How spectrally representative are datasets used to build MIR-based predictive models ? A data-driven study.

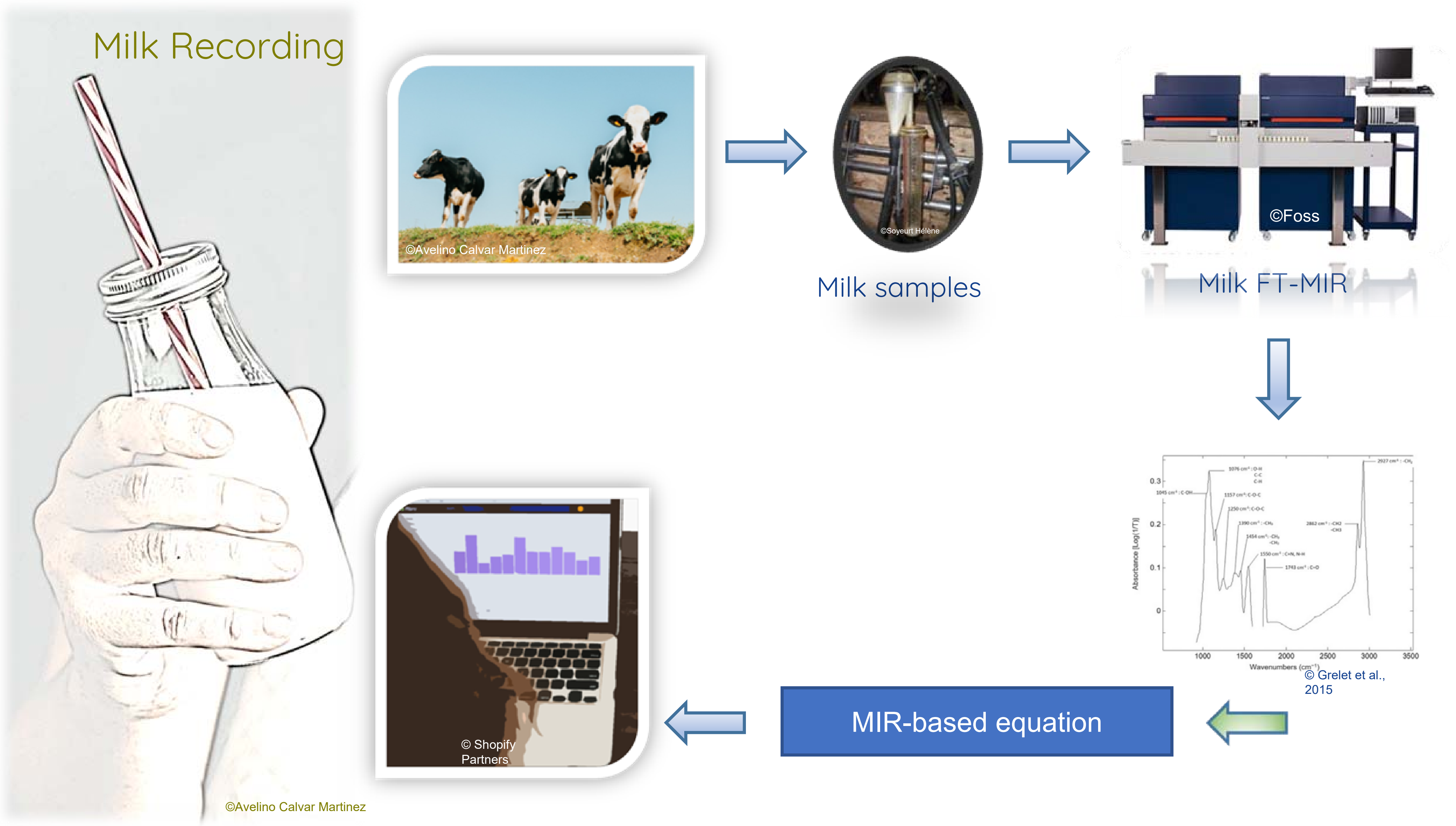
C. Nickmilder[1], J. Leblois[2], O. Christophe[3], C. Grelet[3], Holicow Consortium, H. Soyeurt[1]

[1]Gembloux Agro-Bio Tech, University of Liège, Belgium
[2]Eleveo, Walloon Breeding Association, Ciney, Belgium
[3]Walloon Research Centre (CRA-W), Gembloux, Belgium

