Uncertainty in site investigation and the conceptual site model

Jonathan Welch

15 July 2015

AECOM

Contents

Definition of a Conceptual Site Model (CSM)

"*A conceptual model represents the characteristics of the site in diagrammatic or written form that shows the possible relationships between contaminants, pathways and receptors.*"

CLR11 Model Procedures (Defra/Environment Agency 2004)

"A CSM is a representation of the nature, fate and transport of discharges, wastes or contaminants that allows assessment of potential and/or actual exposure to contaminants. It is an hypothesis that can be tested and refined"

ANZECC 2000

AECOM

Type of information in a CSM

- General site information
- Site characteristics
- Actual/potential receptors, and release and transport mechanisms
- Soil contaminant source characteristics

USEPA corrective action workshop (online 2015)

"*All uncertainties need to be noted..*"
(in the risk assessment)

CLR11 Model Procedures (Defra/Environment Agency 2004)

AECOM

Typical objectives for Site Investigation

- Check presence of contamination at a known potential source
- Measure extent of a known area of contamination
- Find an unknown contamination ‘hotspot’
- Compare an average concentration to a threshold
- Calculate an area, volume or mass for treatment
- Validation testing e.g. remediation process control
- Verification testing e.g. regulatory compliance
- Investigate properties of potential migration pathways
- Find secondary lines of evidence to develop the conceptual model

AECOM

Site Investigation

- Site Investigation is generally a type of survey to draw conclusion for the general population from data samples
- Surveys require planning to obtain representative results

  (1) Identify the population of interest

  (2) Estimate the amount of variability expected

  (3) Decide on the level of confidence required

- Preliminary investigation and CSM are a prerequisite for SI design
- The common purpose of a site investigation is to reduce uncertainty in the CSM to an acceptable level for decision making

Sources of uncertainty in Site Investigation data

- Incomplete or incorrect CSM

    failure to investigate the significant pollutant linkages

- Sampling error

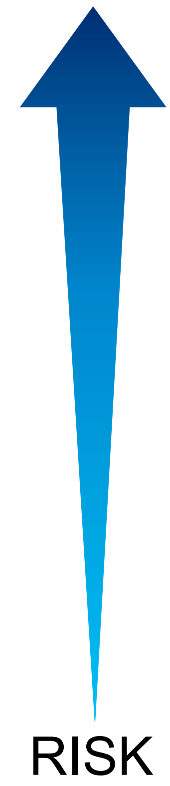
    sample properties not representative, too few samples

- Handling, storage and transport

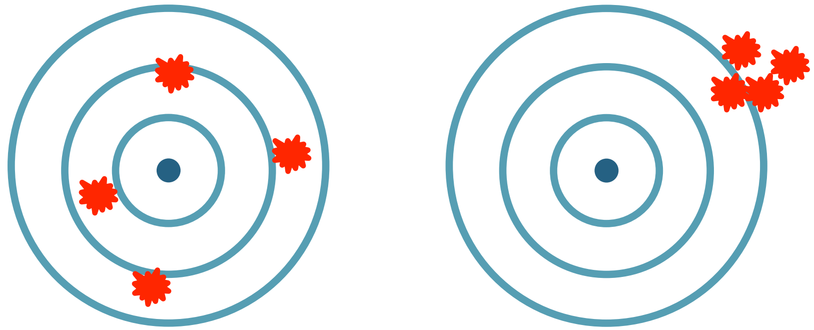
    cross-contamination, miss-allocation, degradation, loss

- Laboratory specific

    sub-sampling, loss of in preparation and extraction, equipment accuracy and precision



Accuracy and Precision



Lack of Precision   Lack of Accuracy

- Precision defines how close you can get to the target   *random error*

  try enough times you can get there (or take an average);
  use of duplicates and blanks may aid evaluation of precision

