Site-ASSESS: A tool for visualising and interpreting *a priori* information for the assessment of contaminated sites

Colin C. Ferguson[1], Peter Tucker[2], Ammar Abbachi[1] and Paul Nathanail[1]
[1]Centre for Research into the Built Environment
The Nottingham Trent University,
Nottingham NG1 4BU, UK.
Tel: (44) 115 9418418, fax: (44) 115 9486510
E-mail: *paul.nathanail@ntu.ac.uk*

[2]University of Paisley
High Street, Paisley PA1 2BE, UK
Tel: (44) 141 8483205, fax: (44) 141 8483204

Abstract
The investigation of potentially contaminated sites requires the integration of diverse spatial information to develop a conceptual model of the ground conditions and history of previous site use. This information is used to develop spatial sampling strategies.

An integrated visualisation system called *Site-ASSESS* has been developed to help site assessors compile and display *a priori* information collected during desk study and walkover surveys of potentially contaminated sites. A digital image of the site is used to provide a framework within which indicators of potential contamination (e.g. leaking tanks, waste storage areas or stressed vegetation) are mapped. A statistical approach based on Bayesian theory is used to design sampling strategies (i.e., number of samples and sample locations) to locate hotspots of a given size with prescribed degree of confidence. The total number of sample locations is then computed, locations are optimised to minimise overlapping areas of influence and distributed over the site to reflect the prior information. The user is presented with a map of the site showing the recommended sampling locations.

The performance of the system in a number of case histories has shown that it facilitates consistency and rigour in designing sampling strategies and assists in reporting the basis on which the investigation was carried out.

Introduction
Pressure to redevelop former industrial land and to avoid developing greenfield sites has increased the need for better site investigation tools. An important component of an environmental risk assessment is determining the nature, extent and significance of soil contamination, and how to manage it. Sufficient information of acceptable quality is required to arrive at defensible decisions based on the likely positions of contaminant hotspots and the estimated spatial distribution of contaminants in soil. Compilation and visualisation of available information is of paramount importance in designing spatial sampling strategies for site investigation. Usually, large amounts of information gathered during desk study and walkover surveys are not effectively used in sampling design. Information technology methods, such as GIS and knowledge-based techniques, provide flexible and integrated tools for visualising and analysing large volumes of data.

The *Site-ASSESS* decision support system, developed under contract for the UK Department of the Environment, is designed to streamline site investigation planning and increase the likelihood of collecting appropriate data in a cost effective way.

The system aggregates information relevant to possible soil contamination and thus indicates the areas most likely to contain hotspots. The indicators are then used to calculate prior probabilities that contaminant hotspots exist and hence to design spatial sampling strategies.

A case study illustrating the use of *Site-ASSESS* to design a first stage sampling strategy is presented. A desk study report (ICC 1995) and site walkover survey were used to compile a list of weighted indicators of likely contamination. The user enters this preliminary information into *Site-ASSESS* in order to help develop an initial hypothesis on the location of suspected hotspots. Then a cost effective spatial sampling strategy is designed using this initial hypothesis. The number of samples required is expressed as a function of hotspot size, and sampling locations reflect the spatial distribution of the weighted indicators.

Using Prior Information
Regular sampling designs, which assume that all parts of a site have an equal probability of containing a hotspot, can lead to large numbers of sampling points (Ferguson 1992). More cost-effective sampling strategies can be devised when there are grounds for relaxing the equiprobability assumption and concentrating investigative efforts in areas suspected of being contaminated.

Experienced assessors will often suspect that some parts of a site are more likely to contain a hotspot than others. They may then wish to design a sampling strategy that reflects their degree of suspicion or strength of belief about the

target's likely location. One approach to this type of sampling design is to partition the site into subareas and then to score subareas (say, on a 1 to 10 scale) to reflect strength of belief as to where the target is likely to be.

Figure 1 shows an example of this type of scoring, and the corresponding sampling plan. The scoring scale is arbitrary, the highest score being used as a normalising factor so that strength of belief is expressed relative to that in the most favoured subarea. Sampling density for the most favoured subarea is calculated using the Monte Carlo method to give 0.95 probability of hitting a target if it exists in that subarea (Ferguson 1992; Ferguson & Abbachi 1993; Department of Environment 1994a). Other subareas carry lower sampling point densities to reflect the assessor's lower expectation that the target is located in these areas.

If the assessor's judgement is correct the 0.95 probability of success is thus achieved with fewer sampling points and lower cost. But if the assessor is wrong the penalty is a reduced probability of success in locating the target (see probability values adjacent to Figure 1B).

The problem with this approach is that it requires a site investigator to convert a variety of disparate information into a score reflecting his or her strength of belief about hotspot location. To overcome this difficulty a computer-based decision support system has been designed so that specific items of information derived from a review of site history and from a preliminary walkover survey can be used directly to optimise sampling designs. *Site-ASSESS*, described below, has been developed to help professionals make informed judgements about sampling designs.