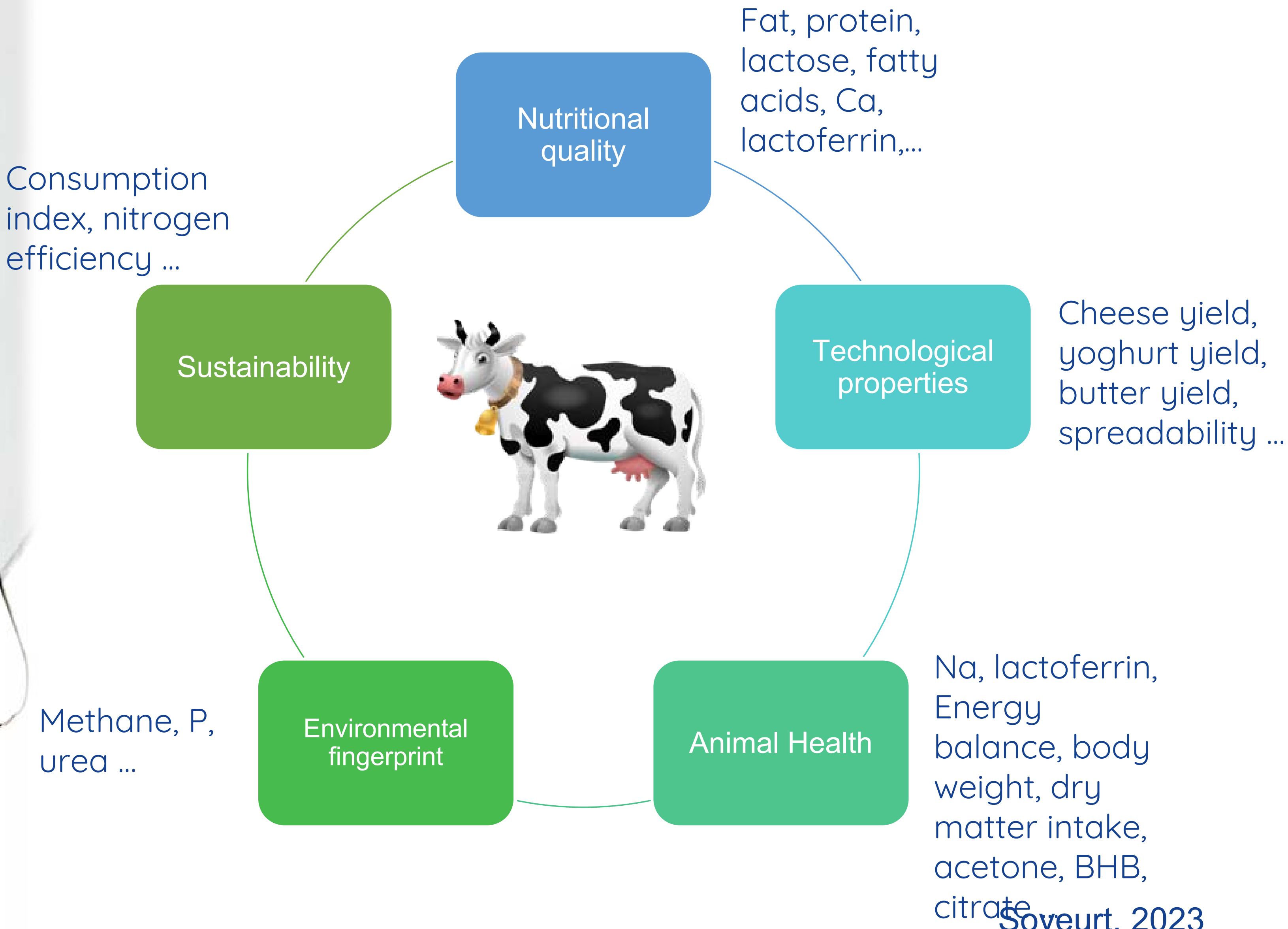
Interreg North-West Europe
Co-funded by the European Union
HoliCow

ICAR meeting, bled, May 2024

LIÈGE université Gembloux Agro-Bio Tech
Wallonie recherche CRA-W
eleveo



Milk Recording
©Avelino Calvar Martinez
©Soyeurt Hélène
Milk samples
©Foss
Milk FT-MIR
Absorbance [Log(1/T)]
Wavenumbers (cm⁻¹)
© Grelet et al., 2015
MIR-based equation
© Shopify Partners
©Avelino Calvar Martinez



Soyeurt, 2023

Interreg
Co-funded by the European Union
North-West Europe

HoliCow

Can we applied all MIR equations on those data as they were built on different calibratio set ?

Partner	Country	N° spectra	N° herds	N° cows	Nbreed
Elevéo	BEL	5,813,993	1,508	317,674	11
Qualitas	CHE	3,672,804	1,890	214,404	8
LKV BW	DEU	2,137,394	503	99,814	15
LKV NRW	DEU	2,142,664	2,388	260,941	17
LKV SH	DEU	8,882,066	2,142	432,598	16
Eliance	FRA	38,353,951	20,824	2,555,698	22
ICBF	IRL	27,610	153	10,663	12
CONVIS	LUX	1,248,653	556	34,629	8

+/- 63,000,000 spectra
6 countries
22 different breeds

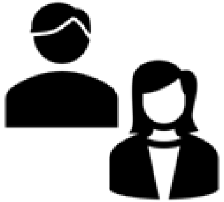
Interreg
Co-funded by the European Union
North-West Europe

HoliCow

Reference values

Fatty acids, minerals, BHB, acetone, methane …

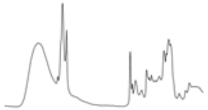
Spectral data

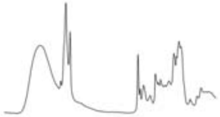


Equation designer

MIR-based equation

?