- Accuracy defines whether you have your aim correct   *systematic error*

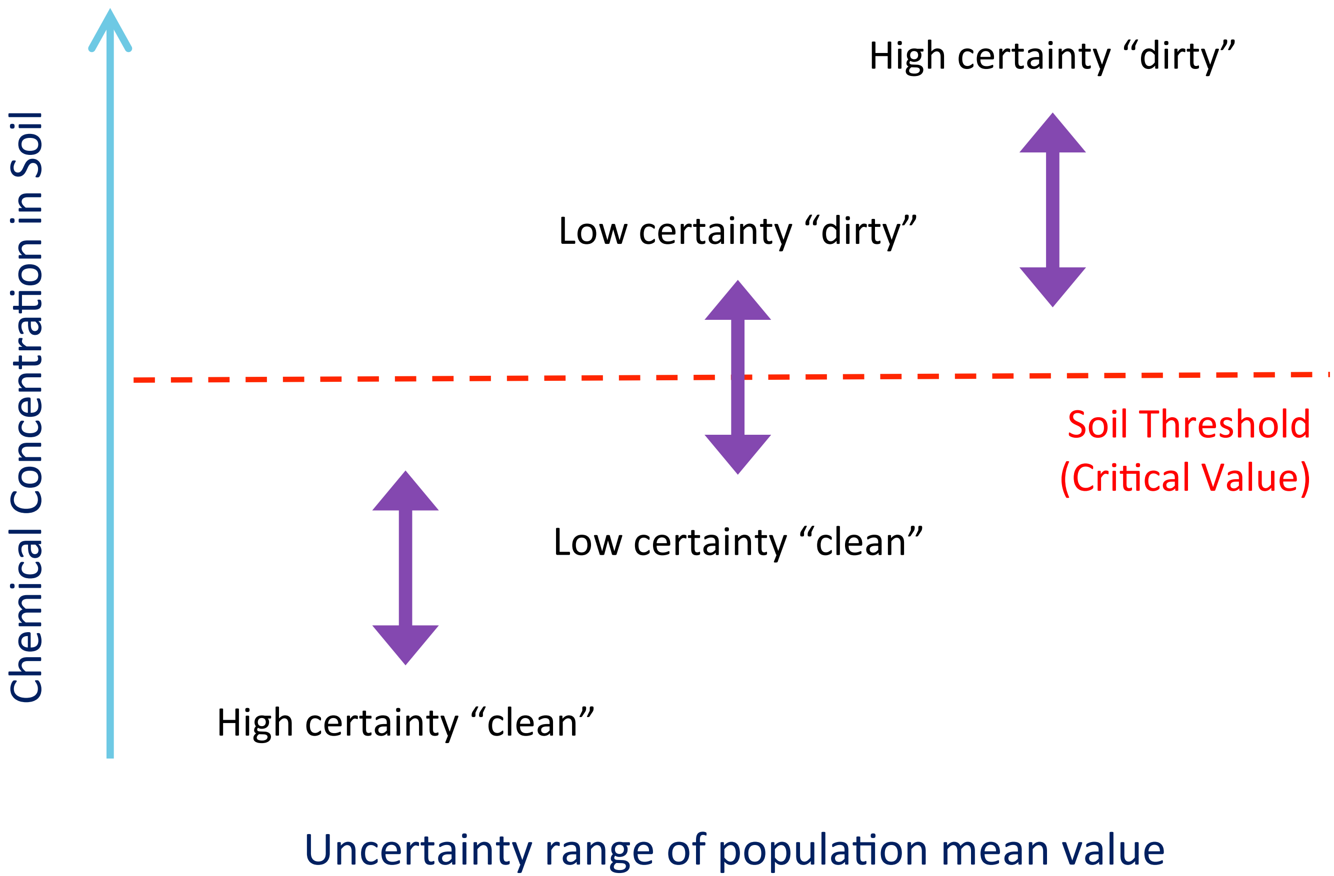
  poor accuracy may result in never hitting the target

- CSM and multiple lines of evidence are main defence against systematic errors and loss of accuracy

AECOM

Uncertainty

Uncertainty arises when data are close to a decision point



Statistical Error

- Under Planning Legislation
  - The assumption is that the land is contaminated.
  - A high level of confidence is required to prove the land is safe.
  - High uncertainty leads to a failure to reject the initial assumption.
  - This may lead to unnecessary remediation.
- Under Part IIa Contaminated Land Legislation
  - It is assumed that the land is not contaminated.
  - A high level of confidence is required to prove the land is contaminated.
  - High uncertainty leads to a failure to reject the initial assumption.
  - This may lead to an unacceptable risk from contamination.
- The two regimes work in different ways.
  - Low statistical power to reject the initial assumption is a particular problem for Part IIa because the benefit of the doubt is given to the site.

AECOM

Soil Variability

15 July 2015

AECOM

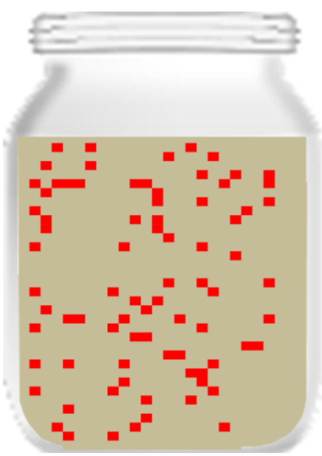
Average (arithmetic mean) soil properties

- Soils can be highly variable at small, intermediate and large scales
- Sample means are unbiased estimates of the population mean, and get closer to the population mean with more (unbiased) samples
- The mean of samples tends to the same as the population mean regardless of the sample physical size (mass)
- Distribution of sample means becomes more **Normal** as the number of samples increases regardless of the underlying population distribution.
- These properties make the mean highly suited to statistical analysis and comparison with a regulatory threshold

