Sampling Size Determination: Application In Geochemical Sampling For Environmental Impact Assessment

Meng Zhou
Southwest Petroleum University https://orcid.org/0000-0002-9753-760X
Elizabeth Chihobve
University of South Africa
Baojin Zhao (baojin_zhao@swpu.edu.cn)
University of South Africa https://orcid.org/0000-0003-4991-1246
Zhen Song
Southwest Petroleum University

Research Article

Keywords: Sample Size Estimation, Central Limited Theory, Environmental Risk Assessment, Geochemical Sampling, Uncertainty.

Posted Date: February 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1234842/v1

Sampling Size Determination: Application in Geochemical sampling for Environmental Impact Assessment

Meng Zhou[1], Elizabeth Chihobve[2], Baojin Zhao[2, 3, 4*], Zhen Song[5,6,7]

1. School of Sciences, Southwest Petroleum University, 8 Xindu Avenue, Chengdu 610500, China.
2. Department of Environmental Sciences, University of South Africa, 28 Pioneer Ave, Florida Park, Roodepoort, 1709, South Africa.
3. School of Geoscience and Technology, Southwest Petroleum University, 8 Xindu Avenue, Chengdu 610500, China.
4. Overseas Expertise Center for Deep Marine Shale Gas Efficient Development Innovation (111Center), Southwest Petroleum University, 8 Xindu Avenue, Chengdu 610500, China.
5. School of Mechanical Engineering, Southwest Petroleum University, 8 Xindu Avenue, Chengdu 610500, China.
6. Sichuan Science and Technology Resource Sharing Service Platform of Oil and Gas Equipment Technology, Southwest Petroleum University, 8 Xindu Avenue, Chengdu 610500, China.
7. Central South University, Changsha 410083, China
*Corresponding author, baojin_zhao@swpu.edu.cn

Abstract

Quantification of the uncertainties associated with environmental geochemical prediction as a function of sample size is always a concern when performing specifically an Environmental Risk Assessment (ERA). While the study of determination of sample size in geochemical field is limited, there is a definite need for the development of a methodology together with a protocol which addresses geochemical sampling (sample sizing and representative) analyses in geochemical predictions. Based on Central Limited Theory, this article proposed a statistical methodology for determining sample size and applied this methodology using the Vaal River and West Wits tailing dams in South African as a case study. It also discussed the factors such as confidence interval, acceptable sampling error etc. that could influence sample size estimation and recommended a tradeoff strategy to reduce the sample size for economic reason.

Key Words: Sample Size Estimation; Central Limited Theory; Environmental Risk Assessment; Geochemical Sampling; Uncertainty.

1. Introduction

An accurate and good environmental impact assessment (EIA) program in geochemical sampling depend on two major factors: 1) The collection of representative samples; 2) How many samples taken. All the above will be guided by sampling and analytical program. In terms of sampling size specifically, most of the time, it is clearly lacking of guideline to determine the sampling size, especially in a mathematic al and statistical manner.

The accuracy of techniques employed during the environmental impact assessment in mining environment depends on the accuracy and precision of the sampling program, and sample size is a key component in a sampling program. It is therefore important that a good sample size determination technique be put in place. However, there has been limited work done on the development of methodologies for determination of sample size in geochemical sampling of EIA. Some examples of good practices can now be found (Mason, 1992; Smith *et al.*,2004; Dominy

and Petersen, 2005), but, the application of correct procedures is far from universal. This, in turn, may result in lack of confidence in the EIA, specifically for the accuracy and precision of the long term water quality prediction and even problems in obtaining approval of precision mining authorizations, water use licenses and mine closure plans. Figure1 illustrates the various stages involved in geochemical sampling and analysis.

