GEOLOGICAL SURVEY OF CANADA

OPEN FILE 3788

Building a lithogeochemical dataset for GIS analysis: Methodology, problems, and solutions

L. Wilkinson[1], J. Harris[2], E. Grunsky[3]

[1]Geological Survey of Canada,
615 Booth Street,
Ottawa, ON
K1A 0E9,
Phone: 613-996-2121,
Fax: 613-995-9273,
Email: lori@gis.nrcan.gc.ca

[2]Geological Survey of Canada,
615 Booth Street,
Ottawa, ON
K1A 0E9

[3]Alberta Geological Survey,
9945-108 St.,
Edmonton, AB
T5K 2G6

1999

Building a lithogeochemical dataset for GIS analysis: Methodology, problems, and solutions

Lori Wilkinson, Jeff Harris, and Eric Grunsky

Abstract

This paper summarizes the methodology and considerations involved in the compilation and preparation of a large lithogeochemical dataset derived from disparate sources into a Geographic Information System (GIS). Approximately 4500 lithogeochemical samples, from 6 sources, have been compiled into a single lithogeochemical dataset for the Swayze Greenstone Belt area of Northern Ontario. Problems dealt with during the compilation process include: missing data, uncertain locations, non-unique sample identifiers, 0's in the data, censored data and analytical uncertainty. The combined lithogeochemical dataset was scrutinized statistically and spatially, so that the data can be reliably used for regional mapping and exploration activities.

Keywords

GIS, lithogeochemistry, Swayze, Compilation, Levelling

1.0 Introduction

A three-year project to compile and analyze a variety of digital data for the Swayze greenstone belt in northern Ontario (Figure 1), using Geographic Information System (GIS) technology, was initiated in 1993 by the Geological Survey of Canada (GSC) and the Ontario Geological Survey (OGS). The project was funded by NODA (Northern Ontario Development Agreement), part of the Canada-Ontario Economic and Regional Development Agreement (ERDA). Project goals included compilation of a digital geoscience database, assessment of the ability of GIS to aid in geologic mapping and mineral exploration activities, and transfer of GIS technology to the geological and mining community (see Harris et al. 1994, Harris et al. 1995a,b). The study is unique in that industry partners (Falconbridge Ltd. and Hemlo Gold, formerly Noranda Exploration Company Ltd.) agreed to supply much needed proprietary digital data and exploration expertise to the project in return for a one-year period of exclusivity over the results generated by the project.

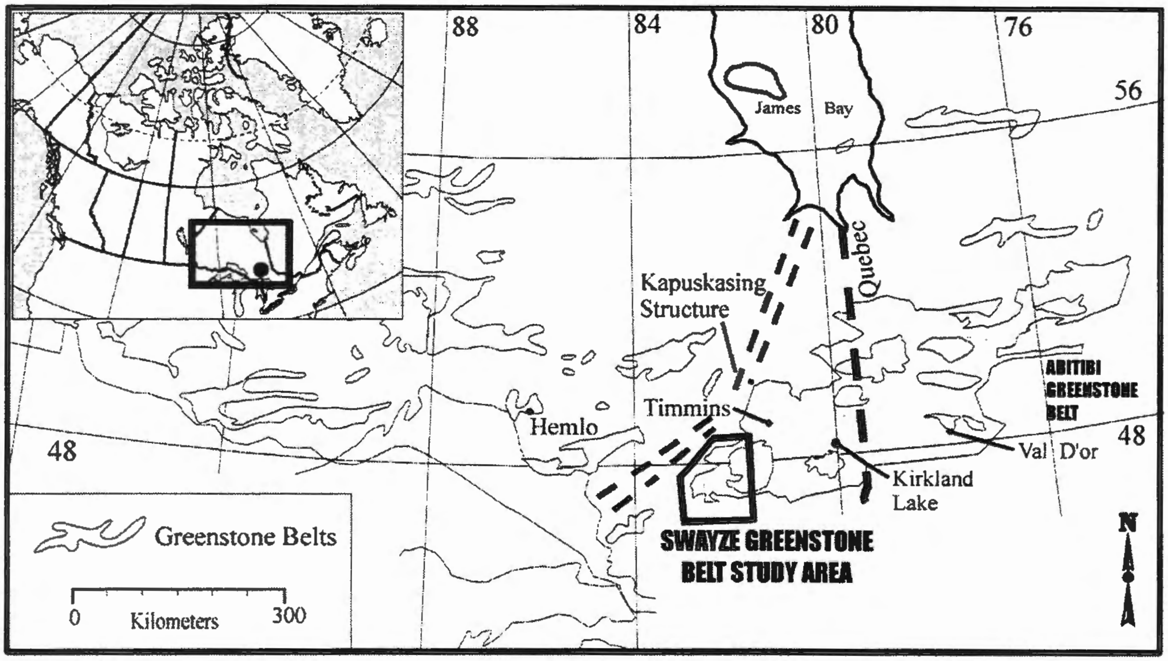


Figure 1: General location map of the Swayze greenstone belt in Ontario.

The Swayze greenstone belt (SGB) is the westernmost extension of the mineral-rich Abitibi greenstone belt (AGB), and has recently been re-mapped by both the OGS and the GSC (Ayer and

Theriault 1993; Heather and van Breeman 1994; Heather et al. 1995). Similar to the AGB, the SGB contains a number of folded 2730-2680 Ma mafic-felsic metavolcanic packages that are unconformably overlain by Timiskaming-type metasedimentary rocks and cut by high-strain zones thought to be extensions of the major "breaks" (Destor-Porcupine and Cadillac-Larder Lake Faults) found in the AGB (Ayer and Theriault 1993; Heather and van Breeman 1994; Heather et al. 1995). Figure 2 is a generalized geology map of the SGB. Unlike, the AGB however, few economic deposits have been found within the SGB, and the exploration level is still relatively low.

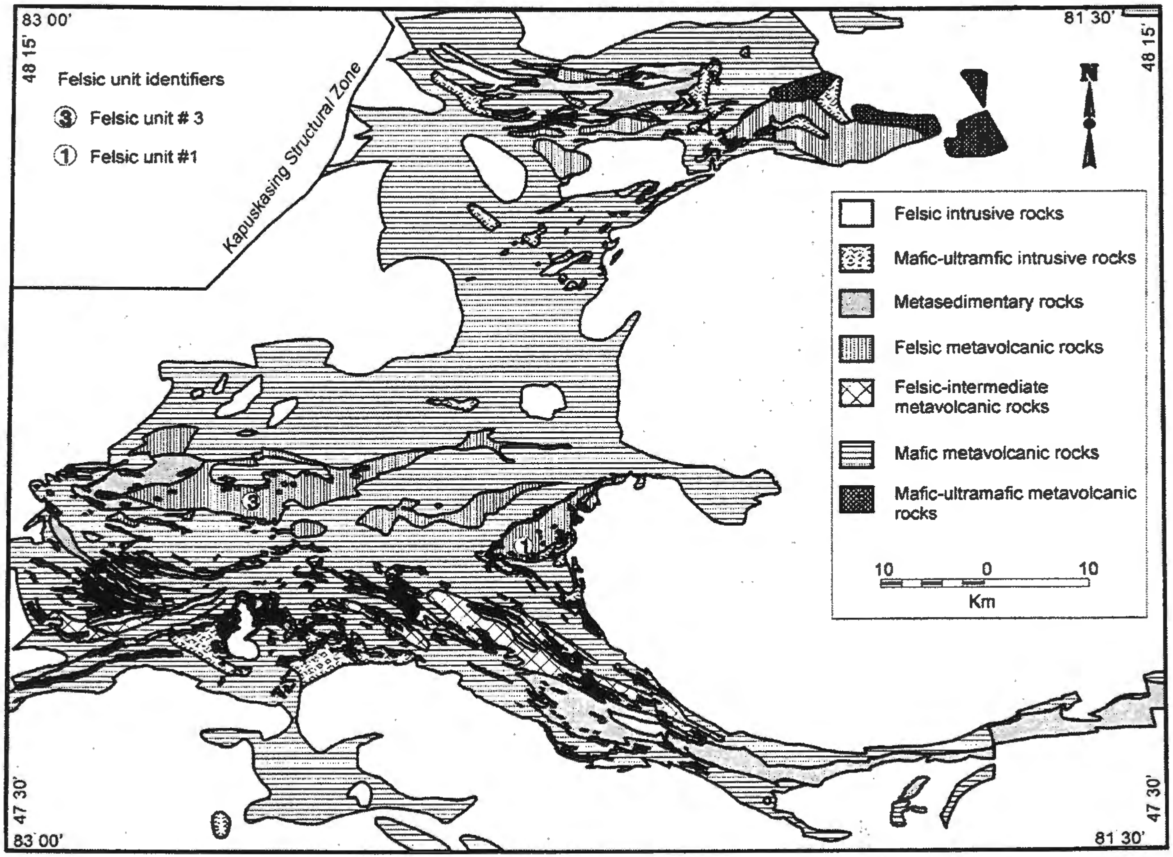


Figure 2: Generalized geology of the Swayze greenstone belt (from 1:50 000 scale compilation provided by Falconbridge Ltd.).

The use of Geographical Information Systems (GIS) is relatively new to the mining industry in Canada. Much like Computer Aided Drafting (CAD) systems, GIS are capable of storing, displaying and plotting georeferenced points (lithogeochemical samples, drill hole locations, etc.), lines (roads, faults, etc.), and polygon data (lithologic units, drainage basins). However, GIS offers a number of advantages over CAD systems in that the "graphic primitives" (points, lines and polygons) can be linked to a database which contains attributes or descriptive information that are associated with each graphic element. Secondly, a GIS provides a wide range of spatial analysis tools with which to display, query, manipulate and analyze the data. A GIS is a useful tool for the mining and exploration industry in that it offers the capability for spatial analysis in a problem-solving environment. GIS provide powerful tools for developing user-specific applications, such as procedures for generating exploration favourability maps (Bonham-Carter et al. 1988; Harris 1989; Bonham-Carter 1994; Harris et al. 1994; Rencz et al. 1994; Wright and Bonham-Carter 1996).

One of the most difficult aspects of compiling a large GIS database involves the integration of data in many different formats. Analog data must be converted to digital form, either by scanning or manual digitization. This process is often time consuming and laborious. Equally challenging in the compilation of a large GIS database is the integration of data from many disparate sources. Creating continuous seamless data is desirable for point data in which a specific location has a number of attributes that record information such as structural or lithologic observations at an outcrop. In exploration, combining various datasets (i.e., separate lithogeochemical whole rock analyses) is desirable to assist in defining regional geochemical trends, which in turn, can yield valuable information on geologic and ore-forming processes. Furthermore, an extensive and seamless lithogeochemical dataset is useful for characterizing regional litho-tectonic patterns and identifying anomalous geochemical populations that may be related to mineralization.

Combining geochemical data which has been collected at different times, by different organizations using different sampling strategies and analyzed by different laboratories using different analytical techniques, is full of potential pitfalls. It is desirable to maintain as much of the original data coding as possible, to retain the lineage information, and to track all the potential sources of error being built into the dataset through the compilation process. Once the dataset has been compiled it is also necessary to verify and evaluate its integrity. This is particularly important for analytical data, acquired from different sources, and hence, potentially different analytical techniques.

This paper emphasizes a regional approach to compiling, evaluating and interpreting lithogeochemical data. It is important to note that the scale of compilation and project objective affects the approach used in subsequent analysis. The objectives of the compilation must be clearly defined. Since compilations and interpretations are scale-dependent, methods of analysis and interpretation must be chosen to suit the scale of study. Mining camp-scale studies require different approaches to both compilation and analysis than do greenstone belt wide studies.

This paper provides a methodology for the creation of a combined lithogeochemical dataset, derived from a variety of sources, using a GIS and associated relational database software. The problems encountered during the data compilation process are grouped into 2 types of errors. Each is discussed and possible solutions are suggested. The integrity of the final combined database is tested using a number of simple statistical comparisons and problems inherent in analyzing the data are discussed.

2.0 Lithogeochemical data

Approximately 4500 whole rock samples were acquired from three principle sources; GSC, OGS and Falconbridge Ltd. Individual datasets and their specifications are listed in Table 1. The proprietary data received from Falconbridge were contributed as part of a legal agreement between the GSC and Falconbridge Ltd. and Hemlo Gold Ltd. for this GIS project (Harris et al. 1994).

DATASET	*# OF SAMPLES*	*YEARS COLLECTED*	*ANALYTICAL METHODS*	*SOURCE*
J. Ayer (JA)	135	1991-1993	91-XRF/ICP-MS/ICP-ES/AA 92-XRF/ICP-ES/ICP-MS	OGS
K. Heather (KH)	348	1992-1995	92-95-XRF, ICPMS, AA, DIONEX	GSC
PETROCH (PT)	646	1976-1993		OGS
S. Fumerton (SF)	1304	unknown (variable)	XRF/ICP-MS/AA/FA	Assessment files
Falconbridge (FA)	1058	1978-1979	XRF	Falconbridge Ltd.
Texas Gulf (TG)	943	unknown (variable)	XRF/ICP-ES	Falconbridge Ltd.

Table 1: Compiled dataset components. XRF - x-ray fluorescence, ICP-MS - inductively coupled plasma mass spectrometry, ICP-ES - inductively coupled plasma emission spectrometry, AA - atomic absorption, DIONEX - Dionex Ion Chromatgraphy Analyzer.

Quality control and the assessment of analytical variability is of crucial importance when analyzing any type of geochemical data (Rose et al. 1979; Fletcher 1981; Thompson 1983). A critical assumption made in this study is that adequate, quality control measures were undertaken by the proprietor of each dataset at the time of original data collection and analysis. This presumably involved inserting split duplicates, field duplicates, and control samples to evaluate analytical variability within each individual dataset. With no control or involvement in these activities, we received each dataset from the organizations listed in Table 1 long after the data had been collected. Therefore, our efforts in this study focus on the compilation and analysis of the data after quality control measures had already been undertaken on each dataset.

Characteristics of each input data source are summarized in Table 2, and are briefly described here. Most major oxides are present in all datasets, with the exception of TiO_2, P_2O_5, K_2O and MnO in the TG dataset. However, volatiles, loss on ignition (LOI) and even "total" are frequently missing in many of the datasets. In addition, not all 4500 samples contained even partial major oxides. Fe is variably measured as FeO_T, Fe_2O_{3T}, or as both FeO and Fe_2O_3 (see Table 3), even within the same dataset (e.g. PT, SF, KH datasets).

Dataset	*Major oxides*	*Minor oxides*	*Duplicate samples*	*# of trace elements*	*Duplicate Elements*	*L. O. I.*	*Total*	*Lithology*
JA	yes	yes	no	42	Nb, Rb, Sr, Y	yes	yes	yes
FA	yes	yes	yes	4	no	yes	yes	no
KH	yes	yes	no	41	Ce, La, Nb, Nd, Rb, Th, Y, Yb, Zr	partial	yes	yes
TG	partial	partial	yes	32	no	partial	no	yes
PT	yes	partial	no	13	no	most	yes	yes
SF	yes	partial	yes	61	no	partial	partial	yes

Table 2: Summary of data source characteristics.

