Prediction of chemical characteristics of fibrous plant biomasses by their near infrared spectrum: comparing local versus partial least square models and cross-validation versus independent validations

B. Godin, **R. Agneessens, J. Delcarte, P. Dardenne**

Centre wallon de Recherches agronomiques (CRA-W)

Valorisation of Agricultural Products Department

Chaussée de Namur, 146 ; 5030 Gembloux ; Belgium

b.godin@cra.wallonie.be

1

1. Outline

- Fibrous plants biomasses
  - Important potential as a source of renewable fuels and chemicals

- Chemical characteristics
  - Difficult to control the variability
  - Need to be analyzed

- Standard wet chemical methods
  - Reliable, primary method
  - Tedious, time and resource consuming, expensive

- Near infrared spectroscopy
  - Simple, fast, cheap, clean, non-destructive, reliable alternative
  - Secondary method

Near infrared spectroscopy

- Partial least square (PLS) regression
  - Linear
  - Uses all the samples of the calibration dataset
  - To improve the accuracy with a large number of samples
    → Split the calibration dataset

- Local method → Shenk at al. 1997
  - Non-linear
  - Specific PLS regression with a low number of samples for each sample by selecting its most similar spectral neighbors
  - Copes with non-linearity and non-homogeneity
    of a large calibration dataset → Especially with multiproduct

Aim of the study

- Reliability of the predictions for chemical characteristics of fibrous plants biomasses

  - Use of multispecies datasets → Larger concentration range

  - Local versus Partial least square (PLS) models

  - Cross-validation versus Independent validations

2. Material and methods

- Analyzed fibrous plant biomasses
  - Fiber corn, Fiber sorghum, Hemp, Jerusalem artichoke leaves and stalks, Miscanthus giganteus, Spelt straw, Switchgrass and Grasses (Tall fescue, Cocksfoot, Immature rye, Immature spelt)
  - Cropping trials → Different years, sites, harvest periods, cultivars and nitrogen fertilization levels

Miscanthus   Switchgrass   Tall fescue   Hemp   Spelt straw

ENERBIOM, 2012

Material and methods

- Presented analyzed chemical characteristics
  - NDF (Neutral Detergent Fiber residue) determined by the Van Soest method
  - ADF (Acid Detergent Fiber residue) determined by the Van Soest method
  - ADL (Acid Detergent Lignin) determined by the Van Soest method
  - Mineral compounds content (MC)

- Van Soest method
  - Cellulose = ADF-ADL
  - Hemicelluloses = NDF-ADF
  - Lignin = ADL

Chemical characteristic (g 100g^{-1} DM)	n	Min.	Max.	Median	Median SD	SEL
NDF	1169	29.59	91.40	66.77	18.04	0.40
ADF	1167	17.15	70.91	42.31	15.08	0.30
ADL	1167	1.13	13.59	6.54	3.91	0.15
Mineral compounds	1377	0.84	20.09	6.53	3.38	0.10

