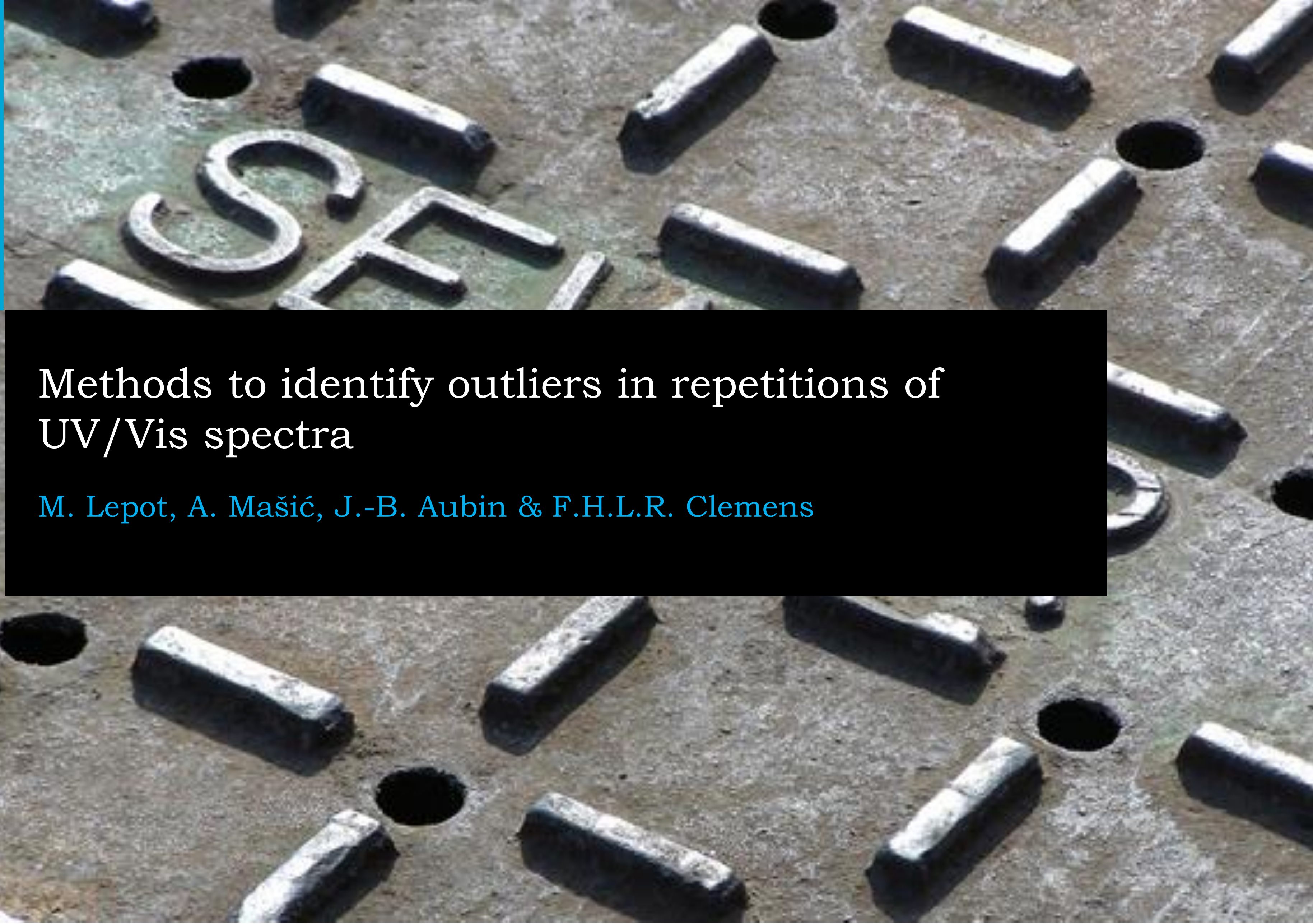
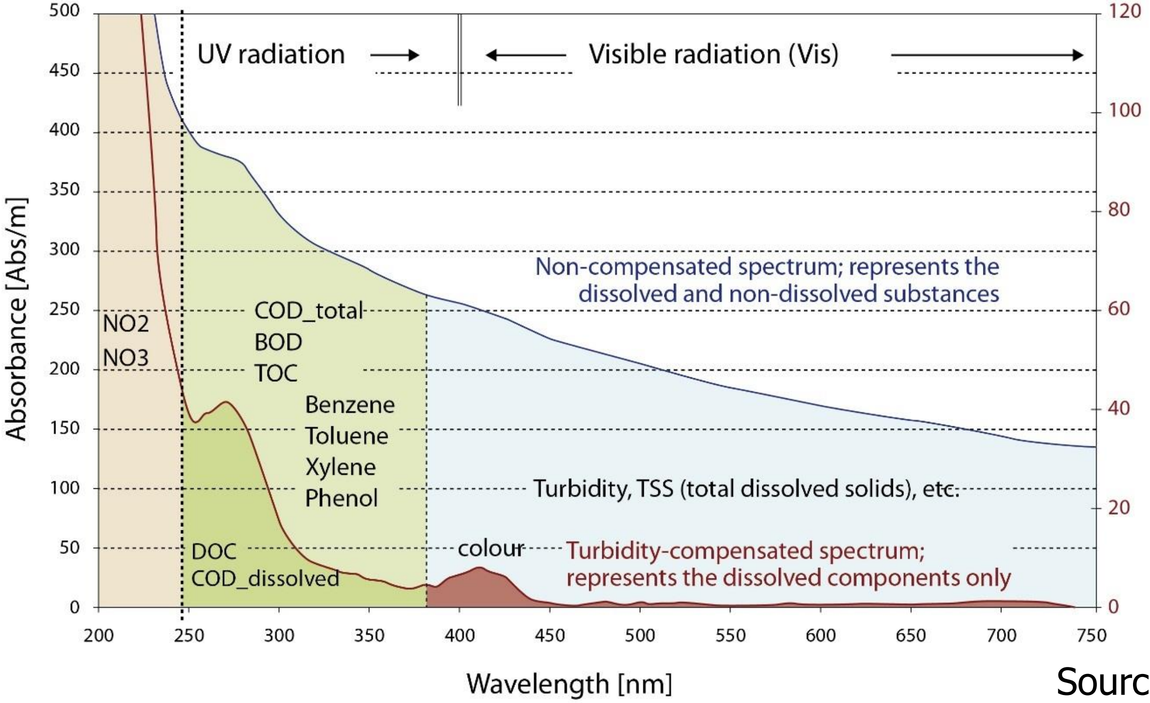


