GEOLOGICAL SURVEY OF CANADA
OPEN FILE 7944

Geochemical Data Quality Control for Soil, Till and Lake and Stream Sediment Samples

M.W. McCurdy, R.G. Garrett

2016

Canada

GEOLOGICAL SURVEY OF CANADA
OPEN FILE 7944

Geochemical Data Quality Control for Soil, Till and Lake and Stream Sediment Samples

M.W. McCurdy, R.G. Garrett

2016

Recommended citation
McCurdy, M.W., Garrett, R.G., 2016. Geochemical Data Quality Control for Soil, Till and Lake and Stream Sediment Samples; Geological Survey of Canada, Open File 7944, 40 p. doi:10.4095/297562

Table of Contents

Table of Figures

Tables

Geochemical Data Quality Control for Soil, Till and Lake and Stream Sediment Samples

Summary

Geochemical data for lake and stream sediments, soils and tills received from commercial and government laboratories need to be evaluated for quality in terms of accuracy and precision before release. Quality control (QC) of National Geochemical Reconnaissance data is carried out using Microsoft Excel® with the Open Source R Project for Statistical Computing software (R Core Team, 2015) and 'rgr', 'The GSC Applied Geochemistry EDA Package' (Garrett 2013a, 2015a), that sits on top of R and provides the functionality required for many QA/QC tasks. Routines in 'rgr' are used to evaluate small (<20 samples) using graphical methods and larger batches using both graphical and statistical methods. Accuracy is estimated by inserting Certified Reference Materials (CRM) with accepted mean concentrations determined by different analytical methods. Precision is estimated using analytical duplicate pairs. For large batches of samples (>30 pairs of observations) an analysis of variance (ANOVA) is carried out using field duplicate pairs.

This document represents the first step in a process with the goal of building a consistent methodology for evaluating the quality of geochemical analytical data before they are released to the public. For the most part, the procedures described below, especially the construction of tables for Accuracy, Precision and ANOVA, are both time-consuming and tedious. However, there are a number of open source programs available that can be used in conjunction with R and Microsoft Excel® that could expedite the generation of statistics and the process of filling in the rows and columns of the tables. These include RExcel, an add-in to Excel that allows the use of R as a 'helper application' for Excel (Heiberger and Neuwirth, 2009). RStudio and Rcmdr are menu interfaces for R, also free and Open Source that could be integrated with R and Excel to create a system for evaluating the quality of geochemical data in a timely and efficient manner.

The reader is advised that the GSC has never had a single institutional approach to geochemical quality control. The procedures described herein have been developed by one group at one location (601 Booth Street, Ottawa). Other groups have instituted other procedures.

Introduction

The goal of this document is to outline the graphical and statistical procedures employed at the Geological Survey of Canada to estimate the quality of lake and stream sediment geochemical data returned from commercial and government laboratories. The accuracy and precision of geochemical data returned from laboratories must be estimated prior to publishing or using these data for research. Incorporating Certified Reference Materials (CRMs), field and analytical duplicate pairs in batches of samples submitted to laboratories provides data for univariate statistical routines used to estimate accuracy, precision and variance. These routines are available in Microsoft Excel®, a spreadsheet, and R, an integrated suite of software facilities for data manipulation, calculation and graphical display

(http://www.r-project.org/). Similar procedures have more recently been applied to the estimation of soil and till data quality.

Evaluating Data Quality

Planning the Survey

The first step in estimating the quality of data from a geochemical survey begins at the design (or planning) stage before samples are collected and analysed, and is usually initiated by a request for budget estimates from project leaders. The cost of collecting and analysing field duplicate samples must be taken into account when estimating the cost of a geochemical survey. Additional funds are required for the analysis of different types of quality control samples that are incorporated into sample batches during preparation and described below and in **Appendix 1**.

Collecting Samples and Preparing a Batch for Analysis

The next step in a strategy for quality control takes place in the field with the collection of duplicate samples at selected sites. After routine and field duplicate samples are dried and during preparation for analysis, suitable samples, preferably one of the field duplicate pair, are identified and split into analytical duplicate pairs. CRMs are inserted into pre-selected positions throughout the batch. A detailed description of the identification and placement of quality control samples is provided in **Appendix 1**: a brief summary is provided below.

A minimum of one aliquot of a suitable Certified Reference Material (**Appendix 1**) is included with each set of one or more samples sent to commercial or government laboratories. This additional sample is packaged and labelled in a manner indistinguishable from Routine samples and is placed in a random location within the sequence. When the number of samples to be analyzed in a study exceeds 20 (including CRMs and duplicates), another CRM is inserted into the sample sequence at a second random location in the following block of 20. A minimum percentage of CRM samples of 5%, or one standard for every 20 samples analysed, are included in a set of samples shipped to a laboratory for analysis. Additionally, each batch of 17 or less routine samples sent to laboratories for analysis includes at least one field duplicate and one analytical duplicate. These are also packaged and labelled similarly to the routine samples. For batches of 18 or more samples, the number of field duplicates and analytical duplicates that should be inserted increases as described in **Appendix 1**.

A spreadsheet table listing the sample numbers('Unique ID'), the National Topographic System map sheet ('NTS Map Sheet'), and type of sample ('Sample Type') must be prepared before the samples are shipped to the lab. This is most efficiently done by preparing a 'key' or master file in Excel® (Fig. 1) that lists, in sequence, all samples shipped to the lab, identifying the Routine samples, CRMs, Field and Analytical Duplicates and Blanks (if included). Separate columns are used to identify the Analytical Split ('AD Split'), the name of the CRM ('CRM ID') and the geographic coordinates and datum ('Latitude (NAD

83); Longitude (NAD 83)') of Routine samples. Only a sequential list of samples is forwarded to the laboratory to ensure that all samples are treated in the same manner.

	A	B	C	D	E	F
1	**Unique ID**	**Sample Type**	**AD Split**	**CRM ID**	**Latitude (NAD 83)**	**Longitude (NAD 83)**
2	021J122001	Analytical Duplicate	021J122005		46.33255	-67.09681
3	021J122002	Routine			46.29986	-67.03677
4	021J122003	Routine			46.28598	-67.00415
5	021J122004	Routine			46.33310	-67.09377
6	021J122005	Routine	021J122001		46.33255	-67.09681
7	021J122006	Routine			46.37011	-67.10072
8	021J122007	Routine (Field Duplicate Site)			46.37254	-67.03888
9	021J122008	Field Duplicate			46.37254	-67.03888
10	021J122009	Routine			46.37414	-67.02966
11	021J122010	Routine			46.33137	-67.01279
12	021J122011	Routine			46.37609	-67.02443
13	021J122012	Routine			46.32285	-67.01801
14	021J122013	Routine			46.29517	-67.03380
15	021J122014	Certified Reference Material		STSD-4		
16	021J122015	Routine			46.35563	-66.99816
17	021J122016	Routine			46.35492	-67.06438
18	021J122017	Routine			46.35583	-67.00738
19	021J122018	Routine			46.36296	-67.04429
20	021J122019	Routine			46.37388	-67.04022
21	021J122020	Certified Reference Material		STSD-2		
22						
23						
24						

Figure 1 A 'key' or 'master' spreadsheet table is prepared by the scientist before samples are shipped to a laboratory and is linked to analytical results returned from the laboratory. The example above is from a study in New Brunswick in 2012. Note that a second CRM has been included at the end of the sequence in addition to the minimum requirement of one CRM per 20 samples (5%). The analytical laboratory only receives the first column.

Formatting Results

Laboratories return analytical data in digital format, usually as a Microsoft Excel® file or in a format such as a 'Comma Separated Value' (csv) file that can easily be imported as a table into a spreadsheet program. As soon as data are received they are merged with a previously prepared spreadsheet table, the 'key' or 'master' table (Fig. 1) of sample numbers, sample type and location coordinates. Small batches of samples can be linked with 'cut and paste' commands (a procedure not without risk of mistakes) but for larger numbers of samples 'Lookup Assistant'®, an add-in product for Excel® has proven useful and effective, https://www.add-ins.com/lookup_assistant.htm. Once laboratory analyses

are linked to a sample type, i.e. Routine, Analytical Duplicate, Field Duplicate, Certified Reference Material, or Blank (if present), data associated with standards can be copied from the laboratory file and into a separate spreadsheet table ('worksheet') for comparison with accepted element concentrations for different analytical methods listed in Lynch (1990, 1996, 1999) and Hechler (2013).

A Microsoft Excel® spreadsheet table provided by Becquerel Laboratories for INAA data from a recent study of streams in New Brunswick (McClenaghan et al., 2015) is shown in Figure 2. Data for each sample are in rows identified by a 'Sample ID' and will be merged with a previously prepared digital key spreadsheet table (Fig. 1) to link Routine samples, Field and Analytical Duplicate samples and CRM positions within the sample suite.

Figure 2 An example of a laboratory spreadsheet table listing data for the concentrations of 35 elements in 20 samples of stream sediment submitted for INA analysis. Note that the Sample IDs are not in the same format as the key table and must be changed before merging. Note also that variables are not listed in alphabetical order.

Figure 3 shows the merged tables with all samples identified by type. The next step is to organize the analytical data into separate csv files for CRM data, Analytical Duplicate and Field Duplicate pairs, and Routine samples for input into R.

4

Figure 3 Tables in Figure 1 and Figure 2 are merged and data for each sample are linked with 'Unique ID' and 'Sample Type'. Columns of element variables are sorted alphabetically for ease of reading.

To create a spreadsheet table in Excel® for input to R, the table is edited so that the first row identifies the element or variable name and method of analysis with the associated geochemical data in the column below each element. Geochemical data for each sample are in rows beginning immediately below the header row. In Figure 4, each column containing element data is identified by the element symbol and the analytical method, separated by an underscore (e.g. 'Ag_INAA').Before exporting the worksheet as a csv file, any empty analytical data fields are replaced with 'NA' (not available).Values below the detection limit are indicated by a '<' followed by the detection limit (e.g. '<2').All occurrences of '<' must be replaced with a minus sign(e.g. '-2') before the data are exported to a csv file. See Appendix 4 for an explanation of how the minus detection limit data are handled in R. Occasionally the concentration of an element, usually Mn, is reported as greater than the detection limit (e.g., >10,000). In these cases, and where the percentage of samples above detection limit is less than 10 percent, the upper detection limit multiplied by a factor of 1.7 replaces the original value reported above detection limit. Thus, a value for Mn of >10,000 ppm would be replaced with 17,000 ppm. An explanation for the reasoning behind this can be found in Sanford et al. (1993).

Figure 4 The data table in Figure 3 has been edited in Excel® and exported as a csv file, ready for sub-setting and entry into R.

Once the data are merged, formatted and exported from Excel as a csv file, separate files for CRM data (Fig. 5), analytical duplicate data and field duplicate data are created by making copies and deleting rows not associated with each set of data. Whether there are two observations (rows) or 100,000 or more, data are organized, identified and sorted in the same fashion. Individual csv files can now be imported into R, where they are stored as data frames.

Figure 5 Analyses of Certified Reference Materials identified in the table in Figure 4 are used to estimate accuracy after being read into R and converted to a data frame. Columns of data not required for estimating accuracy have been deleted.

An example of the script used to import a csv file into R and convert it to an R data frame is provided in **Appendix 4**.

Evaluating Geochemical Data

In general, it has been the experience of the authors over many years of observation that the accuracy and precision of geochemical data returned from both commercial and government laboratories has been satisfactory. However, there have been occasional cases of contamination, sample mix-ups and uncompensated (or poorly or insufficiently compensated) spectra-, matrix- or physical interferences (**Appendix 3**). Contamination or sample mix-ups can happen during the collection, preparation or analysis stages; interferences occur during analysis. Problems such as sample order mix-ups, labelling errors and column shifts can usually be detected by careful examination and comparison of QA/QC sample data.

Accuracy

Accuracy of analytical results received by the GSC from commercial and government laboratories is estimated by comparing data obtained from laboratory analyses of CRMs included in batches of routine samples with accepted average reference values for these standard materials. For a discussion of Certified Reference Materials, see **Appendix 1**. See **Appendix 2** for a more detailed discussion of accuracy and other statistical terms used in this document. The following procedures were developed in the Applied Geochemistry Section, Northern Canada Division, at the Geological Survey of Canada.

