CL:AIRE

TB12

(November 2006)

technical bulletin

CL:AIRE technical bulletins describe specific techniques, practices and methodologies currently being employed on sites in the UK within the scope of CL:AIRE technology demonstration and research projects. This bulletin provides a discussion on the 'Mean Value Test' presented in Contaminated Land Report 7 and suggests alternative techniques to reduce the risk of underestimating the true 95% Upper Confidence Limit of the sample mean.

Statistical Assessment of Contaminated Land: Some Implications of the 'Mean Value Test'

1. BACKGROUND

Statistical methods are commonly used to guide decision-making in many regulatory contexts. For the assessment of potentially contaminated land, it is common practice to use the 'Mean Value Test' approach defined in Appendix A of Contaminated Land Report 7 (DEFRA and Environment Agency, 2002). This defines the estimation of the 95% Upper Confidence Limit of the mean soil concentration of a contaminant (95%UCL, also referred to as US_{95}) and its use as the appropriate value to be compared with the relevant soil guideline value (SGV) or site-specific assessment criterion (SSAC). This 95%UCL is meant to provide a reasonably conservative estimate of whether the measured concentration is acceptable, considering the uncertainty and variability associated with site investigations.

However, when applied to real environmental datasets, the 'Mean Value Test' can actually prove to be a non-conservative choice that may wrongly categorise a potentially contaminated site as 'clean'. This bulletin illustrates how such a situation could arise, highlights the implications and introduces alternative appropriate methods to derive the 95%UCL of the mean soil concentration.

2. INTRODUCTION

In the UK contaminated land regime, the comparison of site data to an appropriate threshold is adopted to support the decision of whether, taking due account of the exposure pathways, the site poses a significant possibility of causing significant harm to human health. This comparison is often influenced by various sources of uncertainty, as discussed in CL:AIRE Technical Bulletin 7 - *Improving the Reliability of Contaminated Land Assessment using Statistical Methods: Part 1 - Basic Principles and Concepts.* As guidance for regulators, consultants and all interested parties, Contaminated Land Report 7 (CLR7) provides indications on the interpretation of site investigation data.

CLR7 suggests, for example, that a careful zoning of a site should be adopted during investigation and that the average concentration in each site zone should be compared to the relevant assessment criterion for a determinand to make a judgement on the risk posed. In addition, CLR7 acknowledges that data interpretation should be guided by statistical principles.

A statistical approach to data interpretation is based on the consideration that it is not practicable to sample and analyse the entire soil mass within a site zone and that the estimate of the mean soil concentration of a determinand therefore has to rely on a (very) limited knowledge of the contaminant distribution. Consequently, in making a decision, one needs to account for the probability that the volume of the unsampled soil may contain concentrations higher than the ones that have actually been measured. In such a case the sample average may underestimate the true average soil concentration at the site (or zone) of interest.

An informed approach to decision-making takes this eventuality into account by including in the estimation both the observed variability of determinand concentrations and information on the distribution type. The decision can then be based on a (conservative) surrogate of the true average concentration, the so-called Upper Confidence Limit (UCL) of the sample average.

The 'Mean Value Test' presented in Appendix A of CLR7 makes use of the 95%UCL (referred to as US_{95} throughout CLR7) as a means to reduce the risk of wrongly classifying a site as clean when it is truly contaminated. This procedure, although theoretically sound, can prove inadequate in some situations and provide insufficient information to the decision maker.

In the following sections, some of the available methods to calculate the 95%UCL of the mean for contaminated land datasets are presented. A brief summary of the advantages and potential pitfalls of the different approaches is illustrated by using examples and literature case studies.

3. THE 'MEAN VALUE TEST' AND THE METHODS TO DETERMINE THE 95%UCL OF THE MEAN

The 95% Upper Confidence Limit is defined as the value such that the true population mean is less than the 95%UCL with a confidence level of 95%.