No spectral extrapolation

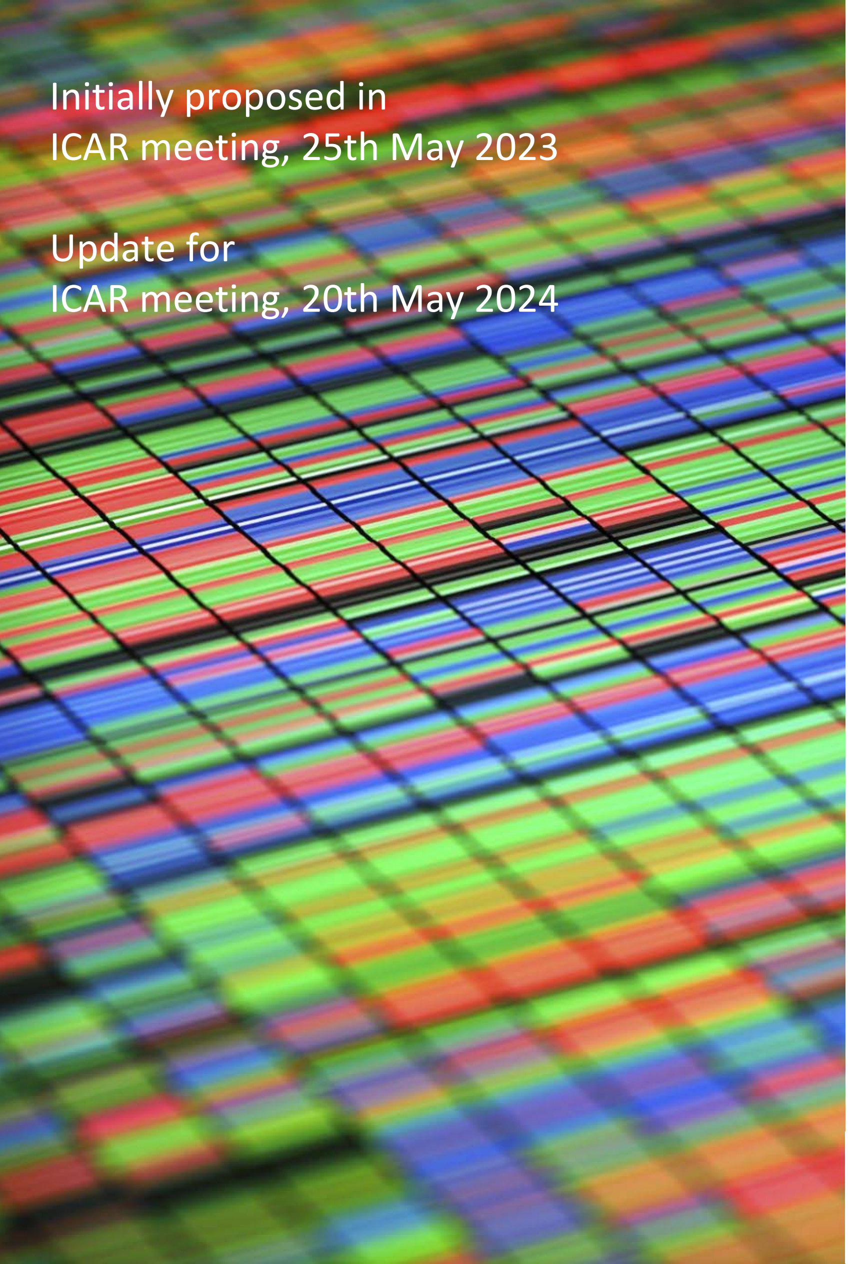
WRSD

Compare your calibration set with the WRSD

Spectral reduction

N = 52,303,184

WRSD reduction
(Soyeurt et al., 2024)