For batches of samples with less than 30 analyses of any individual standard, using statistical methods to estimate accuracy may not be reliable because the values of the sample variance s^2 (**Appendix 2**) may fluctuate considerably from sample to sample if the population distribution is not normal (Walpole, 1982); however, historic CRM data from previous studies can be plotted with new observations using 'rgr' (**Appendix 4**) to provide some assurance of data accuracy. Using the function 'crm.plot.new' in 'rgr', analytical results, produced by using similar sample digestions, of a specific CRM from previous

6

projects have been compiled and are used to make graphical comparisons with more recent results plotted against accepted mean values (Fig. 6). See **Appendix 4** for an example in 'rgr': ('Evaluating the Accuracy of Data with Less than 30 Observations'). Although a statistical estimation of accuracy in the form of a Relative Standard Deviation (RSD %) is not calculated for batches of less than 30 observations, analytical results are plotted with data collected from previous GSC surveys against an accepted mean and, provided that concentrations fall within an acceptable range of variability about the mean (±2 standard deviations), are considered suitable for publication.
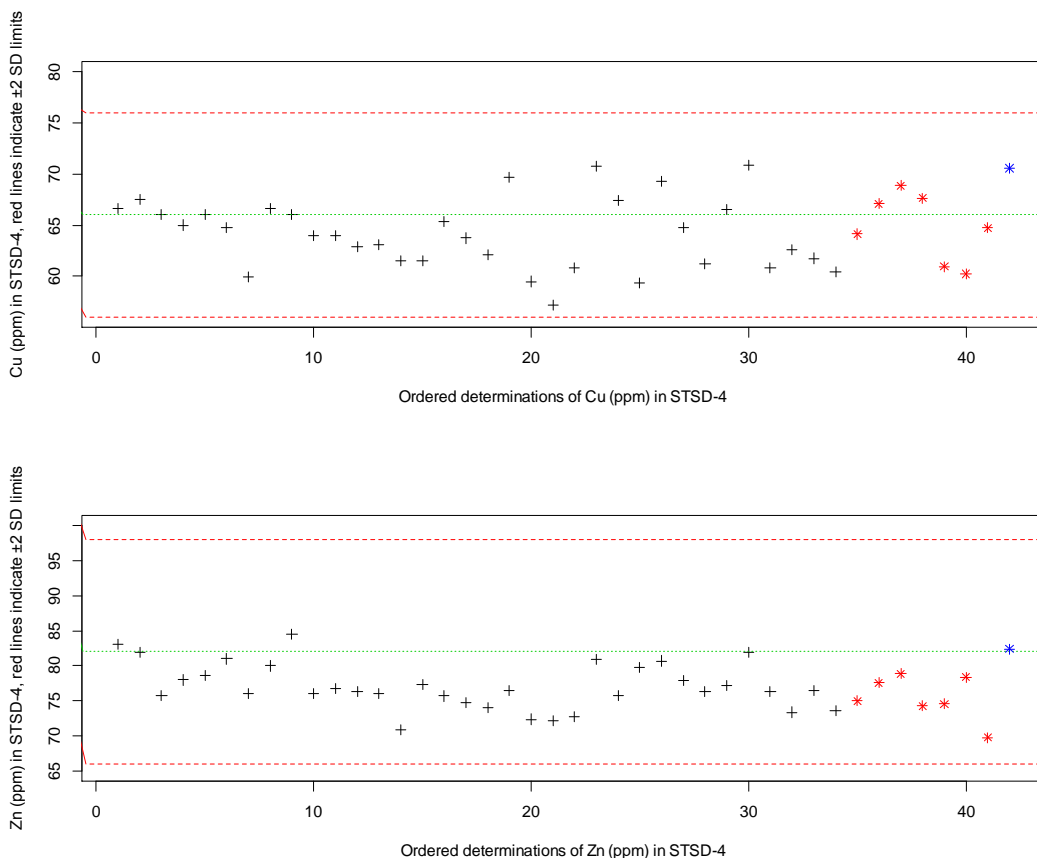
Figure 6 Time charts ('X-charts') for Cu and Zn in STSD-4, a stream sediment CRM analyzed by a 'partial' digestion method. The green lines represent the provisional elemental values from Lynch (1990; 1999) and the red lines mark two Standard Deviation limits about the Mean. Black crosses represent unpublished data from previous surveys, red asterisks are results from a recent survey on Victoria Island, NU (McCurdy, et al., 2012), and the blue asterisks are from a 2012 survey in New Brunswick (McClenaghan et al., in press). See Appendix 5 for a method to prepare charts using R with 'rgr' ('The GSC Applied Geochemistry EDA Package' (Garrett 2013a, 2015a)).

The accuracy of analytical results from batches of samples with 30 or more analyses of one or more standards can be statistically estimated from the mean and standard deviation using 'crm.plot' in 'rgr'. The syntax of the script used for 'crm.plot' is provided in **Appendix 4**.

Lake sediment samples collected in Saskatchewan in 1977 and 1993 (Hornbrook et al., 1978; Friske et al., 1994a & b) were reanalyzed in 2014 and released in a GSC open file in 2015 (McCurdy et al, 2015). Results from 70 analyses of CRM LKSD-2 that were included in a batch of 5,698 samples are listed in Table 1, along with accepted values, where available, for 'LKSD-2'. In this example, results from a number of different analytical methods are summarized by element and method. For convenience, elements are grouped by major categories (metals, metalloids and non-metals) and subcategories (alkali metals, alkaline earth metals, lanthanides, actinides, transition and post-transition metals). Values received from the commercial laboratory are compared with the accepted means and standard deviations from several different sources (Lynch, 1990, 1999; Hechler, 2013). Elements and oxides for which no published mean and standard deviation values were found, list only statistics for the samples analyzed in the Saskatchewan reanalysis project. Relative Standard Deviation (RSD), also known as Coefficient of Variation (CV) is a useful method of comparing the variation of geochemical data measured in different units and with quite different means. Generally, most RSD percentages are below 10%, an indication of relatively low variability. Relatively higher RSD values can be the result of elemental concentrations held within discrete, often refractory, minerals, including spinels, beryl, tourmalines, chromite, zircon, monazite, niobates, tungstates, topaz, and cassiterite (Crock and Lamothe, 2011). Elements such as Au that are difficult to homogenize in samples (Harris, 1982) can also have high RSD values.

Statistics listed in Table 1 can be used to provide an overall estimation of the accuracy of geochemical data. "crm.plot" in 'rgr' also displays the position of individual sample values for each element relative to acceptable limits of variation, two standard deviations above and below the long-term or accepted mean. Figure 7 identifies one Cu value falling outside the acceptable range of variability. Possible reasons for this include switched samples, insertion of the wrong certified reference material, and laboratory errors such as instrument calibration and errors in methodology such as improper acid strength or digestion time. Inspection of the data for the individual sample reveals values lower than two standard deviations below the mean for a number of elements. This is inconsistent with the possibility of a sample digestion error, which should result in all values being either higher or lower. Inconsistent homogenization of the certified reference material is an alternative possible explanation for the outlying values.
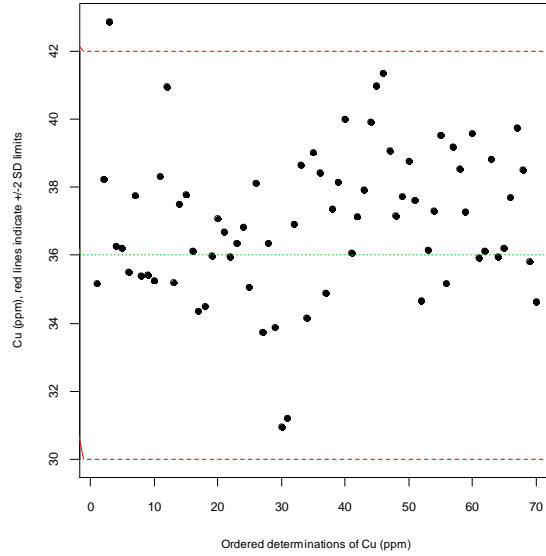
Figure 7 An X-chart generated in 'rgr' using 'crm.plot' displays concentrations of copper (partial digestion) in 70 analyses of certified reference material LKSD-2 in the order of data acquisition. The green dotted line marks the accepted mean for copper in LKSD-1 using a concentrated HCl – concentrated HNO_3 digestion published in Lynch (1999) and the red dotted lines mark the two standard deviations limits about the mean.

Saskatchewan Lake Sediment Reanalysis — Left Half

Element	LKSD-2 - Accepted Values (Mean ± SD)	LDL	Mean	SD	RSD %
Alkali Metals					
Li (AR)	14.3 ± 1.0	0.1 ppm	16.2	0.8	4.9
Na (AR)	-	0.001%	0.025	0.003	13.9
Na_2O (T)	1.93 ± 0.06	0.01%	1.90	0.03	1.7
K (AR)	-	0.01%	0.20	0.01	5.9
K_2O (T)	2.64 ± 0.10	0.01%	2.70	0.06	2.0
Rb (AR)	15.2 ± 1.3	0.1 ppm	16.8	1.2	6.9
Cs (AR)	1.22 ± 0.12	0.02 ppm	1.25	0.08	6.5
Alkaline Earth Metals					
Be (AR)	0.138 ± 0.023	0.1 ppm	0.7	0.2	33.0
Mg (AR) %	0.69 ± 0.06	0.01%	0.60	0.02	3.6
MgO (T)	1.67 ± 0.14	0.01%	1.74	0.02	1.0
Ca (AR) %	0.67 ± 0.06	0.01%	0.55	0.03	4.8
CaO (T)	2.2 ± 0.2	0.01%	2.26	0.02	1.0
Sr (T)	220 ± 41	0.5 ppm	235.0	4.1	1.8
Sr (AR)	32 ± 6	0.5 ppm	26.8	2.1	7.7
Ba (T)	780 ± 75	5 ppm	723	15	2.1
Ba (AR)	211 ± 12	0.5 ppm	213.0	11.9	5.6
Transition Metals					
TiO_2 (T)	0.56 ± 0.04	0.01%	0.55	0.01	1.0
Ti (AR)	1010 ± 166	10 ppm	650	41	6.3
V (AR)	48 ± 10	2 ppm	42	2	5.9
Cr_2O_3 (T)	-	0.002%	0.008	0.001	11.6
Cr (AR)	29 ± 3	0.5 ppm	28.2	1.6	5.8
MnO (T)	0.26 ± 0.01	<0.01%	0.25	<0.01	1.7
Mn (AR)	1840 ± 180	1 ppm	1760	56	3.2
Fe_2O_3 (T)	6.24 ± 0.29	0.04%	6.08	0.08	1.3
Fe (AR) %	3.5 ± 0.3	0.01%	3.19	0.10	3.2
Co (AR)	17 ± 1	0.1 ppm	15.5	0.9	6.0
Ni (T)	26 ± 4	20 ppm	25	7	28.9
Ni (AR)	23 ± 3	0.1 ppm	24.4	1.2	4.9
Cu (AR)	36 ± 3	0.01 ppm	37.00	2.17	5.9
Zn (AR)	200 ± 6	0.1 ppm	197.0	12.2	6.2
Zr (T)	254 ± 33	5 ppm	271	8	3.1
Zr (AR)	3.4 ± 0.3	0.1 ppm	3.8	0.4	11.6
Nb (T)	8 ± 1	5 ppm	10	4	36.8
Nb (AR)	1.29 ± 0.11	0.02 ppm	1.27	0.08	6.3
Mo (AR)	2.0 ± 0.7	0.01 ppm	1.15	0.07	6.3
Pd (AR)	-	10 ppb	<10	-	-
Ag (AR) ppb	800 ± 200	2 ppb	861	66	7.7
Cd (AR)	0.8 ± 0.2	0.01 ppm	0.84	0.06	6.8
Hf (AR)	0.104 ± 0.015	0.02 ppm	0.08	0.02	23.1
Ta (AR)	0.0087 ± 0.0028	0.05 ppm	<0.05	-	-
W (AR)	0.45 ± 0.07	0.05 ppm	0.26	0.03	10.7
Re (AR) ppb	-	1 ppb	1	1	68.7
Pt (AR) ppb	<3	2 ppb	<2	-	-
Au (AR) ppb	3.0 ± 1.1	0.2 ppb	2.8	2.4	85.3
Hg (AR) ppb	168 ± 17	5 ppb	157	12	7.5

