9 juni 2022

Demo Data Science PoC

Nienke Meekel
Frederic Béen
Ton van Leerdam

KWR

Bridging Science to Practice

Doel project

- Ontwikkeling van aanpak om massaspectrometrie data te analyseren

- Bijzondere aandacht voor
  - Langere tijdreeksen
  - Meerdere locaties
  - Identificatie relevante kenmerken/features (prioriteren)
  - Selecteren van relevante features op basis van hun voorkomen en/of temporele trends
  - Multi-platform

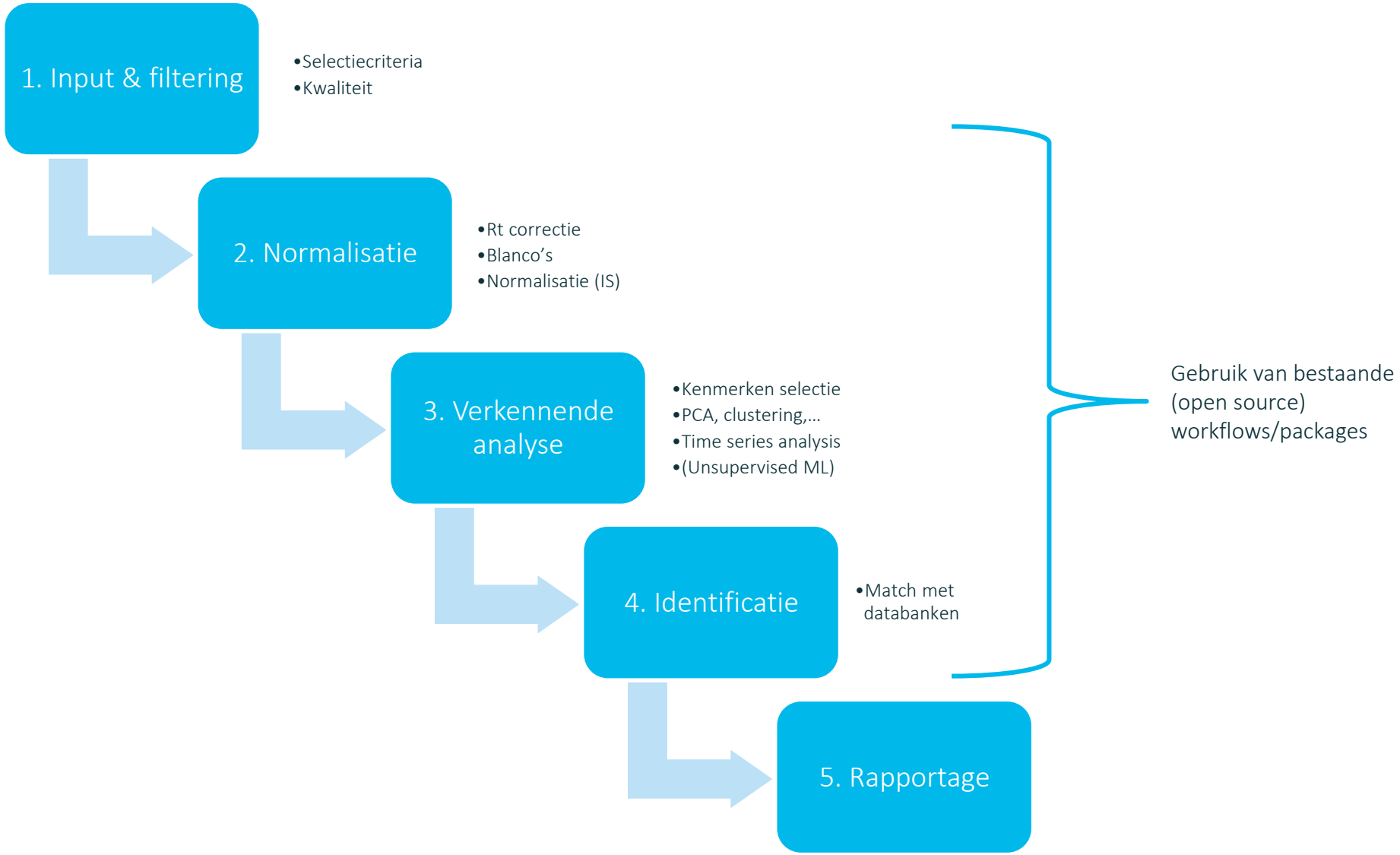
R(Studio)

Programmeertaal voor statistische berekeningen en figuren

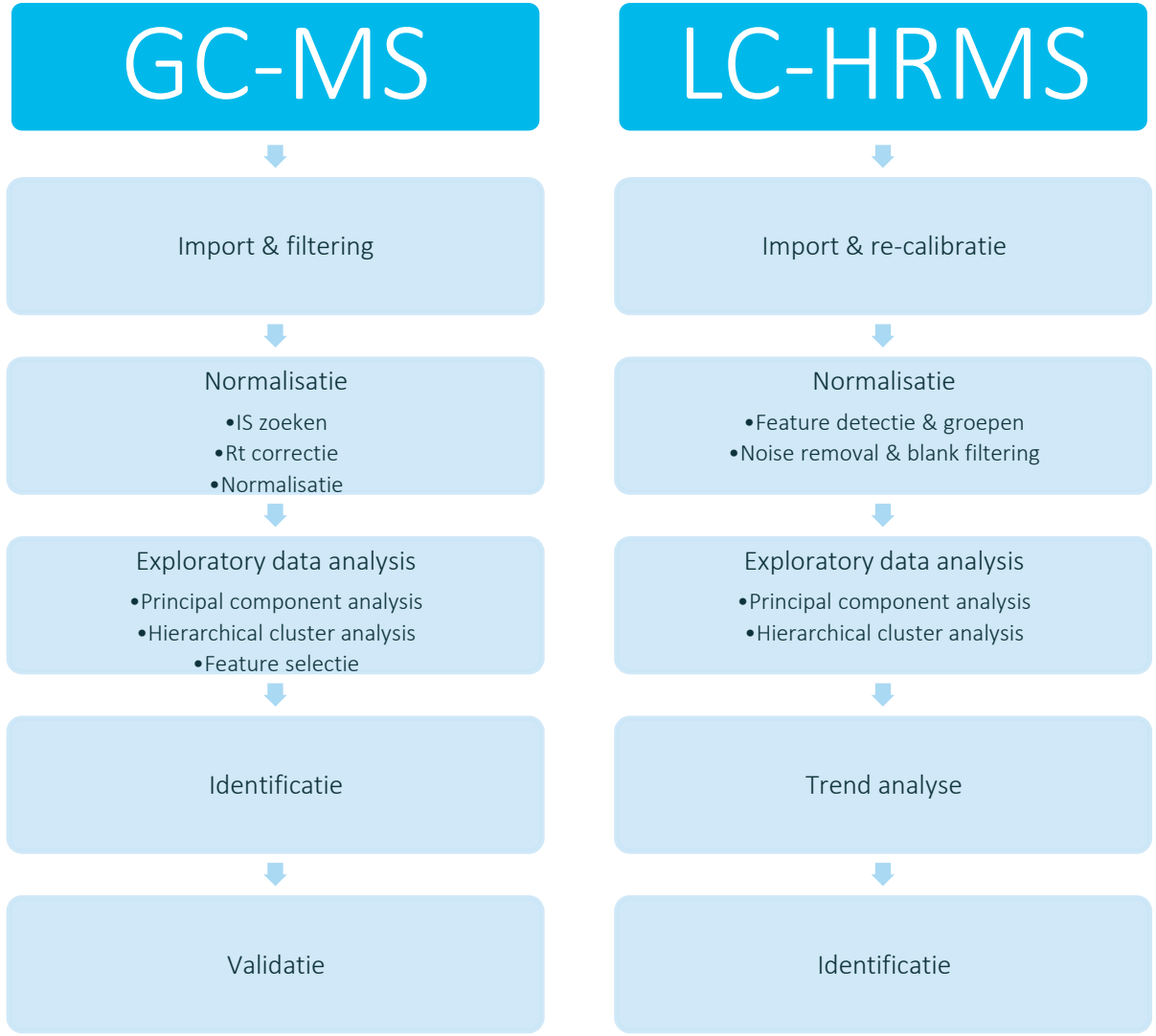
→ Wijdverbreid gebruikt voor de analyse van massaspectrometrie data (MS data)

Ontwikkelomgeving voor R

Strategie

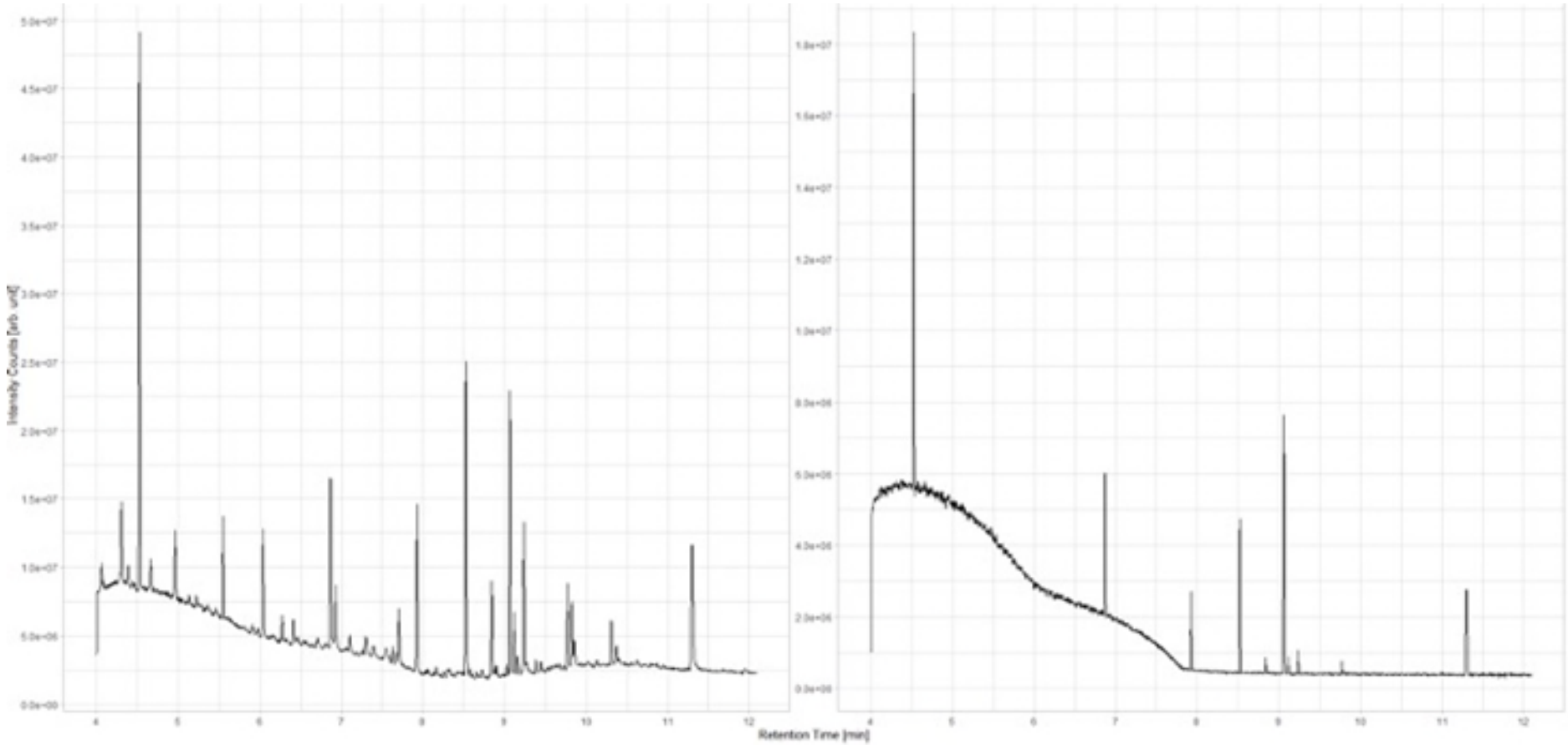

1. Input & filtering
•Selectiecriteria
•Kwaliteit
2. Normalisatie
•Rt correctie
•Blanco's
•Normalisatie (IS)
3. Verkennende analyse
•Kenmerken selectie
•PCA, clustering,…
•Time series analysis
•(Unsupervised ML)
4. Identificatie
•Match met databanken
5. Rapportage
Gebruik van bestaande (open source) workflows/packages

