WP5: Preliminary analysis of the Aquavalens matrix with the ICHNAEA® software for Microbial Source Tracking

Aquavalens Project meeting, Alacant
November 19, 2014

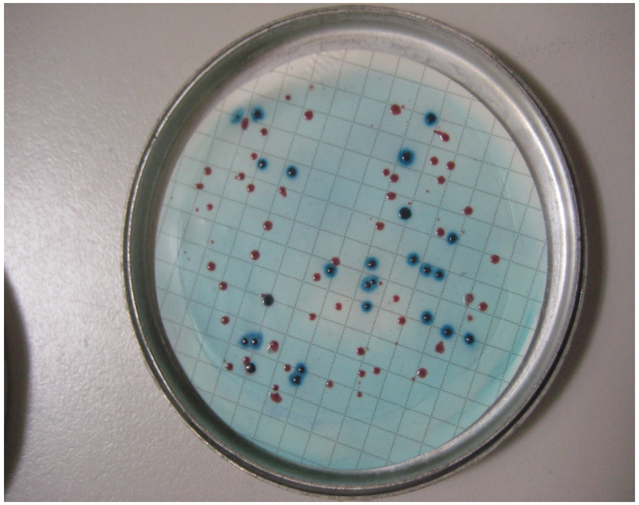
Elisenda Ballesté	eballeste@ub.edu
Lluís A. Belanche	belanche@cs.upc.edu
Anicet R. Blanch	ablanch@ub.edu
Diego Olano	diegoolano@gmail.com

UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

U UNIVERSITAT DE BARCELONA B

Challenges in MST

- Independence of geographical location
- Discriminating ability of different subsets of indicators
- Persistence of indicators in time (including seasonality effects)
- Effects of dilution in watersheds
- Small sample sizes
- Presence of compound mixtures from several distinct animal species



What is ICHNAEA?

- Integrated computer-based prediction system for **Microbial Source Tracking** studies.

- Accepts microbiological measurements:
  - showing different *concentration levels*
  - showing different *environmental persistences*
  - from different *origins* and *geographic areas*
  - from different *users*

WP5 data matrix for MST (1)

118 observations:

Site

DVGW	25
IST	25
TU WIEN	24
UB	26
UH	18

HM	33
PG	24
PL	24
CW	23
(other)	14

Season

Summer	63	(April to September)
Winter	55	(October to March)

WP5 data matrix for MST (2)

Examples in the data matrix are expressed at the **point of source** (non-diluted) and at **zero-time** (fresh)

The data matrix is a **maximal** one:

– only a **fraction** of indicators should be used
– there is an interest to consider **ratios**

WP5 data matrix for MST (3)

Indicators

EC FE CP FeqPCR SomPhg BifTot TLBif AllBac

HMBactPhg CWBactPhg PGBactPhg PLBactPhg

BifSorb HMBif CWBif PGNeo PLBif

BacR Pig2Bac HF183TaqMan

HMMit CWMit PGMit PLMit

Acesulfame Cyclamate Saccharain Sucralose

Adeno Norav

WP5 data matrix for MST (4)

Ratios

```
SomPhg / HMBactPhg
SomPhg / CWBactPhg
SomPhg / PGBactPhg
SomPhg / PLBactPhg

BifTot / BifSorb
TLBif / HMBif
TLBif / CWBif
TLBif / PGNeo
TLBif / PLBif

AllBac / BacR
AllBac / Pig2Bac
AllBac / HF183Taqman
```

Pre-processing of the data matrix

* Identification of observations (104), starting predictive indicators (30), target sources (4), seasons (2), sites (5)

* Harmonization of **volumes**, and **detection limits**

* Dealing with troublesome **special values**: “lower than” (lots of), “higher than” (some), NAs, dnq, nds (some), …

* Calculation of **slopes** for persistences (regression lines, T90, T99, K, %), one for each season (SUMMER/WINTER)

* Creation of **ratios** (12) and log10 of everything:

104 observations, 4 sources, 30+12 indicators

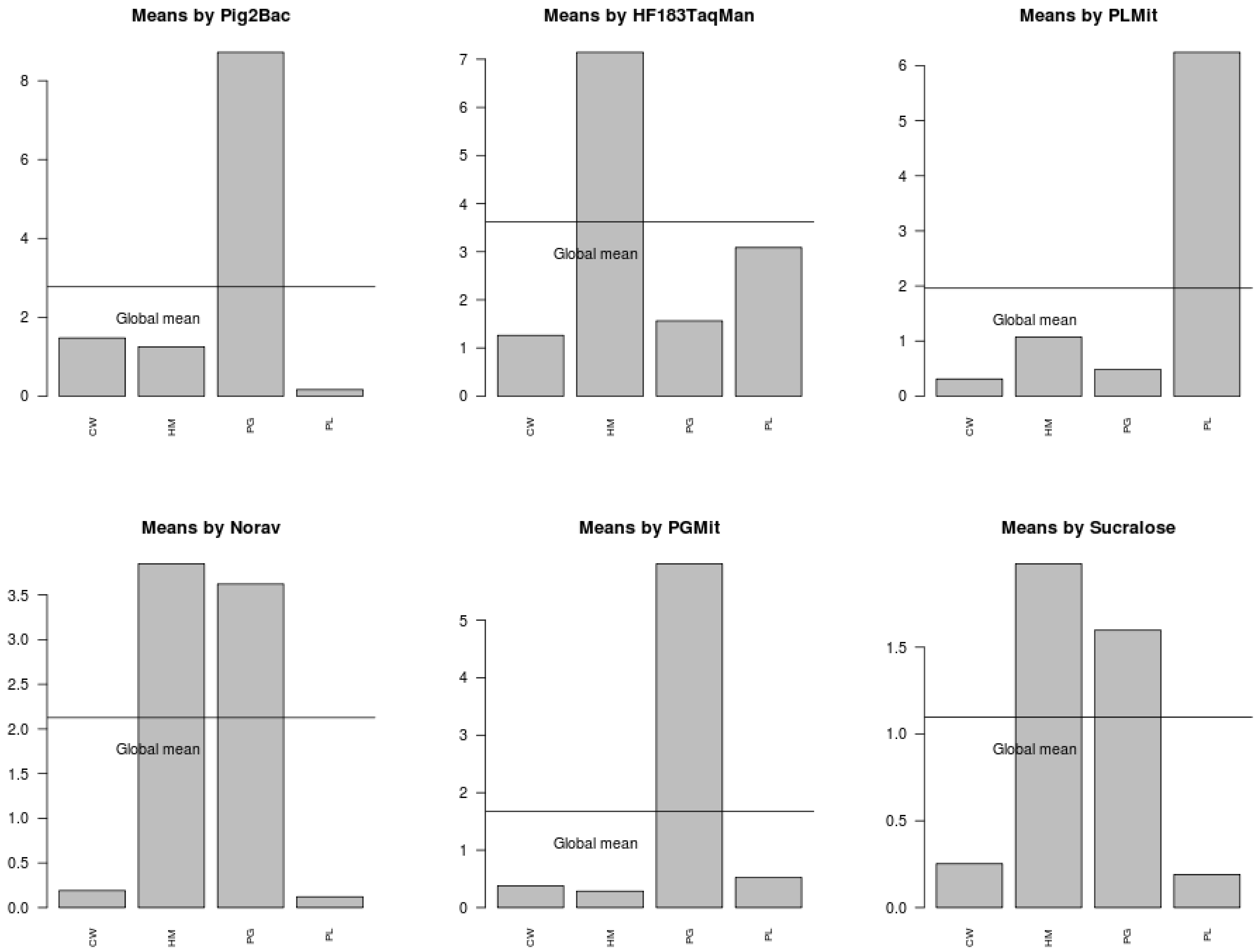
Determining indicator importance (1)

Identification of strongest univariate relationships between **indicators** and **source**

Fisher's F shows that **first 6** are:

- Pig2Bac
- HF183TaqMan
- PLMit
- Norav
- PGMit
- Sucralose

last 3 are FeqPCR, BifTot and CWBactPhg

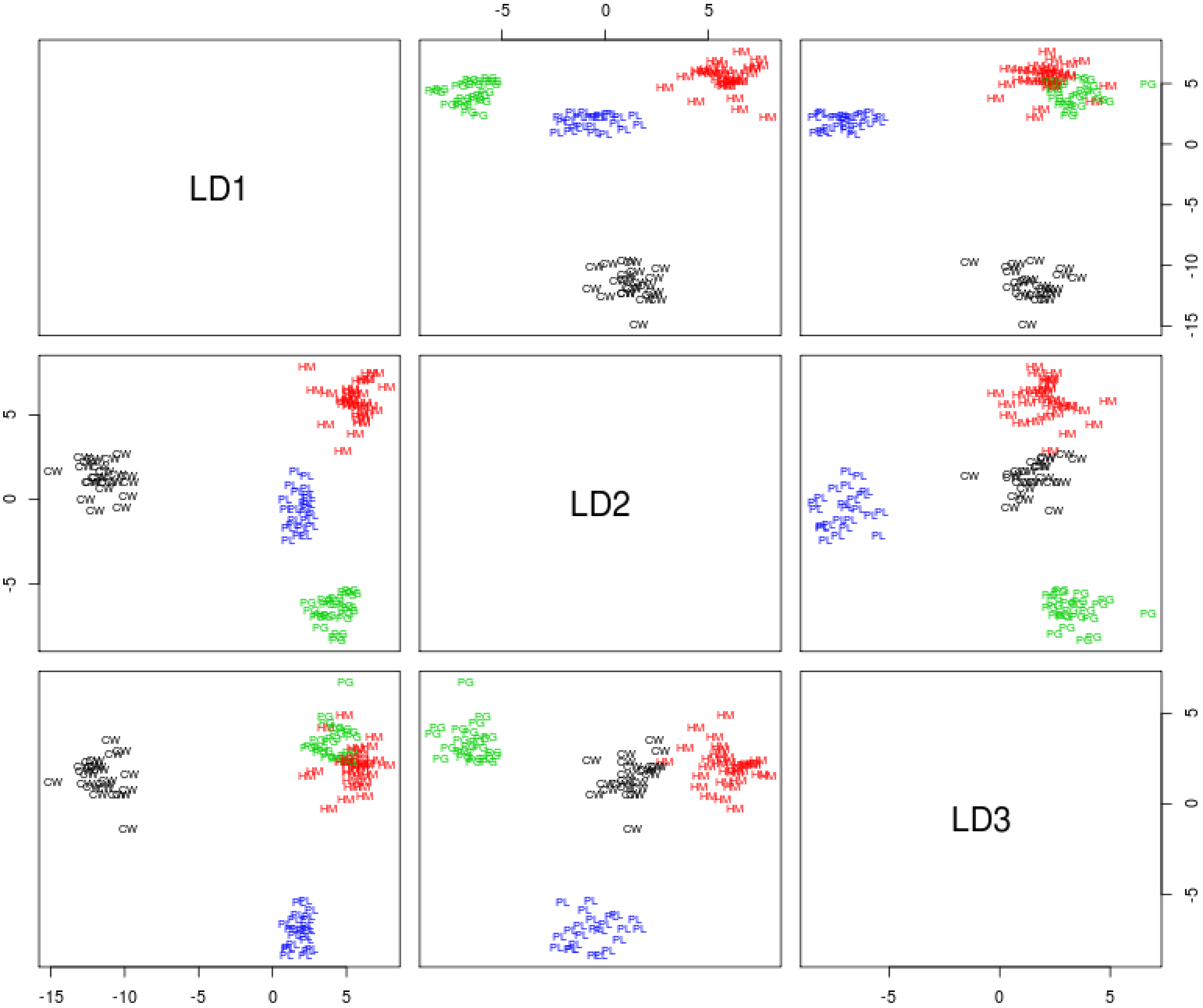

Means by Pig2Bac
0
2
4
6
8
Global mean
CW
HM
PG
PL
Means by HF183TaqMan
0
1
2
3
4
5
6
7
Global mean
CW
HM
PG
PL
Means by PLMit
0
1
2
3
4
5
6
Global mean
CW
HM
PG
PL
Means by Norav
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Global mean
CW
HM
PG
PL
Means by PGMit
0
1
2
3
4
5
Global mean
CW
HM
PG
PL
Means by Sucralose
0.0
0.5
1.0
1.5
Global mean
CW
HM
PG
PL

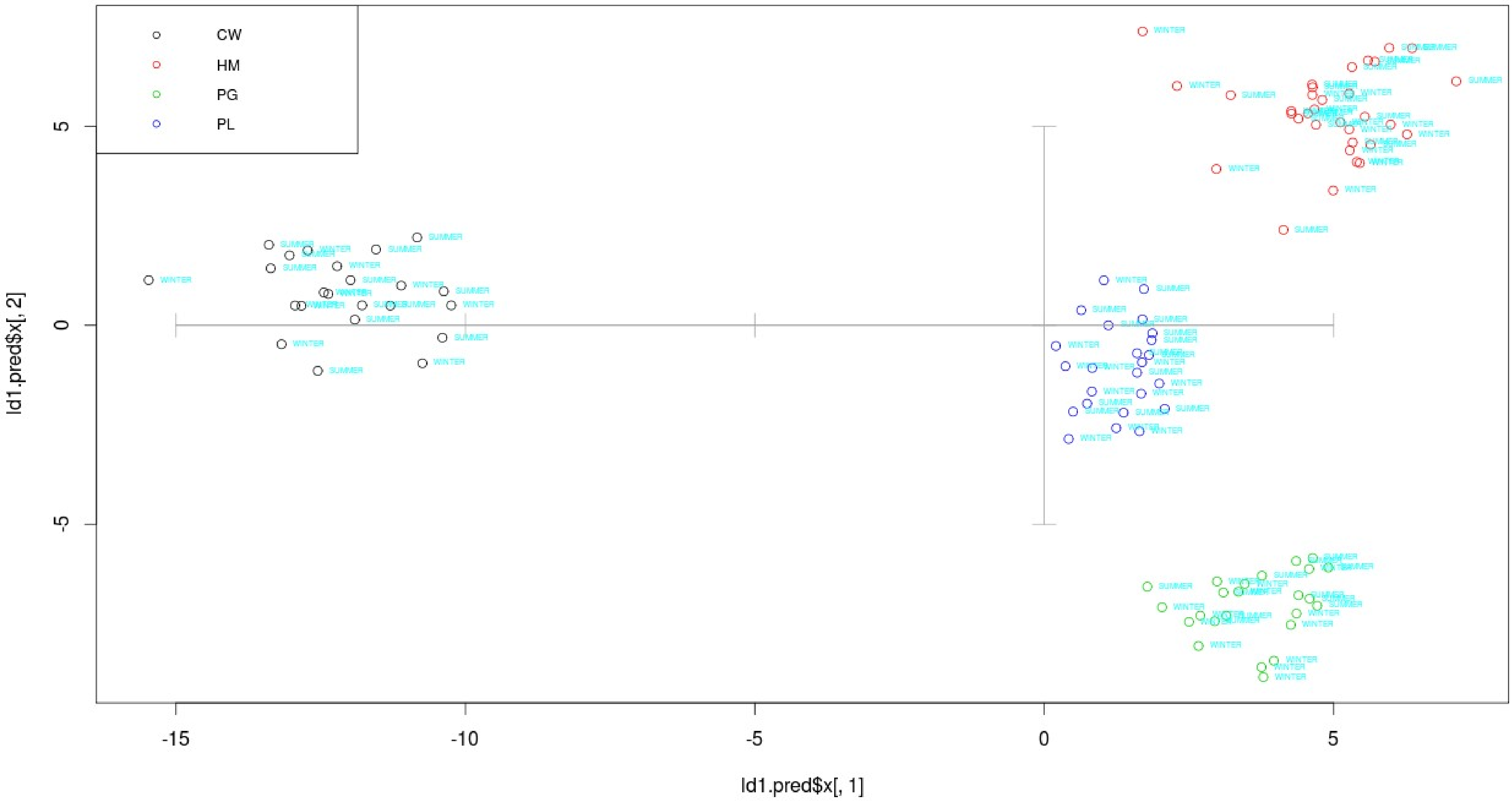
Determining indicator importance (2)

Identification of strongest univariate relationships between **indicators** and **individual sources**

- CW: BacR, CWMit, CWBif
- HM: HMBactPhg,Acesulfame,HF183Taqman,Cyclamate
- PG: Pig2Bac, PGMit
- PL: PLBif, PLMit

(acc. to p-value of test comparing the mean of the group with the global mean)


