CHAPTER FIVE

Data Conditioning of Environmental Geochemical Data: Quality Control Procedures Used in the British Geological Survey's Regional Geochemical Mapping Project

C. C. Johnson,* E. L. Ander,* T. R. Lister,* *and* D. M. A. Flight*

Contents

Abstract

Data conditioning procedures involve the verification, quality control and data levelling processes that are necessary to make data fit for the purpose for which it is to be used. It is something that has to be planned at the outset of any project generating geochemical data. Whether it is in the sampling phase, for example, determining how sites and samples should have a unique identity, or through to the data presentation phase in

* British Geological Survey, Keyworth, Nottingham, NG12 5GG, UK

Environmental Geochemistry
DOI: 10.1016/B978-0-444-53159-9.00005-X

which disparate data sets may have to be joined to form a seamless map. This account describes the methods currently used by the British Geological Survey's regional geochemical mapping project that has been generating geochemical data for various sample media for nearly 40 years. It is important that users of the data are given information that will help them ascertain whether the provided environmental data is suitable for the purpose of its intended use.

1. Introduction

Data conditioning is the process of making data fit for the purpose for which it is to be used. Users of the geochemical data need to have confidence that anomalous results are not an artefact of the sampling or analytical method. The results need to be interpreted in the context of existing data whether it is as a comparison with statutory concentration levels for an element or using the data in a regional context. Failure to condition the data has both time and cost consequences that can easily be avoided if a system of quality assurance is followed. The data conditioning process is a three-stage process.

Initially, data will undergo a series of error checking and verification procedures that relate to the number of samples submitted; the methods used; the elements requested; element ranges; limits of detection; absent, not determined and not detected results; and mis-numbering errors. These procedures are essentially a check of collation errors and that the laboratory has carried out what they were asked to do. This checking and verification phase is generally completed before the laboratory receives payment for their work. This is referred to in subsequent discussions as the 'raw data checking' phase where the term 'raw data' is given to the analytical results as recorded by the laboratory. A second phase of processing involves a series of plots and statistical tests of the data that measure the accuracy and precision of the results and attempts to attribute their sources of variability. Finally, the data may need to be levelled in order to be joined with existing data from earlier batches that may even have been derived using different analytical procedures.

This account of data conditioning describes methods used by the British Geological Survey's (BGS) Geochemical Baseline Survey of the Environment (G-BASE) project (Johnson *et al.*, 2005). This has been a long running high-resolution geochemical mapping programme to establish a geochemical baseline for the land area of Great Britain and Northern Ireland. This is achieved principally by the collection of drainage samples (stream sediments and waters), but other environmental samples such as soils are also collected. Rock samples are not routinely collected but procedures described here would equally apply to a range of sampling media. Quality control throughout the sampling and analytical phases has always been identified as crucial to the production of reliable environmental data and has been an important part of the programme since its inception (Plant, 1973; Plant and Moore, 1979; Plant *et al.*, 1975). These procedures are applicable whether for specific site investigations or regional mapping and are in addition to any internal procedures the analytical laboratory may operate as a condition of its accreditation certification.

Quality control commences in the planning phase of a project and discussion here begins with decisions that need to be made before any samples are collected or analysed.

2. Planning Quality Control—Quality Assurance

Quality assurance is the planned and systematic activities necessary to provide adequate confidence that the service (or product) will satisfy given requirements for quality. Quality control is the combination of operational techniques and activities that are used to fulfil the requirements for quality (Potts, 1997). Such activities must be planned at the outset of any project.

2.1. Appropriate and well-documented procedures

If environmental data are to be fit for the purpose for which they are to be used, then the media and methods of sampling and chemical analysis must be appropriate to the objectives of the project. Additionally, the environmental samples collected and the subsequent analyses are a long-term national asset, particularly when undertaken by public-funded organisations, and every effort should be made to ensure procedures conform to nationally and internationally recognised standards. This is the sentiment expressed by Darnley *et al.* (1995) in the IGCP (International Geological Correlation Project) 259—'A global geochemical database for environmental and resource management'. Although IGCP 259 was concerned with the production of a global database, the recommendations covering all aspects of sampling through to data management form a useful generic guide that should be referenced when planning any environmental sampling programme. At a project level, there should be procedures manuals that should be referenced in the meta-data attached to the database of results. Users of the environmental data can then assess for themselves whether results are appropriate for their needs. The G-BASE project has a field procedures manual (Johnson, 2005), a field database manual (Lister *et al.*, 2005) and a data conditioning manual (Lister and Johnson, 2005). The Geochemistry Database now includes a sampling protocol code which references the version of the field manual current at the time of sampling.

Environmental site investigations in recent years have largely been legislatively driven. In Europe, this has been in response to European Commission environmental directives (e.g., Water Framework Directive and Sewage Sludge Directive (EC, 1986, 2000)) or national legislation such as the United Kingdom's Environmental Protection Act (1990) Part IIa (DETR, 2000). Such legislation means that laboratories produce results to a certified standard to guarantee the quality of the results such as MCERTS accreditation. MCERTS[1] is the England and Wales Environment Agency's Monitoring Certification Scheme (Environment

[1] www.mcerts.net

Agency, 2006). The scheme provides a framework within which environmental measurements can be made in accordance with the Agency's quality requirements and this includes documentation of the sampling and analytical procedures. Internationally, laboratories use the ISO/IEC 17025 Standard (British Standards, 2005) to implement a quality system aimed at improving their ability to produce valid results consistently. As the Standard is about competence, accreditation is simply formal recognition of a demonstration of that competence.

2.2. Sample numbering

A simple but fundamental requirement in any environmental sampling programme is that the sample can be identified with a unique and meaningful identity referred to here as the sample number. Potentially, sample numbering errors present one of the highest risks to the production of high-quality data. A good sample numbering system covering the way in which sample numbers are attributed and labelled is a very important part of ensuring a reliable output of data. When an organisation carries out many discrete projects, the work must conform to a sample numbering system otherwise there is high likelihood of duplicate sample numbers, particularly as the majority of projects will commence sampling using sample number 1.

The BGS uses a combination of field codes to give a sample a unique identification (Fig. 5.1) as prescribed by the rules governing sample numbering for the BGS corporate Geochemistry Database (Coats and Harris, 1995; Harris and Coats, 1992). The complete sample number is composed of (i) a single digit numeric sample numbering system code that distinguishes between the main programme areas of work that generate geochemical results (i.e., mineral exploration, regional geochemical mapping or environmental investigations); (ii) a 1–3 character alphanumeric code known as the project or area code that identifies the area being sampled; (iii) a site number recorded as an integer number from 1 to 99999; and (iv) a single character alphabetic sample type code, for example, 'S' for soil, 'W' for water or 'R' for rock. These field codes can be used to search and retrieve samples from the BGS database that currently holds more than half a million sample records.

The field sampling protocol should ensure that there are quality control measures in place to check that the sample identities attached to samples are both legible and permanent before their dispatch to laboratories. An analyst misreading an ambiguous sample label is a common source of error. Sample-labelling problems can be dealt

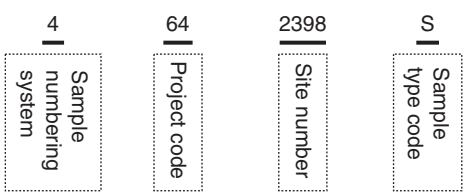

Figure 5.1 Example of a Geochemical Baseline Survey of the Environment (G-BASE) sample number defined by four key fields to give a unique sample identity.

with rapidly if the samples are accompanied by sample lists. Such lists are essential to create an audit trail, so the point at which errors are introduced can be established.