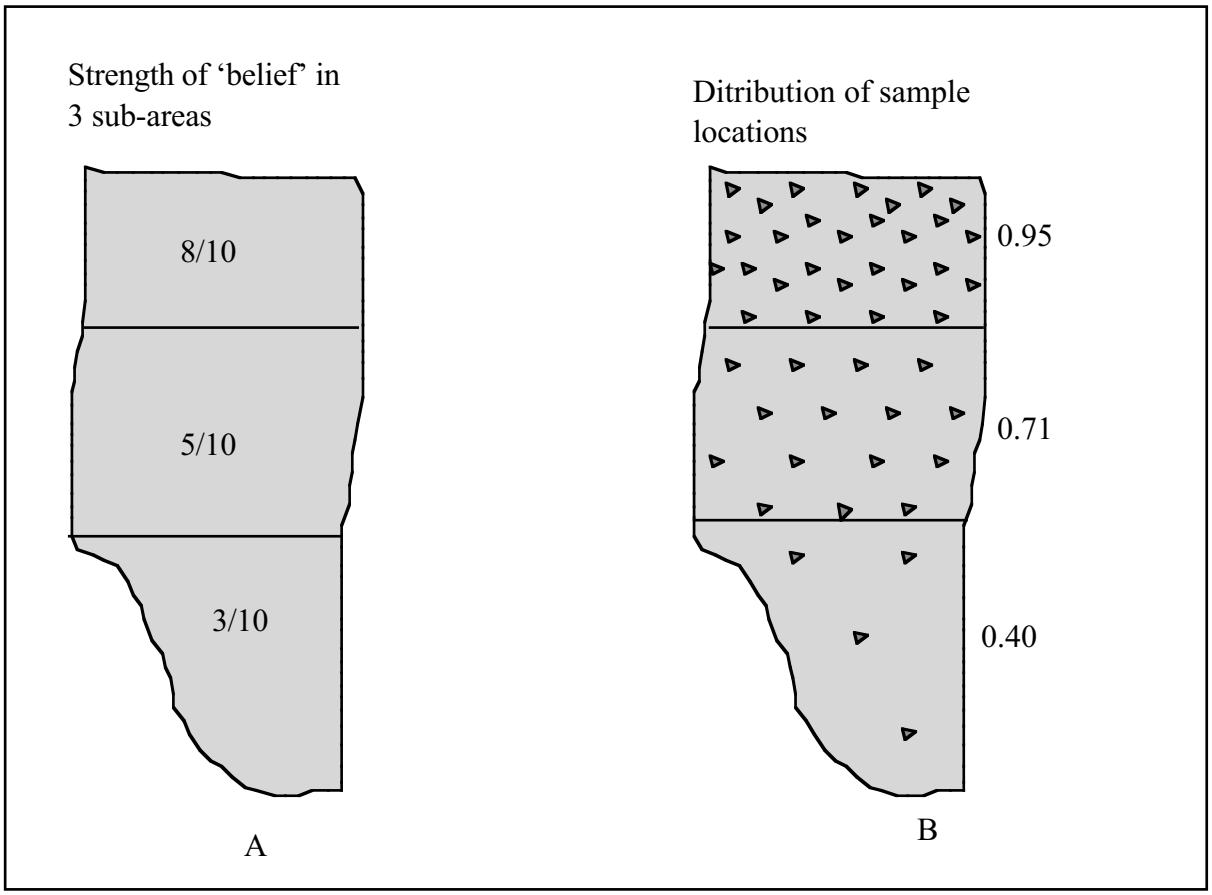
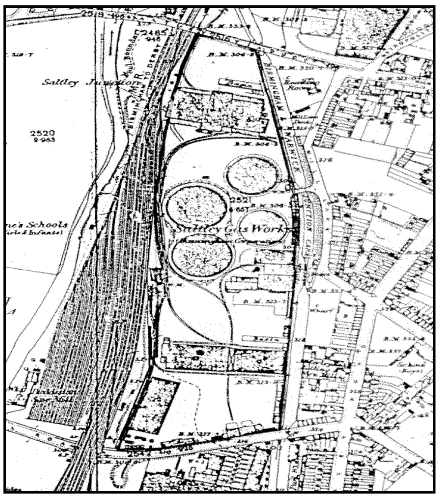


Fig. 1: Variable density sampling for different strengths of belief. The probabilities shown on the right of subareas in B are hit probabilities *if* the specified hotspot exists in the relevant subarea. But all subareas have the same *a posteriori* probability as explained in the text.