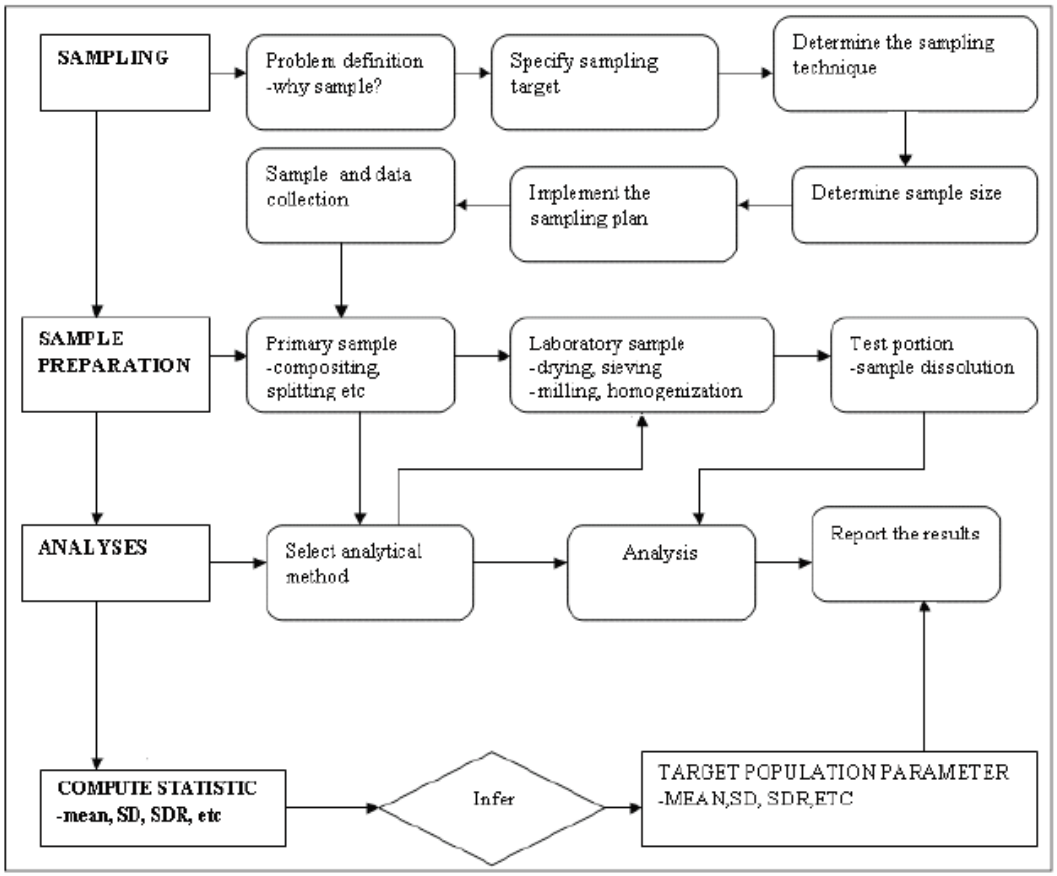

Figure 1: The measurement process in geochemical sampling and analysis

This article addresses this deficiency in sampling by proposing a statistical methodology for determining the optimum sample size for ensuring data quality in the EIA in the mining environment.

The statistical analysis approach was adopted as the best method for sample size determination, which is based on the premise that “the size of the study sample is critical to producing meaningful results”.

The size of the required samples depends on a number of factors including purpose of the study, available budget, variability of the population being sampled, acceptable error and required confidence level.

Findings from this study show that the number of samples needed is influenced by the purpose of the study, size of the target area, nature and type of material, budget, tolerable sampling error and the confidence level required, among other factors. Acceptable sampling error has an inverse relationship with the sample size, which confidence level and standard deviation have a positive correlation with the sample size.

One can minimize error by increasing sample size. While a low value of acceptable sampling error value and high confidence are always desirable, a trade-off among these competing

factors must be found, given the fact that funds and time are normally limited.

2. Methodology

Sample size determination is one of the most crucial and difficult steps in the geochemical investigation process. Studies of geological materials have revealed that variability exists in terms of mineralization, chemical composition and alteration. Under such conditions, it is essential that an adequate sample size accurately characterizes the variability be collected.

In order to successfully determine optimum sample size needed to satisfy the objectives of a study, several factors need to be considered. These include: the purpose of the investigation, type of material to be sampled, cost of sampling, variability of the material and available site information. Additional considerations are the required level of acceptable error, confidence interval and standard deviation. These factors and their effects are explored in detail in the following sections.

2.1 Statistical method and principle

If a sample size n with the values (X_1, X_2,…,X_n) is drawn from the same population that has a mean "μ" and a standard deviation "σ", then the sample mean "$\overline{X}$" is given by

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1}$$

Where $\overline{X}$ (sample mean) is assumed to present the best estimate of μ, and therefore $\mu = \overline{X}$ (Petri Koistinen 2010).