Overzicht van de ontwikkelde PoC

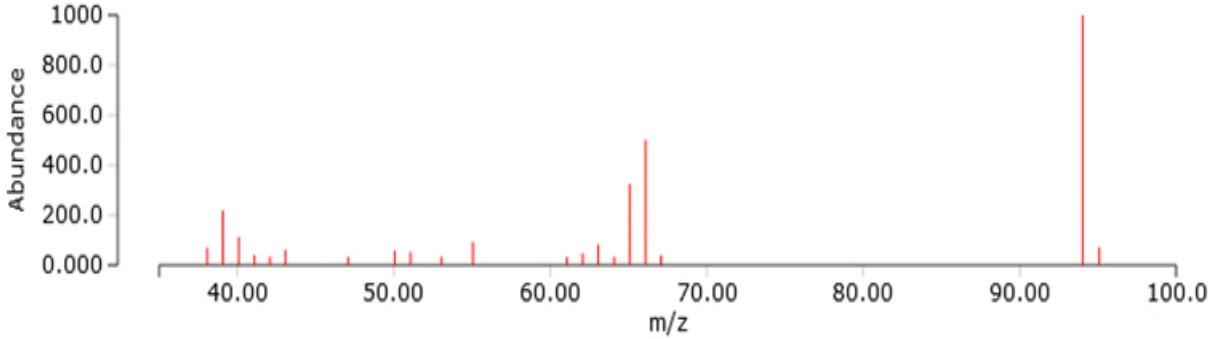


SPE-GC/MS data

SPE-GC/MS data



Massaspectrum

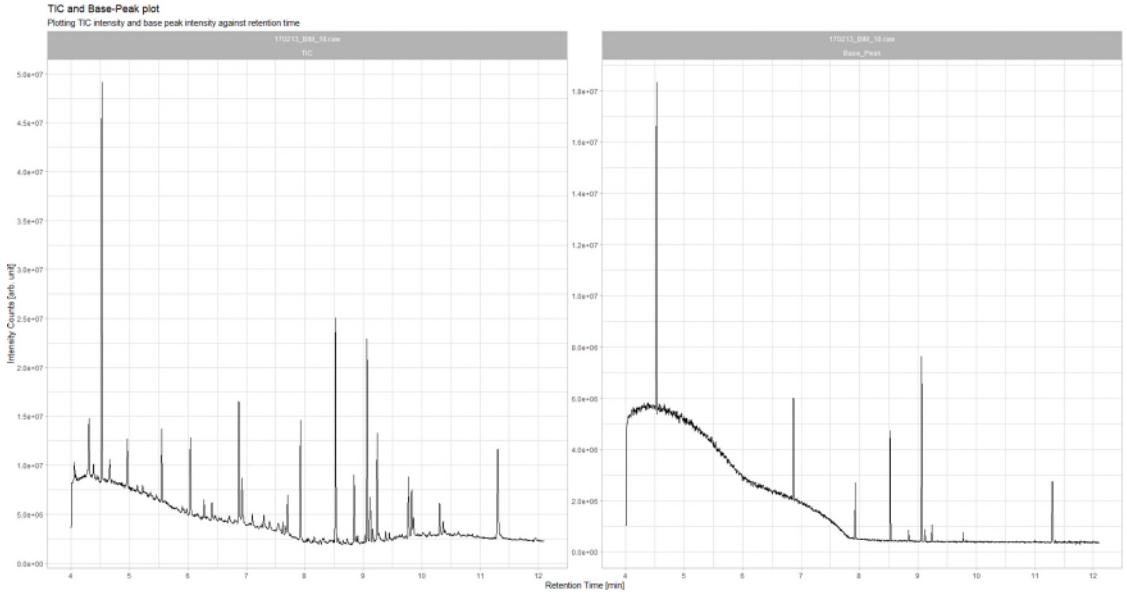


Achtergrond GC-MS data

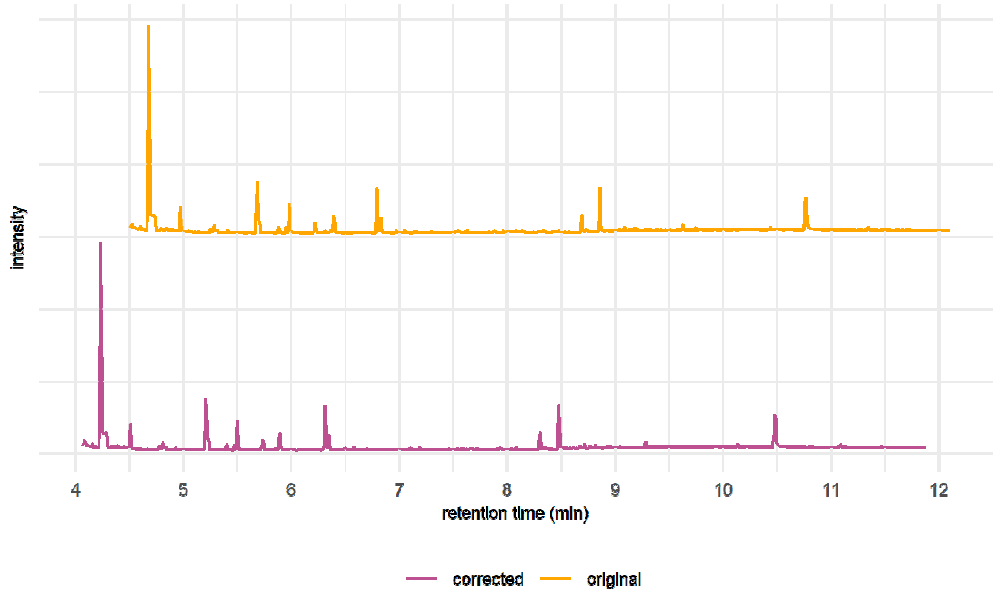
- Elk jaar heeft tot 730 chromatogrammen (periode 2016 – 2021)
- Eén monster per 12u
- Elk chromatogram bevat > 700 variabelen (d.w.z. features)
- Onmogelijk (of zeer onpraktisch) om handmatig naar specifieke pieken (relevante kenmerken) te zoeken

→ Vergemakkelijk (en automatiseer) de detectie van relevante kenmerken (prioriteren)
→ *Exploratory data analysis*

1. Import & filtering



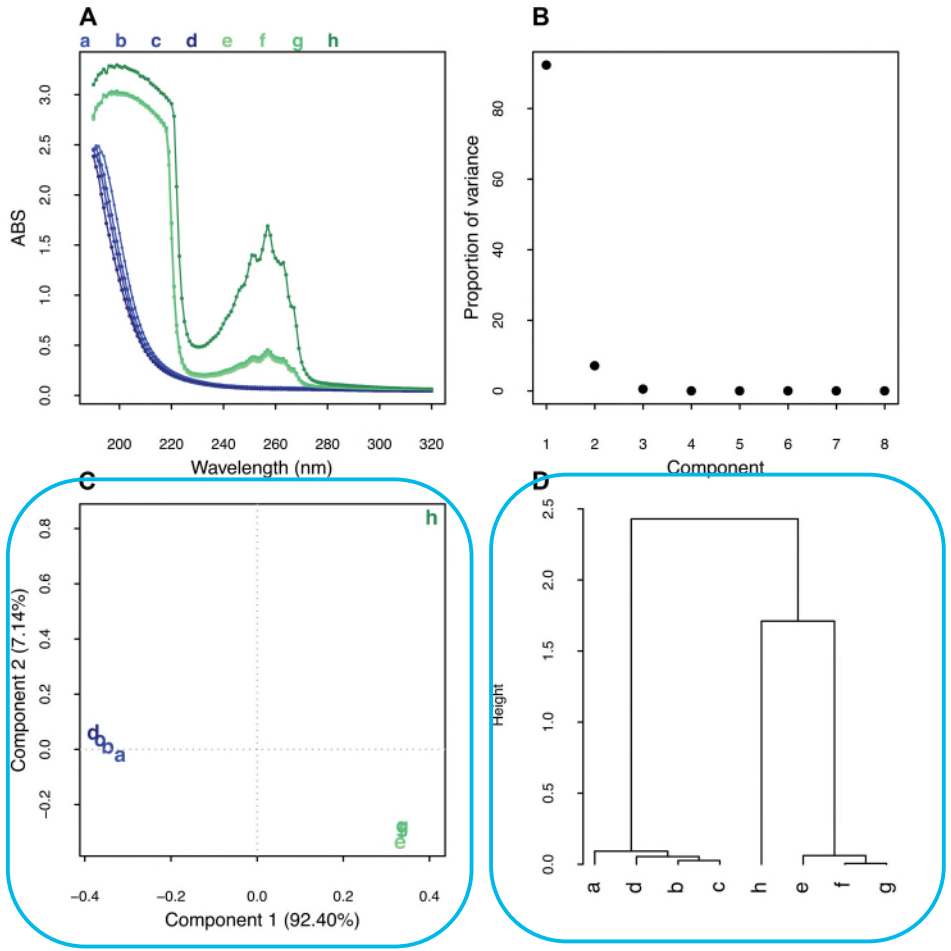
2. Normalisatie



Verkennende analyse

Twee aanpakken:

1. Principal component analysis (PCA)
2. Hierarchical cluster analysis (HCA)



Lima et al., 2011 *PLoS One*, 6 (1)