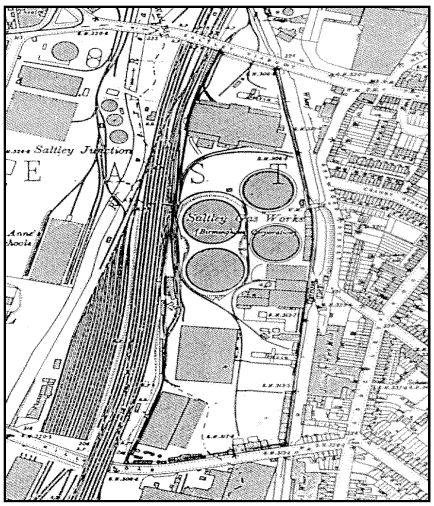
THE *Site-ASSESS* DECISION SUPPORT SYSTEM

Site-ASSESS (Assessment of Sampling Strategies Expert Support System) is a decision support system for the design of sampling strategies for contaminated sites. Eventually the system will comprise the following linked modules:

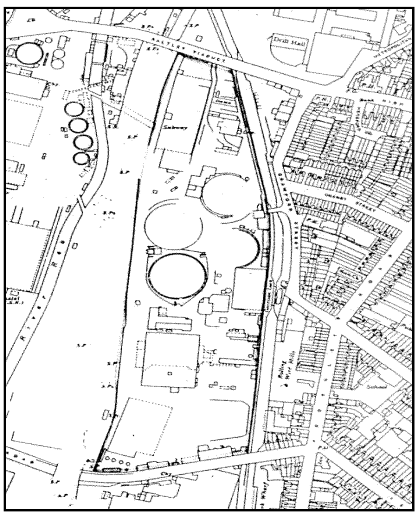
* Likely contaminants
* Location of contaminant hotspots
* Soil gas survey
* Groundwater sampling
* Data analysis of first-stage sampling results
* Design of second-stage sampling



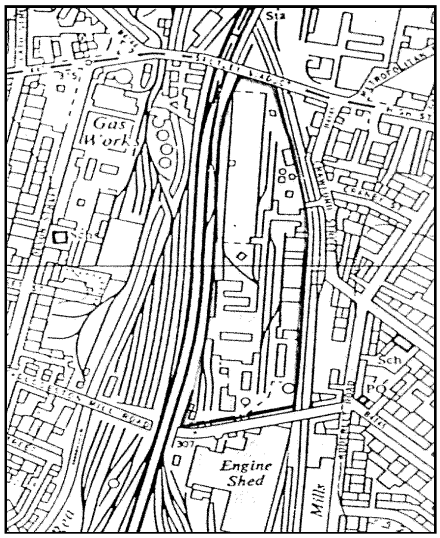
A:1890



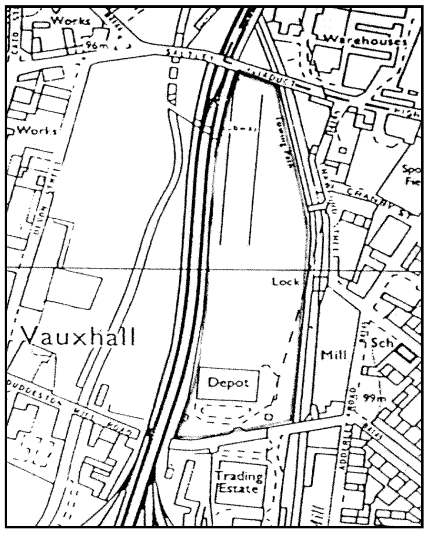
B: 1905



C: 1937



D: 1967



E:1990

Fig. 2: Different Ordnance Survey maps of the site with gas works in central area (A: 1890, B: 1905 and C: 1937) replaced with new infrastructure (D: 1967) and finally with the infrastructure cleared (E:1990)

In version 1.0, which is currently being field-tested, only the hotspot location module (*Site-ASSESS* Hotspot) is fully developed.

The *Site-ASSESS* Hotspot module is intended to help site investigators design the initial, or first stage, sampling of a site. *Site-ASSESS* Hotspot develops an optimised sampling pattern for detecting (though not delineating) hotspots, and provides a statistical justification for the chosen strategy. For the purpose of locating a hotspot it is sufficient to place just one sampling point on the area covered by a hotspot. Therefore if a circle equal to the radius of the putative hotspot is drawn around a sampling point, the circle can be thought of as the zone of coverage for that sampling point.

