Biogeochemical data analysis methods and R implementation in the UltraLIM project

Johanna Torppa and Maarit Middleton

Geologian tutkimuskeskus | Geologiska forskningscentralen | Geological Survey of Finland
Espoo • Kokkola • Kuopio • Loppi • Outokumpu • Rovaniemi
www.gtk.fi • Puh./Tel. +358 29 503 0000 • Y-tunnus / FO-nummer / Business ID: 0244680-7

GEOLOGIAN TUTKIMUSKESKUS

30.01.2017

GEOLOGIAN TUTKIMUSKESKUS **KUVAILULEHTI**

17.01.2017 / 8/2017

Tekijät Johanna Torppa Maarit Middleton	Raportin laji Arkistoraportti Toimeksiantaja
Raportin nimi Biogeochemical data analysis methods and R implementation in the UltraLIM project (UltraLIM projektin biogeokemiallisen datan analysointimenetelmät ja R implementaatio)	
Tiivistelmä Tämä raportti kuvaa UltraLIM projektissa käytetyt laatukontrolli- ja aineiston analyysimenetelmät. Aineisto koostuu kasvinäytteiden kemiallisista analyyseistä kuudelta eri tutkimuskohteelta, joilla tiedetään olevan mineralisaatioita. Menetelmät on implementoitu R ohjelmointiympäristössä. Lähdekoodi, koodin kuvaus, tutkimusaineistot sekä UltraLIM projektin tulokset ovat saatavilla zip-paketissa UltraLIM_biogeochem.zip.	
Asiasanat (kohde, menetelmät jne.) biogeokemia, prospektiivisuus, tilastotiede, klusterointi, implementointi, laadunvalvonta	
Maantieteellinen alue (maa, lääni, kunta, kylä, esiintymä) Lappi ja Pohjois-Pohjanmaa Kuusamo, Kittilä, Sodankylä, Pelkosenniemi, Tervola Hakokodanmaa, Juomasuo, Saivel, Kyörtesselkä, Kersilö, Vähäjoki	
Karttalehdet	
Muut tiedot Julkaisu koostuu tästä raportista sekä liitetiedostoista paketissa UltraLIM_biogeochem.zip	
Arkistosarjan nimi *Arkistoraportti*	Arkistotunnus 8/2017
Kokonaissivumäärä 31 / Kieli Englanti	Hinta / Julkisuus Julkinen
Yksikkö ja vastuualue MIM (Mineraalitalous ja malmigeologia)	Hanketunnus 50404-40045
Allekirjoitus/nimen selvennys Johanna Torppa	Allekirjoitus/nimen selvennys Maarit Middleton

GEOLOGIAN TUTKIMUSKESKUS

30.01.2017

GEOLOGICAL SURVEY OF FINLAND **DOCUMENTATION PAGE**

30.01.2017 / 8/2017

Authors Johanna Torppa Maarit Middleton	Type of report GTK archive report Commissioned by
Title of the report UltraLIM biogeochemical data analysis methods and R implementation	
Summary This report describes the methods of quality control and data analysis used in the UltraLIM project. Data consists of chemical analysis of plant samples from six study sites known to host mineralizations. The methods have been implemented in the R statistical software, and the source code is provided in the UltraLIM_biogeochem.zip archive along with the results for UltraLIM data.	
Keywords biogeochemistry, prospectivity, statistics, clustering, implementation, quality control	
Geographical area Lapland and Pohjois-Pohjanmaa. Hakokodanmaa, Juomasuo, Saivel, Kyörtesselkä, Kersilö, Vähäjoki. Kuusamo, Kittilä, Sodankylä, Pelkosenniemi, Tervola.	
Map sheet	
Other information The publication consists of this pdf report and a number of attachments in archive UltraLIM_biogeochem.zip	

Report serial *GTK archive report*		Archive code 8/2017	
Total number of pages 31	Language English	Price	Publicity Public
Strategic unit MIM (Ore Geology and Mineral Economics)		Project code 50404-40045	
Signature  Johanna Torppa		Signature  Maarit Middleton	

GEOLOGIAN TUTKIMUSKESKUS

30.01.2017

Sisällysluettelo

Kuvailulehti

1 INTRODUCTION

The aim of the 'Ultra low-impact exploration methods in the subarctic – UltraLIM' project was to increase confidence in the analysis of the biogeochemical datasets in order to apply them for prospecting of blind mineralizations in the future. The practical goals were to create guidelines for sampling of soils, plants and snow and to develop analytical methods as well as data analysis for target-scale mineral exploration. The goals were reached by conducting an orientation survey on six well studied mineralizations in northern Finland: two Au mineralizations (Juomasuo and Hakokodanmaa), three base metal mineralizations (Kersilö, Vähäjoki and Saivel) and a P-REE mineralization (Kyörtesselkä). Soil for weak and selective leaches and soil gas hydrocarbon analysis, plant bark and foliage for biogeochemical analysis and snow for element concentrations and soil gas hydrocarbon analysis were sampled. In addition, selective leach analytics of peat was tested on two sites. The geochemical response on the soil, plant organs and snow was interpreted with the help of drill core lithogeochemistry, base of till geochemistry and surface projection of the underlying mineralized zones.

The UltraLIM project was funded from the Green Mining Programme (2011-2016) by Tekes, the Finnish Funding Agency for Innovation. The project was conducted in years 2013-2015. It was managed by the Geological Survey of Finland (Northern Finland Office) with University of Oulu (Oulu Mining School) participating as a project partner. Following companies and agencies also provided funding to the project: Agnico Eagle Finland Oy, AA Sakatti Mining Oy, Dragon Mining Oy, First Quantum Minerals Ltd, Pyhäsalmi Mine Oy and Finnish Forests and Parks Service. Scandinavian GeoPool participated by providing working time. International expertise was provided by Canadian consulting companies and agencies: Heberlein Geoconsulting, Colin Dunn Consulting Ltd. and Geological Survey of Canada.