Dataset	*FeO_T*	*Fe_2O_{3T}*	*$Fe_2O_3 + FeO$*
JA	no	yes	no
FA	yes	no	no
KH	no	yes	yes
TG	no	yes	no
PT	yes	yes	yes
SF	no	yes	yes

Table 3: Summary of Fe data source characteristics.

The range of trace elements analyzed is extremely varied and too numerous to list here. Even within a single dataset, an element may be measured in one sample and not the next. This problem is compounded in the SF and PT datasets, which were obtained from assessment files and various OGS studies over a 20-year period, respectively. In both these datasets, the elements analyzed and the methods used to analyze them have changed over the years, but a systematic study of the possible effects of these factors is beyond the scope of this paper.

An additional problem affecting trace element data, is the occasional instance of multiple trace element analyses by different analytical techniques. Although oxide data are generally analyzed using XRF, it is common for different trace elements to be analyzed by several techniques within the same data source, such as ICP-MS for rare earth elements, and ICP-ES for other elements (see Table 1). Some element concentrations are occasionally determined using several techniques, and are preserved in the source data as duplicated columns e.g., Zr-1 and Zr-2. Since different analytical techniques have different detection limits, there is the possibility that an element may be below detection using one analytical method but above the detection limit on another method. Also, the general problem of censored data (below detection limit) must also be addressed within the combined lithogeochemical dataset. Levelling procedures (Darnley et al. 1995) can be applied to "align" data, however these procedures are time consuming and require careful validation.

The meaning of 0 values in each lithogeochemical database is also problematic. It is often unclear whether a value of 0 has been entered for oxides and/or elements to represent a “no data” value (an unmeasured quantity), or a measured oxide and/or element at which no quantity was observed. The way zeros are treated can have a significant impact on the interpretation of geochemical data that is compositional in nature (Aitchison 1986).

Other problems and inconsistencies observed within the individual lithogeochemical datasets include, “holes” or blanks in the tables where no value was entered, missing, uncertain or obviously wrong geographic locations, non-unique sample identifiers, e.g., duplicated sample numbers and elements analyzed by more than one technique.

Figure 3 is a flow chart that summarizes the methodology used for compiling the unified lithogeochemical dataset. We have divided the problems encountered in this process into two basic types; Type 1 problems of a practical nature, mostly identification problems, involved in building the actual dataset, and Type 2 problems of intrepretation in using the combined database. Figure 3 outlines the order in which data should be evaluated prior to the application of any types of data analysis or interpretation. This paper only deals with the identification of errors and the assembly of data into a dataset suitable for analysis. The identification of altered samples, while a Type 2 problem, is a complex and subjective process that is beyond the scope of this paper.

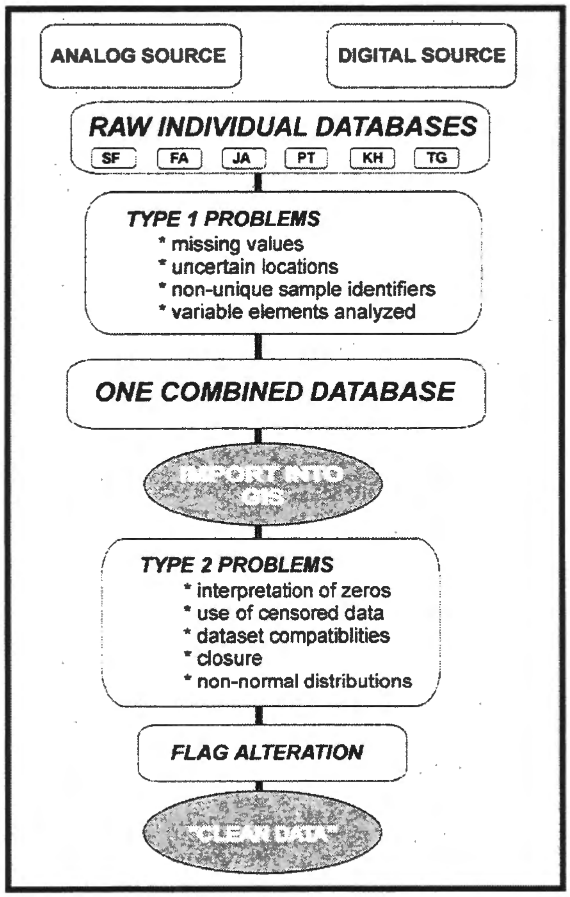


Figure 3: Flow chart summarizing preparation of the geochemical dataset for spatial data analysis. Note that flagging of altered samples was not demonstrated in this paper, but is an important consideration in mineral exploration.

Problems in identification (Type 1) include uncertain locations (geographic coordinates poorly recorded - lack of precision or inaccurate), non-unique sample identifiers (possible source of confusion) and the use of different methods of analysis for a given element (e.g. ICP-MS, XRF for Zr). Problems in the intrepretation of the data (Type 2) include missing values (incomplete analysis or not reported), dataset compatibility (levelling problems), use of censored data (actual value less than detection limit), and the nature of the geochemical distribution.

Type 1 and 2 problems can also be expressed another way. Type 1 problems can require that the data be deleted from the study, or that certain restrictive assumptions be made in order to proceed. On the other hand, Type 2 problems can be accommodated through additional analysis and adjustment. The applications of data analysis, and the use of Geographical Information Systems, can help resolve some of the Type 2 problems.

The importance of metadata, e.g., details of chemical analysis protocols and reporting procedures, is now broadly recognized and its use is becoming increasingly common, which will help to minimize the types of problems faced in this study.

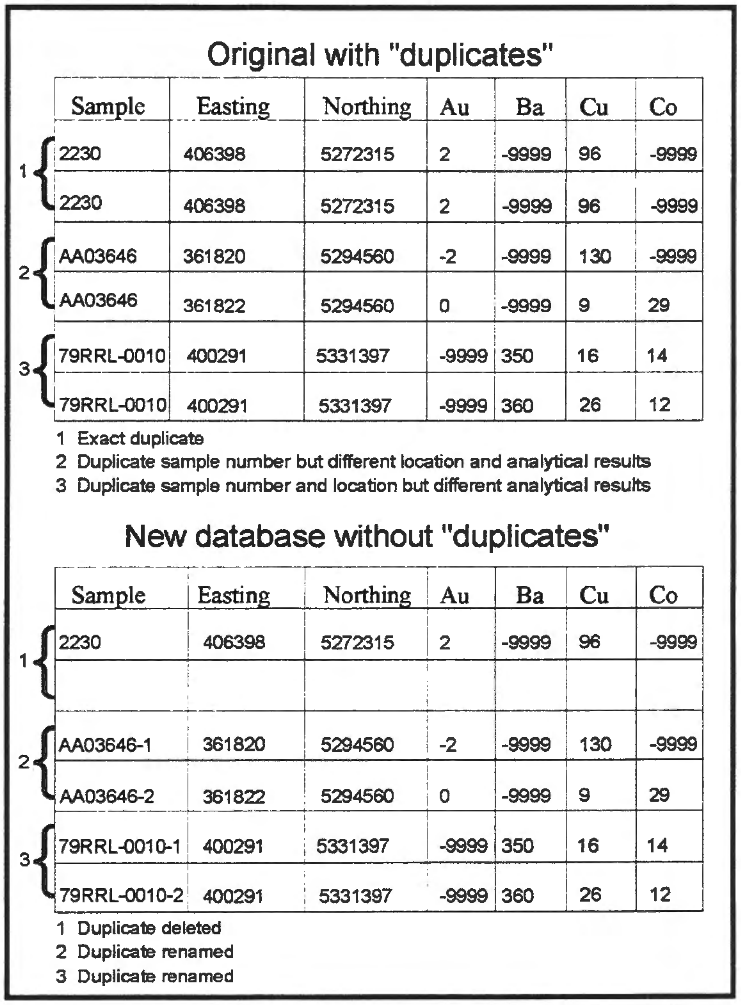
3.0 Data Compilation and Type 1 Problems

Of the 6 individual lithogeochemical datasets, all but one (TG) was in a digital form in spreadsheet or database format (e.g., FoxPro, Access, dBase, Excel). Thus, the first step was to manually enter the TG dataset into a database. Next, a flat (non-relational) table of each dataset, containing sample number, easting, northing, lithology (if available), and elements/oxides available, including duplicated elements by more than one methodology, was created. The 6 tables were then appended to form a single, flat table dataset containing all 6 sources. Numeric missing values (null values) were replaced with a value of -9999, while character missing values (null values) were replaced with "NO DATA". Thus all "holes" in each data source were removed, important for the import of the data into the GIS.

Since lithogeochemical data without a location cannot be interpreted in a spatial context, samples with uncertain locations were eliminated. Samples simply lacking Easting or Northing information or containing obviously wrong coordinates (i.e., 5 digits for Easting, or 6 digits for Northing) were also eliminated from the other datasets, since no means for establishing their correct location was available.

All databases and GIS rely on the concept of a unique identifier (or "key") for each piece of information, whether point, line or polygon. This unique identifier allows linkage of the spatial location of the point, line or polygon with its descriptive information (i.e., attributes) held in the GIS database. In the case of the 6 sources used in this compilation, duplicate sample numbers often occur (see Table 2). In some instances the same record was exactly duplicated within the source, permitting deletion of the repeated record. In other instances, the value of the attributes (oxides or elements) were different, suggesting that either that two samples were run with the same identifier, or that the same sample was run twice, perhaps as a split or field duplicate (Figure 4). This was checked by examination of the geographic coordinates of the two points. If both points had the same location, an assumption can be made that the sample was run twice, perhaps as a split duplicate and one sample can be deleted. In the case of samples with different locations, it is likely that one sample is incorrectly labeled. However, for both cases, without field notes or any other means of determining the source of the slightly different chemistry (and without any metadata descriptions), it is impossible to know which sample to delete. These samples are left in the database but are suffixed with a -1 or -2 to signify the first or second record with the same sample number (see Figure 4 for examples). This preserves the maximum amount of data in the combined dataset, but serves also to flag potential problem samples. The goal is thus to provide the database user with the necessary information to make informed decisions about its use; not to take these decisions out of their hands. Thus, the user may decide prior to spatial and statistical analysis, which of the duplicated samples to use or delete, or may choose to average duplicates.

To facilitate the amalgamation process, and to provide both an archive and a back up of the source data, all 6 original sources were imported into a single PC-based relational database to create a digital archive. In our case, the database serves as an archive of our original data and as a means by which to combine the tables for import into the GIS. All 6 tables were appended to form a single, large table of samples, maintaining all elements including duplicates due to alternative analytical methodologies.



Original with "duplicates"

	Sample	Easting	Northing	Au	Ba	Cu	Co
1	2230	406398	5272315	2	-9999	96	-9999
	2230	406398	5272315	2	-9999	96	-9999
2	AA03646	361820	5294560	-2	-9999	130	-9999
	AA03646	361822	5294560	0	-9999	9	29
3	79RRL-0010	400291	5331397	-9999	350	16	14
	79RRL-0010	400291	5331397	-9999	360	26	12

1 Exact duplicate
2 Duplicate sample number but different location and analytical results
3 Duplicate sample number and location but different analytical results

New database without "duplicates"

	Sample	Easting	Northing	Au	Ba	Cu	Co
1	2230	406398	5272315	2	-9999	96	-9999
2	AA03646-1	361820	5294560	-2	-9999	130	-9999
	AA03646-2	361822	5294560	0	-9999	9	29
3	79RRL-0010-1	400291	5331397	-9999	350	16	14
	79RRL-0010-2	400291	5331397	-9999	360	26	12

1 Duplicate deleted
2 Duplicate renamed
3 Duplicate renamed

Figure 4: Example of problem of duplicate samples within a data source (top table), and the solution used (bottom table). Note a value of –9999 indicates a NULL value for a numeric value field.

The import of the data to the GIS (Arc/Info version 6.1.2) was accomplished using a comma-delimited, ASCII file format. A macro (AML) was written within the GIS environment to convert the ASCII file to an Arc/Info "coverage". A new field was added to each dataset to identify the original source of a particular sample in the final combined dataset. The new attribute, called "dataset" in this case, was then filled with the appropriate initials, (e.g., TG for Texas Gulf, SF for S. Fumerton data etc.) thus preserving the original source of each sample in the compiled dataset.

4.0 Type 2 problems: Using the combined dataset

A variety of difficulties are inherent in using geochemical data compiled from a number of sources. Problems which include 0's and incomplete analyses within the dataset, censored data, closure and dataset compatibility problems are addressed here. It should be noted that a number of assumptions must be made when building and using the combined datasets. These assumptions are summarized in Appendix A.

A number of the datasets, especially TG, contained incomplete oxide analyses as mentioned previously. Although individual oxide elements can still be used for comparative evaluations, incomplete chemical analyses will preclude their use in chemical classification (i.e., ternary classification) and statistical analysis (i.e., classification, cluster and principal component analysis) procedures. The evaluation of individual oxide elements can only be used if they are used "as reported". Thus, they cannot be recalculated to 100%, and then used for classification or the creation of a lithogeochemical index such as an alteration index. Recalculating the data compounds the closure problem, and can often provide results that may be misleading or difficult to interpret. In some cases, ternary diagrams can be constructed and indices can be calculated providing that none of the essential oxide elements are missing. However, the interpretation of these analyses can be misleading if the values of the missing oxide elements do not fall within "acceptable" limits. Since these values are unknown, the use of incomplete analyses are suspect.

4.1 Zero values

It is uncertain in most cases, whether a value of 0 actually represents zero concentration or if it was entered to reflect a null value or unanalyzed element. Grunsky et al. (1992) in a lithogeochemical study using PETROCH (PT) data, considered 0's as missing values for major element oxides (excepting Na_2O, K_2O, TiO_2, P_2O_5, and MnO), and therefore eliminated these data from further consideration. Table 4 summarizes the minimum values that were observed for the major element oxides. It can be seen from the values below that there are as many as three lower limits of detection in the data. This is due to varying detection limits in each source dataset. In addition, within each original source dataset, multiple detection limits were possible if the data were collected over a period of time and analyzed using a variety of techniques.

Element	SiO_2	Al_2O_3	Fe_2O_3	FeO	MgO	CaO	Na_2O
Weight %	NA	0.01	.01/.20	0.2	0.01	0.01	.01/.02/.03
Element	K_2O	TiO_2	MnO	P_2O_5	CO_2	H_2O^-	H_2O^+
Weight %	.01/.02/.05	0.01	0.01	0.01	.01/.10	0.01	NA

Table 4: Lower limits of detection observed for the combined datasets. Note: NA means that no value close to zero was observed for the data. In some cases (Na_2O, K_2O, Fe_2O_3, CO_2), mutliple lower limits of detection were observed.

Alternatively, any entry of 0 can be considered to be an unlikely geochemical value, and thus be treated as a censored value (Miesch 1976). One of the difficulties in treating zero values as censored values is due to the fact that there may be more than one detection limit in a dataset that has been derived by different analytical procedures. In many cases, the true lower limit of detection is not available. In practice, the replacement value suggested by Aitchison (1986, p. 268), which is based on the minimum precision (resolution) of the data (0.01%), is a default lower limit of detection and thus zero values can be treated as censored data.