The "Mean Value Test", as presented in CLR7, calculates the 95%UCL of the mean determinand concentration. This is then compared to the relevant SGV or SSAC, as appropriate. The theory behind this method is based on the assumption of normal distribution of the data but is inadequate to estimate the UCL for a dataset with a distribution other than normal. Since many environmental datasets are non-normal, alternative approaches were developed.

Land (1975) and Gilbert (1987), for example, describe methods to estimate the upper confidence limit for a population that is lognormally distributed (the method known as H-statistics). These and several other methods, referred to as parametric and non-parametric, are available to properly estimate this key value. So, the theory behind the "Mean Value Test" as presented in Appendix A of CLR7 is no longer the only possible way to estimate the 95%UCL of the population mean nor is it appropriate when data are not normally distributed.

3.1 THEORETICAL BACKGROUND

All the formulae to estimate the $[1-\alpha]$%-UCL of a population mean can be expressed in the same way. These formulae, in fact, can be easily linked to the definition of pivotal quantity (see, for example, Casella and Berger, 1990) and be adapted to a variety of cases provided that the appropriate meaning (and value) is assigned to the symbols. Although formula [1] will appear familiar to those who use Appendix A of CLR7, a family of methods is based on it, dependent upon the meaning assigned to some of the key parameters:

If you have any questions about this Technical Bulletin or would like further information about other CL:AIRE publications please contact us at CL:AIRE
Email: enquiries@claire.co.uk Web site: www.claire.co.uk

$$(1-\alpha)\%UCL = \bar{x} + t_\alpha\left(\frac{s}{\sqrt{n}}\right) \qquad [1]$$

In [1], $\bar{x}$ and s represent an estimation of the population mean and standard deviation, respectively, n represents the number of samples and t_α a statistical quantity with a meaning depending on the distribution to which the dataset can be fitted.

An intuitive review of formula [1] suggests that:

- s: the estimated UCL depends on the variability of the dataset under examination (i.e. the larger the standard deviation of the dataset, the more the UCL is likely to differ from the mean of a given dataset);
- n: the estimation of the UCL is more accurate for larger datasets (i.e. the more data are available, the less the UCL is likely to differ from the sample average for a given standard deviation);
- regardless of the variability or the sample size, the UCL of the mean differs from the sample average by a factor called t_α in [1].

Correct estimation of the values used in formula [1] is paramount for appropriately estimating the UCL. Once this is done, however, the key factor that will enable an appropriate UCL estimate is simply the term t_α. The rest of this bulletin will focus on t_α to show the implications of assigning a correct or an incorrect value to this term.

The factor t_α is a 'distribution specific' term. In the case of a sample drawn from a normal population, it stands for the upper α^{th} percentile of the Student's t distribution with n-1 degrees of freedom. In such a case, formula [1] assumes the same meaning as that given in Appendix A of CLR7 to carry out the "Mean Value Test".

For a population which is non-normally distributed, which is the case for many contaminated land datasets, t_α must have a different meaning as the Student's t distribution only applies to a normally distributed population. The approach indicated in Appendix A of CLR7, consequently, is not appropriate for many datasets.

For non-normally distributed datasets, however, formula [1] can still be used, provided that one can appropriately estimate the multiplier of the quantity:

$\left(\frac{s}{\sqrt{n}}\right)$ Estimating this multiplier for a non-normal population is not always straightforward, however.

3.2 PARAMETRIC VERSUS NON-PARAMETRIC METHODS TO ESTIMATE THE (1-α)%UCL OF THE MEAN

In statistical terminology a 'parametric approach' is a method based on distributional assumptions. In order to apply such a method, therefore, the practitioner is required to demonstrate with sufficient confidence that a dataset follows a specific distribution.

Available parametric methods to estimate the UCL of a population mean exist for normal, lognormal and gamma distributed data. Some discussion and references on these methods are given in Masi and Morgan (2005).

