**SIFT**

**Background:** SIFT is a user-friendly sediment fingerprinting tool based upon the “r” platform and a Shiny user interface. It assumes a basic knowledge of the fundamentals of how sediment fingerprinting works (see Collins et al., 2017 for a recent review) but does not require knowledge of statistical or modelling techniques. This tool steps end-users through all critical steps in processing fingerprinting data, providing a simple interface for each stage of the methodology, guiding the user through the key choices which need to be made, and automatically doing any calculations, statistical tests or modelling.

There are four main concepts behind the structure of this tool.

1: Ensuring sediment source groups best fit the tracers and catchment in question by running models based on land use, geology and cluster analysis derived source groups.

2: Removing any tracers which are poor discriminators or are potentially non-conservative to minimise this component of uncertainty in the procedure.

3: Extracting maximum value from tracers that are particularly good discriminators and are most likely to be conservative by forcing them into composite fingerprints, weighting them in apportionment models and using specialised composite fingerprints.

4: Using three different composite fingerprints in each model to reduce the potential for a non-conservative tracer or poor discrimination to impact on model results for source apportionment.

By combining these concepts and running up to 15 separate un-mixing models, a more complete understanding of what the tracers being used reveal about sediment provenance can be established than with the use of a single pre-determined source grouping scheme and composite fingerprint. A flow diagram of each key step in the SIFT software is provided below.



**Instructions**

**Installation:**

Before running the app, you must install R and RStudio. You must also install JAVA and ensure that your operating system, version of R and java version are all either 32-bit or 64-bit. E.g. a 32-bit version of java will cause an error on windows 10.

Extract the folder contained in the SIFT .zip archive into a folder somewhere on your computer and click the app.R icon to run it. The first time the model is run any required packages will be automatically installed. This may take some time so be patient. For subsequent runs the software will open immediately.

**Data Input:**

This app comes with pre-loaded tracer data for overbank sediments retrieved from the River Nene basin in the UK after a major flood event which can be run to familiarise you with the app and its functions.

Using the test dataset, the pre-assigned source groups are:

Land use

1 = Urban

2= Cultivated

3 = Grassland

4 = Channel Banks

Geology

1 = Urban

2 = Diamicton

3 = Marlstone

4 = Mudstone

5 = Sands and Gravels

6 = Limestone

7 = Ironstone

Sediment samples were collected in sequence from the floodplains along the River Nene’s main channel. They start in the east and progress through the towns of Northampton and Wellingborough in its centre and west.

To use your own data, simply overwrite the cells in the input.xls spreadsheet file in the tool’s source folder. The input sheet contains your source tracer data and the sediments sheet contains your sediment data. There is an optional key sheet to record source group numbers, but this is not used by the software. In the input sheet, columns A:E must not be moved or deleted and must contain data. If you are missing any data just repeat “1” in every cell. From column F onwards, rows can be any tracer data you have. Similarly, in the sediments sheet, column A must be the sample names. The tracer data must also match that in the input sheet and be in the same column order with the same names on the top row. Make sure there are no words, errors or missing values in either sheet or the model will reject any affected samples.

Once you have added the data simply save and close the sheet.

You may also upload a map of your study area. This is recommended as it will make interpreting the model outputs simpler. The map must be saved into the “GISInputs” folder and named map.tif. The map must be in a georeferenced .tif format with the same coordinate system as the source samples (which for both must be in a numerical format e.g. decimal degrees). It is recommended that this map is kept simple as the resolution will be reduced in the software to minimise loading times. It is also recommended that the locations of sediment samples are marked on the map.

**Running the model:**

Once app.R has opened in RStudio and you have added your data to the input.xls sheet click the run app button in RStudio to start it.

Basic instructions are provided in each sheet.

## Page 1: Setup

This page allows you to set up the parameters for the model. It is often optimal to run the model multiple times using different settings. You can only set these parameters once at the beginning, so if you wish to change them later on, you will have to restart the software from the beginning.

The yellow sliders allow for parameters associated with the **tracer variability ratio** to be set. The tracer variability ratio is the ratio between the percentage difference in tracer concentration between two source groups and the mean within-source group coefficient of variation for the two source groups. Users can select values for the following:

The minimum mean variability ratio for any pair of sources.

