APPENDIX I

REVIEW OF CIEH/CL:AIRE STATISTICS GUIDANCE

APPENDIX I

REVIEW OF CIEH/CL:AIRE STATISTICS GUIDANCE

I-1 INTRODUCTION

This appendix provides a review of the CIEH/CL:AIRE document entitled "Guidance on Comparing Soil Contamination Data with a Critical Concentration" (CIEH/CL:AIRE, 2008). It has been prepared by Roy MacArthur of the Food and Environment Research Agency (Fera).

I-2 OUTLINE OF THE CIEH/CL:AIRE GUIDANCE

The CIEH/CL:AIRE document outlines a methodology for comparing measured concentrations of contaminants against assessment criteria, in a way that takes the variation inherent within the measured data into account. On a basic level, the methodology can be summarised as follows:

1. Take samples from random (unbiased) locations across the site, or from random locations in a zone (or "averaging area") within the site, about which a decision must be made.
2. Measure the concentration of each contaminant in each sample.
3. Look for potential outliers and exclude them from the representative data set, if this can be justified with reference to the site conceptual model.
4. Test the remaining data set for normality.
5. If the results pass the normality test, use a t-test to evaluate whether the mean concentration of the contamination on the land is, with sufficient confidence (95%)[1], above or below (depending on the purpose of the test) an assessment criterion.
6. If the results do not pass the normalcy test, then use a one-tailed Chebychev inequality in place of the t-test for the same purpose.

In relation to steps 5 and 6, above, the CIEH/CL:AIRE guidance recommends that the 95% upper confidence limit (UCL) of the mean measured concentration should be less than the relevant assessment criterion (or "critical concentration") under a land-use planning scenario while the 95% lower confidence limit (LCL) of the mean measured concentration should (ideally) be greater than the relevant assessment criterion (or "critical concentration") under a Part 2A scenario (a "balance of probabilities" approach is also suggested for use under a Part 2A approach, in some circumstances). The use of a 95% UCL for decision-making is consistent with aspects of the methodology recommended in the withdrawn CLR 7 document (Defra/EA, 2002), where it is referred to as the "mean value test" (as opposed to the "maximum value test", which is recommended for outlier testing).

A particular feature of the CIEH/CL:AIRE guidance is that it gives relatively simple guidance that can be applied with a minimum amount of specialist statistical knowledge, so long as a sufficient amount of sampling has been undertaken in an unbiased way. The guidance is focussed on reliably estimating the true mean concentration of a contaminant and the the upper and lower confidence limits of the

[1] This is the traditional confidence level used in current guidance. The extent to which this level of confidence meets all stakeholders needs is not explored here.

mean, in a way which is: 1) relevant to the exposure that may be experienced by receptors; and 2) does not rely on any detailed knowledge about the particular *statistical* distribution of results (other than whether or not they fit a normal distribution). The guidance forms the basis of an Excel™-based spreadsheet application, which is available commercially (ESI, 2013).

There are five main observations regarding the CIEH/CL:AIRE approach that have been made, as follows:

1. The test used to assess normality could be too strict, leading to the t-test being less frequently employed, and the more conservative Chebychev inequality being more frequently employed, than needs be;
2. The use of the Chebychev inequality may lead to confidence intervals that are too wide (in cases where normality has been too strictly assessed and the t-test could have been used) or not reliably wide enough (in cases where the normal distribution cannot be applied);
3. Additional techniques might need to be considered when assessing potential outliers and the use of statistics for an overall assessment of land with different zones of contamination; and
4. The concept and phraseology of "hypothesis testing" used in the document may be difficult for some practitioners to understand.

These are described individually below.

I-3 THE T-TEST AND THE DISTRIBUTION OF DATA

The CIEH/CL:AIRE guidance states that sample results should be confirmed to be normally distributed before a t-test is applied. However, because the t-test is applied to the mean, the assumption it relies upon is that the probability distribution that describes the uncertainty about the mean is normally distributed, rather than the variation displayed by results. While it is true that normally distributed data *guarantees* that the uncertainty about the mean is normally distributed, the central limit theorem (Rice, 2007) says uncertainty about the mean will be approximately normally distributed as long as the mean is derived from a large enough number of samples (the theorem says that normality is achieved when an infinite number of samples are used).