When it comes to real environmental datasets, however, it is often hard to prove that the data follow a particular distribution. Real datasets, in fact, include values below the limit of detection or concentrations that are outliers. In the case of geochemical datasets, outliers tend to assume very high values and make the distribution of the data extremely skewed.

In many cases, the normal and many other distributions can prove inadequate to fit the available data. When there is not sufficient evidence to eliminate outliers from a dataset, or sufficient scope or budget to re-sample the suspected locations, it is difficult to argue that the assumption of a parametric distribution (normal, lognormal or gamma) still holds. If in these cases a parametric distribution was forced to represent the pattern of the data, the analysis would produce unreliable estimates increasing the risk of a wrong site classification.

The most interesting aspect of non-parametric approaches and the reason for their being widely adopted in applied statistics is that they constitute a set of distribution-free methods. Obviously this broader applicability comes at a cost, which is a generally higher degree of conservatism. It could, however, be argued that this is not excessive given the enormous difference between the total soil mass and the sampled mass of a potentially contaminated site.

The most common non-parametric methods that can be applied for the estimation of the 95%UCL are: the Central Limit Theorem (which can be adjusted for skewed samples; Chen, 1995); the non-parametric version of the (one-sided) Chebyshev Theorem; and the Bootstrap technique, in all its different formulations (e.g. Davison and Hinkley, 1997).

4. UNDERSTANDING DATA DISTRIBUTION: ADVANTAGES AND POTENTIAL REPERCUSSIONS

The benefits of the application of parametric and non-parametric methods to real datasets can be easily understood from simple case studies in which different UCL formulae are applied to some typical datasets.

Let us consider a potentially contaminated site that has undergone site investigation. Ninety-nine soil samples have been taken and analysed by a certified laboratory for two contaminants of concern, one organic compound (contaminant A) and one heavy metal (contaminant B). These two datasets will be used to show some of the potential implications of inappropriate use of the UCL based on the default assumption of normal distribution.

The summary statistics of the two complete datasets are reported in Table 1.

Table 1. Summary statistics for the two contaminants represented in the hypothetical dataset

	Contaminant A (organic)	Contaminant B (inorganic)	Units
Minimum	1.87	0.01	mg/kg
Maximum	1316.67	2873	mg/kg
Mean	134.26	181.99	mg/kg
Standard deviation	223.60	440.75	mg/kg

Although the data themselves span a range of three to five orders of magnitude, initial review of the results did not show enough evidence for the presence of multiple populations. In addition, scattering of high concentrations across the site did not support any zoning of the site area.

The proprietary data analysis package *CLR7_Stats.xls* (ESI, 2006) has been used to analyse the data, calculate the summary statistics, identify potential outliers and the appropriate distributional assumption and correctly estimate the 95%UCL.

As suggested in Appendix A of CLR7, the Grubb's test (the so called "Maximum Value Test" in DEFRA & Environment Agency, 2002; first reported in Grubb, 1969) has been applied to the two datasets of contaminant A and B concentrations to screen for outliers. The test excluded the presence of any outlier.

The two datasets have then been checked to identify possible underlying distributions adopting a coupled visual and numerical assessment. The visual assessment has been implemented through a q-q plot (e.g. Isaaks and Srivastava, 1989) which is a simple and intuitive qualitative approach (see Masi & Morgan, 2005, for an example application). The chosen numerical assessment was performed by the Shapiro-Francia test (Shapiro and Francia, 1972) which has been carried out at a level of significance of 5%. Both the methods have been implemented through *CLR7_Stats.xls*.

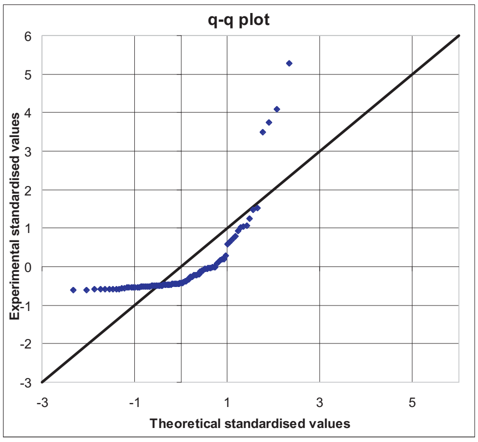

Figure 1. Normal distribution q-q plot of data for contaminant A.

