

SedSAT User's Manual

Arash Massoudieh

Allen C. Gellis

Cara L Peterman-Phipps

April 2023

Executive Summary

The Sediment Source Assessment Tool (SedSAT) User's Manual provides comprehensive guidance for users of the SedSAT3 software, a tool developed for sediment source fingerprinting. The software facilitates the analysis of sediment sources and their contributions to a particular target site by utilizing elemental composition data. Key functions of SedSAT3 include data preparation, statistical analysis, and sediment source attribution using various methods such as Levenberg-Marquardt Maximum Likelihood Estimation, Genetic Algorithm Estimation, and Bayesian Inference. The manual outlines the steps required for importing and preparing raw data, followed by pre-analysis tasks like correlation matrix generation, outlier detection, and organic matter and particle size corrections. It also provides tools for advanced statistical analysis, including variance analysis and discriminant function analysis, to aid in selecting the most effective elements for fingerprinting. The manual culminates in detailed instructions for executing source attribution models and performing uncertainty analysis. A mathematical appendix explains the theoretical foundations for the modeling methods used in the software, offering users both practical and technical insights for sediment fingerprinting.

Contents

1	Raw data preparation	4
1.1	Importing the data	4
2	Deterministic Fingerprinting tools	7
2.1	Outlier analysis	7
2.2	Organic matter and particle size correction	9
2.3	Bracketing Analysis	13
2.3.1	Stepwise Discriminant Function Analysis (DFA)	14
2.4	Multi-way discriminant function analysis	15
2.5	Fingerprinting	19
2.6	Fingerprinting (Batch)	21
3	Bayesian Sediment Fingerprinting	23
3.1	Bayesian Chemical Mass Balance Analysis	23
3.2	Bayesian Chemical Mass Balance Analysis (Batch)	27
4	Other Statistical Tools	31
4.1	Elements' correlation matrix	31
4.2	Analysis of Variance	33
4.3	Auto-select elements	35
4.4	Two-way DFA	38
4.5	One vs. the rest DFA	42
4.6	Distribution fitting	44
4.7	Elements' Discriminant Power (Multi-way)	45
4.8	Elements' Discriminant Power (Two-way)	47
4.9	Error Analysis	47
4.10	Kolmogorov-Smirnov for individual group/element	49
4.11	Kolmogorov-Smirnov for a group	50
4.12	Optimal Box-Cox parameters	51
4.13	Source Verification	51
4.14	Genetic Algorithm estimation	52
5	Appendix A: Mathematical basis	56
5.1	Mass balance formulation	56
5.1.1	Stable Isotopes	56
5.2	Maximum Likelihood estimation	57

5.2.1	Treating source elemental composition deterministically . . .	57
5.2.2	Treating source elemental composition as unknown	58
5.3	Bayesian inference	59

Chapter 1

Raw data preparation

The raw data for SedSAT3 must be prepared as a single Excel spreadsheet file with multiple tabs. The first tab in the Excel file contains the elemental profiles of the target samples. The subsequent tabs contain the elemental profile for each source group. In each sheet, each row is assigned to one sample, the first column contains the names of samples, and each other column represents one element. In order for SedSAT3 to function, the data must be properly formatted (links to example datasets given in the example sample data file, downloaded from <https://sedsat.org/wp-content/Data/SampleData.xlsx>). The specific text identifying column/field headers is unimportant, but the relative location of columns in the source and target datasets must be identical and is a core requirement for the program to run successfully. Column/field headers should be unique; no field name should match another field name within the same dataset. Note that each sheet must contain exactly the same elements named in exactly the same way.

1.1 Importing the data

To import the data, choose **File→Import Data from Excel** from the menu bar located at the top left of the program, then navigate to the Excel file.

Before the Excel file will be imported, you need to select the tabs you want SedSAT3 to import.

When the Excel file is imported, a dialog box will appear asking you to identify the sheet that contains the target samples (Fig. 1.1). Select the radio button that corresponds to the target sample sheet, double-check that each subsequent Excel sheet is your source samples, and then select Ok.

After identifying the target sample spreadsheet, a dialog box will appear to prepare for including/excluding samples, and then the constituent properties window will appear (Fig. 1.2). Each window panel and individual columns can be resized by hovering your mouse between column headers; a double arrow will appear that allows the user to resize for better viewing.

In the middle window panel, there are five columns, the constituents are listed in the first column, and in the second column (Constituent Type), you can double-click on each constituent type to get a drop-down menu from the following list:

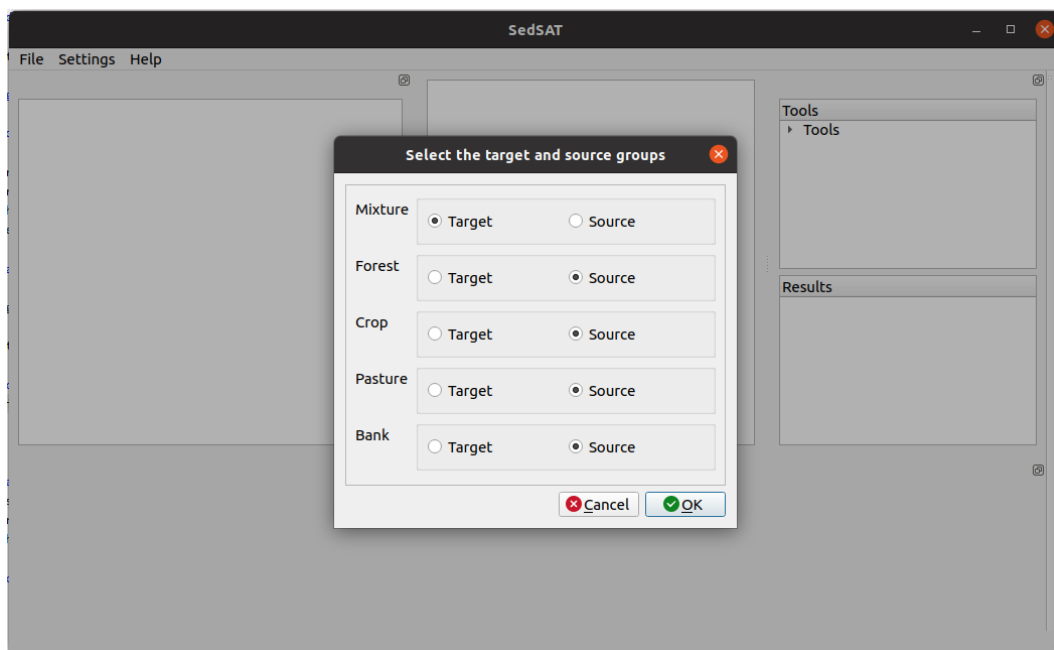


Figure 1.1: Indicating the target sample

- Element
- Isotope
- Particle size
- Organic carbon
- Exclude

Additionally, you can also exclude an element from the analysis by checking off the value in column five of the table. For isotopes, you need to choose the base element and the standard isotope ratio. This is needed for converting δ values to the elemental content of the isotope. For example, for ^{15}N , the base element is N and the standard isotope ratio is $(^{15}\text{N}/\text{N})_{std} = 0.003676$. If a constituent is indicated as particle size, organic matter, or excluded from the analysis, it will not be used directly for sediment fingerprinting.

You can always revisit the constituent properties through the top menu item **Settings**→**Constituent properties**

Now is a good time to save your project using **File**→**Save** or pressing **CTRL+S**. The file will be saved in a readable JSON format with a file extension **.cmb** (CMB Source file). You can open it using any text editor and inspect the contents. You can load the project anytime in the future.

You can visualize the elemental profiles using the tree view on the left side of the screen labeled data (Fig 1.3).

Choosing samples to be discarded

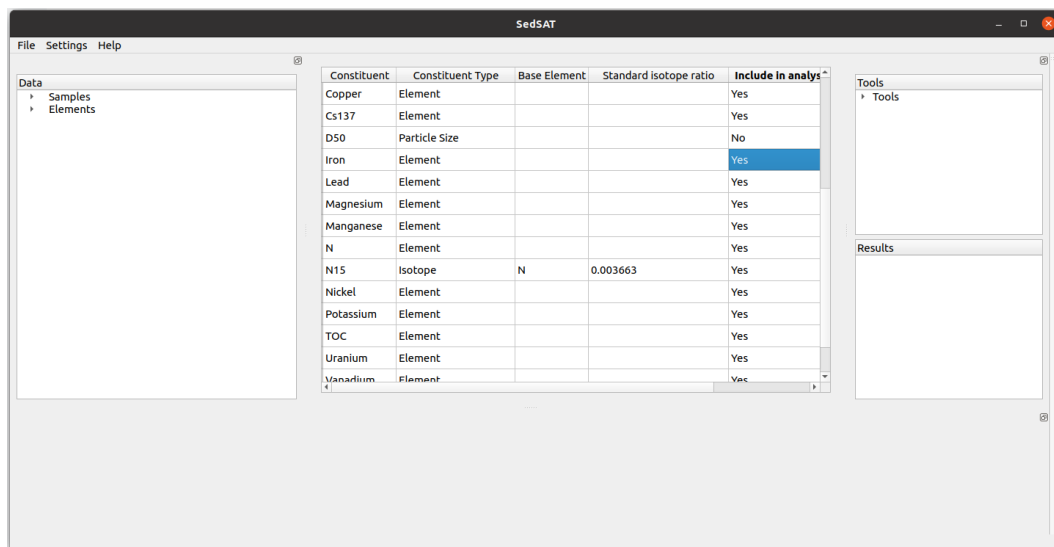


Figure 1.2: Specifying Constituents' properties

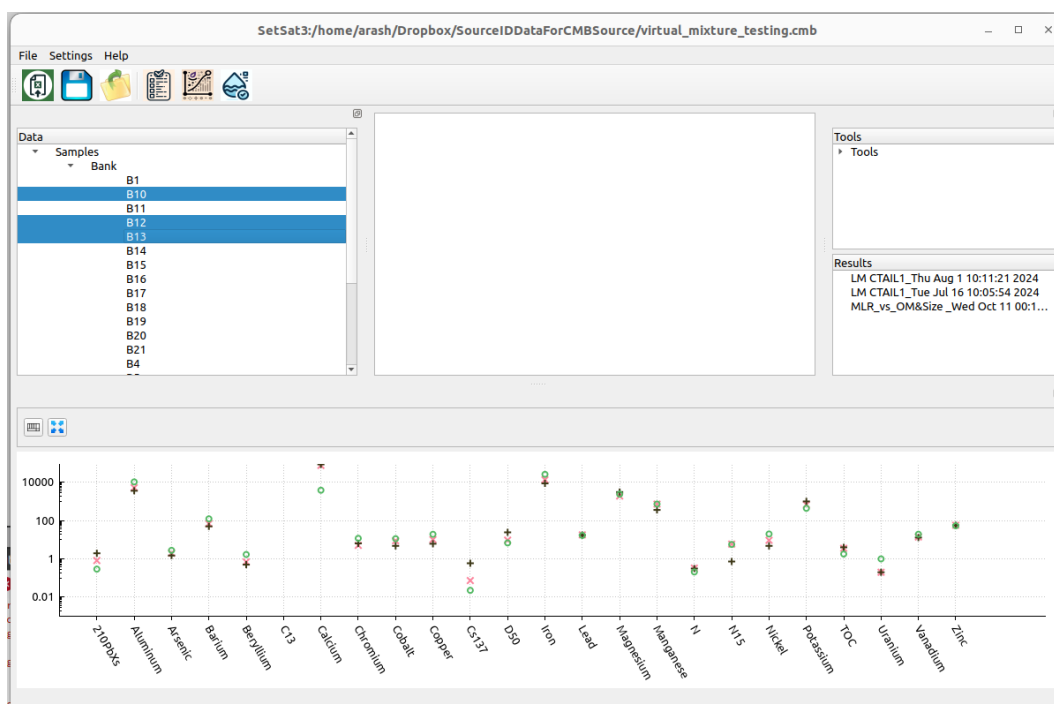


Figure 1.3: Visualizing the raw data

Chapter 2

Deterministic Fingerprinting tools

This chapter outlines the complete sequence of steps and tools required to perform maximum-likelihood (deterministic) sediment fingerprinting. The process includes *Outlier Analysis*, *Organic Matter and Particle Size Correction*, *Bracketing Analysis*, *Stepwise Discriminant Function Analysis*, and concludes with the *Fingerprinting* procedure.

2.1 Outlier analysis

The outlier analysis detects possible outliers in the elemental contents of individual samples. It first maps the concentration of each element in a source group using the Box-Cox transformation to the closest distribution to a Gaussian distribution.

$$\hat{y}_{i,j,k} = \begin{cases} \frac{\left(\frac{y_{i,j,k}}{\sigma_{i,j}}\right)^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log\left(\frac{y_{i,j,k}}{\sigma_{i,j}}\right) & \text{for } \lambda = 0 \end{cases} \quad (2.1.1)$$

where $\sigma_{i,j}$ is the standard deviation of element j in source group i .

The outlier analysis module first finds the optimal value of λ that results in the closest transformed elemental concentrations to a normal distribution and then transforms the data using the optimal λ . The normality level is measured based on the Kolmogorov–Smirnov test. To view the optimal values of λ , use **Tools→Other Statistical Tools→Optimal Box-Cox parameters**.

After the Box-Cox transformation, the outlier analysis feature reports the deviation from the mean for each element in each sample of the selected source group. The deviation measure is calculated as:


$$z = \frac{y_i - \mu}{\sigma} \quad (2.1.2)$$

where μ and σ are the mean and standard deviations of the transformed element content in the group.

The program will automatically run the outlier test when data is imported. However, the user must go to Settings - Include/Exclude samples and

in the Include column uncheck 'Yes'. The outlier test uses a standard deviation of 3 as a default. To view samples that fail the outlier test the user can look at the Include/Exclude Samples tab. In the Notes column, samples that have an element as an outlier are highlighted in red. Should the user want to change the standard deviation of the outlier test they will need to re-run the test using the directions below.

To perform outlier analysis double-click on **Tools→Pre-Analysis Tools→2-Outlier Analysis**. From the form that appears, select the source group you would like the outlier analysis to be performed on. The threshold indicates the value which, if the absolute value of z exceeds, the item will be highlighted in red. The default value of the threshold is 3, which means the elements outside of the 99% bound will be highlighted in red. If you check the *Use only selected elements* item, the Outlier Analysis will be generated only for the elements selected in **Settings→Constituent properties**. Also, checking the *Use only selected samples* checkbox only includes the samples that are selected in the analysis in **Settings→Include/Exclude samples**.

Click the Ok button at the bottom of the form and then click on the table  on the window that appears. A table like Figure 2.1 will appear. The values outside of the specified threshold will be highlighted in red.

At this point, if the user wants to remove a sample based on the outlier analysis, you must go to the left panel and select Settings, Include/Exclude samples and under the Include column, double click and check off the Sample. It will appear as a 'No' (Fig. 2.2).

Outlier Analysis

Export to CSV


	210PbXs	Aluminum	Arsenic	Barium	Beryllium	C13	Calcium  g/g	Chromium	Cobalt
B1	-0.372509	-0.526901	-2.07196	-0.0218383	0.069718	-0.457691	1.04649	-0.374478	-0.4120
B10	1.01517	-0.804048	-0.166108	-0.667989	-0.890575	-1.96728	1.1047	-1.60253	-0.2652
B11	-0.389519	0.713483	-0.166108	0.520545	0.997632	1.23901	-1.28993	-0.0520896	0.35245
B12	2.05418	-1.96584	-0.754337	-1.76005	-2.43904	-1.56471	1.2491	-1.02029	-1.4646
B13	0.0698679	1.99086	0.510373	2.1538	1.91773	0.382139	-1.17556	0.698117	0.47036
B14	-1.0563	1.14458	2.39658	1.3409	0.997632	-1.52334	-0.2616	1.54164	2.3039
B15	1.07239	-0.356825	0.510373	0.215186	1.22728	-2.60121	0.215079	0.882702	0.55886
B16	-1.9566	-0.0411799	-0.06725	0.0734182	0.069718	0.00979567	0.627666	-0.0520896	-0.5587
B17	-0.0773437	0.41088	0.580218	-0.496752	-0.890575	-1.01256	-1.31367	-0.0520896	0.11694
B18	0.674323	-0.502318	0.362433	-0.0750235	0.069718	-1.73417	0.850761	0.279102	0.08752
B19	0.399075	0.149978	-0.754337	-0.869667	0.069718	-1.16929	-0.807986	0.337001	-0.2652
B20	-0.128452	1.67412	0.952264	1.04528	0.069718	-1.85669	-0.884676	2.3184	1.32657
B21	-0.113818	-1.53111	1.00767	-2.09454	-0.890575	-1.71428	-1.19287	-1.17816	-1.0563

Figure 2.1: Outlier table

Choosing samples/elements to be discarded

The user can decide which samples or constituents will be used in the sediment fingerprinting analysis based on the correlation, outlier, discriminant function, and bracketing analyses. To select the constituents to be included in the analysis from the top menu, choose **Settings→Constituent Properties**. Follow the instructions in section 1.1 to include or exclude constituents in the analysis. Choose

Settings→Include/Exclude Samples to exclude individual samples from the analysis. By clicking on this menu item, the program does the following:

- Performs outlier analysis on the source groups with a deviation threshold of 3.0 if it has not been done before.
- Performs bracketing analysis on the target sample.

A window like Fig 2.2 will appear. Select the source or target sample group from the top drop-down menu. Under the "Notes" column, the outlier and bracketing analysis results will be shown. The items that exceed the threshold given when outlier analysis was done or based on a default threshold of 3.0 if the outlier tool was not triggered before. Under the "Include" column, you can choose which samples to include or exclude in the analysis from this point on.

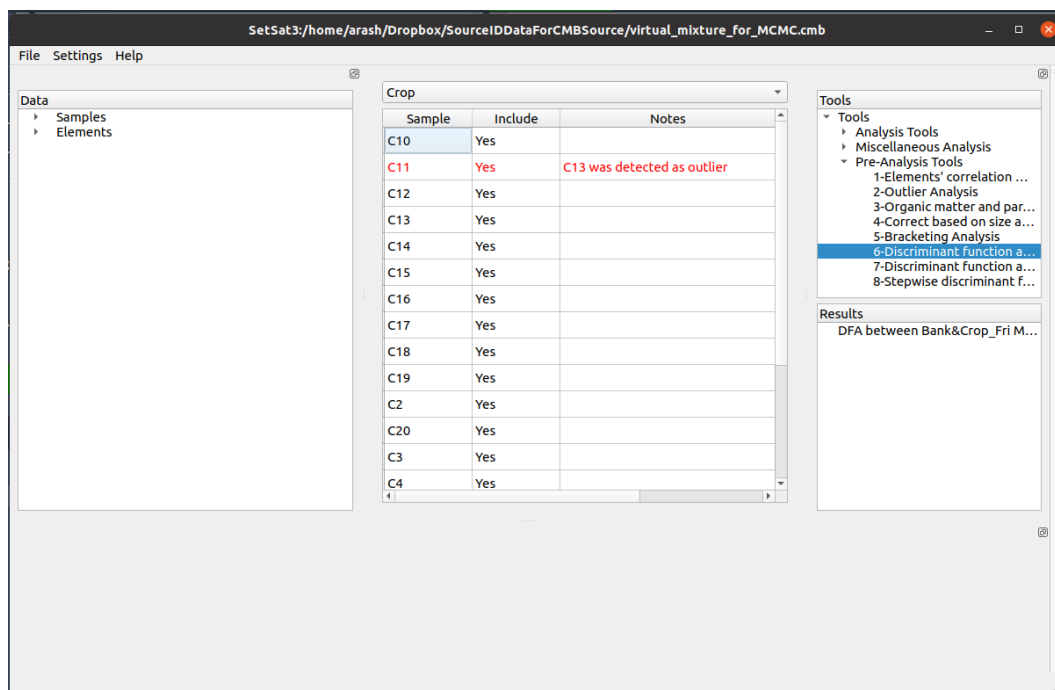


Figure 2.2: Including or excluding samples from the analysis

2.2 Organic matter and particle size correction

In sediment source fingerprinting, organic matter, and size correction are important considerations when attempting to identify and apportion sediment sources within a watershed or catchment area. These corrections help account for the effects that different sediment sources, such as soil erosion, agricultural activities, or construction can have on an area and can vary in organic matter content and grain size distribution. Organic matter and size correction effectively account for the fact that the portion of particles mobilized from a source can have a different organic

matter and size distribution than the samples collected to represent the source. A multiple linear or non-linear regression relationship between the elemental contents of different elements and the organic matter content and mean or effective size of the particles is performed. The relationship is then used to correct the source samples' elemental content based on the target samples' organic content and size. Note that this approach assumes that all sources contributing to the target sample are the same size and have the same organic content as the target sample.