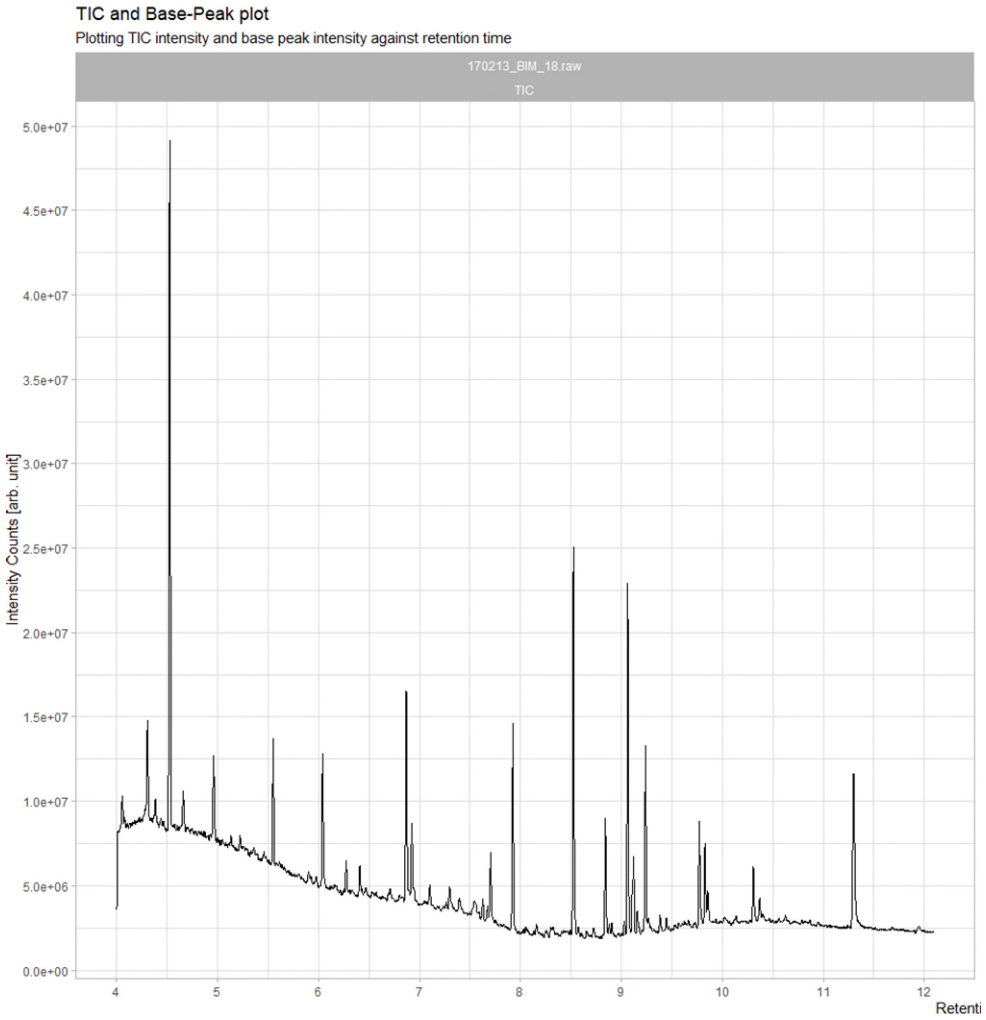
Verkennende analyse

Exploratory data analysis

Doel: patronen identificeren uit een grote dataset

Achtergrond

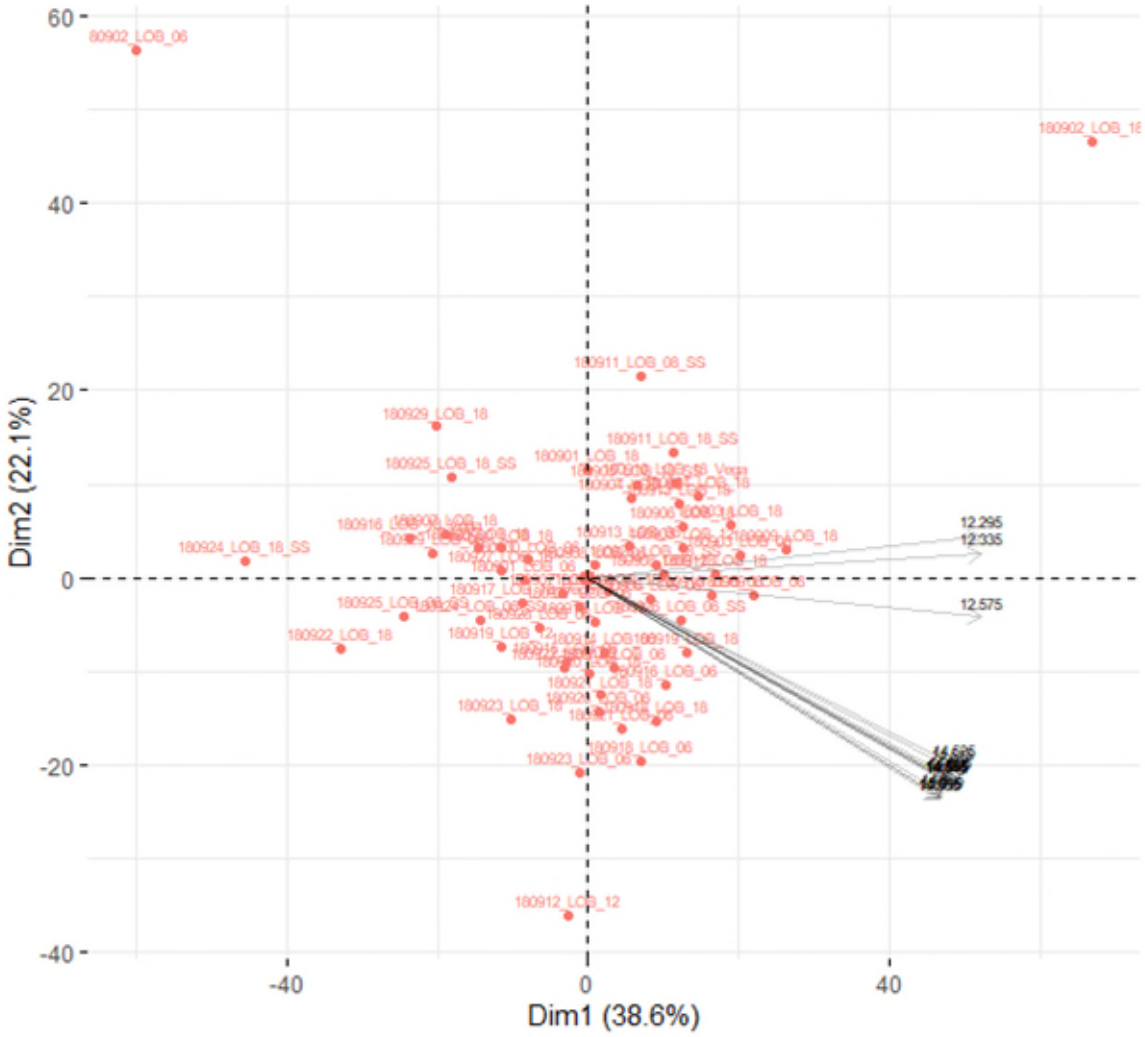
- Er is weinig of geen informatie beschikbaar over patronen
- Gegevens zijn multivariaat
  - d.w.z. dat elke observatie (monster) wordt beschreven door talrijke variabelen
  - In ons geval:
    - y = Intensiteit
    - x = massa (30-330 m/z)
    - z = Retentie tijd (min)



Principal component analysis (PCA)

- Nuttig voor snelle beoordeling
- Detecteert snel uitschieters (d.w.z. calamiteiten)

Lobith, Sept. 2018



Validatie

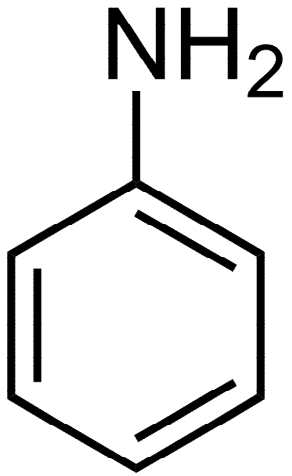
- Gebruik beschikbare informatie over calamiteiten
- Neem willekeurig chromatogrammen op
  1. Test of PCA en/of HCA monsters van calamiteiten kan markeren
  2. Test of (tentatieve) identificatie succesvol is

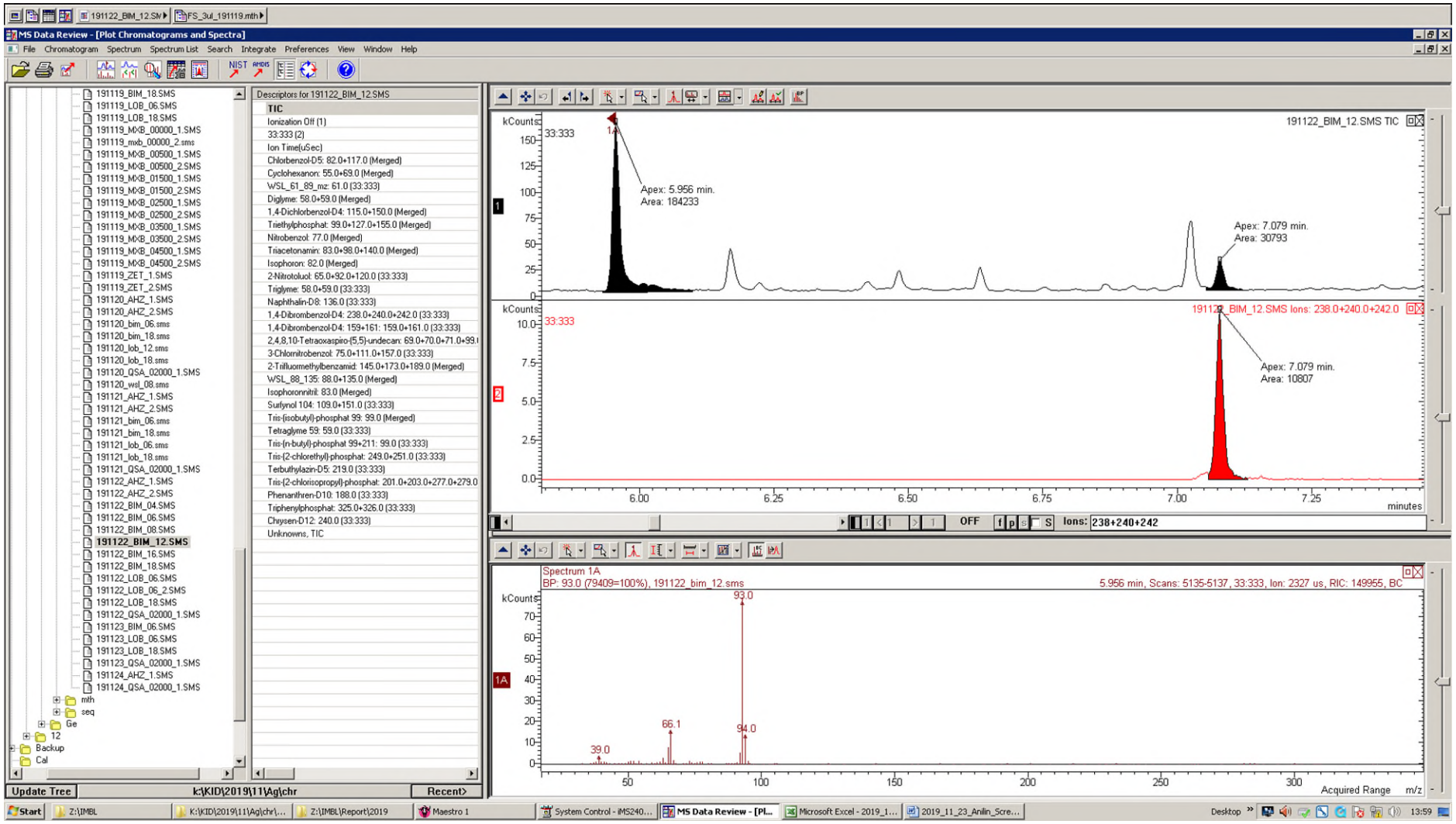
Validatie (I)

Zoeken van een bekende calamiteit

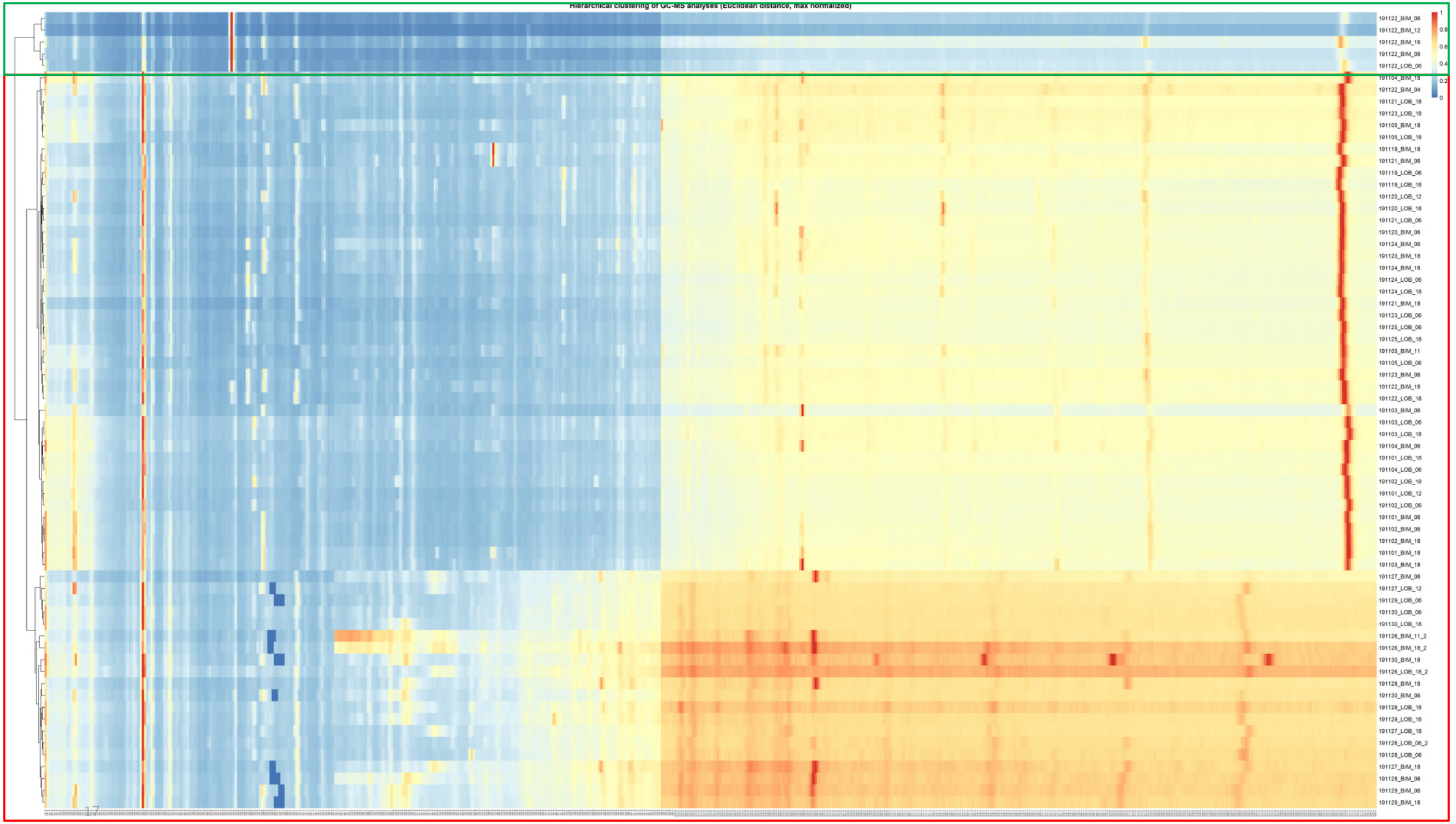
- Aniline (6 µg/L)
- Gedetecteerd in Nov 2019 in Bimmen
- Exact mass: 93.057849228 Da
- RT: 5.96 min