In this report, we present methods and tools developed in the UltraLIM project for the data quality control (QC) and for analyzing and visualizing biogeochemical sample data. The tools were implemented in the free statistical and graphics language R. The report is part of the UltraLIM project archive *UltraLIM_biogeochem.zip*, which also contains the source code as well as the UltraLIM plant data files and the results of QC and data analysis. A summary of the plant data is given in the archive file *readme_UltraLIM_bio.txt*. A thorough discussion on the UltraLIM project's results and their application to mineral prospectivity mapping in the future is provided in the final report of the UltraLIM project (Middleton et al., 2017).

In Sec. 2 we present the QC procedure, in Sec 3 the statistical analysis workflow and in Sec 4 the procedure for the spatial representation of the results.

2 DATA QUALITY CONTROL

We evaluated the data uncertainty originating from field sampling, sample preparation and laboratory analyses mostly following Reimann et al. (2008), and abbreviate the procedure as QC (quality control). In this paper, we describe the general workflow, while the source code is

in directory /src/QC. UltraLIM project's results are provided in *UltraLIM_biogeochem.zip* archive directory /QualityControl.

We have divided QC into five steps, each having its own script, as

QC 1 Checking the periodicity, blockiness and drift in the analysis sequence
QC 2 Defining the trueness (bias) and accuracy of the data using reference samples
QC 3 Defining the laboratory contamination using laboratory blank samples
QC 4 Checking the precision of the laboratory analysis using laboratory (pulp) duplicates
QC 5 Checking the precision of field sampling using field duplicates

In literature, the most often discussed quantities from QC are laboratory accuracy (QC2), laboratory precision (QC4) and field precision (QC5). These values are used to evaluate the uncertainty of the data, and the most significant sources of uncertainty.

Trueness (QC2) is often also relevant. However, in biogeochemical mineral exploration the spatial variation (i.e. anomaly patterns) is usually of interest rather than the absolute concentration levels and the bias can be disregarded if it is constant for the entire data set.

In an optimal case, reference samples should be of the same material as the project samples, but this is rarely possible. However, if the external reference concentrations are of high quality, comparing the distributions of external and project reference concentrations provides an estimate of the project's laboratory analysis quality. In this report, will use the following terminology when referring to reference sample concentrations:

External reference concentrations	Concentrations of the reference samples analysed externally in different sample batches
Project reference concentrations	Concentrations of the reference samples analysed in the project sample sequence
External reference sample	Reference samples collected from other locations than the project sampling sites and inserted to be analysed with the project analysis sequence
Project reference sample	Reference samples collected specifically for the project and inserted in the project sample analysis sequence

2.1 Source code prerequisites and I/O

The source code is written in R, and requires installation of R with packages *tcltk, tcltk2* and *tools*.

The package consists of the main executable script *QC.r* and the following functions in the /src/QC/Functions-folder: data input script *DataIn.r,* script *QC_choose.r* for choosing the QC steps to run, and separate scripts *QC1.r*, *QC2.r*, *QC3.r*, *QC4.r* and *QC5.r* for each QC step.

The following items should be noted when using the scripts for other projects:

- The *DataIn.r* script (Sec 3.1.2) is written for UltraLIM data specifically, and should be replaced if the data structure of format is different.
- External reference concentrations file should follow the format given in Sec 2.1.1 (*/AshControl.csv*)
- Reference sample identification file should follow the format given in Sec 2.1.5 (e.g., */2013/ReferenceSamples_2013.csv*)
- QC2 script, and the related steps in QC script should be adjusted to the reference sample selection for each project.
- The element concentrations in different datasets have to represent similar samples, i.e., expressed either in concentrations in ash or in dry tissue, not mixed. In this study, we used concentrations in ash.

In the following sections, we describe the functionality of each script.

2.1.1 QC

QC.r is the main executable script that calls all functions *DataIn, QC,_choose, QC1*, *QC2*, *QC3*, *QC4* and *QC5*.

User input

As input, *QC* prompts for the

1. *Location of the source code files*, i.e., directory where the R scripts are located (directory browser). In the *UltraLIM_biogeochem.zip* archive, the directory is /src/QC.
2. *Output folder*, i.e., the directory, where the output is to be written (directory browser). Since QC is done for one data entity at a time, folders for each entity should be created prior to running QC. UltraLIM QC output is written in each year's folder /2013 and /2014.
3. *External reference concentrations file*. Samples from this particular reference material, used in QC2, should be included within the project data. Format of this file should be the same as in the UltraLIM's external reference concentrations file */AshControl.csv.*

Output files

There are no output files from the main script.

2.1.2 DataIn

Datain.r script reads the input data, and arranges it in the format used by the QC, data analysis and plotting procedures. Thus, in this procedure, QC should always be run prior to further data analysis. The *DataIn.r* script included in the package is built to read UltraLIM data provided in *UltraLIM_biogeochem.zip* in directories 2013/Data and 2014/Data.

User input

As input, *DataIn.r* prompts for the following files:

1. *Location of the data files,* i.e., location of the laboratory analysis data files. All the files in the given folder with a ".CSV suffix" are combined and analyzed as one dataset. Thus, the folder should contain all the data files of one entity and no other files with a ".CSV" suffix.

UltraLIM project's laboratory data files are in folders /2013/Data and /2014/Data. QC has to be run separately for each year's data set.

2. *Field data file*, which contains relevant information about the field samples. UltraLIM field data file is */FieldData_2013_2014.csv.*
3. *Reference sample identification file,* i.e., reference sample types in the laboratory data file. UltraLIM reference sample type files are */2013/ReferenceSamples_2013.csv* and */2014/ReferenceSamples_2014.csv.*

Function input

1. Path to output file *alldata.csv* = <output folder>+"alldata.csv"
2. Path to output file *elInfo.csv* = <output folder>+"elInfo.csv"

Function output

If the *DataIn.r* script is replaced, it is important that the output of the new script coincides with the output defined below.

DataIn returns a list, with elements defined in the table below.