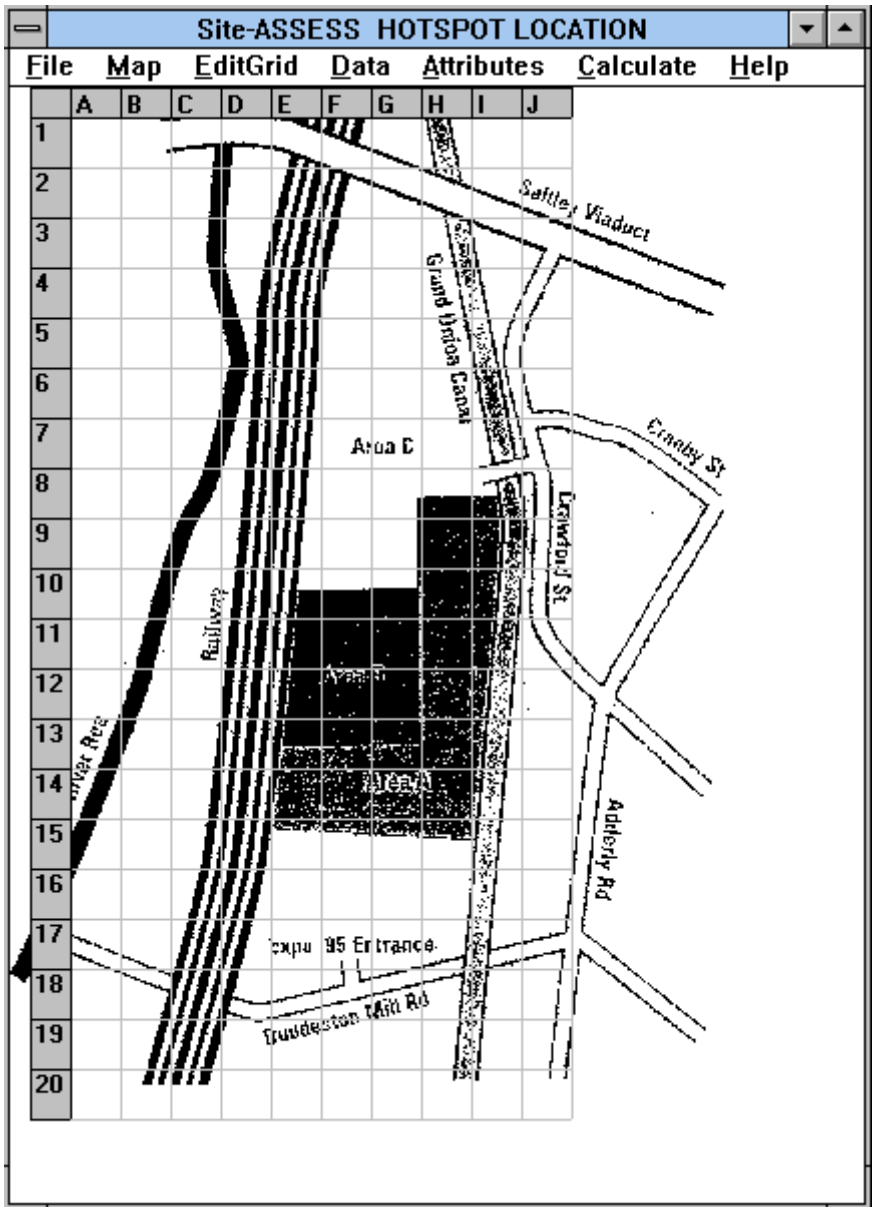


Fig. 3: Scanned map of the site imported into *Site-ASSESS* with 25 x 25 m information cells superimposed.

The total sampling coverage is the sum of all zones of coverage excluding overlaps. The ratio of the total coverage to the total site area can also be thought of as the probability of hitting a hotspot if it exists. *Site-ASSESS* is developed around this concept of coverage and hit probability. The computational part calculates an optimum sampling coverage, covering areas with the strongest prior evidence for the existence of a hotspot.

Before running *Site-ASSESS* Hotspot the user should have already completed a preliminary survey of the site and have compiled:

(i) data on the historic use of the site
(ii) a record of any visual, or other, indications of potential contamination gained during a site walkover (Department of Environment 1994b)

The user may additionally have some preliminary chemical analysis data from previous investigations or from *ad hoc* sampling of the site.

Archive data on previous site use (chemicals and processes) should, when possible, relate to the length of time in use, time since last use, chemicals present, quantities of chemicals handled and the properties of these chemicals, specifically toxicity, biodegradation potential and leaching potential. It is recognised that these data are often very difficult to obtain precisely. In *Site-ASSESS* high precision is not required. The user may simply make a best judgement according to a 3-point scale “High”, “Low” or “Unknown”.