If the SGB lithogeochemical database is intended for regional mineral exploration, then the detection of anomalous values (e.g., > 90th percentile) is important. Therefore, it is desirable to preserve as many samples as possible to obtain a reliable estimate between regional background and truly anomalous data. Thus, 0 values were included in the study to maximize the number of sample points available. This has the advantage of maximizing the number of samples used, and preserving the lower abundance oxide information (e.g., P_2O_5) but can result in a cluster of data around zero. Such an effect can distort any parametric statistical procedures, and creates the possibility of under-estimating thresholds for anomalous populations. Such problems can be minimized by visually examining the probability plots of the data for a threshold that is not necessarily percentile based.

For procedures requiring normally distributed populations, only results greater than 0 were selected to allow for log-transformation. This is particularly important in statistical procedures that require homogeneity of variance/covariance. Although this invokes the assumption that 0 is an unlikely geochemical value, it is a mathematical necessity. To minimize the elimination of samples, it is recommended that the modified procedure of Grunsky et al. (1992) be used, whereby zero values for oxides Na_2O, K_2O, TiO_2, P_2O_5, and MnO are treated as censored data and replaced. Alternatively, if the lower limit of detection is known for a specific group of samples, then the method of Sanford et al. (1993) can be applied as discussed below.

4.2 Data Compatibility

The most important problem remaining in the dataset is one of compatibility between sources. The combined dataset is compiled from 6 sources, 4 of which are themselves compilations of other data sources (i.e., SF is a compilation of assessment files). This can result in significant variations in chemistry between samples of different sources or as multiple populations within a single oxide. To assess the degree to which these variations may affect final geochemical results, we have examined oxide population distributions both statistically and spatially.

4.2.1 Interchanged Na_2O and SiO_2

Probability plots and frequency histograms for Na_2O revealed a distinct sub-population of Na_2O > 40 weight percentage, and > 60 cation volume percentage, representing 198 samples (Figure 5a). Figure 5b also shows a probability plot and frequency histogram for SiO_2, in which an anomalous population comprising values < 20 wt% SiO_2, can be clearly seen. Figure 5c shows a scatterplot of Na_2O vs. SiO_2 where the clusters of data are obvious. The lack of any spatial correlation between the anomalous Na_2O samples, and the fact that a rock with > 40 % Na_2O and < 20% SiO_2 would be highly unusual, indicates a data transcription error. All 198 apparently switched Na_2O and SiO_2 values came from the same dataset, indicating that the error is not random within the combined lithogeochemical database, but rather a

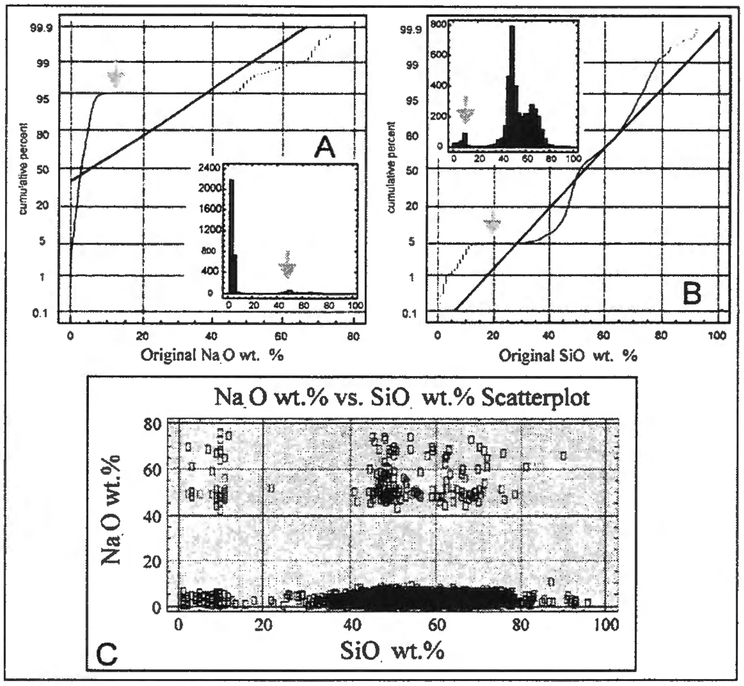


Figure 5: Comparison of Na_2O and SiO_2 populations in original combined dataset (A, and B, respectively. Gray arrows indicate breakpoint in anomalous population distribution. Scatterplot of SiO_2 and Na_2O (C) demonstrates distinct clusters between the anomalous populations of these two oxides.

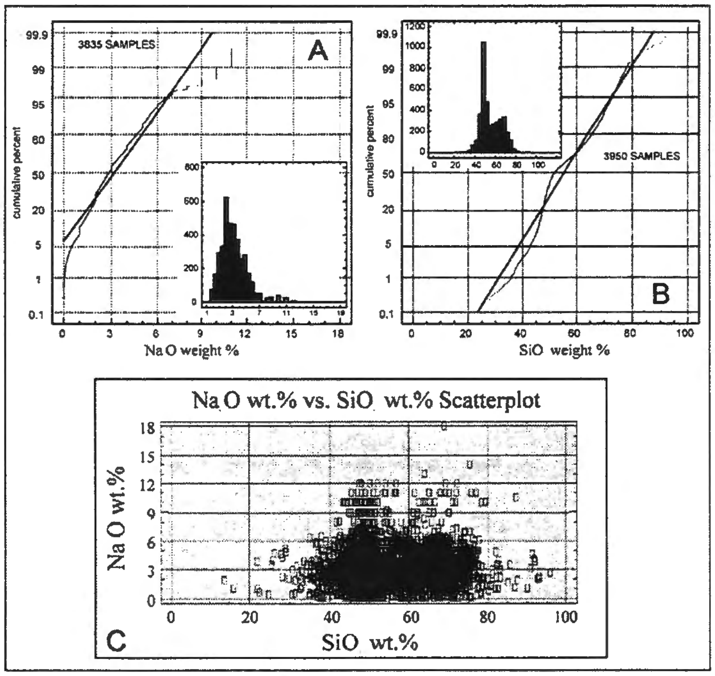
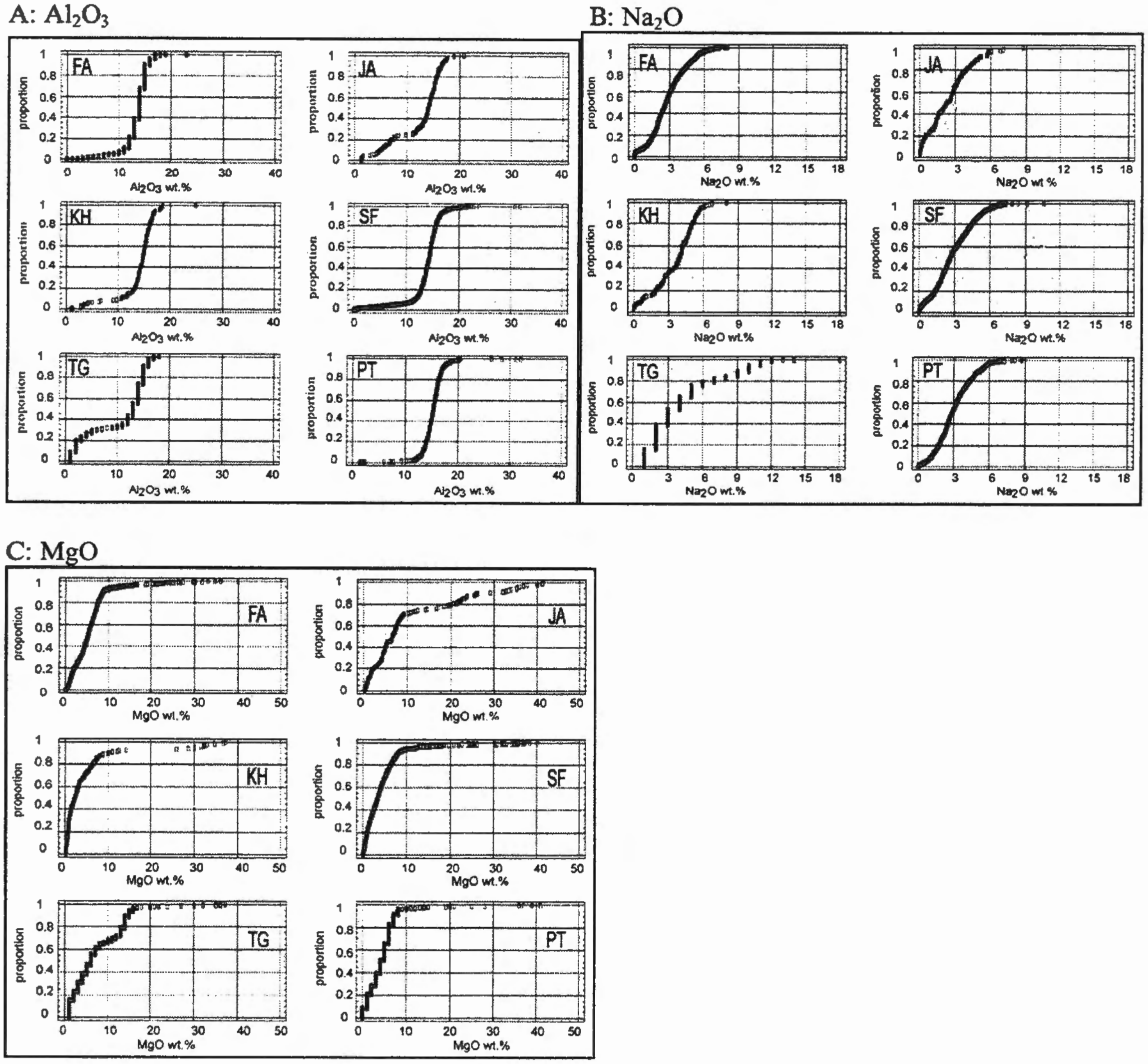


Figure 6: Comparison of Na_2O and SiO_2 populations (A, and B, respectively) in the corrected and combined dataset. Notice SiO_2 versus Na_2O scatterplot (C) now shows a single larger cluster.

systematic recording error within the TG dataset. SiO_2 and Na_2O values for the 198 samples were therefore switched, and the probability plots and histograms recalculated. These are shown in Figure 6a-c. Both SiO_2 and Na_2O now display single population distributions with meaningful geochemical values. Had this error not been detected, any subsequent statistical procedures, such as percentile calculation and principal component analyses, would have been significantly affected.

4.2.2 Levelling of datasets

Examination of oxide populations across the belt by dataset revealed sub-populations of Al_2O_3 < 5, Na_2O >5 and MgO >10 weight % within the TG dataset (Figures 7a, b and c, respectively). These sub-populations were not duplicated or were much smaller within the other datasets, suggesting a problem with the TG data, and possible need to eliminate it from the combined dataset. To test whether the problem represented simply "bad data", or if it was due to the location of individual samples with respect to varying lithologies, the spatial characteristics of each dataset were examined. Although all datasets, except KH, are dominated by mafic metavolcanics, examination of Figure 7d, displaying the location of each point in the combined data by dataset origin, as well as a breakdown of each dataset displayed graphically as pie charts, indicates a number of sampling trends. KH data is located predominantly within granitoids; JA within the northern Swayze mafic metavolcanics; PT within mafic metavolcanics in the southeast high strain area; FA in N-S oriented grid lines within the center (undeformed) portion the belt dominated by mafic metavolcanics; SF strongly clustered in mafic metavolcanics in the north and south, and TG somewhat scattered within primarily undeformed mafic metavolcanics. In addition, although mafic metavolcanics typically account for 34-62% of samples within a dataset, the remaining 38-66% of samples are quite variable between datasets, mafic-ultramafic intrusive rocks being strongly sampled within JA, PT, KH and TG, while metasedimentary rocks were strongly sampled within KH, TG and FA (Figure 8). It is therefore possible that the TG dataset sub-populations represent a spatial or sampling trend rather than a ubiquitous, but non-quantifiable dataset error. To test this, Al_2O_3, Na_2O and MgO populations were broken down into specific ranges and plotted statistically using box and whisker plots, and spatially on an existing geology map (1:50,000 OGS compilation, supplied by Falconbridge Ltd.) by dataset (Figures 9, 10, and 11, respectively).


A: Al_2O_3

B: Na_2O

C: MgO

Figure 7: Q-plots of Al_2O_3 (A), Na_2O (B), and MgO (C) populations by dataset.

Samples with Al_2O_3 < 10 weight % are plotted by dataset on Figure 9. Examination of the accompanying box and whisker plots demonstrates that Al_2O_3 values for the TG dataset are significantly lower compared to the other datasets. The majority of the TG samples are clustered about the Isaiah Lake