Initially proposed in
ICAR meeting, 25th May 2023

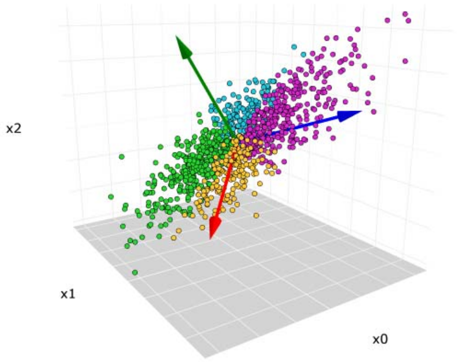
Update for
ICAR meeting, 20th May 2024

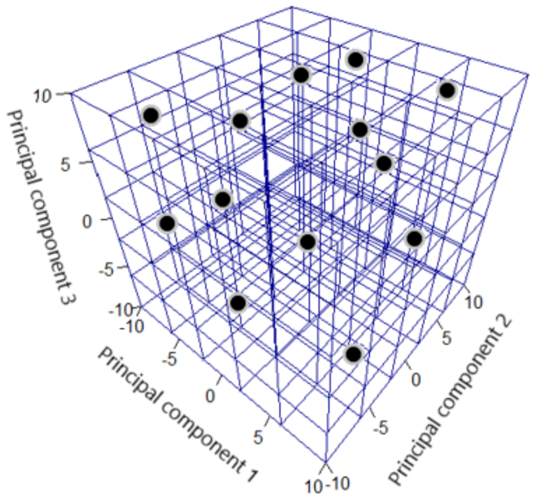
World representative spectral database (WRSD)

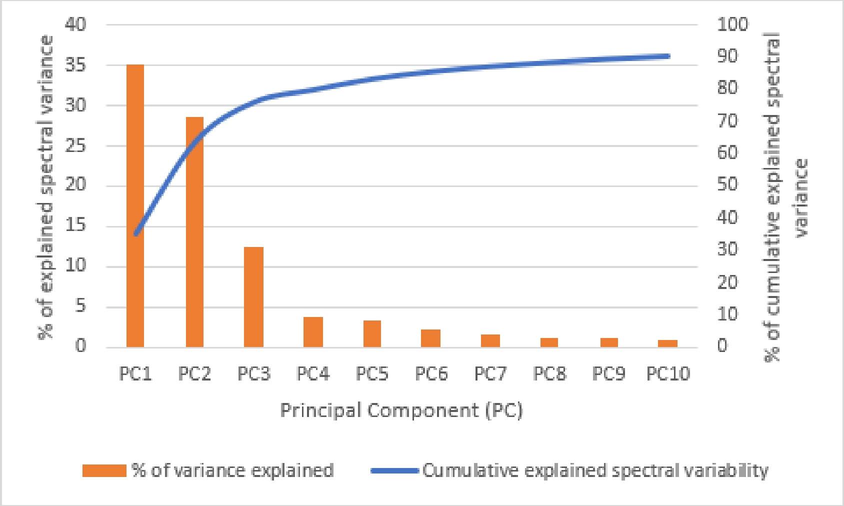
H. Soyeurt, C. Nickmilder, S. Franceschini, M. Whittaker, F. Dehareng, M. Bahadi, J. Leblois, L. Dale, K. Sanders, C. Grelet

Use the scores on the PC

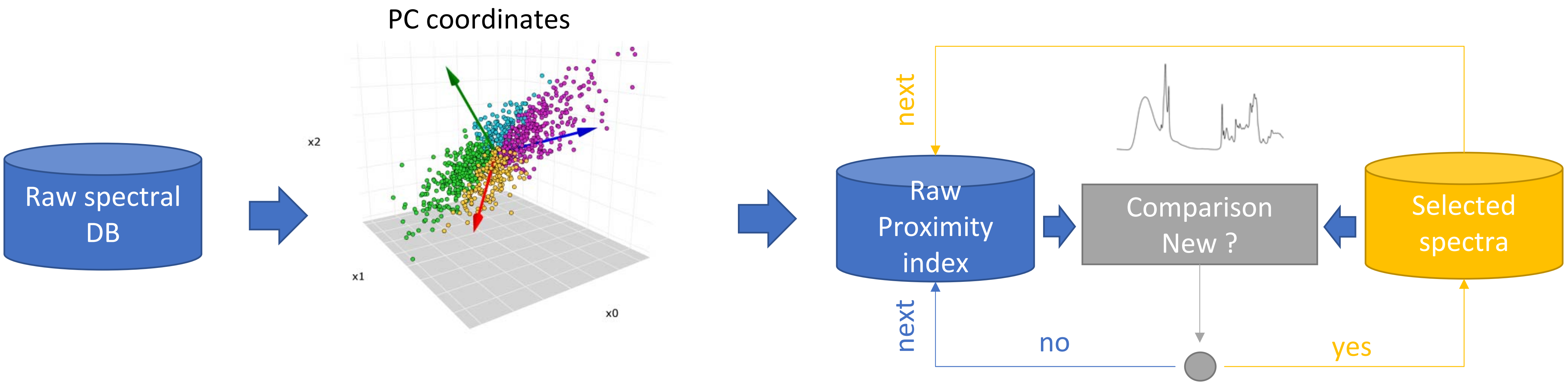
- Scores on the Principal Components
- Create a grid based on the scores on PC
- Only one iteration