As an example, Figure 1 shows the probability distribution of the concentration of a contaminant in samples taken from two areas of land. Both contain a mean of 1000 $\mu g.kg^{-1}$ and, in both cases, the between-location relative standard deviation (RSD; the standard deviation across results produced by samples taken at different location divided by the mean concentration) is 200%. In one area the concentration follows a log-normal distribution and in the other the concentration follows a gamma distribution (ie, the distributions of individual results are very different and neither are close to the normal distribution). However, as shown in Figure 2, the probability distribution of the mean of 100 samples from each of the distributions is approximately normal.

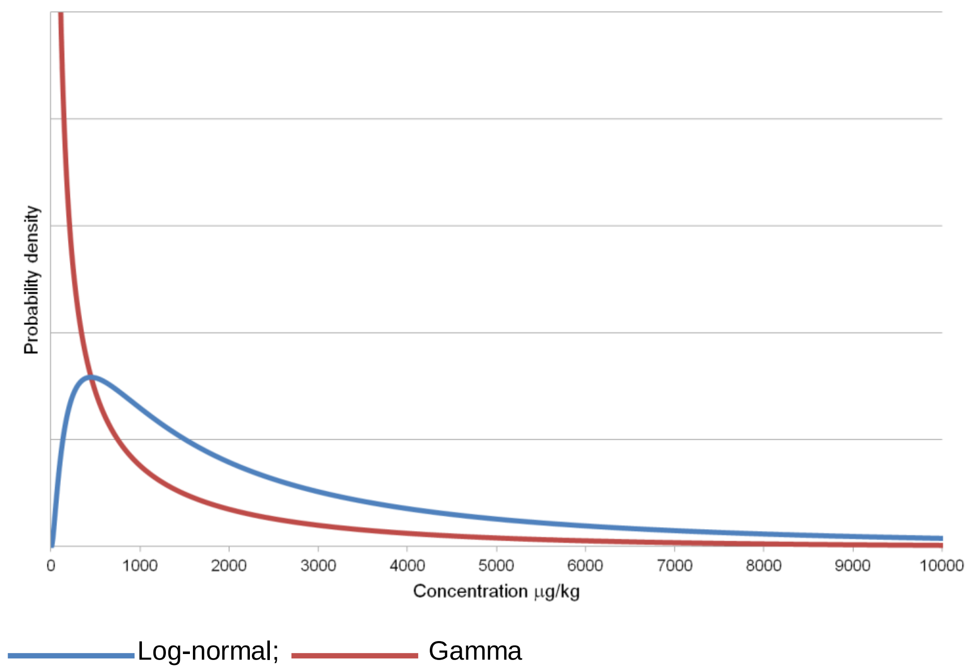

Figure 1: Probability distribution for individual samples taken from contaminated land with the same mean and variation but two different distributions

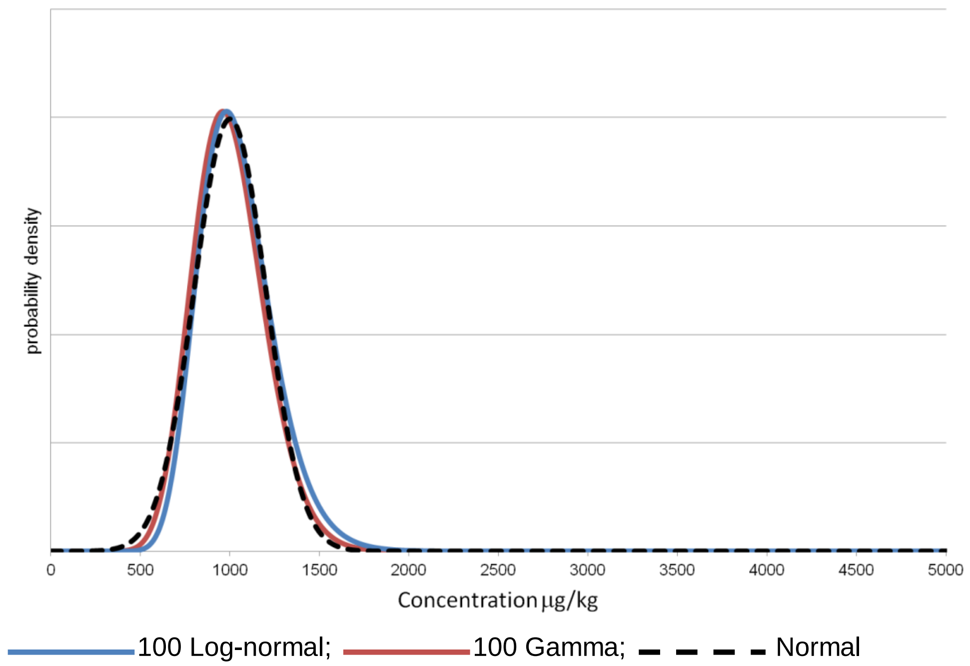

Figure 2: Probability distribution for the mean of 100 samples taken from contaminated land with the same mean and variation but two different distributions.