Near-infrared analysis

- Near infrared reflectance spectra
  - 1100 to 2498 nm by step of 2
  - Spectra normalization by a standard normal variate (SNV) transformation followed by a first order derivation (1, 4, 4, 1)
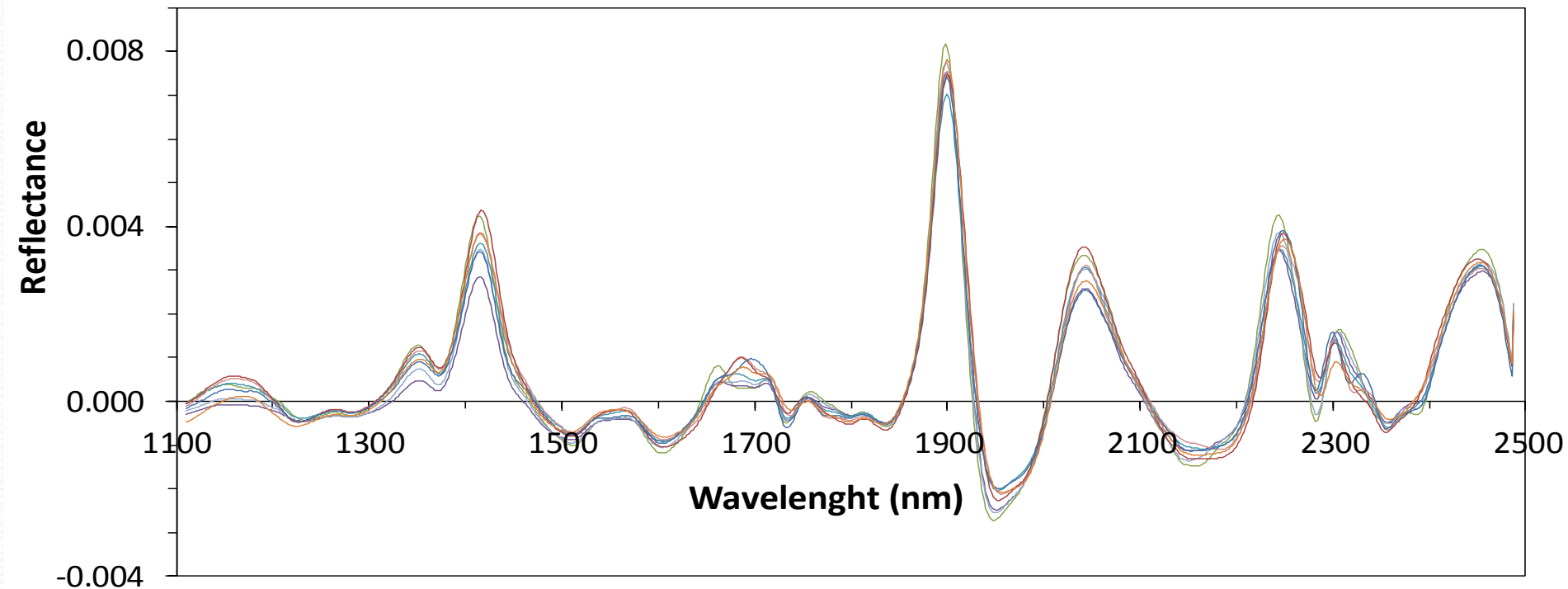
Mean spectra of each type of plant species

Optimization of the models

- Partial least square (PLS)
  - Number of PLS components
  - Software → WinISI 4.6.8

- Local method → Shenk at al. 1997
  - Number of selected samples (from 25 to 400 by steps of 25)
  - Minimum and maximum number of PLS components (between 1 and 40) for the specific regression
  - Software → WinISI 4.6.8

| Chemical characteristic	PLS	LOCAL		
	Number of PLS components	Optimum number of selected samples	Minimum number of PLS components	Maximum number of PLS components
NDF	15	150	7	31
ADF	15	225	8	31
ADL	15	125	7	25
Mineral compounds	16	300	13	34

Performances of the models

- Cross-validation CV-LOO
  - Leave-one-out full cross-validation

- Validation V1
  - Contains approximately 20% of total samples (approximately 20% per plant species group)

- Validation V2
  - Each of the 8 other independent validation datasets V2 only contained the samples of one plant species group

Performances of the models

- Based on medians to be robust
  - MedRE: Median standard residual error of prediction
    - MedRE=1.4826*MAD (median absolute deviation)
  - R²Med: Coefficient of determination of prediction based on medians
    - $R^2Med=(SDMed^2-MedRE^2)/(SDMed^2)$
  - RPDMed
    - $RPDMed=SDMed*MedRE^{-1}$
  - GHMed: Median spectral distance of Mahalanobis

