

Technical Specifications of the siMPleR script: post-analysis of microplastic data

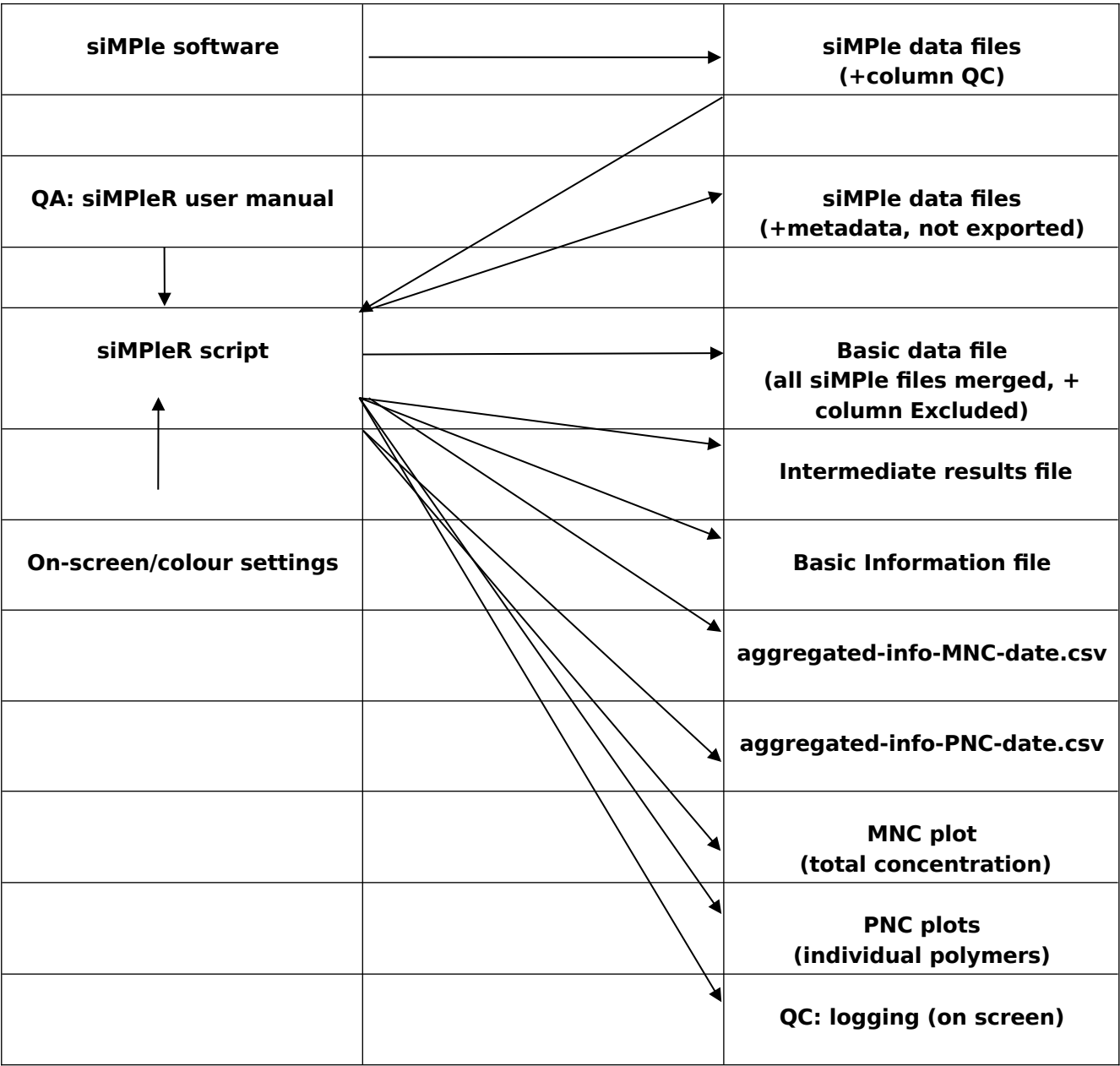
Willem van Loon (RWS) and Dennis Walvoort (WUR)
Versie 13, 14-01-2025

Introduction

The identification and post-analysis of microplastic FTIR data is a complex process, due to the complex mixture of microplastics and natural particles, and the multi-dimensionality of the microplastic mixture regarding polymer types, sizes and shapes. Therefore, one of the functional aims for the post-analysis tool siMPleR is to keep the software and its use as simple as possible.

siMPleR is designed for the post-analysis of siMPle results files. siMPle is well-known software for the automated analysis of FTIR filter files (+link)

The overall process and products of siMPleR are illustrated below.