Saskatchewan Lake Sediment Reanalysis — Right Half

Element	LKSD-2 - Accepted Values (Mean ± SD)	LDL	Mean	SD	RSD %
Post-Transition Metals					
Al_2O_3 (T)	12.3 ± 0.4	0.01%	12.10	0.09	0.8
Al (AR) %	1.87 ± 0.25	0.01%	1.46	0.06	4.0
Ga (AR)	4.6 ± 0.3	0.1 ppm	4.8	0.4	7.4
In (AR)	0.0578 ± 0.0029	0.02 ppm	0.06	0.01	15.8
Sn (AR)	1.25 ± 0.07	0.1 ppm	1.4	0.1	7.9
Tl (AR)	0.264 ± 0.014	0.02 ppm	0.26	0.01	5.0
Pb (AR)	40 ± 7	0.01 ppm	40.60	3.11	7.7
Bi (AR)	1.07 ± 0.04	0.02 ppm	1.19	0.10	8.1
Metalloids					
B (AR)	-	20 ppm	<20		
SiO_2 (T)	58.9 ± 0.6	0.01%	58.80	0.33	0.55
Ge (AR)	-	0.1 ppm	0.1	0.0	53.4
As (AR)	9 ± 3	0.1 ppm	9.0	0.6	7.2
Sb (AR)	1.2 ± 0.5	0.02 ppm	0.46	0.05	10.0
Te (AR)	0.062 ± 0.007	0.02 ppm	0.05	0.03	52.1
Non-Metals					
Total C (LECO)	-	0.02%	4.95	0.12	2.4
LOI (1000°C)	13.6 ± 0.2	0.01%	13.10	0.34	2.6
P_2O_5 (T)	0.28 ± 0.03	0.01%	0.29	0.01	4.9
P (AR)	0.1290 ± 0.0099	0.001%	0.115	0.006	5.1
Total S (LECO)	-	0.02%	0.15	0.02	9.9
S (AR) %	-	0.02%	0.15	0.01	4.9
Se (AR)	0.78 ± 0.30	0.1 ppm	0.8	0.3	42.2
Rare Earths (Lanthanides + Sc and Y)					
Sc (T)	13 ± 2	1 ppm	12	0	2.5
Sc (AR)	5.3 ± 0.4	0.1 ppm	5.4	0.4	6.7
Y (T)	44 ± 7	3 ppm	40	1	2.8
Y (AR)	26.6 ± 1.4	0.01 ppm	28.50	1.46	5.1
La (AR)	>50	0.5 ppm	58.6	2.9	5.0
Ce (AR)	>70	0.1 ppm	96.3	4.7	4.9
Pr (AR)	13.2 ± 0.7	0.02 ppm	14.60	1.02	7.0
Nd (AR)	49.2 ± 2.6	0.02 ppm	52.30	2.69	5.1
Sm (AR)	8.7 ± 0.5	0.02 ppm	9.20	0.56	6.1
Eu (AR)	1.28 ± 0.07	0.02 ppm	1.41	0.12	8.7
Gd (AR)	7.6 ± 0.4	0.02 ppm	6.65	0.59	8.8
Tb (AR)	0.97 ± 0.05	0.02 ppm	0.98	0.09	9.2
Dy (AR)	4.94 ± 0.24	0.02 ppm	5.19	0.36	6.9
Ho (AR)	0.95 ± 0.05	0.02 ppm	1.03	0.09	8.5
Er (AR)	2.67 ± 0.14	0.02 ppm	2.81	0.22	7.7
Tm (AR)	0.366 ± 0.018	0.02 ppm	0.42	0.04	9.4
Yb (AR)	2.34 ± 0.13	0.02 ppm	2.55	0.17	6.6
Lu (AR)	0.361 ± 0.021	0.02 ppm	0.40	0.04	9.4
Actinides					
Th (AR)	7.7 ± 0.7	0.1 ppm	7.4	0.8	11.4
U (AR)	6.23 ± 0.23	0.1 ppm	6.6	0.5	7.7

Table 1 The Relative Standard Deviation (RSD) % provides an indication of the average relative repeatability for each element or oxide from 70 samples of LKSD-2 inserted into Saskatchewan lake sediments reanalyzed in 2014. Each element or oxide can be linked to an associated plot of concentrations ordered by sample number generated in R (Fig. 7). Text colours indicate the source of estimated mean concentrations. Colours indicate the source of estimates of the means and standard deviations (red=Lynch, 1990, 1999; blue=Hechler, 2013; black= no data).

Precision

Precision is a measure of the ability of an analytical method to reproduce the same value each time a sample is analyzed. The methods outlined below are used to estimate random error because random error affects precision (Reimann et al., 2008). Systematic errors in a process produce results that can be precise but not accurate. A more detailed discussion can be found in **Appendix 2**. Before submitting samples for analysis, one of the 17 Routine samples, preferably one of the Field Duplicates, in each block of 20 samples is subdivided into two subsamples and the split is numbered and placed into the first position in each block of 20 (**Appendix 1**). Sample pairs are independently analyzed by the same method, using the same equipment, within short intervals of time. Precision, in the sense of repeatability (**Appendix 2**), of the analytical data received from laboratories is estimated by comparing the variability of results between the analytical duplicate pairs.

In order to provide a measure of analytical precision for each element, the squared difference between two analytical duplicates is calculated for N analytical duplicate pairs. The sum of these values is divided by the number of samples, $2N$, to provide a measure of variability (variance). A Standard Deviation is then calculated as the square root of this variance (Garrett, 1969). Relative Standard Deviation, expressed as a percentage (RSD %), is calculated by dividing the Standard Deviation by the overall mean of the duplicates and multiplying by 100 (Reimann et al., 2008). This value provides an estimate of the relative repeatability of the analytical method for each element or variable (Thompson, 1983). A lower percentage corresponds to relatively low variability and a higher percentage indicates a more variable dataset.

It is important to remember that there is an established relationship between concentration and precision (Thompson & Howarth, 1978). Relatively poor precision is predicted at concentrations near the lower determination limit but improves as concentrations increase to an optimum level for the instrument being used, and then falls off at the upper limit of concentration that the instrument is capable of measuring (Fletcher, 1981). The values listed under Precision (RSD %) in Table 2 thus provide a measure of the average precision over the ranges of concentrations provided in Table 2.

Precision, measured as a percentage Relative Standard Deviation (RSD %), can be calculated with the function `ad.plot2` in the 'rgr' package available within the R system (Garrett, 2015a): an example of the output is shown in **Appendix 4**. Pairs of duplicates are removed from the calculation if one or both values are below detection. Numerical estimates of precision calculated in `adplot2` are listed in Table 2. Shown with the element or analyte and method of analysis are the Lower Detection Limits (LDL), the total percentage of observations for each element below the Lower Detection Limit (% Below LDL (Total)), the number of duplicate pairs removed because one or both values are below detection (Duplicate Pairs Removed), the Range and the Mean of all values. This background information provides context for the estimates of Precision (RSD %) in the last column of the table.

In terms of context, knowing the lower and upper detection limits and the mean value of the data for each element or oxide is essential because the precision of the instruments used to measure concentrations vary from relatively low at the lower and higher detection limits established for the instrument to relatively high precision within the optimal range of the instrument (Thompson and

Howarth, 1978). Range is an important parameter, because if the duplicate data span more than 1.5 orders of magnitude the estimate of precision is unreliable due to heterogeneity of variance considerations (heteroscedasticity) (Garrett, 2015a).

Element (Method)	LDL	% Below LDL (Total)	Duplicate Pairs Removed	Range	Mean	Precision (RSD %)
Alkali Metals						
Li (AR)	0.1 ppm	0		0.5 - 64.6	9.2	6.9
Na (AR)	0.001%	0.4	2	<0.001 - 0.071	0.013	9.6
Na2O (T)	0.01%	0		0.04 - 3.51	1.00	5.1
K (AR)	0.01%	0		0.01 - 0.49	0.08	4.9
K2O (T)	0.01%	0		0.05 - 4.48	1.14	6.0
Rb (AR)	0.1 ppm	0		1.0 - 51.4	9.3	6.2
Cs (AR)	0.02 ppm	0		0.14 - 9.90	1.04	9.6
Alkaline Earth Metals						
Be (AR)	**0.1 ppm**	**3.3**	**17**	**<0.1 – 2.9**	**0.6**	**29.1**
Mg (AR)	0.01%	0		0.04 - 0.79	0.23	4.2
MgO (T)	0.01%	0		0.08 - 2.26	0.57	6.1
Ca (AR)	0.01%	0		0.11 – 1.56	0.45	4.8
CaO (T)	0.01%	0		0.35 - 2.80	1.21	3.9
Sr (T)	2 ppm	0		14 - 374	116	4.0
Sr (AR)	0.5 ppm	0		6.1 - 97.2	23.1	5.2
Ba (T)	1 ppm	0		41 - 1032	352	4.8
Ba (AR)	0.5 ppm	0		24.4 – 650.5	90.2	6.4
Transition Metals						
TiO2 (T)	0.01%	0		0.02 - 0.65	0.19	6.1
Ti (AR)	0.001%	0		0.005 - 0.156	0.032	4.7
V (AR)	2 ppm	0		4 - 97	26	5.1
Cr2O3 (T)	**0.002%**	**14.7**	**61**	**<0.002 - 0.024**	**0.004**	**20.5**
Cr (AR)	0.5 ppm	0		3.6 – 86.0	22.9	5.1
MnO (T)	0.01%	1.0	3	<0.01 - 2.17	0.07	15.6
Mn (AR)	1 ppm	0		28 - >10000	473	11.8
Fe2O3 (T)	0.04%	0		0.40 - 19.85	3.3	7.6
Fe (AR)	0.01%	0		0.25 - 13.0	1.99	9.2
Co (AR)	0.1 ppm	0		1.3 – 33.8	7.8	6.9
Ni (T)	20 ppm	58.1	196	<20 - 99	20	6.5
Ni (AR)	0.1 ppm	0		2.4 - 93.9	20.0	6.5
Cu (AR)	0.01 ppm	0		1.28 – 226.75	27.65	5.8
Zn (AR)	0.1 ppm	0		15.4 - 289.3	99.9	6.7
Zr (T)	5 ppm	0		7 - 381	99	8.4
Zr (AR)	0.1 ppm	0.4	1	<0.1 - 14.0	1.9	11.9
Nb (T)	**5 ppm**	**60.0**	**213**	**<5 - 23**	**5**	**27.6**
Nb (AR)	0.02 ppm	0		0.23 - 3.88	1.28	6.9
Mo (AR)	0.01 ppm	0		0.23 - 45.27	3.74	10.1
Pd (AR)	**10 ppb**	**99.3**	**289**	**<10 - 13**	**<10**	
Ag (AR)	2 ppb	0		9 - 519	98	9.8
Cd (AR)	0.01 ppm	0		0.01 - 1.50	0.43	12.3
Hf (AR)	0.02 ppm	48.3	162	<0.02 – 0.33	0.04	16.3
Ta (AR)	**0.05 ppm**	**100**	**289**	**<0.05**	**<0.05**	
W (AR)	**0.05 ppm**	**0.7**	**2**	**<0.05 – 7.21**	**0.61**	**30.5**
Re (AR)	**1 ppb**	**50.5**	**199**	**<1 - 20**	**2**	**41.2**
Pt (AR)	**2 ppb**	**81.1**	**277**	**<2 - 5**	**<2**	
Au (AR)	**0.2 ppb**	**12.8**	**58**	**<0.2 – 47.2**	**1.5**	**175.1**
Hg (AR)	**5 ppb**	**1.8**	**4**	**<5 - 292**	**82**	**23.5**

Element (Method)	LDL	% Below LDL (Total)	Duplicate Pairs Removed	Range	Mean	Precision (RSD %)
Post-Transition Metals						
Al2O3 (T)	0.01%	0		0.79 - 14.67	6.06	3.0
Al (AR)	0.01%	0		0.20 - 2.83	1.07	5.0
Ga (AR)	0.1 ppm	0		0.3 - 8.4	2.3	5.9
In (AR)	**0.02 ppm**	**88.2**	**269**	**<0.02 – 0.05**	**0.01**	
Sn (AR)	0.1 ppm	5.2	19	<0.1 - 1.0	0.3	15.9
Tl (AR)	0.02 ppm	0		0.02 - 0.76	0.18	10.0
Pb (AR)	0.01 ppm	0		0.81 - 12.55	3.79	13.3
Bi (AR)	0.02 ppm	0.7	3	< 0.02 – 3.08	0.15	19.7
Metalloids						
B (AR)	**20 ppm**	**98.3**	**286**	**<20 - 35**	**<20**	
SiO2 (T)	0.01%	0		6.45 - 79.39	51.58	1.8
Ge (AR)	**0.1 ppm**	**85.8**	**265**	**<0.1 – 0.5**	**0.1**	
As (AR)	**0.1 ppm**	**16.4**	**76**	**<0.1 - 15.1**	**0.8**	**44.8**
Sb (AR)	**0.02 ppm**	**9.0**	**32**	**<0.02 - 0.95**	**0.05**	**66.6**
Te (AR)	**0.02 ppm**	**69.4**	**246**	**<0.02 - 0.08**	**0.02**	**41.4**
Non-Metals						
Total C (LECO)	0.02%	0		0.13 - 50.05	16.16	3.7
LOI (1000°C)	0.1%	0		1.0 - 88.7	34.3	2.5
P2O5 (T)	0.01%	0		0.06 – 0.95	0.31	7.1
P (AR)	0.001%	0		0.022 - 0.406	0.117	9.1
Total S (LECO)	0.02%	1.4	4	<0.02 - 1.40	0.32	10.1
S (AR)	0.02%	1.4	4	<0.02 - 1.48	0.32	7.1
Se (AR)	0.1 ppm	5.5	22	<0.1 – 2.6	0.9	15.5
Rare Earths (Lanthanides + Sc and Y)						
Sc (T)	1 ppm	1.7	6	<1 - 14	5	8.4
Sc (AR)	0.1 ppm	0		0.2 - 6.0	2.0	8.3
Y (T)	3 ppm	0.4	1	<3 - 129	19	6.2
Y (AR)	0.01 ppm	0		1.66 - 132.44	16.32	9.1
La (AR)	0.5 ppm	0		2.2 - 272.6	47.8	6.7
Ce (AR)	0.1 ppm	0		4.7 - 546.5	88.7	10.2
Pr (AR)	0.02 ppm	0		0.53 - 69.26	11.82	6.5
Nd (AR)	0.02 ppm	0		2.23 - 277.80	41.66	7.7
Sm (AR)	0.02 ppm	0		0.41 - 42.99	6.72	7.8
Eu (AR)	0.02 ppm	0		0.04 - 6.45	0.99	9.2
Gd (AR)	0.02 ppm	0		0.26 - 26.94	4.37	11.2
Tb (AR)	0.02 ppm	0		0.05 - 3.74	0.60	8.7
Dy (AR)	0.02 ppm	0		0.35 - 19.43	3.08	8.4
Ho (AR)	0.02 ppm	0		0.05 - 3.91	0.58	8.9
Er (AR)	0.02 ppm	0		0.15 - 11.01	1.58	9.8
Tm (AR)	0.02 ppm	0		0.02 - 1.74	0.23	10.9
Yb (AR)	0.02 ppm	0		0.17 - 11.18	1.46	10.1
Lu (AR)	0.02 ppm	0		0.02 - 1.56	0.23	12.0
Actinides						
Th (AR)	0.1 ppm	0.9	4	<0.1 - 21.9	3.1	11.8
U (AR)	0.05 ppm	0		0.58 - 295.72	13.19	8.6