According to the Central Limit Theory, if the if the sample size is reasonably large (usually greater than 30), the sample mean will tend to be normally distributed. And, we have

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu \tag{2}$$

$$D(\bar{X}) = D(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{\sigma^2}{n} \tag{3}$$

where $E(\bar{X})$ is the expectation of the sample mean $(\bar{X})$, and $D(\bar{X})$ is the variance of the sample mean $(\bar{X})$. So that we could write:

$$\bar{X} : \ N(\mu, \frac{\sigma^2}{n}) \tag{4}$$

which can be further standardized to:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} : \ N(0,1) \tag{5}$$

The difference between μ and $\bar{x}$ is sampling error (d), which is defined as:

$$d = \bar{X} - \mu \tag{6}$$

Hence, we can write:

$$\frac{d}{\sigma / \sqrt{n}} \sim N(0,1) \tag{7}$$

With a given confidence level ($1-\alpha$), the sampling error (d) equals:

$$d = Z_{\alpha/2} \times (\sigma / \sqrt{n}) \tag{8}$$

Where $Z_{\alpha/2}$ represents Z value which is the function of normal distribution related to confidence levels.

Table 1.1: Z values for different confidence levels

Confidence levels	75%	85%	95%	99%
Z values	1.15	1.44	1.96	2.57

From the Equation (8), it can be seen that for a given confidence level, the estimated sampling error is dictated by the standard deviation of the population and the sample size. On the other hand, the wider the population spreads, the larger the estimated sampling error is, while the bigger the sample size, the smaller the estimated sampling error is. For a given confidence level, the estimated sampling error is determined by the sample size. If there is a desire to reduce the error, sample size should be increased. However, increasing sample numbers means spending more money, i.e., increasing the cost, time and many more involvement. So a balance being competing factors should be thought. Rearranging Equation (8), we get the following formula:

$$n = (\frac{Z_{\alpha/2} \times \sigma}{d})^2 \tag{9}$$

This is the basic model for determination of a sample size. Usually the population standard deviation "σ" is unknown, but which can be estimated by taking the sample standard deviation "S" as its unbiased estimator, and the "S" is defined as:

$$S = \sqrt{\sum_{i=1}^{n} (X_i - \overline{X}) 2 / (n-1)} \tag{10}$$

So that equation (9) becomes

$$n = (\frac{Z_{\alpha/2} \times S}{d})^2 \tag{11}$$

This formula is a more convenient way for estimating the sample size for continuous data.

2.2 Parameters that affect sample size

In order to illustrate the factors that influence sample size determination, randomly chosen pH results from the West Wits tailings dams were used as shown in Table 2.1.

Table 2.1: pH data (Pulles *et al.,* 2002)

9.00	8.80	8.10	7.60	8.20
9.00	8.30	8.40	8.10	8.10
8.10	8.40	8.70	8.50	8.20
8.30	8.30	8.60	8.20	8.30
8.20	8.30	8.50	8.20	8.10
8.30	7.90	7.60	8.40	8.50
8.30	8.40	7.50	8.00	8.40
8.30	8.40	8.40	8.50	8.60
7.80	8.30	8.20	7.70	8.90
8.20	8.30	8.10	8.50	8.70
8.20	8.20	8.60	8.80	8.40
8.20	8.40	8.50	8.00	8.50
8.10	8.30	8.50	8.60	8.70
8.00	8.30	8.40	8.70	8.60
8.20	8.30	8.40	8.60	9.00
8.30	8.20	8.20	8.70	8.60
7.50	8.50	7.80	8.40	8.80
7.70	8.50	8.20	8.20	8.60
8.20	8.40	8.30	8.30	8.60
8.40	8.40	7.50	8.40	8.60
8.20	8.30	7.90	8.50	8.10
8.00	8.50	7.90	8.30	9.10
7.60	8.70	8.20	8.50	8.50
8.10	8.50	8.10	8.50	8.50
8.60	8.50	8.30	8.60	8.40
8.70	8.60	8.30	8.50	8.70
8.40	8.40	8.30	8.20	8.80
8.20	8.50	8.30	8.10	8.40
8.50	8.50	8.40	8.20	8.40
8.60	8.60	8.40	8.20	8.70
8.30	8.40	8.40	8.30	8.50
8.20	8.50	8.20	8.00	8.40
8.90	8.50	8.10	8.30	8.30
8.70	8.40	8.30	8.60	8.10
8.50	8.10	8.70	8.60	8.80
8.80	8.40	8.90	8.30	8.50
8.30	8.50	7.20	7.80	8.70
8.50	8.60	8.60	7.70	8.50
8.00	8.00	7.90	8.20	8.20
8.60	8.30	8.20	8.60	8.40
Parameter	Sample numbers	Sample Mean	Sample Variance	Sample Standard deviation
Value	200	8.35	0.090	0.302

2.2.1 Confidence level

Statistically, confidence level describe the probability that the true parameter falling into a specific interval. Based on equation (11), an increase in required confidence level means more samples will be required. This relationship can be illustrated using the dataset from Table 2.1.