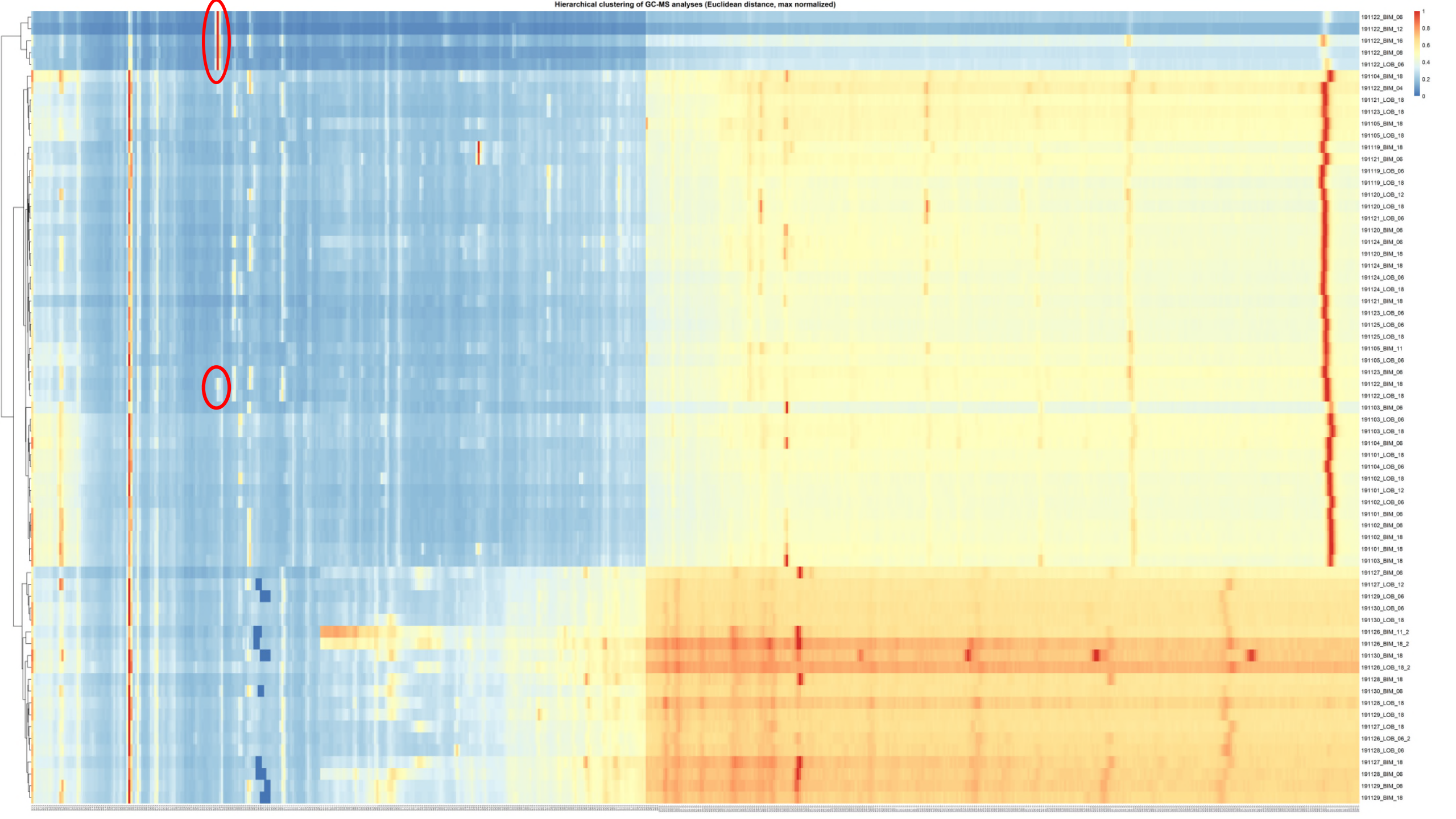
- De hele maand november 2019 met metingen van Bimmen en Lobith geselecteerd: 67 chromatogrammen





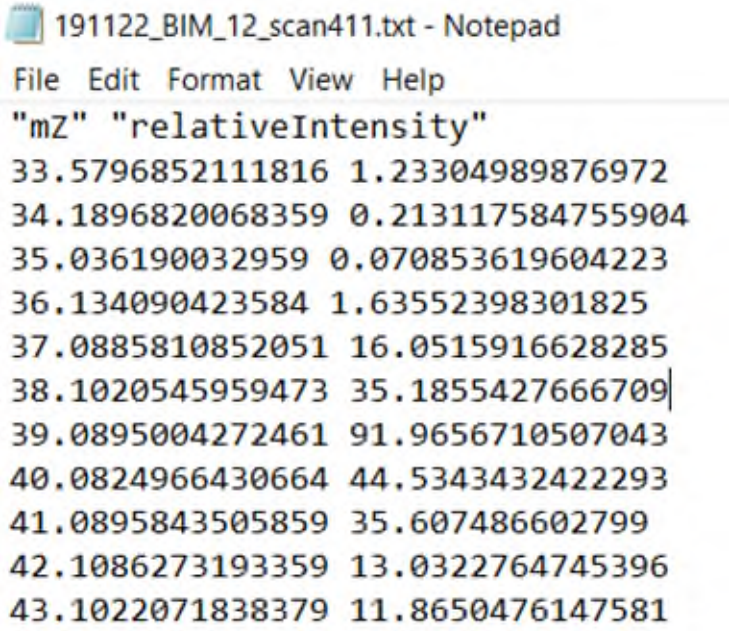
Bron: 2019_11_23_Anilin_Screenshot.doc


Hierarchical clustering of GC-MS analyses (Euclidean distance, max normalized)
Groepen op basis van kenmerken
Retentie tijd index (min)


Hierarchical clustering of GC-MS analyses (Euclidean distance, max normalized)
191122_BIM_06
191122_BIM_12
191122_BIM_16
191122_BIM_08
191122_LOB_06
191104_BIM_18
191122_BIM_04
191121_LOB_18
191123_LOB_18
191105_BIM_18
191105_LOB_18
191119_BIM_18
191121_BIM_06
191119_LOB_06
191119_LOB_18
191120_LOB_12
191120_LOB_18
191121_LOB_06
191120_BIM_06
191124_BIM_06
191120_BIM_18
191124_BIM_18
191124_LOB_06
191124_LOB_18
191121_BIM_18
191123_LOB_06
191125_LOB_06
191125_LOB_18
191105_BIM_11
191105_LOB_06
191123_BIM_06
191122_BIM_18
191122_LOB_18
191103_BIM_06
191103_LOB_06
191103_LOB_18
191104_BIM_06
191101_LOB_18
191104_LOB_06
191102_LOB_18
191101_LOB_12
191102_LOB_06
191101_BIM_06
191102_BIM_06
191102_BIM_18
191101_BIM_18
191103_BIM_18
191127_BIM_06
191127_LOB_12
191129_LOB_06
191130_LOB_06
191130_LOB_18
191126_BIM_11_2
191126_BIM_18_2
191130_BIM_18
191126_LOB_18_2
191128_BIM_18
191130_BIM_06
191128_LOB_18
191129_LOB_18
191127_LOB_18
191126_LOB_06_2
191128_LOB_06
191127_BIM_18
191128_BIM_06
191129_BIM_06
191129_BIM_18
1
0.8
0.6
0.4
0.2
0

Export MS

1. Operator kiest RT vanuit HCA plot
2. Script zoekt originele RT en scannummer
3. Schrijft naar .txt file in bestandsformaat voor MassBank EU / MoNA

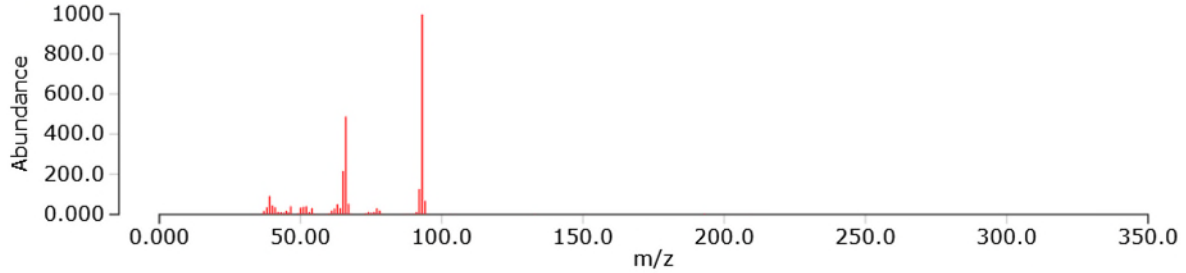


Zoekresultaten

Quick Search Results

MassBank Search Contents Download Accession Go Documentation About MassBank News Archive

Query :

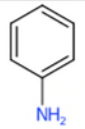
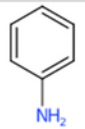


Edit / Resubmit Query

Results : 20 Hit.

