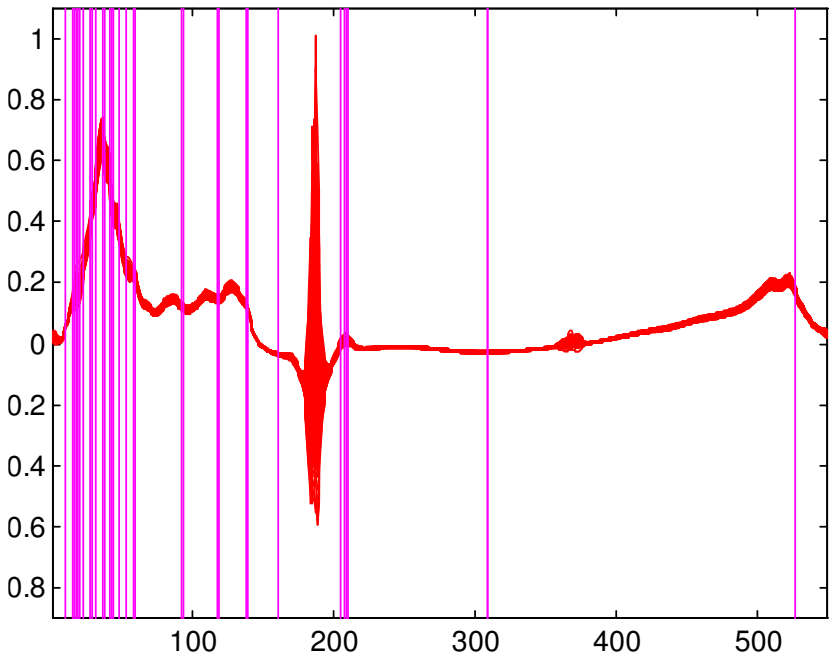


Wine Scan

FOLIN C INDEX

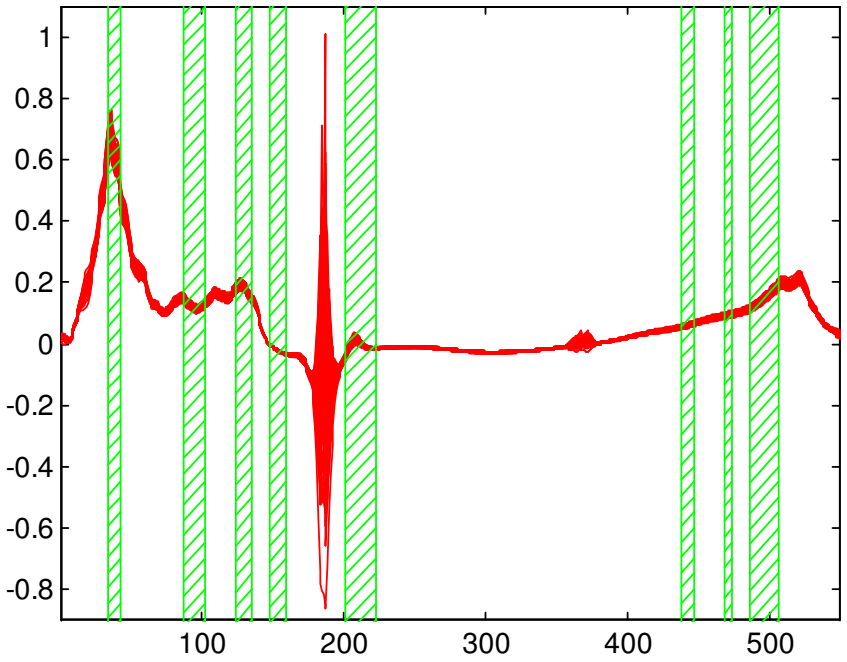


GA

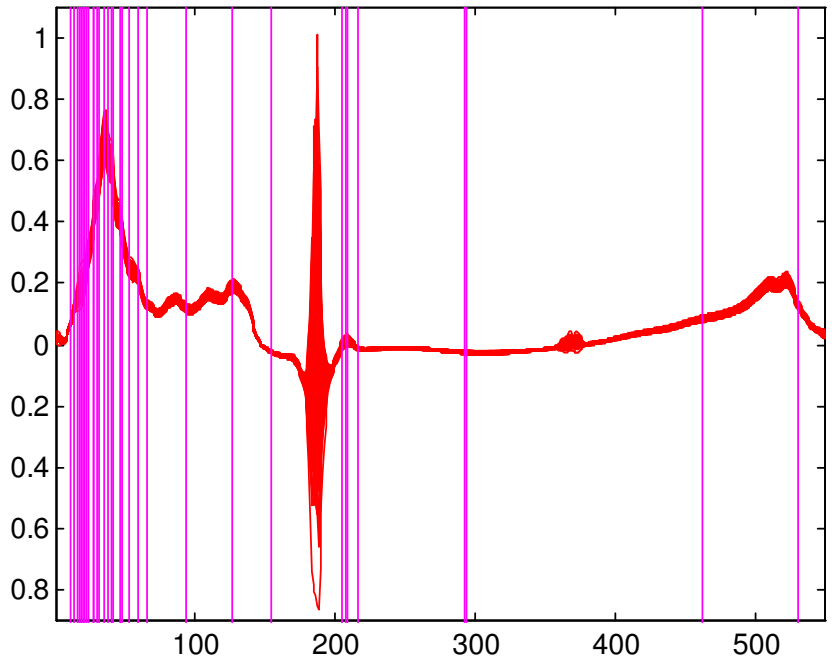


Wine Scan

TEAC

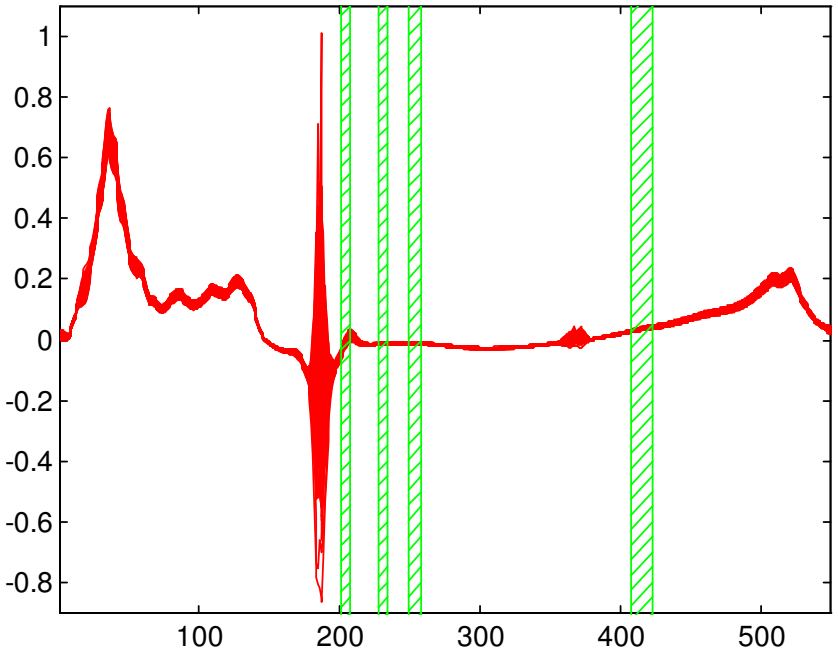


GA

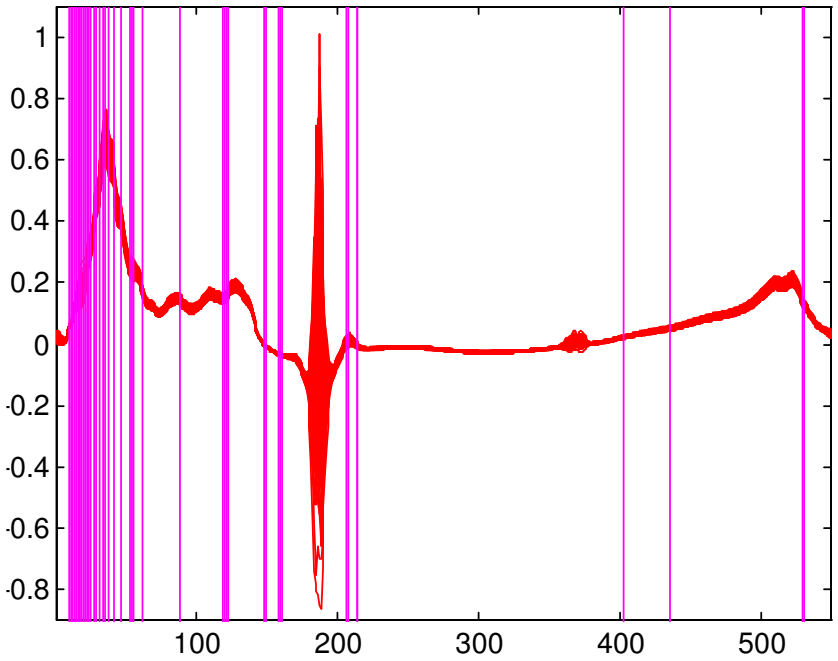


Wine Scan