The mean ($\overline{X}$) of the data set is 8.35 that is the estimation of the population mean (true mean).

If it is necessary to know with 75% confidence level and the sampling error of 0.022 that 75 out of 100 samples are within the true population mean of 8.35±0.022, sample size (n) required is:

$$n = (\frac{1.152 \times 0.302}{0.022})^2 \approx 250 \quad (12)$$

Therefore, the required sample size is 250.

But, if the confidence level is increased to 95%, the required sample size is:

$$n = (\frac{1.962 \times 0.302}{0.022})^2 \approx 725 \quad (13)$$

The calculated results show that if it is necessary to increase the confidence level, it becomes necessary to increase the sample numbers. However, the relationship is not linear, as doubling the sample size does not halve the confidence interval (Israel, 1992). This relationship is shown in Table 2.2.

Table 2.2: Sample numbers for different confidence levels

Confidence levels	75%	80%	85%	90%	95%	99%
Sample numbers	250	309	391	507	725	1245

2.2.2 Sampling Standard Deviation

Sampling Standard deviation defines how the population is spread. This parameter, like sampling error and confidence level also affects sample size. Essentially the wider the population spreads, the larger the number of samples needed. The relationship can be demonstrated by the example below. The sample standard deviation of the dataset in Table 1.2:

S=0.302

Given a 95% confidence and the standard error of 0.02, the sample size for dataset (Table 4.3) can be calculated as follows:

$$n = (\frac{1.962 \times 0.302}{0.022})^2 \approx 725 \quad (14)$$

If the sample standard deviation is increased to 0.40 with 95% confidence level and sampling error of 0.02, the required sample size is:

$$n = (\frac{1.962 \times 0.40}{0.022})^2 \approx 1270 \quad (15)$$

From the results, it is clear that the sampling standard deviation of the sample affects sample size, with an increased standard deviation requiring more samples to obtain the same confidence levels.

2.2.3 Sampling Error

Equation (9) shows that a relationship exists between sample size and the sampling error **(** d **)**. This relationship is exponential and an increase in sample size is followed by a decrease in sampling error. The data in Table 4.5 illustrates this relationship. Using the data values in Table 4.3, s =0.302, C.I = 95% and sampling errors are 0.01, 0.022, 0.03, 0.04, 0.05; then the required sample size n is:

Table 2.3: Sample numbers for different sampling error

Sampling error	0.01	0.022	0.03	0.04	0.05
Sample numbers	3 511	725	390	220	140

The results above indicate that a small sampling error value is associated with a large sample size. In addition, the formula shows that to halve the error, the sample size needs to be quadrupled.

2.2.4 Cost of taking the samples

Financial issues assist in determining how precise the estimates should be. If the decision to be made will incur significant financial outlay, then more samples should be collected in order to have increased confidence in the decision. However, a large sample size means a higher sampling budget and this needs to be balanced against the financial consequences of making a wrong decision based on data with lower confidence levels.

3. Procedure for determining sample size

A step-wise procedure for determining sample size is shown in Figure 2 below.