Multiple Display | Spectrum Search

▼ Results End

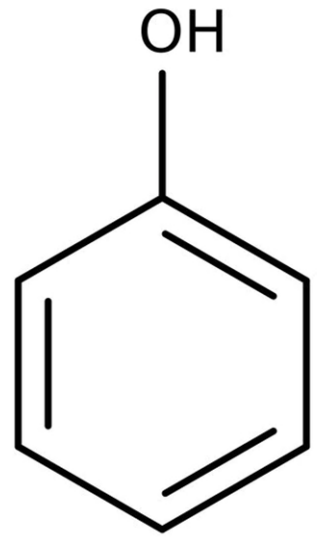
	Name	Formula / Structure	Hit	Score
☐	ANILINE; EI-B; MS	C6H7N 	23	0.9817
☐	ANILINE; EI-B; MS	C6H7N 	24	0.9720

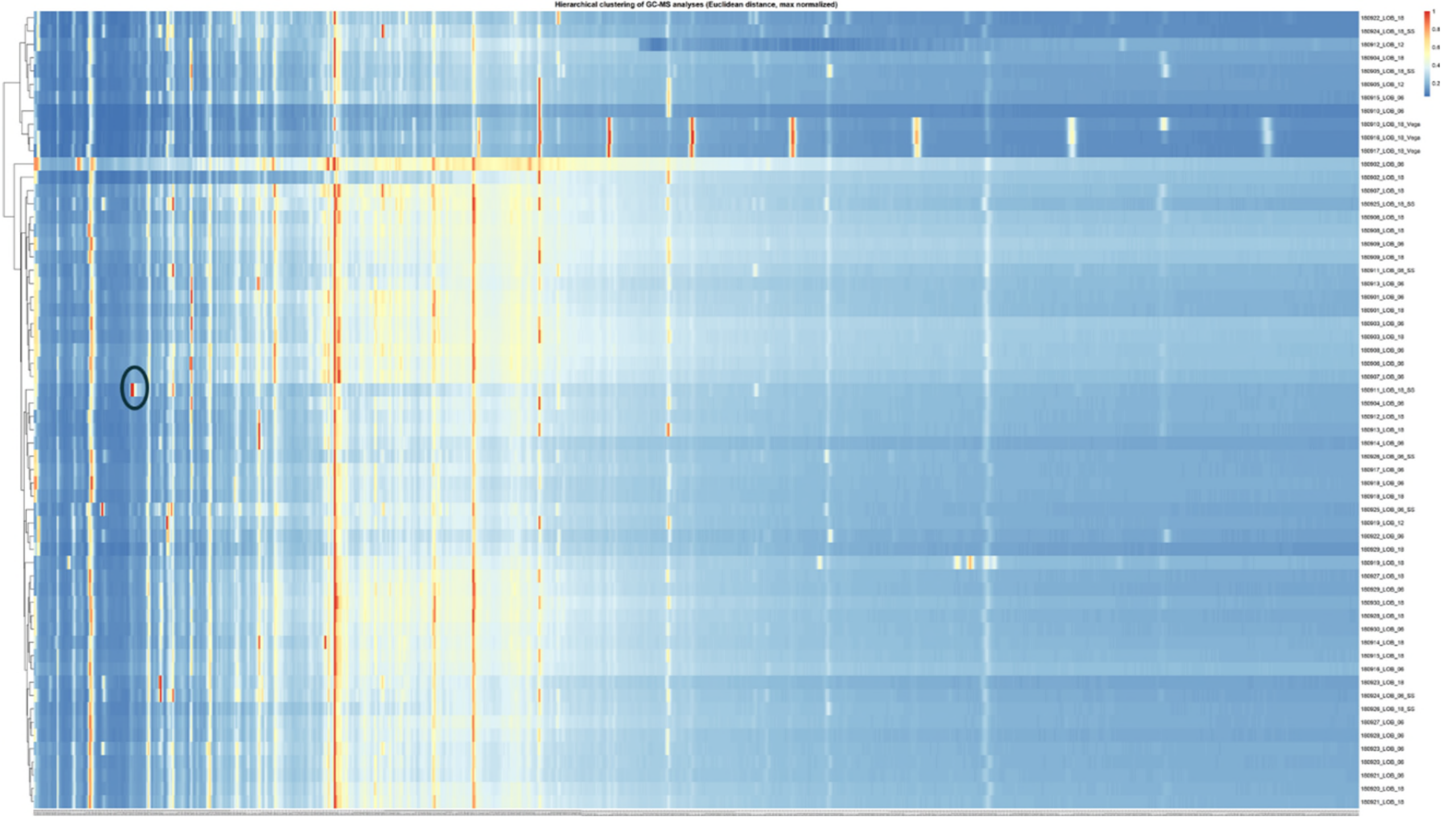
Validatie (II)

Zoeken van een bekende calamiteit

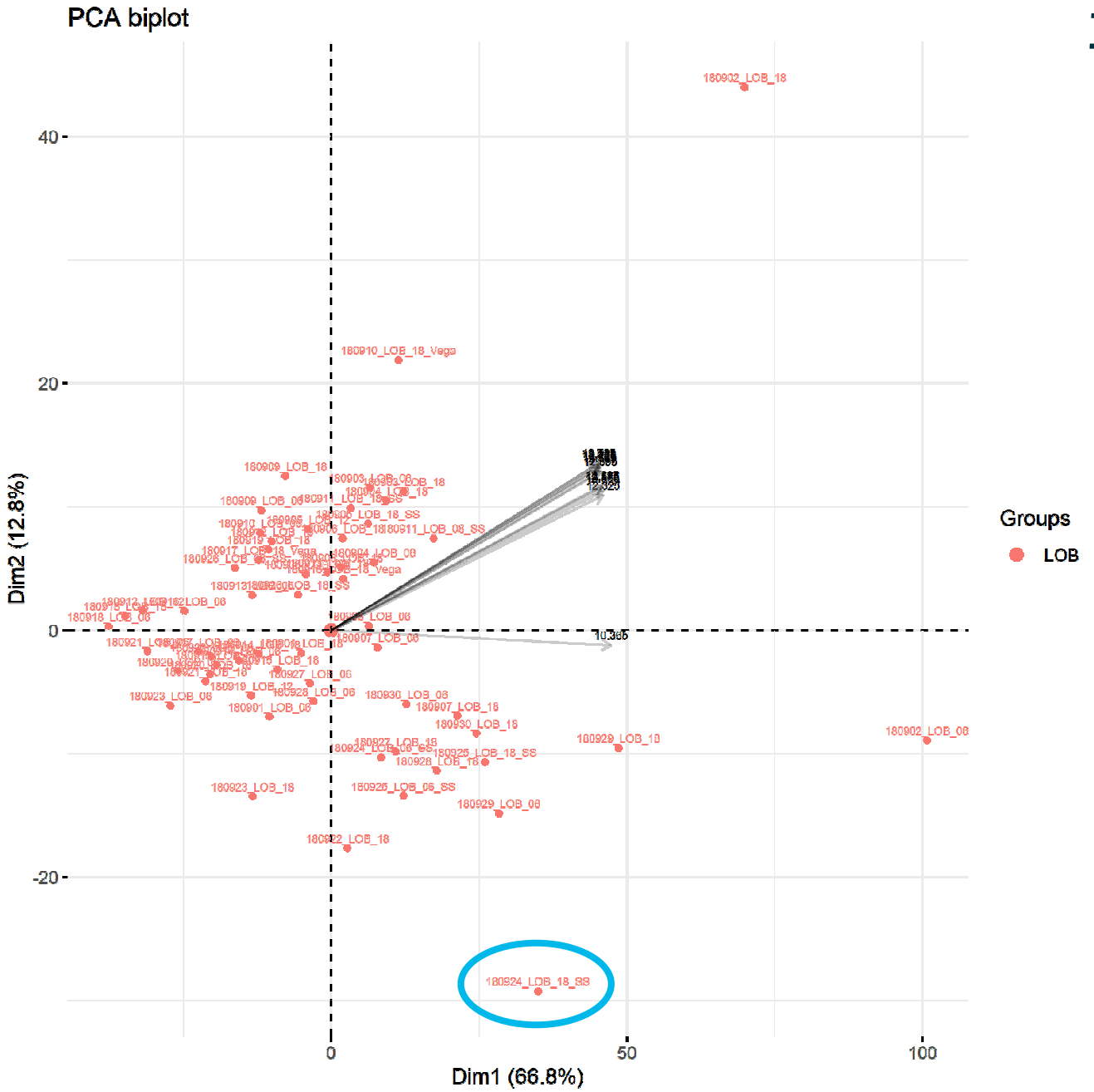
- Phenol (1.4 µg/L)
- Gedetecteerd in maar 1 dag in september 2018
- Exact mass: 94.0418648 Da
- RT: 5.50 min

- De hele maand september 2018 met metingen van Bimmen en Lobith geselecteerd: 60 chromatogrammen




Hierarchical clustering of GC-MS analyses (Euclidean distance, max normalized)
150922_LOB_18
150924_LOB_18_SS
150912_LOB_12
150904_LOB_18
150905_LOB_18_SS
150905_LOB_12
150915_LOB_06
150910_LOB_06
150910_LOB_18_Vega
150916_LOB_18_Vega
150917_LOB_18_Vega
150902_LOB_06
150902_LOB_18
150907_LOB_18
150925_LOB_18_SS
150906_LOB_18
150908_LOB_18
150909_LOB_06
150909_LOB_18
150911_LOB_06_SS
150913_LOB_06
150901_LOB_06
150901_LOB_18
150903_LOB_06
150903_LOB_18
150908_LOB_06
150906_LOB_06
150907_LOB_06
150911_LOB_18_SS
150904_LOB_06
150912_LOB_18
150913_LOB_18
150914_LOB_06
150926_LOB_06_SS
150917_LOB_06
150918_LOB_06
150918_LOB_18
150925_LOB_06_SS
150919_LOB_12
150922_LOB_06
150929_LOB_18
150919_LOB_18
150927_LOB_18
150929_LOB_06
150930_LOB_18
150928_LOB_18
150930_LOB_06
150914_LOB_18
150915_LOB_18
150916_LOB_06
150923_LOB_18
150924_LOB_06_SS
150926_LOB_18_SS
150927_LOB_06
150929_LOB_06
150923_LOB_06
150920_LOB_06
150921_LOB_06
150920_LOB_18
150921_LOB_18
1
0.8
0.6
0.4
0.2

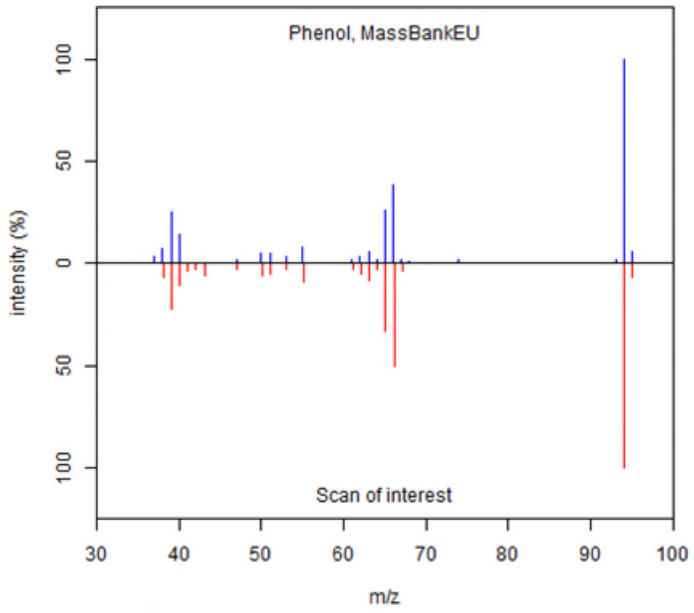
Validatie (II)



180911_LOB_18_SS_scan448_top20.txt - N...

File Edit Format View Help

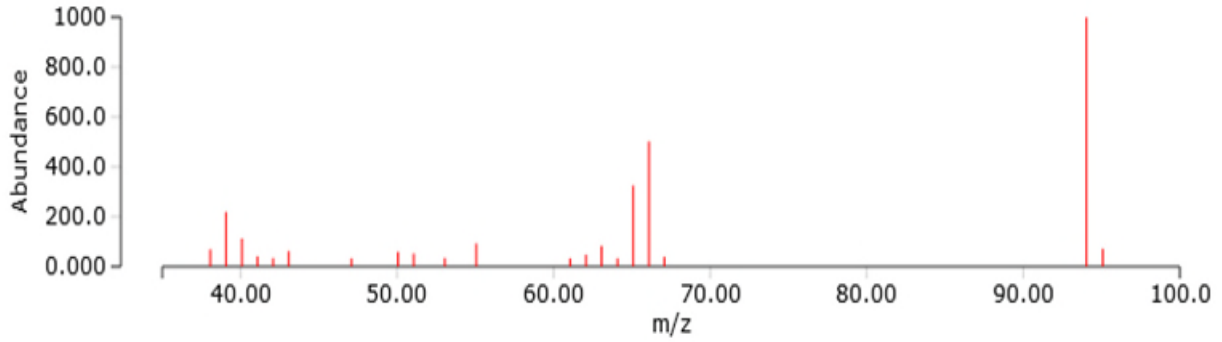
```
"mZ" "relativeIntensity"
94.0363540649414 999
66.0865707397461 501.471662824693
65.0850448608398 324.9251206881
39.0748443603516 218.905783047957
40.0861167907715 112.411232799178
55.063346862793 92.5873029682407
63.058780670166 82.6580493989069
95.0739|4609375 70.9722685968784
38.0677337646484 69.1094138740611
43.0699272155762 61.6783299488952
50.0615539550781 58.6246847801935
51.0659332275391 52.0862762206337
62.0642166137695 47.4076306946086
41.0919952392578 40.7941533781524
67.0805740356445 38.6526140254112
53.0500755310059 34.3016875633431
42.0831565856934 33.9024724284416
64.0824203491211 33.3499978391019
47.0799369812012 33.0574822860859
61.0610122680664 32.8169748760706
```



Quick Search Results

MassBank Search Contents Download Accession Go Documentation About MassBank News Archive

Query :

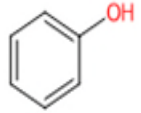


Edit / Resubmit Query

Results : 20 Hit.