In SetSat3 two regression options are available: linear regression and a power-law regression:

Linear regression:

$$y = a + \sum b_i e_i \quad (2.2.3)$$

Power regression:

$$\ln(y) = a + \sum b_i \ln(e_i) \quad (2.2.4)$$

where a is the intercept, b_i is the coefficient for the independent variable i , and e_i is the value of independent variable i . The independent variables can be particle size, organic matter content, or both. After the estimation of the regression coefficients and intercept, the corrected value of each element in each of the source samples is corrected as:

Linear regression:

$$y_c = y + \sum b_i (e_{i,t} - e_{i,s}) \quad (2.2.5)$$


Power regression:

$$y_c = y \prod \left(\frac{e_{i,t}}{e_{i,s}} \right)^{b_i} \quad (2.2.6)$$

where $e_{i,t}$ is the value of independent variable i in the target sample, and $e_{i,s}$ is the value of independent variable i in the source sample.

Performing Organic matter and Particle Size Correction:

In the tools window, click on **Tools→Steps in Sediment Fingerprinting→2-Organic Matter and Particle Size Correction**.

Select the regression equation using the combo box labeled Equation, and the constituents representing organic matter and particle size on the form that appears. If the constituent representing organic matter or particle size is unavailable or you only want to correct based on one measure, leave the constituent for the dependent variable not to be used for correction blank. Click on the Ok button at the bottom of the form. The results window that appears contains the regression results for all of the constituents in all the sources. To visualize the regressions, click the graph button . A window like Fig. 2.4 will appear. Select different constituents and independent variables from the combo boxes at the top.

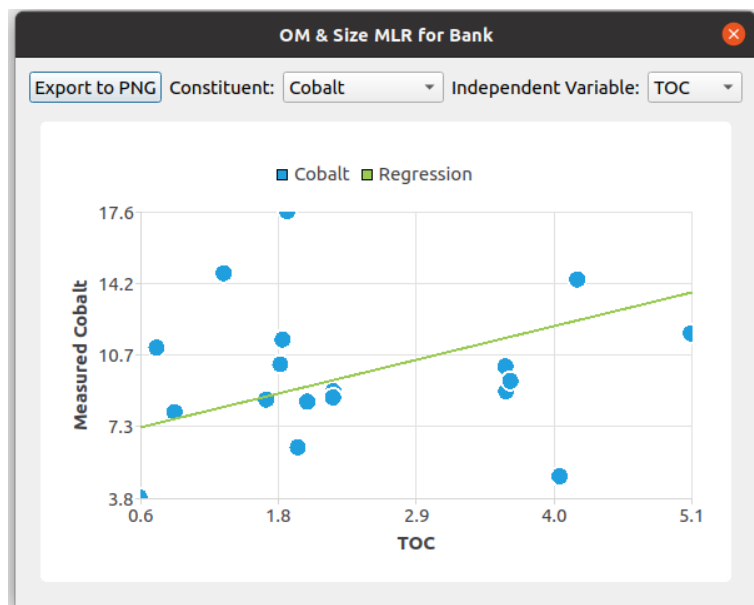



Figure 2.3: Graphical representation of OM and size correction regressions

To see the regression results in tabular form, click on the table button  on the side of each source group in the results window.

	Intercept	OM coefficient	OM P-value	size coefficient	Size P-value
210PbXs	-0.589427	0.0653904	0.469416	0.0797763	0.000518995
Aluminum	9093.24	342.86	0.538302	-293.37	0.0202086
Arsenic	2.71721	0.253133	0.682008	-0.0795158	0.668779
Barium	104.501	5.84089	0.204918	-3.38497	0.00970241
Beryllium	1.2376	0.0592961	0.500015	-0.0411199	0.0473312
C13	-24.6594	-0.678625	0.00288406	-0.0126556	0.968598
Calcium_☿g/g	1775.44	4213.32	0.767584	2362.33	0.425526
Chromium	10.4417	0.928027	0.395582	-0.268651	0.434356
Cobalt	11.7785	1.46898	0.0800614	-0.513863	0.0424314
Copper	16.2161	1.70939	0.255085	-0.697032	0.101981
Cs137	-0.203465	-0.0275829	0.0719158	0.0335078	1.49248e-07
D50	8.88178e-16	-6.66134e-16	nan	1	4.32776e-237
Iron	23313.8	2096.82	0.217999	-1000.43	0.0371425
Lead	15.1165	1.71822	0.0828356	-0.14704	0.79846
Magnesium	1160.46	333.069	0.0577624	13.9466	0.938182
Manganese	844.505	113.33	0.0798747	-36.5049	0.0635967

Figure 2.4: OM and size correction results

The table contains the P-values that indicate the statistical significance of the OM and size. The p-values below 0.05 are in red.

Viewing the corrected elemental profiles


Based on the p-values for the coefficients of OM and size, you can decide whether to correct based on each of those. From the top menu, left-hand side select **Set-**

tings→Organic Matter and size correction. A form including the regression results will appear as in figure 2.5 will appear in the middle panel.

Bank							
Constituent	Intercept	efficient for TOC		efficient for D50	rect based on		
210PbXs	-2.50438	2.10588	0.00138168	Yes	-0.264997	0.95193	No
Aluminum	10.2831	0.154011	0.303189	No	-0.693915	0.00394884	Yes
Arsenic	1.21223	0.225186	0.620229	No	-0.256419	0.824849	No
Barium	5.7129	0.18938	0.145901	No	-0.633258	0.00522085	Yes
Beryllium	1.34889	0.171457	0.309783	No	-0.679652	0.0110126	Yes
C13	nan	nan	nan	No	nan	nan	No
Calcium	7.5767	0.694446	0.466192	No	0.783498	0.736864	No
Chromium	2.62752	0.0521972	0.946624	No	-0.189521	0.803667	No
Cobalt	4.10916	0.54608	0.00417343	Yes	-0.98625	0.0042393	Yes
Copper	3.9474	0.226607	0.456526	No	-0.707071	0.11804	No
Cs137	-6.63684	0.643611	0.103267	No	1.39785	0.0451896	Yes
D50	2.10942e-15	4.44089e-16	nan	No	1	2.15896e-228	Yes
Iron	11.6427	0.343856	0.0664506	No	-0.939073	0.00544183	Yes
Lead	2.82248	0.201983	0.120562	No	-0.0523041	0.951766	No
Magnesium	7.23132	0.272796	0.228475	No	0.0699556	0.96803	No
Manganese	8.66727	0.828482	0.00134521	Yes	-1.19672	0.00835124	Yes
N	-1.71697	0.660705	3.5527e-10	Yes	-0.104319	0.440478	No
N15	4.96234	0.1709	0.802639	No	-1.5856	0.0123261	Yes
Nickelgg	4.93124	0.623023	0.00942157	Yes	-1.29631	0.00325068	Yes
Potassiumgg	5.49503	0.216644	0.102951	No	0.277441	0.295508	No
TOC	3.46945e-17	1	1.21305e-238	Yes	-1.21431e-17	nan	No
Uraniumgg	-0.148805	-0.258327	0.583832	No	-0.265572	0.837223	No
Vanadiumgg	3.14419	0.0402233	0.957978	No	-0.216349	0.689682	No
Zinc	1.93755	0.443856	0.000763766	Yes	-0.199876	0.479144	No

Figure 2.5: Indicating OM and size correction

On the top combo box, you can navigate between various source groups. In the column titled P-value for TOC and the column titled P-value for D50, the p-values for the regression coefficients for organic matter and size are shown, respectively. In column titled Correct based on TOC and column titled Correct based on D50, you can indicate whether to apply the corrections for particular elements based on organic matter and size. The items with a p-value of less than 0.05 are automatically selected to be applied.

To view how the correction affects the elemental profiles of a particular source group based on a specific target sample from the tools panel, choose **Tools→ 2 Organic Matter and Particle Size Correction→ 2a Corrected results (after size and organic correction)**. In the form in the middle panel, choose the target sample on which you want the correction done. The results window that appears will contain panels for each of the source groups. To see the corrected elemental profiles of any source groups, click on the table button  on the side of the panel (Fig. 2.6).



Elemental Profiles for Bank									
Export to CSV									
	B1	B10	B11	B12	B13	B14	B15	B16	E
210PbXs	0.640359	0.958083	4.42903	1.81011	1.46273	0.230923	0.488236	0.0703237	1.236
Aluminum	5750.75	5310.81	7970.59	3681.55	10700.6	8840.75	6031.09	6570.71	7390.
Arsenic	0.8	2	2	1.5	2.8	7.4	2.8	2.1	2.9
Barium	81.9687	70.1245	92.2217	51.9948	125.667	108.77	86.7863	83.7298	73.17
Beryllium	1.65405	1.51294	1.80039	2.03948	2.34749	1.9553	2.39294	1.61313	1.461
C13	-25.76	-27.6124	-23.678	-27.1184	-24.7294	-27.0677	-28.3904	-25.1863	-26.44
Calcium  g/g	74000	77900	3060	88200	3920	17900	32200	49900	2900
Chromium	8.3	5.1	9.3	6.5	11.9	15.3	12.6	9.3	9.3
Cobalt	12.7189	10.0084	26.7158	6.56913	17.8496	27.1153	11.225	17.8119	16.25
Copper	8.7	10.2	14.2	6.2	19.4	22.6	14	9.9	11.7
Cs137	1.83502	1.60454	2.88392	1.00274	2.46738	1.82133	0.900943	2.0205	1.847
D50	8.59	9.595	6.143	24.551	6.865	8.611	14.831	7.913	8.717
Iron	14800.7	13800.8	23300.5	8931.82	26600.5	29400.7	16501.1	15600.6	19500

Figure 2.6: Corrected elemental profiles based on size and organic matter

You can export the corrected elemental profiles to a CSV file to compare them with the original values.

2.3 Bracketing Analysis

A bracketing analysis will check to see if there are elements in a target sample whose value is higher than all of the source samples. If the elemental content of a particular element in a target sample is higher than all source samples, no combination of source samples can reproduce an elemental profile close to the target sample. To perform bracketing analysis from the tools window, go to **Tools→Steps in Sediment Fingerprinting→3-Bracketing Analysis**. Select the target sample to be evaluated and press the Ok button. The results window indicates which elements passed and which did not pass the bracketing analysis. To see the results in tabular form or to export the result to a CSV file, click on the table button . To see the bracketing analysis result on all target samples from the top menu, choose **Settings→Include/Exclude samples**. Choose the target group from the top drop-down menu at the top of the central form. The detailed results from the bracketing test indicates which element(s) violated the criteria for each sample and will be shown in the column titled "Notes" (Figure 2.7).

Mixture		
Sample	Include	Notes
CTAIL1	Yes	
CTAIL10	Yes	
CTAIL11	Yes	
CTAIL12	Yes	
CTAIL13	Yes	
CTAIL14	Yes	
CTAIL15	Yes	
CTAIL16	Yes	
CTAIL17	Yes	
CTAIL2	Yes	
CTAIL3	Yes	
CTAIL4	Yes	
CTAIL5	Yes	
CTAIL6	Yes	
CTAIL7	Yes	
CTAIL8	Yes	
CTAIL9	Yes	

Figure 2.7: Bracketing test results

2.3.1 Stepwise Discriminant Function Analysis (DFA)

Discriminant function analysis (DFA) is a statistical technique used to predict the group membership of individuals or objects based on a set of predictor variables. It is a multivariate method that aims to find a linear combination of variables that maximally discriminate between two or more groups.

The primary goal of the DFA analysis is to determine a discriminant function that can effectively separate observations from different groups by reducing within-group variability and increasing between-group variability. In the context of sediment fingerprinting, DFA can be used to determine which elements most effectively differentiate between two or more sources. The linear discriminant function provides a scalar value which is calculated as a linear combination of the elemental contents for each sample:

$$D(\tilde{\mathbf{y}}_k) = \mathbf{w}^T \tilde{\mathbf{y}}_k + c \quad (2.3.7)$$

where $D(\tilde{\mathbf{y}}_k)$ is the discriminant score of sample k , \mathbf{w} is a vector of weights, $\tilde{\mathbf{y}}_k$ is the vector containing the elemental profile of a source sample k , and c is a constant.

SedSat3 uses Fisher's linear discriminant method [Fisher, 1936]. The discriminant functions are found by solving an optimization problem that maximizes the Fisher criterion, which is the ratio of between-class variance to within-class variance. Mathematically, this is expressed as:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \quad (2.3.8)$$

where S_b is the between-class scatter matrix, which measures the variance between the means of the groups, and S_w is the within-class scatter matrix, which measures the variance within groups.

This criterion is maximized by finding the eigenvectors of the matrix $S_w^{-1} S_b$. The eigenvectors corresponding to the largest eigenvalues are the directions (weights) that define the discriminant functions.

$$S_w^{-1} S_b \mathbf{w} = \lambda \mathbf{w} \quad (2.3.9)$$

The weight vector is obtained as the eigenvector corresponding to the eigenvalue with the largest absolute value.

Wilks' Lambda Λ measures the ratio of within-group variability to total variability. The formula is:

$$\Lambda = \frac{|S_w|}{|S_b + S_w|} \quad (2.3.10)$$

The χ^2 can be calculated as:

$$\chi^2 = - \left(n - 1 - \frac{p + g}{2} \right) \ln(\Lambda) \quad (2.3.11)$$

with the degree of freedom of $df = p \times (g - 1)$.

where n is the total number of samples in all source groups, p is the total number of elements, and g is the number of groups.

An F-test can also be used to test the hypothesis that the centroids of discriminant scores (Eq. 2.3.7) of each group is the same.

2.4 Multi-way discriminant function analysis

Multi-way DFA performs the discriminant analysis between all source groups holistically. It provides the test-statistics based on the weight vectors that will maximize the separation between all source groups.

The results are similar to those for two-way and one-vs-the-rest stepwise discriminant function analysis (located in the 'Other Statistical Tools' menu), but here the statistics are calculated between the source group indicated and all other source samples. Figure 2.8 shows the χ^2 p-values. Significant tracers are selected one at time, beginning from the highest Chi-Squared p-value (left side of Fig. 4.7) to the lowest Chi-Squared p-value. The graph indicates this as the tracers selection from C13 to N (fig. 4.7).

To see the separation as a result of selecting a certain set of elements, select **Tools→Miscellaneous Analysis tools→Discriminant Function Analysis (Mutliway)**. Check "Use only selected elements". Also indicate whether the

Box-Cox transformation is to be conducted and whether the OM and size correction is to be performed before the analysis. Click on the "Ok" button. The result window contains three panels. The first panel shows the χ^2 p-values for separation between the samples in each source from all the rest of the samples. A higher p-value indicates a lower separation between the source group and the rest. Select the graph icon in the top panel to display the graph as shown in Figure 4.7; which shows the significant tracers.

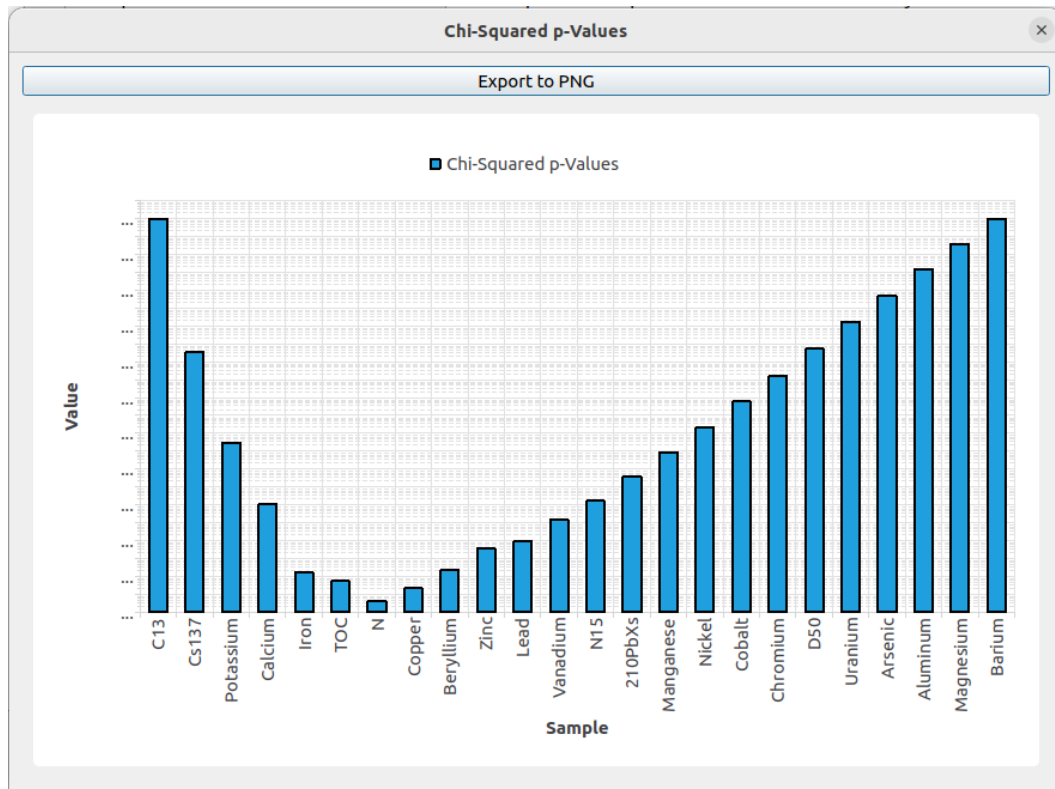




Figure 2.8: χ^2 p-values vs. the number of elements included for multi-way DFA

Click on the table button  on the side of any panel to see the values in tabular format (Fig. 2.9a). For example Fig. 2.9a shows a weaker separation between Pasture samples and the rest of the samples compared to other source groups.

The second panel contains the F-test p-values. Click on the table button  on the side of the panel to see the values in tabular format (Fig. 2.9b). As it can be seen, the F-test p-values also indicate a weaker separation between pasture samples and other samples in this example.

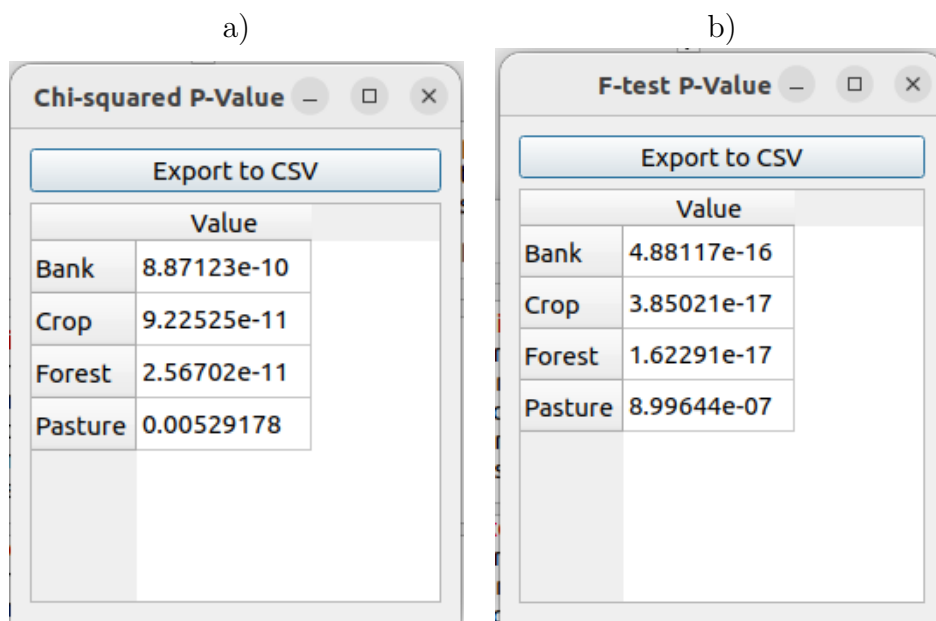



Figure 2.9: a) χ^2 p-values between each source group and the rest of samples for each source group b) F-test p-values between each source group and the rest of samples for each source group when Beryllium, ^{13}C , ^{137}Cs , Manganese, N, Nickle, and Uranium are included in the analysis

Panel three contains the discriminant scores based on the eigenvectors obtained for each source group. Click on the graph button  on the side of the panel. Above the graph window that appears, select the two source groups the plot is intended to be shown for (Fig. 2.10). The x-values show the discriminant scores based on the eigenvector maximizing the separation between the first selected source group and the rest of the samples while the y-values show the discriminant scores based on the eigenvectors obtained from the second selected source group.