1. Import of siMPle data files

The siMPle software processes FTIR instrument files and produces a list of MP identifications. The format of this list is:.

Identifier	Coord. [pix]	Coord. [µm]	Max score	Group	No. of pixe	Area on me	Major dim	Minor dim	Feret min	Volume [µm	Mass [ng]
MP_1	[133,228]	[2783,4770]	0.602	polyamide	23	10067.8	153.2	83.6	120.2	336863.7	369.6313

Data preparation

1.1 In the folder INPUT, all the individual siMPle files (each from 1 filter) are placed.

The root folder of the script can be chosen freely. The INPUT folder is placed in this root folder, as well as the OUTPUT folder. This root folder can be set using the `setwd()` command or using the RGui command "File/Change dir...".

1.2 Each siMPle file must have the following header format: `@Loc.code_Y1234_Rx_Ey_G12_freetext`. For example: `@NW2_Y2023_R1_E2_G20`. The order of the following metadata is flexible, but a fixed and logical order as described here is recommended for quality assurance. Additional text is allowed in the filename, after these essential metadata. Note that the codes have to be connected using underscores.

Note that for each location 3 replicates are measured per year. For each replicate 2 extracts are measured. G denotes the sample mass analyzed in gram. Because we use only one sampling date per year, only Year is used to simplify the format.

Note 2: it is allowed that only 1 or 2 replicates per location in the input data. In these cases, siMPleR calculates mean MP concentrations using the number of replicates in the input data.

In addition, it is allowed that only 1 extract per replicate is present in the input data. This input file may even have zero records, which sometimes occurs.

siMPleR anyway calculates the mean values of the microplastics in the 2 (default) or 1 extract(s) per replicate.

1.2 In each siMPle file (see format above), a QC column (with header QC) must have been added manually.

1.3 This QC column may contain the following QC-codes (added by the analyst or by the script):

- ppring (if a PP is a part of the Anodisc support ring; by analyst)
- <50 um (or any other lower length limit used. Note: this code is added by script)
- duplicate (determined by script as some polymer at same coordinates. The particle with the largest length is chosen)
- natural (if the second database shows the identification is a natural particle; code added by analyst)
- = (if the second FTIR database gives the same identification as siMPle; code added by analyst)
- plastic (if the second database gives a different MP identification as siMPle. So the correct polymer identification is uncertain. Code added analyst)
- the QC field may often be empty, if no QC has been performed on that record.
- Additional QC codes are allowed in the siMPle file, and are not processed by the script.

2. Internal data processing

The script performs the following actions when started:

2.1 It asks if duplicate siMPle records must be removed.

2.2 It asks for the minimum reporting length of microplastic particles. Default is 50 um.

2.3 It checks the presence of, and reads all the siMPle data files in the folder ./input.

2.4 It checks the presence of the ./output folder.

2.5 The script performs many format checks, among others:

- If the field separators are commas, and the decimal signs points (UK country setting). In case of semi-colons and field separators and decimal comma's, they are converted to UK setting.
- If the essentially needed data columns are present.
- If the QC column header is present.

2.6 The script performs the following actions:

- The metadata in the filename are added in the basic data format.
- QC codes are added automatically if necessary (e.g. <50 um, duplicate, etc)
- The following QC-codes are excluded from data analysis: <lower length limit, duplicate, ppring, natural.
- The script adds an additional column, "Excluded". Records which do not pass the selection criteria, e.g. <50 um and duplicates, are marked with an "yes" (otherwise "no"). This is very convenient for manual post-analysis.
These excluded records are however included in the basic-data file for transparency (see Ch 3).
- It analyses the number of extract files per replicate and empty files (no MPs found). The script accepts if only 1 extract file per replicate is provided. This can occur during testing, or if only 1 extract per replicate is used. If an extract file is empty, 0 MPs are reported.
- It updates a specific polymer name to "plastic", if the second database indicates a different polymer type. Note that the original polymer name is retained in the basic-data file.