LD1
LD2
LD3
-5
0
5
-15
-10
-5
0
5
-5
0
5
5
0
-5
-10
-15
5
0
-5
5
0
-5
CW
HM
PG
PL

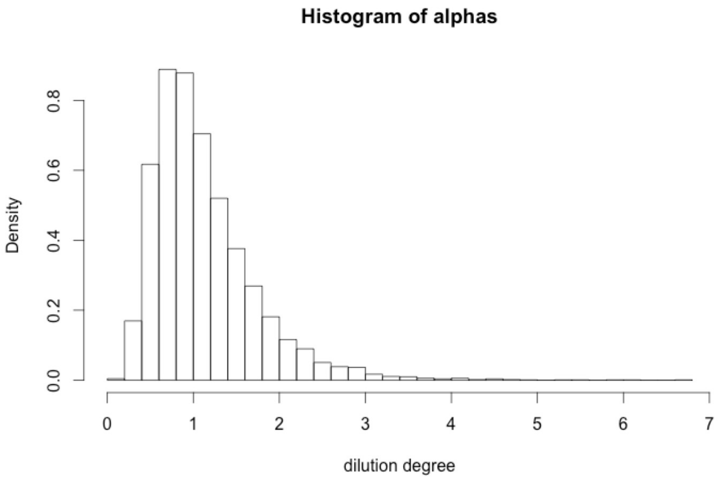


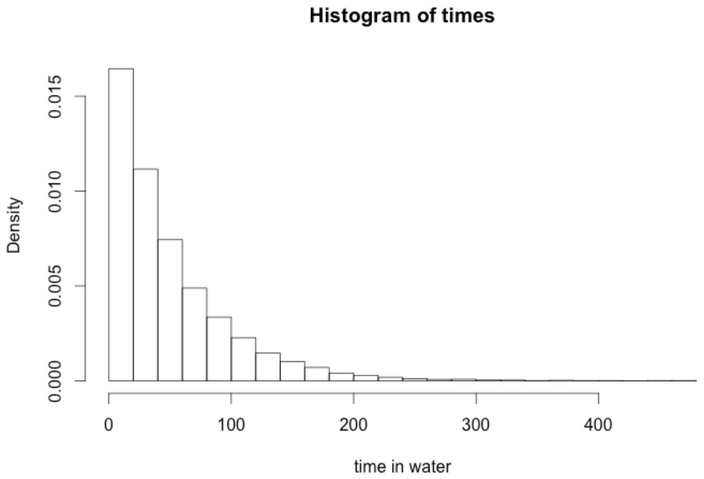
It turns out that {HMBif, BacR, Pig2Bac} suffice to get 100% LOOCV

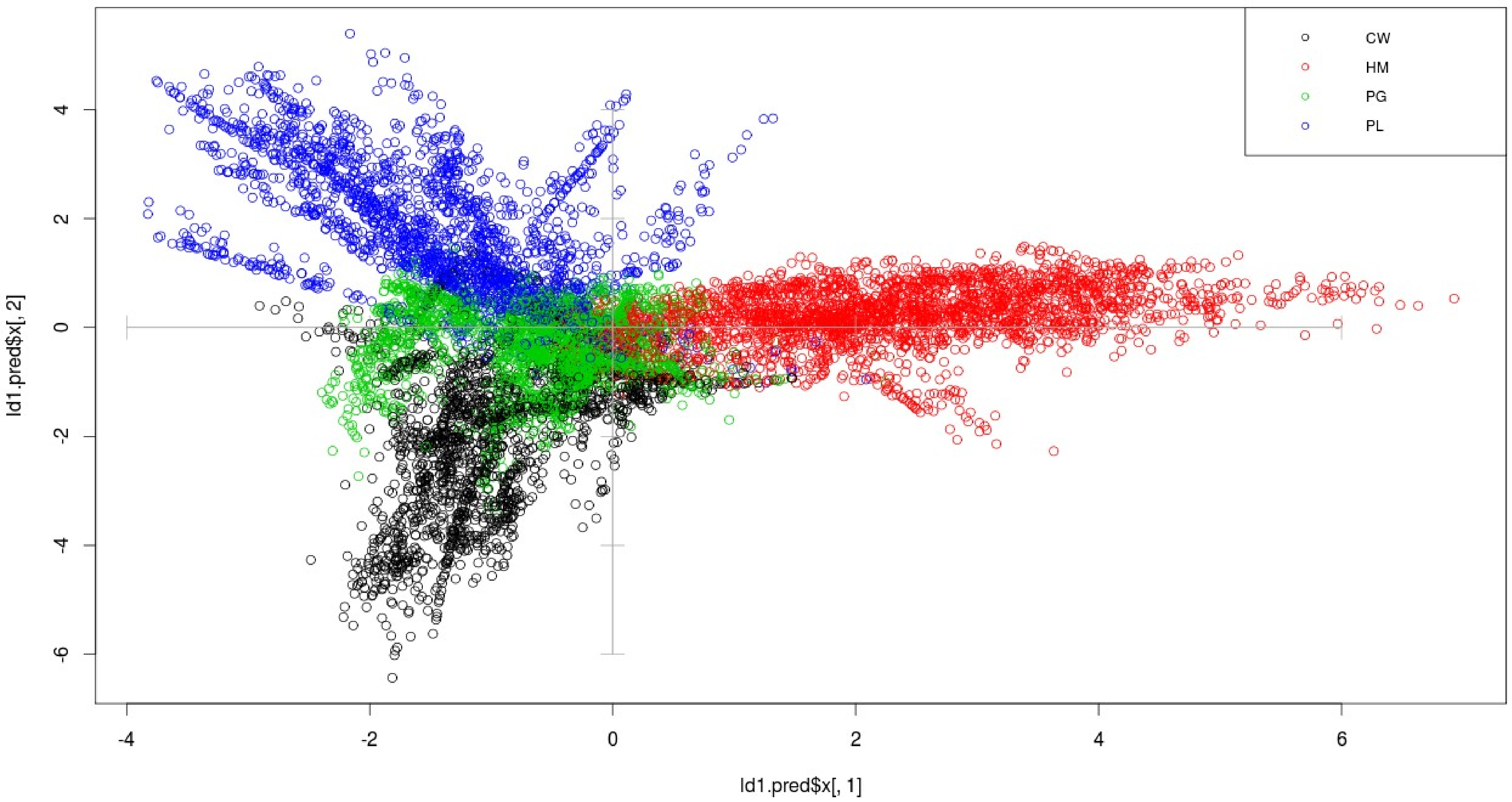
Extended data matrix

We create a realistic scenario of dilution/aging:

- 10.000 observations are created by sampling the data matrix randomly
- Dilution degree is *lognormal*
- Time in water is *exponential*

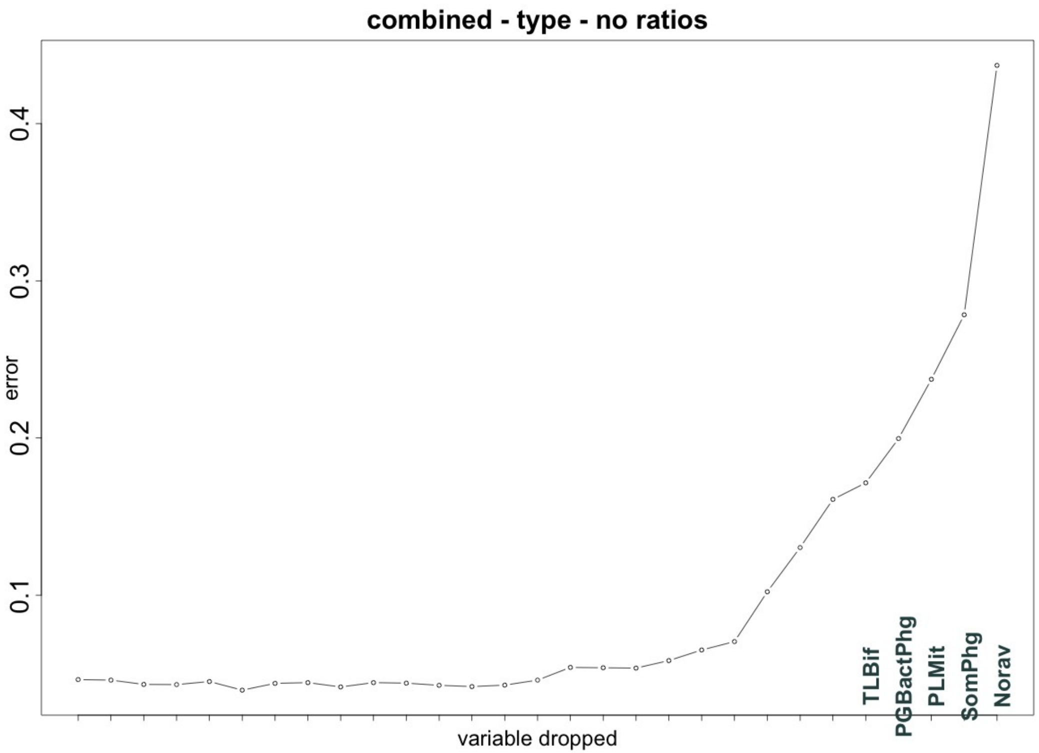


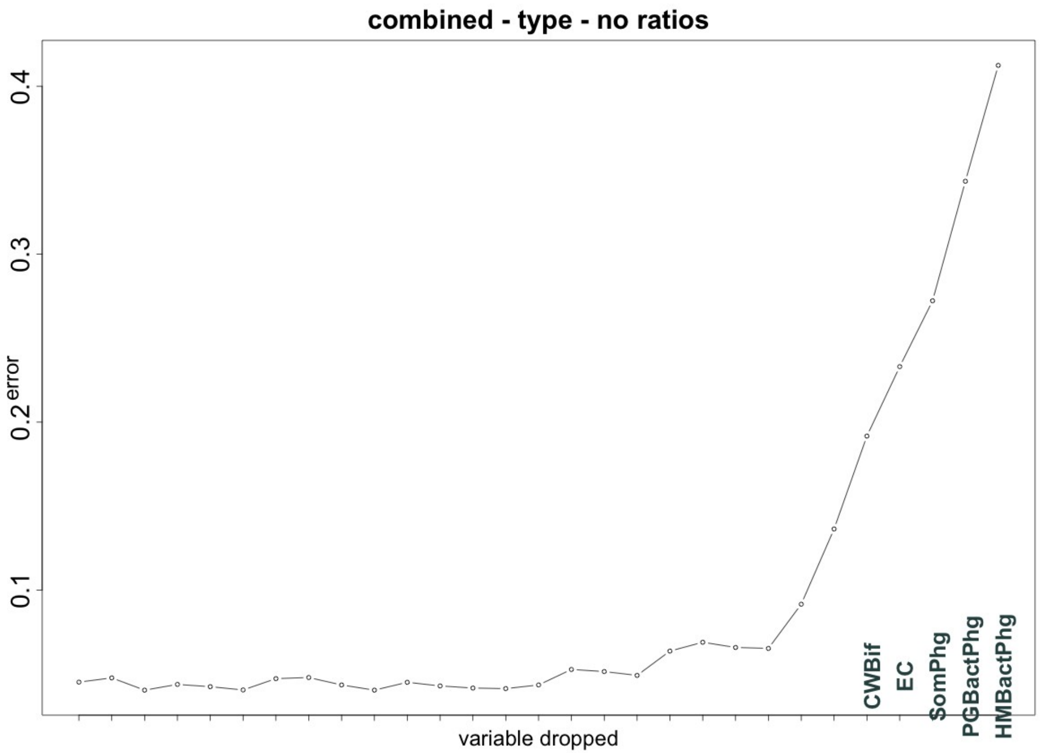