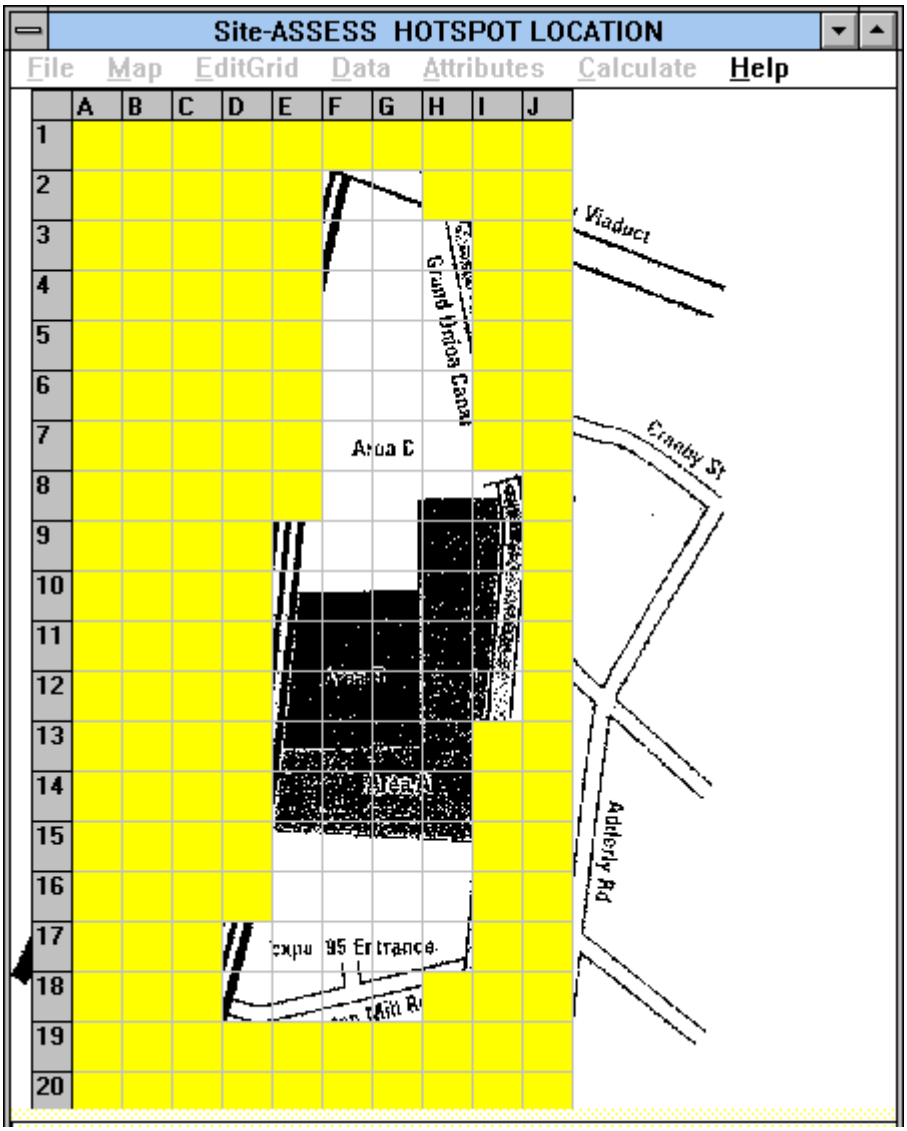


Fig. 4: Information cells outside the site eliminated

It is helpful when conducting a walkover survey to subdivide the site into a number of square cells and to compile the visible or other indicators for each cell. These cells represent an information grid (see Figures 3 and 4) into which all prior information is aggregated. Users must specify the size of the information cells, balancing spatial resolution with the time required to input all the information cell by cell. Too fine a grid rarely leads to a significantly improved solution but substantially increases data entry and computation time. On the other hand , too coarse a grid may fail to resolve some of the spatial complexity of contamination.

The presence or absence of an indicator (or attribute) in a given information cell is registered in terms of a score allocated to that cell; scores are then summed for each of the individual cells.

These total scores can be viewed on the screen (Figure 6), their values being rounded to the nearest integer for display. A high score indicates a high *a priori* probability (i.e. strong evidence that a hotspot exists) and therefore the need for a relatively high sampling density in order to locate it with confidence. Attribute scores are converted to *a priori* probabilities by the user specifying the probabilities he or she thinks most appropriate for the highest and lowest scoring information cells; intermediate scoring cells are scored proportionately. Default values are provided in *Site-ASSESS* as initial guesses. The resulting sampling strategies are not usually very sensitive to the choice of initial guesses.

Estimating Local Sampling Densities

Analysis of the prior information provides a spatially distributed set of *a priori* probabilities that a hotspot exists within specified subareas of the site defined by the information grid cells. The following approach is used to convert the *a priori* probability assigned to each information cell (grid square) into a target number of samples for that grid square.