3.1. Local versus PLS models

- Median standard residual error of prediction (MedRE)
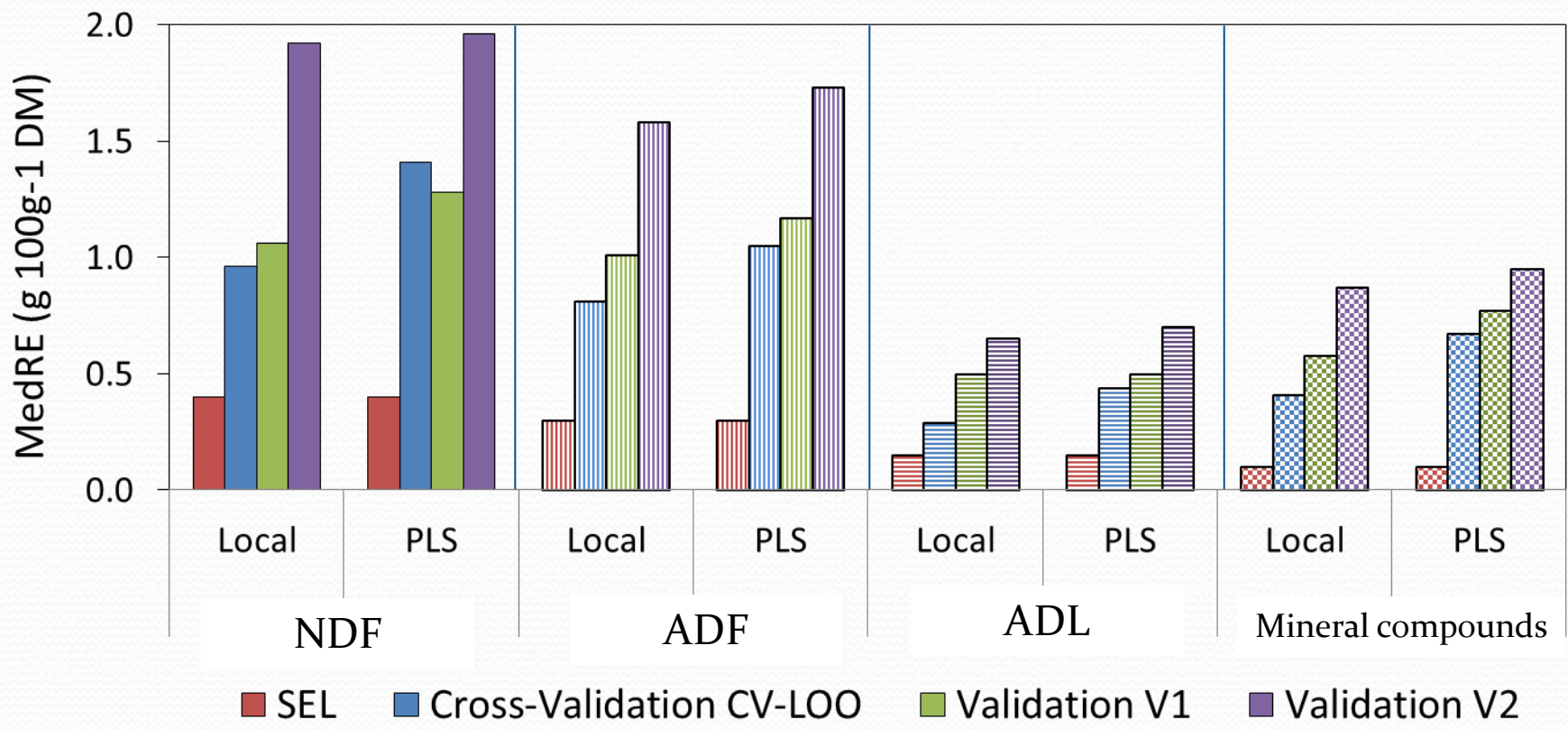
- For each type of characteristic and validation, the local models are more reliable in terms of prediction error with multispecies compared to the PLS models