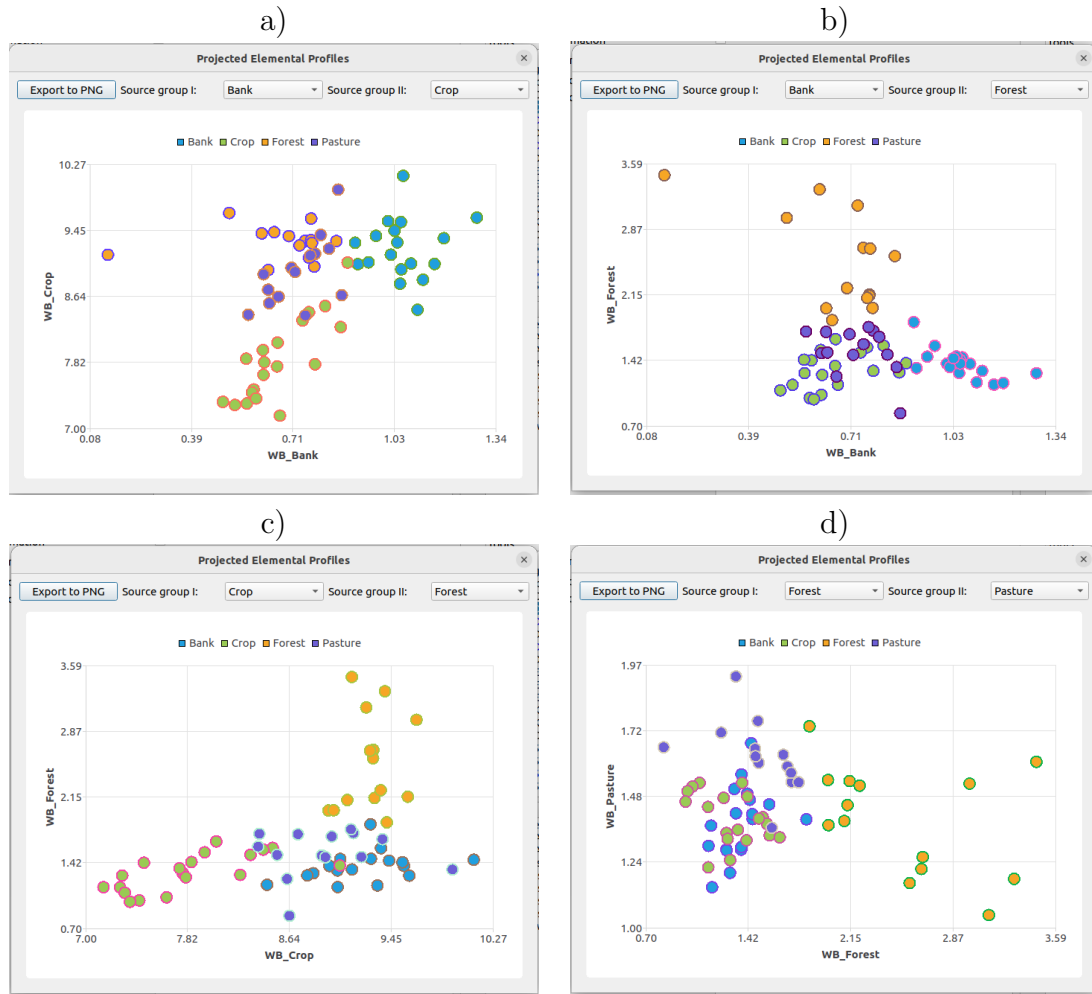
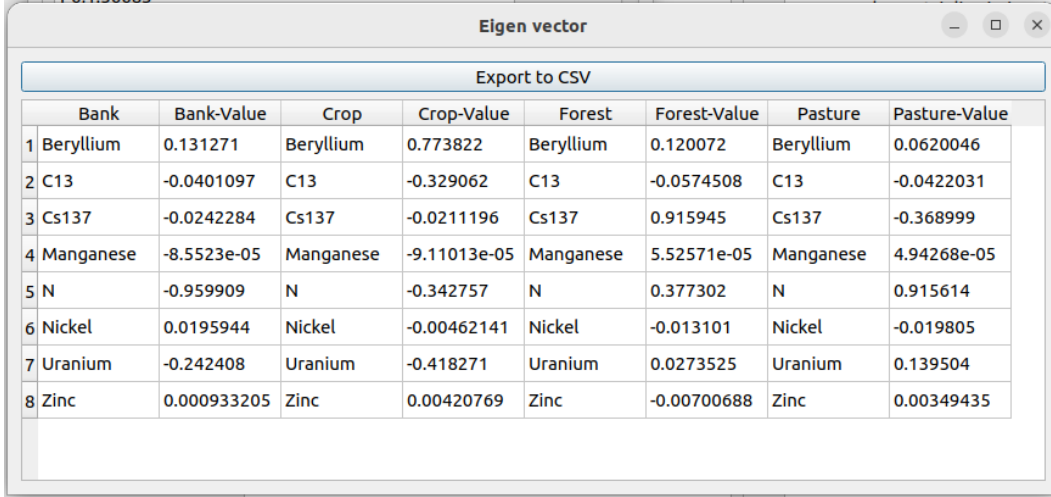


Figure 2.10: Scatter plots showing discriminant scores based on the eigenvectors obtained from a) Bank and Crop, b) Bank and Forest, c) Crop and Forest, and d) Forest and Pasture

Finally, the forth panel contains the eigenvectors obtained based on maximizing the separation between each source groups and the rest. Click on the table button on the side of the panel to see the eigenvectors in tabular form (Fig. 2.11).



	Bank	Bank-Value	Crop	Crop-Value	Forest	Forest-Value	Pasture	Pasture-Value
1	Beryllium	0.131271	Beryllium	0.773822	Beryllium	0.120072	Beryllium	0.0620046
2	C13	-0.0401097	C13	-0.329062	C13	-0.0574508	C13	-0.0422031
3	Cs137	-0.0242284	Cs137	-0.0211196	Cs137	0.915945	Cs137	-0.368999
4	Manganese	-8.5523e-05	Manganese	-9.11013e-05	Manganese	5.52571e-05	Manganese	4.94268e-05
5	N	-0.959909	N	-0.342757	N	0.377302	N	0.915614
6	Nickel	0.0195944	Nickel	-0.00462141	Nickel	-0.013101	Nickel	-0.019805
7	Uranium	-0.242408	Uranium	-0.418271	Uranium	0.0273525	Uranium	0.139504
8	Zinc	0.000933205	Zinc	0.00420769	Zinc	-0.00700688	Zinc	0.00349435

Figure 2.11: Estimated eigenvectors based on maximizing the separation of samples in each source with all other samples.

2.5 Fingerprinting

The Levenberge-Marquardt Maximum likelihood estimation (LMMLE) is the simplest method implemented in SedSAT3. LMMLE uses the Levenberge-Marquardt optimization method to maximize the log-likelihood in Eq. (5.2.11). The transformation function g is assumed to be log-normal, and the scale factor σ of the log-normal distribution is assumed to be the same for all the elements.

Modeled target sample elemental profile, \mathbf{C} is calculated based on the arithmetic average of the elemental profiles of all samples in each source group (i.e., Eq. (5.2.8)).

To perform LMMLE from the tools window, select **Tools→Fingerprinting tools→Levenberg-Marquardt optimization**.

Checking the "Apply size and organic correction" will perform size and organic correction on the source samples if size and organic correction is done previously. It will use the latest size and correction model selected. The size and correction will be applied to the elements, size, and/or organic matter based on what is indicated in **Settings→Organic matter/Size correction**.


If no organic matter/size correction has been performed on the data, an error message will be shown asking it to be performed. You have the option to check off the OM and size correction and proceed or to perform OM and size correction.

Choose the target sample you would like to be analyzed using the drop-down menu labeled Target Sample. The soft-max transformation maps the elemental contributions from $[0, 1]$ domain to $[-\infty, \infty]$ domain through the Softmax transformation:

$$x_i = \frac{e^{\zeta_i}}{\sum_{i=1}^n e^{\zeta_i}} \quad (2.5.12)$$

And the optimization is done for ζ values rather than x values. Using the

Softmax transformation implicitly imposes the constraint of the summation of contributions being 1.0 (Eq. (5.1.1)), and the search in the infinite domain is more efficient than the search in $[0, 1]$ domain.

Click the Ok button at the bottom of the form. The progress window will show the Mean Squared Error (MSE) value as the Levenberge-Marquardt iterations proceed. After the iterations converge, the result window will appear. The result window contains four panels. The first panel contains the inferred contributions of the source groups. You can click on the graph button  on the side of the panel to see the results as a pie chart (Figure 2.12).

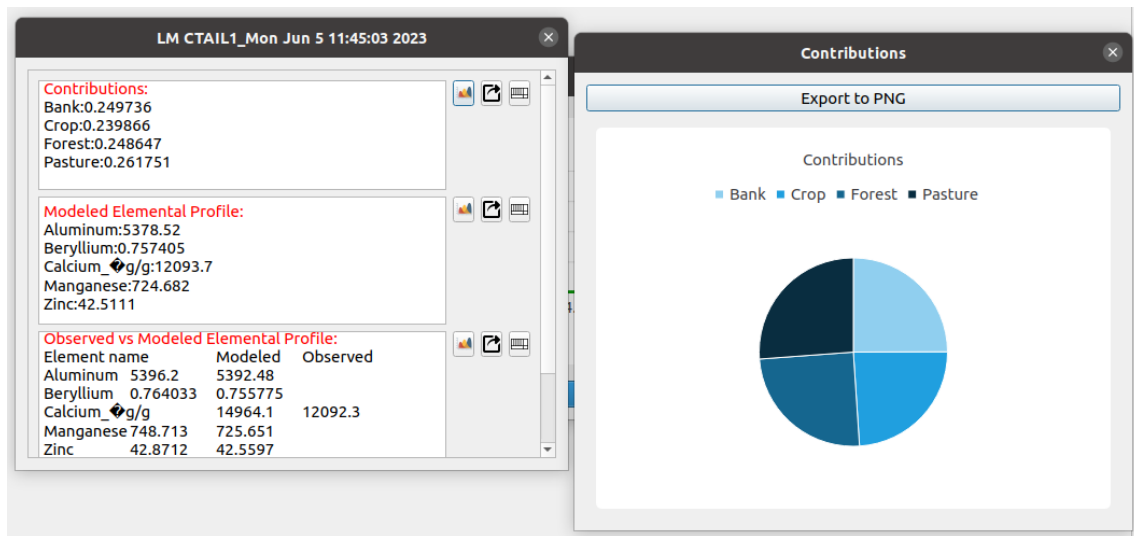



Figure 2.12: Source contributions obtained from the LMMLE method

The second panel shows the modeled target sample elemental profiles, **C**. The third panel contains modeled and observed elemental profiles. We can use this to evaluate how close the elemental profile as a result of the mixing of the source groups is to the observed sample's elemental profile. Click on the graph button  on the side of the panel to see the matching between modeled and observed elemental profiles graphically (Figure 2.13).

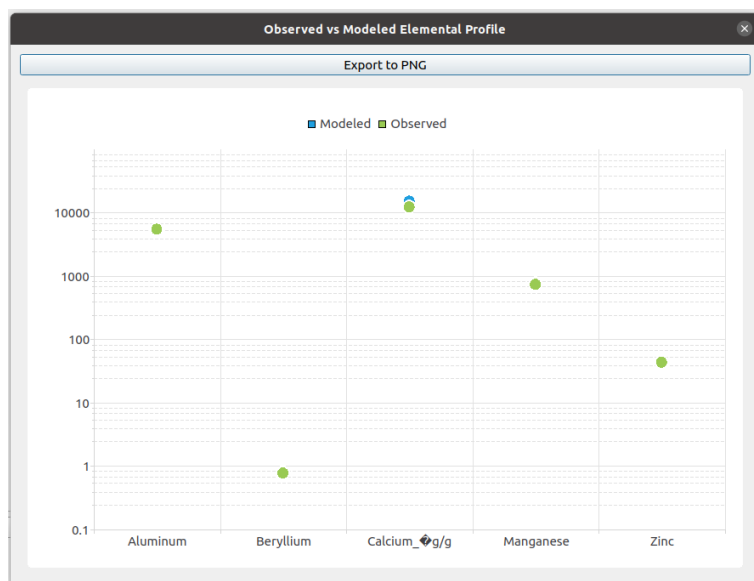



Figure 2.13: Modeled vs. observed Elemental profiles

And finally, the fourth panel contains the modeled vs. measured results for the isotope ratios in case constituents contain some isotopes.

2.6 Fingerprinting (Batch)

This tool performs the Levenberge-Marquardt maximum likelihood fingerprinting for all target samples simultaneously. To perform batch Levenberge-Marquardt fingerprinting, select **Tools**→**Fingerprinting tools**→**Levenberg-Marquardt optimization (Batch)**. Select whether you want the OM and size correction to be applied to the source samples and whether the soft-max transformation is to be used. Click on the "OK" button. Wait until the process is completed. The batch version of the LM fingerprinting provides fewer details than when performing the analysis on each target sample individually. The resulting panel only shows the estimated source contribution for each target sample. Click on the graph button  to see the contribution of each source to each target sample graphically (Fig. 2.14) and the table button to see the results in a tabular form (Fig. 2.15).

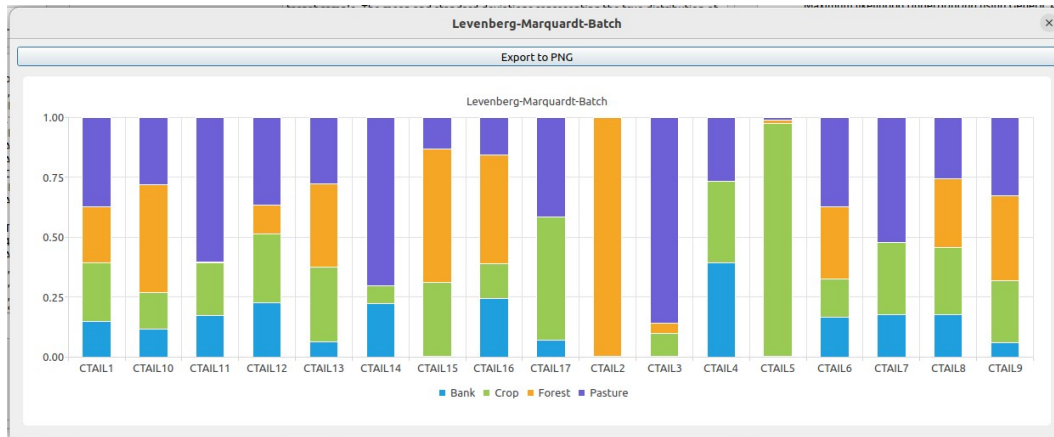


Figure 2.14: The graph showing the Levenberg-Marquardt batch results

Levenberg-Marquardt-Batch				
Export to CSV				
	Bank	Crop	Forest	Pasture
CTAIL1	0.148043	0.245241	0.232533	0.374182
CTAIL10	0.114002	0.153651	0.450558	0.281788
CTAIL11	0.171622	0.22199	9.2401e-05	0.606295
CTAIL12	0.224481	0.289088	0.119733	0.366698
CTAIL13	0.0629001	0.309733	0.348542	0.278825
CTAIL14	0.221136	0.0732738	0.000370614	0.70522
CTAIL15	0.000135971	0.311794	0.554996	0.133073
CTAIL16	0.243109	0.143723	0.456175	0.156992
CTAIL17	0.0701931	0.511904	5.6232e-06	0.417897
CTAIL2	0.000448922	0.000380801	0.998582	0.000588173
CTAIL3	0.000948134	0.0976136	0.0410428	0.860395
CTAIL4	0.391512	0.33904	4.84759e-05	0.2694
CTAIL5	0.000374446	0.971695	0.0172017	0.0107287
CTAIL6	0.166219	0.157171	0.302857	0.373753
CTAIL7	0.175009	0.300933	0.000227389	0.52383
CTAIL8	0.176907	0.279217	0.285309	0.258567
CTAIL9	0.0601827	0.255807	0.356086	0.327924

Figure 2.15: Levenberg-Marquardt batch results in tabular format

Chapter 3

Bayesian Sediment Fingerprinting

This chapter covers the Bayesian fingerprinting method implemented in *SedSat*. The pre-analysis steps required for Bayesian fingerprinting—namely, *Outlier Analysis*, *Organic Matter and Particle Size Correction*, *Bracketing Analysis*, and *Step-wise Discriminant Function Analysis*—were presented in the previous chapter. The focus here is on the Bayesian sediment fingerprinting procedure, which produces the final results of the analysis.

3.1 Bayesian Chemical Mass Balance Analysis

As it was mentioned in section 5.3, there are various sources of uncertainty in sediment fingerprinting. These include but are not limited to measurement errors, lack of representativeness of source samples, presence of unidentified sources, or equifinality. The Bayesian approach provides a probability distribution for the source contributions rather than providing point estimates based on the maximum likelihood estimation.

To perform Bayesian fingerprinting from the tools window, select **Tools**→**Bayesian Sediment Finger Printing**→**Bayesian Chemical Mass Balance Analysis**.

Below are the descriptions of the setting items appearing in the central form:

- **Apply size and organic matter correction:** indicates whether the analysis is performed on the corrected elemental profiles based on size and organic matter. Note that in order to size and organic correction to be done, a size and organic matter correction step (section 2.2) should have been performed before.
- **Dissolve Chains:** This option indicates that during the Markov chain Monte Carlo (MCMC) sampling, chains with significantly lower posterior probability will be discarded and replaced by chains with higher posterior probability. This will improve the convergence speed of the algorithm.
- **Number of Chains:** The number of MCMC chains to be used in parallel.
- **Number of Samples:** Indicates the total number of samples to be generated. For a full convergence, typically, 100,000 samples or more are necessary.

- **Sample:** The label of the target sample to be analyzed.
- **Samples file name:** Indicate the text file where MCMCM samples information will be saved in.
- **Samples to be discarded (burnout):** Indicate the number of samples to be discarded from the beginning of MCMC chains. If the total number of samples is 100000, it is recommended that the burnout be specified as 20000.

After setting the parameters, click on the Ok button to start the analysis. The analysis duration depends on the number of samples indicated to be produced.

The progress window contains some information that can shed light on suitable parameters for MCMC sampling.

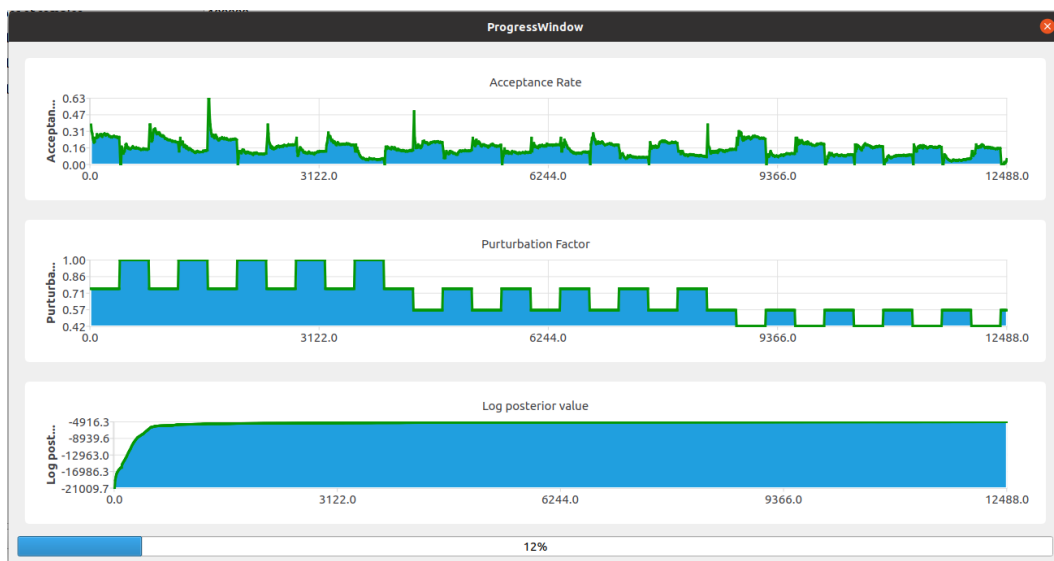



Figure 3.1: Progress window of Bayesian chemical mass balance analysis

The top panel shows the average acceptance rate (i.e., the fraction of accepted new samples). The perturbation factor (panel 2) is adjusted to maintain an acceptance rate of around 15%. The bottom panel shows the maximum numerator of the posterior distribution (Eq. (5.3.20)). Initially, the values are lower when the samples are far from the maximum likelihood estimates. As the algorithm proceeds, the values increase and eventually are stabilized. The number of discarded samples can be indicated based on the number of samples that it will take for the value shown on this chart to stabilize.