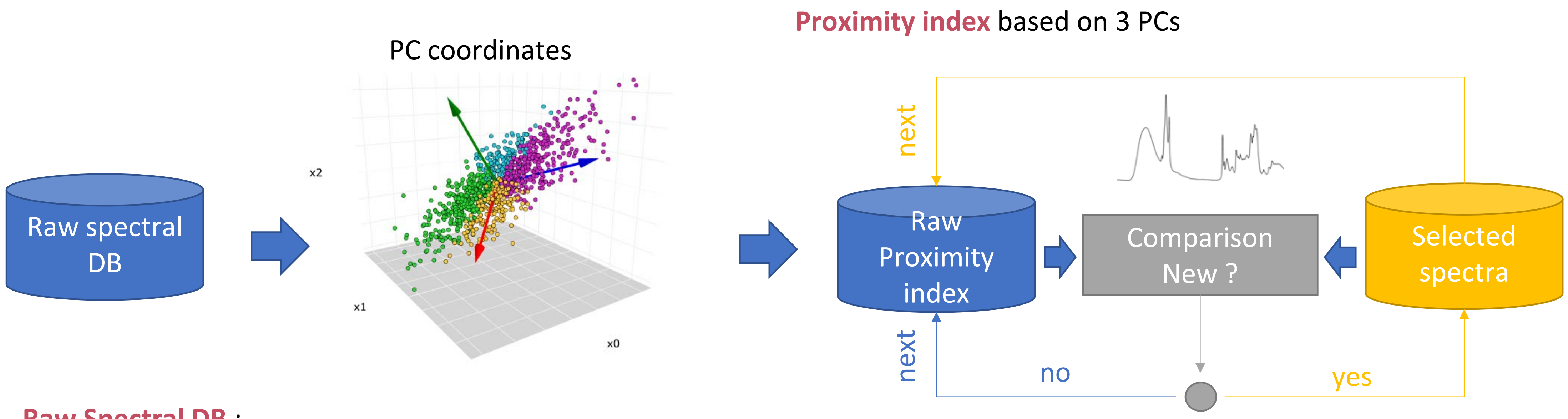
More PCs ➔ More explained spectral variability



The code was written in Python in order to be run easily in each data center.

Proximity index :

•Paste(round(PC1;0),round(PC2;0),round(PC3;0))



Raw Spectral DB :

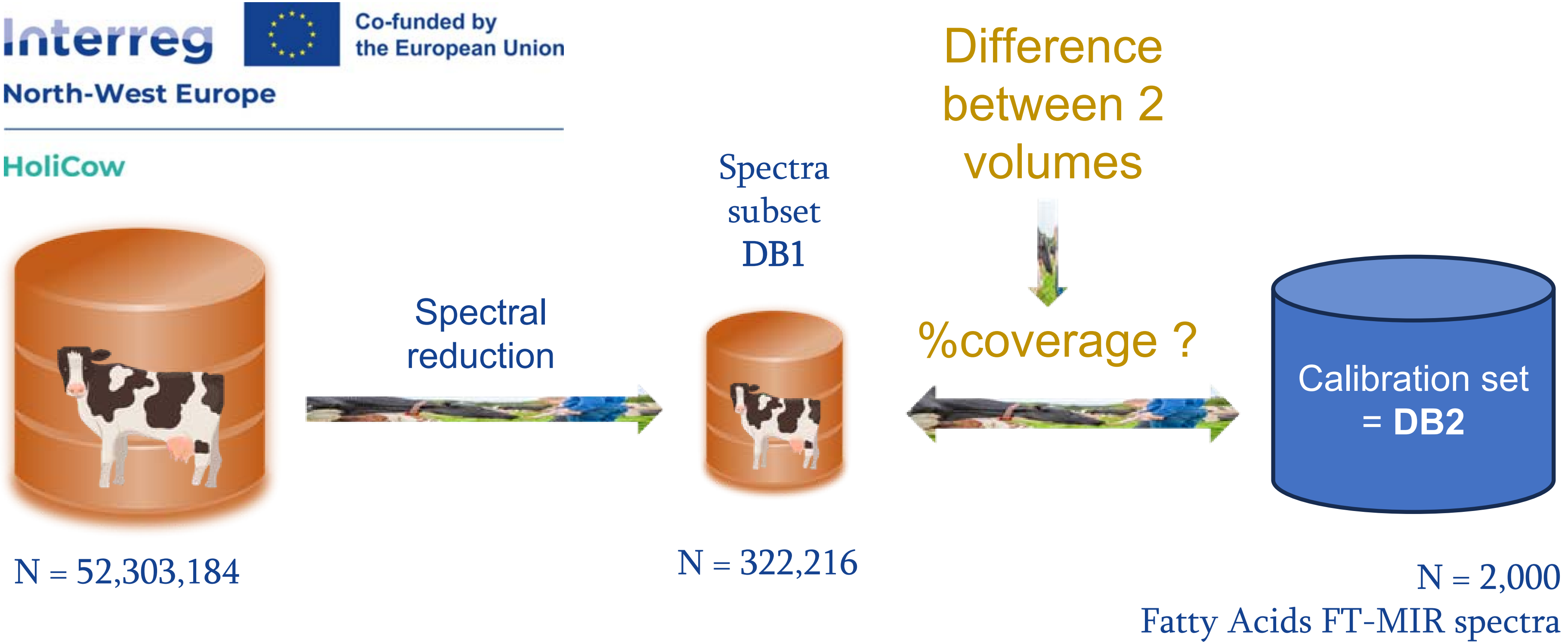
- Foss Database (DB)
  - Eleveo Spectral DB : ± 8,000,000 records
  - Lactanet Spectral DB : ± 10,000,000 records
- Bentley Database:
  - LIC Spectral DB : ± 2,000,000 records
  - LKV-BW Spectral DB : ± 10,000,000 records