2.7 It performs the calculations to produce the table below.

Note that MNC and PNC concentrations are performed by taking the total number of replicates + extracts per location-year into account, even if a polymer was not found in specific replicates/extracts.

2.8 The script exports the tables described below, and the barplots for MNC and PNCs.

3. Update and export the basic-data file

3.1 Export the basic-data file containing all the original siMPle columns in the same order, the additional metadata from the file name in the left side of the file, the QC column and the column Excluded at the right side of the file, and optional columns with Comments which are present in the siMPle input files at the right side of the file.

3.2 Also note that all the original siMPle records are maintained in this exported file for transparency.

The manual use of this data file is easy because of the metadata "excluded", which can exclude all undesired records in 1 action, just as in the script.

3.3 The filename of the exported filename is: basic-data-yyyy-mm-dd, in which the date is the analysis date.

4. Calculate and Export Total Count per Replicate/Extract

This Intermediate-Results file is valuable to monitor the counts from the two different extracts per replicate. It is normally expected that the count of the second extract is lower than of the first extract.

The format is:

location	year	replicate	sample mass [g dw]	extract1 [# MP]	extract2 [# MP]
VD5	2023	1	10	34	0

The exported filename is: intermediate-results-yyyy-mm-dd.csv

5. Calculate and Export the Polymer-QC-info file

The script uses the polymer QC codes, present in the siMPle records, to construct the following table format:

Group	total	# QCs	# =	# plastic	# natural	% false pos
APU	7	6	0	6	0	0
EVA	3	3	0	3	1	25
PA	2	2	2	0	0	0

The following classification rules are applied to construct this table:

5.1 The siMPle records with the QC codes “ppring”, “<50um” and “duplicate” are excluded from the calculation because they fall outside the valid dataset (results are not valid).

These records can still be found in the basic-data file for information/QC.

5.2 A QC action using an external database may result in 3 results:

5.3 if the external database reports the same polymer, the QC result is “=”.

5.4 if the external database reports a different polymer, but still a plastic, the result is “plastic”

5.5 if the external database reports a natural material (inorganic or organic material), the result is “natural”.

5.6 the % false positives is calculated as: $[\# \text{natural} / \text{total}] * 100\%$, in which “total” is the valid total number of microplastic particles.

6. Calculate and export the basic information file

This file contains for each location-year-replicate the microplastic concentrations (total and per polymer) per kg dry sediment. This detailed information is necessary for statistical calculations of differences between locations, e.g. using the Wilcoxon test. This also shows that microplastic counts are not suitable to make these statistical comparisons, because the different sample amounts hinder this comparison. Also note that only the essential information is presented in this file, and other data has been removed

The format of this basic information file:

Location	Year	Replicate	Parameter	Concentration	Unit
VL1	2023	3	MNC	1500	kg-1
VL1	2023	3	polyethylene	300	kg-1

The exported filename is: basic-information-yyyy-mm-dd.csv.

7. Aggregate and export MNCs per location-period

7.1 Calculate the mean microplastic concentrations using the individual location-replicate-year/period results. If 2 or 3 years of data are available for the same location, these data are aggregated.

The maximum amount of years to aggregate (recommended: 3) is steered via the data input.

7.2 Calculate for each location the Standard Error (SE) of the MNC as a measure of the precision of the MNC; and as an indication of analytical repeatability. The derived Relative Standard Error (RSE) may also be useful to calculate (Bauerlein et al., 2023).

<https://www.sciencedirect.com/science/article/pii/S0141113622002495>

7.3 Report the table format shown below:

location	period	# samples	mnc [kg-1]	mnc se [kg-1]	# MP
NW2	2023-2023	3	1300	50	78

7.4 Export this file with the filename: aggregated-info-MNC-yyyy-mm-dd.csv

8. Aggregated PNCs per location-period

8.1 Calculate the mean polymer concentrations using the individual location-replicate-year/period results.

If 2 or 3 years of data are available for the same location, these data are aggregated.

The maximum amount of years to aggregate (recommended: 3) is steered via the data input.