Table 2 Measures of precision for 289 analytical duplicate pairs from a northern Saskatchewan lake sediment survey. AR is aqua regia digestion and T is a total method of analysis using a lithium metaborate fusion. LDL is the Lower Detection Limit. Mean and Precision (RSD %) are calculated with 'ad.plot2' in 'rgr'. Precisions of elements in bold black type are not reported because there are less than 30 duplicate pairs with both values above the detection limit with which to estimate precision. The RSD of elements in bold red type exceeds 20 %.

The overall quality of a sample batch is assessed by the RSD: individual analytical pairs are monitored and assessed for acceptability using a ratio plot (Fig. 8) that is a variant of a Youden plot (Garrett, 2015b). The *x-axis* of the ratio plot spans the range of duplicate data pair means and the *y-axis* plots the ratio of the differences. Logarithmic scaling is employed to give equal graphical 'weight' to ratios above

and below unity. The nature of the ratio leads implicitly to a statistical test equivalent to a paired t-test on the logarithms of the duplicate ratios (Garrett, 2015b). In addition, the plot graphically illustrates the estimation of 95% limits on the expected range of the duplicate ratios across the range of the data to provide a useful estimate of the analytical precision.
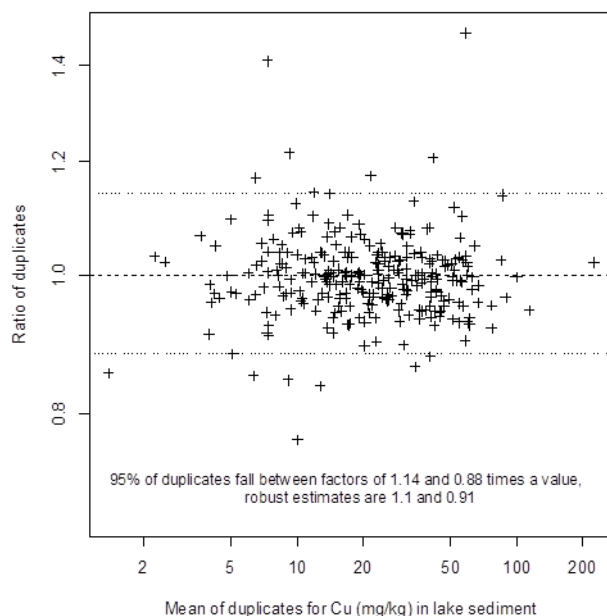
Figure 8 A ratio plot of the $\log_{10}$ mean of Cu in 289 lake sediment analytical duplicate pairs from northern Saskatchewan against the difference from the 1:1 line, expressed on a logarithmic scale. Dotted lines are an estimate of the 95% limits on the expected range of the duplicates across the range of the data.

The RSD assumes that random errors in the data have a normal or Gaussian distribution about their mean concentration (Fletcher, 1981). When censored (below detection) values exceed 12%, the distribution is skewed and is no longer normal (Fletcher, 1981). This tends to occur in variables with low concentrations in a sample or with relatively high lower detection limits such as Cr_2O_3, Ni and Nb by total methods as well as Pd, Hf, Ta, Re, Pt, In, B, Ge, As, and Te. These data need not necessarily be rejected: we simply can't say whether the data are 'good' or 'bad', based on the RSD. RSD calculations listed for elements or variables with values below detection are derived only from those duplicate pairs with both values above detection. RSD is not estimated for elements or variables with less than 30 duplicate pairs having both observations above detection (Pd, Ta, Pt, In, B, Ge). Elements listed in bold red type in Table 2, such as Be, Cr2O3, Nb, W, Re, Au, Hg, As, Sb and Te, have RSD exceeding 20 % and may be the result of heterogeneity in sample material (Au) or inability of the digestion to dissolve some minerals, resulting in apparently low concentrations. The remaining variables have a RSD below 20%, implying that these data are reliable. A value of ±20% indicates that, on average, 95 out of 100 analyses should fall within two standard deviations on either side of the mean (Fletcher, 1981). A RSD exceeding 20% combined with a relatively narrow range of concentrations could result in overlooked anomalies (Howarth and Martin, 1979).

Analysis of Variance (ANOVA)

Field duplicate samples (**Appendix 1**) are collected at sample sites to provide a means of comparing the combined variability of geochemical data introduced by field sampling procedures, sediment heterogeneity and analysis with regional variability. When a sufficient number of field duplicate pairs, at least 30, have been analysed, a *one-way random effects model Analysis of Variance* (**Appendix 2**) can be carried out. Functions 'anova1' or 'anova2' in the 'rgr' package running under the R system (Garrett, 2015a) are used to estimate the combined sampling and analytical variability ('within' sites) and regional variability ('between' sites) in sets of duplicate field samples. This combined variability is more important than analytical variability alone because if the combined sampling and analytical component is not significantly smaller than the field survey component, it cannot be stated that there are statistically significant spatial patterns in the data, and thus the data may be misleading, or at worst, meaningless (Thompson & Howarth, 1973; Garrett, 1983; Reimann et al., 2008). When the analytical duplicates are systematically split from the field duplicate, or the routine sample associated with the field duplicate, it is possible to separately estimate both the sample site and analytical variability (Garrett, 2013b).

The variance ratio, or F statistic (**Appendix 2**), is calculated to determine whether the variance 'within' is significantly smaller than the variation 'between'. As a 'rule of thumb' this ratio should exceed 4.0 for sampling and analytical errors to be significantly smaller at the 95% confidence level. The p-value (**Appendix 2**) is a measure of the exact level of confidence in the F statistic. Generally an acceptable p-value is less than 0.05 (>95th percentile), i.e. there is a <5% probability that the observed F ratio could have occurred due to chance alone. ANOVA results using data from 288 field duplicate pairs from the reanalysis of northern Saskatchewan lake sediments (McCurdy et al., 2015) are shown in Table 3. Data for duplicate pairs in which one or both values are below detection are omitted from the calculations. No results are shown for Pd, Ta or B for this reason. Results in Table 3 indicate that for most elements the component of variability introduced by field sampling and laboratory procedures is significantly less than the regional variability and these data are suitable for mapping and statistical purposes. Based on the percentages of variation and combined parameters of F statistic and p-value, elements for which caution regarding statistical calculations or mapping should be advised include Nb (T), Pt, Au, In and Te. However, anomalously high levels of such elements of economic interest should be investigated even though the data may not be suitable for presentation as contour maps.

A logarithmic transformation of the geochemical data is carried out in recognition that the data are drawn from constant sum compositions (Filzmoser et al., 2009; Garrett, 2013a), and in order to minimize the effects of heteroscedasticity, i.e. where the variance of the duplicate pairs varies as a function of the means of the pairs across the range of the data. Normality and homogeneity of variance (lack of heteroscedasticity) are fundamental assumptions of the ANOVA procedure (Garrett, 1983).

Element (Method)	n-pairs	Between Sites	At Sites	F	p-value
Alkali Metals					
Li (AR)	289	97.3	2.7	73.60	<0.0001
Na (AR)	287	90.8	9.2	20.79	<0.0001
Na$_2$O (T)	288	95.6	4.4	44.93	<0.0001
K (AR)	289	95.5	4.5	43.52	<0.0001
K$_2$O (T)	288	96.1	3.9	50.87	<0.0001
Rb (AR)	289	96.3	3.7	52.7	<0.0001
Cs (AR)	289	96.8	3.2	61.72	<0.0001
Alkaline Earth Metals					
Be (AR)	277	66.3	33.7	4.93	<0.0001
MgO (T)	288	96.9	3.1	63.93	<0.0001
Mg (AR)	289	96.6	3.4	58.26	<0.0001
CaO (T)	288	95.8	4.2	46.4	<0.0001
Ca (AR)	289	94.9	5.1	38.26	<0.0001
Sr (T)	288	95.5	4.5	43.39	<0.0001
Sr (AR)	289	95.7	4.3	45.53	<0.0001
Ba (T)	288	96.1	3.9	50.94	<0.0001
Ba (AR)	289	91.3	8.7	22.09	<0.0001
Transition Metals					
TiO$_2$ (T)	288	96.9	3.1	63.48	<0.0001
Ti (AR)	289	96.1	3.9	50.79	<0.0001
V (AR)	289	95.2	4.8	40.81	<0.0001
Cr$_2$O$_3$ (T)	214	67.8	32.2	5.21	<0.0001
Cr (AR)	289	94.3	5.7	34.04	<0.0001
MnO (T)	284	89.3	10.7	17.7	<0.0001
Mn (AR)	289	89.5	10.5	18.06	<0.0001
Fe$_2$O$_3$ (T)	288	93.2	6.8	28.48	<0.0001
Fe (AR)	289	92.1	7.9	24.22	<0.0001
Co (AR)	289	92.3	7.7	24.98	<0.0001
Ni (T)	95	83.7	16.3	11.3	<0.0001
Ni (AR)	289	95.7	4.3	45.74	<0.0001
Cu (AR)	289	96.8	3.2	62.05	<0.0001
Zn (AR)	289	86.5	13.5	13.82	<0.0001
Zr (T)	288	95.8	4.2	46.97	<0.0001
Zr (AR)	288	92.8	7.2	26.6	<0.0001
Nb (T)	53	34.1	65.9	2.03	0.0057
Nb (AR)	289	95.6	4.4	44.8	<0.0001
Mo (AR)	289	93	7	27.43	<0.0001
Pd (AR)	0				
Ag (AR)	289	93.1	6.9	27.94	<0.0001
Cd (AR)	289	87.3	12.7	14.75	<0.0001
Hf (AR)	127	77.7	22.3	7.96	<0.0001
Ta (AR)	0				
W (AR)	285	92.3	7.7	25.03	<0.0001
Re (AR)	112	71.9	28.1	6.12	<0.0001
Pt (AR)	16	42.5	57.5	2.48	0.0406
Au (AR)	231	39.5	60.5	2.31	<0.0001
Hg (AR)	288	87.4	12.6	14.82	<0.0001