Methods to identify outliers in repetitions of UV/Vis spectra

M. Lepot, A. Mašić, J.-B. Aubin & F.H.L.R. Clemens

Introduction

- UV/Vis spectrophotometers
  - Used since ca. 2000
  - Record absorbance spectra
  - Concentrations (TSS, COD, NO_3, *etc.*)



Source: s::can

Introduction

- A local calibration is needed
    - Using the provided global calibration
    - Using the spectral data

- Spectrum measurement & laboratory analysis

- Repeated measurement of spectra
    - To avoid artefact and potential bias
    - To assess uncertainty

Problems

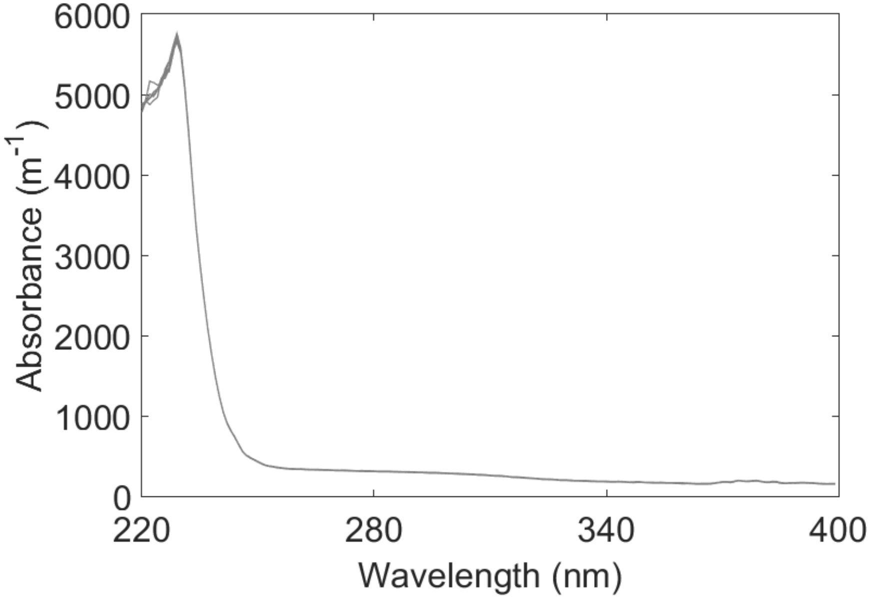
- Is there, at least, one outlier for this sample?