List element	Description	Data type	Variable size
1, *ndat*	Total number of analysed samples	Int	Scalar
2, *nel*	Number of analysed chemical elements	Int	Scalar
3	Data file ID for each sample (dataset may be divided into multiple files)	Int	[*ndat*]
4	Data type for each sample 1 = field sample without duplicate 10 = field duplicate 11 = field sample with duplicate 12 = project reference sample 22 = blank sample 31 = pulp duplicates first pulp 32=pulp duplicates second pulp -9999 = other samples, undefined sample types	Int	[*ndat*]
5	Analysed chemical elements in alphabetical order	String	[*nel*]
6	Units of chemical element concentrations (PPM / PPB / %), arranged in agreement with list element 5.	String	[*nel*]
7	Concentrations matrix, columns with chemical elements arranged in agreement with list element 5.	Float	[*ndat,nel*]
8	Lower analysis detection limit, arranged in agreement with list element 5.	Float	[*nel*]
9	Higher analysis detection limit, arranged in agreement with list element 5.	Float	[*nel*]

10	Project reference sample types	String	[number_of_ reference_samples]

Output files

Files *alldata.csv* and *elInfo.csv* are output in the output directory provided as input to QC. The structure of the files should be as described below.

alldata.csv file consists of a header row and a number of data rows, each of which corresponds to one field sample. Columns of the matrix are as follows:

Column	Description	Data type
1	Sampling year	Int
2	Sampling line	Int
3	Order, in which the samples on a line were collected: the first sample is 1, the second is 2, etc.	Int
4	Data point number, unique for the entire data set	Int
5	Sampling area ID	Int
6	Species ID	Int
7	Organ ID	Int
8	Sample/duplicate	Int as 0=sample 1=duplicate
9	x coordinate of the sampling location	Float
10	y coordinate of the sampling location	Float
11	Data type for each sample 1 = field sample without duplicate 10 = field duplicate 11 = field sample with duplicate 12 = project reference sample 22 = blank sample 31 = pulp duplicates first pulp 32=pulp duplicates second pulp -9999 = other samples, undefined sample types	Int
12	Data file ID for each sample (dataset may be divided into multiple files)	Int
13	Sample type	String
14	Rec. weight	Float
15	Pre ashed weight	Float
16	Ashed weight	Float
17-	Element concentrations	Float

elInfo.csv file is a four-column matrix and consists of a header row, and a number of data rows, each of which corresponds to one element. Column contents are the following:

Column	1	2	3	4
Description	Element	Lower detection limit	Upper detection limit	Unit as % / PPB / PPM
Data type	String	Float	Float	String

2.1.3 QC_choose

The QC_choose function shows a user interface for selecting which QC steps to run.

User input

1. Selection of the QC steps (QC1-QC5) to run (checkboxes).

Function output

1. The indices of the QC steps to run

2.1.4 QC1

The first step of quality analysis is checking whether there is periodicity, drift or blockiness in the data. These features can originate from varying conditions in sampling, sample processing or laboratory analysis. To suppress the effect of sampling and sample processing, the samples should be randomized into the analysis sequence prior to delivering them to the laboratory. This way the possible artifacts deriving from inappropriate laboratory analysis procedures are directly revealed. Also natural geochemical outliers and outliers by erroneous assays can be detected from these graphs. In the UltraLIM project the samples were not randomized, and the QC result shows variation that is due to sampling at different sites and from different plant species and their organs.

Function input

1. Field sample concentrations
2. Analyzed element names
3. Output folder
4. Data file index of each data point
5. Unit of each element concentration
6. Lower detection limit of each element
7. Upper detection limit of each element

Output file

QualityControl/QC1/Data.pdf
In the QC1 result plot, the data are plotted in the order of analysis for visual inspection to find possible periodicity, drift and blockiness in the data. The lower and upper detection limits are shown on the plot. Also, if the dataset has been divided in multiple sets (multiple files), the limits of different sets are shown. Units shown are those used in the data file.

As can be seen in the UltraLIM result file, project data is varying abruptly due to switching from one sampling location or plant species to another, since the data was not randomized prior to delivery to the laboratory.

2.1.5 QC2

The second quality control step is to check the validity of the laboratory analyses by examining the concentration distributions of the reference samples. A quantity called *trueness* is defined as the difference between the levels (e.g., means) of the project reference concentrations and the external reference concentrations. *Accuracy* is the scatter (e.g., standard deviation) of the project reference concentrations. The format of the external reference sample identification file (input 1. below) should obey the format of the UltraLIM file, e.g., */2013/ReferenceSamples_2013.csv.* QC2 function input steps 3-5 and the related steps in QC script should be adjusted for each project's reference data set.

UltraLIM reference samples were prepared and sent directly to the laboratory by Colin Dunn Consulting Ltd. Samples were extracted from a large-volume of fully homogenized ashed sample material representing three different plants: ashed Eucalyptus leaves (Western Australia, Ash-1), pine bark (Central BC, Canada, Ash-2) and pine twigs (Ontario, Canada, V6a). External reference concentrations provided with the reference samples was compiled from analysis conducted in different laboratories. Reference samples were analyzed in the project sample analysis sequence in evenly distributed locations requested by sending an electronic sample list of all samples to the laboratory. In 2013, 4.9% (n=55) of all tissue samples (1112) and in 2014 10.5% (n=75) of total 715 samples were reference samples.

User input

1. External reference concentrations. UltraLIM external reference concentrations file is */AshControl.csv.*

Function input

1. Reference sample identification in the project sample set. UltraLIM identification files are */2013/ReferenceSamples_2013.csv* and */2014/ReferenceSamples_2014.csv.*
2. Output folder
3. Concentrations of reference samples of type ASH-1
4. Concentrations of reference samples of type ASH-2
5. Concentrations of reference samples of type V6a

6. Unit of each element concentration
7. Analyzed element names
8. Number of analyzed elements

Output files

*QualityControl/QC2/Ash*True.pdf,* where *=[1,2,3]
Trueness, or bias, is visualized in plots, where the project reference concentrations are plotted along with the lines showing the mean as well as 1σ and 2σ regions of the external reference concentrations. The three types of reference material, ASH-1, ASH-2 and V6a, are plotted separately and correspond to file name indices 1, 2 and 3, respectively.