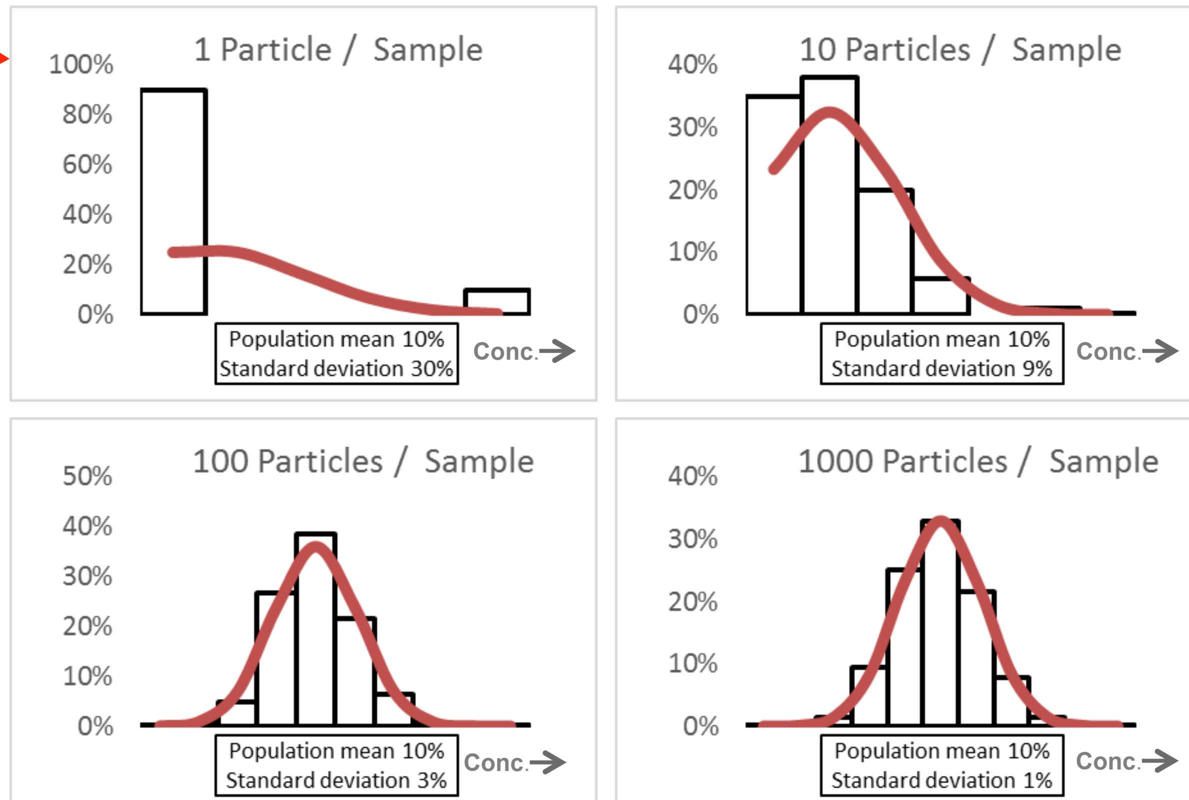
Averages of right-skewed distributions

Assume a population ratio for particles in the soil of:
1 x red (100% contaminated) to 9 x brown (uncontaminated).
Randomly extract 10,000 samples of a given number of particles and average. Plot the distribution of average concentrations in a histogram.

Note that 100x more particles gives 10x less spread (s.d. ∝ √n)



Right-Skewed →



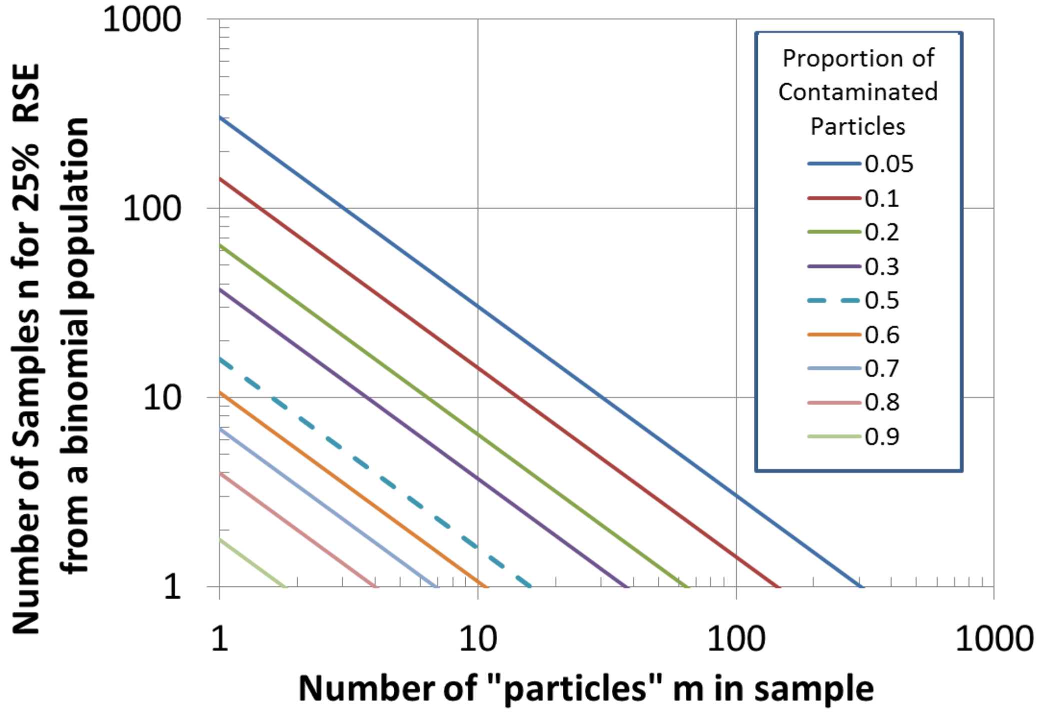
← Normal

AECOM

Sample number and sample physical size

For an expected (target) data reliability expressed as RSE (relative standard error) for the mean of 25% we can test the relationship between numbers of particles and samples of a binomial distribution. Based on a binomial model we might need 100's of samples!