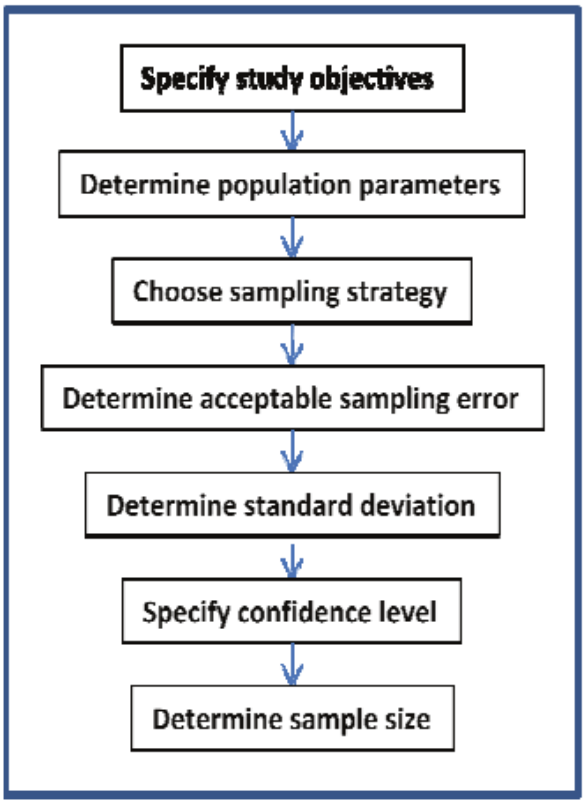

Figure 2: Procedure for determining sample size

3.1 Specify the sampling objectives

In order to correctly determine sample size, it is essential to clarify the objectives of the investigation. One should ask the question “what do I want to know?” and “what information do I expect to gain?” Only when these questions are addressed can one decide on the number of samples to collect. Study objectives vary and hence sample sizes vary as well. In a geochemical investigation, the nature of the questions that need to be answered will determine the type of assessment tool to be used and consequently the data requirements for that assessment tool. Generally speaking, geochemical assessments undertaken in support of mine closure applications will require the highest level of confidence in the results and will require the more extensive sampling programs. It is important to discuss and agree on acceptable confidence levels with key stakeholders prior to designing the sampling program in order to ensure that sufficient samples

are taken to support the decisions to be made.

Similarly, decisions relating to placement and design of waste material facilities such as waste rock dumps, tailings disposal facilities, in-pit disposal systems, etc. require a very high level of confidence as decisions relating to these facilities cannot easily be changed at a later date.

3.2 Determine the population parameters

Determination of the population parameters is the next step after definition of the study objectives. In this process the following questions should be investigated:

- Is the population homogeneous?
- What are the statistical properties of the sample target in terms of mean, standard deviation, median, etc.?
- What is the spatial distribution and physical size of the sample target?

The critical parameters such as mean, standard deviation, acceptable error can potentially be acquired from various sources of information such as:

- Previous publications and documents;
- Similar previous studies;
- Pilot sampling study of the sample target.

3.3 Choose the sampling strategy

Once the above information has been gathered, the next step is to select a sampling strategy. Care must be taken when choosing the sampling strategy. The analyst should take into account such aspects as spatial continuity, correlation of the geological properties, cost and variability. If the sample material is homogeneous, a random sampling strategy may be adopted. If the sample area is heterogeneous, stratified or systematic sampling strategies are preferable. With a stratified sampling strategy, the area is divided into several subpopulations and random samples are collected from each stratum and the sample size of each subpopulation is calculated separately. In addition, stratified design is more economical in that the number of samples is reduced since the target area is divided into different stratum with each stratum having similar properties. The advantage of the systematic grid method is that it takes spatial continuity into account (US EPA, 1999).

3.4 Determination of the acceptable sampling error

In order to calculate the sample size, it is necessary to define the study sampling error **(** d **),** which defines the difference between the true population mean and the sample mean. This is the acceptable error that the stakeholders are willing to take. The following guidelines can be used to decide on the specific value of the sampling error (Yang, 2006):

- The value should be greater than the analysis error
- The value can be calculated as 5% of the sample mean
- The value can be taken as the maximum allowable value for the ERA purpose.

3.5 Determination of standard deviation

Standard deviation is a critical component of the sample size formula. Although the analyst has less control over variance, it must be incorporated into the design. Methods that can be used to estimate the standard deviation value as an input for sample size determination, include use of a pilot study, use of data from previous studies of similar populations, estimates assisted by some logical mathematical results (Bartlett et al., 2001).

3.6 Specifying confidence level

Confidence interval provides a range of values that has a specified probability of containing the

parameter being estimated. The frequently used and recommended confidence level is 95% and its alpha value is 0.01 (Yang and Zhao, 2005). The z value associated with this confidence level is 1.96. This value can be obtained from standard normal distribution tables. If a 95% confidence level has been chosen it simply means that it is believed that the population mean lies within this range 95% of the time.

One major factor affecting the choice of confidence level is the purpose of the study or investigation. Higher confidence levels for example 99% confidence level are often employed in cases where decisions based on the assessment are critical and errors may cause substantial harm (Bartlett, Kotrlik, and Haggins, 2001). Lower confidence levels for example 75% may be used for instance when identifying trivial differences or other statistical phenomena as an antecedent to further studies.

3.7 Determination of sample size

When the various parameters necessary to determine sample size have been estimated, the parameters are combined using the formula (equation (9)) below:

$$n = (\frac{Z_{\alpha/2} \times \sigma}{d})^2 \quad (9)$$

There are two types of sample sizes to be decided on (Yang, 2006):

Minimum sample size: When we use the formula to calculate the sample size we assume that the samples are normally distributed, although based on the Central Limit Theorem, the assumption is correct only when the sample numbers are reasonably large. In statistics, 30-50 samples can be taken as reasonably large. Considering the costs of the sampling 30 is considered as a reasonable minimum sample size.