The primary motivation for a sampling scheme is to ask; “What is the probability of locating a hotspot *if it exists*?” If, however, the sampling scheme fails to locate a hotspot, the question then becomes: “What is now the probability that a hotspot exists given that the sampling scheme has failed to find one?” This (after the event) probability is termed the

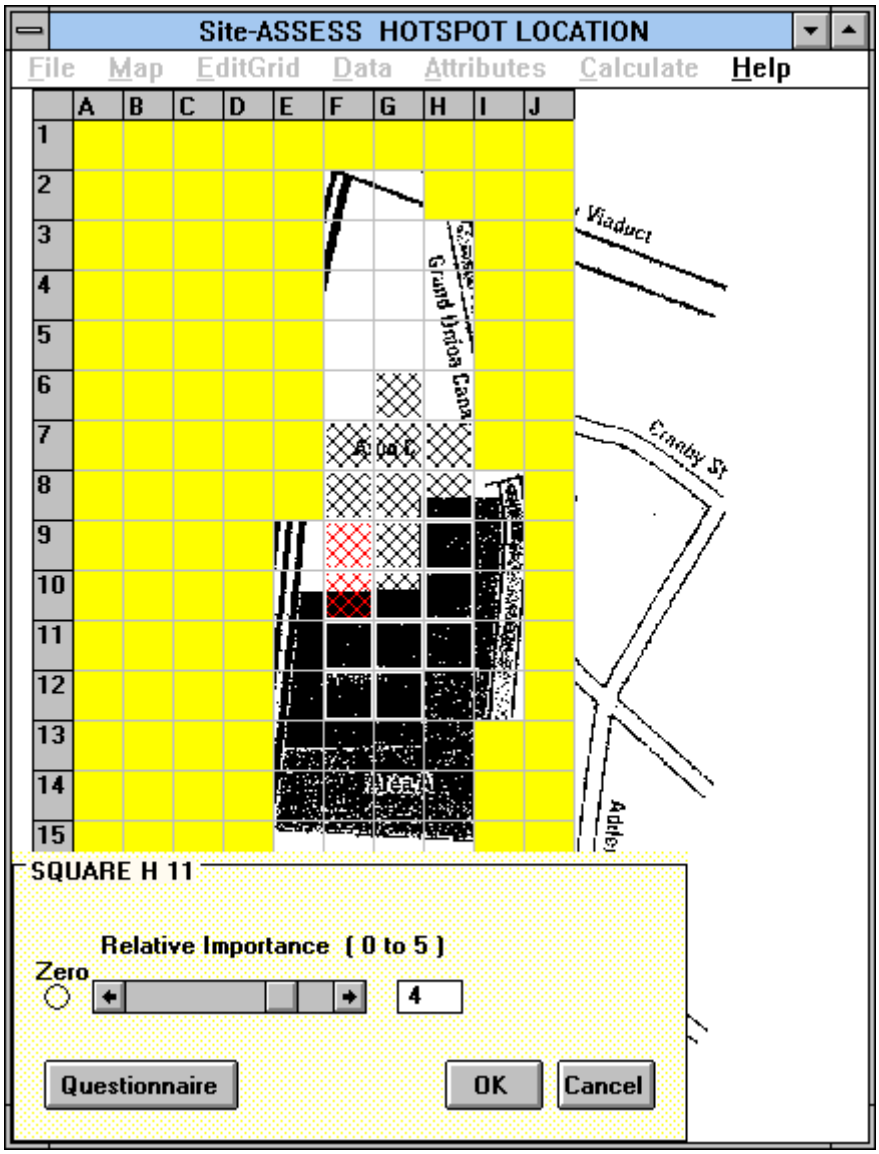


Fig. 5a: Attributes allocation and different scoring systems (expert judgement)

a posteriori probability. The probability of locating the hotspot, if it exists, can be considered as the hit probability which is equal to the sampling coverage as discussed above.

Bayes Theorem (e.g. Johnson 1994) has been adapted to relate the above probabilities:

$$PrH_i = 1 - PrA_i\,(1 - P_i)/(P_i\,(1 - PrA_i))$$

where PrH_i is the hit probability for grid square i, P_i the *a priori* probability and PrA_i the *a posteriori* probability. By setting one of the sampling objectives (i.e. PrH_i or PrA_i) to a fixed target value, the above equation can be used to compute the other probability. We generally set all PrA_i to 0.05, i.e. if, after sampling, a hotspot has not been located in any given square, there is 95% confidence that a hotspot *does not exist* within that square. The value 0.05 is not prescriptive although it provides a confidence level with which most users seem to feel comfortable.

The hit probabilities PrH_i for each grid square are used to calculate the nominal number of samples N_i allocated using the relationships between sampling density and hit probability established by Ferguson (1992). The problem then is how to distribute the total number of samples $N_T = \Sigma N_i$ over the site such that each individual requirement on N_i is satisfied as nearly as possible.

Optimisation of Sample Locations

The given number of samples, N_T, needs to be placed such that their aggregate weighted coverage is maximised. Equivalently, the problem is to minimise the aggregate *a priori* probability score of all parts of the site that fall outside the zones of coverage of the sampling points. This is a well-defined optimisation problem that can be solved using the Quasi-Newton method (Fletcher 1987). However, this is a local optimisation procedure and in practice, a good initial estimate was found to be essential.