This is an example of the central limit theorem in action. If enough samples have been taken, then uncertainty about the mean concentration is described by a normal distribution whatever distribution the individual results have. If enough samples have been taken the distribution that describes uncertainty about the mean depends *only* on the mean concentration and the size of between-location variation (as estimated by the observed standard deviation), not the shape of the variation.

An obvious question is 'How many samples do we need before we can assume that the probability density is close enough to normal?' An answer is that it depends on how precise the estimates of the 5th and 95th percentiles of the interval that describes the mean concentration need to be. In general, lower relative standard errors (relative standard deviation of the individual results divided by the square root of the number of samples) for the mean are more normally distributed and estimates of percentiles made assuming a normal distribution are more reliable. In our example the between-location RSD associated with individual results is 200% and the number of samples (n) is 100. Hence the relative standard error (RSE) associated with the estimated mean concentration is 20% ($RSE = RSD/\sqrt{n}$). Estimates of the 5th and 95th percentiles of the mean concentration assuming a normal distribution are within 5% of true percentiles for the mean concentrations based on individual results with log-normal or gamma distributions (see Table 1).

Table 1: Percentiles of a normal distribution and the mean of 100 samples taken from a gamma distribution and a log-normal distribution

Distribution	5th percentile	95th percentile
Normal	671	1329
100 Gamma	695	1350
100 Log-normal	708	1358

As illustrated above, a t-test can be applied while assuming normality as long as the RSE of the estimated mean concentration is sufficiently small. An assessment of the distribution of individual measurement results is not necessary for the purposes of estimating the mean concentration if this condition is met. For example if we set the condition that the RSE must be no larger than 25% then the relation between variation expressed as an RSD and the number of samples required to provide an essentially unbiased estimate of the average concentration is given in Table 2. In addition the variation is described as the approximate ratio between high concentration (97.th percentile) and low concentration (2.th percentile) parts of the area by assuming a log-normal distribution.

Table 2: Variation and the number of samples needed to give an unbiased estimate of the average concentration

Variation expressed as a RSD	Variation expressed as the ratio of concentrations in high concentration parts against low concentration parts of the area	Number of samples for an unbiased estimate of the average concentration
61%	10	6
79%	15	10
97%	20	15
112%	25	20
137%	30	30
177%	45	50
217%	60	75
250%	70	100

I-4 THE CHEBYCHEV INEQUALITY

The CIEH/CL:AIRE guidance recommends the use of the Chebychev inequality where the variation displayed by measurement results is non-normal. The Chebychev inequality gives an upper limit for the quantile of any distribution given a true mean and a true variance (or standard deviation). Hence, the true quantile of any particular distribution will always be below the value given by the Chebychev inequality.

Although the Chebyshev inequality is recommended in the CIEH/CL:AIRE document for estimating confidence intervals of the mean concentration from sample data, a note on its use in USEPA (2002) guidance says that there is an underlying assumption in doing so that the true mean and true variance are estimated well by the mean and variance of the sample data. It should be noted that this condition is only met in the sense that estimates are likely to be unbiased where the data is likely to be sufficient to apply the normal approximation described above. As a consequence, its use is considered to be valid only in cases where it would be better to apply the normal approximation.

Where there relative standard error is too large to apply the normal approximation the use of the Chebychev inequality, or any one-size-fits-all approach can lead to estimates of ‘mean plus potential error’ and ‘mean minus potential error’ that are too high or too low. In these cases taking more samples to reduce the relative standard error, or fitting a different statistical distribution to measurement can lead to estimates that are more reliable.

If it is that case that it is not possible to take enough samples to reduce the relative standard error of the mean to a sufficiently low figure AND sufficient expertise to assess which statistical distribution best describes measurement results is not available then estimates of the mean concentration of contaminant may be biased.

I-5 OUTLIER TESTING AND ZONES

The purpose of outlier testing is to check the assumptions used when estimating the mean concentration and its level of uncertainty, are consistent with the observed results. If there is an inconsistency then we may need to revisit our assumptions, and modify our conceptual site model. Assumptions concerning the sample results may (for example) include:

1. That the concentration of contaminants across the site is the result of the same or sufficiently similar processes, so they may be considered to be a single statistical population;
2. Samples are taken at random, or in a way that does not lead to an unrepresentative number of samples with a high or low concentration.