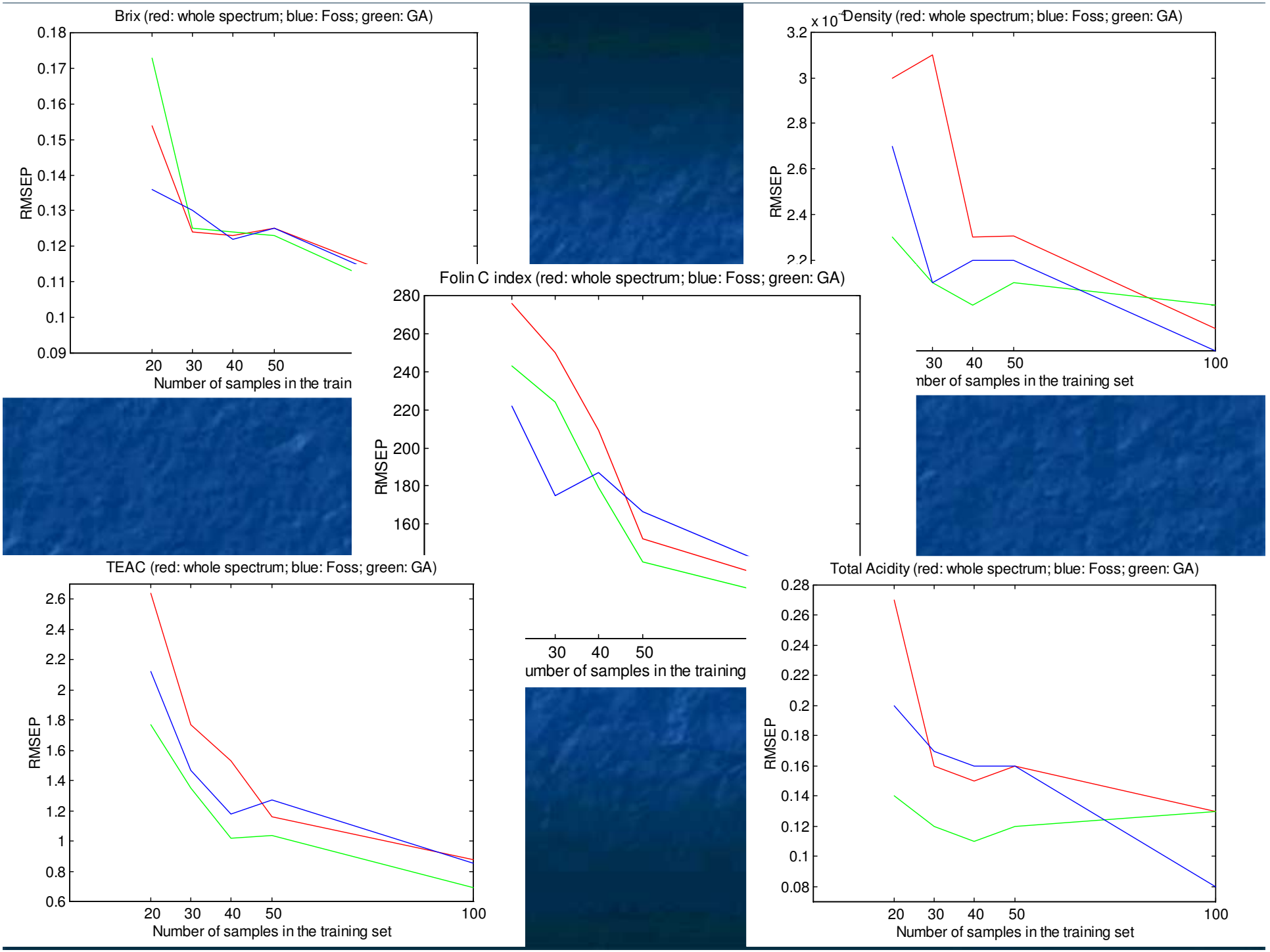
TOTAL ACIDITY AS TARTARIC



GA



Wine Scan


Brix (red: whole spectrum; blue: Foss; green: GA)
RMSEP
Number of samples in the train
Density (red: whole spectrum; blue: Foss; green: GA)
RMSEP
nber of samples in the training set
Folin C index (red: whole spectrum; blue: Foss; green: GA)
RMSEP
umber of samples in the training
TEAC (red: whole spectrum; blue: Foss; green: GA)
RMSEP
Number of samples in the training set
Total Acidity (red: whole spectrum; blue: Foss; green: GA)
RMSEP
Number of samples in the training set