Multiple Display | Spectrum Search

▼ Results End

	Name	Formula / Structure	Hit	Score
☐	PHENOL; EI-B; MS	C6H6O 	17	0.9776

Uitdagingen tijdens de ontwikkeling

- Afwezigheid van blanco monsters
- Kováts retentie index
- Herhalingen

LC-HRMS data

LC-HRMS dataset

Oppervlaktewater, 3 locaties

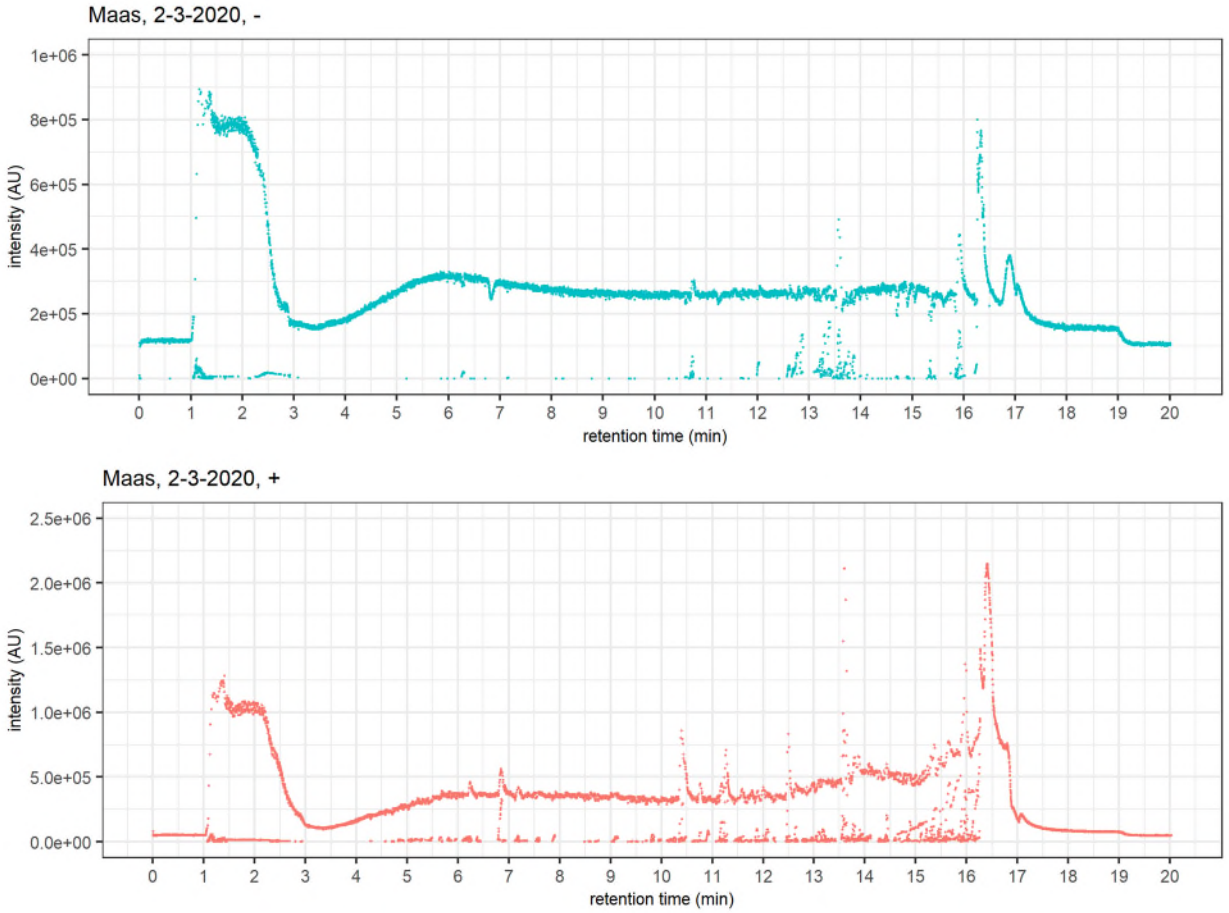
- (Lekkanaal-Rijn, Maas, IJsselmeer)

Over 2 jaar

Blanco’s

Triplo metingen

Inclusief interne standaarden



Import & re-calibratie

- Bruker gegevens moeten opnieuw gekalibreerd worden (externe kalibrant) vóór de uitvoer naar onbewerkt formaat
- Uitgevoerd door UvA
- Na kalibratie kunnen de gegevens getransformeerd worden naar universeel formaat: mzML

Normalisatie

- *patRoon*
- Ruwe gegevens (van leveranciers) importeren
- Pre-processen, deconvolueren, pieken opsporen en feature-lijsten genereren

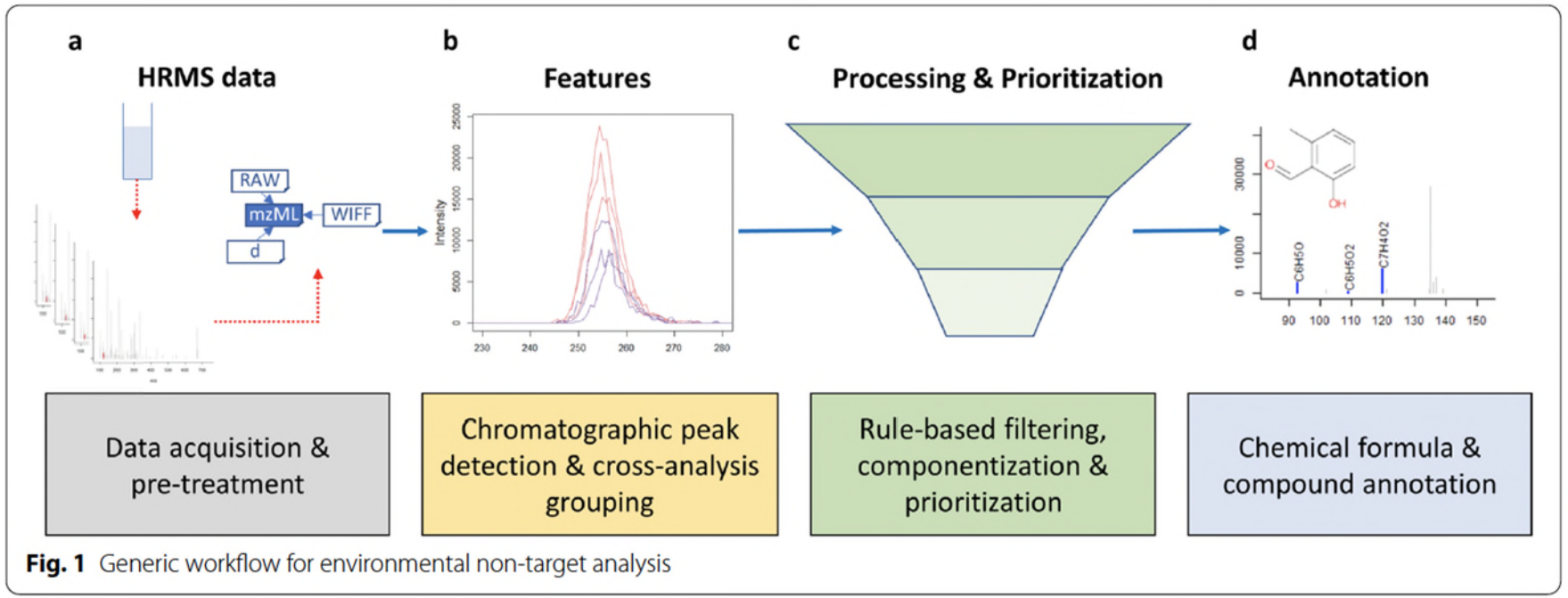
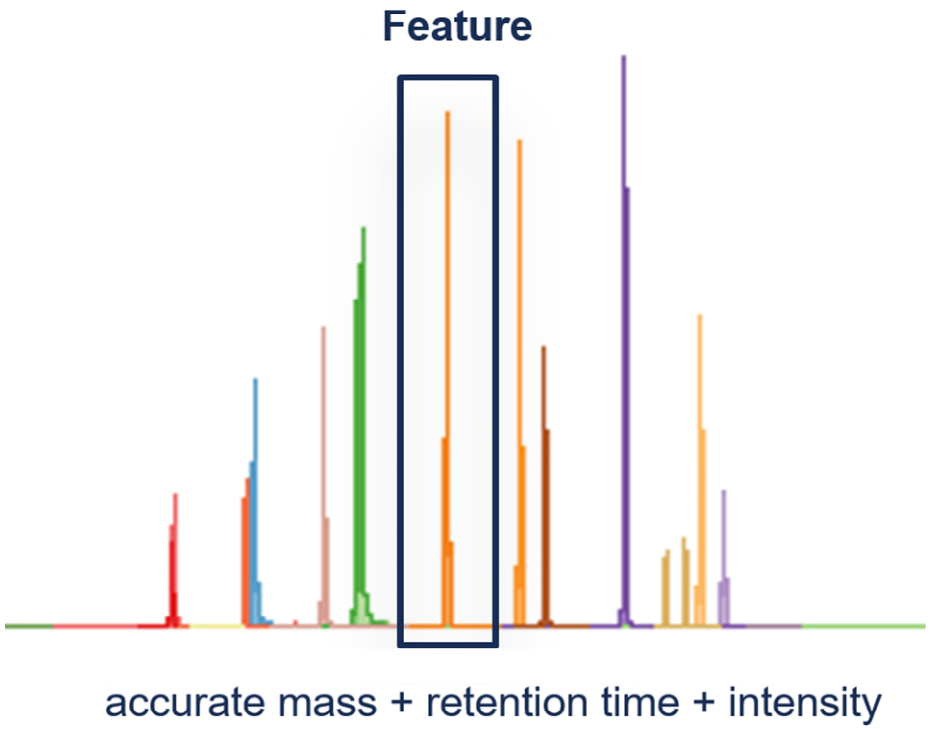


Fig. 1 Generic workflow for environmental non-target analysis

Helmus et al. 2021

Features & Feature Groups



Instrumentele fouten kunnen leiden tot variaties in retentietijd en m/z → groeperen van features na uitlijning

Normalisatie (ii)

- Ruisfiltering
- Verwijdering van signalen uit blanco's

→ Beide cruciale stappen met HRMS gegevens
→ Anders zijn er te veel kenmerken, waarvan de meeste niet van belang zijn

Normalisatie (iii)