After the MCMC sampling is finished, a result window will pop up. The result window contains five panels.

The first panel contains MCMC samples for each parameter, including source contributions, μ , and σ values of log-normal distributions considered for the elemental content of each source group and the error standard deviations (Figure 3.2). Click on the graph button  to see the visualization of the samples (Figure 3.2). Use the drop-down menu at the top to see the results for different parameters. Note

the convergence of the results. For example, based on figure 3.2, it seems that the algorithm takes around 25000 samples to converge as indicated by the bank and crop contribution parameter.

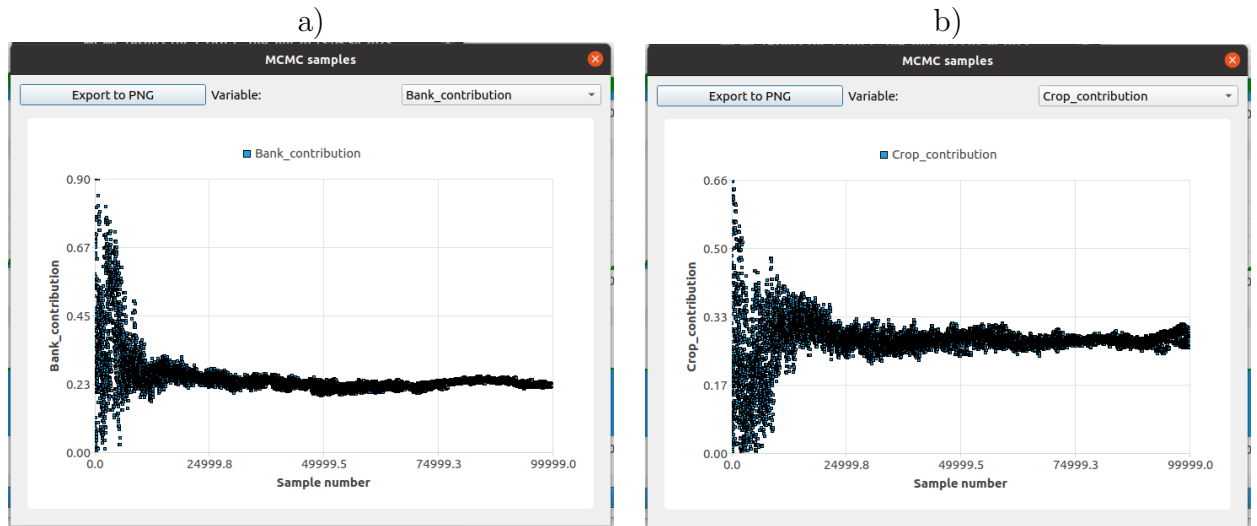



Figure 3.2: Visualization of MCMC samples a) bank contribution samples. b) crop contribution samples

Panel two can be used to visualize the posterior distribution based on the samples drawn. Click on the graph button  to see the posterior distributions for all the parameters.

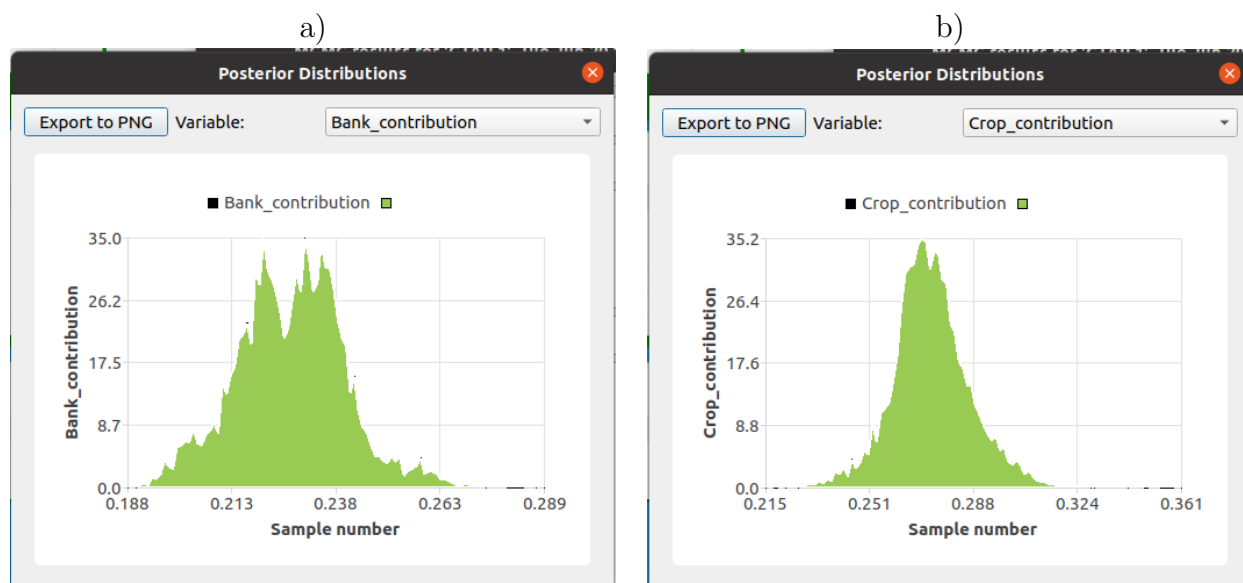




Figure 3.3: Visualization of the posterior distribution for a) bank contribution and b) crop contribution

The third panel contains the 95% credible intervals of source contributions. The credible interval is the range the source contribution falls in with 95% probability (Figure 3.4). In addition to the low and high ranges of the credible intervals, the table also shows the expected value (mean) and median from the posterior distributions.

Click on the table button  on the side of the panel to see the range in tabular form. Click on the graph button  to see the credible intervals graphically.

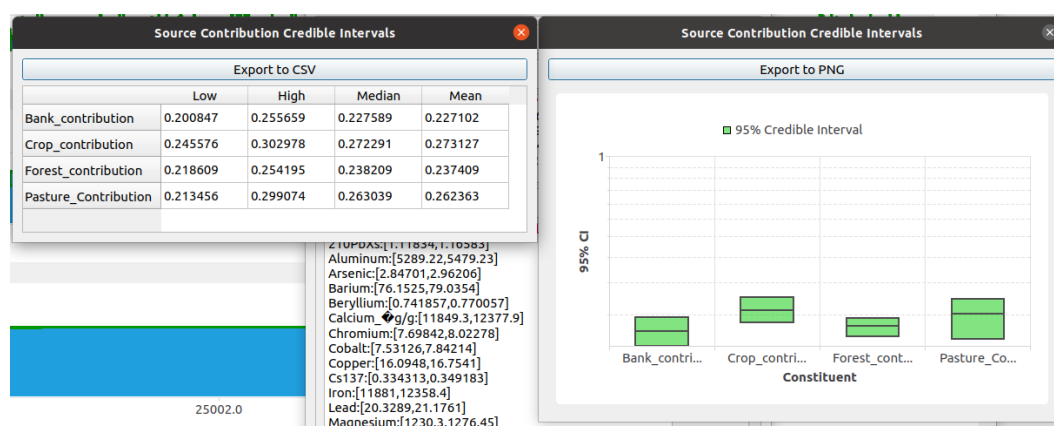



Figure 3.4: Credible intervals for source contributions

The fourth panel in the Bayesian analysis result window contains the predicted (posterior) distribution of element contents in the target sample. Click on the graph button  to see the distributions graphically. Use the drop-down menu at the top of the graph to see the predicted distribution for each element (Figure 3.5). The distribution is obtained by calculating the target elemental profile for each MCMC

sample. The vertical dashed line in the graph shows the measured concentration in the target sample.

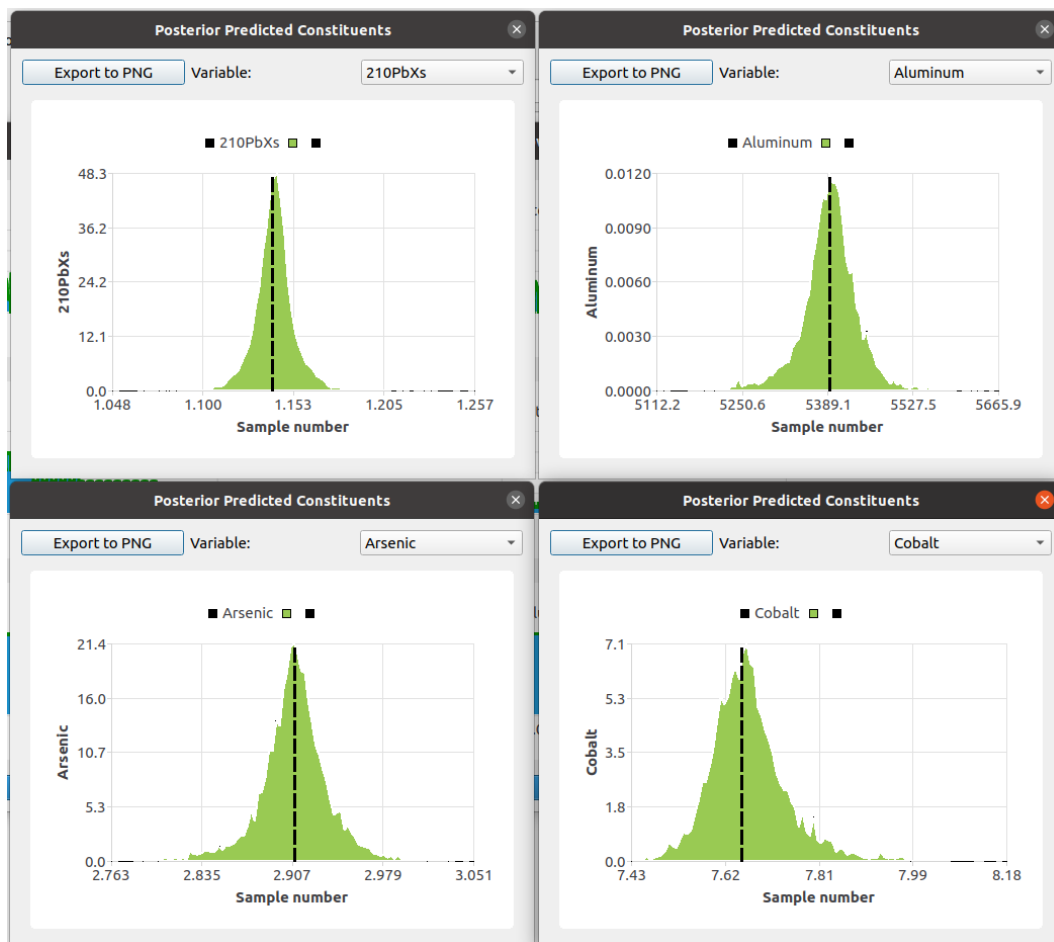




Figure 3.5: Posterior distribution of elemental contents in the target sample

The 95% credible intervals of predicted elemental contents in the target sample can be found in panel five of the result window. Click on the table button  on the side of the panel to see the range in tabular form. Click on the graph button  to see the credible intervals graphically. The graph will also include the measured elemental content in the target sample for each element with dot symbols. The difference between the 95% bracket and the measured value will give us an insight into how well each elemental content can be captured based on the combination of the sources.

Panels six and seven contain similar information to panels four and five but for isotopes if isotopes are included in the analysis.

3.2 Bayesian Chemical Mass Balance Analysis (Batch)

The batch version of the Bayesian Chemical Mass Balance tool conducts Bayesian fingerprinting on all target samples, saving users from selecting samples one by

one and manually saving the analysis. Instead of displaying detailed results as tables and graphs, the analysis outputs are saved in text files that can be easily plotted using standard spreadsheet programs or graphing software applications. The settings are the same as described for the Bayesian Chemical Mass analysis tool.

To perform batch Bayesian chemical mass balance analysis, from the tools window, select **Tools**→**Bayesian Sediment Fingerprinting**→**Bayesian Chemical Mass Balance Analysis (Batch)**.

Note that a large number of samples ($>100,000$) is required for the results to converge, and approximately 10% of the total samples should be designated as burn-in. The analysis can take considerable time, depending on the number of target samples, source groups, and MCMC samples, as well as your hardware's processing power.

After the analysis is complete, folders named after your target samples will be created in the same directory where the project is saved (Figure 3.6). Each folder will contain seven text files (Figure 3.7).

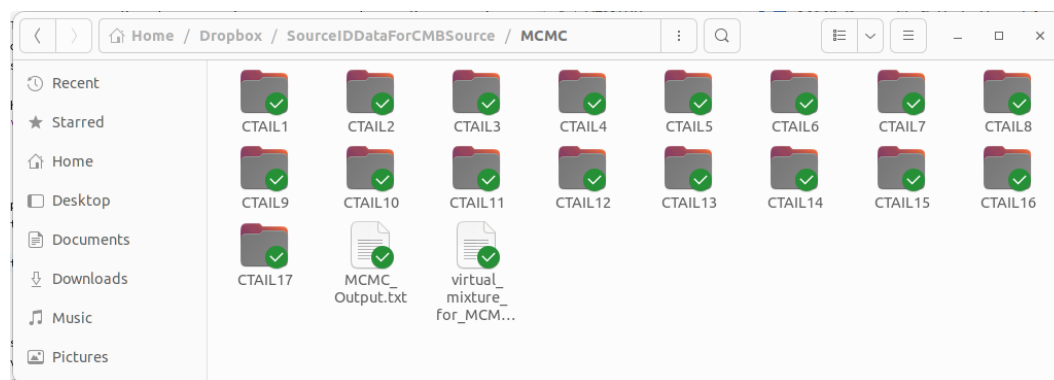


Figure 3.6: The folders created after performing batch Bayesian chemical mass balance modeling

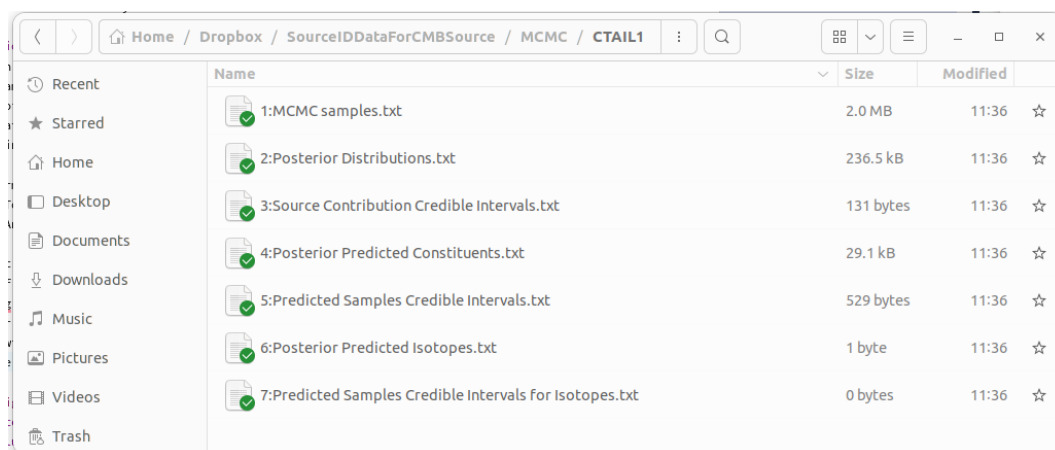


Figure 3.7: Text files containing the batch Bayesian chemical mass balance modeling results

The files include:

- **1:MCMC samples.txt:** This file contains all the MCMC samples generated. The columns represent the estimated parameters, including the source contribution values for each source, the mean and standard deviation of each element in each source group, and the likelihood values for each parameter set.
- **2:Posterior Distributions.txt:** This file contains the MCMC samples converted into frequency distributions. The columns come in pairs, the first one containing the values of the parameters and the second containing the value of the frequency distributions. For example, in order to plot the posterior distribution of "Bank Contribution", column B should be plotted vs. column A and so forth.
- **3:Source Contribution Credible Intervals.txt:** This file contains the 95% credible intervals of each parameter extracted from the posterior distribution (Figure 3.8).

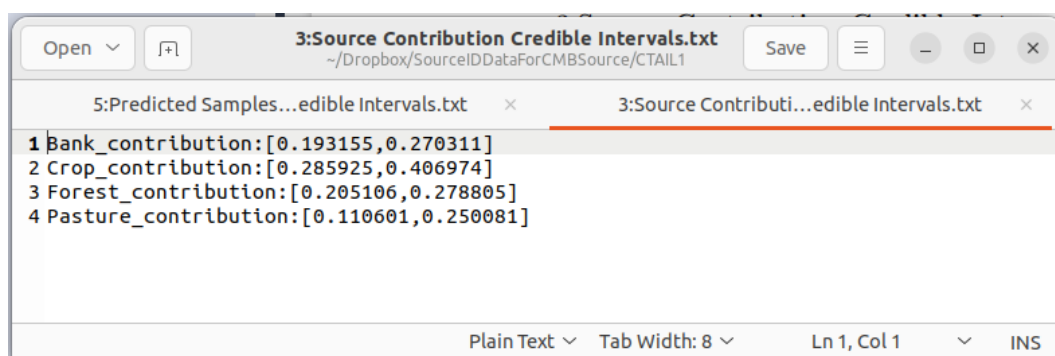


Figure 3.8: Text files containing 95% credible intervals of model parameters

- **4:Posterior Predicted Constituents.txt:** This file contains predicted target sample concentrations based on the drawn MCMC samples converted into frequency distributions. The columns come in pairs, the first one containing the values of predicted elemental content and the second containing the value of the frequency distributions. For example, in order to plot the frequency distribution of ^{210}Pb , column B should be plotted against column A.
- **5:Predicted Samples Credible Intervals.txt:** This file contains the 95% credible intervals of the predicted target sample elemental profile based on the parameters extracted from Posterior Predicted Constituents (Figure 3.9).

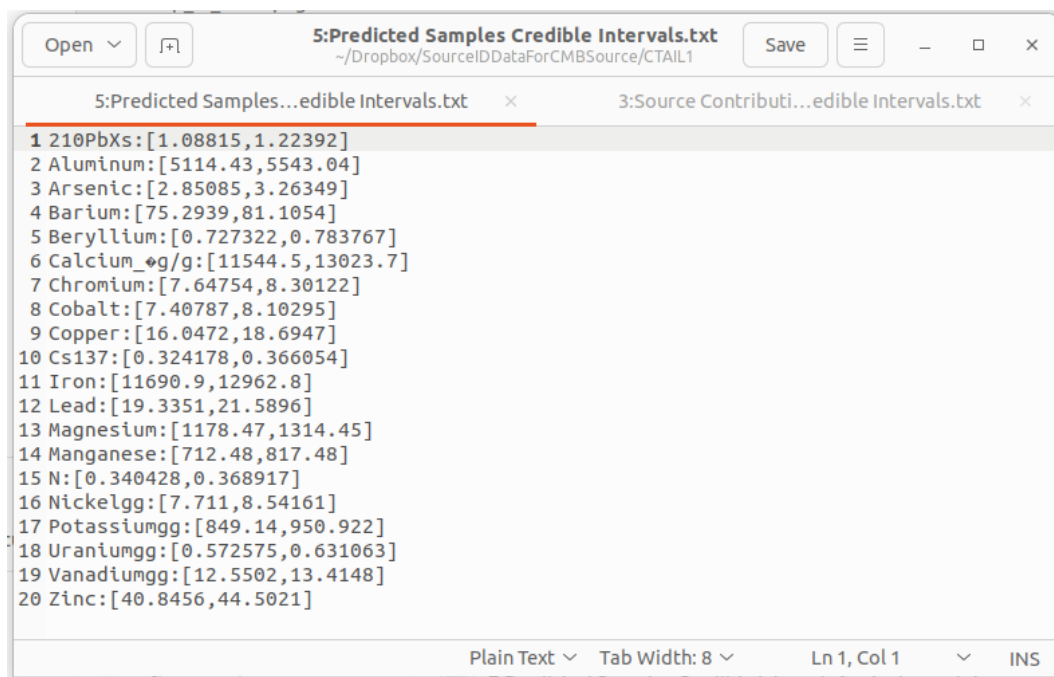


Figure 3.9: Text files containing 95% credible intervals of predicted target sample elemental content

- **6:Posterior Predicted Isotopes.txt:** This file contains predicted target isotope ratios based on the drawn MCMC samples converted into frequency distributions. The columns come in pairs, the first one containing the values of predicted isotope ratios and the second containing the value of the frequency distributions. This file will be empty if your analysis contains no isotopes.
- **7:Predicted Samples Credible Intervals for Isotopes.txt:** This file contains the 95% credible intervals of the predicted target sample isotope ratios based on the parameters extracted from Posterior Predicted Isotopes.