*QualityControl/QC2/Ash*Acc.pdf,* where *=[1,2,3]
Accuracy is visualized in plots, where the project reference concentrations are plotted along with the lines showing their mean as well as 1σ and 2σ regions. The three types of reference material, ASH-1, ASH-2 and V6a, are plotted separately and correspond to file name indices 1, 2 and 3, respectively.

QualityControl/QC2/ControlStats.csv,* where *=[1,2,3]
The trueness and accuracy are summarized in tables with a header row, chemical elements arranged in subsequent rows, and columns containing

Column	Header name	Description
1	Element	Chemical element
2	RefMean	Mean of the external reference concentrations
3	RefSD	Standard deviation of the external reference concentrations
4	RefRSD	Relative standard deviation of the external reference concentrations (RefSD/RefMean). RefMean and RefSD are rounded to 10^{-2}, which causes the disagreement RefRSD ≠RefSD/RefMean in the table.
5	ProjMean	Mean of the project reference concentrations
6	ProjSD	Standard deviation of the project reference concentrations
7	ProjRSD	Relative standard deviation of the project reference concentrations (ProjSD/ProjMean). ProjMean and ProjSD are rounded to 10^{-2}, which causes the disagreement ProjRSD ≠ProjSD/ProjMean in the table
8	Bias	Relative deviation of the project reference concentrations mean from the external reference concentrations mean: |RefMean-ProjMean|/RefMean.

A separate table for each reference material, ASH-1, ASH-2 and V6a, is output, and correspond to file name indices 1, 2 and 3, respectively.

2.1.6 QC3

The third QC step is conducted to check if the laboratory procedures cause contamination in the analysis. This QC step is usually entirely conducted by the laboratory, which uses its own blank samples, clean of the analyzed elements. The significance of contamination depends on the concentration levels of project samples because small concentrations are more easily disturbed by contamination than high concentrations.

Function input

1. Blank sample concentrations
2. Field sample concentrations
3. Number of analyzed elements
4. Analyzed element names
5. Output folder
6. Unit of each element concentration

Output file

QualityControl/QC3/Kdistr.pdf
The effect of the laboratory contamination of each element is visualized using the kernel distribution (Gaussian function) of the blank-sample concentrations along with the mean and 1σ region of the field-sample concentrations. In optimal case the blank concentrations are zero.

2.1.7 QC4

The fourth step of QC checks the precision of the laboratory analysis by taking laboratory duplicates of selected samples, and checking how much the analyzed concentrations differ.

Function input

1. Concentrations of the first each duplicate
2. Concentrations of the second of each duplicate
3. Number of analyzed elements
4. Analyzed element names
5. Output folder
6. Unit of each element concentration
7. Lower detection limit of each element
8. Upper detection limit of each element

Output files

The effect of ICP-MS uncertainty is visualized using plots describing the percentual difference of two subsequent pulps from the same sample. Since different samples are used for each pulp pair, the absolute difference is divided by the mean of each pulp pair, not by the mean of the whole set of pulp duplicates, to get the percentual difference.

QualityControl/QC4/THPlot.pdf
Thomson and Howard plots (mean concentration of the pulp duplicate pair vs difference of the duplicate pulp concentrations) are shown for each element separately. Lines showing a relative difference of 10% and 20% are shown.

QualityControl/QC4/PDupDev.pdf
A plot showing the relative difference (difference divided by the mean of the pair) of the duplicate pulp concentrations for all the elements combined in one plot. Lines showing a relative difference of 10% and 20% are shown.

QualityControl/QC4/PulpDupTable.csv
A table of quantities describing the laboratory precision. Column contents are as follows:

Column title	Description
Element	The chemical element
Unit	Unit of the concentrations
LDL	Lower detection limit
UDL	Upper detection limit
meanRDev	Mean of the relative differences of all pulp pair concentrations
>50%	Percentage of the pulp duplicate pairs, for which the relative difference is >50%.

2.1.8 QC5

The fifth step of QC checks the precision of field sampling by taking two subsequent samples of the same plant, and checking how much the analyzed concentrations differ. In the UltraLIM project, field duplicates were collected at every 10th station.

Function input

1. Concentrations of the first samples of each duplicate
2. Concentrations of the second samples of each duplicate
3. Number of analyzed elements
4. Analyzed element names
5. Output folder
6. Unit of each element concentration

7. Lower detection limit of each element
8. Upper detection limit of each element

Output files

The effect of field sampling to the uncertainty is visualized using plots describing the percentual difference of two field samples taken successively from the same plant. Since sample pairs are taken from different locations, plants and plant parts, the absolute difference is divided by the mean of each sample pair, not by the mean of the whole set of field duplicates, to get the percentual difference.

QualityControl/QC5/THPlot.pdf
Thomson and Howard plots (mean concentration of the field duplicate pair vs difference of the field duplicate concentrations) are shown for each element separately. Lines showing a relative difference of 10% and 20% are shown.

QualityControl/QC5/PDupDev.pdf
A plot with the relative difference (difference divided by the mean of the pair) of the field duplicate concentrations for all the elements combined in one plot. Lines showing a relative difference of 10% and 20% are shown.

QualityControl/QC5/PulpDupTable.csv
A table of quantities describing the field precision. Column contents are as follows:

Column title	Description
Element	The chemical element
Unit	Unit of the concentrations
LDL	Lower detection limit
UDL	Upper detection limit
meanRDev	Mean of the relative differences of all field duplicate pairs
>50%	Percentage of the field duplicate pairs, for which the relative difference is >50%.

3 STATISTICS

The UltraLIM plant samples were analyzed for 64 chemical elements, and represent different tissue types (organs) of various plant species from six different sampling locations. Exploratory data analysis (EDA) methods are used to help finding possible spatial correlation between mineral deposits and element concentrations, and is a potential suite of data analysis techniques for detection of unknown mineralizations.