The minimum highest variability ratio found for any pair of sources.

The default values for these are set at 1 and 2. If your dataset contains tracers which are very good discriminators increasing these values may result in a lower range of uncertainty in un-mixing model results. Lowering the values may be necessary if your tracers are poor discriminators; however, lowering values below 1 will likely result in a large increase in uncertainty being propagated through the remainder of the data processing. Setting values too high may cause all tracers to be rejected by the model.

The red sliders control the **range / bracket tracer conservatism test**. This test, remains a fundamental component of data processing and evaluates if the tracer concentrations in the sediment samples fall within different ranges (3 range descriptors are provided in the tool) found in the source groups. The sliders set the percentage of sediment samples required to fall within each range. If any tracer fails any one of these tests it is removed from further use. The tracers falling within medians is, by default, set to 0, as it is possible for tracers which are conservative to fall outside the medians if one sediment source dominates. It may be beneficial to increase these ratios if your subsequent un-mixing model results have very poor goodness-of-fits or, if different composite fingerprints produce very different sediment provenance results.

The purple slider allows you to set the **mean tracer measurement error**. This value will change the value input for sediments and virtual mixtures into the subsequent un-mixing model by random amounts within the range set.

The first blue slider controls the **correlation conservatism test** which uses bi-plots of correlated tracers with the source and sediment data to determine if relationships between tracers in the source samples are maintained in the sediment samples. This is now used as an addition to the above range / bracket tests. The set value (default 0.8) is the r2 required for two tracers to be classed as correlated.

The second blue slider ensures that tracers which are correlated with each other in each source group have Monte Carlo values which are also correlated in the un-mixing model. The slider sets the minimum r2 value required for a tracer to be classed as correlated. This ensures uncertainty analysis for un-mixing modelling using Monte Carlo iterations maintains correlations between tracers in both source and sediment samples.

The first green slider sets the minimum improvement required in source discrimination for the inclusion of an additional tracer in the linear discriminant analysis (\*100 = percentage improvement).

The second green slider specifies the number of Monte Carlo iterations to use in each un-mixing model. The default is set at 500 so that the model can be trialled without calculations taking too long. When you have run through the model a few times to determine the optimal settings for your data it is recommended to set the value at 3000 – 5000 for the final run, since many existing publications use this range of iterations.

The final green slider sets the goodness-of-fit threshold required for a model result to be accepted. Setting this value higher can reduce model uncertainty; however, it can cause very few Monte Carlo iterations to pass the threshold creating a number of potential errors. If you increase the goodness-of-fit threshold substantially, you will also need to increase the total number of Monte Carlo iterations substantially (likely range – 1000 to 5000).