A further quality control measure that needs to be initiated at the planning phase of the project is the use of a random numbering system when collecting the samples. Systematic sample preparation and analysis errors can more easily be identified if the samples are collected in a random number order then prepared and analysed in sequential number order (Plant, 1973). Figure 5.2 illustrates this point showing that when collecting samples in sequential number order it is more difficult to distinguish

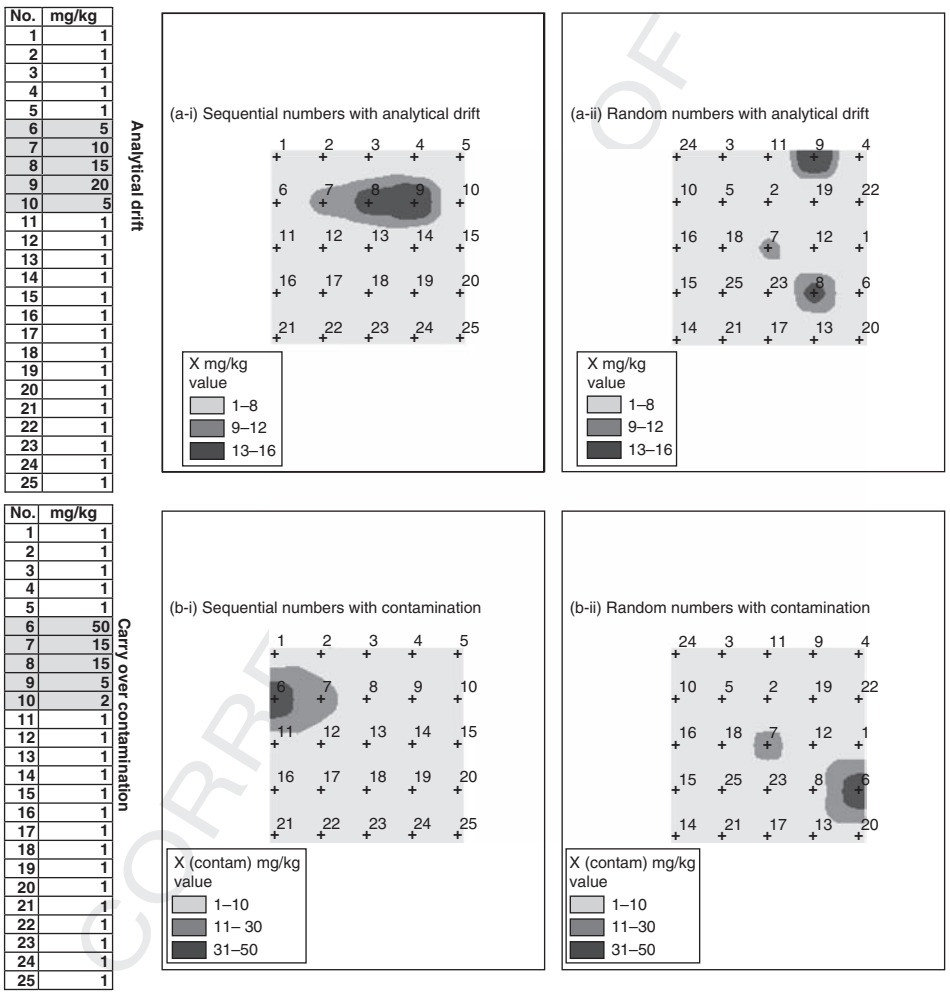

Figure 5.2 Examples of the laboratory systematic errors when samples are collected in numeric and random number order. (a) Shows error due to analytical drift and (b) 'carry over' contamination during sample preparation after the preparation of a 'high' sample. The random numbering of samples makes it easier to identify systematic laboratory errors. If the samples are collected and analysed in sequential order the results look like naturally occurring anomalies while the random numbering method tends to distinguish such errors as isolated highs.

systematic laboratory errors from naturally occurring trends than if the sample numbering was random. When using a random number sampling method it is important that the laboratories are requested to analyse the samples in numerical order.

A key to quality control of environmental analyses is the insertion of 'blind' (hidden) control samples (duplicates, replicates and reference materials) among the routinely collected samples. The control samples need to be allocated sample numbers that make them indistinguishable from the normal samples when submitted for analysis. This can be achieved by the use of sample number list sheets such as that illustrated in Fig. 5.3.

There should also be a field in the field database that indicates which sample numbers are control samples, though this suffix to the sample number should not be given to the analysts. In the G-BASE field database, there is a field called SAMP_STD which is populated with the codes shown in Table 5.1. This field can be used to quickly retrieve all control samples from the data set when data are stored in a database.

2.3. Control samples

A control sample is a sample that is inserted into a batch of samples during the process of sampling or analysis for the purpose of monitoring error, precision and accuracy. There are four main categories of control samples. These are

(i) *Duplicate*: A duplicate sample is collected from the same site as another sample in a manner defined by the project's sampling procedures manual. It is a control sample that can be used to show the amount of variability in results that can be attributed to the process of sampling by collecting two samples from the same location. A duplicate sample collected in the field is also referred to as a 'field duplicate'.

(ii) *Replicate*: This is a control sample created in the laboratory by dividing a sample into two identical parts according to a well-defined protocol. It is used to help define laboratory error. It can also be referred to as a 'sub-sample' (Fig. 5.4).

(iii) *Reference material*: This is a sample that has been prepared and analysed to acceptable documented procedures to give analytical results that through repeated analysis become accepted values. They can be used to indicate the precision and accuracy of results. When an international certified reference sample is determined, results can be published in the context of a sample for which results are recognised internationally. Reference materials should be of a similar composition to the samples being analysed. Reference materials can be subdivided into primary and secondary reference materials (PRMs and SRMs). The PRM is an international reference standard, whereas the SRM will generally be a project created reference material which is submitted at more frequent intervals and blind to the analyst.

(iv) *Blank*: This is a control sample generally only submitted for water analysis where better than 18 MΩ quality water is handled and included with a batch of normal samples. This helps to indicate any trace contamination that may be introduced during the handling, bottling, acidification or storage parts of the

Random number list 1 sediment

Project code.....

Number range.....

18		01		36		29	
49		99		70		73	
46		03		59		43	
41		38		88		82	
32		91		66		55	
45		67		64		14	
94		07		52		87	
98		34		79		06	
56		89		05		12	
15		83		60		92	
26		95		08		02	
19		21		96		63	
39		84		25		31	
28		93		47		53	
54		40		100		27	
62		71		24		30	
80		57		77		11	
16		61		09		76	Duplicate A
17		48		85		81	Duplicate B
72		35		50		86	Sub sample A
65		13		33		78	Sub sample B
37		51		42		68	Standards
04		97		20		22	
90		74		58		10	Blank waters
69		23		44		75	

Figure 5.3 Example of a Geochemical Baseline Survey of the Environment (G-BASE) random number list used for issuing site numbers with 8% of the numbers reserved for duplicates, replicates (sub-samples), standards and blanks. The blank column is used to record details of the date and sampling pair assigned to each number.

water sampling and analysis procedures. G-BASE project submits two blank waters in every hundred water samples, one of which is filtered (as the samples are) and one which is unfiltered blank water.

Table 5.1 Table showing the codes used in the Geochemical Baseline Survey of the Environment (G-BASE) field database to identify control samples

Code	Sample description
DUPA	Duplicate A (original sample)
DUPB	Duplicate B (collected at same site as Dup A)
DUPC	Duplicate C (original sample)
DUPD	Duplicate D (collected at same site as Dup C)
REPA	Sub-sample A (laboratory replicate of DUPA)
REPB	Sub-sample B (laboratory replicate of DUPB)
REPC	Sub-sample C (laboratory replicate of DUPC)
REPD	Sub-sample D (laboratory replicate of DUPD)
STD	Secondary reference material (SRM)
BW	Blank water used only for W

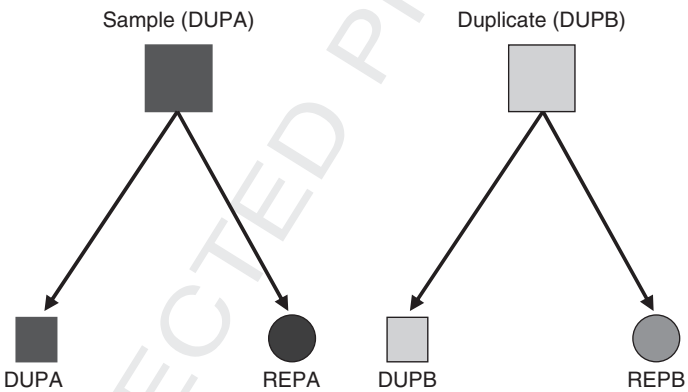

Figure 5.4 Figure showing the relationship between field duplicates and the laboratory replicates.