- How to identify spectrum(a) that can be outlier(s)?

- Among the remaining ones, which ones to choose for subsequent calculation (e.g. calibration)?

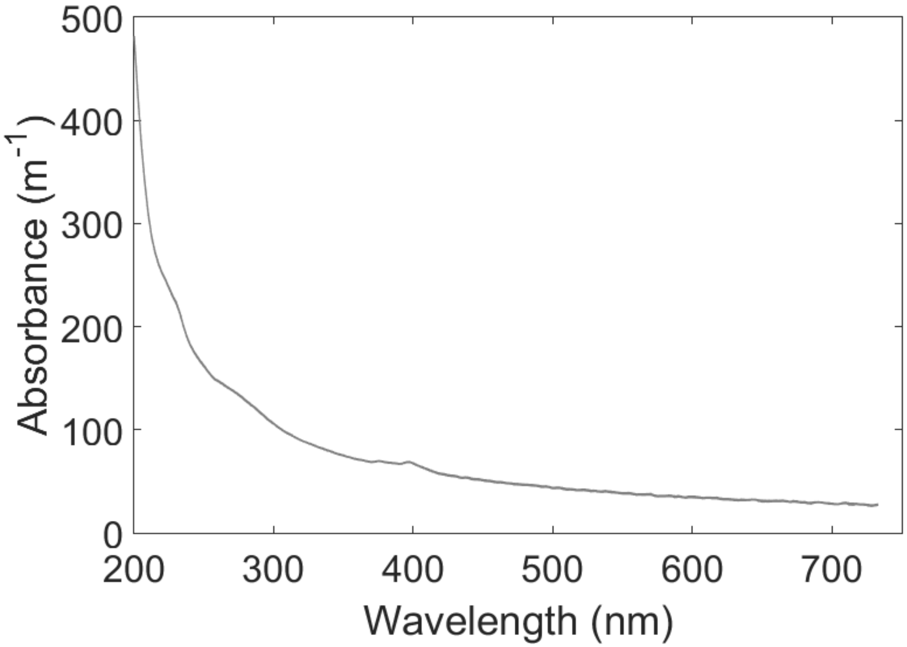
Materials and methods

- Sensors and data sets
  - EAWAG (Switzerland)
    - 4 x 30 samples of (un)filtered/(un)diluted urine
    - 5 spectra recorded per sample
    - (s::can, spectro::lyser, 0.5 mm, UV, 1nm)



Materials and methods

- INSA Lyon (France)
  - 94 samples (inlet of the WWTP, dry weather)
  - Up to 25 spectra recorded per sample
  - (s::can, spectro::lyser, 2 mm, UV/Vis, 2.5nm)



Materials and methods

Methods – two approaches: PCA & DDT/ED

Principal Component Analysis (PCA)

Scores of the first principal component in PCA

- Step 1: Data preprocessing – mean-centering
- Step 2: Singular value decomposition
- Step 3: Score matrix

- Outlier detection
  - PCA_Expert: visual inspection of the PC1 scores
  - PCA_2: automated selection based on mean±2std
- Identification of the MRS
  - PCA_2: smallest distance between PC1 score and median

PCA_Expert

PCA_2

Materials and methods

Data Depth Theory (DDT) & Euclidean Distance (ED)

- Outlier detection (DDT or ED)
  - ED

$$ED_j = \frac{1}{N_T}\sqrt{\sum_{i=1}^{n_x}\left(Abs_{j,i} - Abs_{k\neq j,i}\right)^2}$$