Chapter 4

Other Statistical Tools

The miscellaneous analysis tools provide some non-essential statistical tools for analyzing fingerprinting data. Below, each tool is described.



4.1 Elements' correlation matrix

Knowing the correlation between constituents used for fingerprinting within source groups is useful information in deciding which constituents are to be used in the fingerprinting analysis.

The value of the correlation between each two elements j and j' in a source group i is calculated as:

$$r_{j,j'} = \frac{\sum_k (y_{j,k} - \bar{y}_j)(y_{j',k} - \bar{y}_{j'})}{\sqrt{\sum_k (y_{j,k} - \bar{y}_j)^2 \sum_k (y_{j',k} - \bar{y}_{j'})^2}} \quad (4.1.1)$$

where $y_{j,k}$ and $y_{j',k}$ are respectively the measured elemental contents of elements j and j' in sample k in the source group, and \bar{y}_j and $\bar{y}_{j'}$ are the average elemental contents of elements j and j' respectively in the source group.

From the "Tools" menu in the far right window panel, double-click on **Pre-Analysis tools**→**Elements' correlation matrix**. Choose the source/target group for which you want the correlation matrix calculated from the drop-down menu labeled Source/Target group. The text box labeled threshold indicates the threshold above which the values will be highlighted. If you check the *Use only selected elements* item, the correlation matrix will be generated only for the elements indicated to be included in the analysis in **Settings**→**Constituent properties**. Also, checking the *Use only selected samples* checkbox only includes the samples that are indicated to be included in the analysis in **Settings**→**Include/Exclude samples**. Click the Ok button at the bottom of the form, and a window, like in Figure 4.1, will appear. Click on  button to see the matrix in a tabular form (Fig. 4.2). You can export the results to a .csv file by clicking on . At this point, based on the correlation matrix, the user can make a decision on which tracers to remove from further analysis. To remove a tracer go to settings in the upper left hand

corner of the screen, click on Constituent Properties and double click on "include in analysis" the element you want to remove. To conduct the correlation analysis on the source elemental profiles adjusted for size, organic matter, or both, select the target sample on which you would like the correction to be based. Under the results window panel located beneath the Tools panel, your results will be stored for you. You can go back and re-review any.

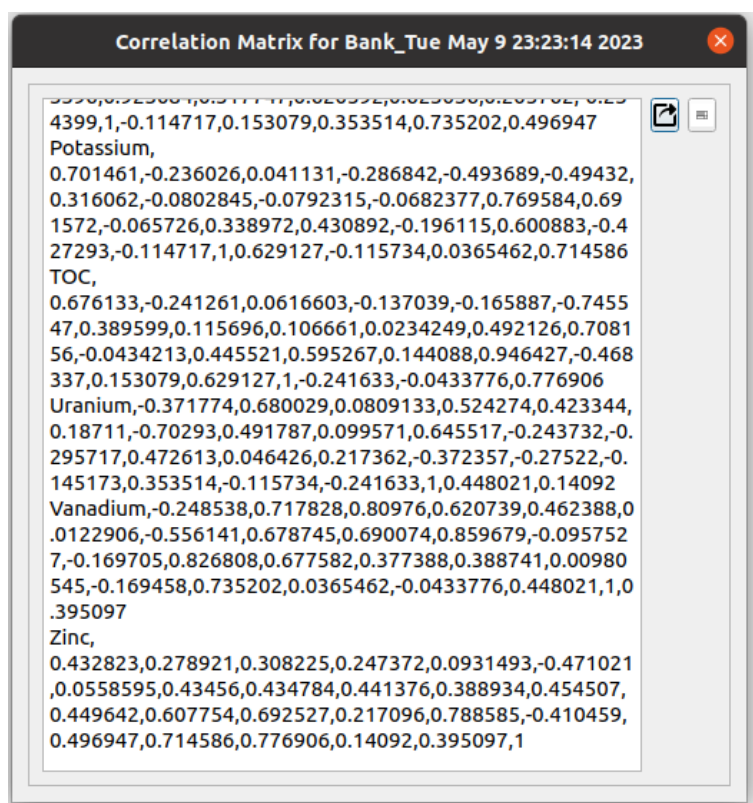


Figure 4.1: Correlation matrix results

	210PbXs	Aluminum	Arsenic	Barium	Beryllium	C13	Calcium
210PbXs	1	-0.539165	-0.232535	-0.447917	-0.370929	-0.420819	0.537695
Aluminum	-0.539165	1	0.406129	0.876982	0.710468	0.374584	-0.617548
Arsenic	-0.232535	0.406129	1	0.360564	0.278942	-0.273451	-0.400609
Barium	-0.447917	0.876982	0.360564	1	0.797212	0.334355	-0.294421
Beryllium	-0.370929	0.710468	0.278942	0.797212	1	0.420961	-0.424384
C13	-0.420819	0.374584	-0.273451	0.334355	0.420961	1	-0.190294
Calcium	0.537695	-0.617548	-0.400609	-0.294421	-0.424384	-0.190294	1
Chromium	-0.354381	0.743453	0.512155	0.637467	0.517518	-0.0525882	-0.491529
Cobalt	-0.323661	0.678492	0.659093	0.649616	0.570985	0.0903911	-0.292219
Copper	-0.386759	0.879143	0.585519	0.814688	0.564344	0.0117585	-0.533372
Cs137	0.855098	-0.425168	-0.127077	-0.47392	-0.487549	-0.319978	0.367844
D50	0.830231	-0.514183	-0.071636	-0.476988	-0.427649	-0.549663	0.430745
Iron	-0.416049	0.927146	0.576243	0.814365	0.62141	0.235477	-0.514117
Lead	0.0697277	0.297079	0.785376	0.154808	0.0376274	-0.433445	-0.306377
Magnesium	0.321382	0.335649	0.144768	0.382361	0.0606308	-0.319581	0.160936
Manganese	-0.227035	0.328713	0.581292	0.477441	0.409513	0.00380118	0.135985

Figure 4.2: Correlation matrix table

Based on the correlation analysis results, the user can decide if a tracer should be eliminated from further analysis. Usually, this decision is made for a tracer that has a high correlation to the same tracer(s) for all source groups. Once a decision is made the user should go to the Setting tab and under constituent properties remove that tracer.

4.2 Analysis of Variance

Analysis of Variance (ANOVA) facilitates the evaluation of the effectiveness of different elements in distinguishing between source groups by comparing the variations in elemental content both between and within these groups. This analysis aids in identifying elements that possess a greater capacity for fingerprinting by highlighting those with significant differences in their distribution across the groups.

The one-way ANOVA (Analysis of Variance) is a statistical technique used to compare means of three or more samples (assuming equal variances) to determine if at least one sample mean is significantly different from the others. It does this by analyzing the variance within each group and the variance between the groups. The formula for one-way ANOVA involves calculating the F-statistic, which is the ratio of the variance between the groups to the variance within the groups. The formula for the F-statistic in one-way ANOVA is:

$$F = \frac{MS_{between}}{MS_{within}} \quad (4.2.2)$$

where:

F is the F-statistic, $MS_{between}$ (Mean Square Between) is the variance between the groups, and MS_{within} (Mean Square Within) is the variance within the groups.

The Mean Square Between ($MS_{between}$) is calculated as:

$$MS_{between} = \frac{SS_{between}}{df_{between}} \quad (4.2.3)$$

And the Mean Square Within (MS_{within}) is calculated as:

$$MS_{within} = \frac{SS_{within}}{df_{within}} \quad (4.2.4)$$

where:

- $SS_{between}$ (Sum of Squares Between) is the sum of squares between the groups,
- $df_{between}$ (degrees of freedom between) is the number of groups minus one ($n - 1$),
- SS_{within} (Sum of Squares Within) is the sum of squares within the groups, and
- df_{within} (degrees of freedom within) is the total number of observations minus the number of groups ($l - n$).

The Sum of Squares Between ($SS_{between}$) and Sum of Squares Within SS_{within} are calculated using the following formulas:

$$SS_{between} = \sum_{i=1}^n n_i (\bar{y}_i - \bar{y})^2 \quad (4.2.5)$$

$$SS_{within} = \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad (4.2.6)$$

where:



- n_i is the number of observations in group i ,
- \bar{y}_i is the mean of group i ,
- \bar{y} is the overall mean of all observations,
- $y_{i,k}$ is the k^{th} observation in the i^{th} group, and
- n is the number of groups.

The result, F , is then used to calculate a p-Value for the element in question.

$$p = \Xi^{-1}(F, n - 1, l) \quad (4.2.7)$$

where Ξ is the cumulative F distribution with degrees of freedom of $n - 1$ and $l = \sum_{i=1}^n n_i$ (the total number of samples in all groups).

To perform analysis of variance in SedSat3, from the **Tools→Pre-Analysis Tools→Analysis of Variance**. You have the option to conduct the analysis using either log-transformed elemental contents, Box-Cox transformed elemental contents or their raw values. To conduct the correlation analysis on the source elemental profiles adjusted for size, organic matter, or both, select the target sample on which you would like the correction to be based. By selecting "Modify the included elements based on the results," elements that have p-values exceeding the specified "P-value threshold" will be automatically removed. This threshold is utilized to identify elements with a low capacity for discrimination. Upon finalizing your selections, click the "Ok" button to proceed. After the analysis is complete the result window will appear.

Click on the graph button  to see the p-Values for each element graphically. You can also see the values in tabular form by clicking on the table  button. For example, Fig. 4.3 shows that the discriminatory capability of Arsenic and Cobalt are low in the demonstration dataset.

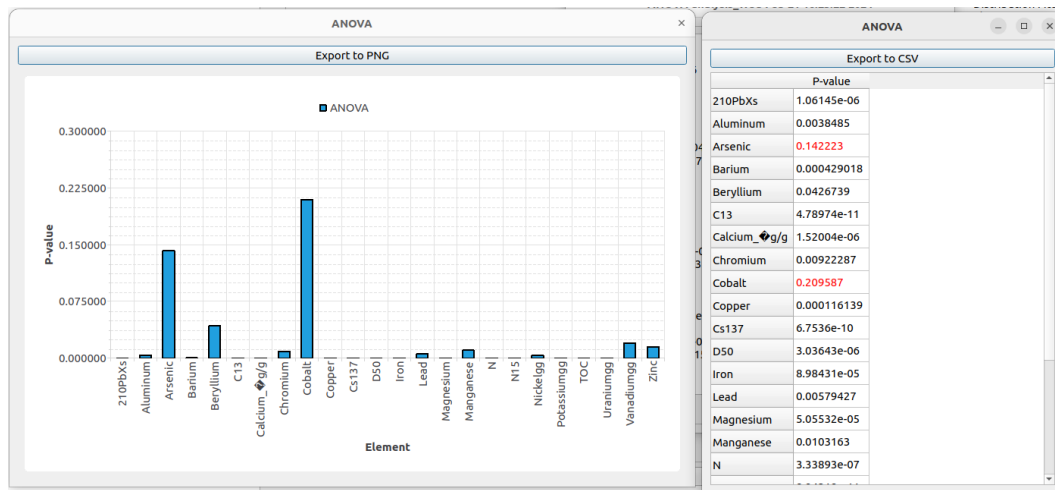


Figure 4.3: Results of the ANOVA analysis

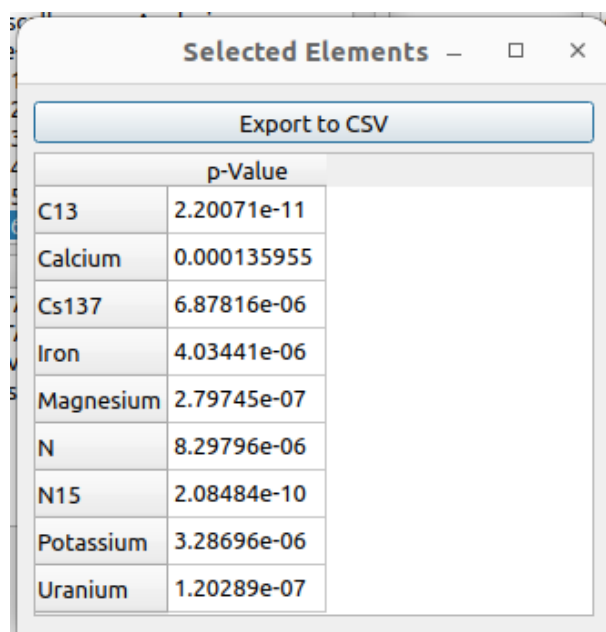
4.3 Auto-select elements

The auto-select tool identifies the minimum number of elements required to effectively distinguish between all pairs of source groups. It first performs a T-test for each combination of source groups, then selects a predetermined number of elements with the lowest p-values from each pair.

To use this tool select **Tools→Pre-Analysis Tools→Auto-select elements**. You may decide to include isotopes in the analysis and adjust the source elemental profiles based on size and organic matter. To apply the selection automatically, check the box labeled as "Modify the included elements based on the results". The value of "Number of elements from each pair" indicates the number of elements with the lowest p-value that will be selected from each pair of source groups. A larger number results in a larger total number of selected elements. Click "Ok" to perform the analysis. The result window will contain two panels. The top panel shows the p-value for each element resulting from the T-test performed on each pair of source groups (Fig. 4.4). The second panel contains the selected elements and the p-values associated with the pair of source groups based on which they are selected (Fig. 4.5).

Multi-way discriminat p-value						
Export to CSV						
	Bank and Crop	Bank and Forest	Bank and Pasture	Crop and Forest	Crop and Pasture	Forest and Pasture
210PbXs	0.60858	0.000273424	0.00596117	0.000303959	0.00173776	0.00206669
Aluminum	0.00140777	0.00120017	0.00875885	0.461871	0.753638	0.362702
Arsenic	0.546699	0.9874	0.115257	0.549437	0.192843	0.11821
Barium	0.00010552	0.107951	0.00157678	0.00298236	0.37665	0.00734065
Beryllium	0.000308693	0.139737	0.114811	0.18141	0.151759	1
C13	2.21611e-07	0.00245663	0.553333	2.20071e-11	3.38635e-07	8.76326e-05
Calcium	6.94102e-05	0.000136101	0.000135955	0.200324	0.221703	0.98646
Chromium	0.00498753	0.00247332	0.205418	0.290358	0.12963	0.0411133
Cobalt	0.077569	0.0179345	0.0684522	0.622818	0.948599	0.671852
Copper	0.00106885	0.0024112	0.142911	2.64138e-05	0.022415	0.000958089
Cs137	0.104037	6.87816e-06	0.00397332	1.29466e-05	0.00461177	5.23953e-05
Iron	0.000531894	4.03441e-06	0.013853	0.0257831	0.402353	0.011976
Lead	0.143433	0.00182232	0.233663	0.000282315	0.0829332	0.278035
Magnesium	2.79745e-07	9.1426e-05	0.0160995	0.465705	0.0501667	0.218889
Manganese	0.00504726	0.0915773	0.346981	0.00668937	0.426828	0.0449863
N	0.588552	6.38683e-05	1.17091e-05	7.77494e-05	8.29796e-06	0.129189
N15	6.60204e-05	2.87576e-06	0.00100433	2.08484e-10	0.315586	3.92535e-09
Nickel	0.000227974	0.0117775	0.00921358	0.38867	0.292505	0.939916
Potassium	0.0151933	0.120719	3.28696e-06	0.63693	0.000462496	0.0392849
Uranium	4.29723e-05	0.145725	0.0438323	1.20289e-07	0.00107108	0.000128985
Vanadium	0.0486777	0.00156114	0.767908	0.0845018	0.152998	0.0107438
Zinc	0.787344	7.08593e-05	0.44987	0.00353593	0.391914	0.00400046

Figure 4.4: Pairwise T-test for auto-selection of elements





p-Value	
C13	2.20071e-11
Calcium	0.000135955
Cs137	6.87816e-06
Iron	4.03441e-06
Magnesium	2.79745e-07
N	8.29796e-06
N15	2.08484e-10
Potassium	3.28696e-06
Uranium	1.20289e-07

Figure 4.5: Selected elements by the auto-select tool

4.4 Two-way DFA

The two-way DFA performs DFA between two given source groups. The first step in all DFA analysis is to determine the elements to be included in the analysis through evaluating the χ^2 statistics and Wilks' Lambda. The stepwise DFA starts from including the element that provides the highest discriminant power between sources and then adds elements one by one so that at each step the discriminant power is maximized. To perform a two-way DFA from the tools window, choose **Tools**→**Miscellaneous Analysis tools**→**Stepwise discriminant Function Analysis (Two-way)**.

In the central form, indicate the two source groups between which the DFA is intended to be performed. Checking the Box-Cox transformation box will indicate that the program will do the DFA on Box-Cox transformed data with an optimal λ to achieve normality. Click Ok at the bottom of the page.

The result window includes three panels. The top panel contains the p-values for the χ^2 statistics versus the number of elements included. You can see the tabular result by clicking on the table button  or see the graphic representation of the S-values by clicking on the graph button  on the side of the panel.

For example, in Figure 4.6 ^{13}C has the highest discriminatory power which the elements Lead through Vanadium did not increase the discriminatory power substantially and cannot be used to differentiate between the two source groups selected. Note that the element with low discriminatory power between two groups may still have discriminatory power between other groups.

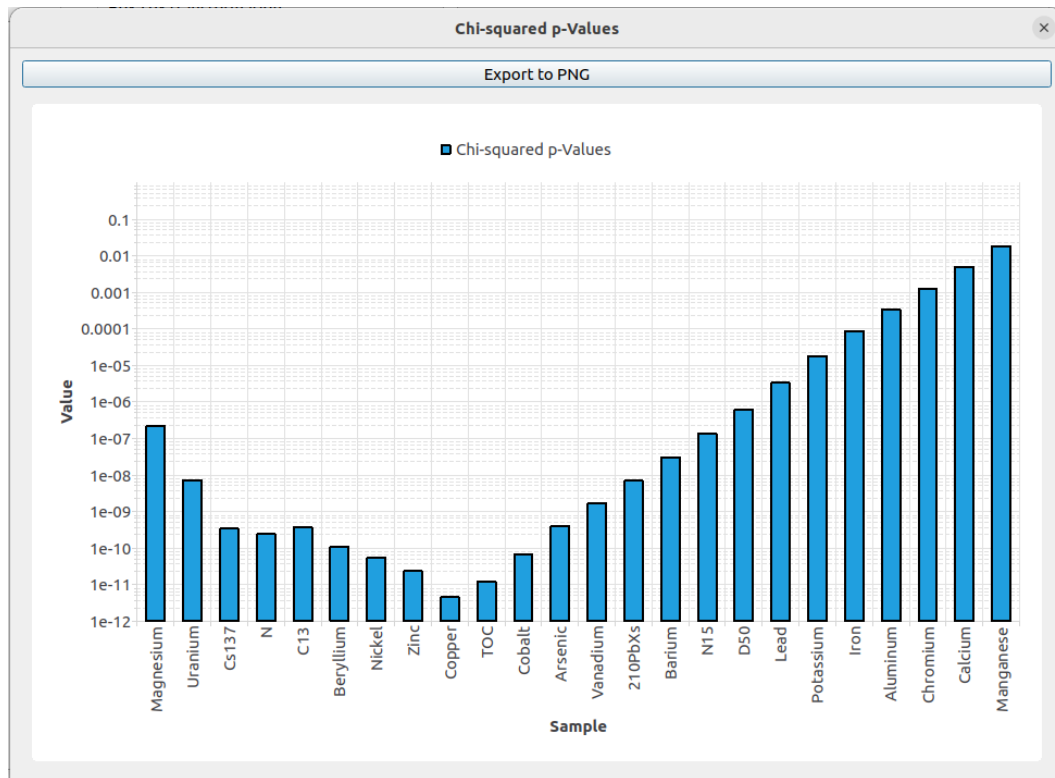


Figure 4.6: χ^2 - p-values for two way stepwise DFA

The graph shows that the χ^2 drops by including all the elements up to copper and then start rising. Including all elements results in a χ^2 - p - value of smaller than 0.05.

The second panel shows the Wilks' Lambda values as a function of number of elements included (Fig. 4.7).