The number of control samples required for each batch of analyses depends on a number of factors, including the number of samples to be analysed in relation to the number of analytical batches, whether analytical batches are analysed over a continuous short period or intermittently over a longer period and the data quality requirement to which the data is being applied. For the global geochemical database for environmental and resource management, Darnley *et al.* (1995) recommend the inclusion of 3% duplicate samples and 4% SRMs. Eight percent of the samples the G-BASE project submits for analysis are control samples (see Fig. 5.3). This project generally collects 1000s of samples each field campaign and the number of control samples is sufficient to give substantial quality control data, particularly for graphical plots that require a range of concentrations to give meaningful linear regressions. When

the project has been involved in smaller sampling projects collecting only tens or hundreds of samples the number of duplicates and replicates is increased to 8 per 100.

3. Raw Data Checking

3.1. Data checking

The initial phase of data checking consists of simple and obvious procedures that need to be carried out as the first stage of acceptance of the results from the laboratory. This needs to be done systematically and as soon after the results are received as is possible. The original digital file of results as received from the laboratory should always be archived unchanged and the subsequent modifications described in the following sections performed on a copy of the original data file.

(i) Has the laboratory reported all the analytes it was contracted to do by the methods agreed in the contract?

(ii) Does the number of samples submitted for analysis correspond with the number of data records received and do all the sample numbers correspond to the sample number list submitted?

(iii) Are the results given with the correct units, absent, outside limit or not detected analyses reported correctly? Have elements reported as oxides been reported as percentages? Reporting results as a mixture of percentage element oxide and element milligram per kilogram can be misleading, particularly for non-geochemists interpreting the results.

(iv) Has the laboratory correctly reported control samples that were not blind to the analyst?

(v) Does the range of values for each element reported look sensible for the area sampled?

(vi) If the samples were collected in random number order but analysed in sequential number order is it possible to identify any analytical drift or carry over contamination from high samples?

3.2. Dealing with missing, semi-quantitative and unreliable data

During the initial data checking phase missing, non-numeric, coded and unreliable results need to be dealt with. To correct bad data you first must find them (Albert and Horwitz, 1995). At this stage, it is simply a case of looking at the data to identify obvious misfits such as results that are way out of expected ranges or control samples that clearly do not conform to expected values. If there is a problem with control samples, then the reliability of the complete data set must be questioned and the matter taken up with the analyst. There is little point proceeding with any data conditioning unless the results can be deemed acceptable. However, the acceptance of the results at this stage is conditional on precision and accuracy tests discussed later in this account. Suspicious analyte determinations can be tagged at this stage with a qualifier or removed altogether until the interpretative phase of the project is complete. The G-BASE project qualifies such

analyte determinations with a '*' in the final database so users can see that there is a documented data quality issue associated with the result. Outlying results may have no explanation as to their cause but are still of great value to the data user.

Different laboratories use different codes to represent missing or semi-quantitative data, examples from the BGS XRFS (X-ray fluorescence spectroscopy) laboratory are shown in Table 5.2. The Analytical Methods Committee of the Royal Society of Chemistry makes a distinction between the *recording* of results and the *reporting* of results (AMC, 2001a, 2001b). Recording of results just as they come, including negative and below detection values, is how the laboratory produces the data. The reporting of results, that is, the data passed onto the customer, is a contractual matter between the analyst and the customer. Most types of statistical processing of data sets containing low concentrations of analyte should be undertaken on the unedited data, that is, the data should be reported in the same way as it is recorded.

For a site investigation, where the question might be whether or not a result falls above or below a statutory level, the problems of below detection limit values may seem of little consequence if the method of analysis can satisfactorily determine results to the level of the statutory value. However, if an assessment is to be based on the average value of a data set it is important that no bias is added to the results by substituting below detection limits or over-range values with a constant that may introduce a high or low bias into the censored (that is modified) data. The laboratory must produce a statement of analytical uncertainty that would usually include a value for each element below which results have a high uncertainty, that is, a detection limit. It is important that such information is passed on to the data user. With modern analytical instrumentation delivering digital results via complex analytical software it is difficult to often pin the analyst down to a single limit of detection. For example, some elements determined for the G-BASE project are done by ED (energy dispersive)-XRFS, and this method will record variable detection limits for each element for every sample depending on the compositional matrix of the sample medium making it difficult for the analyst to cite a single detection limit for the entire analytical batch.

The low-detection limits as reported by analysts are generally more conservative than are the actual limits. This is suggested by the slope of the curve on the data

Table 5.2 Table showing the codes used by the British Geological Survey X-Ray fluorescence spectroscopy (BGS XRFS) laboratory to represent absent or semi-quantitative results

Code	Comment
–94	Insufficient sample (e.g., sample collected but not enough to analyse)
–95	Not determined because of high concentration, but exceeds calibration limit
–96	Not determined because of interference; probably of high concentration
–97	Not determined because of interference; probably of low concentration
–98	Not determined because of interference; no estimate
–99	Absent data (e.g., not requested, sample not submitted or sample lost)

shown in Fig. 5.5. This data represents Al concentrations in the same 10,000 samples, determined by two different methods of analysis. The slope of the graph suggests no sharp change in curvature until a concentration well below the reported detection limit. This indicates that there may be some regional pattern which can be usefully obtained from this data, and that the practical detection limit may be lower than has been reported by the laboratory. However, it should be recognised that the actual concentrations have much higher uncertainties associated with them when concentrations are low with respect to the detection limit. There are statistical plotting methods that can be used to formally determine more realistic levels of detection. These include a method that plots duplicate difference against duplicate means (Thompson and Howarth, 1973, 1976, 1978) and determining limits of analytical detection by identifying bottom truncation on cumulative probability plots (Sinclair, 1976). However, these graphical methods do require a good range of element concentrations with at least some values approaching or below detection, such as is shown in Fig. 5.5 where bottom truncation is indicated by the curve becoming a horizontal line.

Historically, the G-BASE project has replaced values recorded as 'detection' by a value one half the detection limit for soils and stream sediments and for consistency with older data. This practice continues. There is no sound basis for such a remedy

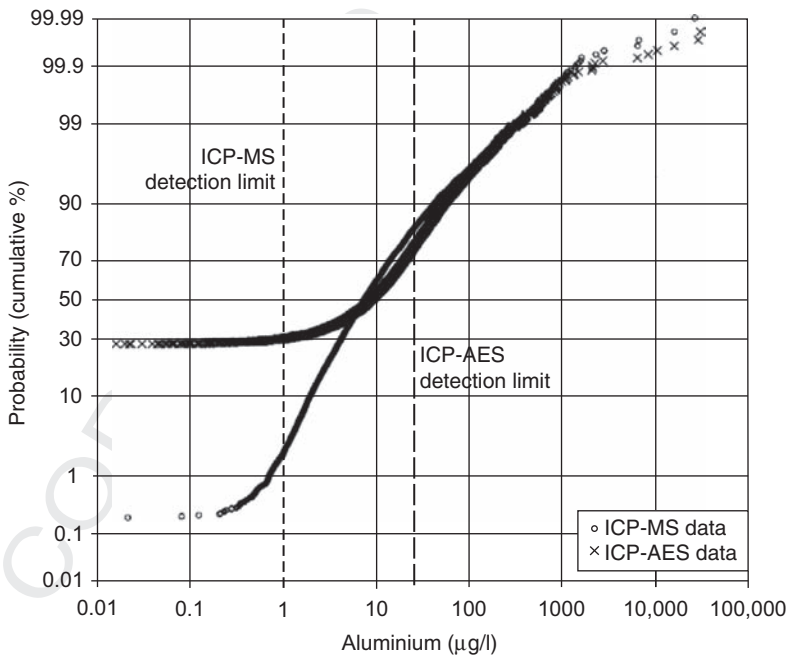

Figure 5.5 Cumulative probability plot indicating true detection limits for samples determined by two different analytical methods. Such plots can demonstrate that the real detection limits are often much lower than those cited by the analyst (the cited detection limits are marked on the plot). Data for 10,000 samples from part of central and eastern England.

except the acceptance that the real value probably lies between the detection limit and zero and the data needs to have numeric rather then semi-quantitative or alpha-numeric values in order to use the data in statistical analyses. For stream waters, for which there is a shorter history of collection and analyses for a wide range of analytes, data are not modified by the application of the detection limits before storage in the geochemistry database. The benefit of this lies in the lack of bias introduced into any statistical tests for which the data are used. Clearly, this does not preclude later censoring of the data where required by the user. Inherent in this approach is the requirement to have a field in the database specifying the detection limit for each analyte in each analytical batch. For instance, the approaches described above to assess the reported detection limit against the data supplied could not possibly be undertaken on censored data.