Final Selection :

- Foss Database
  - Eleveo Spectral DB : **167,015** records
  - Lactanet Spectral DB : **172,469** records
- Bentley Database:
  - LIC Spectral DB : **81,080** records
  - LKV-BW Spectral DB : **91,494** records

An article was written about the methodology and submitted to the Journal of Dairy Science

Interreg
Co-funded by the European Union
North-West Europe
HoliCow

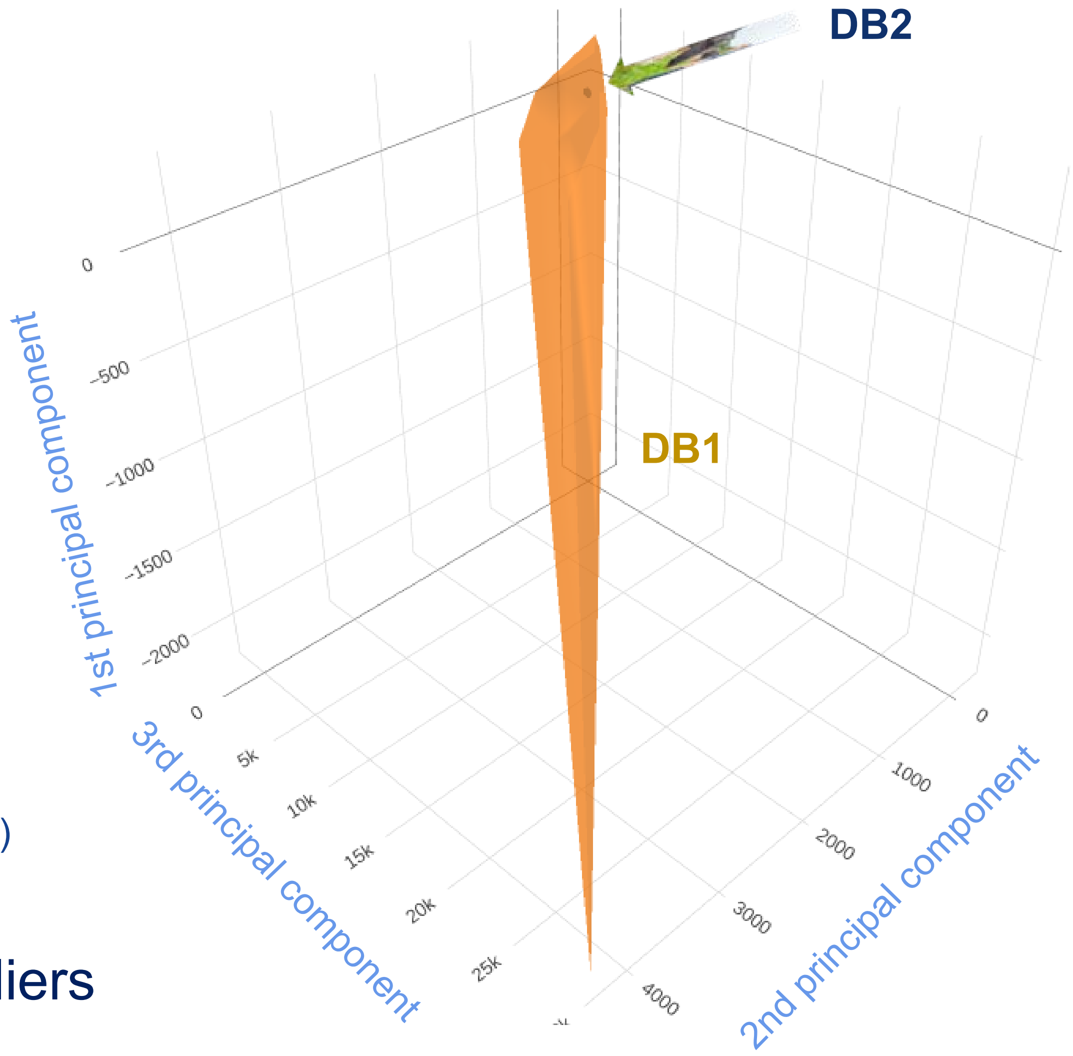


To conserve the information about the spectral **distribution in the studied cow population**, the frequency of the selected spectra is calculated