In this paper, we describe the general workflows of the statistical and visualization techniques that were used, while the UltraLIM results are provided in *UltraLIM_beiogeochem.zip* archive in directory /Statistics and the source code in directory /src/Statistics.

Input for the statistical calculations is read from the files *alldata.csv* and *elInfo.csv* that have been output from the *DataIn.r* script of the QC procedure (Sec 2.1.2). UltraLIM *alldata.csv* and *elInfo.csv* files can be found in the *UltraLIM_biogeochem.zip* archive root. Two input files can be provided corresponding to two different data sets that should be compared. In the UltraLIM project, we used data sets from two subsequent years.

3.1 Data preprocessing

In the original data files, concentrations are given in ash. For data analysis, all the concentrations are transformed to concentration in dry tissue using the ashed and dry tissue sample weights found in the *alldata*.csv file. The ratio of the ashed sample weight (w_a) and the dry tissue sample weight (w_d) is called the ash yield (y)

$$y = \frac{w_a}{w_d}$$

Dry tissue concentration (C') is obtained from ashed concentration (C) using the formula

$$C' = \frac{C}{y}$$

To clearly distinguish samples with concentrations below the lower detection limit (LDL) and above the upper detection limit (UDL), samples with values below LDL are given the constant values of

$$C'_{LDL} = \frac{LDL}{2 * min(\boldsymbol{y})} \qquad (1)$$

and values greater than UDL are given the constant values of

$$C_{UDL} = UDL * 1.2 * \max(\boldsymbol{y}), \qquad (2)$$

where vector $\boldsymbol{y}$ contains the ash yields of all the samples.

Total concentration of rare-Earth elements (REE) is generally given as the sum of the REE element concentrations. However, if the total REE concentration is very low, and there is a significant number of REE elements with <LDL concentration, total REE is addressed as being <LDL, and similar for >UDL values.

3.2 Boxplots

Boxplots are computed following Tukey's ideas for the length of the whiskers:

- the low-end whisker ends in the point that is within Q1-(Q3-Q1)*1.5

- the high-end whisker ends in the point that is within Q3+(Q3-Q1)*1.5

where Q1 and Q3 are the first and third quartiles, respectively.

3.3 Output file naming convention used in the UltraLIM project

Sampling parameters considered in grouping the data for the statistical computations are

1. location
2. species
3. organ
4. sampling year

In the case when separate output files are produced for each location, species and/or organ, specific acronyms for identifying the samples are provided in the UltraLIM output file names (defined in constants.r file):

1. Locations
    a. Juomasuo = Juom
    b. Hakokodanmaa = Hako
    c. Saivel = Saiv
    d. Kyörtesselkä = Kyört
    e. Kersilö = Kers
    f. Vähäjoki = Vähä
2. Species
    a. Rhododendron Tomentosum = RhoTom
    b. Picea Abies = PicAbi
    c. Pinus Sylvestris = PinSyl
    d. Juniperus Communis = JunCom
    e. Pleurozium Schreberi = PleSch
    f. Vaccinium Uliginosum = VacUli
    g. Salix Caprea = SalCap
    h. Betula Pubescens = BetPub
    i. Empetrum Nigrum = EmpNig
    j. Vaccinium Myrtillus = VacMyr
    k. Vaccinium Vitis-idaea = VacVit
3. Organ
    a. Leaf or needle = LeaNee
    b. Twig or stem = TwiSte
    c. Bark = Bar

File naming convention for different datasets, which in UltraLIM were the two data sets from years 2013 and 2014, is hardcoded, and a running integer index is assigned for each data set in the order of input. In UltraLIM, the indices for the years in the file names are

1. 2013 = 1
2. 2014 = 2

3.4 Source code I/O and prerequisites

The code is written in R, and requires installation of R with extra packages *tcltk*, tcltk2, *tools, fields* and *StatDA*.

The package consists of the main executable script *Statistics.r*, and the following functions in Functions-folder: *constants.r, reorder.r, Statistics_choose.r, organDistr.r, organPerElDistr.r, StatColl.r, doTable.r, MannWhitneyU.r, organLocDistr.r, corSets.r* and *RR_stat_all.r.*

Function *constants.r* contains information on the elements that are used for the analysis, location, species and organ names, location of the mineralizations etc. This function should be adapted to the problem and study areas at hand.

The different datasets that are compared may have a different number of analysed elements. Input data has to be consistent, and the input files *alldata.csv* and *elInfo.csv* have to be unified before re-running the program. If an element with no observations is added to a dataset, the data values should NOT be ==0 to avoid problems with logarithms, for instance. Also negative values are not allowed, nor are NaNs. Unity is a good choice.

3.4.1 Statistics.r

This is the main executable script that reads the input data and calls the functions.

User input

The following directories and files are prompted for when the *Statistics.r* script is executed:

1. *Location of the source code files*, i.e., the directory where the R scripts, are located.
2. *Location of the alldata.csv and elInfo.csv files*, i.e., the data and element information files created by the *DataIn.r* script of the QC-package. Two input file locations can be provided subsequently corresponding, for instance, to two different sampling years or other entities.
3. *Location of output files*, i.e., the folder where the output is to be written.

Output

There is no output from the main script, but all the results are produced in the functions.

3.4.2 constants.r

This function defines the sample types, mineralization location, plotting layout etc. All the functions included must be present with the given input and output parameters, even if the

contents of the functions are changed. The function provided in the *UltraLIM_biogeochem.zip* archive is specific for UltraLIM data.

Contains multiple functions, which are not described here.

3.4.3 reorder.r

Reorders the points using the x,y coordinates so that the subsequent points are always closest to each other.

3.4.4 Statistics_choose.r

The input checkboxes for choosing which statistical computations to run.