Data set PINE SEEDS:

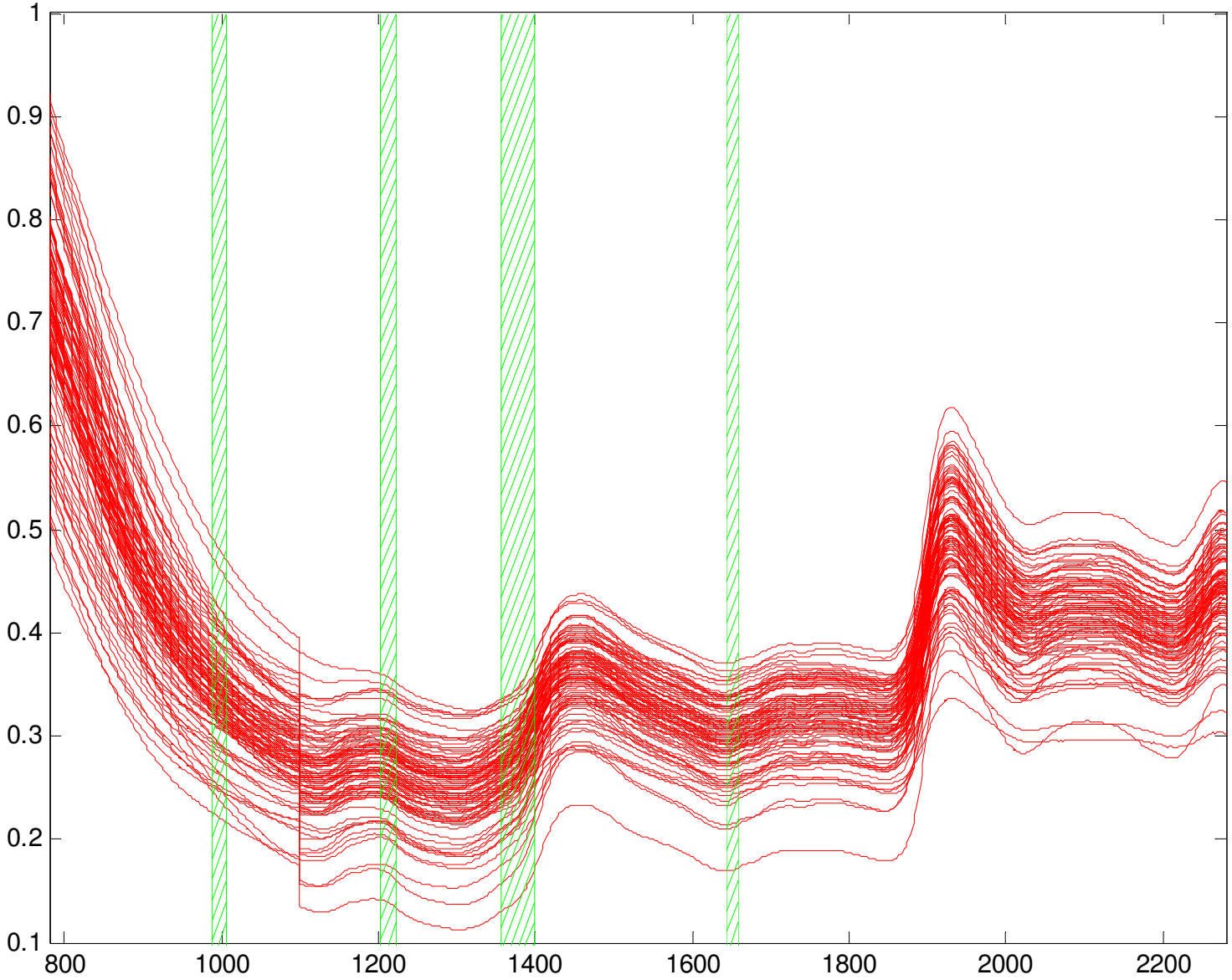
- **Moisture measured on 155 single seeds of Scots pine (*Pinus sylvestris* L.)**
- **Training set: 103 samples**
- **Validation set: 52 samples**
- **NIR spectra (751 wavelengths in the range 780-2280 nm) by NIRS 6500 (NIRSystems, Silver Spring, MD, USA)**

Torbjörn Lestander (Dept. of Silviculture, Swedish University of Agricultural Sciences, Umeå) and **Paul Geladi** (Unit of Biomass Technology and Chemistry, Swedish University of Agricultural Sciences, Umeå)