$$ED_j > k_M \times median\left(\left[ED_1 \colon ED_{N_T}\right]\right)$$

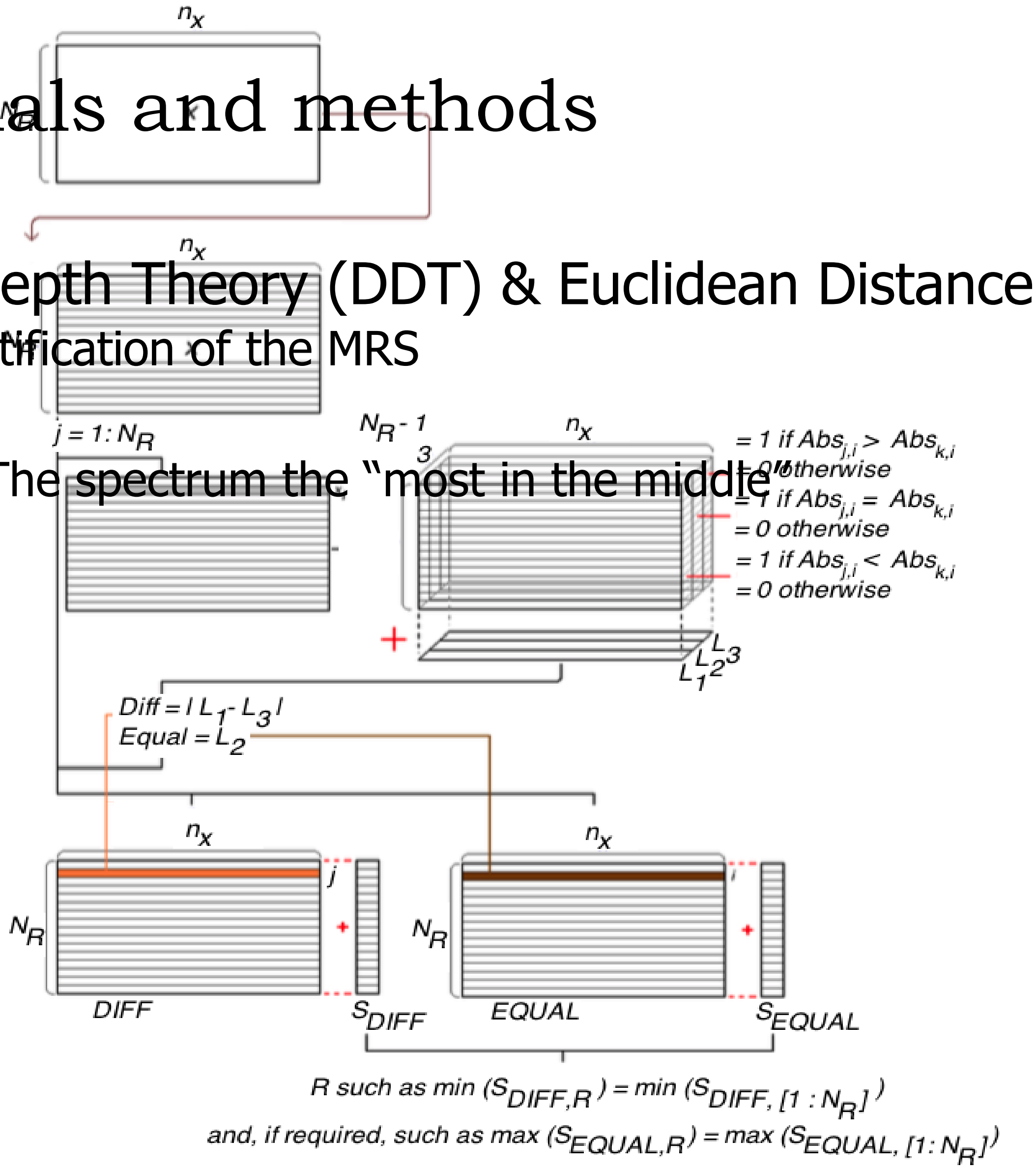
DDT_ED_1
DDT_ED_2
DDT_ED_3

  - DDT
    - Removal of every spectra always below or above the other ones

DDT_DDT

Materials and methods

- Data Depth Theory (DDT) & Euclidean Distance (ED)
  - Identification of the MRS
    - The spectrum the "most in the middle"



n_x
N_R
n_x
N_R
$j = 1 : N_R$
$N_R - 1$
3
n_x
$= 1$ if $Abs_{j,i} > Abs_{k,i}$
$= 0$ otherwise
$= 1$ if $Abs_{j,i} = Abs_{k,i}$
$= 0$ otherwise
$= 1$ if $Abs_{j,i} < Abs_{k,i}$
$= 0$ otherwise
+
L_3
L_2
L_1
$Diff = | L_1 - L_3 |$
$Equal = L_2$
n_x
n_x
j
N_R
+
N_R
+
DIFF
S_{DIFF}
EQUAL
S_{EQUAL}
R such as min $(S_{DIFF,R}) = \min (S_{DIFF, [1 : N_R]})$
and, if required, such as max $(S_{EQUAL,R}) = \max (S_{EQUAL, [1 : N_R]})$