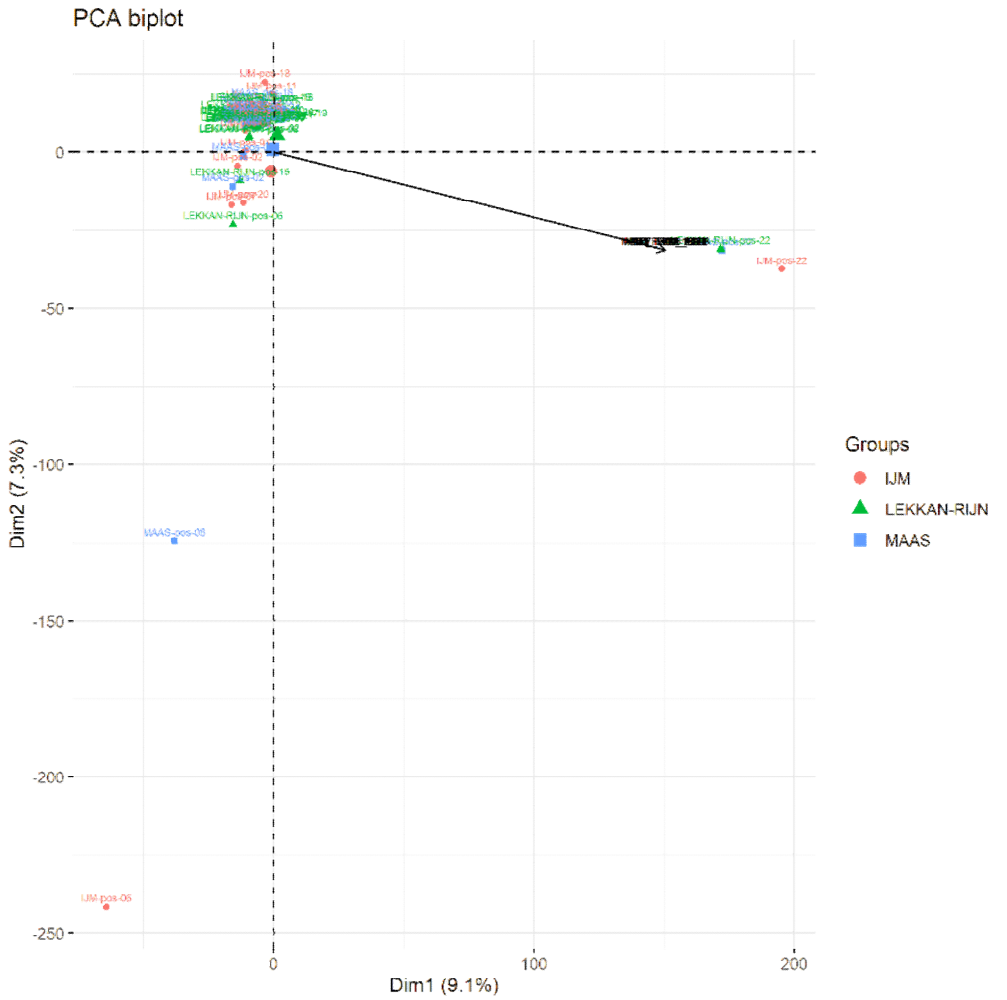
	analysis	ID	ret	mz	area	intensity	retmin	retmax	mzmin	mzmax	isocount
1	IJM-pos-01	f_15686429643064123212	952.747	38.96301	21408.630	3314	946.458	955.898	38.96294	38.96319	1
2	IJM-pos-01	f_5081088241392461239	215.984	45.03313	46542.120	2978	205.478	229.963	45.03289	45.03345	1
3	IJM-pos-01	f_2541063847962680870	959.046	45.04448	25555.330	4010	889.968	983.455	45.04423	45.04472	1
4	IJM-pos-01	f_9695041755674200979	306.745	46.06472	17876.050	3259	302.271	308.855	46.06457	46.06486	1
5	IJM-pos-01	f_16724282151736260192	750.683	57.06952	54050.870	12244	747.273	754.860	57.06944	57.06972	1
6	IJM-pos-01	f_16228712282045501977	78.742	68.01027	46193.950	3978	68.608	89.061	68.01004	68.01057	1
7	IJM-pos-01	f_6655480887537865596	387.553	69.06956	16634.320	5534	383.913	391.097	69.06929	69.06975	1
8	IJM-pos-01	f_9391797144393831603	819.185	73.02820	29930.100	6997	817.095	822.336	73.02817	73.02824	1
9	IJM-pos-01	f_16250092542303639385	343.620	73.06443	15763.180	3012	341.010	349.495	73.06411	73.06481	1
10	IJM-pos-01	f_8734255446469361998	413.263	73.06443	18239.270	2829	405.668	424.284	73.06406	73.06465	1
11	IJM-pos-01	f_11749505120381814919	149.239	80.94731	785986.500	45818	93.420	173.452	80.94707	80.94757	1
12	IJM-pos-01	f_18211038451167438953	81.084	80.94734	650015.700	49553	71.736	93.146	80.94720	80.94749	1
13	IJM-pos-01	f_14511912294683096663	914.495	81.06962	15319.030	5623	912.401	916.450	81.06949	81.06970	1
14	IJM-pos-01	f_11538822027010294325	921.828	81.06968	10403.890	4967	919.739	922.668	81.06948	81.06994	1
15	IJM-pos-01	f_16230200802772026742	149.380	82.94435	273130.900	21142	124.546	168.018	82.94404	82.94474	1
16	IJM-pos-01	f_6047272366943447675	81.084	82.94439	227197.500	17578	71.736	124.280	82.94403	82.94472	1
17	IJM-pos-01	f_12531529348156130242	174.386	83.05996	208554.500	60352	168.995	182.495	83.05961	83.06036	1
18	IJM-pos-01	f_10300111777406107830	173.452	84.06308	9486.903	4259	172.115	175.470	84.06267	84.06347	1
19	IJM-pos-01	f_14354815329452025538	82.674	84.95920	4174191.000	228695	62.316	434.915	84.95869	84.95952	2
20	IJM-pos-01	f_7847823545275183475	411.198	88.02114	118757.700	35542	397.545	421.164	88.02092	88.02168	1

Normalisatie (iv)

	group	ret	mz	IJM-pos-01	IJM-pos-01	IJM-pos-01	MAAS-pos-01	MAAS-pos-01	MAAS-pos-01
1	M39_R1014_1	1013.8415	38.96303	0	0	0	0	0	0
2	M39_R923_2	923.0452	38.96305	0	0	0	0	0	0
3	M39_R954_3	953.9436	38.96306	3314	3259	2684	3489	3293	3492
4	M39_R884_4	883.7007	38.96308	0	0	0	0	0	0
5	M43_R283_8	282.7633	43.01745	0	0	0	0	0	0
6	M43_R294_10	293.9756	43.01748	0	0	0	0	0	0
7	M43_R376_11	376.4058	43.01758	0	0	0	0	0	0
8	M43_R288_13	287.9704	43.01760	0	0	0	0	0	0
9	M43_R424_14	423.8749	43.01761	0	0	0	0	0	0
10	M43_R501_19	501.1905	43.01797	0	0	0	0	0	0
11	M43_R298_22	298.3022	43.01806	0	0	0	0	0	0
12	M45_R439_26	439.2912	45.03324	0	0	0	0	0	0
13	M45_R222_28	222.0241	45.03334	2978	2897	3991	4634	4349	5542
14	M45_R270_29	270.4260	45.03336	0	0	0	0	0	0
15	M45_R302_30	302.2260	45.03356	0	0	0	0	0	0
16	M45_R928_34	927.9138	45.04444	0	0	0	0	0	0
17	M45_R919_35	918.5295	45.04444	0	0	0	0	0	0
18	M45_R1010_36	1009.6403	45.04448	0	0	0	0	0	0
19	M45_R997_37	996.5864	45.04450	0	0	0	0	0	0
20	M45_R893_38	893.0938	45.04451	0	0	0	0	0	0

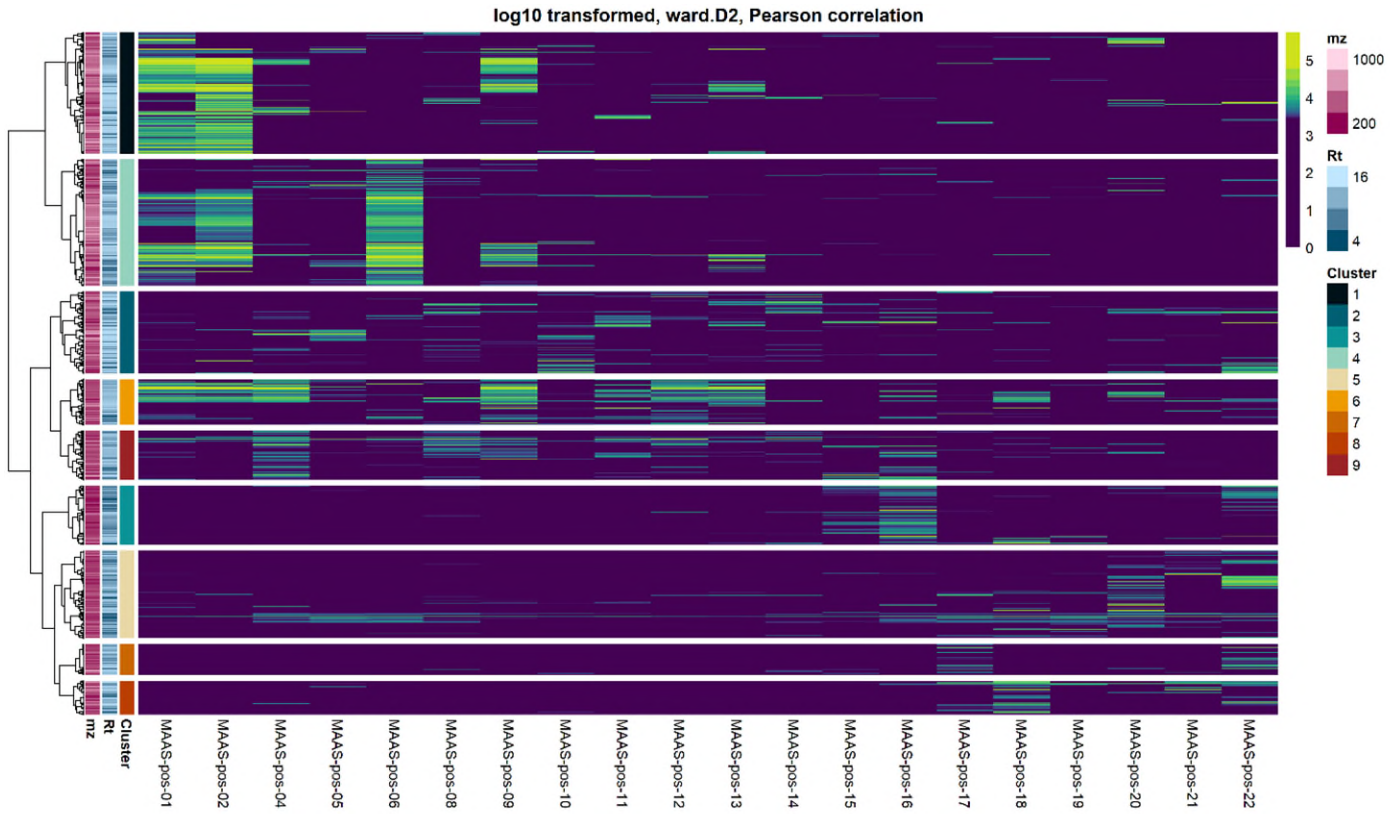
Principal component analysis

- Net als bij GC-MS data, nuttig om **uitbijters op te sporen**
- En voor een **snelle analyse** van de data