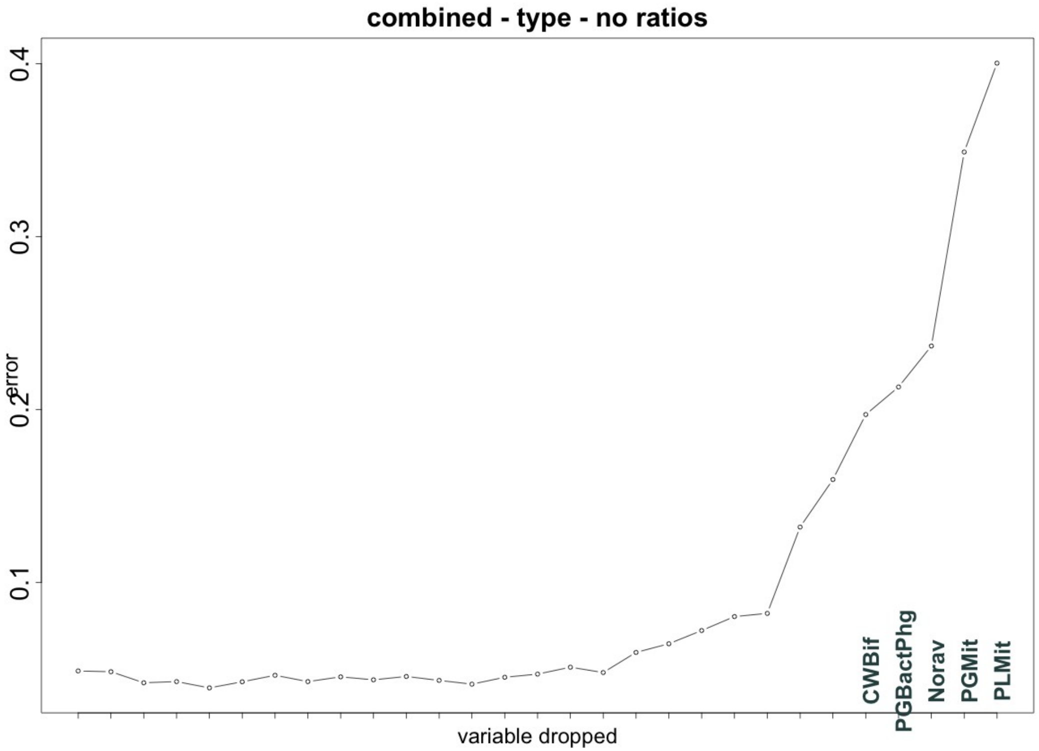


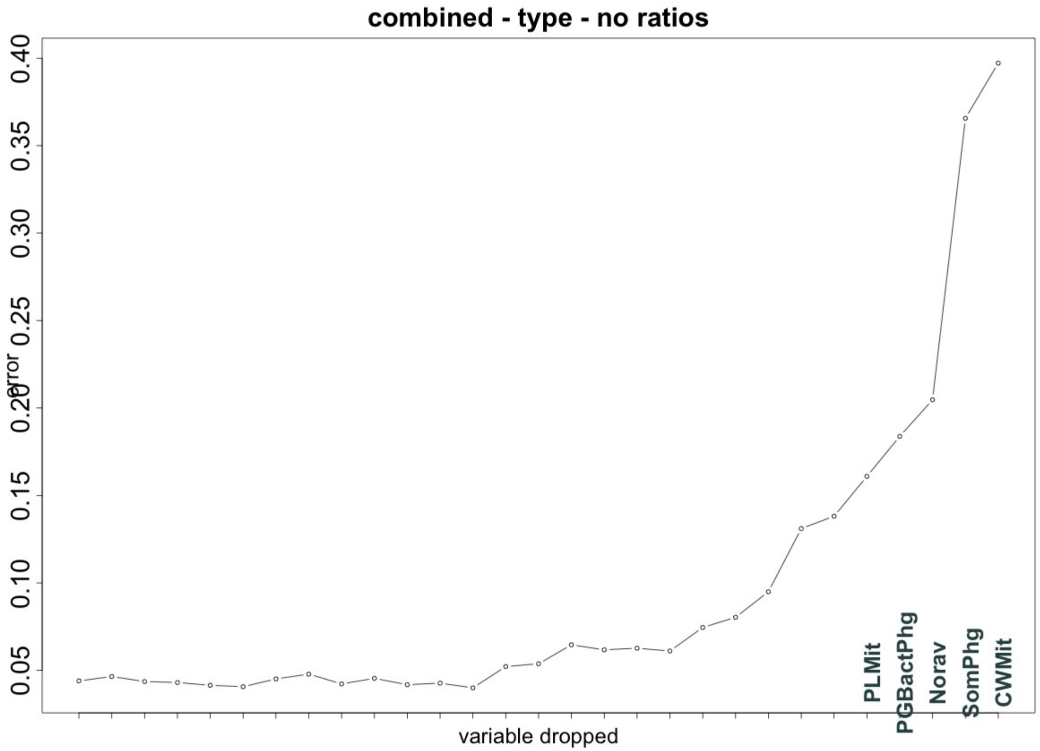
```
        true
pred      CW    HM    PG    PL
  CW    82.6   1.1   6.1   1.8
  HM     2.8  91.4   3.0   2.6
  PG     6.3   7.5  84.0   4.6
  PL     8.3   0.0   6.9  91.0
```