Results

- Identification of sample containing outlier(s): ca 75% of consistency (TP – TN / FP – FN)

WWTP (94 samples)						
Method	DDT_ED_1 94	DDT_ED_2 89	DDT_ED_3 69	DDT_DDT 39	PCA_Expert 82	PCA_2 60
DDT_ED_1	-	89 - 0 / 5 - 0	69 - 0 / 25 - 0	39 - 0 / 55 - 0	82 - 0 / 12 - 0	60 - 0 / 34 - 0
DDT_ED_2		-	69 - 5 / 20 - 0	41 - 5 / 48 - 0	81 - 3 / 9 - 1	60 - 6 / 28 - 0
DDT_ED_3			-	39 - 25 / 29 - 1	66 - 8 / 4 - 16	60 - 22 / 10 - 2
DDT_DDT				-	38 - 11 / 1 - 44	34 - 29 / 5 - 26
PCA_Expert					-	58 - 11 / 23 - 2
PCA_2						-

Results

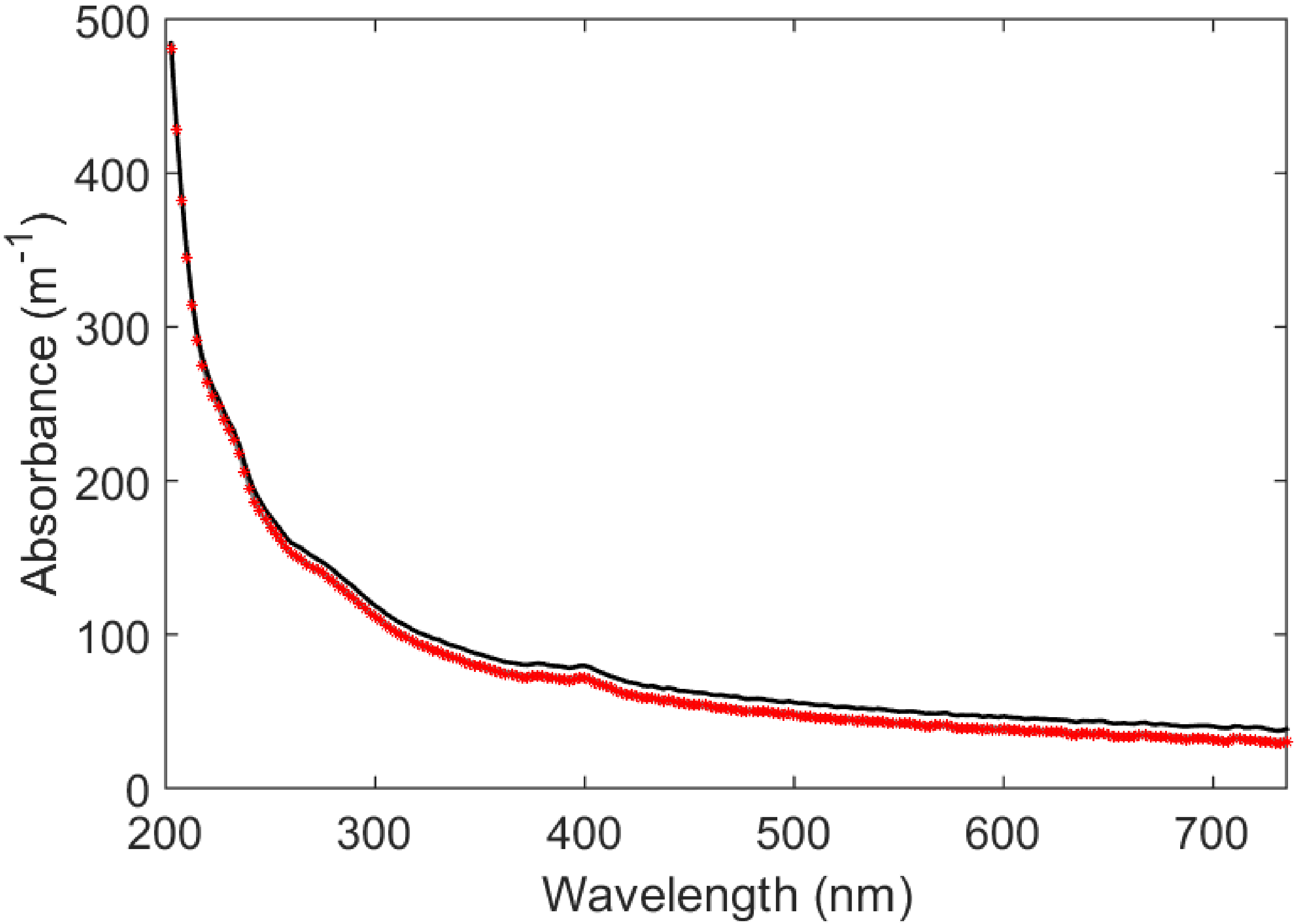
FU (30 samples)						
Method	DDT_ED_1 30	DDT_ED_2 7	DDT_ED_3 4	DDT_DDT 0	PCA_Expert 6	PCA_2 0
DDT_ED_1	-	7 - 0 / 23 - 0	4 - 0 / 26 - 0	0 - 0 / 30 - 0	6 - 0 / 24 - 0	0 - 0 / 30 - 0
DDT_ED_2		-	4 - 23 / 3 - 0	0 - 23 / 7 - 0	5 - 22 / 2 - 1	0 - 23 / 7 - 0
DDT_ED_3			-	0 - 26 / 4 - 0	4 - 24 / 0 - 2	0 - 26 / 4 - 0
DDT_DDT				-	0 - 24 / 0 - 6	0 - 30 / 0 - 0
PCA_Expert					-	0 - 24 / 6 - 0
PCA_2						-