User input

1. Selection of the statistical computations to run (checkboxes).

Function output

1. The indices of the statistical computations to run.

3.4.5 BoxHisKer.r

This function computes concentration distributions (boxblots, histogram and kernel function) for data of one organ of one species combined from all the sampling areas. One element is plotted in a frame, and six elements are shown on one page.

Function input

1. Concentrations in dry tissue.
2. Output folder
3. Index of the input data file (multiple files can be given in 2. user input of *Statistics.r*)
4. Indices of plant species for each sample
5. Indices of plant organs for each sample
6. Column names of the *alldata.csv* table
7. Units of element concentrations
8. Lower detection limits multiplied by mean(***y***) (mean ash yield) of the samples
9. Upper detection limits multiplied by mean(***y***) (mean ash yield) of the samples

Output files

The asterisks in file names represent the dataset index and acronyms of the plant and organ (Sec 3.3).

*Statistics/BoxHisKer/box*_*_*.pdf*
Boxplot distribution.

*Statistics/ BoxHisKer /ker*_*_*.pdf*
Gaussian kernel fitted distribution.

*Statistics/ BoxHisKer /his*_*_*.pdf*
Histogram distribution.

3.4.6 BoxOneDataset.r

This function generates boxplots of data distributions for one organ of one species. All the boxplots of for element are shown in two subsequent frames so that the first frame shows the distributions of data combined from all the sampling areas and the second separately for each sampling area. The number of data points for each organ are indicated in the plots.

Separate documents are provided for concentration, logarithm of concentration, response ratio and logarithm of response ratio

Function input

1. Concentrations in dry tissue.
2. Output folder
3. Index of the input data file (multiple files can be given in 2. user input of *Statistics.r*)
4. Indices of plant species for each sample
5. Indices of plant organs for each sample
6. Indices of sampling locations for each sample
7. Column names of the *alldata.csv* table
8. Units of element concentrations
9. Index telling, which form of data to use (concentration, logarithm of concentration, response ratio and logarithm of response ratio)

Output file

The asterisks in file name represents the form of the data (1=concentration, 2=response ratio, 3=logarithm of concentration and 4=logarithm of response ratio) and dataset index.

*Statistics/BoxOneDataset/*_set*.pdf*

3.4.7 StatColl.r

This function generates histogram and box-, ECDF- and CP-plots on one page for each element. Statistics are calculated for each organ of each plant separately, and both for combined locations and for each location separately. Number of data points in each category is indicated in the plots.

Function input

1. Concentrations in dry tissue/ashed samples.
2. Index of input 1 representing ashed (=1) or dry tissue (=2).
3. Output folder
4. Index of the input data file (multiple files can be given in 2. user input of *Statistics.r*)
5. Indices of plant species for each sample
6. Indices of plant organs for each sample
7. Indices of sampling locations for each sample
8. Column names of the *alldata.csv* table
9. Units of element concentrations

Output files
The asterisks in file names represent the dataset index and the acronyms of the plant, organ and location (Sec 3.3).

*Statistics/StatColl/org/ set*_*_*.pdf*
For dry tissue concentrations with all sampling areas combined.

*Statistics/StatColl/orgLoc/ set*_*_*_*.pdf*
For dry tissue concentrations, and each sampling location separately.

*Statistics/StatColl/Ashorg/set*_*_*.pdf*
For ashed sample concentrations with all sampling areas are combined.

*Statistics/StatColl/AshOrgLoc/ set*_*_*_*.pdf*
For ashed sample concentrations, and each sampling location separately.

3.4.8 doTable.r
This function creates three tables for the statistics by location, species, and organ.

Function input
1. Concentrations in dry tissue.
2. Concentrations in ashed samples.
3. Output folder
4. Index of the input data file (multiple files can be given in 2. user input of *Statistics.r*)
5. Indices of plant species for each sample
6. Indices of plant organs for each sample
7. Indices of sampling locations for each sample
8. Lower detection limits for the elements
9. Upper detection limits for the elements
10. Column names of the *alldata.csv* table
11. Units of element concentrations

Output files
Table structure of each output file is shown below. The total number of samples used indicated in the first cell, i.e., “Element,N”, where N is the number of samples. Values are rounded to four decimal places, and for further use original numbers have to be retrieved.

Element, N	Element and total number of samples used for statistics computation
%<LDL	Percentage of samples below lower detection limit
N>LDL	Number of samples below lower detection limit
%<UDL	Percentage of samples above upper detection limit
N>UDL	Number of samples above upper detection limit
Min	Minimum concentration in dry tissue

Q0.05	0.05 quantile value of the dry tissue concentrations
Q0.25	0.25 quantile (1. quartile) value of the dry tissue concentrations
Mean	Mean value of the dry tissue concentrations
Median	Median value of the dry tissue concentrations
Q0.75	0.75 quantile (3. quartile) value of the dry tissue concentrations
Q0.87	0.87 quantile value of the dry tissue concentrations
Q0.93	0.93 quantile value of the dry tissue concentrations
Q0.95	0.95 quantile value of the dry tissue concentrations
Q0.96	0.96 quantile value of the dry tissue concentrations
Max	Maximum concentration in dry tissue
Range	Max-Min
StD	Standard deviation
RDS	Relative standard deviation (StD/Mean*100)
MAD	Median absolute deviation
CVR%	Coefficient of variation (MAD/Median)
IQR	Inter-quartile range
Unit	Unit of element concentration
LDL	Lower detection limit
UDL	Upper detection limit

The asterisks in file names represent the dataset index and the acronyms of the plant, organ and location (Sec 3.3).

*Statistics/StatTables/org/set*_*_*_*.csv*
Table of statistical properties of a single organ from all the sampling areas. Elements are organized in groups as: 1. Major elements (K, Ca, Mg, P, S, Mn), 2. Minor elements (Zn, Fe, B, Cu, Co), 3. Other elements and 4. Typically below detection limit.

*Statistics/StatTables/orgLoc/set*_*_*_*.csv*
Table of statistical properties of a single organ from each sampling area separately.