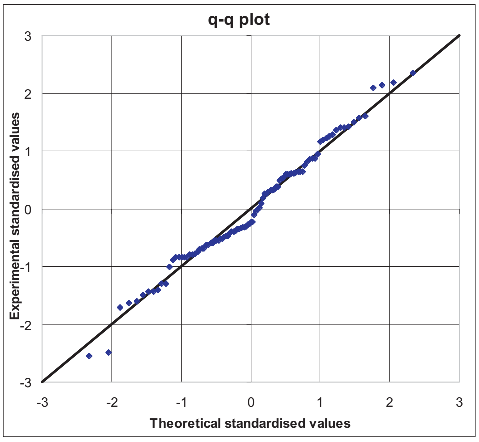

Figure 2. Lognormal distribution q-q plot of data for contaminant A.

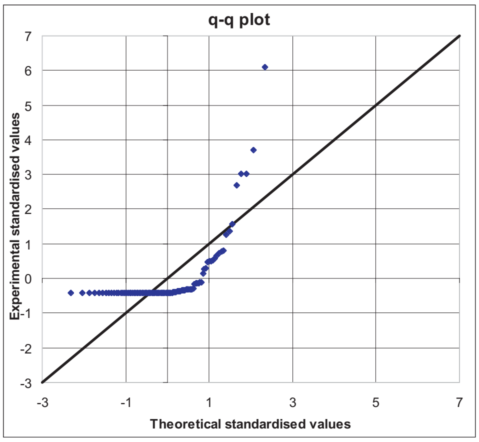

Figure 3. Normal distribution q-q plot of data for contaminant B.

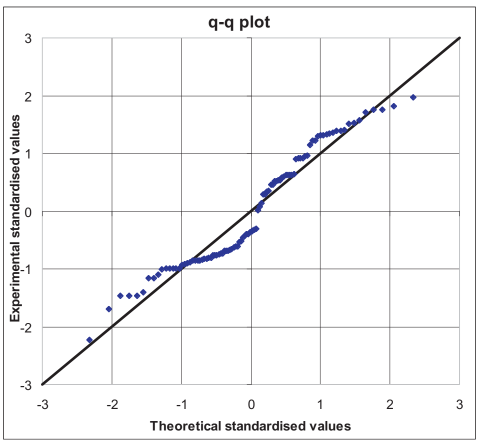

Figure 4. Lognormal distribution q-q plot of data for contaminant B.

Figure 1 shows the q-q plot to assess normality of contaminant A distribution. Normally distributed data would plot as a straight line; this is clearly not normally distributed. Figure 2 shows an equivalent q-q plot to assess the lognormality of the contaminant A dataset and shows that it is likely that this follows a lognormal distribution.

The same visual approach (Figures 3 and 4) seems to suggest that contaminant B does not follow a normal distribution; however it might fit a lognormal distribution.

The q-q plot results have then been confirmed by carrying out a Shapiro-Francia test. The results are reported in Tables 2 and 3.

Table 2. Results of the Shapiro-Francia test for normality on contaminant A and contaminant B

	Shapiro-Francia Statistics	Critical Value	Level of Significance	Normality rejected
Contaminant A	0.5519	0.976	0.05	Yes
Contaminant B	0.4545	0.976	0.05	Yes

Table 3. Results of the Shapiro-Francia test for lognormality on contaminant A and contaminant B

	Shapiro-Francia Statistics	Critical Value	Level of Significance	Lognormality rejected
Contaminant A	0.9846	0.976	0.05	No
Contaminant B	0.9443	0.976	0.05	Yes

Closer examination of the Shapiro-Francia test statistics for contaminant B (Table 3) shows that the value is close to the critical value of the statistical test. This condition suggests that a second look is taken at the q-q plot (Figure 4) where it can be seen that the data points actually follow an s-shaped pattern around the straight line. The closer visual assessment therefore supports the results of the statistical test in rejecting the hypothesis of lognormality of the contaminant B dataset.