3. The distribution of contaminant concentrations follows a particular statistical distribution; and
4. All samples were taken and analysed correctly, and no significant sampling or analytical errors or biases exist.

A particular *statistical* issue is the choice of the basis for the outlier test. Given the likelihood that a relatively small number of results will be available, non-parametric tests or tests based on data-clusters are not likely to be useful. It is unlikely that data will be sufficient to discriminate between different right-skewed distributions that may describe measurement results. Hence selecting a particular statistical distribution against which to test can be problematic.

Where a statistical outlier is detected there are three likely explanations for the observation:

1. Different areas of land may have been subject to different sources of contamination, or may have been subject to different levels of disturbance, remediation or other process since becoming contaminated. Note that this can result in either more contaminated areas within a site (hotspots), or cleaner areas within a more contaminated site (cleanspots).
2. The sample may not have been taken correctly.
3. The analysis may not have been undertaken correctly.

If there is insufficient evidence to confirm either case 2) or 3), then the presence of the outlier may be indicative of an area with a different contamination profile, and which may be considered as a separate zone if there is sufficient evidence to modify the conceptual site model.

A potential limitation of the current CIEH/CL:AIRE guidance is that, while outlier tests are described and the separate zoning of land is mentioned, guidance is not given on how to estimate the mean concentration of a contaminant across zones or crucially how to estimate the upper and lower intervals. This may encourage the assumption that separate zoning is not required. This problem has been noted by Nathanail (2004) and the use of the mean of intervals determined separately for each zone weighted for the area of each zone has been suggested. Such an approach is simple but somewhat conservative, becoming more conservative as more zones are added. A less conservative alternative is to estimate an upper and lower limit for the mean concentration based an observed mean concentration, an observed standard deviation and degrees of freedom as described in current guidance, but to use a combined mean and standard deviation for the two zones.

I-6 HYPOTHESIS TEST OR MEASUREMENT UNCERTAINTY?

The current CIEH/CLAIRE guidance describes the use of hypothesis tests to assess the sample results, including an approach based on the null hypothesis test that *the true mean confidence is not above the assessment criteria*.

An alternative, and statistically equivalent approach, is to estimate upper and lower one-tailed 95% confidence intervals. These can then be used to ascertain and illustrate whether an assessment criteria is exceeded, as shown in Figure 3 and Tables 3 and 4 below.