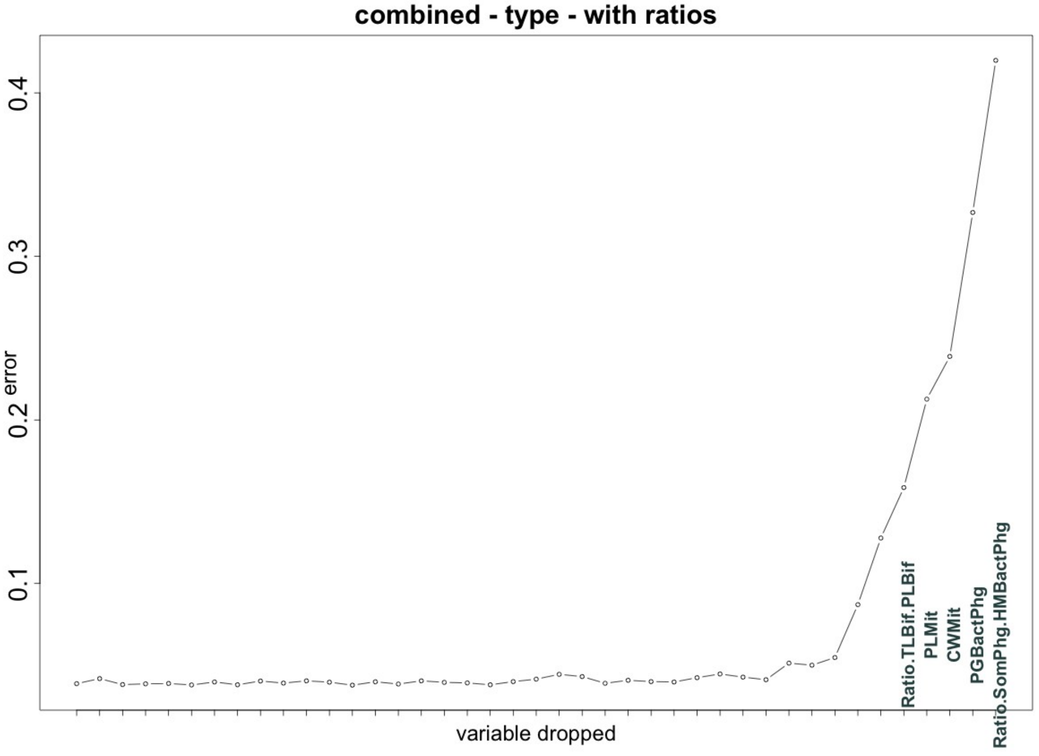
Full set: 87.7%; SomPhg HMBactPhg Norav CWMit PGMit PLMit: 80.4%
(the use of ratios does not make this any better)

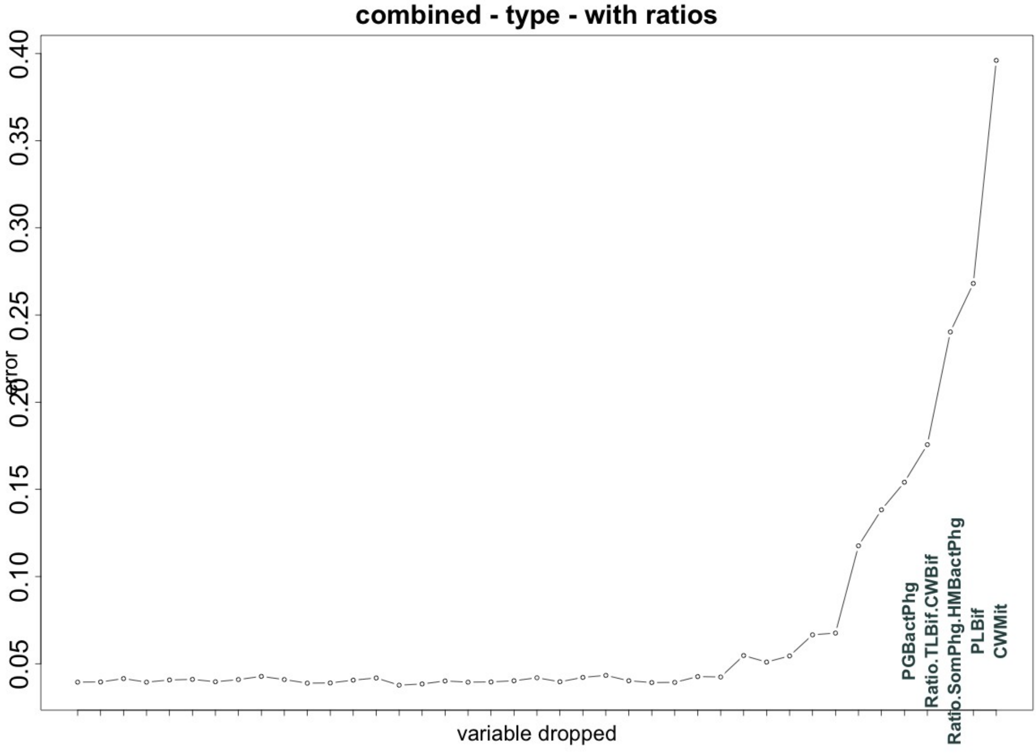

combined - type - no ratios
error
0.1
0.2
0.3
0.4
variable dropped
TLBif
PGBactPhg
PLMit
SomPhg
Norav