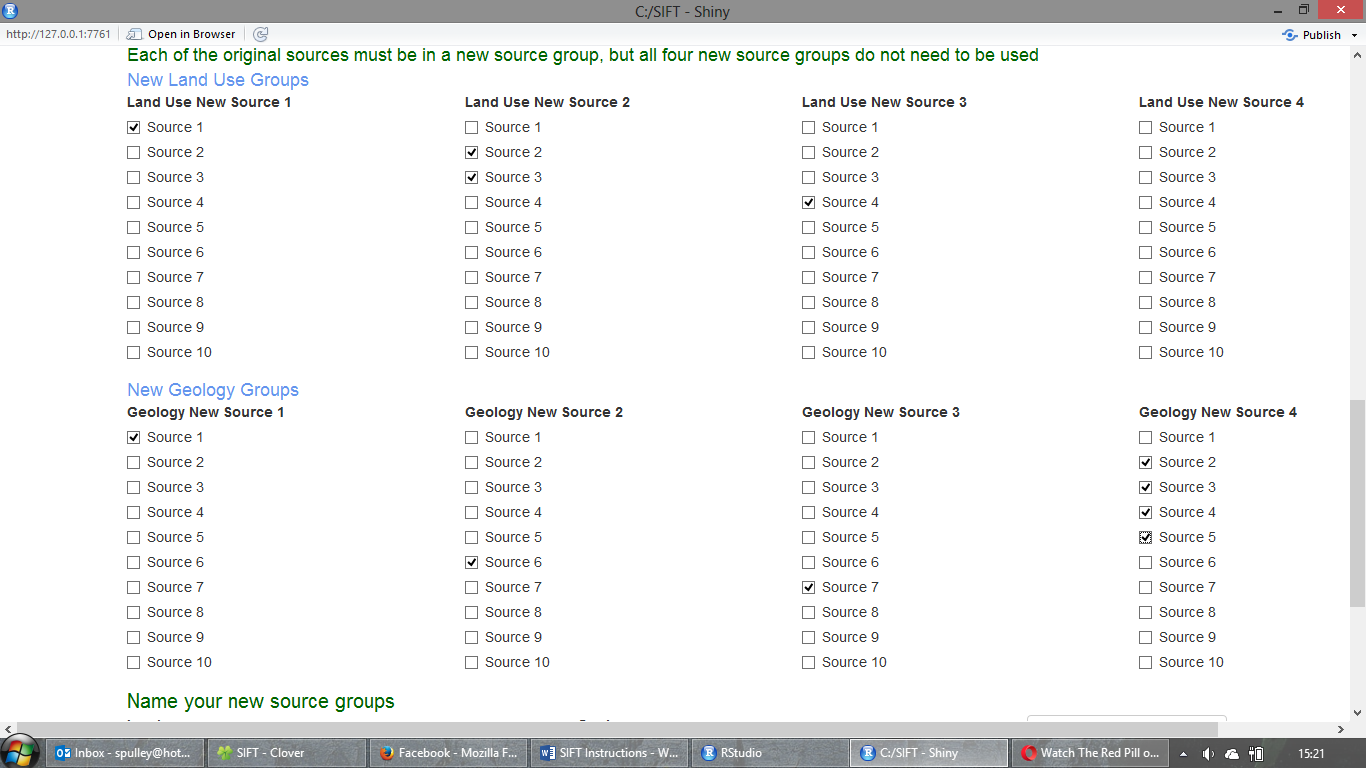
## Page 2: Sediment Samples

This page tests if each tracer in each sediment sample falls within the full minimum – maximum range found in the source groups. If multiple tracers fail for any sediment sample it may suggest that dissolution, organic matter or particle size effects are controlling tracers in that sample. If this is the case (indicated by there being multiple FALSE cells) then tick the sample in the list at the bottom of the page to delete it from further analysis. Leaving in non-conservative samples can cause problems when a high percentage of samples are needed to pass the range / bracket test.

## Source Group Classification

It is likely that the available tracers will not be able to discriminate all source groups. It is also the case that using too many source groups will result in equifinality problems in source apportionment. For this reason, your pre-assigned land use and geology source groups must be assigned into one to four new source groups based upon the ability of the tracers to discriminate. Each of the original source groups must be in a new source group, but all four new source groups do not need to be used.

Use the tick boxes to assign the new groups in the following way. There are four sets of boxes for land use and four for geology underneath them. Below the boxes you may name each of the new source groups. These names will be carried through each page of the modelling procedure.



## Misclassified Samples

The misclassified samples screen runs a new linear discriminant analysis and identifies any samples which are potentially misclassified using bi-plots showing the two largest discriminant functions as well as maps of the catchments; both showing samples identified by the LDA to be misclassified. Points are coloured as their current source group, labels are coloured according to the group the sample best fits into if that is not the current group.

Potentially misclassified samples can either be left as they are by ticking no box, reclassified, or deleted. It is important to only reclassify or delete a sample if there is robust justification to do so. Reclassifying too many samples may remove natural variability within the source groups you have selected. Many samples identified as misclassified will likely reflect poor discrimination between the source groups.

The remaining steps of the model can be completed by just clicking the buttons on each screen to move to the next. It is, however, recommended that you read the instructions on each page and follow the suggested steps. A publication has been submitted for peer review and should shortly be available to assist end-users with this software.

Collins, A.L., Pulley, S., Foster, I.D.L., Gellis, A., Porto, P. and Horowitz, A.J. (2017). Sediment source fingerprinting as an aid to catchment management: a review of the current state of knowledge and a methodological decision-tree for end-users. *Journal of Environmental Management* 194, 86-108.