Objective RSE (standard deviation / mean) ÷ √sample size ≤ 25%



NOT FOR DESIGN
This graph relates only to the highly simplified model of a soil as a binomial distribution. In reality soils exhibit variability at different scales including variation within particles in addition to other sources of error and uncertainty,

AECOM

Averaging areas

- Risk models assess average exposure for the sensitive receptor therefore the physical site dimension is important
  e.g. single garden
- This is an important distinction because we are concerned with individuals and not averages; contaminants have threshold health effects
- Uncertainty of the sub-plot means > uncertainty of site-wide mean
- At least one sub-plot will have a true mean value > site-side mean
- Statistical support for the mean value should be the averaging area (corrections for the support are site specific)

Investigation Strategy

15 July 2015

AECOM

Types of sampling

- Convenience sampling

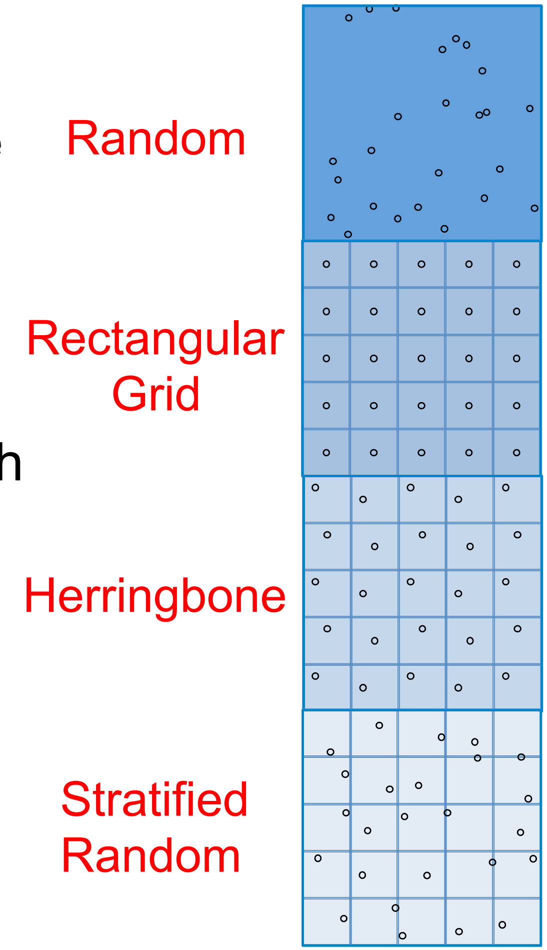
  generally no properties of the sample can be inferred to the general population.

- Judgemental sampling

  non-random sample selected by an expert: results dependent on individual skill. Use with care due to risk of bias and legal challenge.

- Statistical based sampling

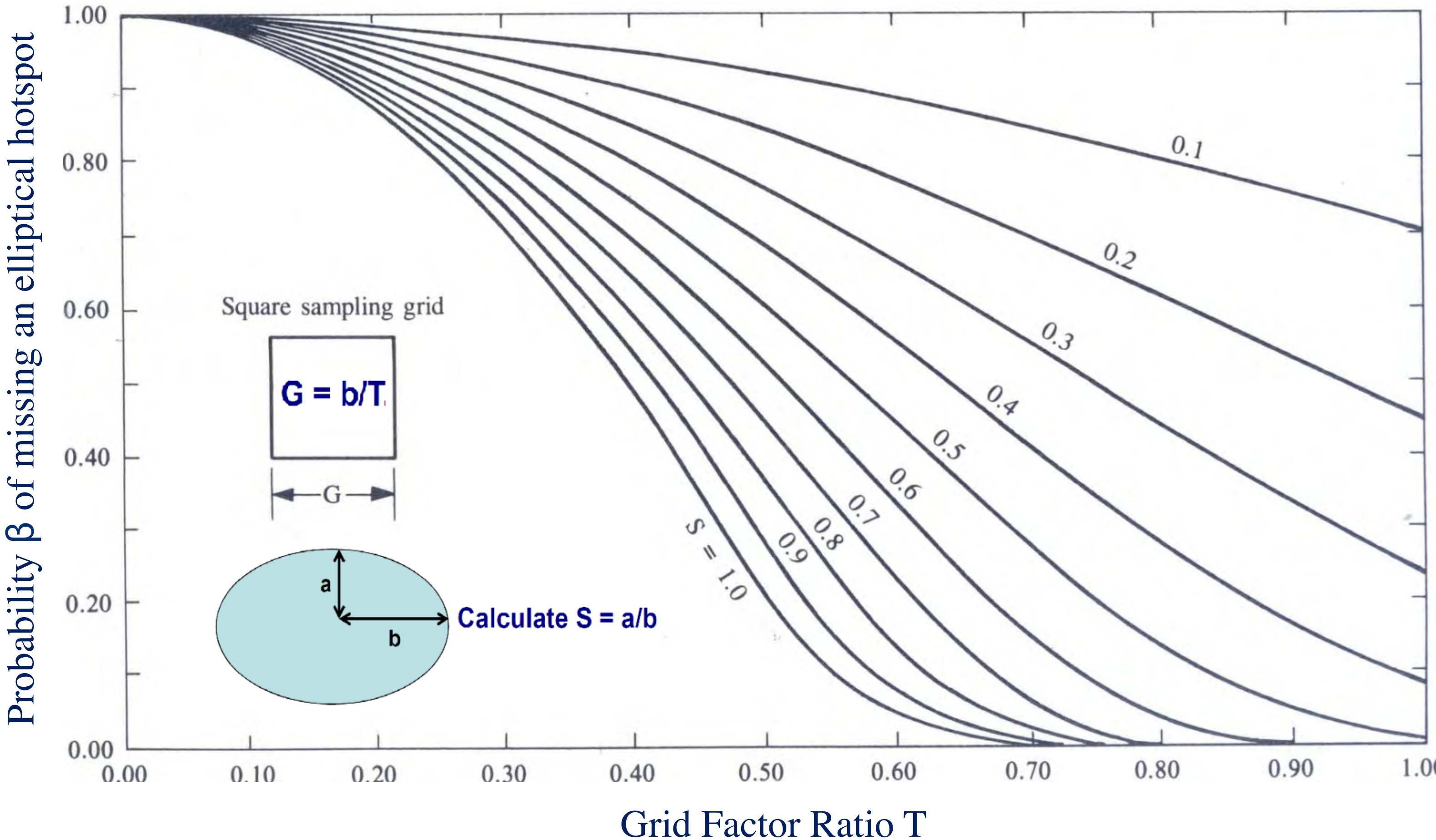
  every member of the population has an equal probability of being selected; suitable for calculating probabilities