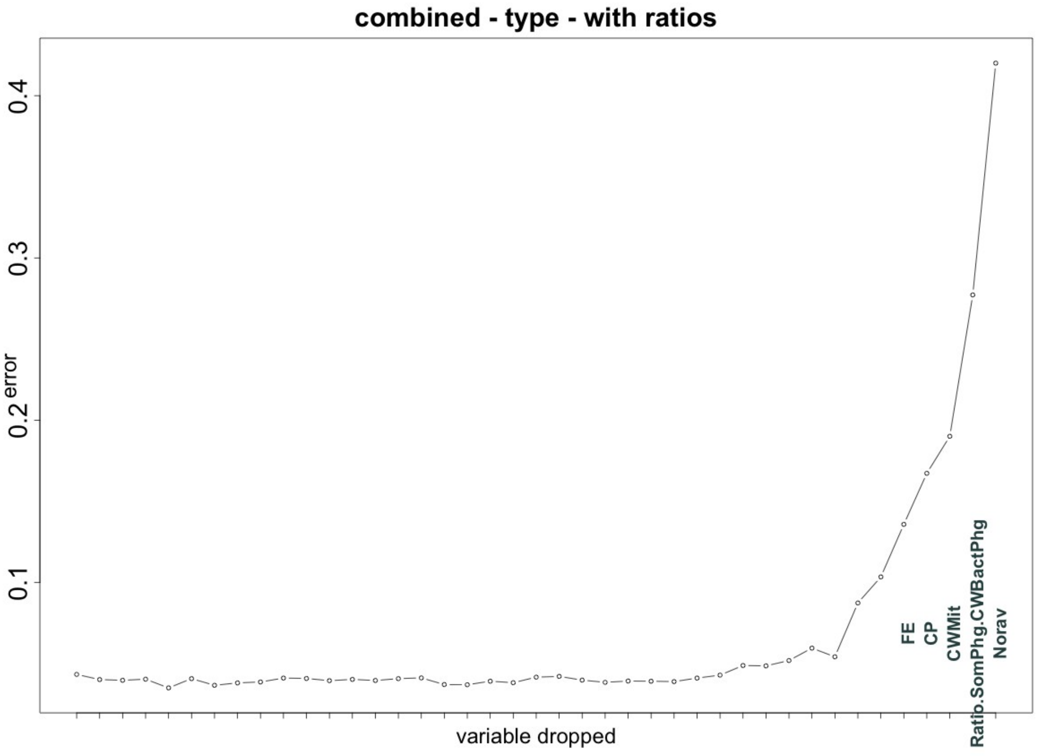

combined - type - no ratios
error
0.1
0.2
0.3
0.4
variable dropped
CWBif
EC
SomPhg
PGBactPhg
HMBactPhg

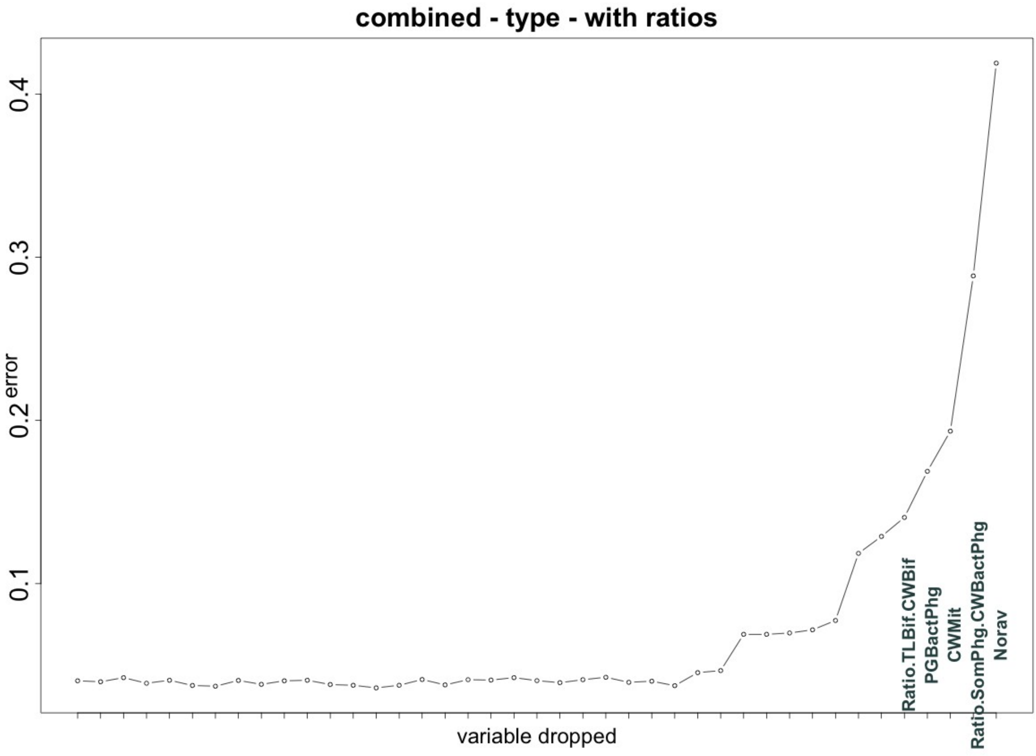

combined - type - no ratios
error
0.1
0.2
0.3
0.4
variable dropped
CWBif
PGBactPhg
Norav
PGMit
PLMit


combined - type - no ratios
error
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
variable dropped
PLMit
PGBactPhg
Norav
SomPhg
CWMit


combined - type - with ratios
error
0.1
0.2
0.3
0.4
variable dropped
Ratio.TLBif.PLBif
PLMit
CWMit
PGBactPhg
Ratio.SomPhg.HMBactPhg


combined - type - with ratios
error
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
variable dropped
PGBactPhg
Ratio.TLBif.CWBif
Ratio.SomPhg.HMBactPhg
PLBif
CWMit


combined - type - with ratios
error
0.1
0.2
0.3
0.4
variable dropped
FE
CP
CWMit
Ratio.SomPhg.CWBactPhg
Norav


combined - type - with ratios
error
0.1
0.2
0.3
0.4
variable dropped
Ratio.TLBif.CWBif
PGBactPhg
CWMit
Ratio.SomPhg.CWBactPhg
Norav

Preliminary assessment

SomPhg / HMBactPhg   SomPhg / PGBactPhg

CWMit   PGMit   PLMit

HMBactPhg   Norav

Estimated prediction error around 5%

WP5 Scenario

Next goal is to provide the final subset of indicators

Final goal is to provide WP5 with the software:

- Estimation of **dilution degree** (*concentration level)*
- Estimation of **age** (*time since contamination)*
- **Origin** of the sample (*prediction of source*) + probabilities for all 4 sources (*confidence in the prediction)*

Thank You
Mahalo
Kiitos
Tack
Toda
Grazie
Obrigado
Thanks
Takk
Merci
Gracias