Element (Method)	n-pairs	Between Sites	At Sites	F	p-value
Post-Transition Metals					
Al$_2$O$_3$ (T)	288	96.7	3.3	59.15	<0.0001
Al (AR)	289	94.6	5.4	36.31	<0.0001
Ga (AR)	289	96.2	3.8	51.22	<0.0001
In (AR)	14	0	100	0.75	0.6949
Sn (AR)	265	81.1	18.9	9.56	<0.0001
Tl (AR)	289	90.6	9.4	20.29	<0.0001
Pb (AR)	289	92.2	7.8	24.57	<0.0001
Bi (AR)	285	91.2	8.8	21.79	<0.0001
Metalloids					
B (AR)	0				
SiO$_2$ (T)	288	95.4	4.6	42.39	<0.0001
Ge (AR)	27	68.2	31.8	5.29	<0.0001
As (AR)	201	74.3	25.7	6.79	<0.0001
Sb (AR)	268	73.4	26.6	6.52	<0.0001
Te (AR)	49	13.7	86.3	1.32	0.1681
Non-Metals					
Total C (LECO)	288	67.9	32.1	5.24	<0.0001
LOI (1000°C)	285	83.2	16.8	10.91	<0.0001
P$_2$O$_5$ (T)	288	92.9	7.1	27	<0.0001
P (AR)	289	93.8	6.2	31.46	<0.0001
Total S (LECO)	288	88.5	11.5	16.37	<0.0001
S (AR)	288	90.5	9.5	19.96	<0.0001
Se (AR)	283	83.9	16.1	11.43	<0.0001
Rare Earths (Lanthanides + Sc and Y)					
Sc (T)	279	92.2	7.8	24.7	<0.0001
Sc (AR)	289	94.2	5.8	33.21	<0.0001
Y (T)	285	96.9	3.1	63.92	<0.0001
Y (AR)	289	96.7	3.3	59.12	<0.0001
La (AR)	289	98.1	1.9	103.81	<0.0001
Ce (AR)	289	98	2	97.15	<0.0001
Pr (AR)	289	97.9	2.1	95.05	<0.0001
Nd (AR)	289	97.9	2.1	92.28	<0.0001
Sm (AR)	289	97.7	2.3	84.42	<0.0001
Eu (AR)	289	96.9	3.1	62.85	<0.0001
Gd (AR)	289	96.5	3.5	56.22	<0.0001
Tb (AR)	289	97.1	2.9	66.8	<0.0001
Dy (AR)	289	96.6	3.4	57.65	<0.0001
Ho (AR)	289	95.9	4.1	47.34	<0.0001
Er (AR)	289	95.9	4.1	47.86	<0.0001
Tm (AR)	289	95	5	38.68	<0.0001
Yb (AR)	289	95.5	4.5	43.89	<0.0001
Lu (AR)	289	95.6	4.4	43.99	<0.0001
Actinides					
Th (AR)	282	95.2	4.8	41.09	<0.0001
U (AR)	289	98.1	1.9	104.15	<0.0001

Table 3 ANOVA table derived from field duplicate pair geochemical data for variables in northern Saskatchewan lake sediments. 'AR' is aqua regia digestion and 'T' is a total (lithium metaborate) digestion.

Acknowledgements

References

Alfassi, Z.B., 2000. Instrumental Neutron Activation Analysis; *in* Encyclopedia of Analytical Chemistry, (ed.) R.A. Meyers; John Wiley & Sons, Chichester, pp. 12497-12526.

Burnham, O.M., Schweyer, J., 2004. Trace element analysis of geological samples by inductively Coupled Plasma Mass Spectrometry at the Geoscience Laboratories; *in* Summary of Field Work and Other Activities 2013; Ontario Geological Survey, Open File Report 6145, p. 54-1 to 54-20.

Crock, J.G., Lamothe, P.J., 2011. Inorganic chemical analysis of environmental materials – A lecture series: U.S Geological Survey Open-File Report 2011-1193, 117 p.

Dallal, G.E., 2007. The Little Handbook of Statistical Practice (http://www.jerrydallal.com/LHSP/LHSP.htm). Accessed 2015-01-20.

Exell, R.H.B., 2001. Practical Mathematics (http://www.jgsee.kmutt.ac.th/exell/PracMath/ErrorAn.htm) , King Mongkut's University of Technology Thonburi. Accessed 2015-01-20.

Filzmoser, P., Hron, K., Reimann, C., 2009. Univariate statistical analysis of environmental (compositional) data: problems and possibilities; Science of the Total Environment, 407, pp. 6100-6108.

Fletcher, W.K., 1981. Analytical methods in geochemical prospecting. In: Govett, G.J.S. (Ed.), Handbook of Exploration Geochemistry, Vol. 1; Analytical Methods in Geochemical Prospecting. Elsevier, Amsterdam, 255 p.

Friske, P.W.B., McCurdy, M.W., Day, S.J.A., Gross, H., Balma, R.G., Lynch, J.J., Durham, C.C., 1994a. National Geochemical Reconnaissance lake sediment and water data, northeastern Saskatchewan (parts of NTS 64L, 64M and 74P); Geological Survey of Canada Open File 2857.

Friske, P.W.B., McCurdy, M.W., Day, S.J.A., Gross, H., Balma, R.G., Lynch, J.J., Durham, C.C., 1994b. National Geochemical Reconnaissance lake sediment and water data, northwestern Saskatchewan (parts of NTS 74N and 74O); Geological Survey of Canada Open File 2858.

Garrett, R.G., 1969. The determination of sampling and analytical errors in exploration geochemistry; Economic Geology, 64, pp. 568-569.

Garrett, R.G., 1983. Sampling Methodology. In: R.J. Howarth (Ed.) *Statistics and Data Analysis in Geochemical Prospecting*. Handbook of Exploration Geochemistry, Vol. 2, (G.J.S. Govett, series ed). Elsevier, Amsterdam, Chapter 4, pp. 83-110.

Garrett, R.G., 2013a. The 'rgr' package for the R Open Source statistical computing and graphics environment - a tool to support geochemical data interpretation. Geochemistry: Exploration, Environment, Analysis, 13, pp. 355-378.

Garrett, R.G., 2013b. Assessment of local spatial and analytical variability in regional geochemical surveys with a simple sampling scheme. Geochemistry: Exploration, Environment, Analysis, 13, pp. 349-354.

Garrett, R.G., 2015a. The GSC Applied Geochemistry EDA Package, http://cran.r-project.org/web/packages, 17 March 2015.

Garrett, R.G., 2015b. A comparison of Shewhart, Thompson and Howarth, and Youden Plots - Advantages and Disadvantages; Explore – Assoc. of Applied Geochemists Newsletter, 167, pp. 1, 5-11.

Geboy, N.J., Engle, M.A., 2011. Quality assurance and quality control of geochemical data: a primer for the research scientist; U.S. Geological Survey Open-File Report 2011-1187, 28 p.

Grunsky, E.C., 2010. The interpretation of geochemical survey data; Geochemistry: Exploration, Environment Analysis, 10, pp. 27-74.

Harris, J.F., 1982. Sampling and analytical requirements for effective use of geochemistry in exploration for gold; In: A.A. Levinson (Ed.), Precious Metals in the Northern Cordillera, proceedings of a symposium sponsored by the Association of Exploration Geochemists and the Cordilleran Section of the Geological Association of Canada, pp. 53-67.

Hechler, J.H., 2013. QA/QC: Summary of 2012-2013 quality-control data at the Geoscience Laboratories; in Summary of Field Work and Other Activities 2013; Ontario Geological Survey, Open File Report 6290, pp. 48-1 to 48-6.

Heiberger, R.M., Newwirth, E., 2009. R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis and Graphics; Springer, London, 342 p.

Helsel, D.R., 2005. Nondetects and Data Analysis: Statistics for Censored Environmental Data; Wiley Interscience, Hoboken, 250 p.

Hornbrook, E.H.W., Coker, W.B., Lynch, J.J., 1978. National geochemical reconnaissance release NGR 19-1977, regional lake sediment and water geochemical reconnaissance data, northeastern Saskatchewan; Geological Survey of Canada Open File 508.

Horowitz, A.J., 1991. A Primer on Sediment-Trace Element Chemistry, 2nd Edition; United States Geological Survey Open File Report 91-76; 136 p.

Howarth, R.J., Martin, L., 1979. Computer-based techniques in the compilation, mapping and interpretation of exploration geochemical data. In: P.J. Hood (Ed.), Geophysics and Geochemistry in the Search for Metallic Ores; Geological Survey of Canada, Economic Geology Report 31, pp. 545-574.

Krumbein, W.C., Graybill, F.A., 1965. An Introduction to Statistical Models in Geology; McGraw-Hill Book Company, New York, 475 p.

Lynch, J.J., 1990. Provisional elemental values for eight new geochemical lake sediment and stream sediment reference materials, LKSD-1, LKSD-2, LKSD-3, LKSD-4, STSD-1, STSD-2, STSD-3 and STSD-4; Geostandards Newsletter, 14, pp. 153-167.

Lynch, J.J., 1996. Provisional elemental values for four new geochemical soil and till reference materials, TILL-1, TILL-2, TILL-3 and TILL-4. Geostandards Newsletter 20, pp. 277-287.

Lynch, J.J., 1999. Additional provisional elemental values for LKSD-1, LKSD-2, LKSD-3, LKSD-4, STSD-1, STSD-2, STSD-3 and STSD-4; Geostandards Newsletter, 23, pp. 251-260.

McClenaghan, M.B., Parkhill, M.A., Pronk, A.G., McCurdy, M.W., Boldon, R., Pyne, M., Rice, J.M. In press. Geochemical and indicator mineral data for stream sediments and waters around the Sisson W-Mo deposit, New Brunswick (NTS 21J/06, 21J/07); Geological Survey of Canada, Open File 7756.

McCurdy, M.W., McNeil, R.J., and Day, S.J.A., 2012. Geochemical, Mineralogical and Kimberlite Indicator Mineral Data for Silts, Heavy Mineral Concentrates and Waters from Two Geochemical Surveys (2010 and 2011) on Victoria Island, Northwest Territories (NTS 87G, 87H, 88A and 88B); Geological Survey of Canada, Open File 7198, 1 CD-ROM, doi:10.4095/291509

McCurdy, M.W., Berman, R.G., Kerr, D.E., Vaive, J.E., 2013. Geochemical, Mineralogical and Kimberlite Indicator Mineral Data for Silts, Heavy Mineral Concentrates and Waters, Duggan Lake Area (NTS 76-H and 76-I South); Geological Survey of Canada, Open File 7471, doi:10.4095/293044

McCurdy, M.W., Pehrsson, S.J., Campbell, J.E., Adcock, S.W., 2015. Regional Lake Sediment and Water Geochemical Data, Northern Saskatchewan, (NTS 64-L, 64-P, 74-N, 74-O and 74-P); Geological Survey of Canada, Open File 7746, 1 zip file, doi:10.4095/295977

R Core Team, 2015. The R Project for Statistical Computing. https//www.r-project.org. Accessed October 20, 2015.

Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Statistical Data Analysis Explained. John Wiley & Sons, Ltd., 343 p.

Scheuren, F., 2005. "What is a Margin of Error?", Chapter 10, in "What is a Survey?", American Statistical Association, Washington, D.C. Accessed 2015-01-19.

Thomas, R., 2004. Practical Guide to ICP-MS. Marcel Dekker, Inc., 324 p.

Thompson, M., 1983. Control procedures in geochemical analysis. In: Howarth, R.J. (Ed.), Handbook of Exploration Geochemistry, Vol. 2; Statistics and Data Analysis in Geochemical Prospecting. Elsevier, Amsterdam, 437 p.

Thompson, M., Howarth, R.J., 1973.The rapid estimation and control of precision by duplicate determinations; The Analyst, 98, pp. 153-160.

Thompson, M., Howarth, R.J., 1978.A new approach to the estimation of analytical precision; Journal of Geochemical Exploration, 9, pp. 23-30.

Walpole, R.E., 1982. Introduction to Statistics, 3rd Edition. Macmillan Publishing Company, New York, 521 p.

Van der Linden, W.E., 1989. Definition and classification of interferences in analytical procedures; Pure and Applied Chemistry, 61, pp. 91-95.

Appendix 1

Quality Control Samples

Routine, Field and Analytical Duplicate Samples

Routine Sample – A 'Routine' sample is a sample of material (lake sediment, stream silt, soil, till, etc.) collected at a specific location for the purpose of a scientific investigation, from which physical characteristics (weight, colour, grain size, etc.) and geochemical data (concentration of elements, oxides, organic material or some other chemical constituent) are determined by observation and analysis.

Field Duplicate – A Field Duplicate sample is a second sample taken at or within a few metres of the site, depending on the scale of the survey, of a Routine sample. It is collected, prepared and analyzed using the same methods as for the Routine sample.

Analytical Duplicate – An Analytical Duplicate sample is a split from a Routine Sample or a Field Duplicate after the samples have been prepared for analysis but before analysis, and analyzed using the same methods as the routine samples.

Certified Reference Material – A Certified Reference Material (CRM) has values for one or more properties certified by a technically valid procedure accompanied by or traceable to a certificate or other documentation issued by a certifying body (Horowitz, 1991). See below for a discussion of the CRMs used at the Geological Survey of Canada.

Certified Reference Materials (CRM)

A suitable Certified Reference Material (Horowitz, 1991) is one whose mineralogical matrix matches the 'Routine' samples (samples collected for a scientific study and/or to prepare maps) submitted for analysis as closely as possible in terms of origin, preparation and composition. Certified Reference Materials collected and prepared by GSC staff (lake and stream sediment, soil and till) are included with real ('Routine') samples submitted for analysis to estimate the accuracy of laboratory results. Several hundred kilograms each of sample material were collected to prepare four lake sediment, four stream sediment, two soil and two till Certified Reference Materials. Bulk samples were separately dried, sieved, ball-milled, homogenized and bottled. Sample aliquots were submitted to many different laboratories for analysis by a number of partial and total extraction procedures. Although insufficient data were returned for certification under International Organization for Standardization (ISO) guidelines, provisional elemental values with standard deviations for CRMs have been published (Lynch, 1990; 1996; 1999). Additional published (Burnham and Schweyer, 2004; Hechler, 2013) and unpublished data for the CRMs have been compiled for many elements and several different extraction procedures. Analytical data from CRMs in a sample batch are compared with accepted or provisional results to estimate the accuracy of the batch samples, and monitor for drift, systematic changes in determinations with time.