8.2 Do NOT calculate and report the Standard Error (SE) of the PNCs, because the Wilcoxon test using the basic information file (Chapter 5) will be used when more data are available to test for statistical differences between locations.

8.3 The reported table format is a similar table as in Chapter 6, but without the SE column

8.4 The exported filename is: aggregated-info-PNC-yyyy-mm-dd.csv

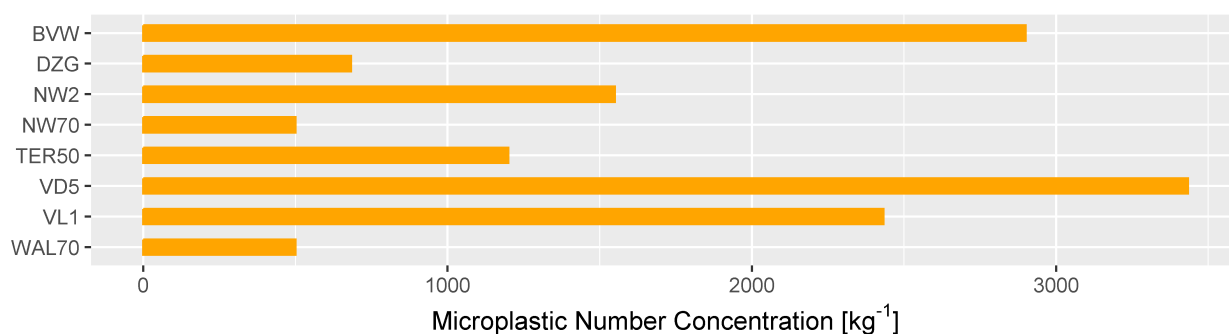
9. Make plots for MNC and PNCs

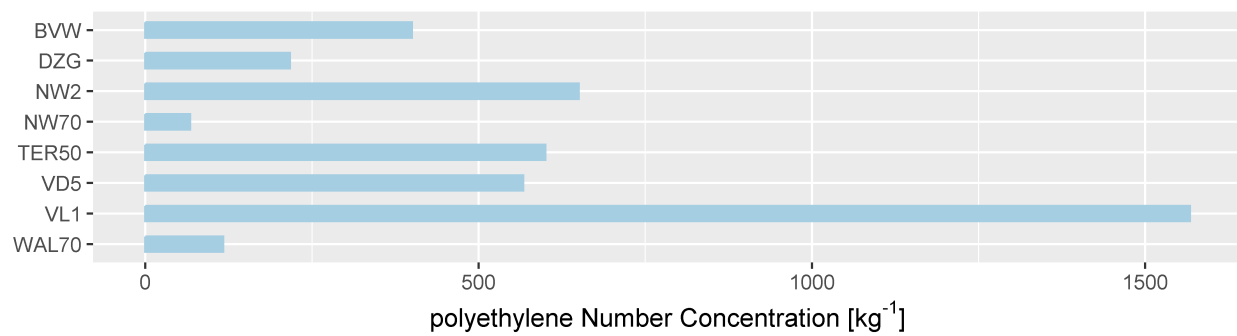
9.1 make a barplots, in horizontal format, for the aggregated MNCs for all locations.

9.2 sort the locations codes alphabetically

9.3 make barplots for all polymers and plastics, same format as MNC

9.4 within these polymer barplots, sort the location codes alphabetically





10. QA/QC

10.1 The QA is organized by providing a user manual and by clear error messages of the script.

10.2 The QC is implemented using on screen logging of specific and clear error messages.

This logging must be checked carefully after a run.

10.3 The file intermediate-results file can be used for QC of the microplastic numbers per replicate and extract.

10.4 The polymer-QC-info provides relevant information if the threshold values are optimized, or can still be improved.

Annex 1: siMPleR import function for DONAR export format.

RWS will probably store microplastic data in DONAR (the Rijkswaterstaat database) in the format: location | date | (replicate nr) | polymer x | length | width, using a MUX (multiplexed data series).

The replicate nr may be stored on three different adjacent dates.

At this moment, this DONAR format import function is not needed, but it may be needed in the future.

Annex 2: color scheme used

ColorBrewer: Color Advice for Maps

<https://colorbrewer2.org/#type=qualitative&scheme=Paired&n=12>