Results

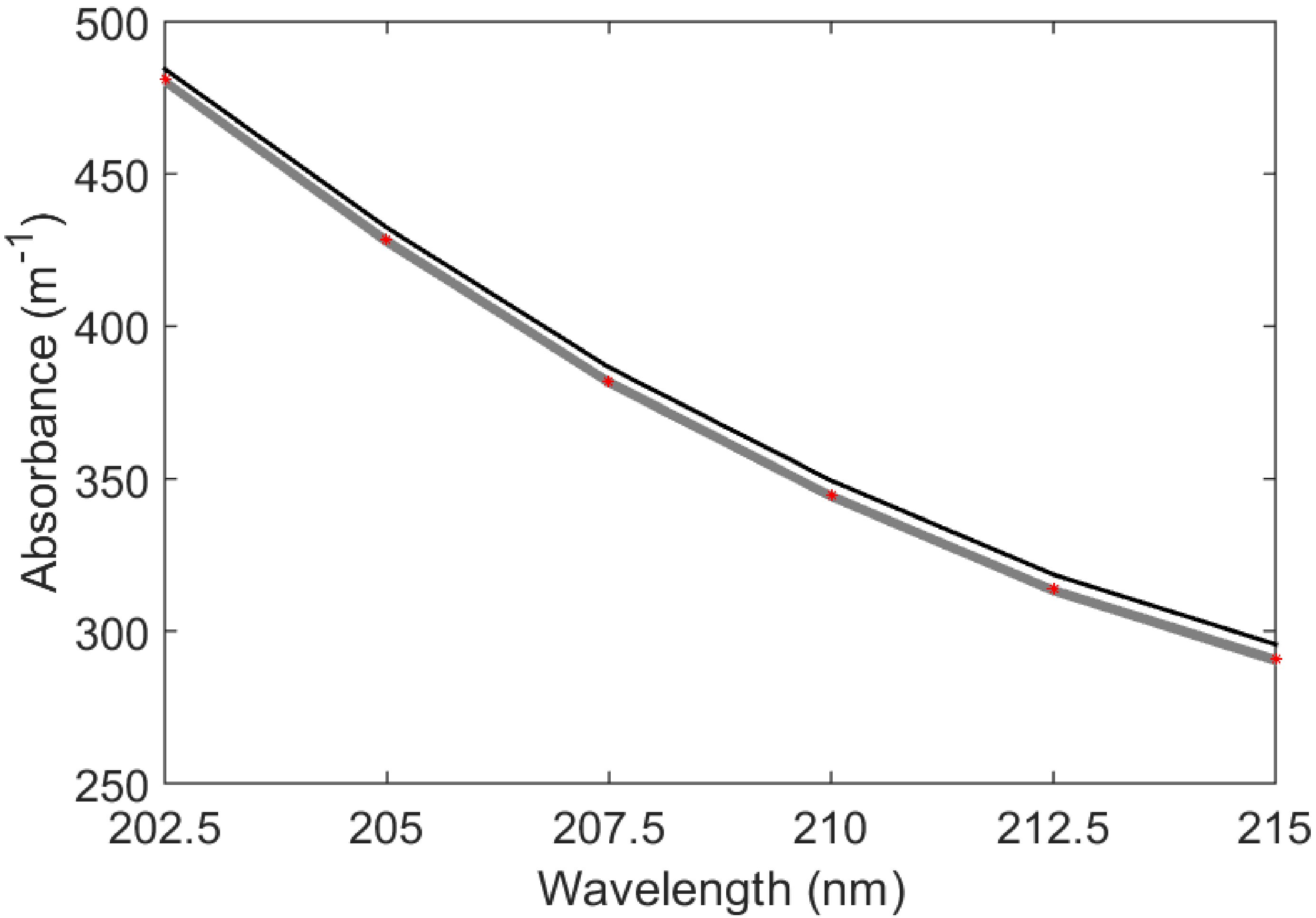
- Identification of the outlier(s): ca 95% of consistency (consistency ratios for at least one outlier in common)