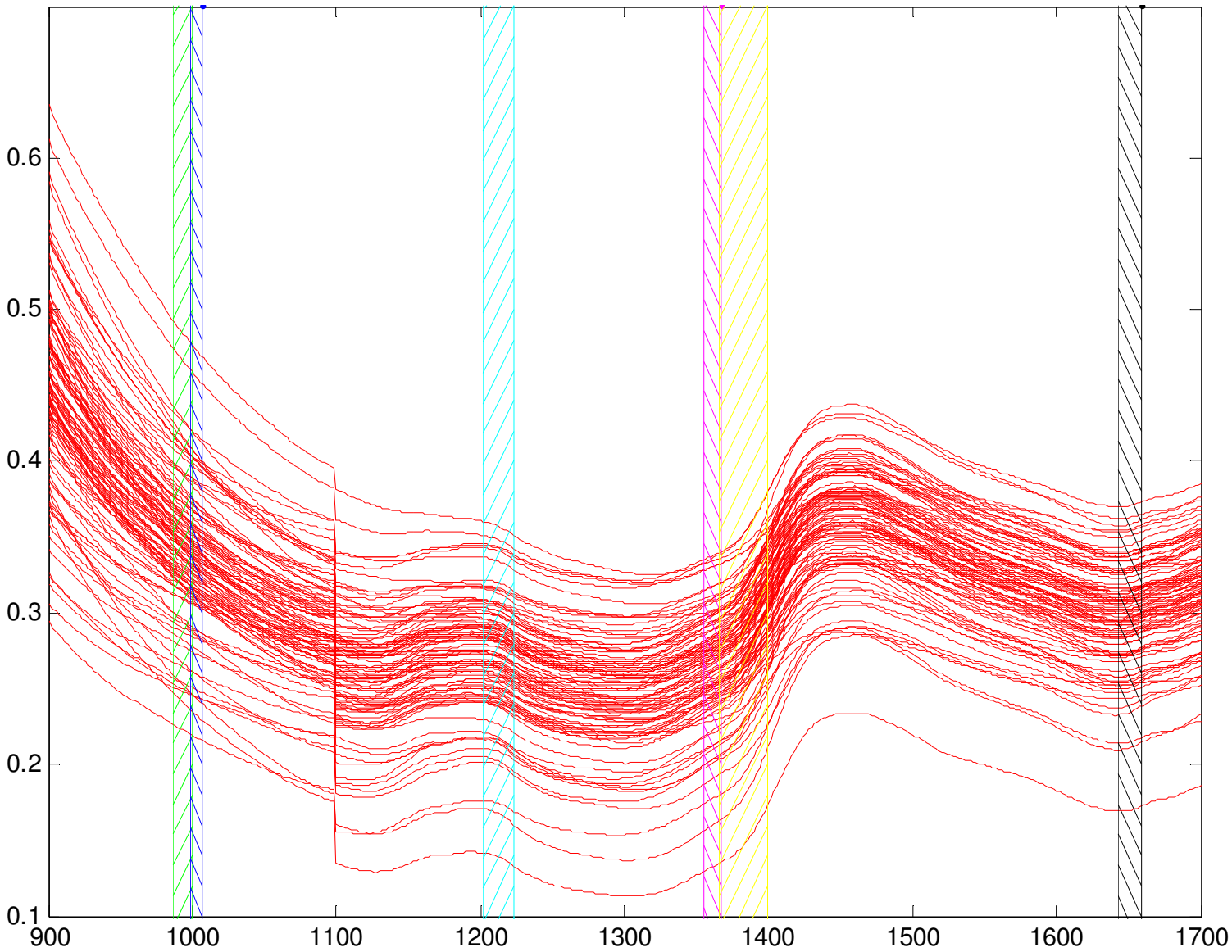
GOAL OF THIS STUDY

Select wavelengths that could be used in a few NIR filter sensors to predict moisture content in single seeds of Scots pine.

The results are of importance to the construction of an apparatus that uses parallel NIR-sensors for automatic and fast moisture determinations of conifer seeds.



RMSEP full spectrum: 1.9; RMSEP selected regions (50 wl.): 1.6



RMSEP full spectrum: 1.9; RMSEP six uniform density filters: 2.1

CONCLUSIONS

The application of GA as a technique of wavelength selection produced models that

- **were able to emulate region choices of experts**
- **gave results better than a well-known commercial software (lower RMSEP, better interpretation of selected wavelengths)**
- **allowed to detect relevant regions for the construction of filter instruments**