Current procedures emphasise the importance of storing the laboratory recorded data ('raw data') and the XRFS analyses are now directly transferred via a Laboratory Information Management System (LIMS) to data tables in the BGS corporate geochemistry database. Conditioned data is loaded to different data tables and importantly every analyte result has an accompanying qualifier field that can be provided to the user to explain any data quality issues.

So far, all outside limit discussion has been concerned with the lower rather than the upper detection limit. Similar procedures apply with the data being qualified with a code and the value set to the upper detection limit. Generally, semi-quantitative 'above detection' values are less of a problem than the 'below detection' values. This is because the analysts are better able to find simple remedies for very high results than they are for very low results. For example, when an extractive procedure is used less sample can be used in the analysis, or the final solution being determined can be diluted. With low concentrations increasing the sample weight for analysis or reducing dilution results in increased interference problems. The upper limit of elemental totals should not exceed 100%, though in reality uncertainty in results for abundant oxide values (e.g., silicon) can lead to totals >100%.

4. Statistical Analyses and Plotting of Control Sample Data

Once the preliminary error checking of the raw data has been done, the control samples should be separated from the normal samples for more detailed examination. This process of separation is greatly aided by the inclusion of the STD_SAMP field in the field database (see earlier) and a comprehensive sample list that identifies control samples and their relationships (Fig. 5.3). Control sample results can then be subjected to a number of statistical and plotting procedures that determine the accuracy and precision of results. These processes give an indication of the levels of uncertainty that are associated with the results, information that is essential to interpret the data and present it in a meaningful manner.

4.1. Control charts

G-BASE uses control charts (Shewhart plots) that are time sequenced quantitative data plotted as a graph that has fixed defining limits (Miller and Miller, 2005). The upper and lower limiting lines are usually defined as the mean (accepted) value for an analyte ±3 standard deviations. Both PRMs and SRMs are repeatedly analysed over a period by being included in every batch of samples submitted for chemical analysis. The results can be plotted on the control chart to see how close the result falls to an accepted value and whether any analytical drift is recognisable over a sequence of successive analytical batches. The G-BASE project has used a number of software packages to plot control charts, the simplest being MS Excel. Specialist software packages such as the MS Excel add-in SPC XL (Fig. 5.6) and QI Analyst (Fig. 5.7) will flag values that exceed the upper or lower limits or whether there is an upward or downward drift in results. If a reference material result is suspect, that is, it plots outside the defined limits or a significant shift or drift in values is shown, then all the results from the analytical batch should be treated as suspect and, in the absence of any satisfactory explanation, the whole analytical batch should be reanalysed.

4.2. Duplicate–replicate plots

An effective but simple way of graphically illustrating the variability associated with the analytical data is to plot x–y plots of the duplicate and replicate pairs. Most statistical packages will have an option for plotting simple x–y plots. The G-BASE project uses MS Excel running a macro that will automatically plot duplicate-replicate and duplicate-duplicate results. Figure 5.8 shows three examples from the G-BASE East Midlands atlas area duplicate–replicate data for soils. This method gives an immediate visual appreciation of any errors present in an analytical batch and an indication of 'within site' variability, as shown by the duplicate pairs, or the 'within sample' variability, as indicated by the replicate pairs that demonstrate

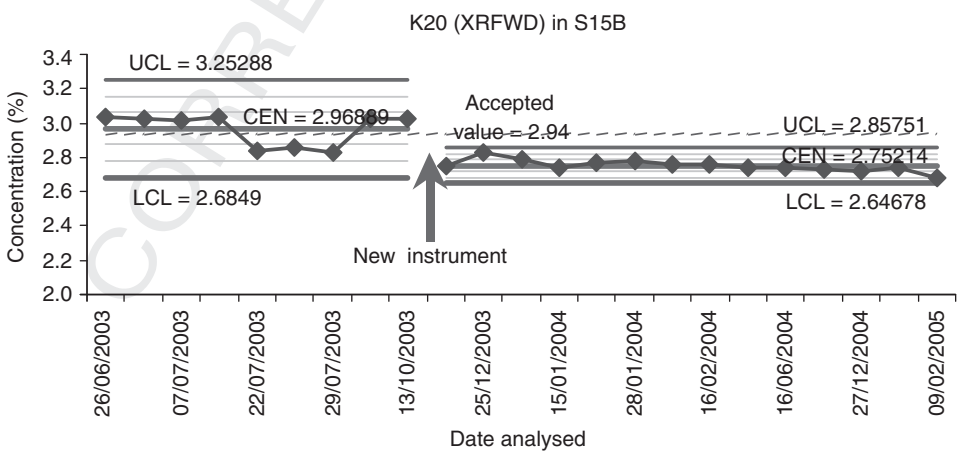

Figure 5.6 Example of output from the SPC XL control chart plotting MS Excel add-in.

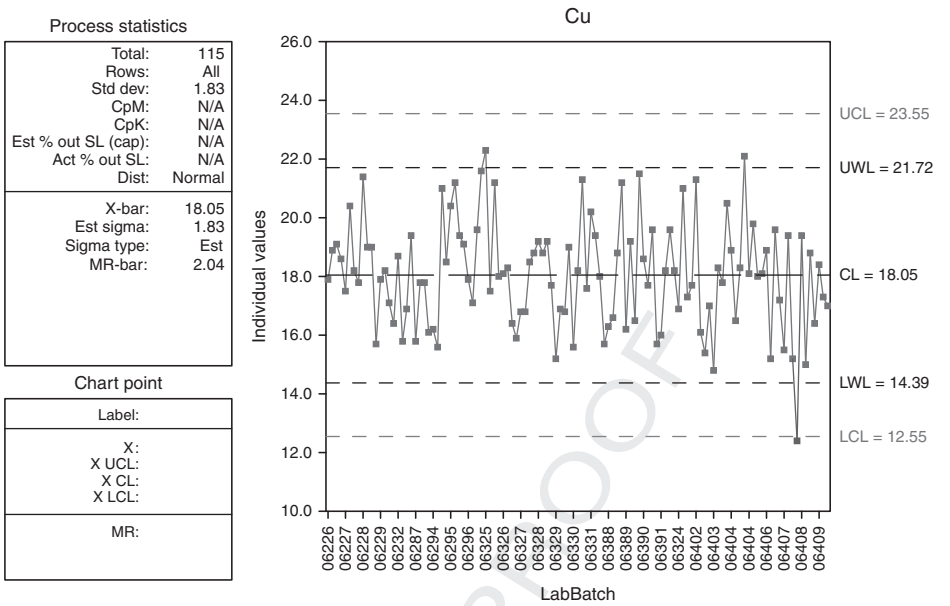

Figure 5.7 Control chart plot using QI Analyst software (Cu in mg/kg on *y*-axis).

analytical uncertainty. In the examples shown in Fig. 5.8, the uranium results show a broad spread about the line of gradient 1 (the lines in Fig. 5.8), particularly around the detection limit, indicating that for this element there is high sampling and analytical variance. The iodine results show generally low analytical variance (shown by the dup-rep plot) but have higher within site variability, particularly at higher concentrations. The copper data shows a good correspondence with the line of gradient 1 suggesting that for this element both the within site and within sample (i.e., analytical variability) are low.

In instances where only a few duplicate–replicate pairs are available, individual results can still be plotted to see where they fall relative to the line of gradient 1. For larger surveys, collecting many duplicate–replicate pairs, a good spread of values can be produced, giving plots such as those shown in Fig. 5.8.

Thompson (1983) and Thompson and Howarth (1978) describe a method of estimating analytical precision using duplicate pairs. This is not a procedure routinely used by the G-BASE project but is a particularly useful way of estimating the analytical precision when no truly representative reference materials are available.