Interreg North-West Europe
Co-funded by the European Union

HoliCow

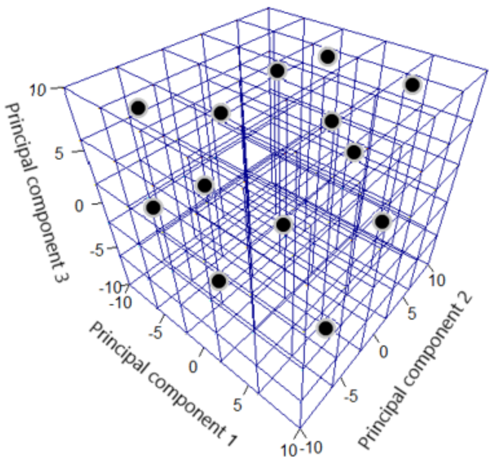
Visualisation of the full Holicow subset (DB1)



Volume drawn by the subset

Volume drawn by the calibration set (DB2)

Presence of outliers



Data Cleaning

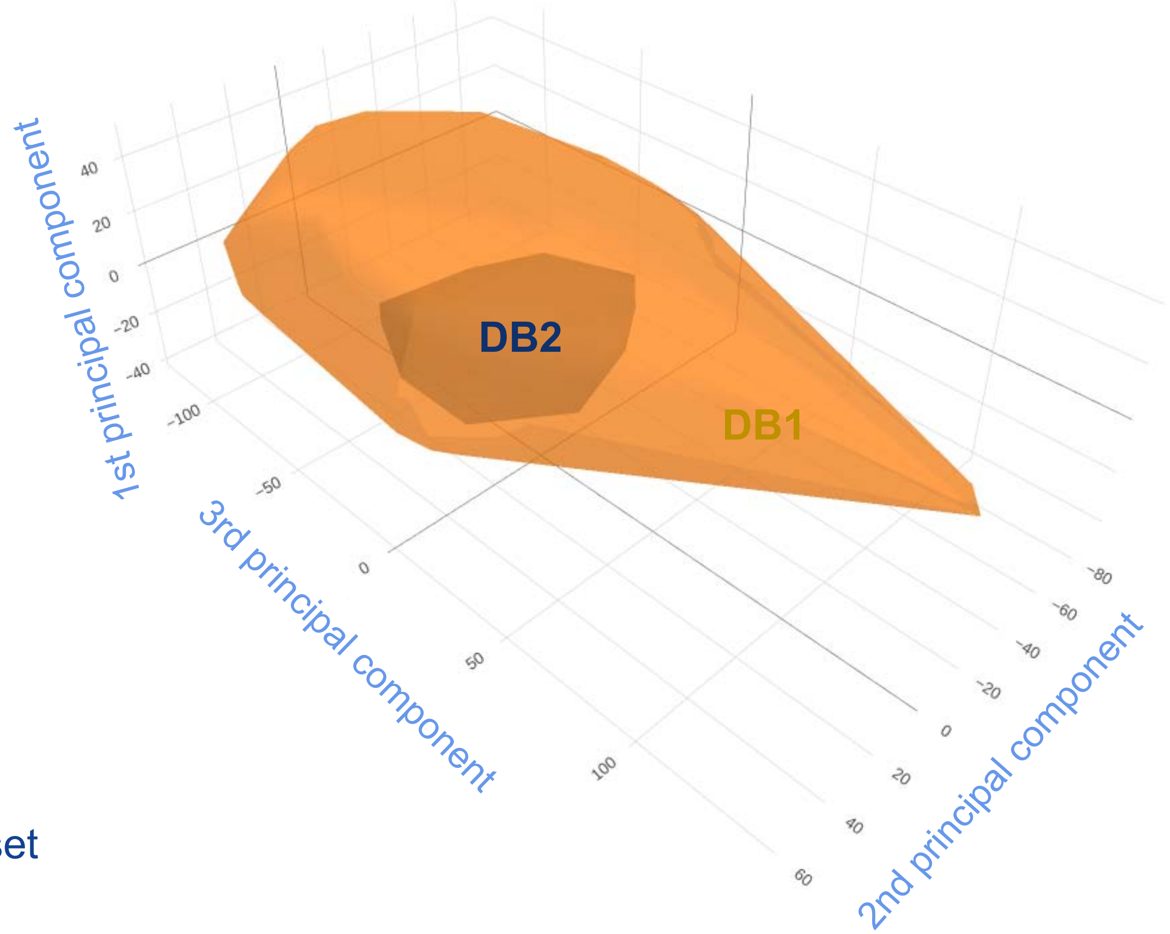
- Standardized Mahalannobis distance < 5