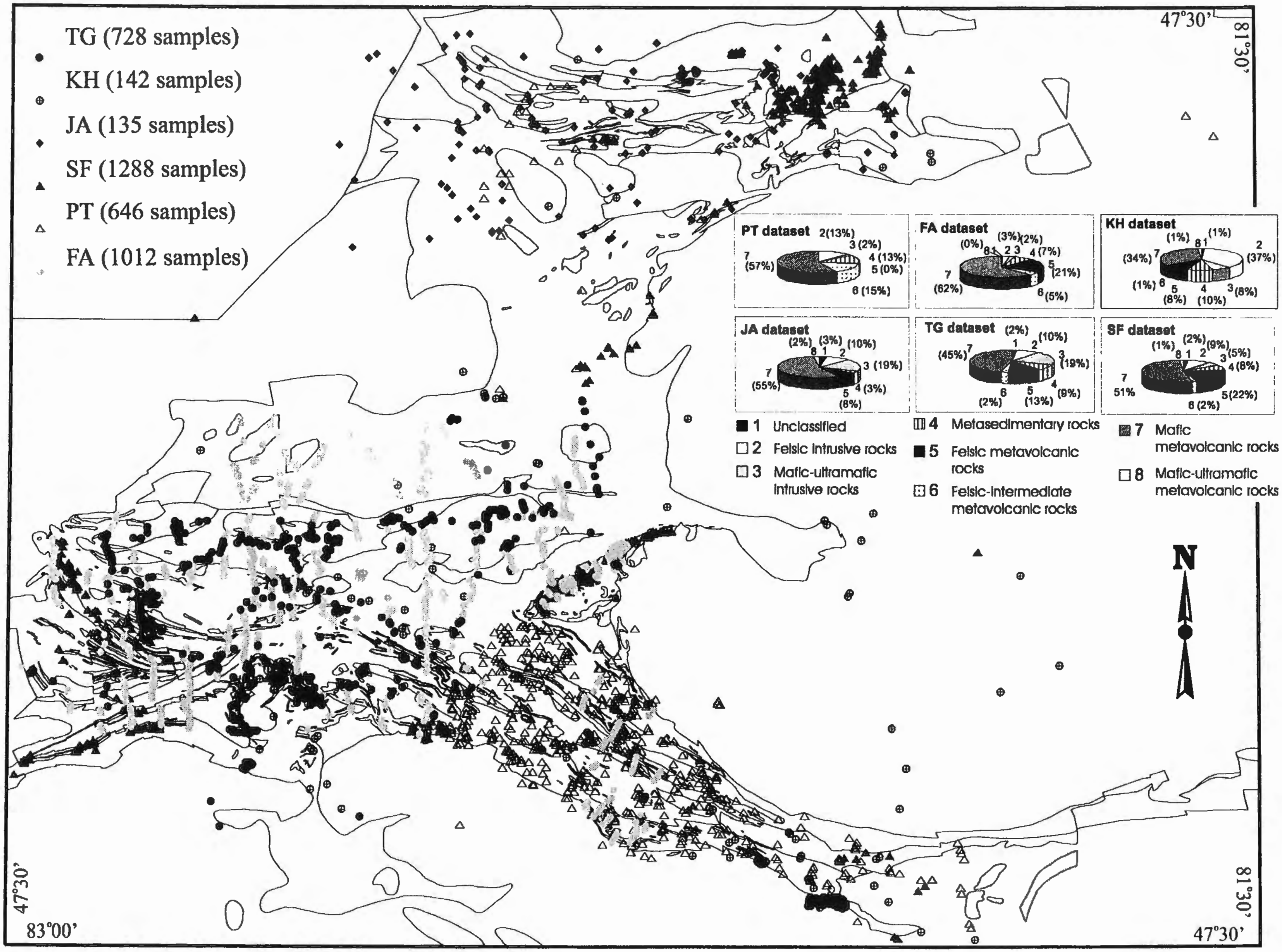

TG (728 samples)
KH (142 samples)
JA (135 samples)
SF (1288 samples)
PT (646 samples)
FA (1012 samples)
47°30'
81°30'
83°00'
47°30'
81°30'
47°30'
PT dataset 2(13%) 3 (2%) 4 (13%) 5 (0%) 6 (15%) 7 (57%)
FA dataset (0%) 8 1 (3%) 2 (2%) 3 4 (7%) 5 (21%) 6 (5%) 7 (62%)
KH dataset (1%) 8 1 (1%) 2 (37%) 3 (8%) 4 (10%) 5 (8%) (1%) 6 (34%) 7
JA dataset (2%) 8 (3%) 1 2 (10%) 3 (19%) 4 (3%) 5 (8%) 7 (55%)
TG dataset (2%) 1 2 (10%) 3 (19%) 4 (9%) 5 (13%) 6 (2%) (45%) 7
SF dataset (1%) 8 (2%) 1 2 (9%) 3 (5%) 4 (8%) 5 (22%) 6 (2%) 7 51%
■ 1 Unclassified
□ 2 Felsic intrusive rocks
□ 3 Mafic-ultramafic intrusive rocks
▥ 4 Metasedimentary rocks
■ 5 Felsic metavolcanic rocks
⊞ 6 Felsic-intermediate metavolcanic rocks
▨ 7 Mafic metavolcanic rocks
□ 8 Mafic-ultramafic metavolcanic rocks
N

Figure 8: Spatial distribution of all oxide samples by dataset.

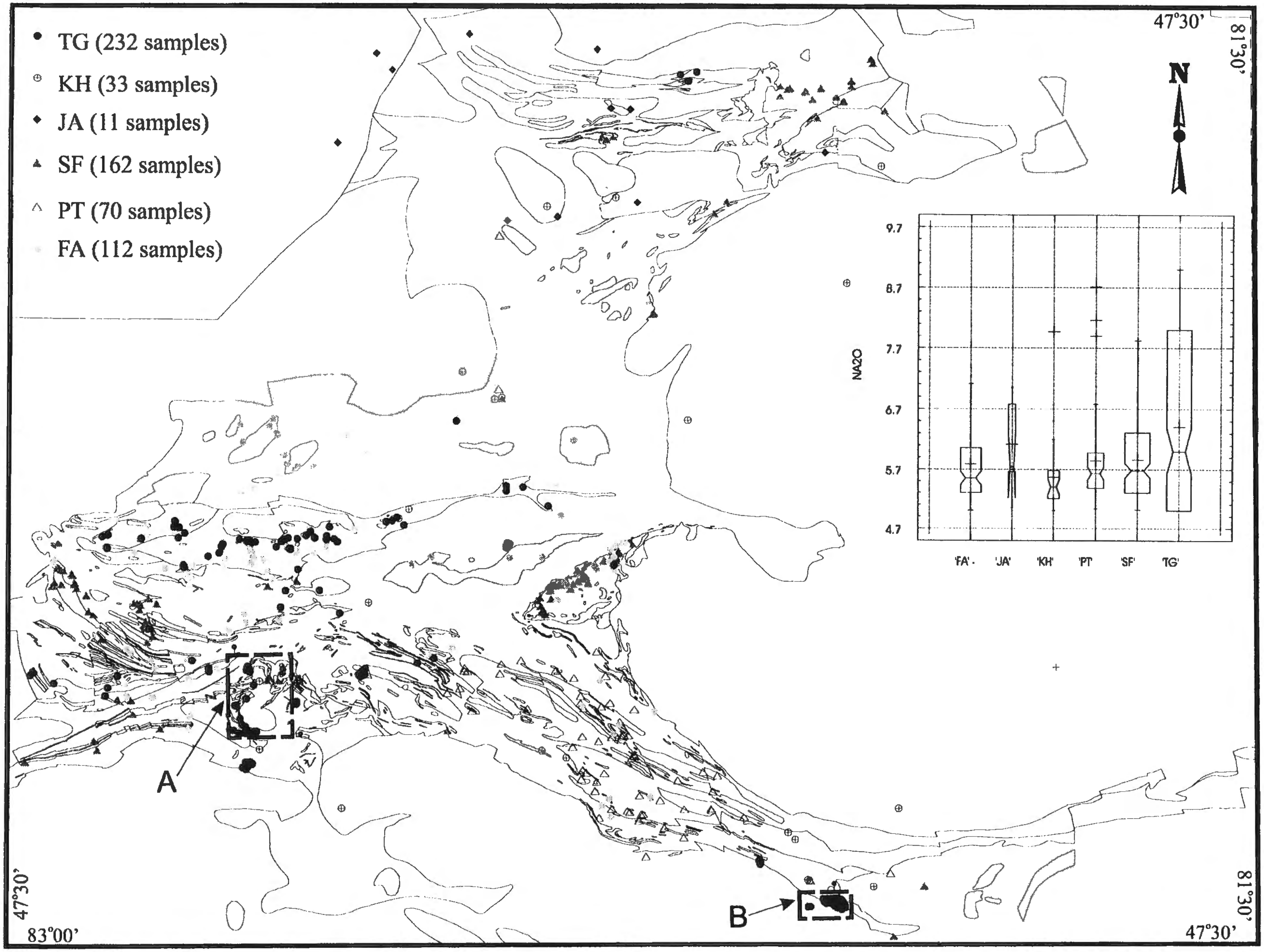


Figure 9: Statistical, and corresponding spatial distribution of Al_2O_3 between 0 and 10 weight %, by dataset.

Stock, indicated by black box (labeled "A") in the lower left portion of the map, and a small sliver of greenstone in the southeast portion of the main belt (lower right of map - labeled "B").

Na_2O between 5 and 10 wt.%, and MgO greater than 10 wt.% samples are plotted by dataset on the geologic map, and statistically as a box and whisker plots in Figures 10 and 11, respectively. These plots show that the TG dataset is characterized by significantly lower MgO and higher Na_2O (with the exception of the JA dataset) values than the other datasets. Again, the majority of TG samples plot around the Isaiah Lake Stock area and the lower right corner of the map area. This spatial bias in the TG dataset indicates that the dataset as a whole is not biased in MgO, Na_2O and Al_2O_3, but that the problem is areally restricted. It is therefore apparent that two areas of the belt (labeled "A" and "B" on Figures 9, 10 and 11), are marked by anomalous chemistry with respect to the TG dataset. There are 3 possibilities to investigate for the source of these apparent anomalies: one, systemic errors associated with the TG dataset, two, lithologic variation of samples, and three, spatially-restricted alteration effects.

Box and whisker plots of the TG samples over the Isaiah Lake Stock area show that the anomalous chemistry is not related systematically to lithology (Figure 12). While Na_2O anomalies occur mainly within mafic metavolcanic rocks, Al_2O_3 and MgO anomalies occur within felsic intrusive rocks, felsic-intermediate metavolcanic rocks, mafic metavolcanic rocks and mafic-ultramafic intrusives rocks. For the TG anomaly to be due to alteration (which does not affect the other datasets), one would expect that the TG dataset to be isolated spatially. This is, in fact, the case for the Isaiah Lake Stock area shown on Figure 13. Although the TG samples are clustered in the local area, a few TG samples occur in close proximity to FA samples within mafic metavolcanic rocks in the northwest. Comparison of oxide values reveals Al_2O_3 is approximately 13 wt.% for the FA sample and only 1-2 wt.% for the TG sample. Conversely Na_2O is 2-3 wt.% in FA and 9-10 wt.% in the TG sample. MgO and FeO values are likewise very different. A similar comparison within felsic intrusive rocks corroborates that the anomalous chemistry is consistent across lithology. This lack of consistency across lithology and datasets, and the unusual chemistry of the anomalous samples, suggests an alteration origin for the anomalous data is not plausible. Therefore, it is likely that the anomalous Isaiah Lake Stock area TG chemistry is due simply

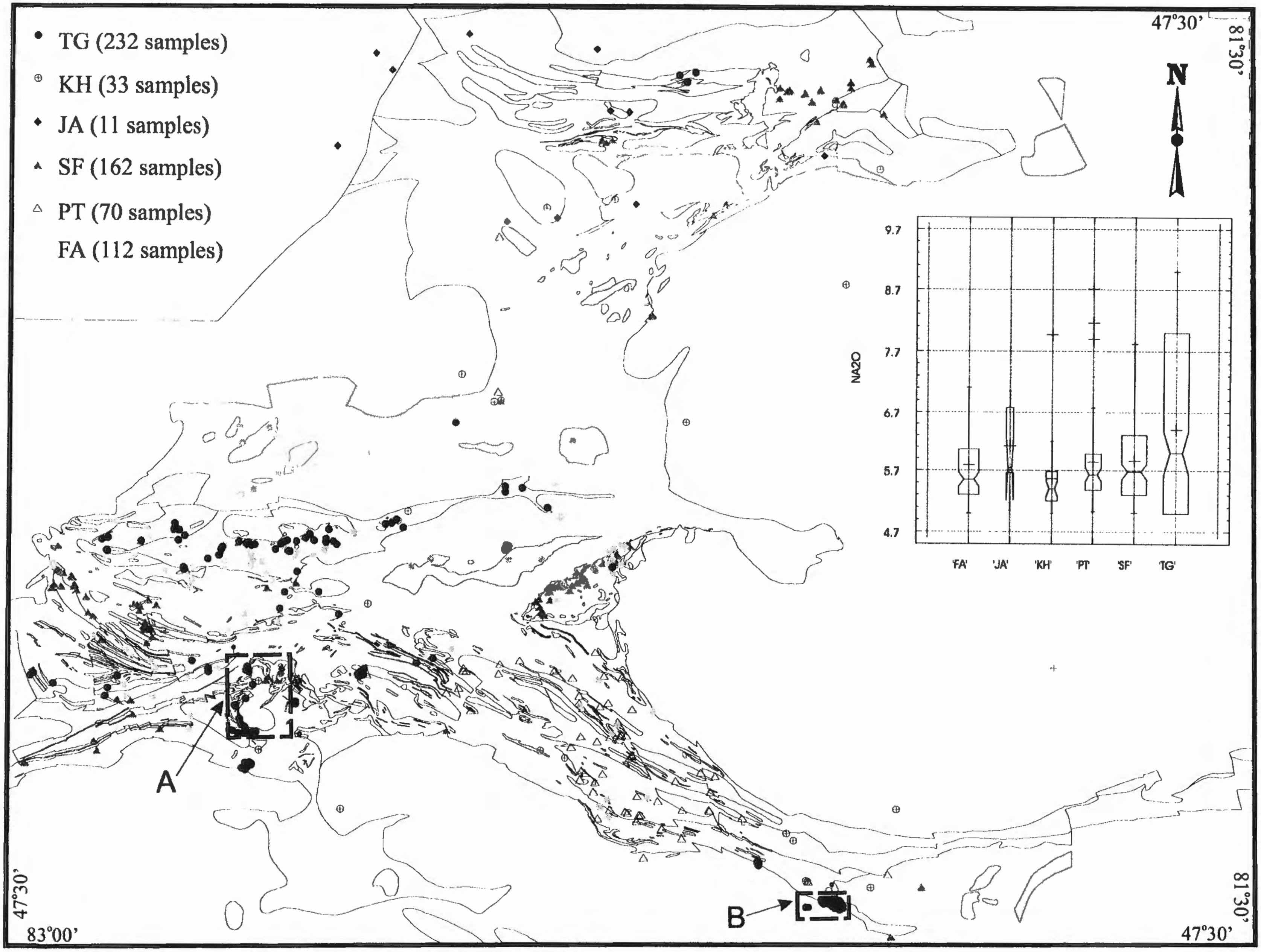


Figure 10: Statistical, and corresponding spatial distribution of Na_2O greater than 5 weight %, by dataset.