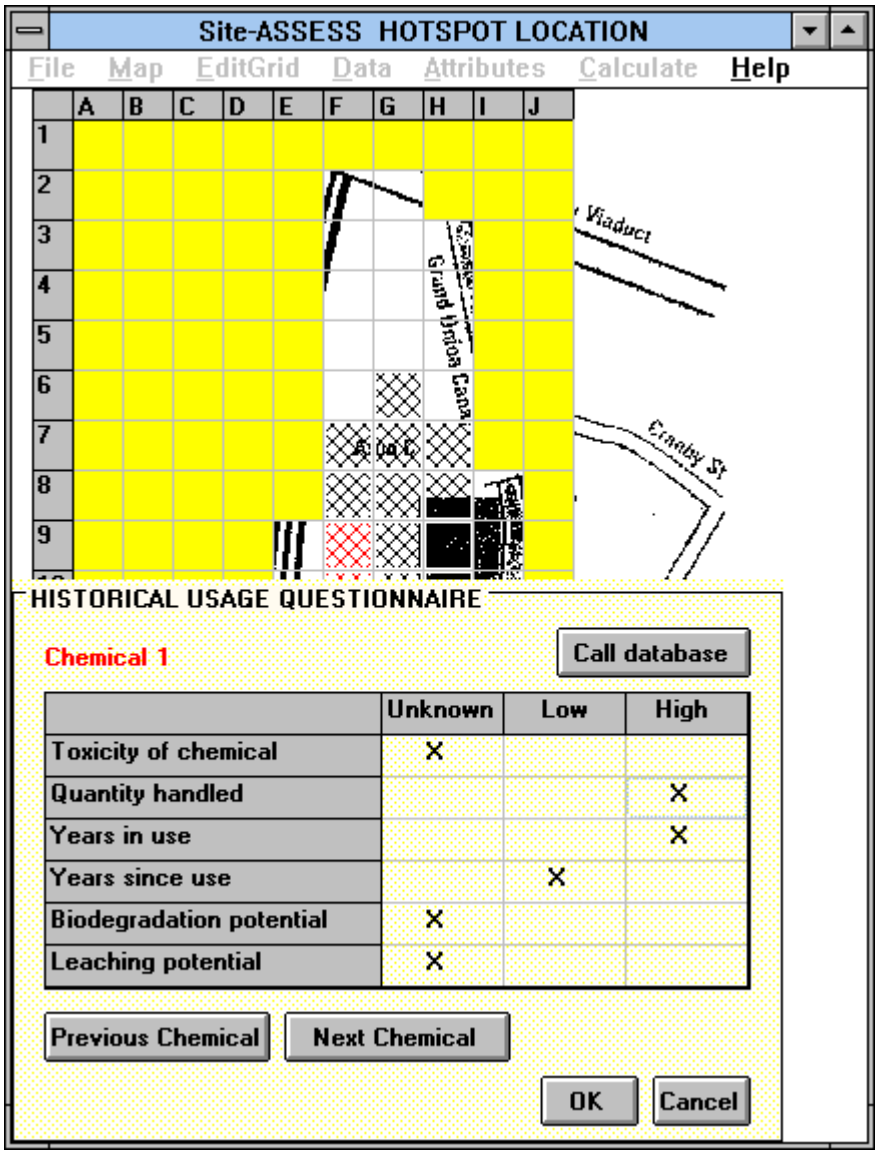


Fig. 5b: Attributes allocation and different scoring systems (B: using knowledge base)

The initial estimate is found using an approximate sequential placement in which samples are placed according to the rank order of the *a priori* probabilities. To improve the accuracy of this discretized approach, each information cell is subdivided into a fine grid, typically comprising 3x3 or 4x4 smaller grid squares. The centres of the fine grid squares define the set of possible sample locations in the approximation. The *a priori* probability scores of all the fine grid squares are first placed in rank order. When fine grid squares have the same score, a possible sample location whose zone of coverage lies wholly within a high probability subarea is ranked higher than one whose zone of coverage overlaps into an adjacent lower probability subarea. More generally, the rank order is based on the average *a priori* probability over the whole zone of coverage rather than the probability at the sample placement point. Any remaining ties in rank order are broken arbitrarily. In practice the sequential solution usually performs almost as well as the optimised solution, which typically improves the overall weighted sampling coverage by less than 5%.

Case Study

The initial objective in the case study was to design a preliminary sampling strategy to locate suspected hotspots and to provide an overall picture of the spatial distribution of soil contaminants within the former industrial site.

The site used to illustrate *Site-ASSESS* extends over approximately 4#ha and has been the location of several past industrial activities. Figure 2 shows survey maps of the site at different periods. The presence of a gas works in the central area (from 1890 onwards) is regarded as a subarea with high potential for containing contaminant hotspots. Loading and off loading areas are not accurately recorded and could be anywhere next to the railway tracks to the west, or the road to the south of the site. The canal junction to the east of the site may also have been used as a loading and off-loading area, as waterways were in use until the middle of this century.