Cleaning	Holicow WRSD (DB1)	%kept
Raw	322,216	100%
Cleaned	275,593	85.53%

Interreg North-West Europe
Co-funded by the European Union

HoliCow

Cleaned Holicow subset



Volume drawn by the subset

Volume drawn by the calibration set

Interreg
Co-funded by the European Union
North-West Europe

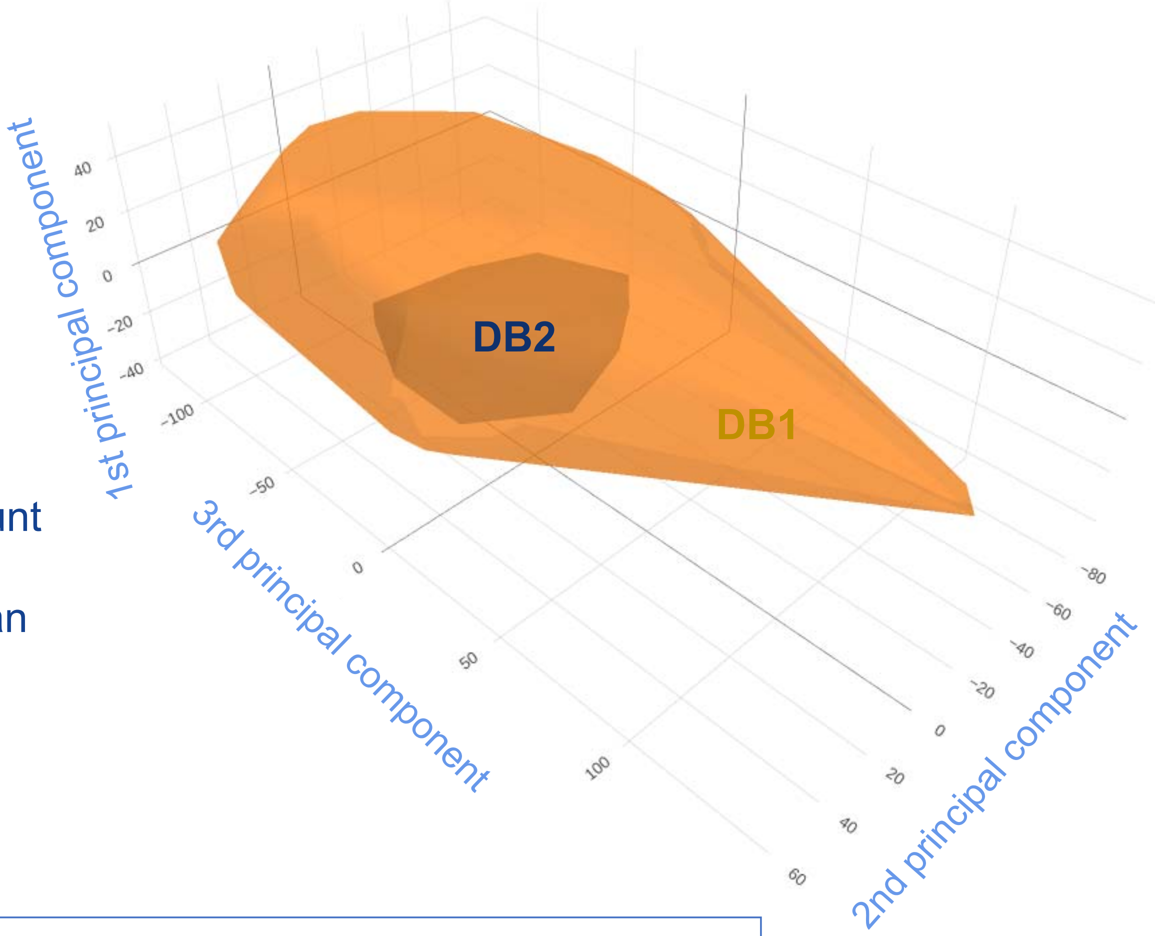
HoliCow

Without taken into account the density
DB2 Volume = **0.12** * DB1 Volume

DB2 covers **99.64%** of the spectral variability of DB1 if we take into account the **density in the population**

➔Some spectra are more present than others

➔More important to be in the area in which the frequency of spectra is the higher.



Fatty acids equations can be applied

Conclusions

- WRSD method selects, as expected, outliers → **data cleaning**
- Need to consider the **frequency** of each spectra to have a correct conclusion.
- Even if the calibration set was limited, it has a variability that allow to cover the variability in the Holicow Database → **FA equations can be applied**.

How spectrally representative are datasets used to build MIR-based predictive models ? A data-driven study.

C. Nickmilder[1], J. Leblois[2], O. Christophe[3], C. Grelet[3], Holicow Consortium, H. Soyeurt[1]

[1]Gembloux Agro-Bio Tech, University of Liège, Belgium
[2]Eleveo, Walloon Breeding Association, Ciney, Belgium
[3]Walloon Research Centre (CRA-W), Gembloux, Belgium

Interreg North-West Europe
Co-funded by the European Union
HoliCow

ICAR meeting, bled, May 2024