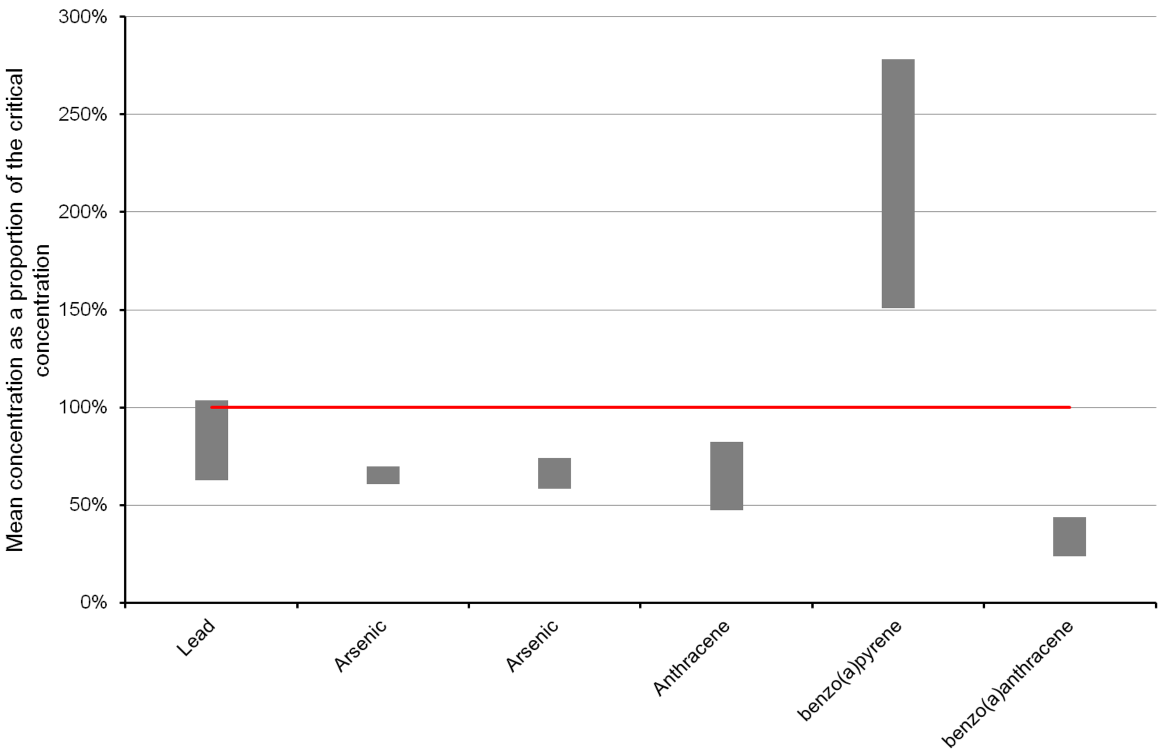

Figure 3: Interval (95% one-tailed) for the mean concentration of contaminant

Table 3: Assessment of concentration of a number of contaminants on an area of land with a hypothesis test

Value	Lead	Arsenic	Arsenic	Anthracene	Benzo(a) pyrene	Benzo(a) anthracene
Critical Concentration	450	24.8	24.8	0.709	1	5.91
Mean Concentration	374.7	16.174	16.417	0.4604	2.1457	1.9957
s.d.	257.68	3.0845	3.704	0.348	1.7822	1.6799
Number of samples	23	23	12	23	23	23
Significance	5%	5%	5%	5%	5%	5%
RSD (%)	0.69	0.19	0.23	0.76	0.83	0.84
RSE of mean (%)	0.14[a]	0.04[a]	0.07[a]	0.16[a]	0.17[a]	0.18[a]
t	1.72	1.72	1.80	1.72	1.72	1.72
Conclusion	Null hypothesis not rejected	Null hypothesis not rejected	Null hypothesis not rejected	Null hypothesis not rejected	Null hypothesis rejected mean > limit	Null hypothesis not rejected

[a] RSE≤20%. Under Recommendation 1 normal statistics can be assumed

Table 4. Assessment of concentration of a number of contaminants on an area of land using the size of the uncertainty associated with the mean concentration

Value	Lead	Arsenic	Arsenic	Anthracene	Benzo(a) pyrene	Benzo(a) anthracene
critical concentration	450	24.8	24.8	0.709	1	5.91
Mean concentration	374.7	16.174	16.417	0.4604	2.1457	1.9957
s.d.	257.68	3.0845	3.704	0.348	1.7822	1.6799
Number of samples	23	23	12	23	23	23
confidence (1-tailed)	95%	95%	95%	95%	95%	95%
RSD (%)	0.69	0.19	0.23	0.76	0.83	0.84
RSE of mean (%)	0.14	0.04	0.07	0.16	0.17	0.18
T	1.72	1.72	1.80	1.72	1.72	1.72
Interval	282.44	15.07	14.50	0.34	1.51	1.39
	466.96	17.28	18.34	0.59	2.78	2.60
Interval as a proportion of	62.7	60.8	58.5	47.4	151	23.6
the critical conc (%)	103.8	69.7	73.9	82.5	278	43.9

The above tables illustrate an alternative description of confidence intervals for mean concentrations, provide all of the information provided by the hypothesis testing approach and, in addition, provide additional information that may inform assessment, such as, in the example above, that further measurements of lead may allow a stronger conclusion to be drawn, as there is only a narrow non-rejection of the null-hypothesis, while in contrast further measurement of the other contaminants are unlikely to result in the change of the conclusions.

I-7 References

CIEH/CL:AIRE, 2008. Guidance on Comparing Soil Contamination Data with a Critical Concentration. Chartered Institute of Environmental Health and Contaminated Land: Applications in Real Environments.

Defra/EA, 2002. Assessment of Risks to Human health from Land Contamination: An Overview of the Soil Guideline Values and Related Research. Defra and the Environment Agency, 2002. R&D Publication CLR7. Withdrawn.

ESI Ltd, 2013. Contaminated Land Statistical Calculator. Available at: http://esinternational.com/contaminated-land-statistical-calculator/. Accessed 13 March 2013

Nathanail, C.P., 2004. Technical Note: The use and misuse of CLR 7 acceptance tests for assessment of risks to human health from contaminated land. Quarterly Journal of Engineering Geology and Hydrogeology, 37, 361–367.

Rice, J.A., 2007. Mathematical Statistics and Data Analysis. Third Edition. Duxbury Brooks Cole Thomson. ISBN 0-534-39942-8.

USEPA, 2002. Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites. OSWER 9285.6-10. December 2002. Office of Emergency and Remedial Response, US Environmental Protection Agency, Washington, D.C. 20460.