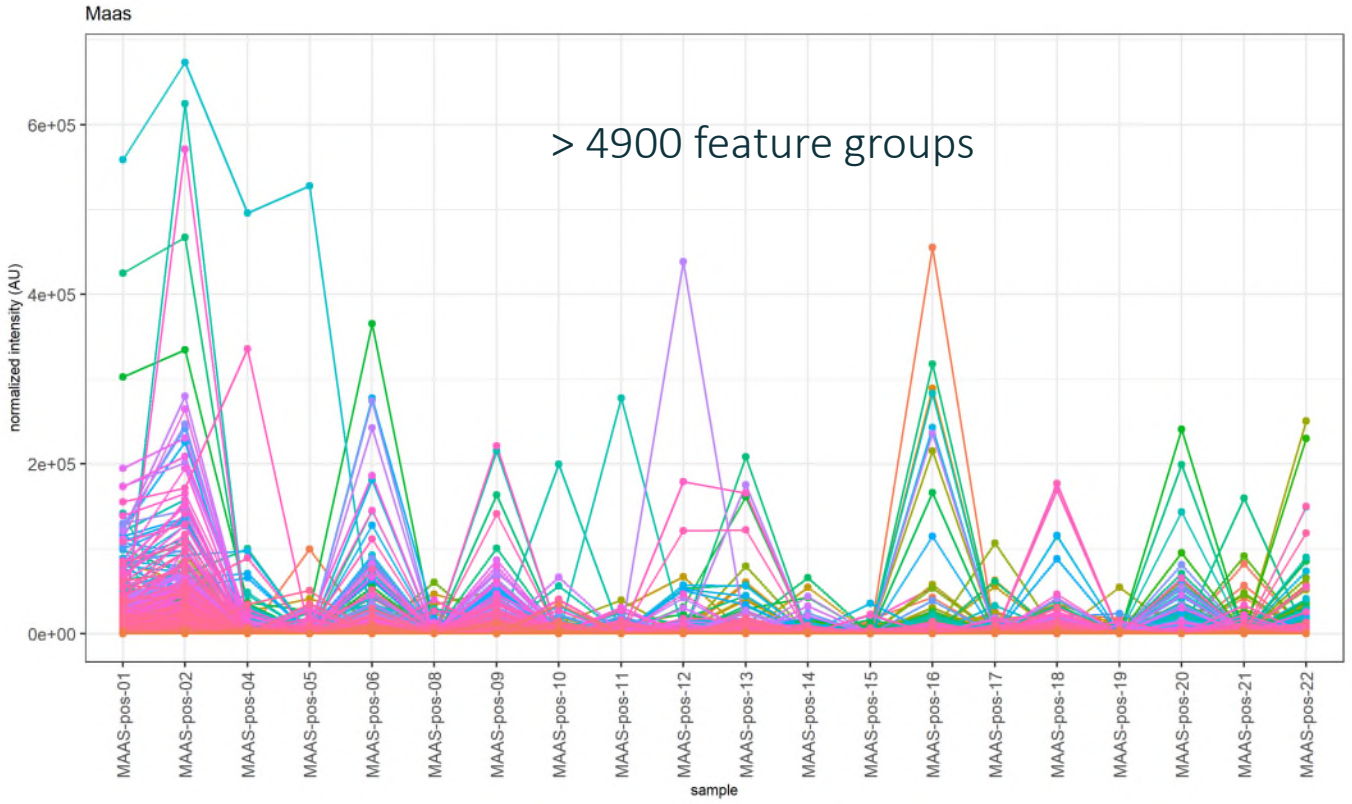
Cluster analyse

- Net als voor GC-MS gegevens, nuttig voor de opsporing van specifieke (groepen van) kenmerken
- In de toekomst: trendanalyse (gecombineerd met correlatieanalyse)



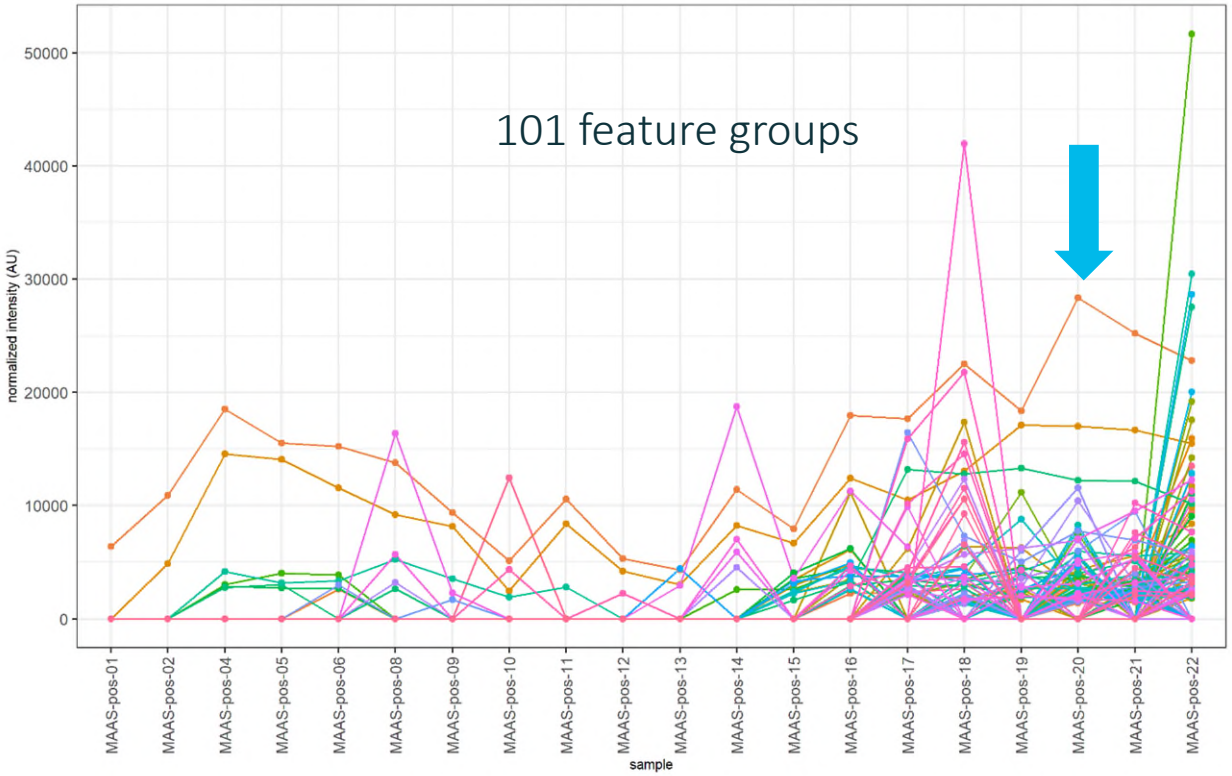
4. Trendanalyse

- Aantal kenmerken te groot voor visuele inspectie
- Filtering (prioritering) op basis van statistische tests/analyse
- 3 opties
  - Regressie-analyse
  - Niet-parametrische correlatieanalysetests
  - Tijdreeksanalyse (niet geïmplementeerd)



4. Trendanalyse

- Selecteer (prioriteer) feature groups op basis van
  - Regressie-analyse
  - Mann-Kendall en Spearman's rank correlatietests

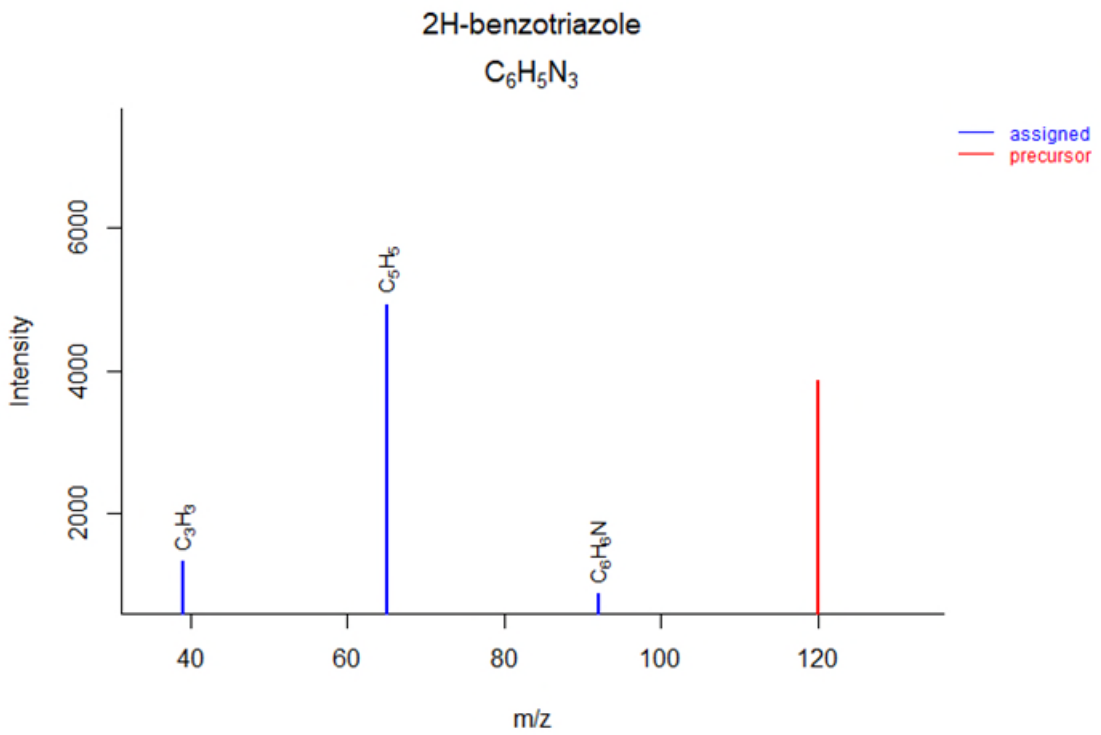


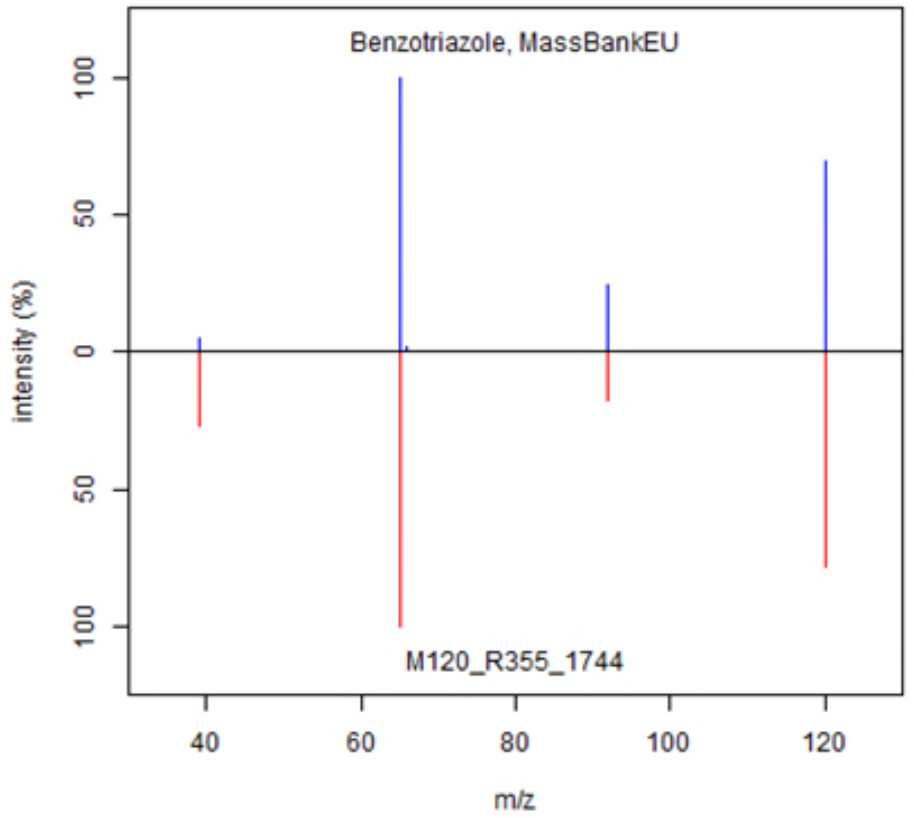
5. Identificatie

- Gebaseerd op MS1 informatie:
  - Mogelijke bruto formule: $C_6H_6N_3$
  - 26 kandidaat-verbindingen. Hoogste score: 2H-benzotriazool
- Kenmerken in kwestie had MS2 informatie ("vingerafdruk")
- MS2 spectra kunnen vergeleken worden met spectrale bibliotheken
  1. MetFrag
  2. MassBankEU

5. Identificatie

MetFrag





Validatie

- Geen informatie over calamiteiten beschikbaar
- Relevante kenmerken kunnen bevestigd worden
- Analyse van referentiestandaarden (of maken reeds deel uit van de routinebewaking)

Vragen?

KWR

Groningenhaven 7
3433 PE Nieuwegein
The Netherlands

T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

@KWR_Water KWR KWR_Water

Nienke Meekel
Nienke.Meekel@kwrwater.nl

Frederic Béen
Frederic.Been@kwrwater.nl

Ton van Leerdam
Ton.van.Leerdam@kwrwater.nl

Colophon

KWR | Juni 2022| 403817

Project number
403817

Project manager
Ton van Leerdam

Client
Rijkswaterstaat

Quality Assurance
Dr. Peter van Thienen, Erik Emke BSc

Author(s)
Nienke Meekel MSc, Frederic Béen PhD

Presented at
- MS Teams
- 9 June 2022

Keywords
Mass spectrometry, trend analysis