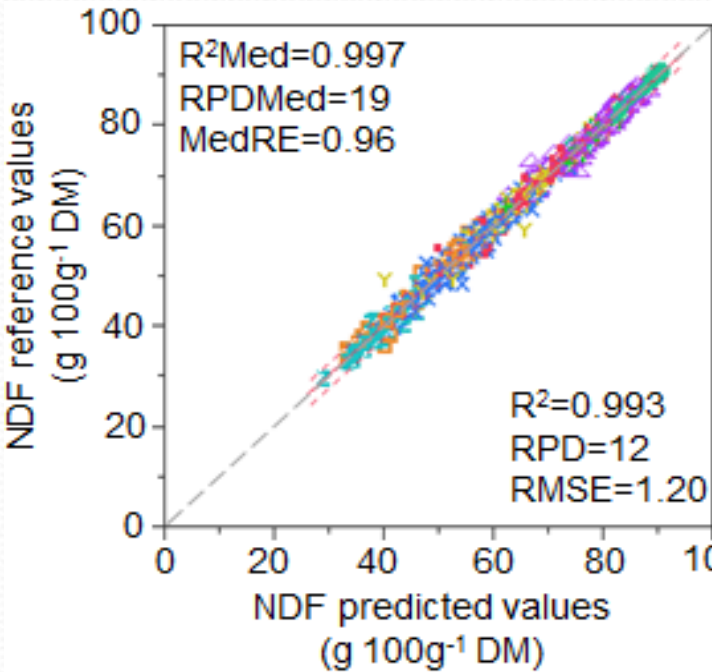# 3.2. Cross-validation versus Independent validations

- R2Med and RPDMed
  - High prediction performances of the multispecies

Chemical characteristic	Cross-validation CV-LOO		Validation V1		Validation V2		Cross-validation CV-LOO		Validation V1		Validation V2	
	Local	PLS	Local	PLS	Local	PLS	Local	PLS	Local	PLS	Local	PLS
	R^2Med						RPDMed					
NDF	0.997	0.994	0.997	0.996	0.989	0.988	19	13	19	15	9.4	9.2
ADF	0.997	0.995	0.996	0.994	0.989	0.987	19	14	15	13	9.5	8.7
ADL	0.995	0.987	0.984	0.984	0.972	0.968	14	8.9	7.9	7.9	6.0	5.6
Mineral compounds	0.985	0.961	0.970	0.947	0.934	0.921	8.3	5.1	5.8	4.3	3.9	3.6