4.3. Hierarchical analysis of variance

In the previous section, the duplicate–replicate control data set was used to give graphical representation of sampling and analytical variability. A statistical procedure referred to as analysis of variance (ANOVA) analysis can be done on the same data set to give a more quantitative statement on variability. Sinclair (1983) describes this method that compares variations that arise from different identifiable sources,

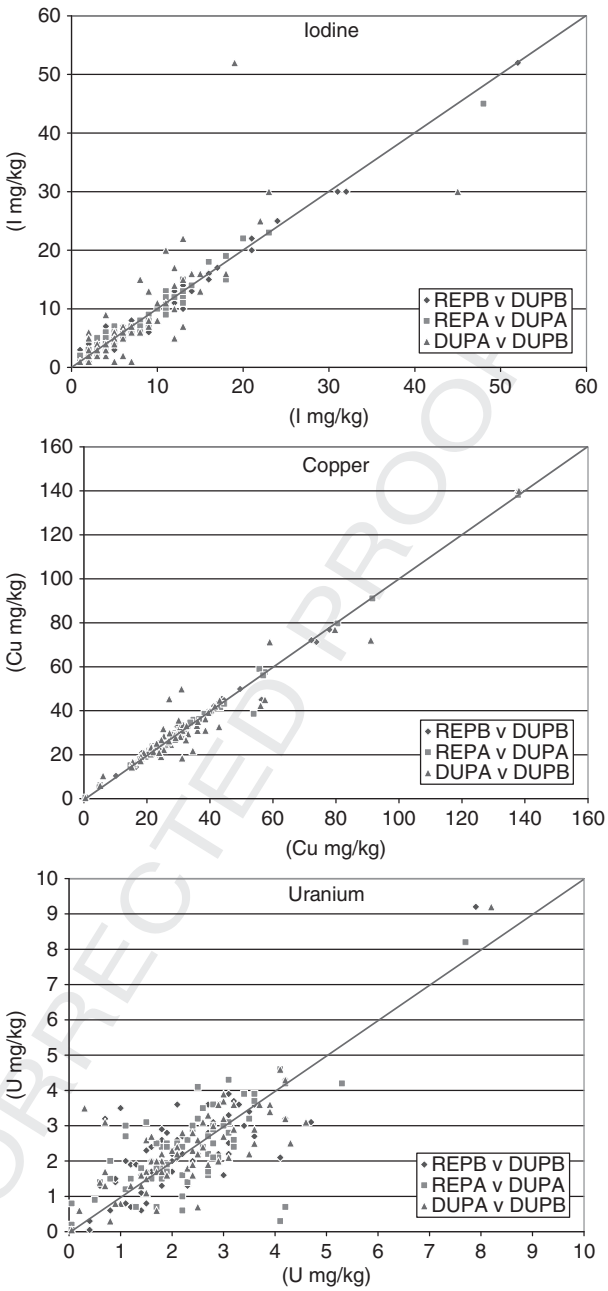

Figure 5.8 Examples of duplicate–replicate plots from soil analyses from the Geochemical Baseline Survey of the Environment (G-BASE) project generated by an MS Excel macro (for the relationship of control samples, see Fig. 5.4). Note that the cited detection limits for Cu, U and I are 1.3, 0.5 and 0.5 mg/kg, respectively.

namely analytical error, sampling error and regional variation. This procedure is therefore an important test of a project's strategy. It would be hoped that the analytical (within sample) variations and the sampling (within site) variations are small relative to regional variations. If not, then the analytical and sampling methodologies need to be reviewed. In many instances, particularly where the natural levels of an element are at or below the lower analytical detection limit, it is not possible to make improvements to the methodology. It is important that users of the data are made aware that for some elements the analytical variance accounts for a high percentage of the total variance in the results. The table that attributes variation generated by the ANOVA (Table 5.3) is an excellent way of disseminating such information.

The G-BASE project has used several statistical packages to perform this nested ANOVA analysis (e.g., Minitab and SAS). It currently uses an MS Excel procedure with a macro based on the equations described by Sinclair (1983) in which the ANOVA is performed on results converted to $\log_{10}$ (Johnson, 2002). Ramsey *et al.* (1992) suggest that the combined analytical and sampling variance should not exceed 20% of the total variance with the analytical variance ideally being <4%.

5. Levelling Data

The final part of the data conditioning procedure is the process of making the data fit with existing data sets. The accuracy of small data sets derived from a single analytical batch, and not used along with other geochemical data, can be put into context simply by publishing the certified international reference materials results. However, if a data set is composed of many analytical batches analysed over a period of months or even years, then the data will need to be levelled between batches to ensure the seamless creation of geochemical images. An example of a geochemical image where results have not been levelled is shown in Fig. 5.9. An annual field campaign boundary can be seen.

Another issue concerning the levelling of results can arise from improvements in the analytical lower detection limits. Such improvements will give much greater resolution in the data at low values which will be absent from older data with a poorer detection limit. In such instances, it is difficult to produce geochemical maps and images when the different data sets have differing degrees of resolution. Indeed, levelling may not be desirable as it would involve degrading the results that have the better resolution. The G-BASE approach to dealing with this is described later in this section. Finally, a problem very common to any project that requires building a national geochemical database is the requirement to combine geochemical data generated by different analytical methods.

A good discussion of the levelling of geochemical data sets using the mathematical process of normalisation is given in Darnley *et al.* (1995). This work describes how the term normalisation is used in a mathematical sense, that is, 'to adjust the representation of a quantity so that this representation lies within a prescribed range (Parker, 1974), or, any process of rescaling a quantity so that a given integral or other functional of the quantity takes on a pre-determined value (Morris, 1991), rather

Table 5.3 Example of nested analysis of variance (ANOVA) of stream sediment control samples from the Geochemical Baseline Survey of the Environment (G-BASE) sampling of east Midlands, UK

Element	Between site %	Between sample %	Within sample %
Na_2O	88.71	6.53	4.76
MgO	97.18	2.43	0.39
Al_2O_3	96.34	3.45	0.21
SiO_2	97.53	2.28	0.19
P_2O_5	91.83	7.98	0.19
K_2O	98.46	1.45	0.10
CaO	96.61	3.16	0.22
TiO_2	97.43	2.45	0.12
MnO	92.79	6.81	0.40
Fe_2O_3	94.25	5.25	0.50
Sc	92.02	2.83	5.15
V	95.55	4.00	0.45
Cr	96.62	2.54	0.84
Co	83.79	7.78	8.44
Ba	96.96	2.71	0.33
Ni	92.73	6.40	0.87
Cu	96.58	3.08	0.34
Zn	94.96	4.88	0.16
Ga	95.83	3.48	0.68
Ge	65.50	8.82	25.68
As	93.86	5.30	0.84
Se	88.84	3.85	7.31
Br	89.50	9.96	0.54
Rb	97.49	1.81	0.70
Sr	96.81	2.84	0.35
Y	93.35	5.96	0.69
Zr	94.97	4.74	0.28
Nb	97.89	1.55	0.56
Mo	76.58	–0.51	23.93
Hf	82.13	9.19	8.68
Ta	29.89	–13.36	83.47
W	58.12	–6.52	48.39
Tl	12.97	6.49	80.54
Pb	97.37	2.04	0.59
Bi	34.61	–12.75	78.14
Th	96.24	1.42	2.35
U	70.81	–0.92	30.11
Ag	70.49	0.29	29.22

(*continued*)

Table 5.3 (*Continued*)

Element	Between site %	Between sample %	Within sample %
Cd	87.71	5.36	6.94
Sn	88.49	5.62	5.89
Sb	65.07	6.89	28.05
Te	–2.65	0.00	102.65
I	84.27	11.79	3.94
Cs	72.25	–1.37	29.11
La	88.98	4.26	6.77
Ce	94.28	1.08	4.64

Note: The table was generated using an Excel procedure (Johnson, 2002). Elements where <80% of the variance is attributable to regional variations are highlighted, and results for these elements should be treated with caution.