Maximum sample size: The sample size calculated based on the formula can be taken as the maximum sample size because it assumes all samples as random and independent and gives each sample an equal weight.

3.8 Limitations

- One disadvantage associated with the method is that a "good" estimate of the population standard deviation is necessary. Often this value is unknown and the unbiased estimator “S” is used instead.
- The method assumes that a simple random sampling scheme is the method of choice and does not apply when other sampling schemes are chosen.
- The method assumes that the sample size is sufficiently large.

4. Application results

This section serves to illustrate how the proposed methodology can be applied. It also serves to assess if the number of samples collected during the case study were enough for ERA. As such, ABA results for Vaal River dam 3 and West Wits dam N will be used to demonstrate the applicability of the method. The data for the two dams is presented in Table 4.1. Key parameters for acid rock drainage (ARD) study are the acid potential (AP) and the neutralizing potential (NP). These parameters are used to illustrate how the proper sample size can be determined.

Table 4.1: ABA data from the Wits Basin (Pulles *et al.*, 2003)

Location	Sample ID	Total S (wt %)	Calcite (kg/t)	AP (kg/t)	NP (kg/t)
Vaal River	VRD3 A1	0.130	8.75	4.06	8.75
	VRD3 A2	0.277	0.25	8.66	0.25
	VRD3 A3	0.326	0.00	10.19	0.00
	VRD3 A4	0.310	0.75	9.69	0.75
	VRD3 A5	0.207	2.00	6.47	2.00
	VRD3 A6	0.320	0.25	10.00	0.25
	VRD3 P1	0.070	2.00	2.19	2.00
	VRD3 P2	0.125	0.00	3.91	0.00
	VRD3 P3	0.283	0.00	8.84	0.00
	VRD3 P4	0.127	0.75	3.97	0.75
	VRD3 P7	0.188	0.00	5.88	0.00
	VRD3 P10	0.247	0.00	7.72	0.00
West Wits	N-1S	0.789	2.75	24.66	2.75
	N-2S	0.656	7.50	20.50	7.50
	N-3D	0.691	9.75	21.59	9.75
	N-5S	0.764	7.50	23.88	7.50
	N-6S	0.716	7.00	22.38	7.00
	N-7D	0.652	9.75	20.38	9.75

The dataset in the table above was derived from results obtained from the case study of Vaal River and West Wits tailing dams as discussed in previous. The dataset is only a small portion of the results that were obtained from the study. Population A represents 12 samples taken from Vaal River ABA results and Population B represents 6 samples extracted from ABA results for West Wits mining area.

For the population A, the sample means ($m_{AP\ (A)}$, $m_{NP\ (A)}$) and the sample standard deviations ($S_{AP(A)}$, $S_{NP(A)}$) are calculated as follows:

$$m_{AP(A)} = 1/12 \times (\sum X_i) = 6.80$$
$$m_{NP(A)} = 1/12 \times (\sum X_j) = 1.23$$
$$S_{AP(A)} = sqrt[(1/11) \times \sum (X_i - 6.80)^2] = 2.77$$
$$S_{NP(A)} = sqrt[(1/11) \times \sum (X_j - 1.23)^2] = 2.48$$

For the population B, the sample means ($m_{AP(B)}$, $m_{NP(B)}$,) and the sample standard deviations ($S_{AP(B)}$, $S_{NP(B)}$) are calculated as follows:

$$m_{AP(B)} = 1/6 \times (\sum X_i) = 22.23$$
$$m_{NP(B)} = 1/6 \times (\sum X_j) = 7.38$$
$$S_{AP(B)} = sqrt[(1/5) \times \sum (X_i - 22.23)^2] = 1.758$$
$$S_{NP(B)} = sqrt[(1/5) \times \sum (X_j - 7.38)^2] = 2.563$$

Here we set that the acceptable sampling error (*d*) of AP and NP equals 1(kg/t) and the required confidence level equals 95%.

Based on the information above, the sample sizes can be calculated as follows:

$$n_{AP(A)} = (Z_{\alpha/2} \times S_{AP(A)} / d)^2 = (1.96 \times 2.77 / 1)^2 = 30$$
$$n_{NP(A)} = (Z_{\alpha/2} \times S_{NP(A)} / d)^2 = (1.96 \times 2.48 / 1)^2 = 24$$
$$n_{AP(B)} = (Z_{\alpha/2} \times S_{AP(B)} / d)^2 = (1.96 \times 1.758 / 1)^2 = 12$$
$$n_{NP(B)} = (Z_{\alpha/2} \times S_{NP(B)} / d)^2 = (1.96 \times 2.563 / 1)^2 = 25$$

5. Discussion

The principle behind this technique is that "the size of study sample is critical to producing meaningful results" (Germishuyse et al., 2002). The approach is based on two main assumptions, which are:

(i) It is assumed that the selection of samples is random and unbiased.