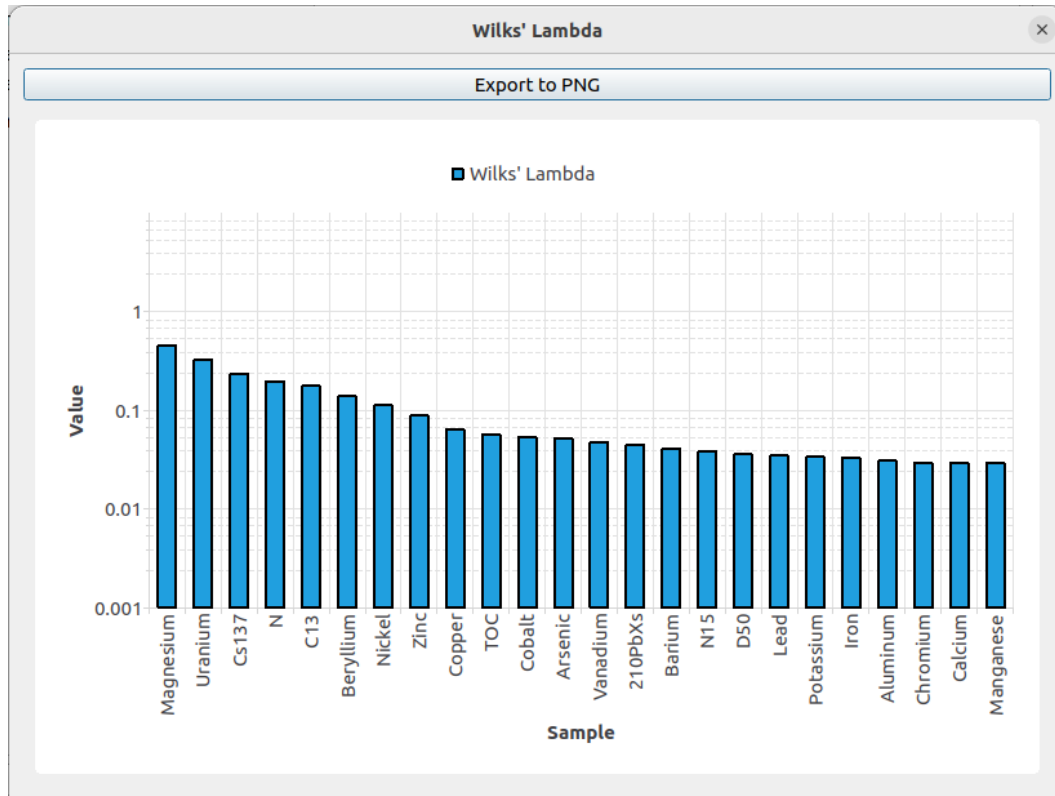


Figure 4.7: Wilks' Lambda values for consecutive addition of elements in two way DFA

As it can be seen, in this case the Wilks' Λ continues to improve to the last element added.

Finally, the third panel shows the p -value for the F-test on the discriminant scores obtained from the two source groups (Fig 4.8).

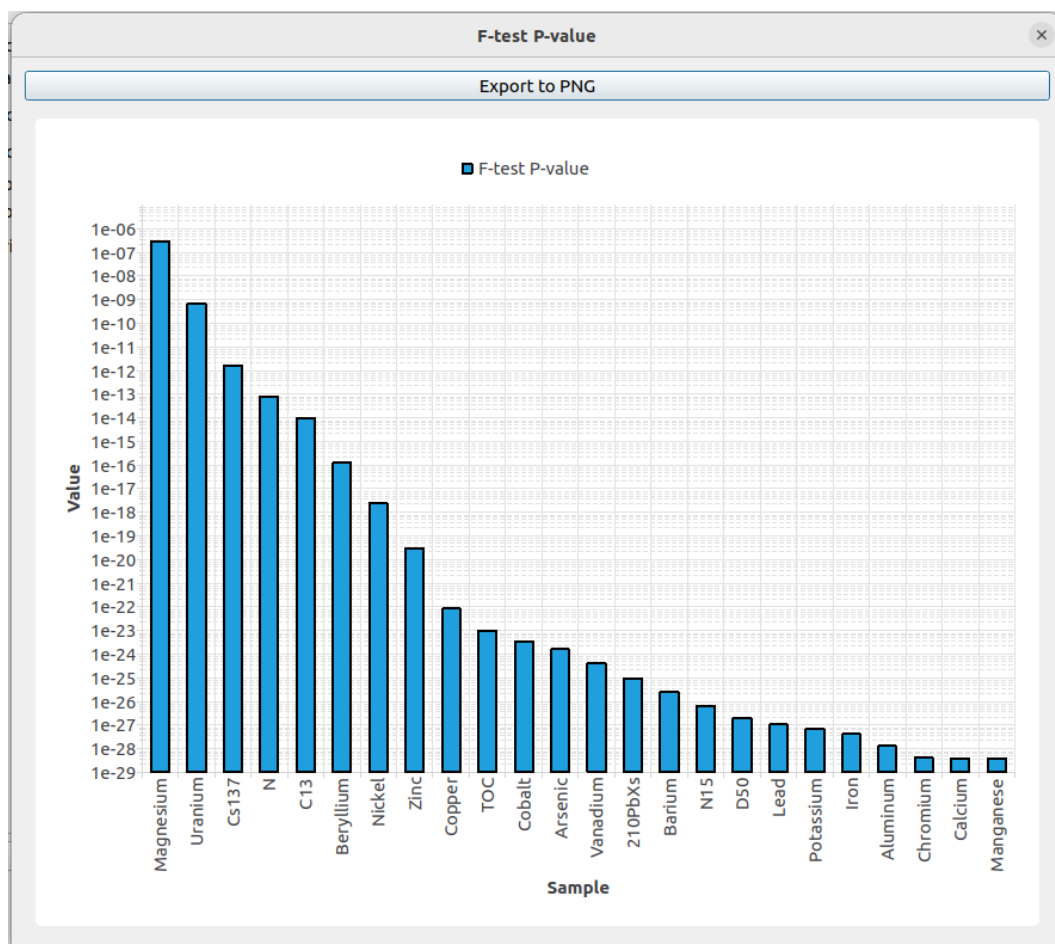



Figure 4.8: F-test p-values for consecutive addition of elements in two way DFA

To see the separation provided as a result of selecting a number of elements, first select the elements to be included in the analysis via the top menu **Constituent Properties**. This can be done based on the result of the step-wise DFA. Then choose **Tools**→**Miscellaneous Analysis tools**→**Discriminant Function Analysis (Two-way)**. Select the two sources for which the DFA is intended to be performed and choose whether the OM and size correction to be performed and whether the Box-Cox transformation is to be performed. Choose *Use only selected element* box so only the elements included in the analysis to be considered. The result window contains the χ^2 p-value, the F-test p-value and the calculated discriminant scores for the two source groups. Click on the graph icon  on the side of the panel titled "Projected Elemental Profiles". The graph shows the discriminant scores for the samples in each source group (Fig. 4.9).

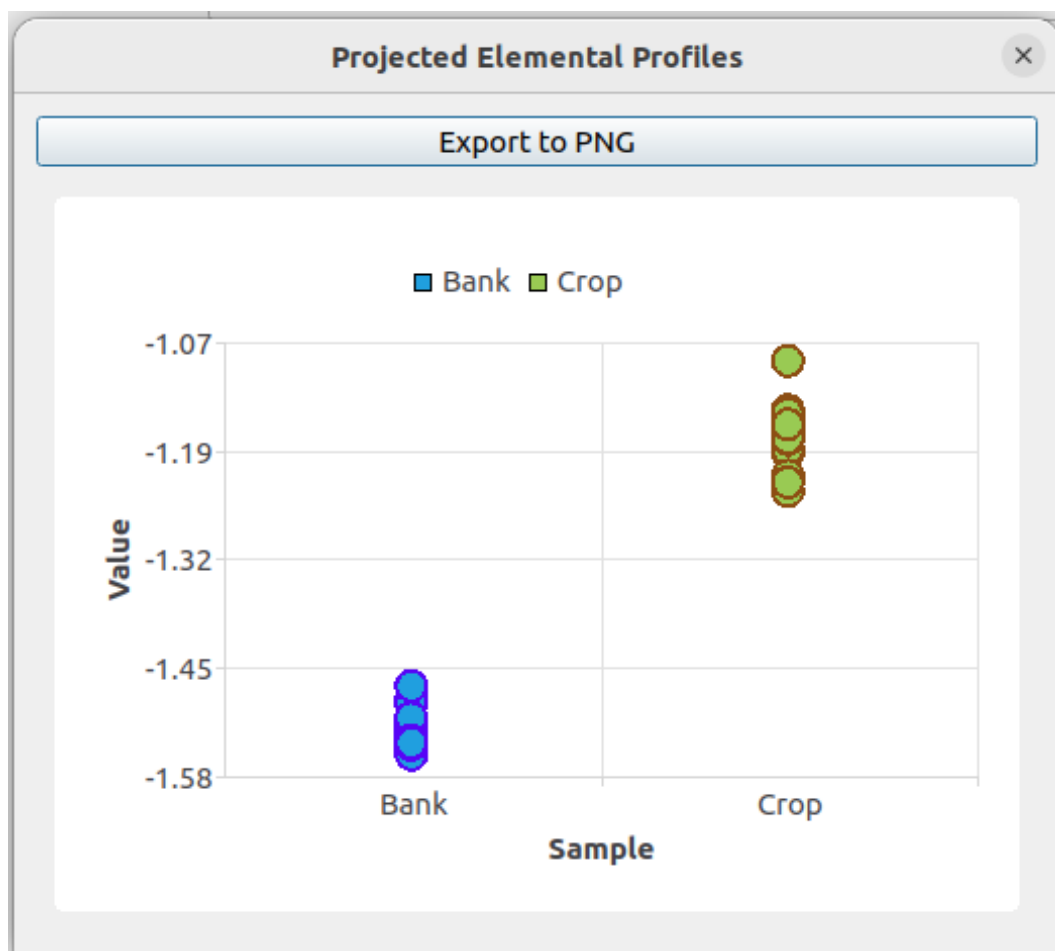


Figure 4.9: Discriminant scores for samples in source groups Bank and Crop when Beryllium, ^{13}C , ^{137}Cs , Manganese, N, Nickle, and Uranium are included in the analysis

we see the separation provided as a result of selecting a number of elements, first select the elements to be included in the analysis via the top menu **Constituent Properties**. This can be done based on the result of the step-wise DFA. Then choose **Tools**→**Miscellaneous Analysis tools**→**Discriminant Function Analysis (one vs. the rest)**.

4.5 One vs. the rest DFA

The one vs. the rest DFA, performs Discriminant Function Analysis between all the samples belonging to one specific source group and all other samples collectively. To perform one vs. the rest step-wise DFA, from the tools panel, choose **Tools**→**Miscellaneous Analysis tools**→**Stepwise Discriminant Function Analysis (one vs. the rest)**. In the central form select the source group the analysis will be done on, whether Box-Cox transformation and organic and size correction to be performed, and click on the Ok button.

The results are similar to those for two-way stepwise discriminant function analysis but this time the statistics are calculated between the source group indicated and all other source samples. Figure 4.10 shows the χ^2 p-values for Bank samples versus the rest of the samples.

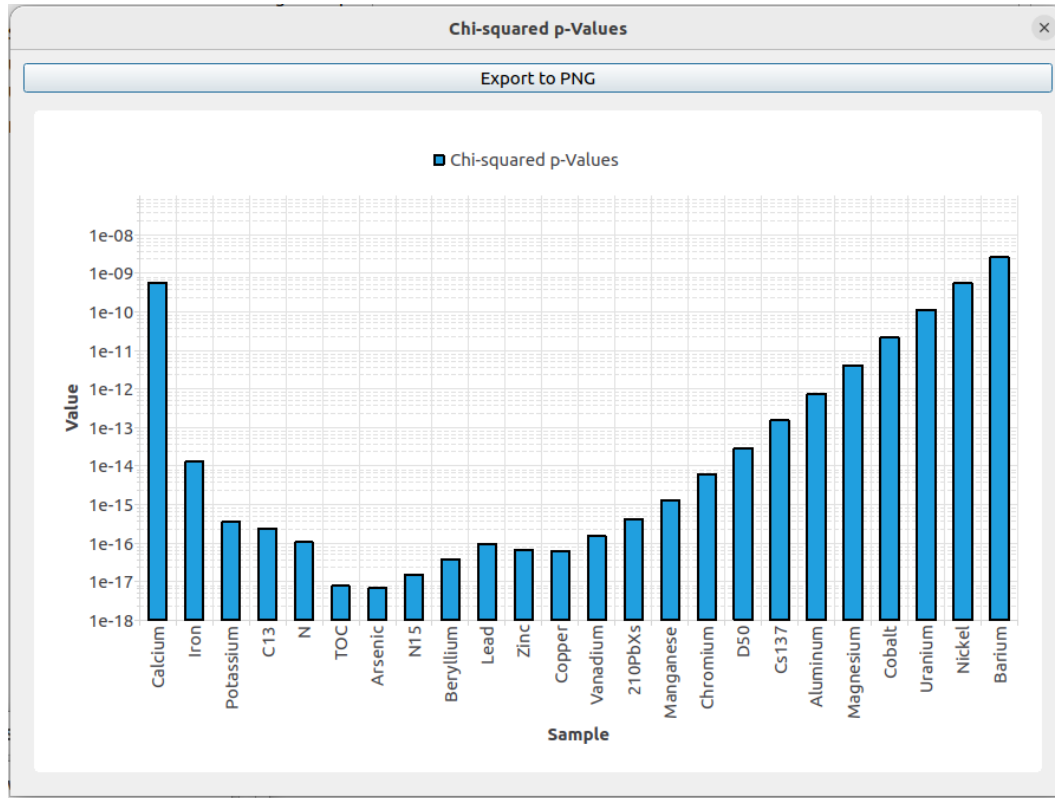



Figure 4.10: χ^2 p-values vs. the number of elements included for DFA between Bank and other sources

After performing stepwise DFA, you can use the results to select the elements that are beneficial in sediment fingerprinting.

To select the elements to be considered in the fingerprinting based on the result of the DFA analysis from the top menu, choose **Settings**→**Constituent properties** and from the column titled "Include in analysis" indicate whether a constituent to be included or excluded.

Select the source group for which the DFA is intended to be performed and choose whether the OM and size correction to be performed and whether the Box-Cox transformation is to be performed. Choose *Use only selected element* box so only the elements included in the analysis to be considered. The result window contains the χ^2 p-value, the F-test p-value and the calculated discriminant scores for the two source groups. Click on the graph icon  on the side of the panel titled "Projected Elemental Profiles". The graph shows the discriminant scores for the samples in each source group (Fig. 4.11). Note that the weight vector \mathbf{w} is determined to maximize the difference between bank samples and the rest of the samples.

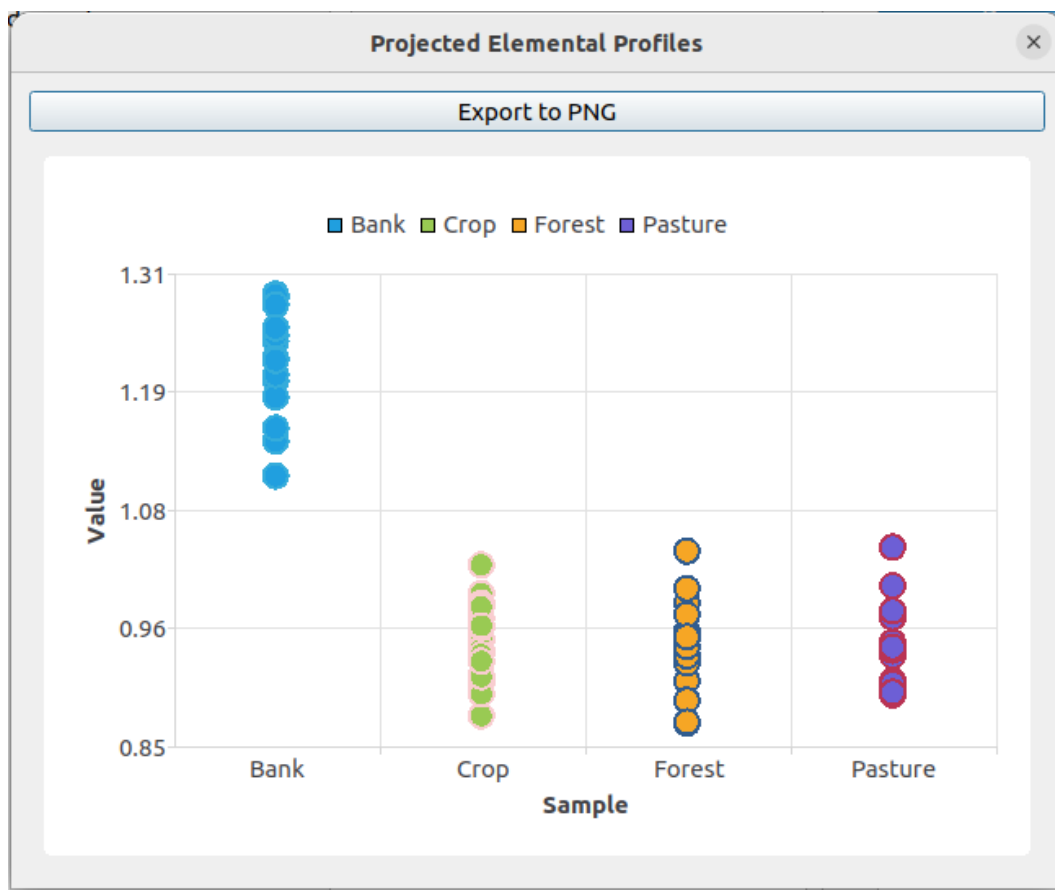



Figure 4.11: Discriminant scores as a result of one-vs-rest DFA analysis for samples in source groups Bank when Beryllium, ^{13}C , ^{137}Cs , Manganese, N, Nickle, and Uranium are included in the analysis

4.6 Distribution fitting

The distribution fitting tool enables visual inspection of the frequency distribution of constituent content in each source, allowing you to determine whether it aligns better with a normal or log-normal distribution. To execute distribution fitting, from the tools menu, select **Tools**→**Miscellaneous tools**→**Distribution fitting**. From the central form, you can choose the element and source group, and decide whether to include all samples or only the selected ones in the analysis. You can also specify if the distribution fitting should be done on the Box-Cox transformed elemental contents or on the raw data. You can also select the target sample for based on which the OM and size correction is done in case you want the fitting to be done on corrected data. If you desire the analysis to be done on uncorrected data, leave the "OM and Size corrected based on target sample" box blank. Click the "Ok" button.

In the results window, click the "Graph" button  next to the PDF (top) panel or the CDF (bottom) panel to view the results. If the Box-Cox transformation is selected, the results will show the Box-Cox transformed elemental contents fitted

with a Gaussian distribution. If not selected, the graphs will include both normal and log-normal fits.

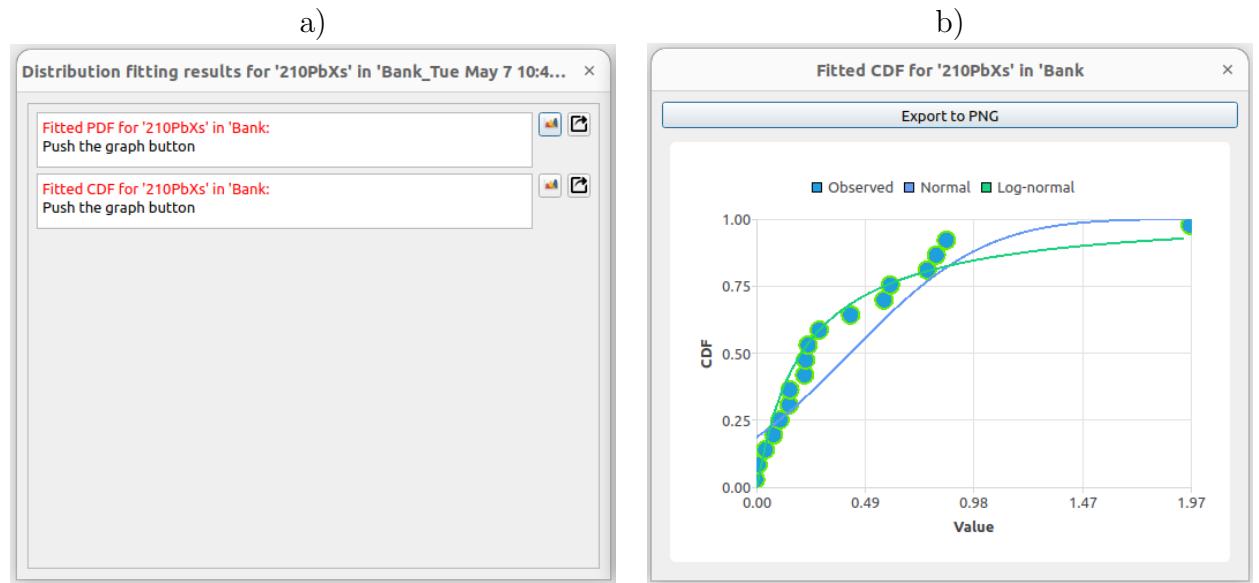


Figure 4.12: Distribution fitting without Box-Cox transformation

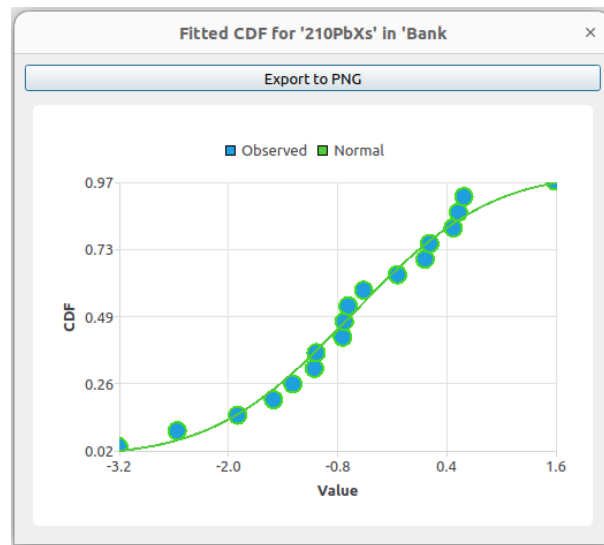




Figure 4.13: Distribution fitting with Box-Cox transformation

4.7 Elements' Discriminant Power (Multi-way)

The Elements' Discriminant Power (Multi-way) tool evaluates each element's ability to distinguish between different source group pairs. It provides three types of results:

1. The ratio of the mean distance between groups to the standard deviation within groups for each element across all source pairs.
2. The "discriminant fraction" indicates the number of sources that can be accurately attributed to one of the two sources in a pair based solely on a single element.
3. A t-test p-value that tests the hypothesis that the mean elemental content between two source groups is the same.