The Block of 20 Samples System

The following example is from a stream sediment and water survey carried out in 2014 in Nunavut (McCurdy et al., 2013). Samples were collected from 92 pre-selected sites on two different National Topographic System map sheets, 76H and 76I; however the number of samples submitted to commercial labs for analysis totalled 110. Additional samples were included as follows: within each block of 20 consecutive sample numbers (e.g. 76I 141001 to76I 141020) there were 17 routine sites, one Certified Reference Material sample, one 'Analytical Duplicate' sample and one 'Field Duplicate' sample (Fig. 1.1).

Ideally, field duplicate sites should be selected randomly. In practice, the choice of which of the 17 sites within a block of 20 to duplicate is left to the judgement of the sampling team: ease of collection is generally the main consideration during a stream (or lake) sediment survey. At the site selected for a field duplicate sample, two separate samples are collected several meters apart. In this example, sample numbers are consecutive within the block of 20 (e.g., 76I 141014 and 76I 141015), but they do not have to be; occasionally the number chosen for a reference material sample intervenes.

A sample number in each block (e.g., 76I 141012) was reserved for a Certified Reference Material (CRM). Ideally, lists of random numbers are computer-generated to select positions for reference materials; however, 'random' locations are usually selected within each block of 20 by the sampling crew.

Sample Number	Sample Type
076I 141001	Analytical Duplicate (split from 76I 141006)
076I 141002	Routine
076I 141003	Routine
076I 141004	Routine
076I 141005	Routine
076I 141006	Routine (split for analytical duplicate)
076I 141007	Routine
076I 141008	Routine
076I 141009	Routine
076I 141010	Routine
076I 141011	Routine
076I 141012	Certified Reference Material
076I 141013	Routine
076I 141014	Field Duplicate 1 (Routine)
076I 141015	Field Duplicate 2
076I 141016	Routine
076I 141017	Routine
076I 141018	Routine
076I 141019	Routine
076I 141020	Routine

Figure 1.1 In addition to 17 'Routine' samples, the first block of 20 samples submitted to a commercial laboratory for geochemical analysis includes a sample of a certified reference material, a pair of samples (field duplicates) collected from the same ('Routine') site and an aliquot of sample split from another 'Routine' sample (analytical duplicate).

A sample number in each block of twenty is reserved for the second member of an analytical duplicate pair. This reserved number is always the first field sample number in the block (i.e. 76I 141001, 76I 141021, 76I 141041, 76I 141061, etc.).

Samples were collected from 7 routine sites in NTS map sheet 76H. A sample number was added to the end of the incomplete sequence in order to provide a place for a control reference sample in this partial block (Fig. 1.2). Samples submitted to a commercial laboratory for geochemical analysis also included a 'Field Duplicate' sample and an 'Analytical Duplicate' sample (see above). It should be noted that including the NTS map sheet number (e.g. 076H) in the sample number can lead to inefficiencies resulting from the insertion of additional control samples in incomplete 'blocks of 20' associated with surveys that span multiple map sheets.

Sample Number	Sample Type
076H 141001	Analytical Duplicate (split from 76H 141008)
076H 141002	Routine
076H 141003	Field Duplicate 1 (Routine)
076H 141004	Field Duplicate 2
076H 141005	Routine
076H 141006	Routine
076H 141007	Routine
076H 141008	Routine (split for analytical duplicate)
076H 141009	Routine
076H 141010	Certified Reference Material

Figure 1.2 A partial 'block' of 7 'Routine' sites submitted to a commercial laboratory for geochemical analysis includes a Field Duplicate sample, a Certified Reference Material sample and an Analytical Duplicate sample split from a Routine sample.

The exact value of an element concentration in a standard cannot be determined (Thompson, 1983). The mean value used as the 'true' concentration is thus referred to as the 'accepted' or 'provisional' value, and consists of all the values for a given element related to a given extraction or analogous group of methods (Lynch, 1990).

Where the Analytical Duplicate can be prepared from one of the field duplicate pair of samples it is possible to separately estimate the sampling variability and the analytical variability (Garrett, 2013b). The availability of this knowledge can inform decisions concerning field and analytical procedures to improve the quality of future surveys.

Appendix 2

Definitions

Populations and Parameters, Samples and Statistics, Descriptive Statistics and Statistical Inference

A **population** is the total number of things from which observations or measurements can be collected. The totality of the population is referred to as the 'target population' about which inferences and statements are to be made, and the physical samples[1] collected for investigation form the 'sample population' (Krumbein and Graybill, 1965). A statistical **sample** is a subset of a 'target population'. A **statistic** is an estimate of an unknown property of a 'sample population'. Descriptive properties of a population are **parameters**, that is, any numerical value describing a characteristic of the population. **Descriptive statistics** are methods that provide information only about the set of collected observations and do not draw inferences or conclusions about the larger population. **Statistical inference** uses statistical procedures to analyze a sample of a population, leading to predictions or inferences about the entire population (Walpole, 1982).

Measures of Central Location

Statistics that are used to estimate the quality of analytical data include the arithmetic mean, median, range, and the sample standard deviation. In order to know how accurately a sample statistic is estimating a population parameter, it is necessary to know the distribution of the observations about a central location. The most important distribution is the normal, or Gaussian, distribution (see below). The most commonly used measures of central location are the mean, the median and the mode. The arithmetic mean (of a finite) population or sample is the sum of the values divided by the number of values. The mode of a set of observations is the value with the highest probability of occurrence. The median of a set of observations arranged in an increasing or decreasing order of magnitude is the middle value when the number of observations is odd. In the case where the number of observations is even, the median is the arithmetic mean of the two central values (Walpole, 1982).

Range

The range of a set of observations is the difference between the minimum value and the maximum value. Geochemical data often include extreme values (outliers) and therefore do not have a normal (Gaussian) distribution. Consequently the range is not usually a good measure of the variation in a set of geochemical observations (Reimann et al., 2008).

Accuracy and Precision

Accuracy is an estimate of how close the mean of a set of observations is to the 'true' or 'accepted' concentration the analyte (Thompson, 1983). **Precision** is a measure of the spread of individual measurements or results about a central value, which is estimated from the standard deviation of the sample: a low standard deviation indicates a high precision (Reimann et al., 2008). Repeatability is the

---

[1]Note this is distinct from the usage of statisticians. To a physical scientist a sample is an individual specimen under study, a collection of which forms a statistical sample.

precision estimated when a single analyst completes the analysis in a single session using the same solutions, equipment and instrumentation. Reproducibility is the precision under any other set of circumstances, including between analysts, or between laboratory sessions for a single analyst. Reproducibility cannot be better than repeatability (Reimann et al., 2008).

Types of Statistical Error

In statistics, an 'error' is not a mistake: sampling error is the estimated difference or variability of a statistic obtained from a sample and the 'true' value of the population parameter (Scheuren, 2005). There are two types of error that have to be considered when evaluating geochemical data. Random errors affect precision: systematic errors affect accuracy (Fletcher, 1981). Systematic error remains constant or varies in a predictable way, with the result that repeated measurements will be consistent: it can be anticipated and corrected. In principle, all systematic error is avoidable, but in practice may remain undetected (Exell, 2001). Bias is a measure of the total systematic error (Reimann et al., 2008).

Systematic error can arise at any stage during the collection, preparation and analysis of geochemical samples. A sampling crew consistently collecting sand-sized material rather than silt from the active part of the stream channel, or water samples preserved using nitric acid from the same bottle that is assumed to be ultra-pure, but is not, will introduce systematic error. A lab technician who consistently forgets to adjust the balance for the weight of the sample paper is making a systematic error. (Small weight changes from sample paper to sample paper add a random component to this last example.)

Random error varies in unpredictable ways. Random error affects precision and is characterized by random variations in magnitude and direction. Because they are random, positive and negative deviations tend to cancel, provided enough measurements are made (Krumbein and Graybill, 1965). Random error can arise during the collection of samples, the preparation and analytical procedures, and while making measurements. At the collection stage, only a small portion of available material is actually collected. During preparation, only a fraction of the original collected material is put into a vial and shipped to the lab for analysis. The chance of small-scale differences, such as unequal distributions of particles of gold throughout the set of samples, for example, increases at each step.

Sample Variance

The variance of a sample, usually denoted by s^2, is a statistic indicative of the spread or dispersion of the distribution of observations in a sample of a population. It is calculated by taking the sum of the squares of the differences between each value and the mean of the sample and dividing by the number of samples less one (*n-1*) for the degree of freedom lost by estimation of the mean (Walpole, 1982).

Sample Standard Deviation

Sample standard deviation (*s*) is the positive square root of the sample variance. The standard deviation, because it is in the same units as the mean, provides an estimate of the variation or dispersion about the average, or mean, value. A relatively small standard deviation indicates that most observations are relatively close to the mean, whereas a higher standard deviation indicates that observations are distributed over a greater range of values (Walpole, 1982).

Normal Distribution

The normal distribution is the most important continuous probability distribution in the field of statistics (Walpole, 1982). For a normal distribution of a population, the median, mode and mean all have the same value. Its graph, the normal distribution curve (Fig. 2.1), is a symmetric bell-shaped curve extending infinitely in both directions. The shape of the curve and its location along the horizontal axis are determined by the population standard deviation σ and mean μ of a normal random variable X. The more variable the set of observations, the lower and wider the curve will be (Walpole, 1982).
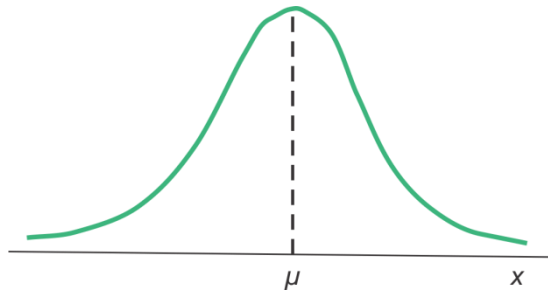
Figure 2.1 The normal curve (Walpole, 1982)

Coefficient of Variation (Relative Standard Deviation)

The coefficient of variation (CV), expressed in percent, is a normalized measure of dispersion of a probability or frequency distribution. It is independent of the magnitude of the data and thus the data measurement units. The variation of data measured in different units can be compared. It is calculated by dividing the sample standard deviation by the sample mean and multiplying by 100 (Walpole, 1982).

Analysis of Variance

Analysis of variance (ANOVA) is a method used to partition the total variation of sample data into components estimating the variation from different sources (Walpole, 1982). The total variation in a sample of lake sediments consisting of field duplicate pairs, for example, can be distributed between a component of regional variation and a component consisting of analytical and sampling variation, and expressed as percentages (Garrett, 1983). The classification of observations on the basis of a single factor, such as the element (e.g., Li, Ag, etc.) is called a one-way classification (Walpole, 1982). Results are typically shown in table form, listing along with the Source (of variation), parameters used to construct the model, i.e. Sum of Squares, Degrees of Freedom and Mean Square. The ANOVA F-test statistic (see below) is used to assess whether or not the variance within each component is equal. The p-value (see below) is used to assess the statistical significance of the F-test statistic.

ANOVA F-test Statistic

The F-test statistic is an integral part of an analysis of variance. The test statistic in an F-test is the ratio of two scaled sums of squares reflecting different sources of variability. It may be used to estimate whether the sample means are within sampling variability of each other (Dallal, 2007). These sums of squares to test the null hypothesis of equal means are constructed so that the statistic tends to be greater when the null hypothesis is not true.

F is the ratio of the 'Between' Mean Square to the 'Within' Mean Square. The null hypothesis is rejected if the F ratio is large (Dallal, 2007).

p-Value

The p-value describes the result of a fixed level test (F-test). The p-value is the probability of obtaining a test statistic (F-test) at least as extreme as the one that was actually observed by chance alone. The fixed level, set by the researcher, is often 0.05 or 0.01. The null hypothesis is rejected when the p-value is less than the fixed level, and the result is said to be statistically significant (Dallal, 2007).

Appendix 3

Notes on Analytical Procedures

Lake or stream sediments and waters, tills, rocks, vegetation, etc. are analyzed to obtain a numerical estimate of the proportion of an element or compound within a known mass of the sample medium. For example, using one analytical method, a concentration of 100 parts per million of copper is detected in 0.5 grams of the less than 177 micron (<0.177 mm) fraction of a sample of dried stream sediment. At this point it is important to remember that the rest of the 0.5 g sample, 999,900 parts or so, by weight (there will also be measurement error), is composed of other things (Grunsky, 2010).