AECOM

Estimating the probability of a Hotspot

Grid size necessary to achieve a probability β of hitting an elliptical hotspot for a square grid.



(ELIPGRID simulations by Zirschky & Gilbert widely quoted by USEPA)

AECOM

Bayesian approach to Hotspot Detection

Bayes method of using investigation data to update
prior knowledge (e.g. CSM) to obtain a revised risk assessment.

(1) Decide the size of hotspot that would be significant

(2) From preliminary investigation assess the probability of a hotspot being present. This is known as the prior probability (PrB).

(3) Decide on the required probability that a hotspot does not exist if the investigation fails to find one. This is the posterior probability (PrA).

(4) Use theorem to determine the hotspot detection probability rate (PrH)

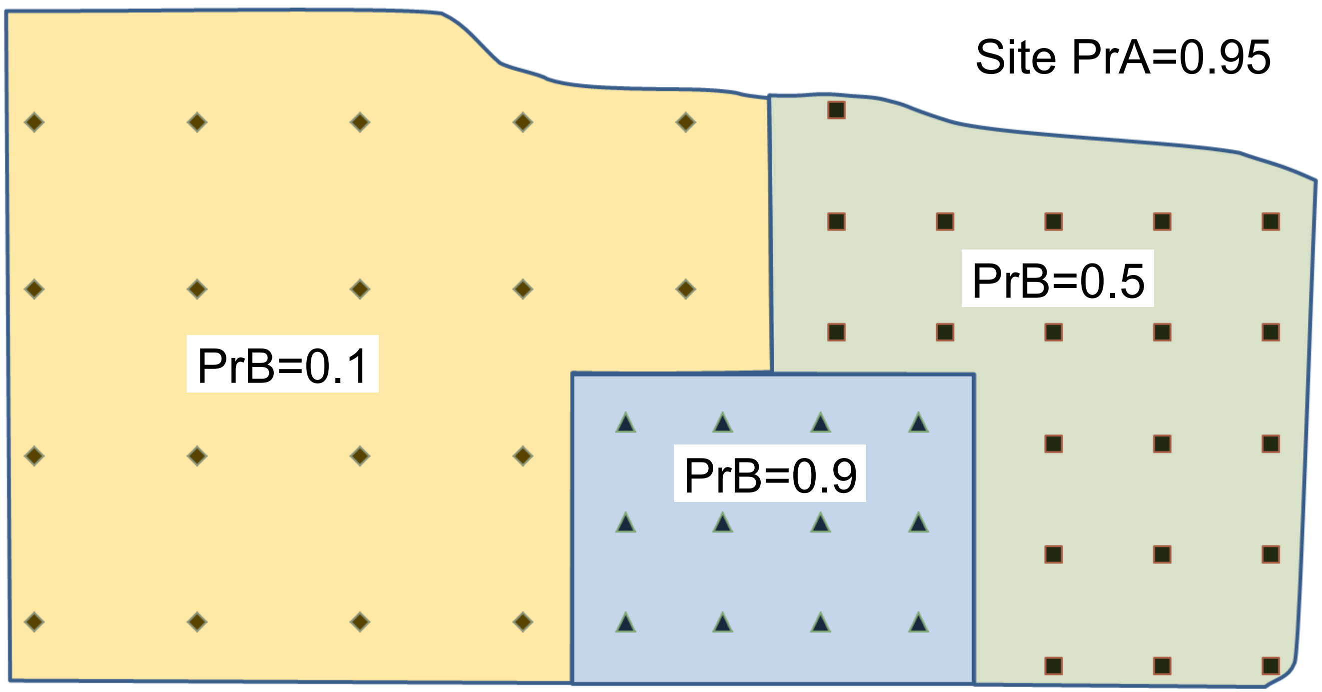
$$PrH = PrB + PrA - 1/PrB.PrA$$

(5) Determine the grid size requirement from PrH.