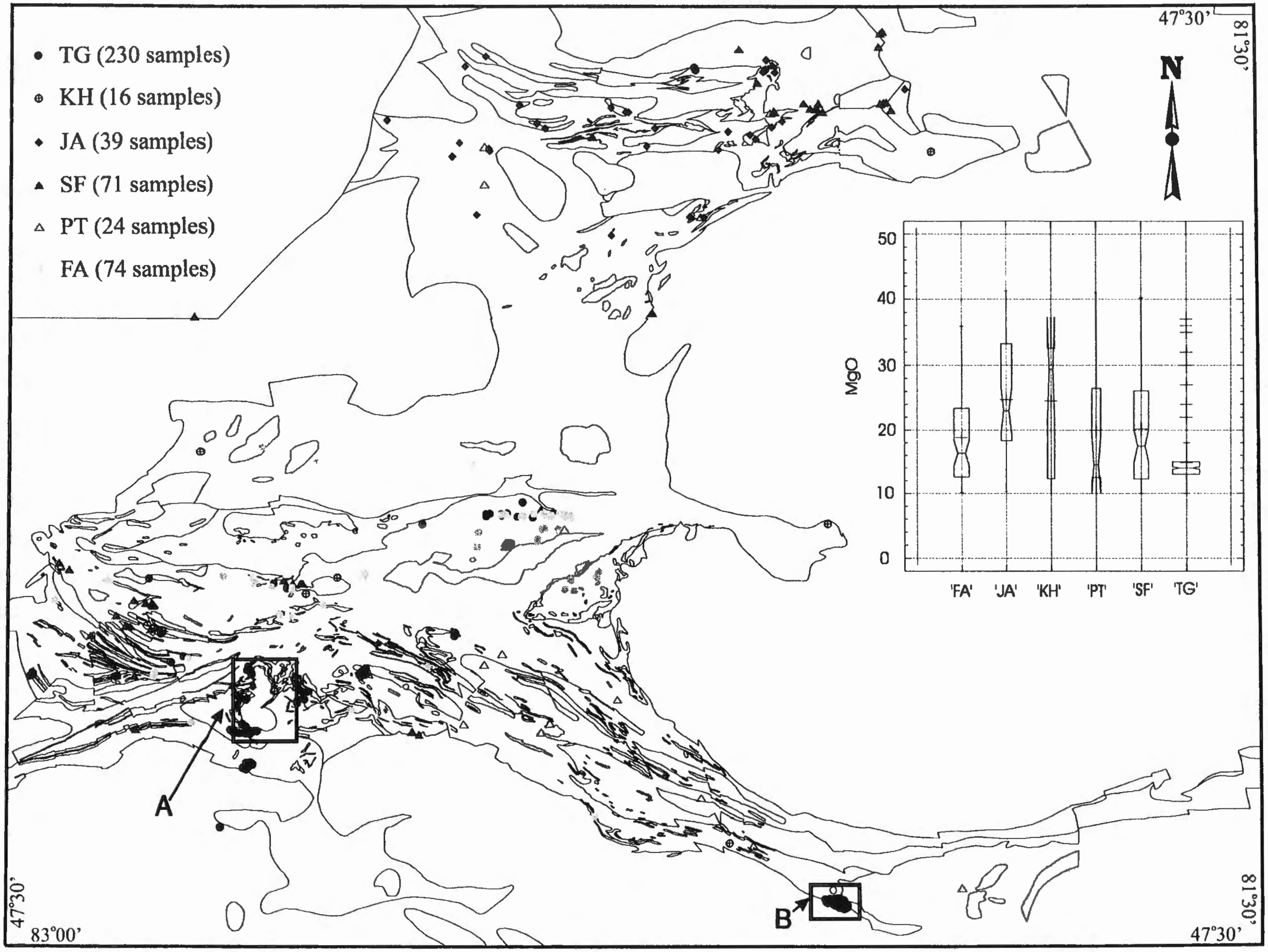


Figure 11: Statistical, and corresponding spatial distribution of MgO greater than 10 weight %, by dataset.

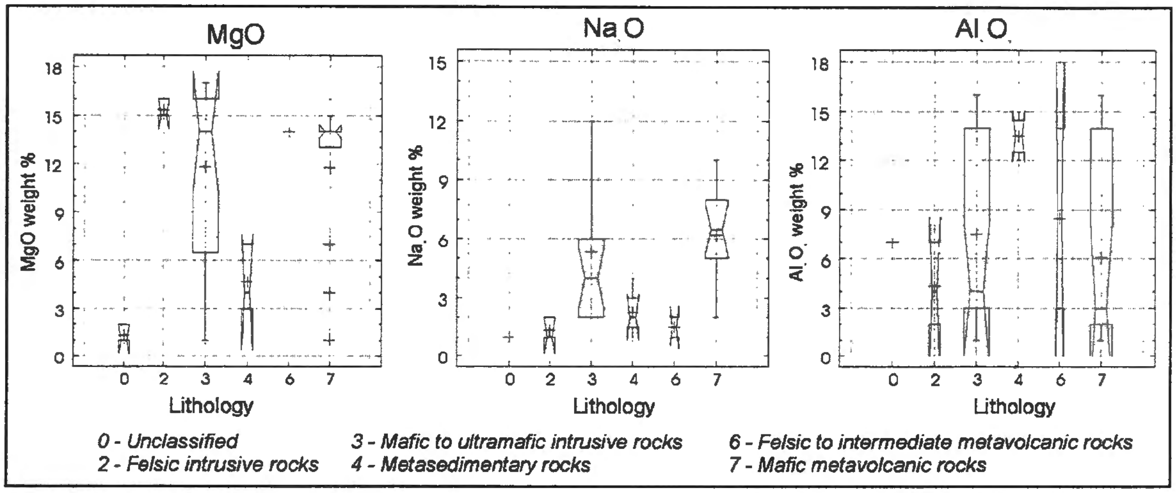


Figure 12: Anomalous oxides (MgO, Na_2O, and Al_2O_3) by mapped lithology over the Isaiah Lake stock area.

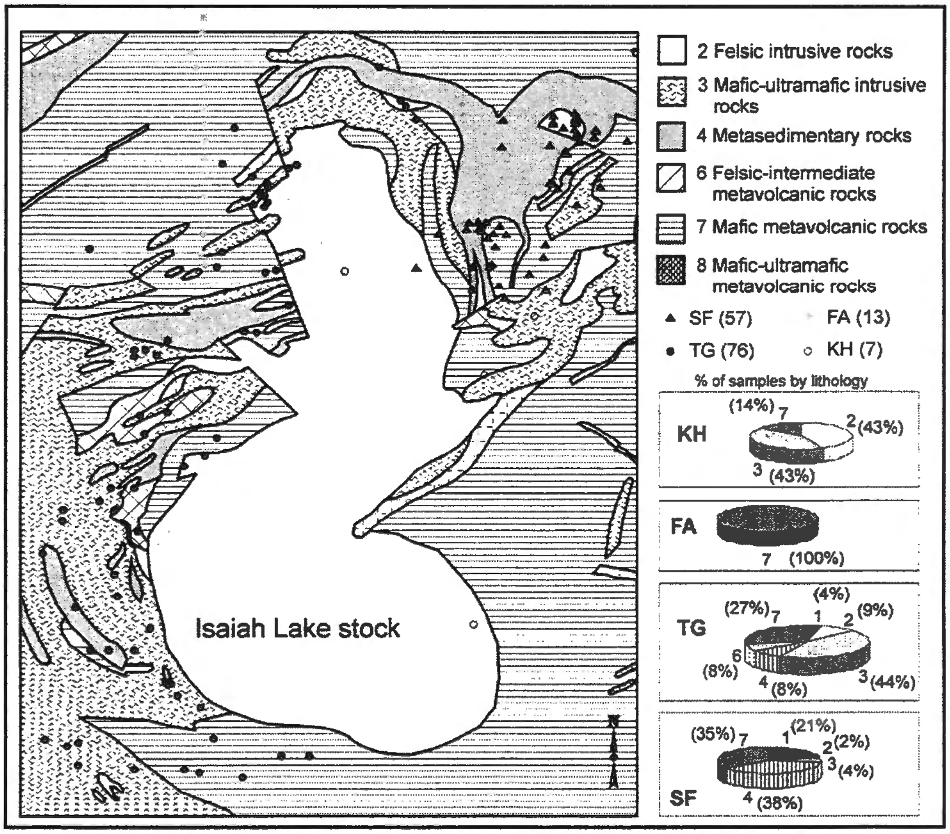
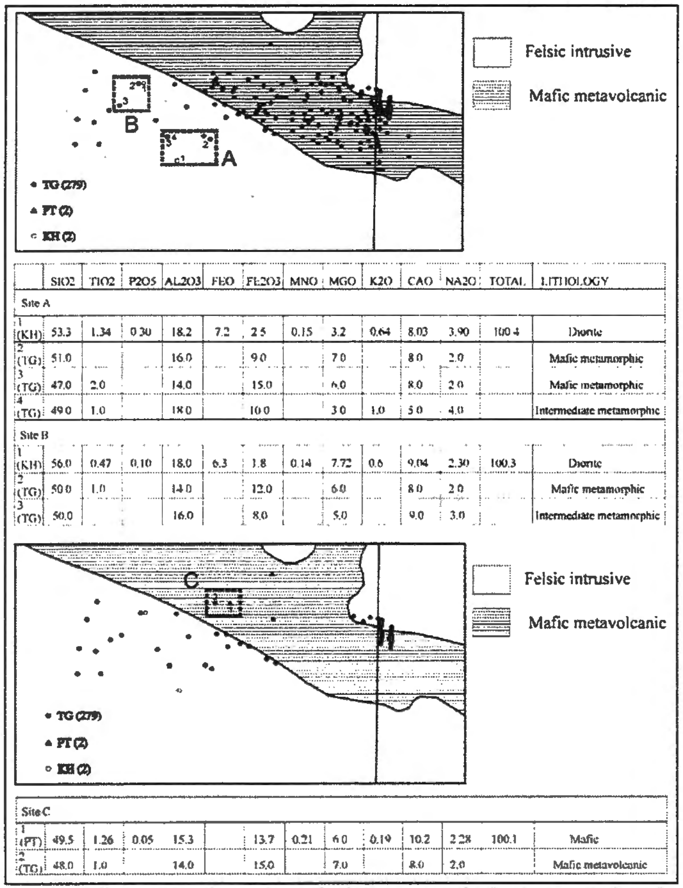


Figure 13: Spatial relationships between samples by dataset in the Isaiah Lake stock area.

to errors in the data. As a result, 76 samples from the TG dataset in this area were discarded.

The area in the lower corner of NTS sheet 41O/9 (area "B" on Figures 9, 10 and 11) is somewhat different in that the majority of samples are TG. This makes the assessment of the origin of the anomalous chemistry more difficult, as there are fewer samples from the other datasets with which to compare the TG samples. Samples occur within only two lithologies; mafic metavolcanic rocks and felsic intrusive rocks (Figure 14). The TG geochemical data within the felsic intrusive data were checked by comparing to two nearby KH samples (see areas "A" and "B" on Figure 14). This comparison showed that the TG and KH values were similar. Within the mafic metavolcanic rocks, TG samples often contained no FeO or Fe_2O_3. Samples with Fe_2O_3 < 5 wt.%, were characterized by other anomalous oxides, namely Al_2O_3 < 5 wt.% and Na_2O > 9 wt.%. All samples lacking Fe or with Fe_2O_3 <5 (a total of 113 samples) were eliminated. The remaining samples were compared with PT samples in close proximity and revealed excellent correlation.



	SIO2	TIO2	P2O5	AL2O3	FEO	FE2O3	MNO	MGO	K2O	CAO	NA2O	TOTAL	LITHOLOGY
Site A													
1 (KH)	53.3	1.34	0.30	18.2	7.2	2.5	0.15	3.2	0.64	8.03	3.90	100.4	Diorite
2 (TG)	51.0			16.0		9.0		7.0		8.0	2.0		Mafic metamorphic
3 (TG)	47.0	2.0		14.0		15.0		6.0		8.0	2.0		Mafic metamorphic
4 (TG)	49.0	1.0		18.0		10.0		3.0	1.0	5.0	4.0		Intermediate metamorphic
Site B													
1 (KH)	56.0	0.47	0.10	18.0	6.3	1.8	0.14	7.72	0.6	9.04	2.30	100.3	Diorite
2 (TG)	50.0	1.0		14.0		12.0		6.0		8.0	2.0		Mafic metamorphic
3 (TG)	50.0			16.0		8.0		5.0		9.0	3.0		Intermediate metamorphic

Site C													
1 (PT)	49.5	1.26	0.05	15.3		13.7	0.21	6.0	0.19	10.2	2.28	100.1	Mafic
2 (TG)	48.0	1.0		14.0		15.0		7.0		8.0	2.0		Mafic metavolcanic

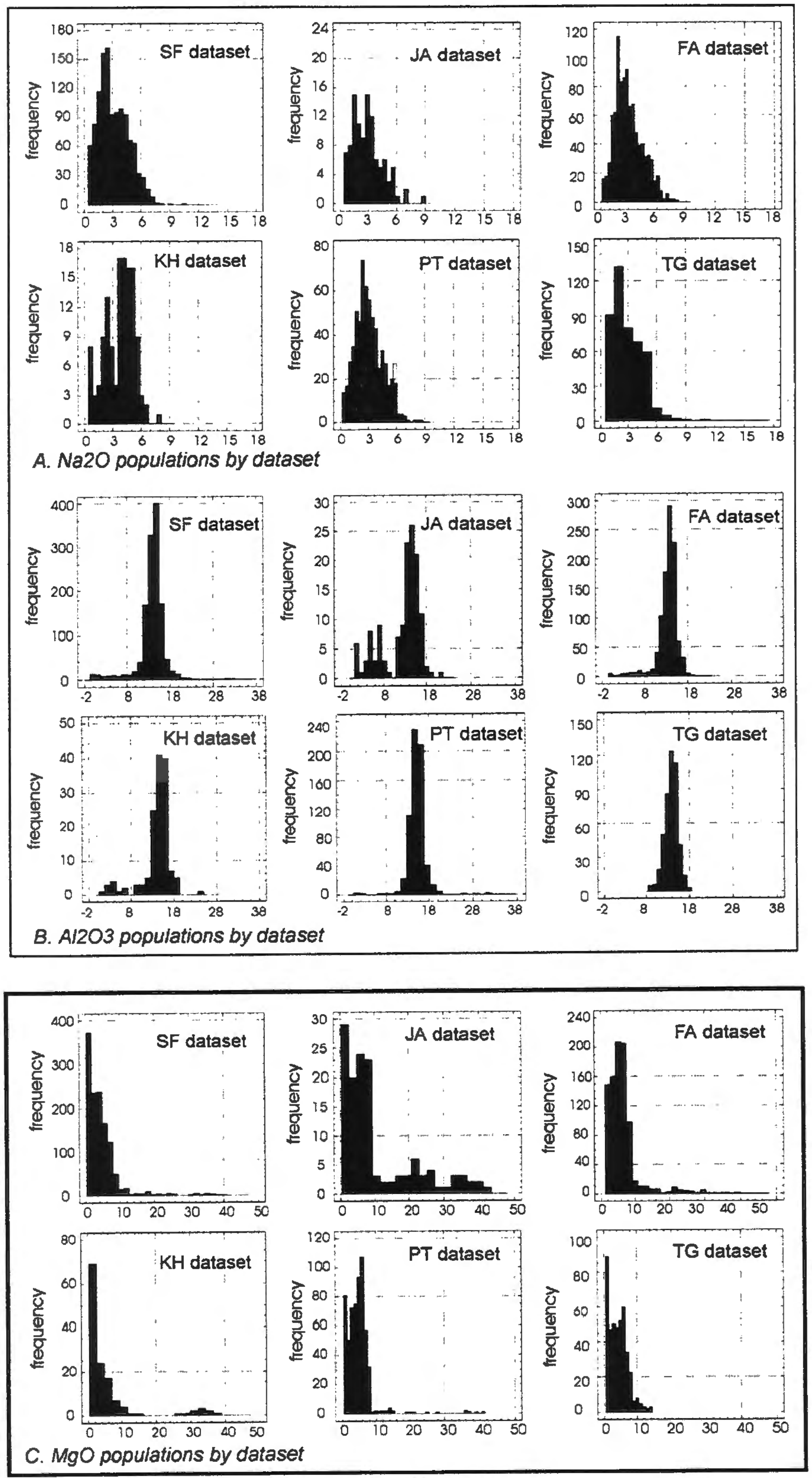
Figure 14: Original spatial relationships between samples by dataset (top), and new spatial relationships with anomalous samples removed (bottom). Charts document similarities in remaining TG data with respect to other datasets.