A majority of the chemical analyses returned to the Geological Survey of Canada from commercial and government labs fall into three broad categories; 'partial', 'near-total' and 'total'. Partial analyses employ a reagent mixture that will only liberate that part of the total amount present held in particular mineralogical forms in the sample material. A frequently used 'partial' extraction is a mixture of hydrochloric and nitric acids - aqua regia. The resulting solution is passed through an instrument to estimate the concentration of elements dissolved in the solution. Examples of measurement systems are Inductively Coupled Plasma Mass Spectrometry (ICP-MS), ICP Atomic Emission Spectroscopy (ICP-AES), and Atomic Absorption Spectrophotometry (AAS). Although the concentrations of some elements estimated in this manner can be 'near-total', a portion of some elements can remain undissolved, locked in minerals resistant to dissolution. An example of a specific 'partial' analysis would be a hydroxylamine hydrochloride extraction which preferentially liberates elements held by Fe-oxyhydroxides.

The second broad category of analysis is 'near-total' and includes multi-acid digestions using hydrofluoric, nitric, hydrochloric and perchloric acids (4-acid digestion). Another 'near total' method, commonly used to dissolve more resistate minerals in samples, is heating with fusion mixtures of fluxes such as lithium metaborate, sodium carbonate or sodium peroxide in a metal crucible (e.g., platinum, silver or nickel) and dissolving the fusion product in a dilute mineral acid to liberate elements held in the insoluble residue (Thomas, 2004).

The third broad category of analysis is 'total' and includes Instrumental Neutron Activation Analysis (INAA) and X-ray fluorescence (XRF).

Minerals resistant to 'near-total' acid digestions include spinels, beryl, tourmaline, chromite, zircon, monazite, niobates, tungstates, topaz and cassiterite (Crock & Lamothe, 2011). Recently, fusion digestion procedures have been employed to analyze tills and lake sediments, using lithium metaborate as the fusion reagent in order to liberate elements held in these minerals.

The importance of recording the details of the method (digestion + analytical technique) used to estimate elemental concentrations cannot be sufficiently stressed.

The most common analytical procedure for dried sediment samples is to dissolve a small aliquot of sample, usually 1 gram or less, in acid or combination of acids (e.g., HNO_3, HCl, HF, $HClO_4$, aqua regia). A method in use that does not involve sample dissolution is instrumental neutron activation analysis

(INAA). Dried samples, usually around 30 g or so, are irradiated with neutrons in a nuclear reactor for a measured time, removed, and the resulting gamma-ray spectrum is used to identify and quantify a range of trace metals (Alfassi, 2006). Whereas most chemical methods of analysis employ small sample weights, e.g.,0.5 grams, the larger sample weight used in INAA provides a more representative aliquot of the sample, so reducing the 'nugget effect' due to the presence, or absence, of elements in discrete mineral forms.

Interference

Interference in analytical chemistry is a type of systematic error resulting from the presence of concomitant material occurring with the chemical constituent (analyte) of interest (Van der Linden, 1989). This 'other stuff' may be present in the original sample or added during the analytical procedure. Analytical results from each of the methods described in the preceding paragraphs may be subject to interference (Geboy and Engle, 2011).

Interferences in ICP-MS may be spectral-, matrix- or physical- based interferences. Spectral interferences or 'overlaps' are often associated with something added to the sample during the analytical process, such as when argon gas used to nebulize a sample solution, combines with components in the digestion of the sample material or even oxygen or nitrogen from the surrounding air. Briefly, two or more ions in the sample can combine to form molecular ions with the same mass as the ion of interest being measured. For example, in a hydrochloric acid medium, $^{40}Ar^+$ combines with the most abundant chlorine isotope at 35 atomic mass units (amu) to form $^{40}Ar^{35}Cl^+$, which interferes with the only isotope of arsenic at mass 75 (Thomas, 2004). Ways to compensate for interference have been developed that include removing the matrix before analysis using chromatography-type equipment, using mathematical interference correction equations and altering the analytical method itself (Geboy and Engle, 2011; Van der Linden, 1989).

Interference in INA analysis is usually caused by the formation of the same radionuclide from two different elements or from two radionuclides having very close gamma lines. In the first case, interference can be compensated for by identifying and measuring another isotope produced in the reaction of an element with neutrons. In the second case, more sensitive spectrometers can separate these peaks if they have similar activities (Alfassi, 2006).

Appendix 4

Using R and 'rgr' to Assess Data Quality

There are many different ways in R and 'rgr' to achieve similar goals. The scripts (in red) outlined below provide some suggested procedures among many that are available to estimate the quality of geochemical data. The output generated by executing the R function or the R command is displayed below the function/command, if applicable, in **Courier New bold font**. An explanation of what the function (or command) does is provided below the function/command or output.

I. Importing Data into R

```
>skcrdata<-read.csv("F:\\R_Work\\R-2.15.3-win\\SK_CR.csv")
```

A table created in Excel® and exported as a csv file ('SK_CR.csv') is read into R from a subdirectory 'R_Work\\R-2.15.3-win\\' on the F: drive to create a data frame named 'skcrdata'. Note that F: drive is where the files are stored for this example. Users can choose any drive or folder to store their R Work files.

```
>skcrdata.fixed<-ltdl.fix.df(skcrdata)

  n = 121 by p = 103 matrix checked, 0 NA(s) present

  3 factor variable(s) present

  865 -ve value(s) set to +ve half the negative value
```

The data frame 'skcrdata' contains negative values (less than detection limit) that are replaced with positive half-detection limit values (e.g., 1 replaces -2).[2] Blank cells and cells containing '0' are replaced with 'NA' (no information). The resulting data frame is called 'skcrdata.fixed.' The data frame consists of 121 rows plus a header row, and 106 columns, of which three are non-numerical.

```
>attach(skcrdata.fixed)
```

The data frame 'skcrdata.fixed' is now available for use by R and 'rgr' functions.

II. Evaluating the Accuracy of Data with Less Than 30 Observations

```
> crm14<-read.csv("F:\\R_Work\\R-2.15.3-win\\STSD1_4.csv")
```

A table created in Excel® of past acceptable analyses of CRMs STSD-1 to STSD-4 and exported as a csv file ('STSD1_4'.csv') is read into R from a subdirectory 'R_Work\\R-2.15.3-win\\' on the F: drive. A data frame, named 'crm14' containing analytical data for four CRMs is created.

```
> crm14.fixed<-ltdl.fix.df(crm14)

  n = 80 by p = 38 matrix checked, 0 NA(s) present

  2 factor variable(s) present

  0 -ve value(s) set to +ve half the negative value
```

---

[2] All data in the file 'SK_CR.csv' have consistent detection limits; however, low sample weights can result in variable detection limits reported for samples analysed by Instrumental Neutron Activation. For a discussion of replacement methods for such values, see Helsel (2005).

The data frame 'crm14' contains no negative values (less than detection limit) and no blank cells or cells containing '0'. The data frame resulting from this check is named 'crm14.fixed.' The data frame consists of 80 rows plus a header row, and 40 columns, two of which are non-numeric.

>names(crm14.fixed)

```
 [1] "Unique.ID" "OF"     "Rep_Stat"  "CR.ID"     "Ag_ppb"    "Al_pct"  "As_ppm" "Au_ppb"  "Ba_ppm"
[10] "Bi_ppm" "Ca_pct" "Cd_ppm" "Co_ppm" "Cr_ppm" "Cu_ppm"     "Fe_pct"  "Ga_ppm"  "Hg_ppb"
[19] "K_pct" "La_ppm" "Mg_pct" "Mn_ppm" "Mo_ppm"  "Na_pct"     "Ni_ppm"  "P_pct"    "Pb_ppm"
[28] "S_pct" "Sb_ppm" "Sc_ppm" "Se_ppm" "Sr_ppm"  "Te_ppm"     "Th_ppm"  "Ti_pct"  "Tl_ppm"
[37] "U_ppm"      "V_ppm"  "W_ppm"     "Zn_ppm"
```

'names' is an R base function that lists the column headings in the first row of a data frame, allowing the user to quickly view them.

>attach(crm14.fixed)

The 'attach' function makes the data frame available for use by R and 'rgr'.

These next two commands set up the parameters for creating a plot in R:

>par(mfrow=c(2,1))

'par' is an R base function that sets a graphical parameter in the viewing window. With the option 'mfrow=c(nrows,ncols)' a matrix of plots is specified by a vector of length 2, with the first argument a number of rows and the second the number of columns of plots. In this case there will be two plots, one above the other.

>par(pin=c(10,2.75))

'pin=c(x,y)' specifies the plot region size in inches.

> new.crm.data.Cu1<-c(64.13,67.08,68.87,67.62,60.96,60.25,64.75)

> new.crm.data.Cu2<-c(70.55)

> new.crm.data.Zn1<-c(75.0,77.6,78.9,74.3,74.6,78.4,69.8)

> new.crm.data.Zn2<-c(82.4)

The four commands above define four numeric vectors using 'c(…)', a generic function to concatenate arguments into a vector. The assignment operator '<-' points to the object (e.g., new.crm.data.Cu1) receiving the value of the expression. The values in the brackets ('arguments') are typed in by the user and are the analytical values for Zn and Cu in STSD-4 included with samples from recent surveys on Victoria Island and in New Brunswick.

>crm.plot.new(Cu_ppm[CR.ID=="STSD-4"],xname="Cu (ppm) in STSD-4",crm.mean=66,crm.sd=5,crm.new=8,cex.lab=1,ylim=c(60,80))

```
  Control Reference Material: Cu (ppm) in STSD-4

                            Mean = 64.1    SD = 3.4        RSD% = 5.3        N = 34

                            Median = 64    MAD = 3.74      rRSD% = 5.85

  CRM mean provided = 66

  CRM SD provided = 5

  Number of new CRM determinations to be plotted = 8

  Number of SDs to be used for tolerance bounds = 2
```

'crm.plot.new' is a function in 'rgr' that plots new CRM data with historic CRM data. In the above example an established mean and SD are provided from Lynch (1991) to estimate tolerance bounds for the new data. The inclusive range of the new data to be plotted, ylim, is provided to ensure that all data and tolerance bounds are plotted.

```
>points(rep(35:41),new.crm.data.Cu1,pch=8,col=2)
```

```
>points(rep(42:42),new.crm.data.Cu2,pch=8,col=4)
```

The new data are plotted into the display once the function has executed using the R plot command 'points' above. Symbols (pch) and colours (col) are specified in the command line. To view the symbols and colours available, use the functions 'display.marks' and 'display.lty', respectively. For example, pch=8 and col=2 plots the Cu1 data points as asterisks in red.

III. Evaluating the Accuracy of Data with 30 or More Observations

```
>skcrdata<-read.csv("F:\\R_Work\\R-3.1.2-win\\SK_CR_data.csv")
```

A table created in Excel® and exported as a csv file ('SK_CR_data.csv') is read into R from a subdirectory 'R_Work\\R-3.1.2-win\\' on the F: drive to create a data frame named 'skcrdata'.

```
>skcrdata.fixed<-ltdl.fix.df(skcrdata)
```

```
  n = 289 by p = 132 matrix checked, 3845 NA(s) present
  4 factor variable(s) present
  2499 -ve value(s) set to +ve half the negative value
```