Cross-validation versus Independent validations

Local method

Cross-validation CV-LOO

R^2Med=0.997
RPDMed=19
MedRE=0.96

R^2=0.993
RPD=12
RMSE=1.20

Validation V1

R^2Med=0.997
RPDMed=19
MedRE=1.06

R^2=0.991
RPD=10
RMSE=1.45

Validation V2

R^2Med=0.989
RPDMed=9.4
MedRE=1.92

R^2=0.976
RPD=6.5
RMSE=2.28

NDF reference values (g 100g^{-1} DM)
NDF predicted values (g 100g^{-1} DM)
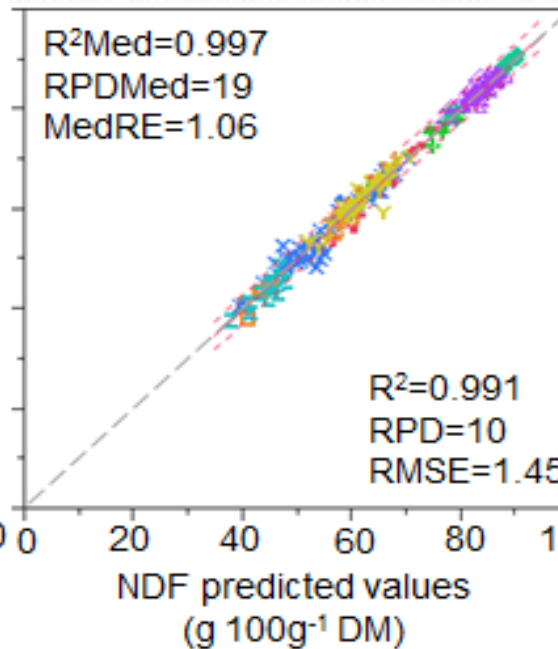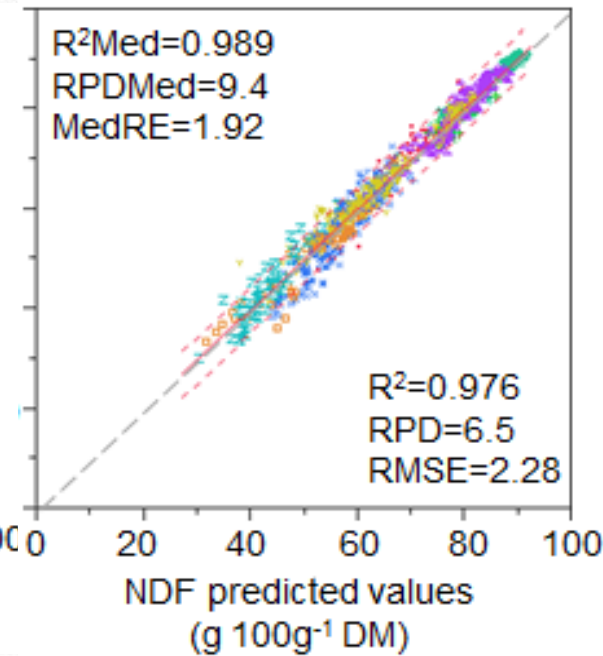
□: Fiber corn
■: Hemp
◊: Miscanthus giganteus
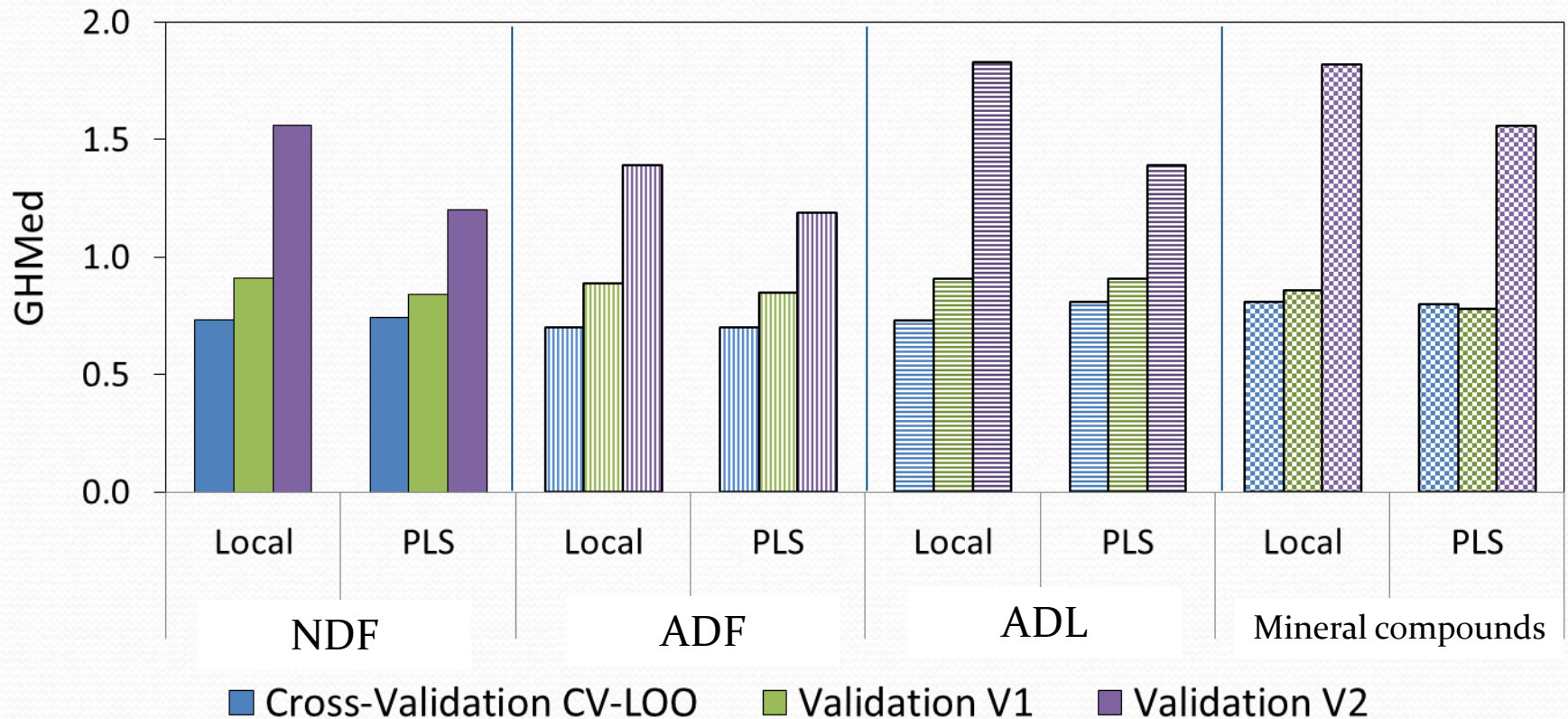Δ: Switchgrass