Identification of systematic problems with the TG dataset in these two areas, discussed above, necessitates an evaluation of TG samples outside these areas. Using the same criteria discussed above ($Fe_2O_3 < 5$ wt.% and $Al_2O_3 < 5$ wt.%), an additional 98 TG samples were flagged as "bad", and eliminated from the dataset. In total, 287 "bad data" samples were eliminated from the TG dataset by using these screening procedures.

It is often helpful to compare data sets graphically by plotting adjacent histograms as shown in Figure 15. Re-plotted Na_2O, Al_2O_3, and MgO (Figure 15a, b, and c, respectively) sample values by dataset demonstrates that the anomalous TG dataset sub-population has been eliminated, and the TG dataset distribution is now similar to the other 5 datasets. Statistics for each dataset source population are tabulated in Table 5, for Al_2O_3, Na_2O and MgO. Standardized skewness and kurtosis values within the -2 to 2 range indicate a normal distribution, and it can be seen from Table 5, that few populations are normally distributed, particularly for the oxides Al_2O_3 and MgO.

	Size	Mean	Std. dev.	Min.	Max.	Skewness	Std. Skewness	Kurtosis	Std. Kurtosis
					Al_2O_3				
KH	153	14.06	3.62	1.0	24.76	-1.83	-8.40	4.36	9.97
JA	135	12.32	4.52	1.32	20.83	-1.02	-4.83	-0.06	-0.14
SF	1286	13.84	3.12	0.0	31.8	-1.76	-25.80	8.02	58.71
PT	646	15.17	2.28	1.44	32.1	-0.13	-1.39	18.90	98.03
TG	460	13.91	1.56	9.00	18.0	-0.39	-3.40	0.45	1.97
FA	1012	13.42	2.57	0.00	23.0	-2.28	-29.62	52.22	52.22
					Na_2O				
KH	149	3.59	1.70	0.00	7.97	-0.44	-18.4	-0.44	-1.61
JA	126	2.46	1.78	0.01	8.71	0.68	3.16	0.32	0.65
SF	1274	2.88	1.72	0.00	10.6	0.43	6.24	-0.15	-1.29
PT	642	3.01	1.51	0.01	8.71	0.45	4.67	0.06	0.36
TG	451	2.92	1.73	1.00	18.00	2.13	18.50	12.86	55.75
FA	1012	2.87	1.56	0.00	8.06	0.43	5.56	-0.08	-0.51
					MgO				
KH	153	4.91	7.81	0.14	37.1	2.91	13.78	7.99	17.99
JA	135	10.37	10.64	0.32	41.29	1.43	6.76	1.01	2.39
SF	1286	4.31	4.99	0.00	40.2	3.65	53.39	18.01	131.62
PT	646	4.56	4.29	0	41.00	4.33	44.90	29.44	152.75
TG	430	4.33	2.77	1.00	14.00	0.77	6.55	0.50	2.11
FA	1012	5.69	4.81	0	39.90	2.90	37.68	12.18	79.00

Table 5: Summary of cleaned oxide database by input dataset source for Al_2O_3, Na_2O and MgO.



A. Na2O populations by dataset

B. Al2O3 populations by dataset

C. MgO populations by dataset

Figure 15: Corrected Na_2O (A), Al_2O_3 (B), and MgO (C) oxide populations by dataset.

Box and whisker plots of each oxide by dataset are shown in Figure 16, and demonstrate that there is some statistical variation between dataset populations, as indicated by non-overlapping notches. This is not surprising however, given the variable spatial distribution and the resulting wide range of lithologies sampled by each data set.

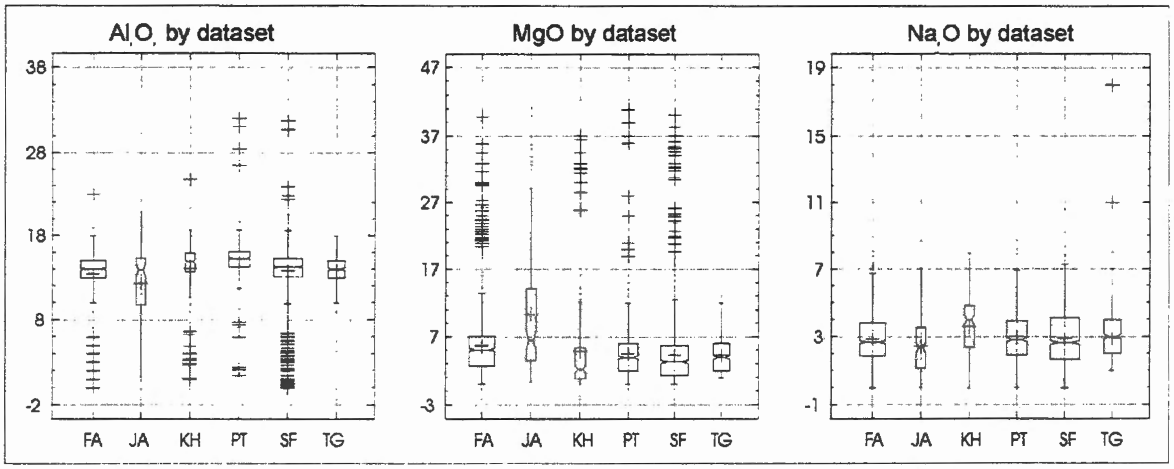
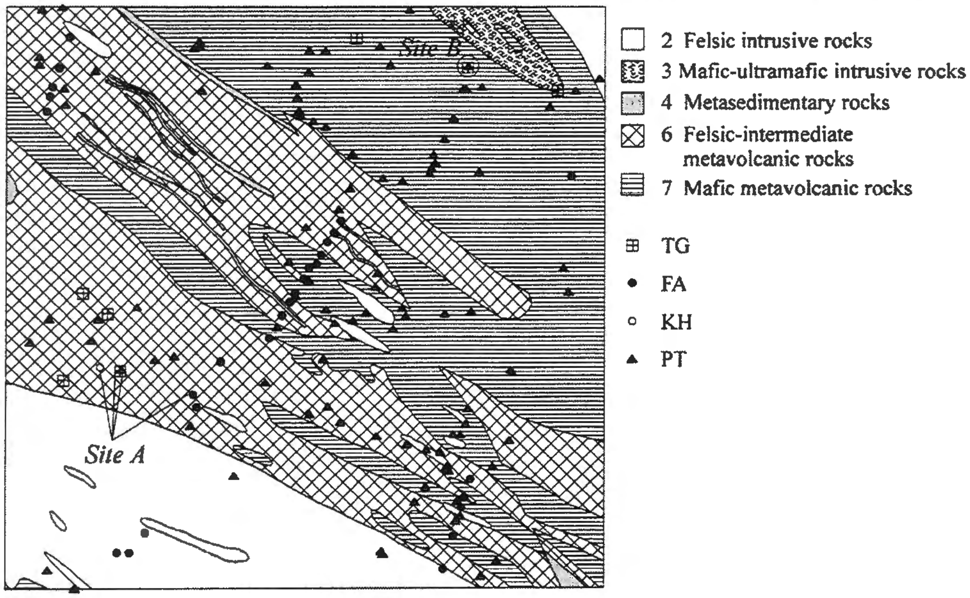


Figure 16: Box and whisker plots showing Al_2O_3, MgO, and Na_2O populations by dataset (for all lithologies). Lack of overlap between notches indicates statistically distinct population (e.g., KH and PT in MgO).

4.2.3 Lack of "total" oxide data

Although some errors in the TG data have been eliminated, the TG dataset is still problematic given the lack of "total" oxide information, and the paucity of oxide analyses for TiO_2, P_2O_5, K_2O and MnO. Therefore, a comparison of TG oxide analyses with analyses from other datasets via various lithologies was undertaken to assess the accuracy of the oxide measurements in the TG dataset. The results of this comparison are shown in Figure 17, which summarizes oxide concentrations in felsic-intermediate metavolcanic rocks and in mafic metavolcanic rocks in two different sites ("A" and "B" - see Figure 17) for different datasets. In general, oxide values for TG correspond well with other, more complete samples from other datasets at these particular sites, and we infer that these data are acceptable for use in further analysis. However, it is not acceptable to use this data in any calculations or classifications that necessitate a complete oxide composition.


Site B
Site A
□ 2 Felsic intrusive rocks
▨ 3 Mafic-ultramafic intrusive rocks
▨ 4 Metasedimentary rocks
☒ 6 Felsic-intermediate metavolcanic rocks
▤ 7 Mafic metavolcanic rocks
⊞ TG
● FA
○ KH
▲ PT

	SIO2	TIO2	P2O5	AL2O3	FEO	FE2O3	MNO	MGO	K2O	CAO	NA2O	LOI	H2O	CO2	TOTAL
Site A															
PT	56.4	0.77	0.15	15.7		8.41	0.11	5.0	1.02	6.15	3.38	2.0		0.11	98.54
TG	57.0			15.0		7.0		4.0		5.0	3.0				
KH	56.33	0.91	0.17	16.04		6.00	0.11	4.91	0.67	5.29	4.53	2.0	2.7	0.2	99.94
FA	56.3	1.02	0.42	15.0	5.22		0.11	3.29	.18	5.59	3.17	6.23			99.7
Site B															
TG	47.0			14.0		12.0		7.0		7.0	1.0				
PT	48.1	1.19	0.12	16.2		12.4	0.21	5.0	0.66	5.7	3.0	6.0		2.39	99.10

Figure 17: Comparison of TG oxide concentrations in felsic to intermediate and mafic metavolcanic rocks, with nearby samples from other sources.

4.2.4 Spatial control of oxide populations

The variability of oxides by datasets and even within datasets is controlled in part by the sampling process with respect to lithology. The effects of this can been seen within "normal" oxide populations. For example, two felsic units (unit 1 and unit 3 shown on Figure 18) were selected for comparison on the basis of dense whole rock sampling (unit 1 > 120 samples, unit 2 > 50 samples). A

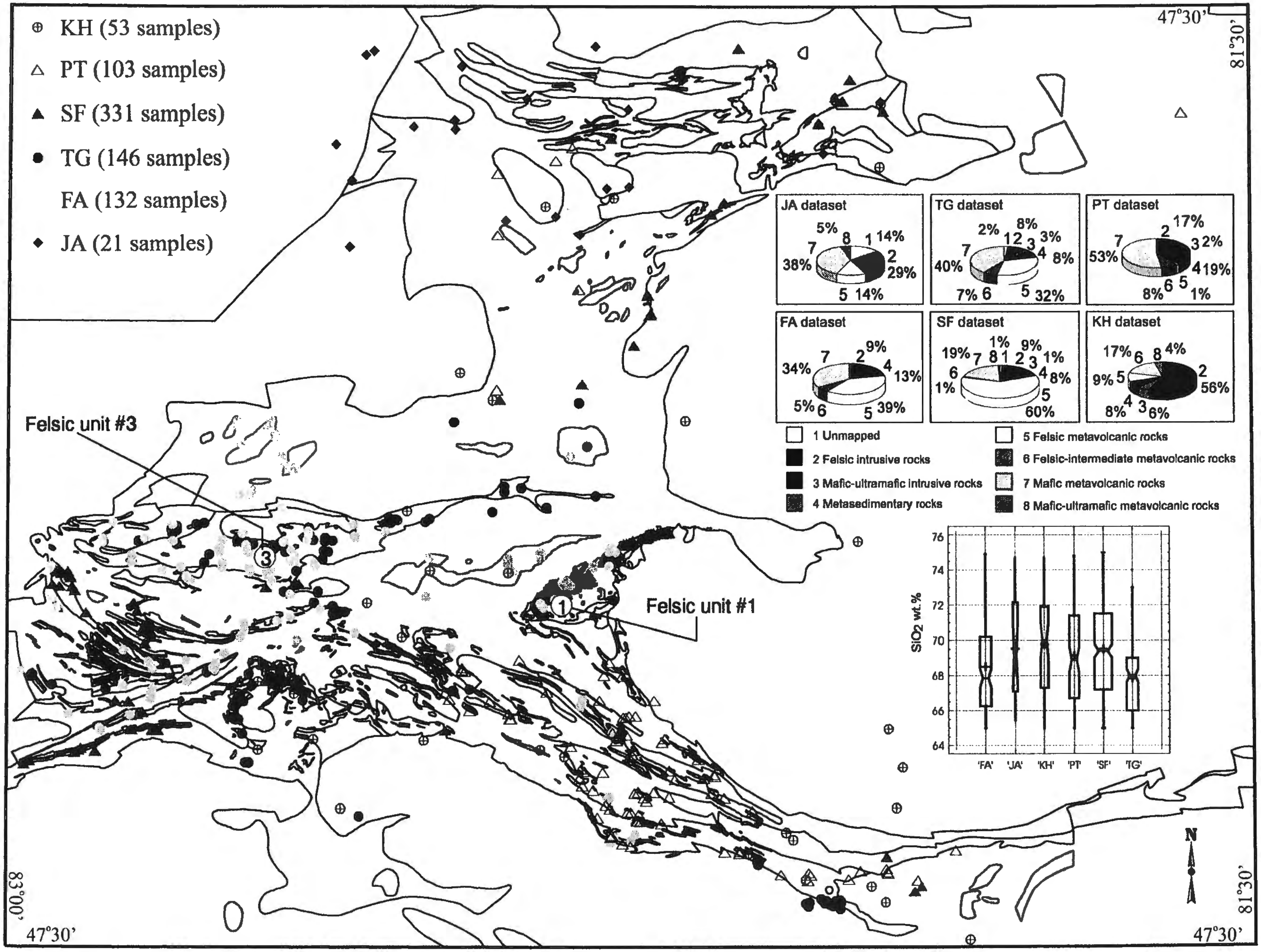


Figure 18: Statistical, and corresponding spatial distribution of SiO_2 between 65 and 75 weight %, by dataset.

"point-in-polygon"[4] operation was undertaken using the GIS to identify the geochemical sample points that fell within each felsic lithology mentioned above. It is recognized that the mapped geology is a generalization of the diversity of lithology identified on the ground, thus our mapped felsic units may have interbedded iron-formation or mafic units. To ensure variability of oxides is a reflection of sampling, rather than lithologic variation, these samples were then categorized into volcanic rock types using the Al-Fe+Ti-Mg cation classification (Jensen 1976) which is a popular scheme for establishing volcanic rock types from chemical analyses. Samples that were not classified as felsic (i.e. calc-alkaline rhyolites - CR) and tholeiitic rhyolites - TR) were eliminated as unreliable as their Al-(Fe+Ti)-Mg classification did not match the mapped lithology, in this case felsic volcanic rocks. Summary statistics for these samples were calculated, the results of which are shown in Table 6. Units 1 and 3 appear to be chemically distinct with SiO_2 and K_2O higher, and all other oxides lower in unit 1 (Table 6). Standardized skewness and kurtosis values calculated for each unit, and each oxide, indicate all oxides are normally distributed in unit 3, with the exception of P_2O_5 and MnO. Unit 1 oxides range from normally distributed (P_2O_5, FeO, Na_2O, SiO_2, MgO, TiO_2 and CaO), to log-normally distributed (Fe_2O_3, MnO), to strongly skewed (Al_2O_3, K_2O). The assessment of the significance of the difference in means between the unit 1 and unit 3 oxide data, using the t-test, requires normal distributions, and no significance difference in the variances at the 95% confidence interval. Only SiO_2, FeO, Na_2O, TiO_2, CaO and MgO populations can be compared on the basis of normal populations, and of these only TiO_2 and CaO do not have a statistically significant difference in variance at the 95% confidence interval. The results from the t-test for these two oxides indicate both have a statistically significant difference in means at the 95% confidence level. In addition, unit 1 appears to be more chemically variable than unit 3, as shown by the typically larger range in oxide values. It must be noted however, that sampling (and by corollary, dataset origin) may be a control on felsic unit chemistry. Figure 18 shows SiO_2 samples, between 65 and 75 wt.%, by dataset origin. Note the concentration of SF samples within unit 1 and FA and TG samples within unit 3. Box and whisker

[4] "Point-in-polygon" operation refers to an intersection operation of the GIS in which the lithogeochemical point cover is overlain on the geological unit polygon map to identify the points which overlap with each polygon. The results of the intersection can then be extracted for each polygon.

plots of SiO_2 between 65 and 75 wt.% (Figure 19) reveal systematically lower SiO_2 content within FA and TG datasets overall. Thus felsic unit 3, dominated by FA and TG datasets, may be characterized by a lower SiO_2 content than unit 1 simply due to the dataset which dominates its sampling.

