Nicola,

As discussed, following experience in a public inquiry at the end of last year, I consider there are strong grounds for updating the statistics guidance. At the inquiry, statistical evaluation of the data became a major issue and the statistics guidance was applied in a black and white fashion. Having been involved in the steering group of the guidance, I recognised how little I really understood about the application of the statistics and therefore sort some help from a board member of the Royal Statistics Society and Operations Research Society. They directed me to Nigel Marriott as one of the leading experts on the application of statistics to sampling programmes. What Nigel did was help me understand that the CL:AIRE statistics guidance was being misused at the inquiry by not adequately interrogating the data to which the statistical tests were being applied.

I have now asked Nigel to take another look at the Guidance and provide a commentary on whether it meets the needs of the contaminated land community. I attach this commentary for review by the Land Forum. Nigel has offered to undertake a re-write of the guidance and thinks this will only require two to three days of his time. However it will clearly take more time to respond to comments on the draft he prepares and finalise a document that meets the needs for the contaminated land community. Could we put this on the Land Forum Agenda and see if there is an appetite for an update?

Many thanks

Peter Witherington

As a Chartered Statistician with 25 years' experience of solving business problems around the world and 15 years' experience of running training courses for non-statisticians in many industries, I have lost count of the many ways people get themselves confused when trying to undertake a statistical exercise. A very common question I get asked is "what software/method should I use to do this analysis?" and invariably my answer is "what are you trying to achieve?" Statistics is a lot like a toolbox with a lot of tools. Sometimes you need to use a hammer, sometimes a spanner, sometimes a saw, etc. If you tried to undertake all building work with just a hammer, you wouldn't get very far and rightly so. Yet for some reason, people seem to think that with statistics, you can do everything with just one or two tools.

When it comes to sampling questions, my experience has shown me that it is all about objectives, objectives, objectives. One person's objective will require a completely different approach from another person's objective. From what I know of contaminated land sampling, I am aware that the following questions are quite common:

1. What is the overall level of contamination across the site?
2. Are there localised hotspots of contamination within the site?
3. Is there a pattern of contamination across the site?
4. Are contamination levels (locally or generally) acceptable for planning purposes?
5. Are contamination levels (locally or generally) acceptable for Part 2A purposes?

I am sure that there are other questions as well but just from this list of questions, I can foresee at least 10 different objectives that could arise for any land sampling exercise each requiring a different tool.

The statistical guidance document you forwarded me describes (in text book fashion) a single tool to be applied to two different objectives i.e. does land meet the criteria as set out by planning/part 2a purposes? My initial opinion is that this guidance overcomplicates things. My understanding from conversations with you is that the principle used by courts is that of "on balance of probability". I would interpret this standard of proof to mean that if the probability of exceeding a certain threshold value is greater than 50% then the court would conclude that the threshold had been exceeded and make judgement accordingly. Yet the CLAIRE guidance initially uses much stricter criteria for drawing this conclusion which is a P-value less than 5%. This is equivalent to a court requiring "proof beyond reasonable doubt" rather than on "balance of probabilities". However, the guidance then goes on to advise that if the site cannot be determined on this strict test, the assessor can re-assess on the basis of a P-value of 51%

I would say that this is an example of p-values being used but not being understood. In March 2016, the American Statistical Association became so concerned about misuse of p-values that they issued this general guidance to the world. This guidance has been endorsed by the Royal Statistical Society as well and you can download this statement at this link http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108 . In this statement (which is preceded by an editorial so skip that), 6 principles are listed for the correct use of p-values and at first glance, I would say that the CLAIRE guidance is in conflict with 4 of them (2, 3, 5 & 6).

In fact, if courts are ruling on the basis of "balance of probabilities", then there is no need for statistical tests at all. For example, let's assume we are measuring a piece of land such as a school playing field where it is expected that children will make use of all the available land. In this instance, the global average of the site is what we want to measure and let's suppose we find that from our unbiased random sample that the global average is 40 units and the Category 2/3 threshold

is 50 units. Straightaway, we can conclude that the probability that we are exceeding this threshold is less than 50% and a court can rule that it has not been breached. Why? Because the average is less than the threshold and this immediately means the probability of exceeding the threshold must be less than 50% and there is no statistical method (*see digression*) that will change the probability to be greater than 50%. It should make no difference to the court whether this probability is 49.9% or 9.9%, both should lead to the same decision. If you then tell me that a court would rule differently in these 2 scenarios, then my response would be that the burden of proof has been misunderstood and the sample has been selected under an incorrect objective. Therefore we would need to repeat the sampling exercise this time with a clearer understanding of the objective. Trust me, I have seen this time and time again!

Digression Actually there is one statistical tool that could end up concluding that the probability of exceeding a threshold of 50 units is greater than 50% even though the sample average is 40 units. This is a method known as Bayesian sampling and can be demonstrated with the following example. Suppose you are and Alien with no understanding of human reproduction. You want to find out what percentage of adult humans are pregnant and which sex carries the babies so you randomly abduct 10 men and 10 women. Your tests show that none of the men and women are pregnant. What conclusions would you draw? An Alien statistician would tell be able to tell you that the general pregnancy rate is unlikely to be more than 12% but it would have no idea which sex carries babies. However, a human that sees this result would know straightaway that 0% of men are pregnant and a human statistician would say that the female pregnancy rate is unlikely to be higher than 20%. The reason why humans would come to a different conclusion from the aliens, even though both have the same data, is that humans are using prior knowledge that men can't have babies which aliens lack. Bayesian statistics is a method whereby prior expert knowledge (such as the layout of previous gas works on reclaimed land) can be fused with new data to produce an updated estimate of contamination. In our example, suppose an expert said that he or she would expect contamination to average 100 units then our sample average of 40 units can be combined with this expert opinion and might be result in a revised estimate of greater than the threshold of 50 units.

If I was asked to assist CL:AIRE in producing revised statistical guidance, my first step would be untangle the myriad list of objectives that drive sampling & measurement of land for contamination. Once this has been achieved, I would then recommend robust and simple to use statistics which would directly result in probabilities for each of the possible decisions available. For example, a method could be used to measure the probability of each category e.g. P(category 1)=2%, P(category 2)=15%, P(Category 3)=60%, P(category 4)=23%. In my opinion, such outputs are easier to work with when making decisions. As far as possible, I would use methods that are in the public domain and applied elsewhere. For example BS6000 and BS6001 are standards I have helped clients use before. My experience of sampling covers many industries including food safety, medical devices, utility asset performance, staff workload, school places planning, labour skills shortages, etc, and I have seen many sampling plans described in these fields. The basic principle I would use is that any proposed method could be used and understood by non-statisticians without calling upon a statistician.

If it was decided that more advanced methods are preferred that would require calculations undertaken by a qualified statistician then Bayesian methods as described in the digression above could be used as well as geo-spatial statistical modelling. The latter would result in a heat map of a site perhaps colour coded according to the category 1,2, 3 & 4 standards.