To perform multi-way elements' discriminant power analysis, from the tools window, select **Tools**→**Miscellaneous tools**→**Elements' Discriminant Power (Multi-way)**. From the central window, you can select whether to conduct the analysis on untransformed or log-transformed elemental contents. To include target samples in the analysis, check the corresponding box. The p-value threshold specifies the cutoff above which values will be highlighted in the resulting table. You can also select the target sample for based on which the OM and size correction is done in case you want the fitting to be done on corrected data. If you desire the analysis to be done on uncorrected data, leave the "OM and Size corrected based on target sample" box blank. After selecting the options, click the "Ok" button. A result window will appear. The top panel contains the ratio of the mean distance between groups to the standard deviation within groups for every pair of source groups. Click on the table button  next to the top panel to see the results in tabular format. A higher value indicates an element's greater ability to distinguish between the two source groups. Scroll down to the second panel in the results window and click the table button  next to it. The values represent the percentage of source samples that can be identified based solely on the element.

A value of 0.5 means that samples with a given elemental content are equally likely to belong to either source group, indicating that the element has no discriminating power. The closer the discriminant fraction value is to 1, the greater the element's ability to differentiate between groups.

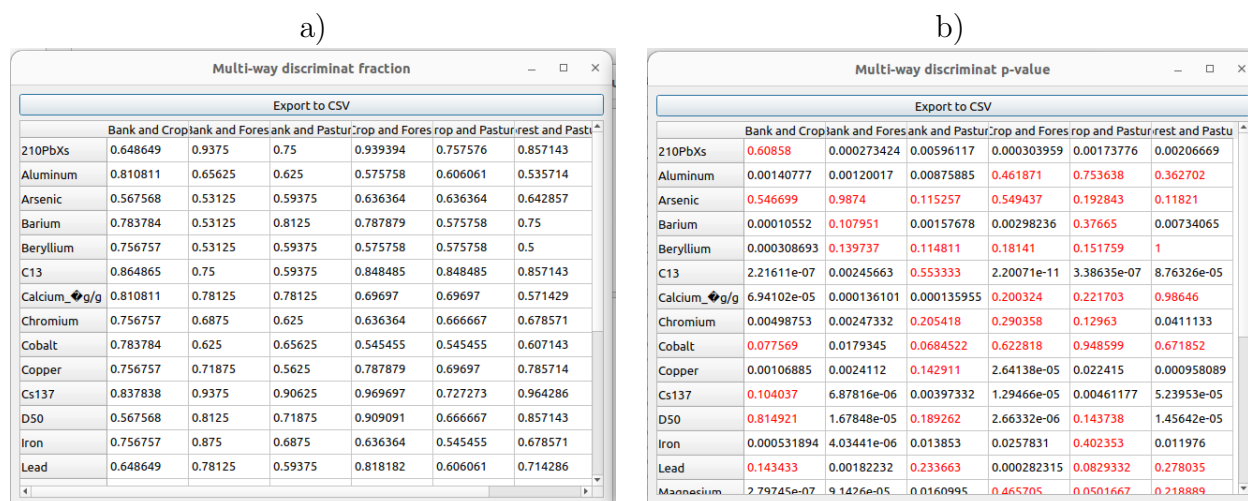



Figure 4.14: a) Discriminant fraction b) Discriminant p-value

Finally, scroll down to the bottom panel and click the table button  next to it to view the p-value for each element and each source pair. Values highlighted in red indicate that the specific element lacks discriminant power between the two source groups.


4.8 Elements' Discriminant Power (Two-way)


This tool functions like Multi-way Discriminant Power, but the analysis is conducted between specific source groups chosen by the user.

4.9 Error Analysis

The Error Analysis tool conducts uncertainty analysis on estimated source contributions through bootstrapping. It randomly excludes a percentage of source samples and then estimates the source contributions using the Levenberg-Marquardt algorithm.

To perform error analysis, from the tools window, select **Tools**→**Miscellaneous tools**→**Error Analysis**. In the central window, choose the target sample, the number of realizations, and the percentage of source samples to be excluded in each iteration. You can also select whether to conduct the analysis on OM and size-corrected source elemental profiles.

Once the analysis is complete, the results window will display three panels. The first panel shows the source contributions estimated for each realization. If the number of realizations is less than 100, you can view the results as a graph. Click the table button  next to the top panel to see the inferred source contribution for each realization.

The second panel provides the frequency distribution of the inferred contributions. Click on the graph button next to the second panel  to see the posterior

distributions for the contribution of each source group (4.15).

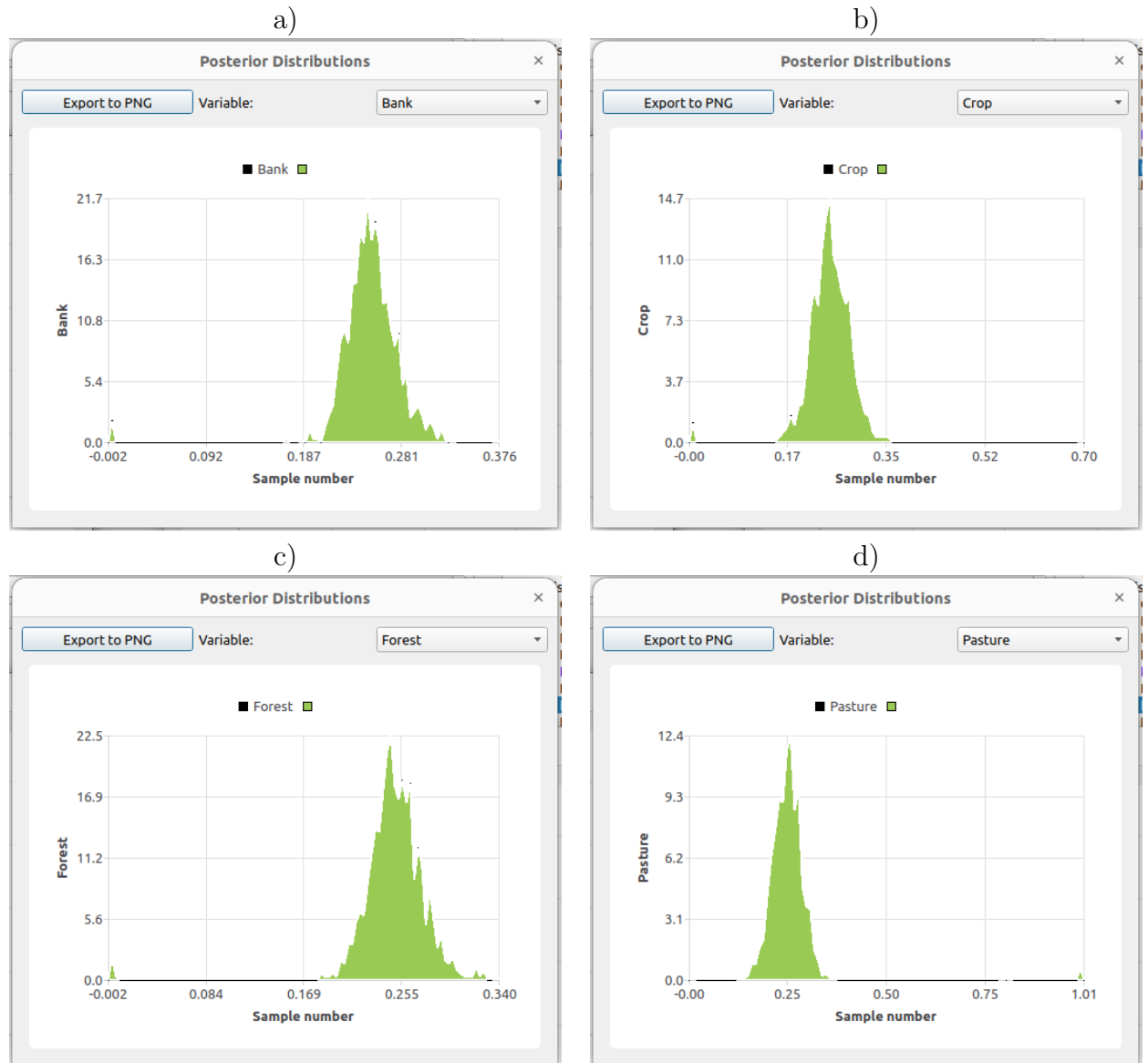



Figure 4.15: Error analysis frequency distributions for a) Bank, b) crop, c) forest and, d) pasture source groups

Finally, the bottom panel contains the 95% credible intervals. Click on the graph button  next to the third panel to see the 95% probability bounds of the source contributions (Figure 4.16).

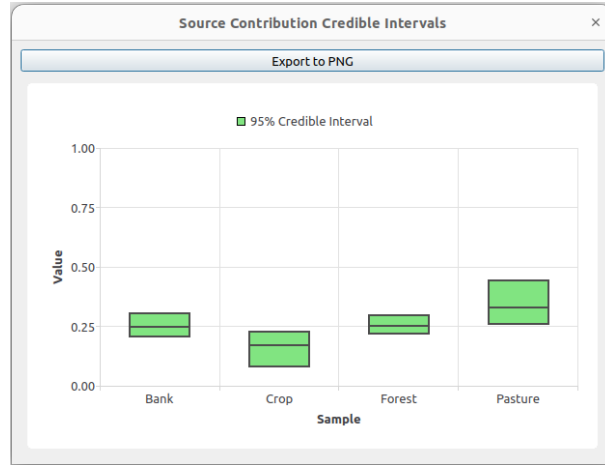



Figure 4.16: 95% bounds of source contributions as a result of error analysis

4.10 Kolmogorov-Smirnov for individual group/element

The Kolmogorov-Smirnov (K-S) tool facilitates the visual inspection of the normality of an element within a source group. It utilizes the K-S test, which determines the normality based on the maximum difference between the empirical distribution function (EDF) of the sample and the cumulative distribution function (CDF) of a specified theoretical distribution, either normal or log-normal. To perform the K-S test, navigate through **Tools**→**Miscellaneous tools**→**Kolmogorov-Smirnov test for an individual group/element**. In the central form, select the element, the source group, and the distribution type (normal or log-normal) to be tested. After clicking the "Ok" button, in the results window, click on the graph button . The graph that appears will display the empirical distribution, the fitted theoretical distribution, and the difference between them (Figure 4.17). The extent of the maximum difference between the empirical and theoretical distributions serves as an indicator of the adequacy of the chosen distribution to represent the element's data.

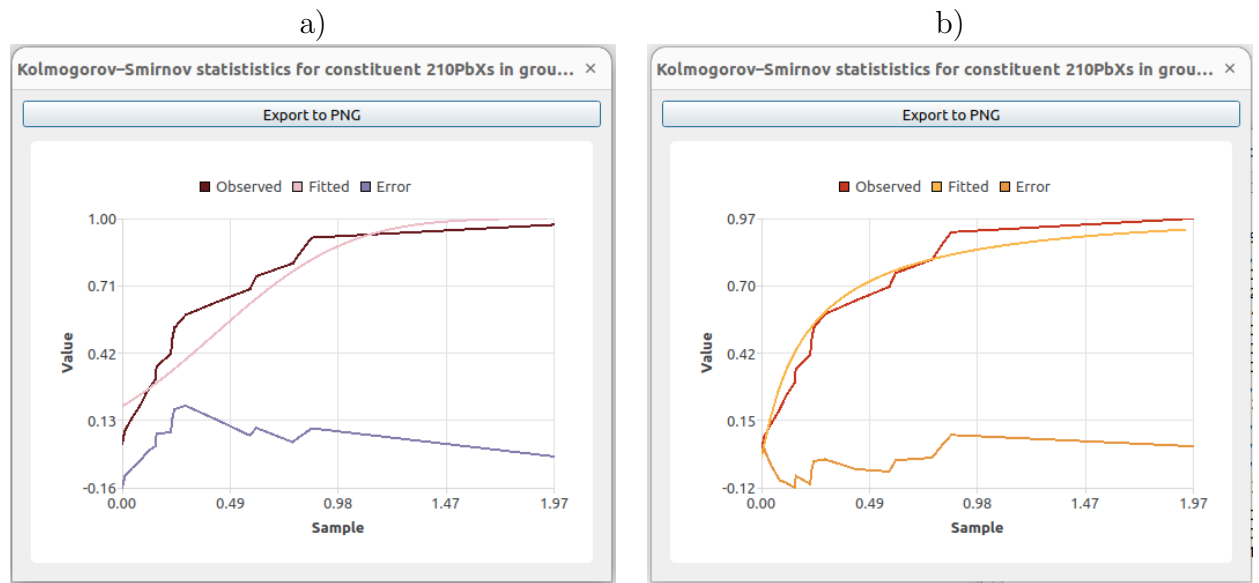




Figure 4.17: K-S test with a) a normal distribution and b) log-normal distribution

4.11 Kolmogorov-Smirnov for a group

The Kolmogorov-Smirnov for a group tool, performs the K-S analysis for all the elements in a particular source group and reports the maximum difference between the empirical and theoretical density functions for each element. To use this tool, navigate through **Tools**→**Miscellaneous tools**→**Kolmogorov-Smirnov test for a group**. From the central form, select the distribution type and the source group and then click on the "Ok" button. On the result window, the maximum difference between the empirical and theoretical cumulative density functions can be viewed in tabular format by clicking on the table  or as a bar chart via the graph button .

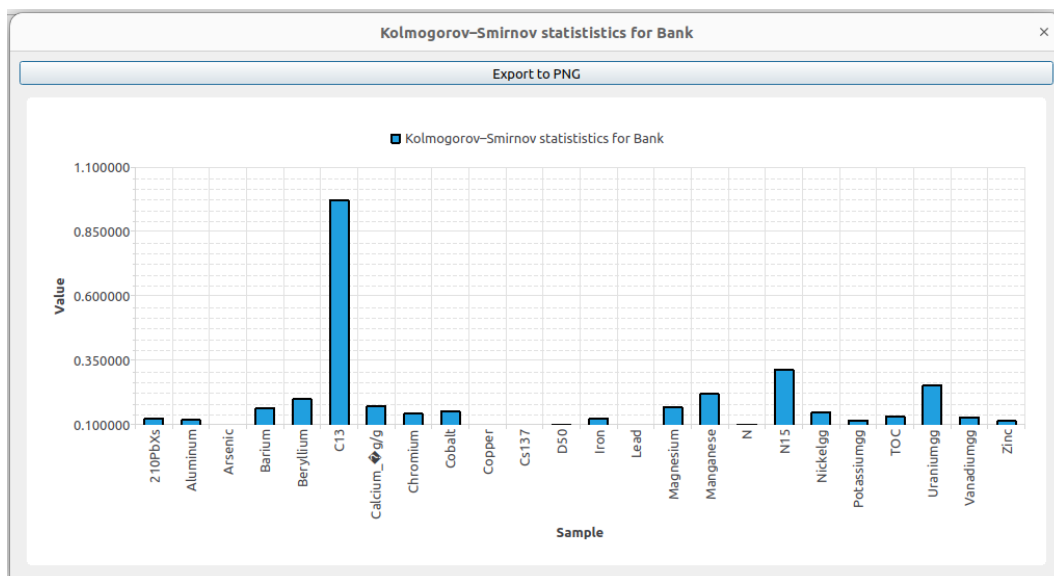


Figure 4.18: K-S results for all elements in a source group


4.12 Optimal Box-Cox parameters

The optimal Box-Cox parameters tool determines the optimal λ that transforms the elemental contents of a specific source group to closely resemble a normal distribution. The Box-Cox transformation is defined as:

$$\hat{y}_{i,j,k} = \begin{cases} \left(\frac{y_{i,j,k}}{\sigma_{i,j}} \right)^{\lambda} - 1 & \text{for } \lambda \neq 0 \\ \log \left(\frac{y_{i,j,k}}{\sigma_{i,j}} \right) & \text{for } \lambda = 0 \end{cases} \quad (4.12.8)$$

The optimization minimizes the maximum difference between the cumulative distribution functions of the data and a fitted normal distribution according to the Kolmogorov-Smirnov test. The closer the optimal value of λ to one means the data is more closely represented based on a normal distribution while the closer the value to zero means the data more closely follows a log-normal distribution.

To obtain the optimal Box-Cox parameters, go to the tools window and select **Tools**→**Miscellaneous tools**→**Optimal Box-Cox parameters**. In the central form, choose the source group you want to analyze, then click the "Ok" button.



The result window will contain the value of optimal Box-Cox parameters for every element for the specified source group. Click on the table button  to see the values in tabular format.

4.13 Source Verification

The source verification tool assesses how well individual source samples align with their overall source group. It treats each source sample within a group as a target sample and performs fingerprinting using the Levenberg-Marquardt algorithm. If

a source sample's elemental profile closely matches that of its source group, the analysis should identify that group as the dominant contributor.

To perform Source Verification, choose **Tools**→**Miscellaneous tools**→**Source Verification** from the tools window. Select the source group you would like the analysis to be performed on. Choose whether size and organic matter correction needs to be performed prior to the analysis. Note that only the elements and samples selected to be included in the analysis in **Settings**→**Constituents Properties** and **Settings**→**Include/Exclude Samples** will be considered in this analysis. Click the "Ok" button.

The results window will contain the estimated contribution of each source into all the samples in the selected source group. Click on the table button  or the  to see the results in tabular or graph format (Fig. 4.19).

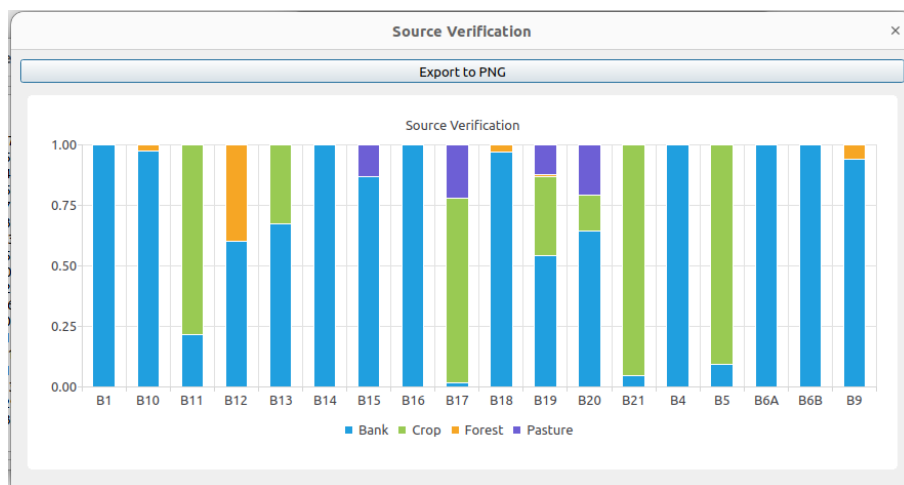


Figure 4.19: Source verification results

4.14 Genetic Algorithm estimation

The Genetic Algorithm estimation does not calculate the mean elemental profiles from the source samples directly but rather estimates a mean value for each constituent in each source as part of the inverse modeling. Each constituent in each source group is assumed to follow a log-normal distribution:

$$\tilde{y}_{i,j} \sim \frac{1}{\sqrt{2\pi}\sigma_j y_{i,j}} \exp\left(-\frac{[\ln(y_{i,j}) - \mu_{i,j}]^2}{2\sigma_j^2}\right) \quad (4.14.9)$$

where σ_j is the shape factor of the log-normal distribution for element j , which is assumed to be the same for all source groups, and $\mu_{i,j}$ is the scale factor. As part of the optimization algorithm, the values of σ and $\mu_{i,j}$ will be estimated along with the contribution of each source. To predict the elemental profile as a result of a given set of σ , μ and the contribution of the source, the mean elemental profile is calculated based on the parameters of the log-normal distribution as:

$$y_{i,j} = \exp\left(\mu_{i,j} + \frac{\sigma_j^2}{2}\right) \quad (4.14.10)$$

Due to the fact that the number of parameters to be estimated in this case is much larger than in the case where the mean elemental profiles of the sources are directly calculated, SedSAT3 uses a genetic algorithm to estimate the parameters. There is also an option to use a hybrid genetic algorithm consisting of a conventional binary genetic algorithm with a stochastic gradient descent method [Alikhani et al., 2017].