	Unit 1			***Unit 3***		
	# of samples	**min-max**	**mean**	**# of samples**	**min-max**	**mean**
SiO_2	184	63.5-81.0	71.99	56	61.6-70.9	65.55
TiO_2	182	0.04-0.52	0.27	28	0.23-0.5	0.33
P_2O_5	182	0.03-0.16	0.08	28	0.08-0.28	0.12
Al_2O_3	184	11.8-20.3	14.54	56	12.3-18.1	15.02
Na_2O	184	0.07-6.92	3.70	56	2.0-7.74	4.51
CaO	184	0.08-6.43	1.99	56	1.0-6.0	2.67
K_2O	182	0.11-8.01	2.34	52	0.43-3.8	1.7
MgO	184	0.01-1.71	0.56	56	0.59-1.47	0.98
MnO	182	0.0-0.3	0.06	28	0.0-0.19	0.05
Fe_2O_3	174	0.29-7.26	2.16	31	0.90-3.0	2.65
FeO	13	0.0-3.98	1.72	26	1.26-2.45	1.78

Table 6: Summary of major oxides for unit 1 versus unit 3 felsic metavolcanic packages.

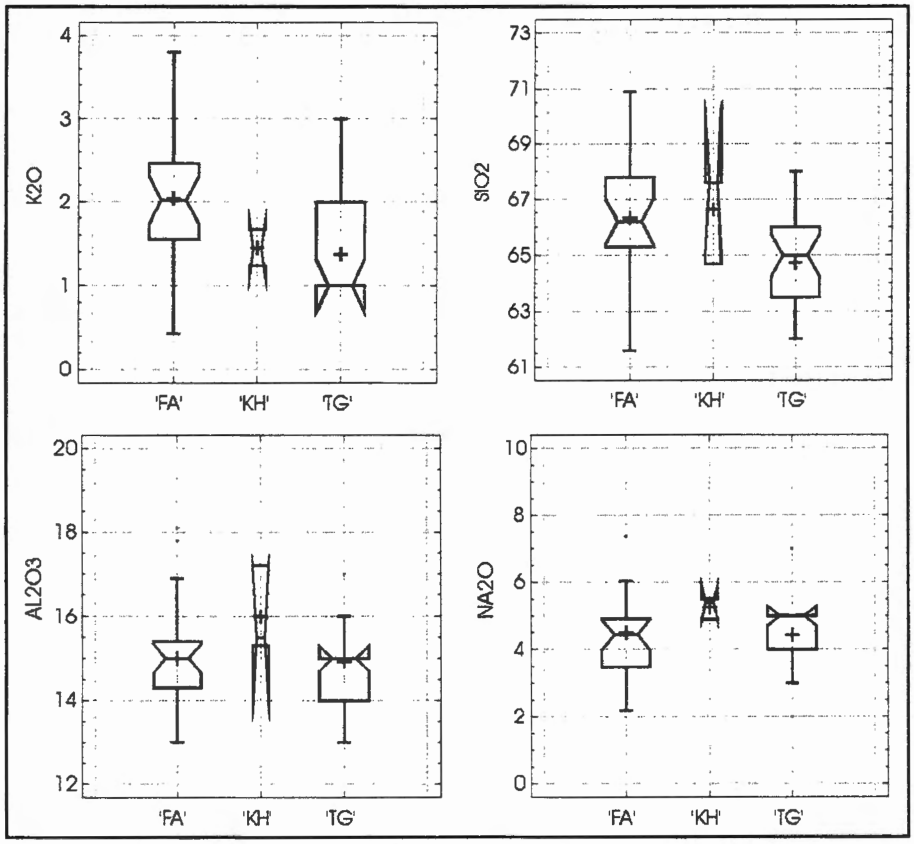


Figure 19: Box and whisker plots of felsic unit three for K_2O, Al_2O_3, SiO_2, and Na_2O concentrations, by dataset source.

A paucity of samples within unit 1 does not permit a rigorous statistical comparison between datasets, as FA has only 10 samples, and TG, 2, with the majority of samples coming from SF (172). Likewise, unit 3 contains samples from 3 different datasets, with FA comprising 25 and TG 28 samples while the KH dataset contains only 3 samples. Figure 19 shows box and whisker plots for several oxide analyses for unit 3. It can be seen that the range of values for SiO_2, Al_2O_3, and Na_2O are similar between the two datasets FA and TG. However, there is a difference in K_2O between TG and FA. A plot of unit 3 samples by dataset (Figure 20) indicates that the central portion of the belt has been more selectively sampled in the FA dataset, which may explain the difference in K_2O values.

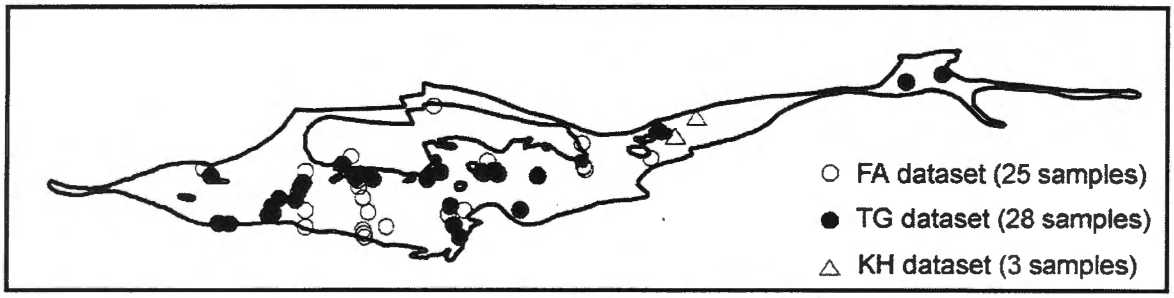


Figure 20: Spatial distribution of felsic samples by dataset origin, with felsic unit 3.

4.3 Censored data

An important consideration for industry is to minimize the amount of time and money chasing after false anomalies (i.e., anomalies not related to mineralization). It is always a difficult task for a geochemist to determine a threshold. If it is too low then there are too many anomalies. If it is too high, then potentially valuable ground may be missed. Thus a careful evaluation of the population distribution, and a clear understanding of the lithologies must be kept in mind when determining suitable thresholds. Censored data (i.e., geochemical analyses that fall below the detection limit of the particular technique used to analyze the data) can bias the computation of distribution parameters (i.e., mean, variance), and potentially affect the determination of preliminary thresholds.

The problem of censored data has been studied by Chung (1985, 1988), and Sanford et al. (1993). Many laboratories have reported data that is less than the detection limit by recording the detection limit value as a negative number (e.g. -5 ppm), or occasionally as less than values (e.g. < 5 ppm). A common method of handling this data is to replace it with a suitable value, affecting the computation of the distribution parameters (i.e., mean, variance) of individual geochemical data distributions. However, if the distribution is assumed to be normal, or can be transformed to be, then the replacement value of the censored data and parameters of the distribution (mean, variance) can be estimated from the portion of the distribution that is not censored. A better estimate of the mean and variance of the distribution can therefore be determined and ultimately, this assists in the identification of atypical (anomalous) values, particularly in multivariate applications (Grunsky 1995; Sanford et al. 1993).

The combined dataset contains varying levels of censorship, identified by negative and zero values. Assignment of an arbitrary replacement value such as 3/4 of the detection limit is not practical given the uncertainty in the origin, method of collection and accuracy of the analytical procedures. Concerns about the accuracy of replacing censored values with simple fractions (i.e. 1/2 the detection limit) have also been raised (Sanford et al. 1993). Calculation of a more suitable replacement value, using the method of Sanford et al. (1993), allows a better estimate of population parameters to be made. However, given the large sample population, mixed lithologic sources and the generally low level of censoring, replacement values are unlikely to affect the final statistical and spatial geochemical interpretation. An example is shown in Figure 21 for Na_2O data, in which a small, censored population has been replaced by 0. Visual comparison of Q-plots (cumulative density function plots) reveals little difference in the shape of the population and no difference in the shape at the high end of the distribution of values. Breakpoints in the distribution of Na_2O values are identical (see arrows, Figure 21), and anomalous high value preliminary thresholds are easily identified on both plots at 7.5 wt. %. Any break in the Q-plot at the censored data range is easily interpreted, and unlikely to mask any other breakpoint. Of more concern, is the stair-step geometry of the breakpoints caused by the restriction of TG dataset values to integer values only. This may be obscuring more subtle breakpoints due to geochemical processes. Removal of the TG data and re-inspection of the Q-plots indicates the stair-step breakpoints are eliminated (Figure 22). Preliminary thresholds for anomalous data possibly related to mineralization were the same for the censored and uncensored datasets at <1.1 wt.% Na_2O (possible Na depletion?) and >7.5 wt.% Na_2O (possible Na metasomatism?). Thus for subsequent analysis using these data, a replacement value of 0 will be used, or the data eliminated if a log-transformation was required to normalize the distribution. However, censored data can be considered important when the characterization of element variation with respect to geochemical processes is being sought. This is particularly important in multi-element geochemical studies where the presence of a censored value may exist for one element, while another element may be "significant" from an exploration point of view.

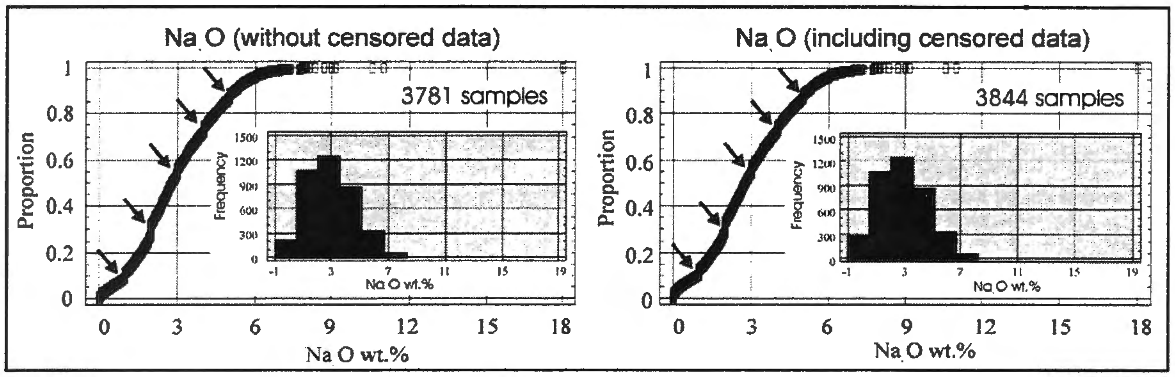


Figure 21:Q-plots, and histograms showing distribution of Na_2O populations without censored data (left), and with censored data replaced by a value of 0 (right). Arrows indicate apparent breaks in population caused by integer TG values of 1, 2, 3, 4, and 5.

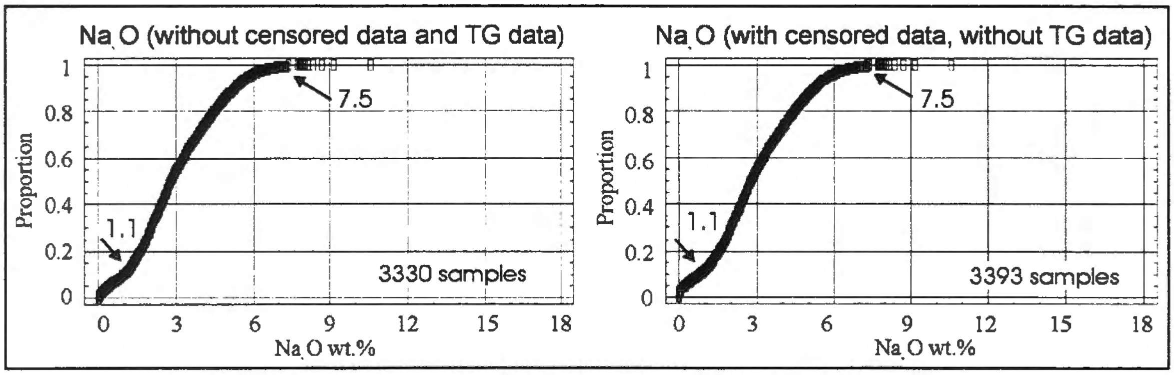


Figure 22: Q-plots, and histograms showing distribution of Na_2O populations without the integer TG dataset, for uncensored data (left), and censored data (right). Breaks are indicated by arrows on the Q-plots.

4.4 Closure

The problem of closure represents an additional type 2 problem that should be considered when analyzing the combined lithogeochemical dataset. Compositional data such as weight percentages reported for oxides sum to 100%. From a statistical point of view this standard form of representing geochemical analyses by a constant sum can result in statistical inconsistencies and spurious correlations. Basically, variables in a closed number system such as data expressed as percentages are not free to vary independently, as a change in the value of any one oxide by definition must affect the value of the rest in order for the total to remain at 100%. The problem of closure was initially noted by Pearson (1897), but has been largely ignored by the geoscience community. Chayes (1960, 1966, 1970) and Chayes and Kruskal (1966) attempted to deal with the problem with limited success for specific applications and only recently Aitchison (1986) has developed a method for dealing with the problem of closure of compositional data based on log-ratio transformations. Rollinson (1993) provides a detailed summary of the closure problem.

The application of statistical procedures on closed data has a number of implications; one, false correlations may be induced between elements, and two, subpopulations (such as using Al_2O_3, MgO and FeO+TiO_2 to plot mafic metavolcanic variability on an AFM diagram) may not reflect chemical relationships within the complete dataset (Rollinson 1993). Meisch (1976) has suggested that the effects of closure are not as significant when both variables being compared occur in only minor amounts but that covariance and hence correlation coefficients may not reflect geochemical relationships in oxides occurring in major abundances. More recently, Madiesky and Stanley (1993) have used closed data successfully in exploration applications by applying Pearce element ratios (Pearce 1968). This approach can be used when a particular cation is known to be conserved in a particular magmatic fractionation sequence. However, such an approach cannot be currently applied to regional geochemical data since there is no one cation that is likely to be conserved. Multiple geochemical processes must be modeled effectively to interpret the data. This can be time consuming and difficult.