WWTP (94 samples)						
Method	DDT_ED_1 94	DDT_ED_2 89	DDT_ED_3 69	DDT_DDT 39	PCA_Expert 82	PCA_2 60
DDT_ED_1	1	1	1	1	0.99	1
DDT_ED_2		1	1	1	1	1
DDT_ED_3			1	0.97	0.99	0.99
DDT_DDT				1	0.98	0.99
PCA_Expert					1	1
PCA_2						1

Results


500
400
300
200
100
0
Absorbance (m^{-1})
200
300
400
500
600
700
Wavelength (nm)

Results

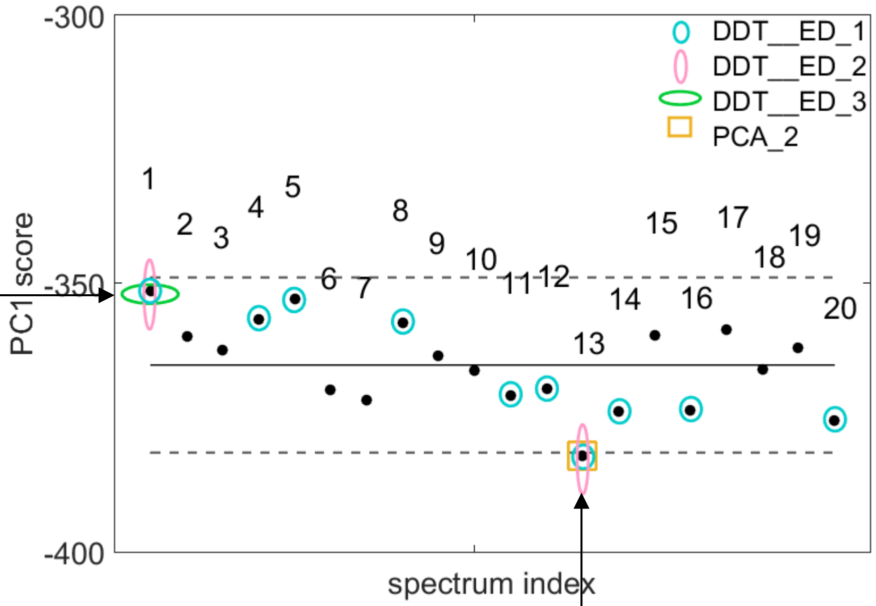

500
450
400
350
300
250
202.5
205
207.5
210
212.5
215
Absorbance (m^{-1})
Wavelength (nm)

Results

UD (30 samples)						
Method	DDT_ED_1 30	DDT_ED_2 4	DDT_ED_3 3	DDT_DDT 0	PCA_Expert 2	PCA_2 0
DDT_ED_1	1	1	1	--	0.5	--
DDT_ED_2		1	1	--	NSWOIC	--
DDT_ED_3			1	--	NSWOIC	--
DDT_DDT				1	--	--
PCA_Expert					1	--
PCA_2						1