y: Fiber sorghum
z: Jerusalem artichoke leaves and stalks
+: Spelt straw
x: Grasses

Cross-validation versus Independent validations

- Median spectral distance of Mahalanobis (GHMed)

Cross-validation versus Independent validations

- Reliability
based on the ratio of MedRE to SEL (standard error of laboratory)
  - Cross-validation CV-LOO
    - Predictions are in median 2.4 times less accurate than SEL
    - Local models are in median 23% more accurate than PLS models
  - Validation V1
    - Predictions are in median 2.9 times less accurate than SEL
    - Local models are in median 14% more accurate than PLS models
  - Validation V2
    - Predictions are in median 4.5 times less accurate than SEL
    - Local models are in median 8.7% more accurate than PLS models

Cross-validation versus Independent validations

- The degree of independence of the validation set in regards to the calibration set has a major impact on the prediction performances of multispecies models
  → Especially for the local method

3.3. Addition of a few samples of the predicted group

- Validation V2
  - Calibration not containing samples of the predicted plant species group
  - Too independent
  - Reliability

- Reduction of the degree of independence of validation V2
  - Addition a few independent samples (5, 10, 15, 20, and 25)
    of the predicted plant species group to the calibration dataset

Addition of a few samples of the predicted group
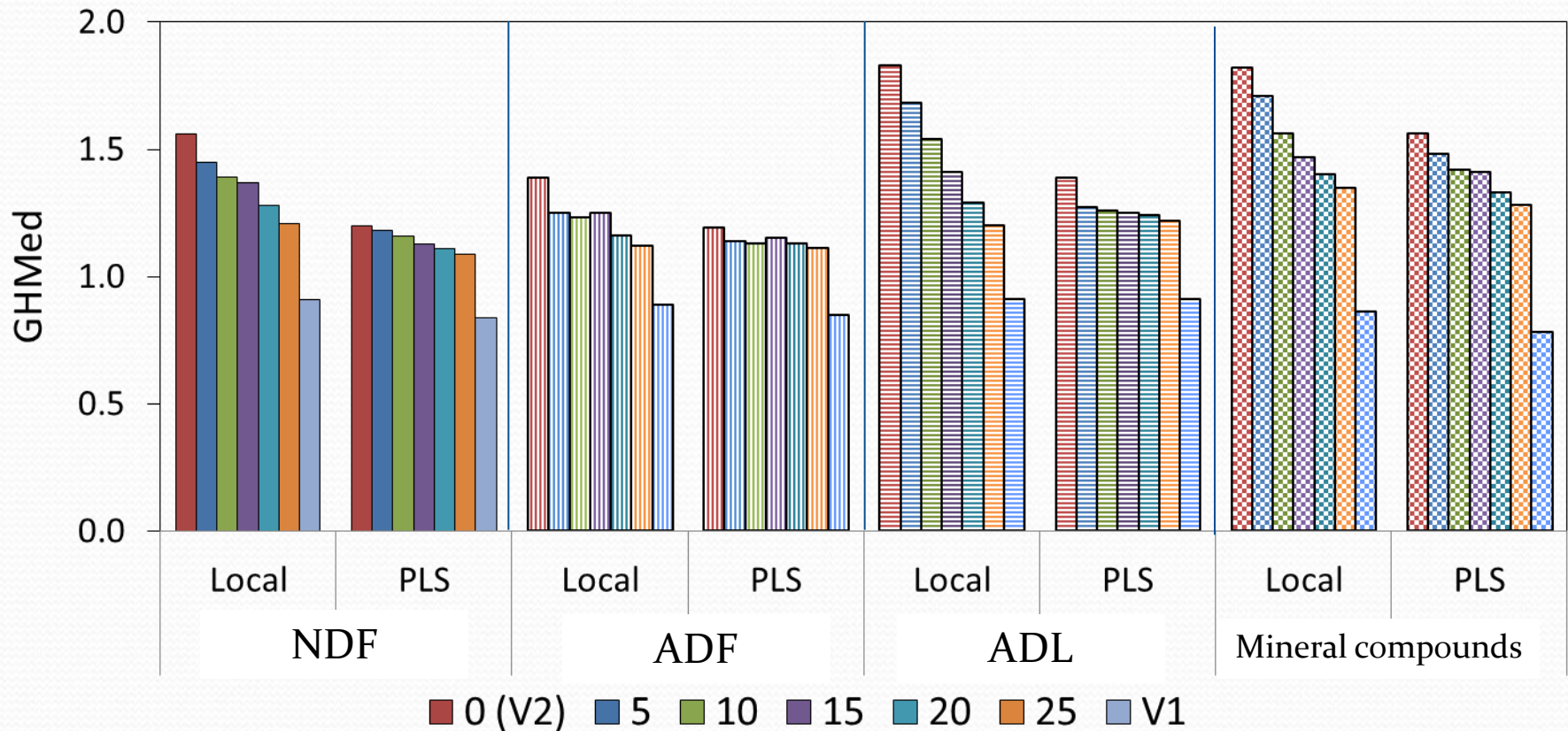
- Median standard residual error of prediction (MedRE)
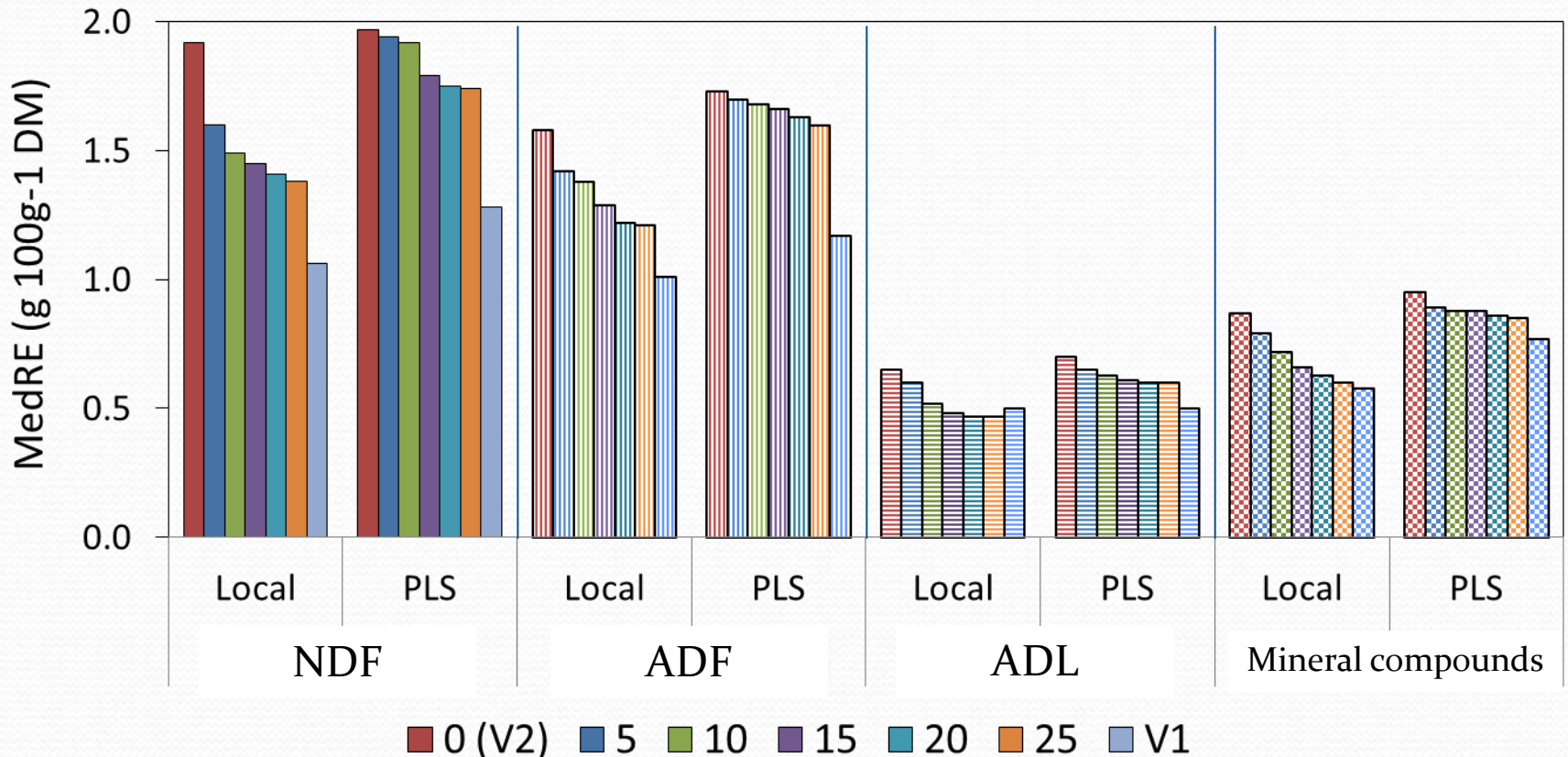
- Improvement of the prediction performances of multispecies models
  →Especially for the local method

Addition of a few samples of the predicted group

- Median spectral distance of Mahalanobis (GHMed)

- Improvement of the prediction performances of multispecies models
→Especially for the local method

Addition of a few samples of the predicted group

- Prediction performances improvement of validation V2
  after the addition of 25 samples
  - Based on the ratio of MedRE to SEL (standard error of laboratory)
    - In mean by 28% for local models
    - In mean by 11% for PLS models

  - Based on the median spectral distance of Mahalanobis (GHMed)
    - In mean by 26% for local models
    - In mean by 12% for PLS models

Addition of a few samples of the predicted group

- The local method can be used for predictions of a given plant species when there are only a few samples of them which are present in a large multispecies dataset of similar plant species samples

  → Fast cost-effective NIR screening, ranking and quantitative analyses of new plant biomasses

Acknowledgements

Programme cofinancé par l'Union européenne
Fonds européen de développement régional
EU-gefördertes Programm
Europäischer Fonds für regionale Entwicklung

Chimiométrie XVI

19-21 January 2015