The conclusion of the analysis of the data distribution is therefore that the lognormal theory could be applied to calculate the 95%UCL for contaminant A but both normal and lognormal calculation methods would be inappropriate for contaminant B. For the latter dataset the large skewness suggests that the non-parametric Chebyshev method is appropriate to supply a reasonably conservative estimate of the 95%UCL (Singh et al., 1999).

The correct estimates of the 95%UCL for the two datasets are reported in Table 4 together with the values that would have resulted from an inappropriate application of the normal theory. The H-statistics estimate and the Chebyshev method are reported as well.

Table 4. Comparison of 95%UCL for contaminants A and B calculated using appropriate statistical approaches and an incorrect assumption of normal distribution

	95%UCL (normal assumption)	95%UCL (H-statistics)	95%UCL (Chebyshev)	Difference*	Units
Contaminant A	171.58	192.14	232.02	60.44	mg/kg
Contaminant B	255.56	N.A.	374.69	119.13	mg/kg

N.A. = Not Applicable
* 95%UCL (Chebyshev) – 95%UCL (normal assumption)

5. DISCUSSION

A close look at Table 4 can be used to exemplify what, in statistical terms, is called a "false negative response". The 95%UCL estimated by the non-normal assumption, be it the lognormal theory (or H-statistics) or the one-sided Chebyshev theorem, are always greater than the ones estimated by means of the normal theory. This is in agreement with theoretical studies (e.g. Singh et al., 1997) showing that non-normal assumptions usually offer a better coverage of the 95%UCL in the case of skewed populations. Since environmental datasets are often characterised by high skewness, it follows that an estimation of the 95%UCL based on the normal assumption is often likely to be inappropriate for sound decision-making.

In the case that contaminant B in the example above had a SSAC of 300 mg/kg, the difference between the results of the normal and the non-normal theory would be large enough to allow an assessor unaware of the limitation of the normal theory to wrongly conclude that the site does not pose a significant risk.

A second and deeper level of observations can be made specifically from the H-statistics. It has been found that contaminant A follows a lognormal distribution, and therefore it should be appropriate to adopt the H-statistics estimation rather than the Chebyshev theorem to estimate the 95%UCL for this contaminant. However, the H-statistics estimation is characterised by some instability (i.e. a little change in the data might imply a large change in the estimated UCL) and can therefore be an unreliable basis for risk management related decisions.

To show this feature of H-statistics, the minimum of the dataset of contaminant A concentrations (1.87 mg/kg) has been replaced with the lower value of 0.01 mg/kg. Since this substitution affects only the lower end of the dataset, it should be harmless and not significantly alter any statistics. However, the H-statistics estimate of the UCL increases from 192.14 mg/kg to 275.72 mg/kg. The effect of a small change to the low end of the dataset has therefore been a significant change in the estimated 95%UCL. This effect, which has been previously observed (e.g., Singh et al., 1997; 1999), suggests that the H-statistics, although based on sound theory, is not always an appropriate basis for informed risk management. It follows that more general methods, such as the Chebyshev theorem, are to be preferred.

Some analyses have been carried out on the potential implications of an incorrect evaluation of the 95%UCL in different contexts. Masi & Morgan (2005) give a specific example of the possible impact of a false negative response with specific reference to CLR7. In that paper, a number of hypothetical site investigations were simulated on a hypothetical contaminated site. The gram by gram soil concentration of a contaminant of concern, described by a distribution with an average greater than the relevant SSAC, was sampled thousands of times to simulate thousands of hypothetical site investigations. The 95%UCL has then been estimated on each one of the datasets using the normal theory, the Chebyshev theorem and the modified Central Limit Theorem, another non-parametric method (Chen, 1995). The number of cases in which each estimate failed in identifying the potential risk posed at the site was recorded.