A new data frame called 'skcrdata.fixed' is created. The data frame contains 2,499 negative values (less than detection limit) that are replaced with positive half-detection limit values (e.g., 1 replaces -2). Blank cells and cells containing '0' are replaced with 'NA' (no information). The resulting data frame is called 'skcrdata.fixed.' The data frame consists of 289 rows plus a header row, and 136 columns, of which four are non-numerical.

```
>names(skcrdata.fixed)
```

```
  [1] "Unique_ID"   "Number"    "Type"      "Old_OF"    "OldCRID"   "NewCRID"   "Batch.Number"
  [8] "Ag"          "Al"        "As"        "Au"        "B"         "Ba"        "Be"
 [15] "Bi"          "Ca"        "Cd"        "Ce"        "Co"        "Cr"        "Cs"
 [22] "Cu"          "Dy"        "Er"        "Eu"        "Fe"        "Ga"        "Gd"
 [29] "Ge"          "Hf"        "Hg"        "Ho"        "In"        "K"         "La"
 [36] "Li"          "Lu"        "Mg"        "Mn"        "Mo"        "Na"        "Nb"
 [43] "Nd"          "Ni"        "P"         "Pb"        "Pd"        "Pr"        "Pt"
 [50] "Rb"          "Re"        "S"         "Sb"        "Sc"        "Se"        "Sm"
 [57] "Sn"          "Sr"        "Ta"        "Tb"        "Te"        "Th"        "Ti"
 [64] "Tl"          "Tm"        "U"         "V"         "W"         "Y"         "Yb"
 [71] "Zn"          "Zr"        "Al2O3_LF"  "Ba_LF"     "CaO_LF"    "Cr2O3_LF"  "Fe2O3_LF"
 [78] "K2O_LF"      "LOI_LF"    "MgO_LF"    "MnO_LF"    "Na2O_LF"   "Nb_LF"     "Ni_LF"
 [85] "P2O5_LF"     "Sc_LF"     "SiO2_LF"   "Sr_LF"     "Sum_LF"    "TiO2_LF"   "TOT_C_LF"
 [92] "TOT_S_LF"    "Y_LF"      "Zr_LF"     "Ag_AAS"    "As_COL"    "As_INA"    "Au_INA"
 [99] "Ba_INA"      "Br_INA"    "Cd_AAS"    "Ce_INA"    "Co_AAS"    "Co_INA"    "Cr_INA"
[106] "Cs_INA"      "Cu_AAS"    "Eu_INA"    "F_ISE"     "Fe_AAS"    "Fe_INA"    "Hf_INA"
[113] "Hg_CV_AAS"   "La_INA"    "LOI_GRAV"  "Lu_INA"    "Mn_AAS"    "Mo_AAS"    "Mo_INA"
[120] "Na_INA"      "Ni_AAS"    "Ni_INA"    "Pb_AAS"    "Rb_INA"    "Sb_INA"    "Sc_INA"
[127] "Sm_INA"      "Ta_INA"    "Tb_INA"    "Th_INA"    "U_NADNC"   "U_INA"     "V_AAS"
[134] "W_INA"       "Yb_INA"    "Zn_AAS"
```

The R command 'names(data_frame)' lists the column headers.

```
>attach(skcrdata.fixed)
```

The command 'attach(data_frame)' is used to put the data frame (or 'object') into the R environment for analysis.

```
>table(NewCRID)
```

```
NewCRID
    9m      9p      9r LKSD-1 LKSD-2 LKSD-3 LKSD-4
    13      15      14      68      70      76      33
```

The R command 'table(variable_name), e.g., NewCRID, lists each reference standard by column and the number of occurrences in the row below, as shown above.

```
>crm.plot(Cu[NewCRID=="LKSD-1"],"Cu
(ppm)",crm.mean=44,crm.sd=5,n.sd=2,pch=19,cex.lab=0.75,cex.axis=0.8)
```

```
 Control Reference Material: Cu (ppm)
                                 Mean = 44       SD = 5          N.SD = 2      Percent
tolerance =
 This project:    N = 68        Mean = 43.1     SD = 3.32       RSD% = 7.7
                                Median = 42.7   MAD = 2.39      rRSD% = 5.59
```

The 'rgr' command 'crm.plot' calculates a mean, standard deviation and relative standard deviation for 68 determinations of Cu in the Certified Reference Standard 'LKSD-1'. A median, median absolute deviation (MAD) and robust relative standard deviation (rRSD%) (Reimann et al., 2008) is also calculated. An input of the accepted mean and standard deviation, from Lynch (1999), is required, as well as an acceptable range of variation, in this case 2 standard deviations from the mean. Commands 'pch' (plotting character), 'cex.label' and 'cex.axis' (character expansion) sets graphical parameters for the plot produced by crm.plot, reproduced in Figure 7 of the text part of this Open File. An additional command, 'crm.tol' can be used to set a percentage tolerance level for the maximum acceptable absolute relative percent difference from the CRM recommended value ('crm.mean'). If no value is set, 'Percent tolerance' is left blank.

IV. Estimating Precision with Analytical Duplicate Data in rgr

Table 2 in the main body of this open file provides an estimate of precision measured as % Relative Standard Deviation (Precision (RSD %)) for analytical duplicate pairs. In order to put this number in context, the percentage of all values below detection (% Below LDL), the number of duplicate pairs removed from the calculation because one or both values are below detection (Duplicate Pairs Removed), the Range and the Mean are listed with the element name and method and the lower detection limit (LDL).

In the example below, a comma-separated value (csv) file of analytical duplicate data is exported from Excel 2010® and imported into R. Because the format of this file has alternating rows consisting of the first field duplicate data followed by the second field duplicate data (rather than alternating columns) ad.plot2 in 'rgr' is used. Negative values representing values below detection limit are replaced with 'NA' by using the function ltdl.fix.df with the argument negs2na=T. ad.plot2removes pairs of values if one or both values equal NA before performing the calculations.

```
>skaddata<-read.csv("F:\\R_Work\\R-3.1.2-win\\skad_data.csv")
```

A table created in Excel® and exported as a csv file ('skad_data.csv') is read into R from a subdirectory 'R_Work\\R-3.1.2-win\\' on the F: drive to create a data frame named 'skaddata'. Note that F: drive is where the files are stored for this example. Users can choose any drive or folder to store their R Work files.

```
>skaddata.fixed<-ltdl.fix.df(skaddata,negs2na=T)
```

```
 n = 578 by p = 89 matrix checked, 2 NA(s) present
 3 factor variable(s) present
 5293 -ve value(s) set to NA
```

The data frame 'skaddata' contains 5,293 negative values (less than detection limit) that are replaced with 'NA' (e.g., NA replaces -2). The resulting data frame is called 'skfadata.fixed.' The data frame consists of 578 rows plus a header row, and 89 columns, of which three are non-numeric.

>attach(skaddata.fixed)
The data frame 'skcrdata.fixed' is now available for use by R and 'rgr' functions.

>ad.plot2(Cu,ad.tol=20,ifalt=T,if.order=F,log=T)

```
 Analytical duplicates for: Cu
 This batch:      N = 289              Mean = 27.7    SD = 1.59        RSD% = 5.76
                                       Median = 21.8  MAD = 0.511      rRSD% = 2.34
```
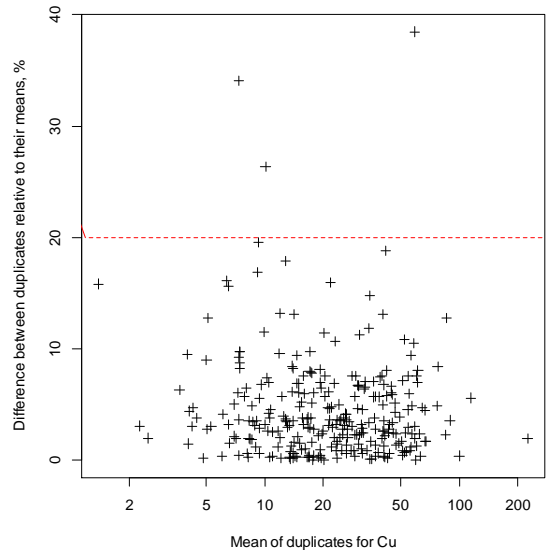

Figure 4.1 Shewhart plot for Cu in lake sediment analytical duplicates produced by the function 'ad.plot2' in 'rgr'.

The estimate of precision (Precision RSD %) is read from the output of 'ad.plot2', shown in the example for Cu. A Shewhart plot (Fig. 4.1) is also created with the function ad.plot2 where the percent relative differences are plotted against the duplicate pair means: in this example a 20% acceptance level is indicated by the red line, and three duplicate pairs require examination.

>ad.plot2(Be,ad.tol=20,ifalt=T,if.order=F,log=T)

```
  17 row(s) with NA(s) removed from matrix

 Analytical duplicates for: Be

 This batch:      N = 272              Mean = 0.643   SD = 0.187       RSD% = 29.1

                                       Median = 0.55  MAD = 0.0741     rRSD% = 13.48
```

In cases where there are analytical values below detection, represented by 'NA' in the data frame, 'ad.plot2' counts the number of pairs of analytical duplicates with one or both values below detection, and removes them from the calculations. In the example above for Be, 17 pairs of values were removed before generating statistics on the remaining 272 pairs of values.

```
>gx.summary2(Be)
```

```
  Extended Summary Stats for: Be
  N and number of NAs:          559 & 19
  Arithmetic Mean and 95% CLs:  0.6331 & 0.597 <-> 0.6692
  SD and CV%:                   0.4345 & 68.64%
  Geometric Mean and 95% CLs:   0.5177 & 0.4909 <-> 0.546
  Log10 Mean and SD:            -0.2859 & 0.2784
  Median and 95% CLs:           0.5 & 0.5 <-> 0.6
  MAD and IQR estimates of SD:  0.2965 & 0.3706
  Percentiles: Min  2   5   10   25 - 50 - 75   90   95   98   Max
 0.1 0.1 0.2 0.2 0.3 - 0.5 - 0.8 1.2 1.5 1.8 2.9
```

The total percentage of values below detection in the example below for Be is derived from the output of function gx.summary2, shown above, by dividing the number of NAs (19) by N (578) and multiplying by 100. The range of the data above the detection limit is equivalent to the minimum (Min) and maximum (Max) values of the percentiles, and the Mean is read from the 'Arithmetic Mean and 95% CLs:' line.

```
>ad.plot4(Cu,"Cu (mg/kg) in lake sediment",ifalt=T,if.order=F,if.text=T)
```

```
  Range of ratios of the 289 duplicates: 0.767 1.48
  Median ratio = 0.9961        MAD of ratios = 0.04919
  Mean ratio = 0.9989    SD of ratios = 0.06956  95% CI for ratios = 0.1369

  Absolute ratio difference from 1 = 0.001091    SE of Mean of ratios = 0.00409  95% CI =
0.00805
  Mean ratio is not different from 1 at the 95% level, no bias

  95% of duplicates will fall between factors of 1.14 and 0.88 times a value
    Robust factor estimates based on the MAD are 1.1 and 0.91
```

The function 'ad.plot4' plots analytical duplicate pair data as ratios against their means on a logarithmic scale. The 95% standard error interval on the mean ratio are estimated and used to test that the mean ratio is not significantly different from 1. The 95% confidence bounds are shown as dotted lines. See text and Garrett (2015b) for a detailed description. In addition, a plot of the duplicate means against the ratios of the means is produced, shown for Cu in northern Saskatchewan lake sediment analytical duplicate pairs in Figure 8.

V. Analysis of Variation (ANOVA) in rgr

In the example below, a comma-separated value (csv) file of field duplicate data is exported from Excel 2010® and imported into R. Because the format of this file has alternating rows consisting of the first field duplicate data followed by the second field duplicate data (rather than alternating columns) anova2 in 'rgr' is used. Negative values representing values below detection limit are replaced with 'NA' by using the function ltdl.fix.df with the argument negs2na=True. Anova2 removes pairs of values if one or both values equal NA before performing the analysis of variation calculations.

```
>skfddata<-read.csv("F:\\R_Work\\R-3.1.2-win\\skfd_data.csv")
```

A table created in Excel® and exported as a csv file ('skfd_data.csv') is read into R from a subdirectory 'R_Work\\R-3.1.2-win\\' on the F: drive to create a data frame named 'skfddata'.Note that F: drive is where the files are stored for this example. Users can choose any drive or folder to store their R Work files.

```
   n = 578 by p = 85 matrix checked, 38 NA(s) present

   4 factor variable(s) present

   5283 -ve value(s) set to NA
```

The data frame 'skfddata' contains 5,283 negative values (less than detection limit) that are replaced with 'NA' (e.g., NA replaces -2). The resulting data frame is called 'skfddata.fixed.' The data frame consists of 578 rows plus a header row, and 86 columns, of which four are non-numerical.

>attach(skfddata.fixed)

The data frame 'skfddata.fixed' is now available for use by R and 'rgr' functions.

>anova2(W,"W (ppm)",log=T,ifalt=T)

```
   4 row(s) with NA(s) removed from matrix

 Combined Sampling and Analytical, or Analytical Variability, Study,
Utilizes Field Sampling or Laboratory Duplicates.  In ANOVA Tables,
the variability:
   Between would be between sampling sites or analysed samples, and
   Within would be at sampling sites or due to duplicate analyses

 Data have been Log10 transformed for the ANOVA

Two-Way Random Effects Model for W (ppm)
Source           SS            df        MS             F        Prob
Between          80.589        284       0.28376     71.24    0.0189
Within           3.1862        1         3.1862      19887.89   0
Residual         0.045499      284       0.00016021
Total            83.821        569       0.14731
```

```
One-Way Random Effects Model for W (ppm)
Source           SS            df        MS             F        Prob
Between          80.589        284       0.28376               25.03    0
Within           3.2317        285       0.011339
Total            83.821        569       0.14731


Source           MS            Var Comp          %age
Between          0.28376       0.13621           92.3
Within           0.011339      0.011339          7.7
  0.14755
```

```
 Summary Statistics for W (ppm)
 Grand Mean =     -0.40545              Variance =      0.14731
 'Error' S^2 =    0.011339              Std. Dev. =     0.10649
 'Error' RSD% =   -26.3
Miesch's V =     12.01         Vm =             24.03
```

The output highlighted in yellow is used to populate the ANOVA table shown in the main text of the open file (Table 3).