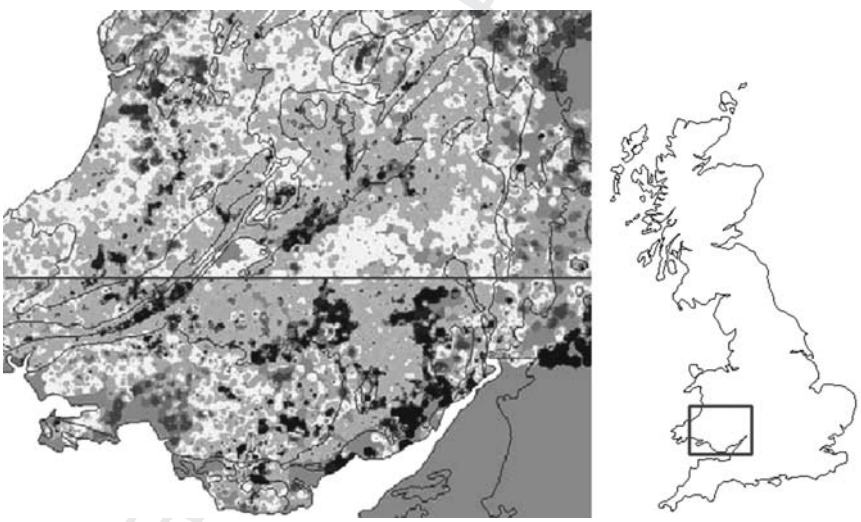

Figure 5.9 **Example of a geochemical image with unlevelled data (copper in stream sediments from Wales). The horizontal E-W line is shown just above the southern extent of the 1988 field campaign boundary. A subtle but noticeable linear feature can be seen coincident with the boundary between samples collected in different years. The results in the south had to be reduced by levelling in order to remove this analytical artefact.**

than in the statistical sense, where it connotes a transformation of a data set so that it has a mean of zero and a variance of one'. Normalisation of the secondary reference material results gives levelling factors that are applied to the data to give, ultimately, a single discrete national G-BASE data set.

The G-BASE project predominantly uses parametric levelling as described in the following section. When comparing geochemical results analysed by different

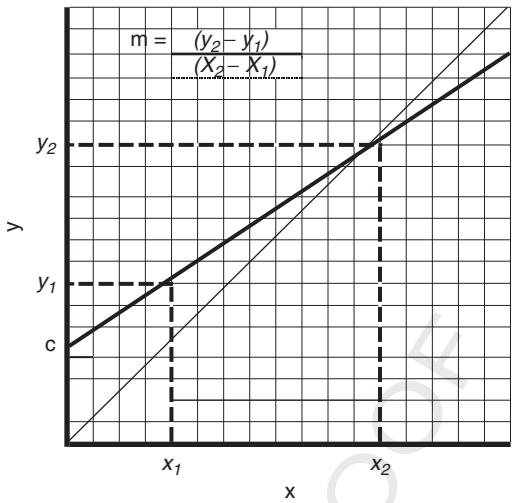

Figure 5.10 Cartesian equation $y = mx + c$ representing a straight line intercepting the y-axis at c with a gradient of m. The line shown passing through the origin has gradient $m = 1$.

geochemical methods, or by the same geochemical method but at a different time one can anticipate that the results will show a proportional relationship between the two results x and y. The simplest model is a directly proportional relationship between x and y represented by a straight line and the Cartesian equation $y = mx + c$ illustrated in Fig. 5.10. Secondary reference materials included in every analytical batch over a long period can be plotted as x–y plots where the 'accepted' value (i.e., a result determined by repeated analysis) is plotted as the y value against the determined value for a particular batch plotted as the x value. G-BASE submits samples in batches of 500 including sufficient SRMs that have a range of elemental values allowing a linear graph to be plotted (Fig. 5.11). From the linear regression modelling of the straight line, the derived Cartesian equation can then be used to determine levelling factors for each element.

5.1. Between batch and between field campaign data levelling

The between batch and field campaign levelling is best illustrated by the use of a real example. A potassium control plot for G-BASE secondary reference material S15B (determined by XRFS) over a period of 19 months is shown in Fig. 5.6. This shows that although all results fell within accepted limits of analytical variability (which is generally high for major oxides determined by XRFS on pellets); there are significant analytical shifts observable during this period. The most noted change came when the XRF machine was replaced, but before that there was also a period when results were consistently lower. The data therefore needs to be divided into three groups: (1) high values/old instrument, (2) low values/old instrument and (3) values from new instrument; and then levelled against the accepted result for K_2O in S15B. Although the results for S15B are shown here, other SRMs were analysed in the same batches and exhibited similar shifts. The data for all reference materials can be

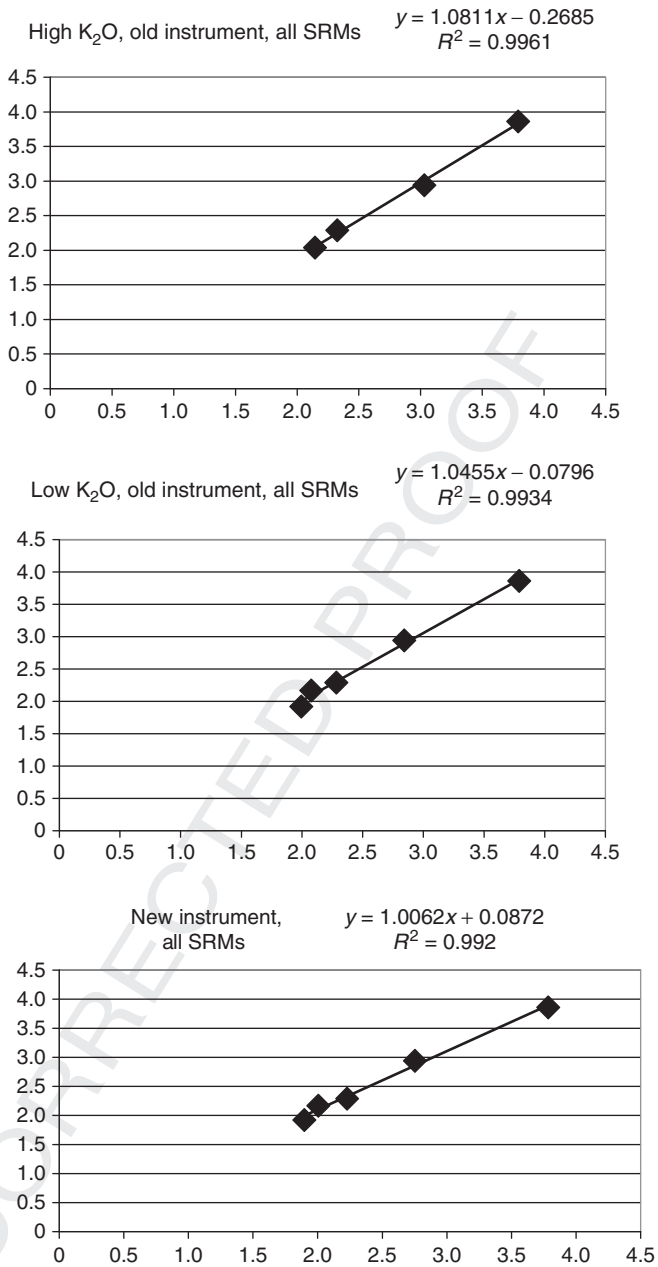

Figure 5.11 Liner regression plots for all secondary reference materials (SRMs) for potassium. The three data subsets relate to shifts indicated in Fig. 5.6. Axis units are % K_2O with the accepted value plotted as *y* and the batch value plotted as *x*.

plotted as shown in Fig. 5.11. The regression analysis (equations shown on Fig. 5.11) gives the levelling factors that need to be applied to bring the results to the level of the 'accepted' SRM results. For example, for the batches of high K_2O results on the old instrument (top graph in Fig. 5.11), all K_2O results should be multiplied by 1.0811 and 0.2685 subtracted. These factors can be applied to their respective data subsets and the results 'levelled' as shown in Fig. 5.12.

5.2. Levelling data with differing lower limits of detection

Variations (usually decrease) in detection limits occur with time, affecting both long-running projects, and the comparison of time-series data or adjacent project areas separated in time. It arises as analytical methods improve, and has its greatest impact on the trace elements, where the natural abundance is low in relation to the lowest measurable concentration. Such improvements can significantly increase the number of sample locations with measurable values in comparison to older data. The ability to make use of all the data acquired to its best potential is important, particularly for national mapping programmes. The data can be levelled as described above, if the standards used fall above the detection limit of the older method.