Statistics/StatTables/statAll.csv*
Table of statistical properties with data from all organs, species and location combined.

3.4.9 MannWhitneyU.r

This function performs the Mann-Whitney test for defining whether the concentrations above the mineralization deviate from the background concentrations. We provide the *p-value*, that can be used to evaluate the probability with which distributions represent the same

populations. *p*-value is derived from the Mann-Whitney test statistic. In the following description, distribution A represents the concentrations above the mineralization and distribution B the background values. The test can be only used when a priori knowledge of the underlying mineralizations is available from bedrock drillings. Results for each sampling locations are given in separate files and the asterisks in the file names represent the dataset index and the acronym of the sampling location (Sec 3.3). Results for each plant species and organ are provided in separate columns.

Statistics/MannWhitney/MWGreater.csv*
Table of MannWhitney-testin *p-values* with null hypothesis “distributions A and B represent the same population” and alternative ”distribution of population A represents higher values than distribution B”.

Statistics/MannWhitney/MWLess.csv*
Table of MannWhitney-testin *p-values* with null hypothesis “distributions A and B represent the same population” and alternative ”distribution of population A represents lower values than distribution B”.

Statistics/MannWhitney/MWTwoSided.csv*
Table of MannWhitney-testin *p-values* with null hypothesis “distributions A and B represent the same population” and alternative ”distribution of population A represents different values than distribution B”.

3.4.10 organLocDistr.r

This function generates boxplots of concentration and response ratio distributions for all the locations, species and organs arranged and coloured according to location, organ and species. Two data sets (two subsequent years in UltraLIM) are compared side by side.

Function input

1. Concentrations in dry tissue.
2. Output folder
3. Index to whether to use response ratios (=1) or concentrations (=2)
4. Indices of plant species for each sample
5. Indices of plant organs for each sample
6. Indices of sampling locations for each sample
7. Column names of the *alldata.csv* table
8. Units of element concentrations

Output files

Boxplots of all the locations, species and organs arranged and coloured according to location, species and organ. Asterisk in the file names refers to location *=*Loc*, organ *=*Org*, and species *=*Spe*. Two observing sets (years) side by side.

Statistics/OrganLocDistr/Ccol.pdf*
Concentrations in dry tissue.

Statistics/OrganLocDistr/RRcol.pdf*
Response ratios of dry tissue.

3.4.11 corSets.r

This function calculates Spearman correlation and 2D plots for visual correlation of data from two different observing sets (two years). Separate plots for each location, species and organ combination. Nine elements on one page. Asterisks in the file names are acronyms of the location, plant and organ (Sec 3.3).

Function input

1. Concentrations in dry tissue.
2. Output folder
3. Indices of plant species for each sample
4. Indices of plant organs for each sample
5. Indices of sampling locations for each sample
6. Column names of the *alldata.csv* table
7. Units of element concentrations
8. Sampling point ID numbers

Output files

*Statistics/Correlation/****.pdf*

3.4.12 RR_stat_all.r

This function generates a table containing the response ratios for the entire dataset (both sampling years in the case of UltraLIM) as well as a table containing response ratio statistics.

Function input

1. Concentrations in dry tissue.
2. Output folder
3. Index of the input data file (multiple files can be given in 2. user input of *Statistics.r*)
4. Indices of plant species for each sample
5. Indices of plant organs for each sample
6. Indices of sampling locations for each sample
7. Column names of the *alldata.csv* table
8. Units of element concentrations

Output files

Statistics/RRStat/RRall.csv
Table containing the response ratios for the entire dataset. Columns represent the elements and rows contain the samples, all in the same order as in *alldata.csv*.

Statistics/RRstat/RRstat.csv
Table containing the mean, median, minimum and maximum response ratios for the entire dataset. Elements are arranged in rows.

4 LINEPLOTS

Data is spatially visualized, showing the spatial variation of concentrations or response ratios for each analysed chemical element. Uni- and multivariate analysis is used to assign biogeochemical data into natural clusters. In the following, the directories containing different representations of spatial plots are listed with summaries of file contents.

Samples with values <LDL and >UDL are treated as described in Sec 3.1.

In this paper, we describe the general workflows, while the UltraLIM results are provided in *UltraLIM_biogeochem.zip* archive in directory /Lineplots and the source code in directory /src/Lineplots.

Input for the calculations is read from the files *alldata.csv* and *elInfo.csv* that have been output from the *DataIn.r* script of the QC procedure (Sec 2.1.2). UltraLIM *alldata.csv* and *elInfo.csv* files can be found in the *UltraLIM_biogeochem.zip* archive root.

4.1 Source code I/O and prerequisites

The code is written in R, and requires installation of R with extra packages *tcltk*, tcltk2, *tools, fields, StatDA, jpeg* and *mvoutlier*.

The package consists of the main executable script *Lineplots.r*, and the following functions in Functions-folder: *constants.r, reorder.r, Lineplots_choose.r, writeCSV.r, plotline.r, plotLineCL.r, plotLineOL.r* and *publishPic.r.*

Function *constants.r* contains information on the elements that are used for the analysis, location, species and organ names, location of the mineralizations etc. This function should be adapted to the problem and study areas at hand.

Functions *plotLineCL.r* and *plotLineOL.r* are adapted to the example input data in terms of the plot layout etc.

4.1.1 Lineplots.r

Main script that reads the input data and calls the functions.

User input

1. *Location of the source code files*, i.e., the directory where the R scripts, are located.
2. *Location of the alldata.csv and elInfo.csv files*, i.e., the data and element information files created by the *DataIn.r* script of the QC-package. Multiple locations can be provided.

3. *Location of output files*, i.e., the folder where the output is to be written.
4. *Field duplicate file* for each data set from quality control

Output

There is no output from the main script, but all the results are produced in the functions.

4.1.2 constants.r

This function defines the sample types, mineralization location, plotting layout etc. All the included functions must exist with the given input and output parameters, even if the contents of the functions are changed. The function provided in the *UltraLIM_biogeochem.zip* archive is specific for the UltraLIM data.