(ii) The parameters used to calculate sample size are assumed to be normally distributed.

These conditions should be met before applying the method.

In order to determine a representative sample size using the presented approach, knowledge of the distribution of the variables is required. Additional factors on which sample size depend, include purpose of the study, confidence interval, spread of the population, homogeneity and the sampling error, among other important factors. Failure to recognize the impact of these factors on sample size may lead to poor results and hence poor decisions.

Adoption of the approach requires a good estimator(S) of the population variance and acceptable sampling error (d). Unfortunately, there is no way of knowing for sure how variable the population is until after a study has been carried out (Townend, 2002). This information can be acquired from:

(iii) Previous research of a similar nature: If, for example, paste pH has been used for geochemical research at a given mine residue deposit and the research was successful, the standard deviation value (σ) could be calculated from the data used in that specific research. This can be justified by the fact that while similar populations may differ in their means, they may differ slightly in terms of spread (standard deviation) (Chou, 1989).The same concept may be used to estimate acceptable sampling error (d). The sampling error defines the error between the true population mean and sample mean, as we normally couldn't know the true population mean, the prior knowledge should be in mind already. Furthermore, it should be borne in mind that the distributions of variables in mine residue deposits is governed by complex geochemical and other factors inherent to that specific geological and mining environment.

(iv) Double sampling: involves sampling in phases. Results obtained from the first phase, mean, variance or standard deviation are used to determine the sample size for the next phase.

(v) Use pilot study results: If no information about the population to be studied is available, a preliminary or pilot study can be carried out. A pilot study aids in establishing procedures, understanding and protecting against things that can go wrong, and in obtaining the variable estimates needed in determining sample size.

5.1 Factors affecting sample size

The results of the application indicate a positive correlation between sample size and its standard deviation. As the population variability, represented by sample standard deviations increases, the number of required samples n follows suit. This effect can be demonstrated by comparison of

Neutralization Potential (NP) results for population A to NP results for population B. Close examination of the NP measurement results for population A and B shows that results for population A vary significantly from population B. Such variations in NP measurements results may be attributed to more non-uniform distribution of neutralizing elements in population A than in population B. Therefore, a large sample size was required for population A in order to capture all patterns existing in the target area. In the case of population B, which exhibited uniformity in terms of the distribution of neutralizing elements, fewer samples were required. The results also highlight an exponential relationship between sample size (n) and acceptable sampling error (d). The impact of sampling error on sample size can be demonstrated by taking AP results for population A. By selecting measurement results at random to create different sample sizes and calculating error associated with each as illustrated in Table 7.1, it can be seen that a low acceptable or sampling error is associated with large sample size. Figure 5.1 illustrates this relationship. The observable trend in this figure is that as the d value gets smaller, sample size increasessignificantly.

Table 5.1: Relationship between sample size and error

Sample size (n)	Mean (m) (kg/t)	Standard deviation (s)	Sampling Error (d) (kg/t)
12	6.8	2.8	2.2
10	7.4	2.7	2.3
9	6.7	2.9	2.8
6	6.3	3.0	3.6
3	7.1	2.9	**5.0**

The results from previous sections also highlight the positive correlation between sample size and confidence level. Any increase in confidence level requires that more samples be collected. The problem with this is that an increase in sample size entails high expenditure. Therefore a trade-off among these competing factors must be reached, and it is necessary to agree on project degree of confidence and acceptable error with relevant stakeholders.