Illustrations of two sets of data with different lower levels of detection can be problematical as one set will show greater resolution of low results than the other. Presentation must be done in such a way as to not downgrade the best data (by applying the highest detection limit to all results), nor giving a false impression (by suggesting that samples below two very different detection limits represent the same data distribution). The example in Fig. 5.5 shows that a substantial amount of useful data occurs below the ICP-AES detection limit of 14 μg/l Al and that the ICP-MS data should not be truncated at that concentration. The graph also shows

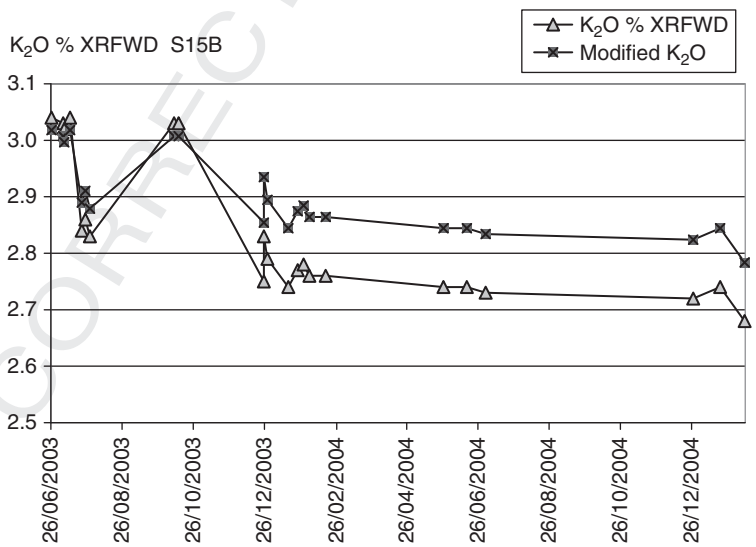

Figure 5.12 A graph showing the effect of levelling K_2O values for secondary reference material (SRM) S15B using the levelling factors determined in Fig. 5.11.

that to represent all the data at <14 μg/l Al in the same way as samples below the ICP-MS detection limit of <1 μg/l Al would be a false representation. This process is best illustrated by the techniques used to combine stream water data for England and Wales, exemplified by a significant lowering of the detection limit (in relation to the natural abundance of Al) caused by a change in analytical method. The earliest data, from Wales, was acquired by Inductively Coupled Plasma Atomic Emission Spectroscopy (ICP-AES). However, the advent of routine Inductively Coupled Plasma Mass Spectrometry (ICP-MS) analysis of stream water samples led to a marked increase in the number of samples with measurable Al, with only ~2% of data below the detection limit, compared to ~50% by the ICP-AES. Figure 5.13 illustrates the method used by the G-BASE project to overcome these very different data distributions, using stream water Al data. The ICP-AES data for Wales is plotted using the same percentile scale as is derived for the ICP-MS data in England, until the detection limit of 14 μg/l is reached. Beneath that concentration, data are shown in a dark grey colour; this is to demonstrate that the measurement of the sample has been made in that location, but that the information it provides is not as detailed as occurs in the more recent data.

5.3. Levelling data determined by different analytical method

Over a long period, it is inevitable that analytical methods will improve and change. The G-BASE project made significant changes in the 1990s when Direct Couple Optical Emission Spectroscopy (DCOES) was replaced as the principal method of analysis for sediments and soils by XRFS. This occurred during the regional geochemical mapping of Wales (BGS, 2000). Since this major change in analytical methodology geochemical images of the United Kingdom are plotted with data levelled with reference to the Welsh XRFS data. The levelling procedure removes features on maps that would solely be attributed to different analytical methods.

The levelling was achieved by a combination of techniques. The first of which involved a similar procedure to that described above to level between analytical batches, namely, regression analysis of SRM samples that had been determined by both DCOES and XRFS. Levelling factors for elements were further refined by plotting elemental results, DCOEC versus XRFS for some 3000 samples analysed by both methods. A final check of the levelling was then done by using a percentile regression method (see Darnley *et al.*, 1995) in which the percentile distributions of both methods are compared and adjusted until they match.

6. Discussion

Data conditioning is a collection of processes that makes data suitable for the purpose for which it is to be used. Some or all of the processes described may be applied to the data, depending on the way it is to be used. Its importance is illustrated by geochemical images such as that shown in Fig. 5.13 where geochemical data collected over a long period, and determined by different analytical methods, can be presented as a single geochemical image without additional analytical or temporal artefacts. At a local or site scale, it is important that geochemical results are subjected

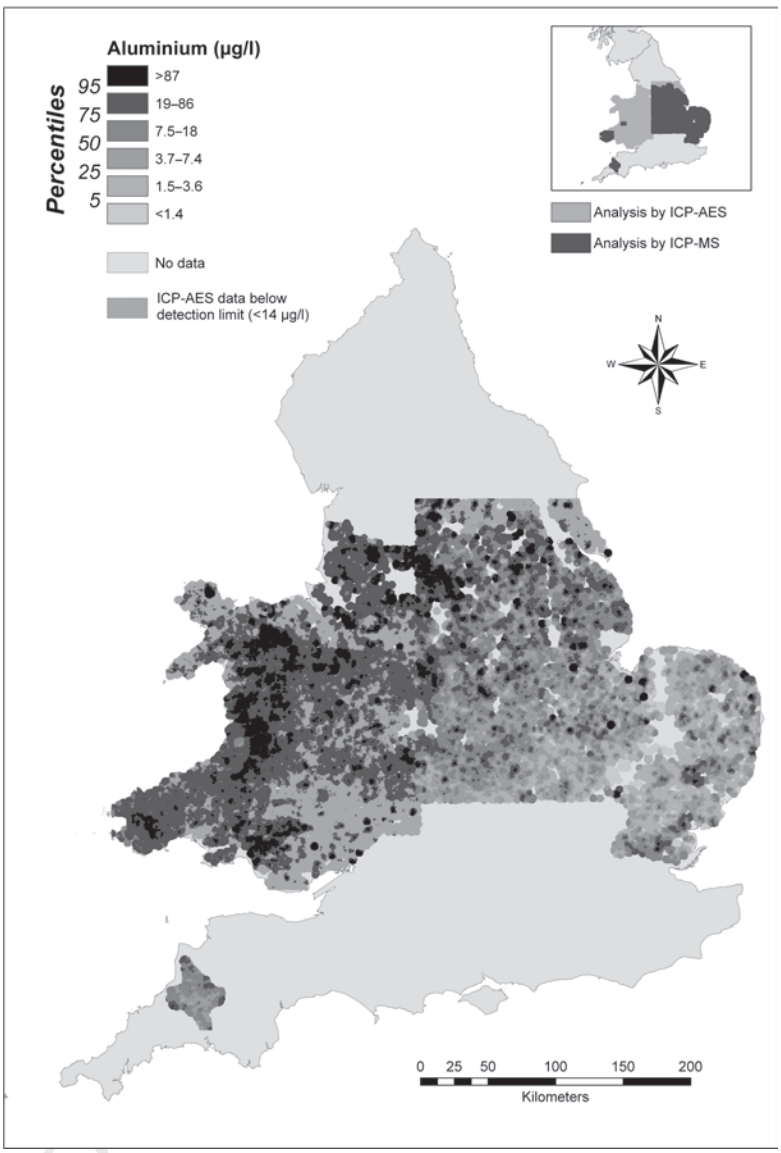

Figure 5.13 Aluminium in stream waters from England and Wales from the Geochemical Baseline Survey of the Environment (G-BASE) project illustrating how data sets with different lower detection limits can be combined.

to a similar level of conditioning, so they can readily be interpreted in the context of a regional setting.

In the preceding sections, the importance of documenting procedures and passing on quality information to the end user has been repeatedly stressed. Two areas not covered by the previous discussions are management responsibilities and

user-appreciation of quality control information. The former requires good coordination and communication from the planning phase to data delivery. Lack of understanding on how the customer uses the data may negate all the effort put into making the data fit for purpose.