Example: assume 20% prior probability of a hotspot (PrB) being present and an overall posterior probability (PrA) that no hotspot exists (if not found) of 95% to obtain a design hit rate PrH of 79%. Using the table or by other means calculate the grid size required for the revised hit rate PrH.

Site Zoning

Prior knowledge from the CSM can be used to subdivide the site into areas having different prior probabilities of hotspots:
use Bayes theory to obtain an efficient grid design for hotspot detection.



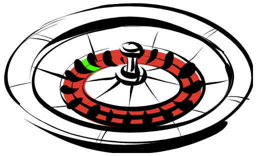
Method after CLR4 Sampling Strategies for Contaminated Land, DETR, 1994.

Geostatistical modelling

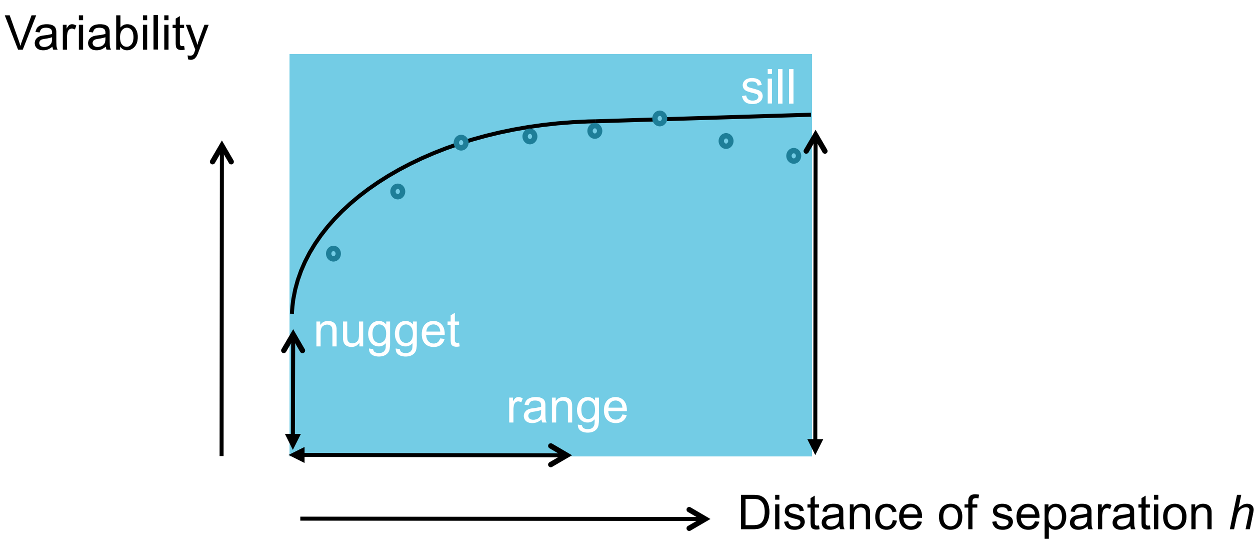
"Everything is related to everything else, but near things are more related than distant things.“