To perform fingerprinting using the genetic algorithm from the tools window, select **Tools→Fingerprinting tools→Maximum likelihood fingerprinting using Genetic Algorithm**.

On the form that appears:

- **Apply size and organic matter correction** indicates whether the analysis is performed on the corrected elemental profiles based on size and organic matter. Note that in order to size and organic correction to be done, a size and organic matter correction step (section 2.2) should have been performed before.
- **Crossover probability** Indicates the cross-over probability in the genetic algorithm. In most cases, the default value of 1.0 works fine.
- **GA output file** indicates the text file where detailed information about the genetic algorithm process will be saved.
- **Mutation probability** is the mutation probability in the genetic algorithm. In most cases, the default value should be good.
- **Number of Generations** is the number of generations used in the genetic algorithm. It can be increased if it is observed that the GA does not adequately converge with the existing number of generations.
- **Number of threads used** is the number of processor's threads used for performing GA optimization. If the number of cores available on your CPU is smaller than the number of threads indicated by the user, the number of threads used all the cores will be used.
- **Numerical cross-over:** If checked, the algorithm will use a numerical cross-over operation rather than a binary cross-over.
- **Population:** is the number population number used in the GA. It can be increased if it is observed that the GA does not adequately converge with the existing number of generations. In most cases, the value of 100 should be adequate.
- **Sample:** The target sample to be analysed.

- **Shake coefficient:** is the coefficient used for numerical perturbation that is utilized when the GA does not improve the likelihood in a few consecutive generations. Leave it at the default value.
- **Shake coefficient reduction factor** is the factor by which the shake coefficient is reduced if the likelihood is not improved after a few consecutive iterations of numerical shaking.
- **Gradient Descent:** Combines the genetic algorithm with a gradient descent method. This option often results in faster and more accurate convergence.

Note that there is a possibility that the optimal solution is not obtained with the number of generations specified. It is a good practice to execute the fingerprinting step multiple times to ensure the inferred source contributions are close enough or to increase the number of generations to see if the inferred source contributions change relative to the one obtained from a smaller number of generations. Getting different results when executing the fingerprinting multiple times may signify equifinality.

After setting the parameters, press Ok and wait for the analysis to complete. The result window will contain several panels. The first panel shows the inferred contribution of sources. Pressing the graph button  on the side will show the inferred contributions as a pie chart (Figure 4.20).

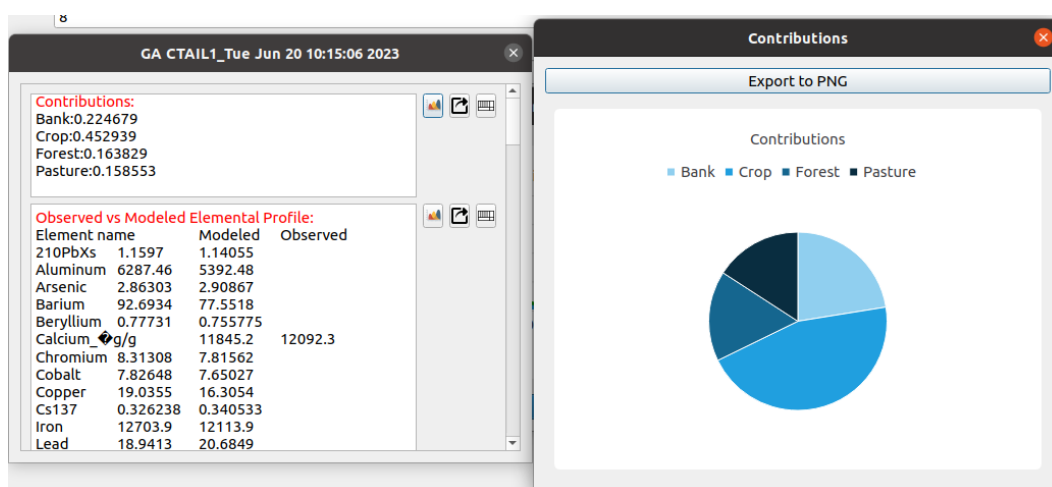


Figure 4.20: Estimated source contribution by the genetic algorithm

The second panel contains the modeled vs. measured elemental contents (Figure 4.21, and the third panel contains the measured vs. modeled isotope δ values if isotopes are included in the analysis.

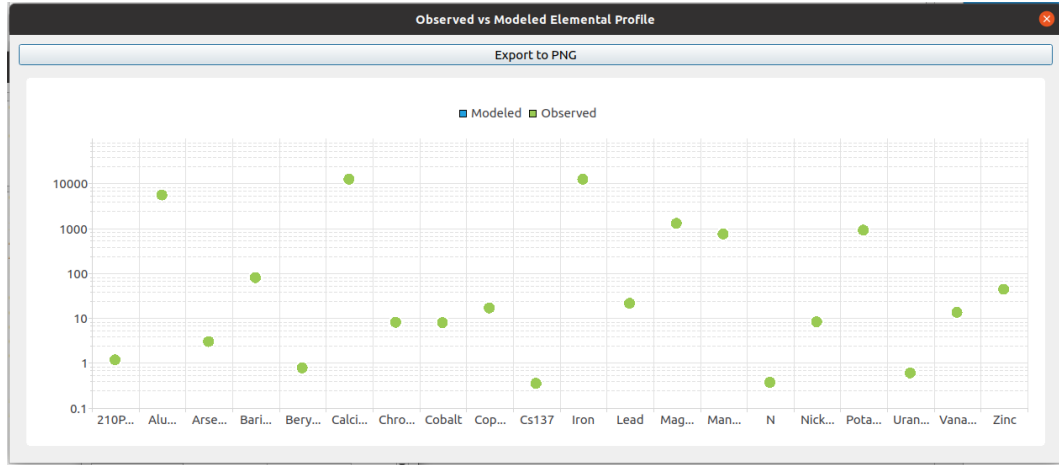


Figure 4.21: Predicted vs. measured elemental contents by the genetic algorithm

The fourth panel contains the calculated element means for each source group, and the fifth panel contains the inferred mean based on the estimated log-normal distribution for the elemental contents in each source. This information can be used for the diagnosis of the results.

The sixth panel shows the calculated scale parameter, μ , of the log-normal distribution of the log of the geometrical mean of the elemental contents in each source, and panel seven contains the estimated value of mu by the genetic algorithm.

Panels eight and nine contain each source group's calculated and estimated shape parameter σ of elemental contents, respectively.

Chapter 5

Appendix A: Mathematical basis

5.1 Mass balance formulation

The basis of the sediment fingerprinting method is the use of tracer groups, where each group is a unique sediment source, to identify statistically significant differences in chemical properties between the source areas and the fluvial sediments (the target sample). That is, if n sources are mixed $y_{i,j}$ [M/M] mass of a particular element j in each source i , the elemental content of element j in the mixture (or target sample) will be calculated as:

$$c_j = \sum x_i y_{i,j} \quad (5.1.1)$$

where x_i is the contribution proportion of source i . Eq. (5.1.1) can be written for all the elements considered in the analysis in the following matrix form:

$$\mathbf{C} = \mathbf{YX} \quad (5.1.2)$$

where $\mathbf{C}_m = [\dots c_j \dots]^T$ is a vector containing the elemental profile of the target sample, $\mathbf{X}_n = [\dots x_i \dots]^T$ is a vector containing the contribution fraction of each source, $\mathbf{Y}_{n \times m} = [y_{i,j}]$ is a matrix containing the elemental content of each element in each source and m is the total number of elements (or signatures) used in the analysis.

Note that the sum of the elements of \mathbf{X} must be equal to 1:

$$\sum x_i = 1 \quad (5.1.3)$$

5.1.1 Stable Isotopes

Stable isotopes are also commonly used for sediment fingerprinting. The stable isotope content of a sample is usually expressed using the δ value, which is defined as:

$$\delta^u E = \left(\frac{\frac{^u E}{E}}{\left(\frac{^u E}{E} \right)_{std}} - 1 \right) \times 1000 \quad (5.1.4)$$

where uE is the isotope content, E is base element. The $\left(\frac{{}^uE}{E}\right)_{std}$ is a standard isotope ratio that is used to calculate the δ value. So the isotope content uE of for sample can be calculated as:

$${}^uE = \left(\frac{\delta {}^uE}{1000} + 1\right) \left(\frac{{}^uE}{E}\right)_{std} E \quad (5.1.5)$$

When several sediment sources are mixed, the isotope content of the mixture can be calculated as:

$$\zeta_j = \sum x_i v_{i,j} \quad (5.1.6)$$

where ζ_j is the isotope content of isotope j in the mixture and $v_{i,j}$ is the mean isotope content of isotope j in source i which is calculated using Eq. (5.1.5).

Substituting Eq. (5.1.5) into Eq. (5.1.6) results in:

$$\delta \zeta_j = \frac{\sum x_i \delta v_{i,j} y_{i,j}}{\sum x_i y_{i,j}} \quad (5.1.7)$$

where δ refers to the δ values of each isotope.

The set of linear equations obtained from isotopes can be appended to Eq. (5.1). Note that in order to include isotopes in the analysis, the base elements of those isotopes are needed.

5.2 Maximum Likelihood estimation

The goal of sediment fingerprint modeling is to estimate the source contributions, \mathbf{X} given some measured values representing the elemental profiles in a number of source groups, \mathbf{Y} and the elemental profile of a target sample \mathbf{C} . Here we denote the measured elemental profiles of sources by $\tilde{\mathbf{Y}}$ and the measured elemental profile of the target sample as $\tilde{\mathbf{C}}$. The goal is to infer \mathbf{X} given $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{C}}$. Note, because multiple samples representing a source group typically exist, we have multiple instances of $\tilde{\mathbf{Y}}$ to work with, so we denote the elemental profiles of sources in an individual source sample with k , subsequently by $\tilde{\mathbf{Y}}_k$.

5.2.1 Treating source elemental composition deterministically

The simplest assumption is to consider \mathbf{Y} to be deterministically known as an average over all samples representing that group.

$$y_{i,j} = \frac{1}{ns_i} \sum \tilde{y}_{i,j,k} \quad (5.2.8)$$

where ns_i is the number of samples in source group i .

Assuming conditional independence of errors, the probability density function of observing a measured target sample elemental profile $\tilde{\mathbf{C}}$, given a contribution vector \mathbf{X} can be calculated as:

$$p(\tilde{\mathbf{C}}|\mathbf{X}) = \prod \frac{g'_j(\tilde{c}_j)}{\sigma_j} \phi \left\{ \frac{g_j[c_j(\mathbf{X})] - g_j(\tilde{c}_j)}{\sigma_j} \right\} \quad (5.2.9)$$

where ϕ is the standard normal distribution, σ_j is the error standard deviation for element j , and g_j is a transformation function that is deemed to map the probability distributions of the error to a normal distribution, which determines the error structure assumed in the method. For example, if the error structure is assumed to be normal, $g(c_j) = c_j$, and the likelihood function will be Gaussian, or if the error structure is assumed to be log-normal and multiplicative, $g(c_j) = \ln(c_j)$ which results in a log-normal likelihood function.

The logarithm of the likelihood function is typically used for maximum likelihood estimation:

$$\ln[p(\tilde{\mathbf{C}}|\mathbf{X})] = \sum \ln[g'_j(\tilde{c}_j)] - \sum \ln(\sigma_j) - \frac{m}{2} \ln(\pi) - \sum \frac{\{g_j[c_j(\mathbf{X})] - g_j(\tilde{c}_j)\}^2}{2\sigma_j^2} \quad (5.2.10)$$

Because the first three terms are independent of \mathbf{X} , in the maximum likelihood approach, we only need to find source contributions, \mathbf{X} that maximize the last terms, which turns the problem into a least-squares problem. An optimization algorithm can be used to estimate \mathbf{X} vector and σ values that will maximize the log-likelihood function.

In the case of a log-normal and multiplicative error structure, it may be reasonable to assume the same error standard deviation to be the same for all of the elements. This implies that the expected relative error for all elements are equal. This assumption will simplify Eq. (5.2.10) to:

$$\ln[p(\tilde{\mathbf{C}}|\mathbf{X})] = \sum_j \ln[g'_j(\tilde{c}_j)] - m \ln(\sigma) - \frac{m}{2} \ln(\pi) - \frac{1}{2\sigma^2} \sum \{g_j[c_j(\mathbf{X})] - g_j(\tilde{c}_j)\}^2 \quad (5.2.11)$$

which means estimating the source contributions \mathbf{X} involves only minimizing the sum of squared error, $\sum \{g_j[c_j(\mathbf{X})] - g_j(\tilde{c}_j)\}^2$.

5.2.2 Treating source elemental composition as unknown

Because the number of samples collected for each source is typically limited, estimation of \mathbf{Y} based on Eq. (5.2.8) is not necessarily accurate. If we assume the elemental content of sources follows a probability density function:

$$\tilde{y}_{i,j} \sim p_{y,i,j}(\boldsymbol{\theta}_{i,j}) \quad (5.2.12)$$

or in matrix form:

$$\mathbf{Y} \sim \mathbf{P}_y(\boldsymbol{\Theta}) \quad (5.2.13)$$

where $\theta_{i,j}$ is the parameters of the distribution for element j in source i , Θ is the matrix containing parameters for all elements and sources and \mathbf{P}_y is a matrix transformation representing elemental content distributions for all elements in all sources.

For the maximum likelihood estimation, the probability of observing the target elemental profile and the source elemental profile of all samples in all source groups given should be maximized:

$$p(\tilde{\mathbf{C}}, \tilde{\mathbf{Y}}|\mathbf{X}, \mathbf{Y}) = \prod_j \frac{g'_j(\tilde{c}_j)}{\sigma_j} \phi \left\{ \frac{g_j[c_j(\mathbf{X}, \Theta)] - g_j(\tilde{c}_j)}{\sigma_j} \right\} \prod_k \mathbf{P}_y(\tilde{\mathbf{Y}}_k; \Theta) \quad (5.2.14)$$

where $\tilde{\mathbf{Y}}_k$ is the observed source elemental matrix for sample k . Note that because the expected value of \mathbf{P}_y depends on the parameters defining it (i.e. Θ), the predicted elemental profiles of the target samples will depend on the contribution of each source and the distribution parameters the elemental contents in each source:

$$c_j(\mathbf{X}, \Theta) = \sum_i x_i E(y_{i,j}; \theta_{i,j}) \quad (5.2.15)$$

where $E(y_{i,j}; \theta_{i,j})$ is the expected value of $y_{i,j}$ based on the source elemental content distribution:

$$E(y; \theta) = \int y p_y(y; \theta) dy \quad (5.2.16)$$

As an example if p_y is assumed to be a log-normal distribution with parameters $\theta \equiv \{\mu, \sigma\}$, the expected value of the elemental content will be $E(y) = e^{\mu + \frac{\sigma^2}{2}}$.

If we assume that the distributions of element contents in the sources are independent, the log-likelihood function will be obtained by taking the log of Eq. (5.2.14):

$$\begin{aligned} \ln[p(\tilde{\mathbf{C}}, \tilde{\mathbf{Y}}|\mathbf{X}, \mathbf{Y})] &= \sum_j \ln[g'_j(\tilde{c}_j)] - \sum_j \ln(\sigma_j) - \frac{m}{2} \ln(\pi) \\ &- \sum_j \frac{\{g_j[c_j(\mathbf{X}, \theta_{i,j})] - g_j(\tilde{c}_j)\}^2}{2\sigma_j^2} + \sum_i \sum_j \sum_k \ln[p_{y,i,j}(y_{i,j,k}; \theta_{i,j})] \end{aligned} \quad (5.2.17)$$

Because the first three terms are independent of \mathbf{X} and Θ , all we need to do to determine \mathbf{X} and Θ is to maximize the last two terms.

5.3 Bayesian inference

There are several sources of uncertainty that can impact the accuracy and reliability of sediment fingerprinting. These sources include measurement errors in both the target and source samples, non-uniform selection or release of sources, model

structural errors due to the changes in elemental composition during transport, the presence of sources not considered, and equifinality.

Measurement errors are an inevitable part of chemical analysis but can also be caused by the natural variability in sediment characteristics. Non-uniform selection of sources can occur due to systematic spatial variation of elemental profiles of a source. This results in the distribution of elements from a particular source contributing to a target sample differing from the distribution obtained from a spatially uniform sampling of sources. Structural errors stem from the fact that the model or the assumption used is a simplification of the real processes. For example, the changes in elemental composition during transport can contribute to model structural error.

Additionally, there may be sources of sediments not considered in the analysis, which can lead to inaccuracies in the fingerprinting results. Equifinality refers to the situation where multiple source combinations can produce the same (or equally good) match with a target sample, making it difficult to identify the true sources of sediment.

It is crucial to quantify these uncertainties to make sound engineering or management decisions based on sediment fingerprinting results., Bayesian inference produces a joint probability distribution of these contributions, in contrast to the maximum likelihood-based techniques.

The Bayesian inference is based on the Bayes theorem:

$$p(\Upsilon|\Xi) = \frac{p(\Xi|\Upsilon)p(\Upsilon)}{p(\Xi)} \quad (5.3.18)$$

where Υ denotes model parameters, Ξ represent observed data, $p(\Xi|\Upsilon)$ is our likelihood function, $p(\Upsilon)$ is the prior distribution of the parameters, and $p(\Xi)$ is a normalizing factor making the integral of the posterior distribution, $p(\Upsilon|\Xi)$ equal to one. For slightly complex problems, the analytical evaluation of the posterior distribution is not feasible, and often, methods such as Markov Chain Monte Carlo (MCMC) is used to generate a large number of samples from the posterior distribution. MCMC method does not need to use the precise value of the posterior distribution and only requires the relative value of the posterior distribution for any two parameter sets. Because in Eq. (5.3.18), the denominator is independent of the parameter values; we can eliminate the denominator and write it as a proportionality.

$$p(\Upsilon|\Xi) \propto p(\Xi|\Upsilon)p(\Upsilon) \quad (5.3.19)$$

In the context of sediment fingerprinting, Eq. (5.3.20) becomes:

$$p(\mathbf{X}, \mathbf{Y}|\tilde{\mathbf{C}}, \tilde{\mathbf{Y}}) \propto p(\tilde{\mathbf{C}}, \tilde{\mathbf{Y}}|\mathbf{X}, \mathbf{Y})p(\mathbf{X})p(\mathbf{Y}) \quad (5.3.20)$$

where $p(\mathbf{X})$ is the prior distribution of source contributions, and $p(\mathbf{Y})$ is the prior distribution of elemental profiles of sources. The most reasonable assumption in the absence of any other information is to assume \mathbf{X} follows a symmetric Dirichlet distribution with a parameter $\alpha = 1$. Because \mathbf{Y} is already included in the

likelihood function, its prior must be considered non-informative. The likelihood function $p(\tilde{\mathbf{C}}, \tilde{\mathbf{Y}}|\mathbf{X}, \mathbf{Y})$ is calculated based on Eq. (5.2.14).

Bibliography

Jamal Alikhani, Imre Takacs, Ahmed Al-Omari, Sudhir Murthy, and Arash Mas-soudieh. Evaluation of the information content of long-term wastewater characteristics data in relation to activated sludge model parameters. *Water Science and Technology*, 75(6):1370–1389, 2017.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.