Worldwide there is an understandable perception that collecting and preparing samples would generally be considered as a 'labouring' job, whereas chemical analysis and data interpretation are more intellectual tasks. This perception must be managed carefully. The reality is that the sampling and sample preparation are the phases at which data quality is at the greatest risk. The efforts and the value of the samplers and sample preparation staff must not be underestimated in the final outcome of the project. Good management requires that all those involved in the project are aware of all the procedures throughout the project. A sampling team which, for very good reason and showing good initiative, may decide to collect an additional sample from a site and label them with identical site numbers with an 'A' and 'B' suffix will probably not appreciate the dilemma this produces for a field database that only accepts unique numeric sample numbers. Everyone must be made aware that small, unimportant tweaks to procedures can have serious impacts later in the programme.

Finally, the effort to assess the quality of the data is wasted if the information is not passed on to the data user. Such users may not be a scientist, so effort must be made to present the data quality information in a manner that they can understand. The G-BASE project when distributing data to users makes use of font formatting, colour coding and highlighting in Excel spreadsheets to pass on information on quality, and Fig. 5.14 is a key to such an Excel document. This gives the user data in a completely numeric format with simple and straightforward advice about how it should be used. It is then up to the data user to choose whether to ignore the advice.

Colour/ format	Signifies	User action
56.7	No quality issues	None
25.4	This result is associated with a qualifier	Be aware that the result is qualified
Red	Data of dubious quality with significant issue(s)	Pay careful attention to what the quality issue is and if necessary don't use results
Yellow	Results <= 0	Be aware that the result could give problems in some statistical or plotting packages
Grey	<null> value	Be aware that no result is present though transferring to some software packages could erroneously reset this to 0
25	Generally a data issue relating to representation of results above or below limits of detection	Be aware that this result has some quality issue but is unlikely to restrict its use
10		
5		
0.6		

Figure 5.14 A key to the colour-coded quality information sent out to data users when the data is distributed in an Excel spreadsheet. Note that when produced in colour the colour-fill is used but is shown above by the text red, yellow and grey in the data cells. Similarly italic font in different colours (only shown in black/grey above) are used to indicate different data quality issues.

ACKNOWLEDGMENTS

This chapter is published with the permission of the director of the British Geological Survey (NERC). The authors also wish to acknowledge the efforts of several generations of geology students who have worked as 'voluntary workers' collecting samples for the G-BASE project. Their efforts working to strict procedures have ensured that the G-BASE project output is based on data of high quality and reliability. We appreciate the constructive reviews from C. A. Palmer and an anonymous referee.

REFERENCES

Albert, R. H., and Horwitz, W. (1995). Incomplete data sets: Coping with inadequate databases. *J. Assoc. Anal. Chem. (AOAC) Int.* **78**(6), 1513–1515.

AMC (2001a). What should be done with results below the detection limit? Mentioning the unmentionable. *In* ''AMC (Analytical Methods Committee)'' Technical Brief No 5. Royal Society of Chemistry, United Kingdom.

AMC (2001b). Measurement of near zero concentration: Recording and reporting results that fall close to or below the detection limit. *Analyst* **126,** 256–259.

BGS (2000). Regional geochemistry of Wales and part of west-central England: Stream sediment and soil. *In* ''British Geological Survey.'' Keyworth, Nottingham, UK. ISBN 0 85272 378 4.

British Standards (2005). General requirements for the competence of testing and calibration laboratories. BS EN ISO/IEC 17025: 2005. ISBN 0 580 46330 3.

Coats, J. S., and Harris, J. R. (1995). Database design in geochemistry: BGS experience. *In* ''Geological Data Management'' (J. R. A. Giles, ed.), No. 97, pp. 25–32. Geological Society of London Special Publication. London, UK.

Darnley, A. G., Bjorklund, A., Bolviken, B., Gustavsson, N., Koval, P. V., Plant, J. A., Steenfelt, A., Tauchid, M., and Xuejing, X. (1995). A global geochemical database for environmental and resource management UNESCO publishing, 19.

DETR (2000). ''Contaminated Land Implementation of Part IIa of the Environment Protection Act.'' HMSO London.

EC (1986). Council Directive 86/278/EEC of 12 June 1986 on the protection of the environment, and in particular of the soil, when sewage sludge is used in agriculture. (Sewage Sludge Directive). Published in the Official Journal (OJL 181/6), 4 July 1986.

EC (2000). Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for the Community action in the field of water policy. EU Water Framework Directive, 23 October 2000. Published in the Official Journal (OJL 327) 22 December 2000.

Environment Agency (2006). Performance standard for laboratories undertaking chemical testing of soil, March 2006, version 3. Environment Agency for England and Wales.

Harris, J. R., and Coats, J. S. (1992). Geochemistry database: Data analysis and proposed design. British Geological Survey Technical Report, WF/92/5.

Johnson, C. C. (2002). Within site and between site nested analysis of variance (ANOVA) for Geochemical Surveys using MS EXCEL. British Geological Survey, UK, Internal Report No. IR/02/043.

Johnson, C. C. (2005). 2005 G-BASE Field procedures manual. British Geological Survey, Keyworth, UK, Internal Report No. IR/05/097.

Johnson, C. C., Breward, N., Ander, E. L., and Ault, L. (2005). G-BASE: Baseline geochemical mapping of Great Britain and Northern Ireland. *Geochem. Explor. Environ. Anal.* **5**(4), 347–357.

Lister, T. R., and Johnson, C. C. (2005). G-BASE data conditioning procedures for stream sediment and soil chemical analyses. British Geological Survey, BGS Internal Report Number IR/05/150.

Lister, T. R., Flight, D. M. A., Brown, S. E., Johnson, C. C., and Mackenzie, A. C. (2005). The G-BASE field database. British Geological Survey, BGS Internal Report Number IR/05/001.

Miller, J. N., and Miller, J. C. (2005). ''Statistics and Chemometrics for Analytical Chemistry,'' 5th edn. Ellis Horwood, Chichester.

Morris, C. G. (1991). ''Academic Press Dictionary of Science and Technology,'' pp. 2432. Academic Press, San Diego.

Parker, S. B. (1974). ''McGraw-Hill Dictionary of Scientific and Technical Terms,'' pp. 2088. McGraw-Hill, New York.

Plant, J. A. (1973). A random numbering system for geological samples. *Trans. Inst. Min. Metall.* **82,** 64–65.

Plant, J. A., and Moore, P. J. (1979). Geochemical mapping and interpretation in Britain. *Philos. Trans. R. Soc.* **B288,** 95–112.

Plant, J. A., Jeffrey, K., Gill, E., and Fage, C. (1975). The systematic determination of accuracy and precision in geochemical exploration data. *J. Geochem. Explor.* **4,** 467–486.

Potts, P. J. (1997). A glossary of terms and definitions used in analytical chemistry. *Geostand. Newsl.* **21**(1), 157–161.

Ramsey, M. H., Thompson, M., and Hale, M. (1992). Objective evaluation of precision requirements for geochemical analysis using robust analysis of variance. *J. Geochem. Explor.* **44,** 23–36.

Sinclair, A. J. (1976). Applications of probability graphs in mineral exploration. *Assoc. Explor. Geochem.* Special Volume No. 4, pp 95.

Sinclair, A. J. (1983). Univariate analysis. *In* ''Statistics and Data Analysis in Geochemical Prospecting'' (R. J. Howarth, ed.). Handbook of Exploration Geochemistry, Vol. 2, pp. 59–81. Elsevier, Amsterdam.

Thompson, M. (1983). Control procedures in geochemical analysis. *In* ''Statistics and Data Analysis in Geochemical Prospecting'' (R. J. Howarth, ed.), Handbook of Exploration Geochemistry, Vol. 2, pp. 39–58. Elsevier Science, Amsterdam.

Thompson, M., and Howarth, R. J. (1973). The rapid estimation and control of precision by duplicate determinations. *Analyst* **98,** 153–160.

Thompson, M., and Howarth, R. J. (1976). Duplicate analysis in geochemical practice. *Analyst* **101,** 690–709.

Thompson, M., and Howarth, R. J. (1978). A new approach to the estimation of analytical precision. *J. Geochem. Explor.* **9,** 23–30.