Contains multiple functions, which are not described here.

4.1.3 reorder.r

Reorders the points using the x,y coordinates so that the subsequent points are always closest to each other.

4.1.4 Lineplots_choose.r

The input checkboxes for choosing which functions to run.

User input

Selection of the spatial plots to generate (checkboxes).

1. Write CSV data file (*writeCSV.r*)
2. Lineplots, clustering by single-element quantiles (*plotLine.r*)
3. Lineplots,clustering by SOM-Kmeans (*plotLineCL.r*)
4. Lineplots, outlier analysis (*plotLineOL.r*)
5. Plots for publication (*publishPic.r*)

Function output

1. The indices of the spatial plots to generate.
2.

4.1.5 writeCSV.r

Transforms the input data file to a CSV file that can be input to other software, for instance, ArcGIS and SiroSOM.

Function input

1. Output folder
2. Concentrations in dry tissue.
3. *x* coordinates of the sampling points
4. *y* coordinates of the sampling points
5. Order in which the samples were collected
6. Indices of sampling locations for each sample

7. Indices of plant species for each sample
8. Indices of plant organs for each sample
9. Units of element concentrations
10. Column names of the *alldata.csv* table
11. Lower detection limits of element concentrations
12. Upper detection limits of element concentrations
13. Point IDs of sampling points

Output files

Asterisks in the file names are the index of the input data set and the acronym of the location (Sec 3.3).

*Lineplots/DataTables/allEl*_**

The entire data set in separate CSV files for each location. All the elements are included. Samples are arranged in rows and columns contain:

Column	1	2	3	4	5	6	7 -
Description	Species index	Organ index	Point order number	Point number	x-coordiinate	y-coordinate	Elements

*Lineplots/DataTables/forMV_****

Data in separate CSV files for each location, species and organ. Only the elements chosen for multivariate analysis are included. Samples are arranged in rows and columns contain:

Column	1	2	3 -
Description	x-coordiinate	y-coordinate	Elements

4.1.6 plotLine.r

This function plots the concentrations for each sampling line, points separated by their distance. Points are coloured by concentration quantiles defined for each element separately.

Function input

1. Output folder
2. Concentrations in dry tissue.
3. Units of element concentrations
4. Column names of the *alldata.csv* table
5. *x* coordinates of the sampling points
6. *y* coordinates of the sampling points
7. Indices of plant species for each sample
8. Indices of plant organs for each sample
9. Indices of sampling locations for each sample

10. Sampling line index
11. Point IDs of sampling points
12. Index of plotted quantity (response ratio/concentration)

Output files

Asterisks in the file names are the index of the input data set and the acronym of the location (Sec 3.3).

*Lineplots/Plotline/**plotlines.pdf*

Lineplot for each location, species, organ and element separately. Colouring of the points by quantiles.

4.1.7 plotLineCL.r

This function plots the concentrations for each sampling line, points separated by their distance. Points are coloured by SOM-Kmeans clusters. SOM clusters must be computed separately using external software, and the SOM cluster file names and table structure must conform to the format of UltraLIM SOM cluster data in `/Lineplots/Input/SOM`., i.e., file names must contain the string with the acronyms for the location, species and organ.

User input

1. Location of the SOM cluster files (for one data set at a time)

Function input

1. Output folder
2. Index of the data set
3. Concentrations in dry tissue.
4. Units of element concentrations
5. Column names of the *alldata.csv* table
6. *x* coordinates of the sampling points
7. *y* coordinates of the sampling points
8. Indices of plant species for each sample
9. Indices of plant organs for each sample
10. Indices of sampling locations for each sample
11. Sampling line index
12. Point IDs of sampling points
13. Field duplicate file
14. SOM cluster file
15. Index of plotted quantity (response ratio/concentration)

Output files

Asterisks in the output file names below are the acronyms of the location, species and organ (Sec 3.3) and the data set index.

*Lineplots/PlotLine/Clusters/****SOM.pdf*

Concentrations plotted along sampling lines.

*Lineplots/PlotLine/Clusters/****SOM_boxes.pdf*

Boxplots of the distributions of each element within each cluster

4.1.8 plotLineOL.r

This function plots the concentrations for each sampling line, points separated by their distance. Points are coloured by outlier/non outlier status based on the Mahalanobis distance. Lineplots are computed for one location, species and organ at a time.

Part of the data sets produce singular matrices, and in these cases the outliers cannot be computed.

Function input

1. Output folder
2. Index of the data set
3. Concentrations in dry tissue.
4. Units of element concentrations
5. Column names of the *alldata.csv* table
6. *x* coordinates of the sampling points
7. *y* coordinates of the sampling points
8. Indices of plant species for each sample
9. Indices of plant organs for each sample
10. Indices of sampling locations for each sample
11. Sampling line index
12. Point IDs of sampling points
13. Field duplicate file
14. Index of plotted quantity (response ratio/concentration)

Output files

Asterisks in the file names are the acronyms of the location, species and organ (Sec 3.3) and the index of the data set.

*Lineplots/Outliers/Plotline/****OL.pdf*

5 FUTURE WORK

- automated anomaly detection
  - moving window mean/std
  - anomaly amplitude vs data uncertainty
- Evaluating the field sampling and laboratory uncertainty by taking lab duplicates from field duplicates -> ANOVA
- add lower detection limit to blank sample plots
- How to evaluate the significance of the effect of the lab contamination

6 REFERENCES

Middleton, M., Sarala, P., Taivalkoski, A., Torppa, J., Kyllönen, E., Lahaye, Y., Lukkari, S., Peuraniemi, V., Pietikäinen, K, Rekilä, J., Rönnqvist, J., and Sutinen, R., 2017. Ultra low impact geochemical exploration methods in the sub-arctic. Report of Investigations, Geological Survey of Finland (in preparation).

Reimann C., Flizmoser P., Garret R. and Dutter R., 2008. Statistical data analysis explained, John Wiley & Sons, Ltd.