Aitchison (1986) overcomes the problem of closure by projecting the closed data to real number space using log-ratio transformations. This method involves dividing each major oxide element by a common divisor (i.e., TiO_2 or P_2O_5) and then performing the logarithmic transformation on the resulting ratio value. Results can be presented in a covariance matrix in which each ratio is compared with every other ratio (see Table 2.5 - Rollinson 1993). Problems with this method include the loss of one element from consideration (i.e., the oxide which is used for the common divisor), and the lack of spatial information this yields. Spatial information is lost because each sample undergoes a unique transformation depending on its original oxide values. For example, a continuous surface map of MgO values may indicate an anomalously high area. However, with transformation, this cluster could be lost depending on the divisor used, and the variability in the value of the divisor amongst the clustered samples.

4.5 Non-Normal Distributions

Typically, many trace element distributions do not display normal distribution characteristics. Therefore, it is often necessary to apply transformations to the data to produce normal distributions. Prior to applying any type of transformation it is necessary to evaluate the nature of the distribution. In many cases, a long-tailed distribution that may have the appearance of a log-normal distribution, may in fact be a mixture of two or more distributions. In such cases, it is better to separate the populations. This can be done on the basis of ancillary information such as lithology or it can be done using a numerical approach (Miesch 1981; Stanley and Sinclair 1987).

If populations cannot be distinguished from a non-normal distribution then a number of transformations are available in which the data can be "normalized". A commonly applied transformation is the logarithmic (monotonic) transformation which is particularly useful for positively skewed data, as it tends to homogenize variance (Miesch 1976) and enhance background trace element associations (Howarth and Earle 1979). This is often applied when estimates of means are required and to trace element data that are almost always log normally distributed (Miesch 1976). However, a number of workers have suggested the log-transform may in fact not be better than non-transformed data (Link and Koch 1975; Miesch 1976; Howarth and Earle 1979), and that it may enhance negative skewness and reduce kurtosis, particularly for already negatively skewed data (Howarth and Earle 1979). An alternate approach is the use of the generalized Box-Cox power transform (Smith et al. 1984; Grunsky et al. 1992). The application of Box-Cox power transformations and the use of log-ratios for compositional data have been studied by Barcelo et al. (1995) and Grunsky et al. (in prep). However, prior to applying any type of transformation, atypical values must be removed from the distribution. The use of robust statistical methods are best employed prior to transforming the data (c.f. Campbell 1980; Garrett 1989a,b), to better define background and accentuate atypical samples.

5.0 Summary and Conclusions

A large lithogeochemical dataset comprising approximately 4500 sample points, after screening procedures, has been compiled from 5 separate lithogeochemical datasets using relational database software and GIS software tools. The data have been standardized with respect to missing values in that all "holes" in the data have been filled.

No studies regarding analytical variability have been undertaken with this dataset, as each dataset was received long past the time of original collection and analysis. It is assumed in this study that quality control procedures were conducted by the proprietor of each individual dataset.

Obvious data compatibility problems have been identified using a variety of comparison techniques. This screening process has resulted in the deletion of approximately 300 samples, principally from the Texas Gulf dataset. The dataset is incomplete in some respects as various oxide elements are missing from, once again, the Texas Gulf dataset. One might conclude the given the plethora of problems discovered with the Texas Gulf dataset, the entire dataset should be discarded. However, we have tried to identify the most obvious problems, and have attempted to normalize the dataset with respect to the other datasets. Therefore, we have left the remaining Texas Gulf data in the compiled dataset. However, the user should be aware of the potential problems with this dataset, and treat the data accordingly, especially if rigorous statistical analysis of the data is to be attempted.

Several methods for calculating replacement values for censored data exist, permitting the calculation of statistical parameters assuming that the data are normally distributed. This results in better estimates of correlation and covariance when multi-element data are evaluated. However, the effects of censored data are minimal on determining population parameters of a large dataset composed of variable lithologies, although the statistical effects are augmented as the size of the censored population increases. Detailed examination of CDF-plots for the determination of anomalous thresholds minimizes the spatial effects of using censored data, and the possibility of the censored data obscuring important breaks in the population.

Non-normally distributed data is often the result of mixtures of two or more populations. Prior to applying transformations to the data, the preferred approach is to separate the populations which aids an appropriate choice of breakpoint or geochemical threshold value, and thus the spatial distribution of anomalous areas. A comparison of maps and distributions for both the non-transformed and transformed data help distinguish different geochemical processes.

We have attempted to illustrate the potential problems and pitfalls in compiling a large geochemical database derived from a number of disparate sources. The relational database and GIS software tools were integral to this exercise as they provided a method for quickly and efficiently assembling, visualizing and comparing the data, both statistically and spatially. When statistically manipulating geochemical data, the effects on the spatial distribution of the data should be closely monitored for it is the spatial patterns that result from data analysis which are the most important for exploration purposes. Therefore, a distinction between the statistical and spatial properties of geochemical dataset should always be considered when manipulating and analyzing the data.

Finally, many problems involved with the compilation of a lithogeochemical dataset can be avoided with proper data collection, recording and documentation. All samples should be properly located, and detection limits recorded at the time of data archival. In addition, the lineage of the data, which should include results of quality control analysis, can be archived within the GIS database for future reference.

Acknowledgments

The authors wish to thank Dr. R.G. Garrett and Dr. G. Bonham-Carter of the Geological Survey of Canada for helpful and constructive comments.

Bibliography

Aitchison, J., 1986. The statistical analysis of compositional data. Methuen Inc., New York.

Ayer, J.A. and Theriault, R., 1993. Geology of Keith and Muskego Townships, Northern Swayze Greenstone Belt. In: N. Wood, R. Shannon, L. Owsiacki and M. Walters (Editors), 1992-1993 NODA Summary Report, pp. 26-33.

Barcelo, C., Pawlowsky, V., and Grunsky, E.C., 1995. Classification Problems of Samples of Finite Mixtures of Compositions. Mathematical Geology, 27(1): 129-148.

Bonham-Carter, G.F., 1994. Geographic Information Systems for Geoscientists: Modelling with GIS. Pergamon (Elsevier Science Ltd.), New York.

Bonham-Carter, G.F., Agterberg, F.P. and Wright, D.F., 1988. Integration of geological datasets for gold exploration in Nova Scotia. Photogrammetric Engineering and Remote Sensing, 54(11): 1585-1592.

Campbell, N.A., 1980. Robust procedures in multivariate analysis. I . Robust covariance estimation. Applied Statistics, 29: 231-237.

Chayes, F., 1960. On correlation between variables of constant sum. Journal of Geophysical Research, 65: 4185-4193.

Chayes, F. , 1966. Alkaline and subalkaline basalts. American Journal of Science, 264: 128-145.

Chayes, F., 1970. Ratio Correlation. University of Chicago Press, Chicago.

Chayes, F. and Kruskal, W., 1966. An approximate statistical test for correlation between proportions. Journal of Geology, 74: 692-702.

Chung, C.F., 1985. Statistical treatment of geochemical data with observations below the detection limit. In: Current Research Part B, Geological Survey of Canada, Paper 85-1B, pp. 141-150.

Chung, C.F., 1988. Statistical analysis of truncated data in geosciences. Science de la Terre. Informatique Geologique 27(1): 157-180.

Darnley, A.G., Bjorklund, A., Bolviken, B., Gustavsson, N., Koval, P.V., Plant, J.A., Steenfelt, A., Tauchid, M. and Xie, X., 1995. A Global Geochemical Database for Environmental and Resource Management, Earth Sciences 19, UNESCO Publishing.

Fletcher, K., 1981. Analytical Methods in Geochemical Prospecting. In: Handbook of Exploration Geochemistry Vol. 1. Edited by G.J.S. Govett. Elsevier, New York.

Garrett, R.G., 1989a. The chi-square plot: a tool for multivariate outlier recognition. Journal of Geochemical Exploration, 32: 319-341.

Garrett, R.G., 1989b. A Robust Multivariate Allocation Procedure with Applications to Geochemical Data. In: F.P. Agterberg and G.F. Bonham-Carter (Editors), Statistical Applications in the Earth Sciences. Geological of Survey of Canada, Paper 89-9, pp. 309-318.

Grunsky, E.C., 1995. Numerical Methods, Techniques and Strategies for the Evaluation and Interpretation of Geochemical Data, Short Course Notes, Curtin University of Technology and Cooperative Research Centre for Australian Mineral Exploration Technologies, April 10-11, 1995, 109 pages, 1 diskette.

Grunsky, E.C, Easton, R.M., Thurston, P.C. and Jensen, L.S., 1992. Characterization and Statistical Classification of Archean Volcanic Rocks of the Superior Province using Major Element Geochemistry in Geology of Ontario. Ontario Geological Survey, Special Volume 4, Part 2, pp. 1347-1438.

Grunsky, E.C., Barcelo, C., and Pawlowsky, V., in prep. Statistcally Based Classification on the Simplex: An example using lithogeochemistry.

Harris, J.R., 1989. Data Integration for Gold Exploration in Eastern Nova Scotia Using a GIS. Proceedings of the 7th Thematic Conference on Remote Sensing for Exploration Geology, Calgary, Alberta, pp. 233-249.

Harris, J.R., Broome, J. and Heather, K.B., 1994. Swayze Greenstone Belt GIS Project. In: N. Wood, R. Shannon, L. Owsiacki and M. Walters (Editors), 1993-1994 NODA Summary Report, pp. 115-121.

Harris, J.R., Wilkinson, L. Broome, J. and Fumerton, S., 1995a. The GIS/Database project: Progress in the Swayze Greenstone Belt. In: L. Owsiacki, M. Walters, H. Brown, R. Shannon (Editors), 1994-1995 NODA Summary Report, pp. 110-116.

Harris, J.R., Wilkinson, L. and Broome, J., 1995b. Mineral Exploration Using GIS-Based Favourability Analysis, Swayze Greenstone Belt, Northern Ontario. In: Proceedings of the 7th International Conference on Geomatics, Ottawa, ON. (CD-ROM).

Heather, K.B. and van Breeman, O., 1994. An interium report on geological, structural and geochronological investigations of granitoid rocks in the vicinity of the Swayze greenstone belt, southern Superior Province, Ontario. In: Current Research 1994-C, Geological Survey of Canada, pp. 259-268.

Heather, K.B, Shore, T.G. and van Breeman, O., 1995. The Convoluted "Layer-cake": An old Recipe with New Ingredients for the Swayze Greenstone Belt, Southern Superior Province, Ontario. In: L. Owsiacki, M. Walters, H. Brown and R. Shannon (Editors), 1994-1995 NODA Summary Report, pp. 94-103.

Howarth, R.J. and Earle, S.A.M., 1979. Application of a Generalized Power Transformation to Geochemical Data. Mathematical Geology, 11(1): 45-62.

Jensen, L.S., 1976. A new cation plot for classifying subalkalic volcanic rocks. Ontario Department of Mines Miscellaneous Paper 66.

Link, R.F., and Koch G.S., 1975. Some Consequences of Applying Lognormal Theory to Pseudolognormal Distributions. Mathematical Geology, 7(2): 117-128.

Madiesky, H.E., and Stanley, C.R., 1993. Lithogeochemical Exploration of Metasomatic Zones Associated with Volcanic-Hosted Massive Sulphide Deposits Using Pearce Element Ratios. International Geology Review, 35: 1121-1148.

Miesch, A., 1976. Geochemical survey of Missouri - Methods of sampling, laboratory analysis and statistical reduction of data. U.S. Geological Survey, Professional Paper 954-A.

Miesch, A.T., 1981. Estimation of the Geochemical Threshold and it Statistical Significance. Journal of Exploration Geochemistry, 16: 49-76.

Pearce, T.H., 1968. A contribution to the theory of variation diagrams. Contributions to Mineralogy and Petrology, 19: 142-157.

Pearson, K., 1897. Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society, 60: 489-498.

Rencz, A.N., Harris, J.R., Watson, G.P., and Murphy, B., 1994. Data Integration for Mineral Exploration in the Antigonish Highlands, Nova Scotia: Application of GIS and Remote Sensing. Canadian Journal of Remote Sensing, 20(3): 257-267.

Rollinson, H.R., 1993. Using Geochemical Data: Evaluation, Presentation, Interpretation. Longman Scientific and Technical, London.

Rose, A.W, Hawkes, H.E. and Webb, J.S., 1979. Geochemistry in mineral exploration. 2nd. Edition. Academic Press, London.

Sanford, R., Pierson, C. and Crovelli, R., 1993. An Objective Replacement Method for Censored Geochemical Data. Mathematical Geology, 25(1): 59-80.

Smith, R.E., Campbell, N.A., and Litchfield, R., 1984. Multivariate Statistical Techniques Applied To Pisolitic Laterite Geochemistry At Golden Grove, Western Australia. Journal of Geochemical Exploration, 22: 193-216.

Stanley, C.R. and Sinclair, A.J., 1987. Anomaly recognition for multi-element geochemical data; a background characterization approach. Journal of Geochemical Exploration 29: 333-353.

Thompson, M., 1983. Control Procedures in Geochemical Analysis. In: R.J. Howarth (editor), Handbook of Exploration Geochemistry, Vol. 2, Statistics and Data Analysis in Geochemical Prospecting. Elsevier, New York, pp. 39-58.

Wright, D.F. and Bonham-Carter, G.F., 1996. VHMS favourability mapping with GIS-based integration models, Chisel Lake - Anderson Lake area. In: by G.F. Bonham-Carder, A.G. Gally, G.E.M. Hall (Editors), EXTECH 1: A Multidisciplinary Approach to Massive Sulphide Research in Rusty Lake - Snow Lake Greenstone Belts, Manitoba. Geological Survey of Canada, Bulletin 426, pp. 339-401.

Appendix A

Many of the assumptions necessary to use a lithogeochemical database compiled from a variety of sources have been dealt with in the body of the paper under the appropriate sections and are not repeated here. This appendix represents a summary of assumptions that were necessarily a priori to the commencement of this research.

1) Volatiles:

Geochemical data compiled for this study came from 6 sources, with little supporting documentation. Volatiles are recorded variably as H_2O, H_2O^+, H_2O^- or LOI. and it must be assumed that these are entered correctly i.e. that H_2O represents total H_2O etc.

2) Missing values for certain oxides:

In cases where oxide values are missing it is assumed that the remaining values have not been recalculated to 100%. The values that are recorded are treated as the original values. This was checked (see "Lack of total oxide data" section).

3) Iron:

Iron is also variably recorded in the component sources, between FeO^T, $Fe_2O_3^T$, and FeO and Fe_2O_3. These must be adjusted in a consistent manner.

4) Locations:

It must be assumed that locations for each sample have been accurately determined and faithfully recorded in a common projection and datum.

5) Sampling

It must also be assumed that all samples represent surface chips rather than from drill-hole, and that some care has been to taken in the selection of representative sample from the least altered areas.