Waldo Tobler’s First Law of Geography

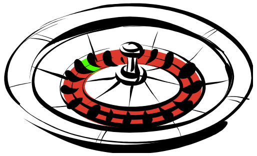
Geostatistics



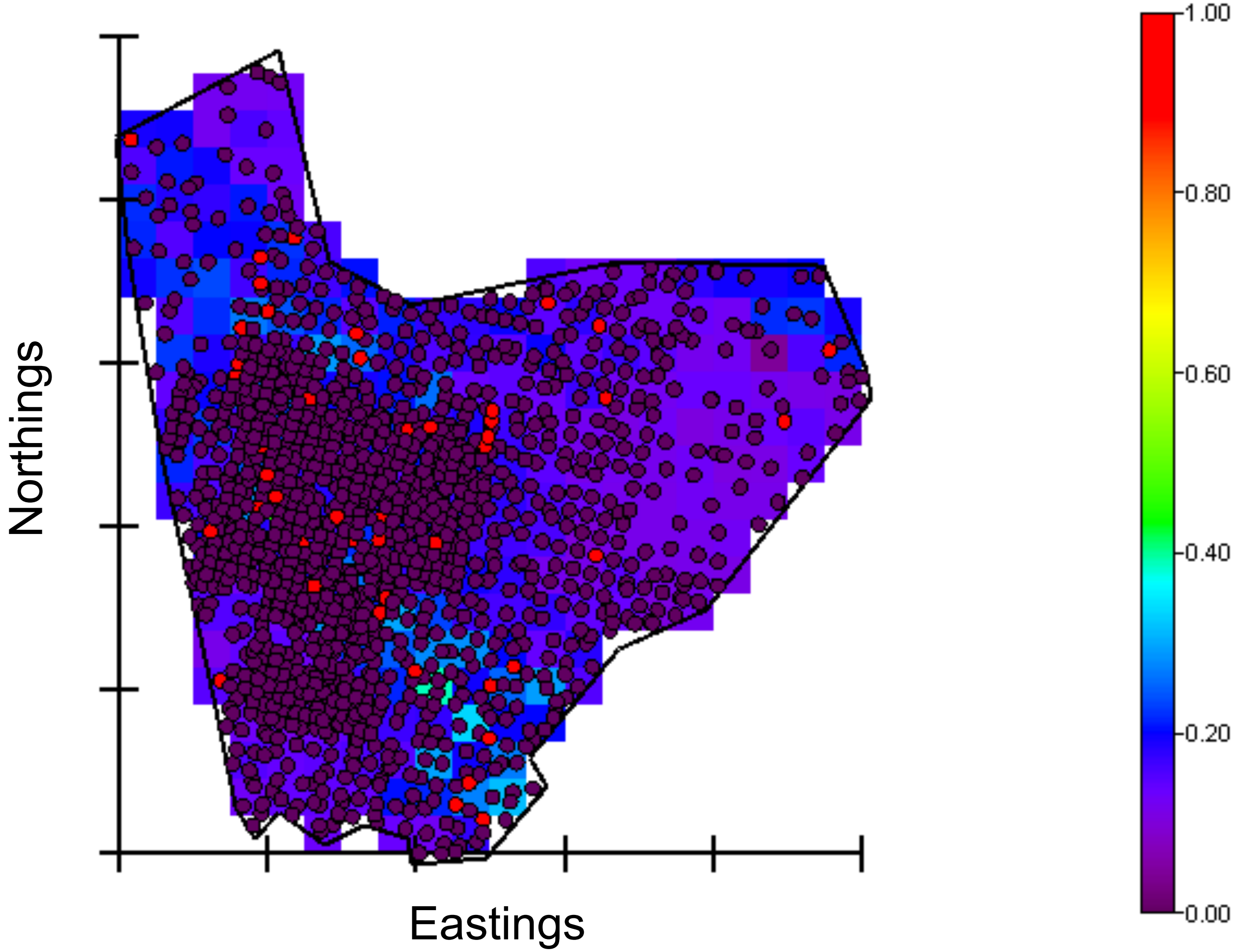
- Spatial analysis of contamination data
  - uses a form of contouring known as Kriging
  - underpinned by a model of the spatial correlation of concentration
  - usually a Bayesian approach to update an empirical or theoretical prior
  - method of estimating the variable, its mean and uncertainty



Geostatistics



Block Kriging of Probability of Mean Value Exceeding a Threshold

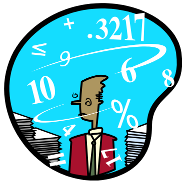


Data exploration

15 July 2015

AECOM

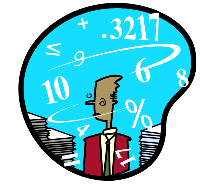
Data exploration - refinement of CSM



- Analysis of population statistics
  - data quality assessment
  - analyse sub-populations spatially, by depth and stratigraphic unit
  - identification of outlier values and possible ‘hotspots’
  - analysis of uncertainty, data modelling and simulation
- Interpretation of contaminant source
  - regression analysis between chemical species
  - forensic fingerprinting and aging of TPH, and PAH double-plots
  - principal component analysis (PCA)
  - cluster analysis e.g. K-means
- Geochemical analysis
  - bioavailability studies
  - sequential extraction (CISED)
  - mineral determination (e.g. X-ray crystallography)