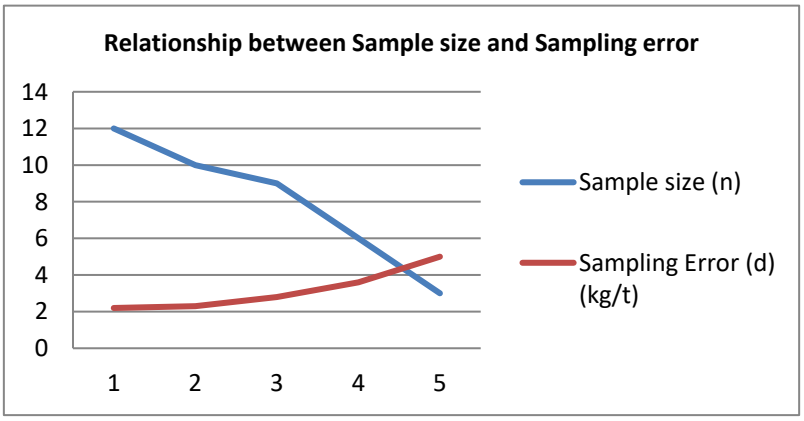

Figure 3: Relationship between sample size and error

Based on the results above, for population A, 30 samples are acceptable for AP and 24 samples are acceptable for NP; for population B, 12 samples are acceptable for AP and 25 samples are acceptable for NP. However, according to the proposed minimum sample constraint, at least 30 samples for each population should be collected.

5.2 Methods for reducing the sample size

The relationship between sample size and the various parameters on which this depends can be used as a means of reducing the number of required samples and therefore cost. For example, one may increase the allowable or acceptable error in order to minimize the number of samples

required. A major risk associated with increasing acceptable sampling error is the possibility of reducing the quality of results and representatively which may lead to poor decisions.

Decreasing material variability within population being sampled is another method that can be adopted in order to reduce the required number of samples. Methods that can be adopted include sample compositing. Mason (1992) recommends taking a large number of small increments and combining them to form a single composite sample which is submitted to the laboratory for analysis.

The use of parameters derived from a previous study of a similar nature is an effective means of minimizing the required sample size. It may be possible to use the standard deviation or acceptable sampling error from a previous study. It must be borne in mind that the accuracy of sample size calculation relies heavily on the accuracy of estimates of the parameters used in the calculation.

6. Conclusions

- Sample size determination is dependent on the purpose of the study, cost, acceptable sampling error (risk that one is willing to accept), variability of the population (standard deviation) and confidence level required.
- Understanding of the geology as well as the distribution of contaminants of concern is crucial in geochemical investigations for ERA.
- The sample size chosen determines the accuracy and representatively of the results and hence decision making. In order to achieve high precision in the values, a large sample size is required.
- Various parameters or variables that are required in calculating sample size in Equation 3.1 may be derived from either historic datasets or conducting a pilot study.
- Use of information from a previous study of a similar nature is crucial, especially when background information of the population to be studied is not available.
- A large sample size is required for a heterogeneous material, while fewer samples are need for homogeneous material.

7. Acknowledgment

The numerical calculations in this paper have been done on the super-computing system (SE000021) in the Super-computing Center for science and engineering of Southwest Petroleum University.

8. References

CHIHOBVU, E. ZHAO, B. and PULLES, W. (2011), WRC Report No. 1624/1/10 ISBN 978-1-4312-0039-9.

MASON B (1992) Preparation of Soil Sampling Protocol: Sampling Techniques and Strategies. University of Nevada-Las Vegas, Environmental Sciences Division,Office of Research and Development. Las Vegas: US EPA.

SMITH R, Jeffree R, John J and Clayton P (2004) Review of Methods of water Quality Assesment Of Temporary Stream and Lake System. Queensland: Australian Center for Mining Environmental Research (ACMER).

DOMINY SC and PETERSEN JS (2005) Sampling Coarse Gold Bearing Mineralisation-Developing

Effective Protocols and a Case Study from the Nalunaq Deposit, Southern Greenland. Second World Conference of Sampling and Blending Conference (pp. 151-165). Queensland: The Australian Institute of Mining and Metallurgy.

Petri Koistinen (2010) Computational statistics, Department of Mathematics and Statistics, University of Helsinki, Chapter 4, pp 43-45.

BARTLETT JE, KOTRLIK JW and HAGGIN CC (2001) Organisational Reserch: Determining Appropriate Sample Size in Survey Reseach. Information Technology, Learning and Performance Journal, 19 (1), 43-50.

GERMISHUYSE T, YIBAS B, and PULLES W (2002) Appendix D : Statistical Assessment of Sample and Data Representivity. Auckland Park: Pulles Howard and De Lange Incorporated

TOWNEND J (2002) Practical Statistics for Environmental and Biological Scientist. West Sussex: John Wiley and Sons.

YANG W and ZHAO B (2005) Geochemical Sampling and Analysis for environmental risk Assessment using the Wits Basin as a Case Study: Task 1 and 2. Auckland Park: Pulles Howard and De Lange incorporated (PHD).

YANG W (2006) Geochemical Sampling and Analysis for ERA using the Wits basin as a Case Study: Determination of Sample Size. Alice: University of Fort Hare.