Table 5 shows a summary of the result of this study in terms of percentage of failure of each method adopted to estimate the 95%UCL. This table clearly shows that the Chebyshev theorem outperformed the other two methods.

Table 5. Percentage failure of each statistical method in drawing a true conclusion from the sampled dataset

Method	Percentage failure
Normal theory (Mean Value Test)	28.2%
Chebyshev Theorem (non-parametric)	7.7%
Central Limit Theorem (adjusted)	23.2%

Table 5 suggests that inappropriate data treatment is more likely to predict that the 95%UCL of the mean is below the screening criterion, even when the actual population mean is higher than the screening criterion.

6. CONCLUSIONS

A rigorous and scientifically sound approach to decision-making is expected from site assessment to ensure that decisions are correctly made and to avoid unnecessary expenditure. Together with a sound conceptual model and site zoning, appropriate statistical treatment of site data is essential. By means of some theory and practical examples, this bulletin has illustrated some possible consequences of incorrect statistical assessment. This can include both under- and overestimating the true mean concentration.

CLR7 outlines a way to carry out the 'Mean Value Test' which is designed for a specific data distribution (normal) and therefore is often inappropriate in real world cases.

Commercial statistical packages make a wide range of statistical tests readily available to the interested practitioner. However, spreadsheets are more than adequate to allow professionals to develop their own tool with little programming effort. Appropriate techniques, including the q-q plot, can easily be implemented in a spreadsheet. As the above examples show, these methods can have major benefits in data interpretation and should be used on a routine basis in the interpretation of site data, as they will significantly increase the confidence in decision-making.

ACKNOWLEDGEMENTS
This bulletin was prepared by CL:AIRE staff from an original text provided by Paolo Masi and Philip Morgan of ESI.

REFERENCES

- Casella, G., and Berger, R. L. 1990. Statistical Inference, Belmont, CA, Duxbury.
- Chen, L. 1995. Testing the Mean of Skewed Distributions, J Am Stat Assoc, 90, 767-772.
- Davison, A.C. and Hinkley, D.V. 1997. Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge.
- DEFRA and Environment Agency. 2002. Assessment of Risk to Human Health from Land Contamination: An Overview of the Development of Soil Guideline Values and Related Research. R&D Publication CLR7.
- ESI. 2006. CLR7_Stats.xls User Manual
- Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring. NY, Van Nostrand Reinhold.
- Grubb, F.E. 1969. Procedures for detecting outlying observations in samples. Technometrics, 11, 1 - 21.
- Isaaks E. H., and Srivastava, R.M. 1989. An introduction to applied geostatistics. Oxford University Press, Inc. NY
- Land, C.E. 1975. Tables of Confidence Limits for Linear Functions of the Normal Mean and Variance. In Sel Tab Mathem Stat, 3, American Mathematical Society, Providence, R.I., pp.385-419
- Masi, P. and Morgan, P. 2005. Appropriate methods for the Mean Value Test for soil contamination datasets. Land Contamination & Reclamation, 13 (4).
- Shapiro, S.S. and Francia, R.S. 1972. An approximate analysis of variance test for normality. J. Amer. Statist. Assoc. 63. 1343 - 1372
- Singh, A.K., Singh, A., Engelhardt, M. 1997. The lognormal distribution in Environmental Applications. Technology Support Center EPA/600/R-97/006
- Singh, A.K., Singh, A., Engelhardt, M. 1999. Some practical aspects of sample size and power computations for estimating the mean of positively skewed distributions in environmental applications. Technology Support Center EPA/600/s-99/006.

For further information please contact Paolo Masi and Philip Morgan at ESI:
New Zealand House,
160-162 Abbey Foregate,
Shrewsbury,
SY2 6BZ, UK
Email: PaoloMasi@esinternational.com; PhilMorgan@esinternational.com