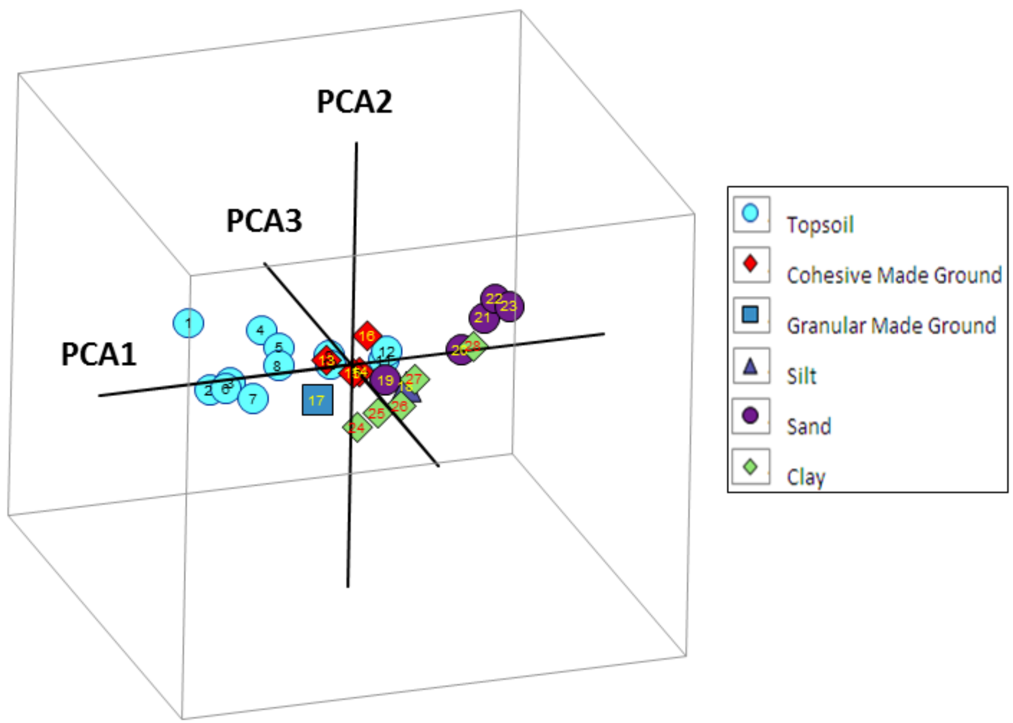
AECOM

Principal Component Analysis



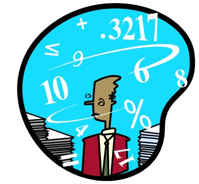
- Statistical procedure to describe the difference between samples in a set of factors or "principal components"
  - each successive factor calculated explains the maximum of the remaining variance in the dataset
  - typically n>3 analytes are reduced to 2- or 3-dimensions
  - enables identification of natural groupings between samples

Analytes (12)
Arsenic
Cadmium
Chromium
Copper
Mercury
Nickel
Lead
Selenium
Zinc
Vanadium
Sulphur
pH



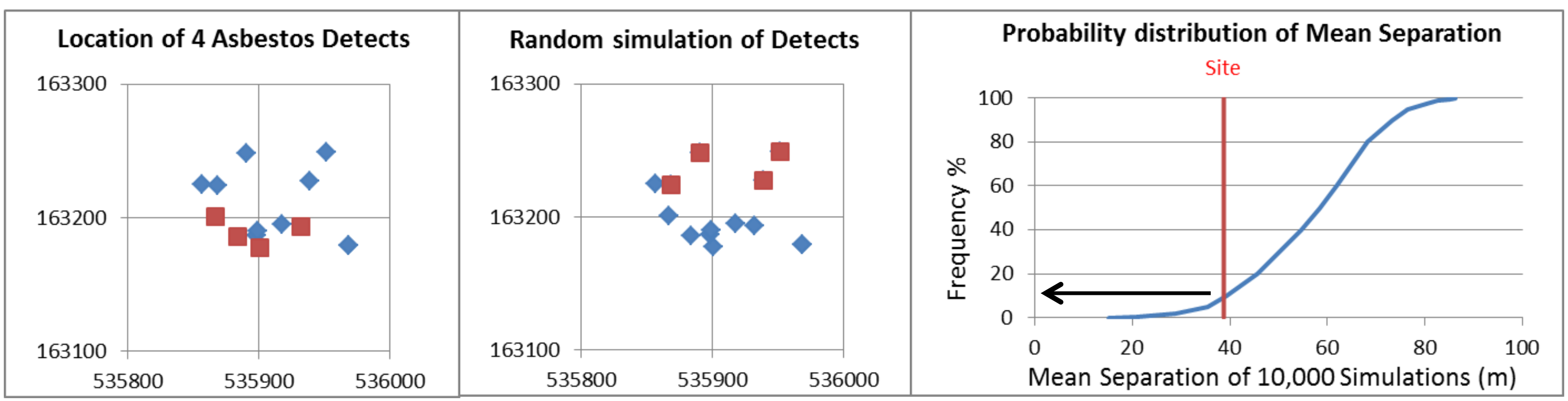
AECOM

Probability and simulation



- Generation of alternative computer models based on the site investigation data to map out the probability space for possible interpretations of the data
  - methods such as sequential gaussian simulation, or simulated annealing
  - provides confidence intervals for predicted values such as means
  - may be useful for assessing alternative CSM

- Simple example to model alternative distributions of asbestos.
  results of 10,000 realisations indicates <10% probability of this cluster occurring by chance.



Advanced methods – software tools

15 July 2015

AECOM

Software for design of sampling plans

- Public domain software is available for the design of statistically sound sampling plans and decision support, examples are:

  SADA Spatial Analysis and Decision Assistance (University of Tennessee)

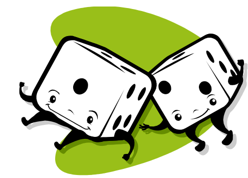
  http://www.sadaproject.net/

  Visual Sampling Plan (Pacific Northwest Laboratories)

  http://vsp.pnnl.gov/

- They include methods for design of grids, confidence limits for population statistics, estimation of uncertainty, geospatial modelling, geospatial simulation, cost benefit analysis, adaptive sampling, judgemental sampling, visualisation etc.

Conclusions



- The CSM and Site Investigation are intrinsically linked
- A preliminary CSM is a requirement for planning a site investigation
- The greatest scope for error in decision making is an incorrect CSM
- Site investigation is necessary to confirm the CSM and reduce uncertainty
- Site investigation is inexact due mainly to soil variability
- Large and complex sites are more efficiently investigated in stages
- Statistical methods can reduce uncertainty but they rely on unbiased data
- The objective of site investigation specifically for exposure risk assessment is to reduce the uncertainty for decision making at the scale of the averaging area.

Thank you for your attention…….

Remember the more you want to get out of statistics the more you have to put in.