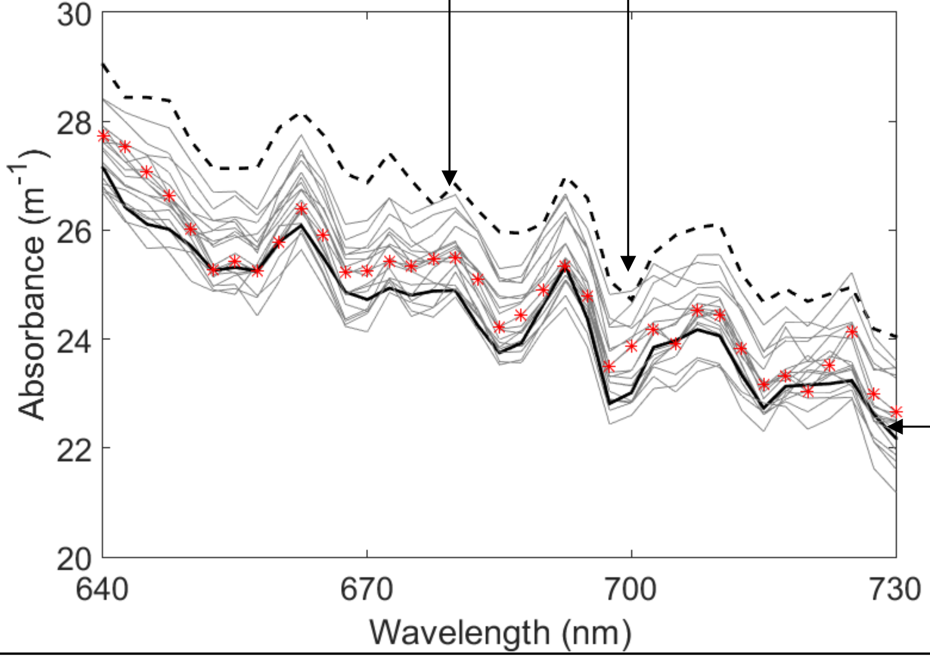
Results

Or by PCA



Identified by DDT_ED

But not by DDT_DDT



And by PCA_2

Identified by DDT_ED_1 and DDT_ED_2

Results

- Identification of the MRS: ca 28% of consistency (consistency ratios for the MRS identification)

WWTP (94 samples)					
Method	DDT_ED_1	DDT_ED_2	DDT_ED_3	DDT_DDT	PCA_2
DDT_ED_1	1	0.35	0.35	0.41	0.11
DDT_ED_2		1	0.81	0.71	0.13
DDT_ED_3			1	0.87	0.24
DDT_DDT				1	0.3
PCA_2					1

Conclusions

- Repeated spectra are required
- If only few spectra per sample: DDT_ED_2
- PCA and DDT are equivalent when more spectra are recorded (> 5)
- PCA and DDT are inconsistent for the identification of MRS
- A voting system?

Acknowledgments

- R2DS programme, Ile de France Regional Council (www.r2ds-ile-de-france.com)
- HURRBIS French network of Urban Hydrology Observatories (www.graie.org/hurrbis)
- OTHU project (www.othu.org)
- FP7 PREPARED research project (www.prepared-fp7.eu)
- MAC-Nut project, Eawag Discretionary Funds (5221.00492.007.10, www.eawag.ch)

- Dr. Kris Villez (Eawag, Switzerland)
- Ana Santos (U. Nova de Lisboa, Portugal)

Acknowledgments: QUICS

The University Of Sheffield.

University of BRISTOL

TU Delft Delft University of Technology

Aquafin

WATERWAYS

RTC4Water Real Time Control of Water Infrastructure

ch2m.

WAGENINGEN UNIVERSITY WAGENINGEN UR

RUHR UNIVERSITÄT BOCHUM RUB

Witteveen + Bos

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

UNIVERSIDADE DE COIMBRA

UNIVERSITÉ LAVAL

CTGA Centro Tecnológico de Gestão Ambiental

TECHNISCHE UNIVERSITÄT DRESDEN

LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY LIST

eawag aquatic research

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607000.

www.quics.eu