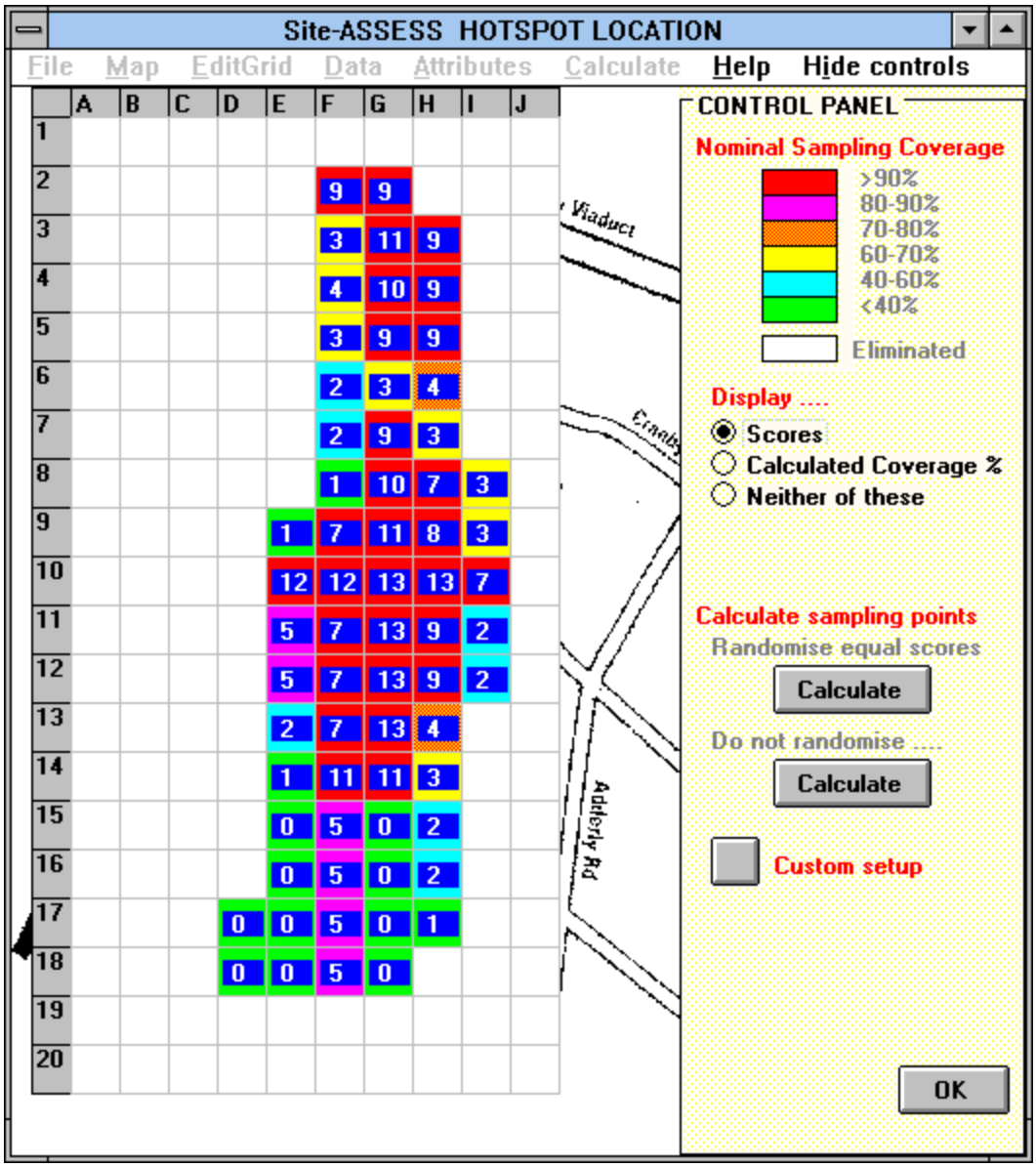


Fig. 6: Scoring results (high scores in central area).

A full search of historical data for this site would be time consuming and reporting it would be beyond the scope of this paper. However, a list of possible attributes has been compiled.

The following are the historical attributes used:

- Process areas
- Storage areas of raw materials
- Waste disposal areas
- Loading and off-loading areas
- Filled areas

In addition, a site walkover indicated the existence of the following attributes:

- Irregular surface
- Poor drainage
- Anomalous soil type
- Oily patches
- Bare areas with sparse vegetation
- Remains of site infrastructure
- Waste tips

The scanned map of the site was imported into *Site-ASSESS* and divided into 66 25m x 25m information cells (Figure 3). Although a greater number of information cells will result in higher resolution it will considerably increase the amount of time required to input all the information. Cells outside the site have been eliminated as shown in Figure 4.

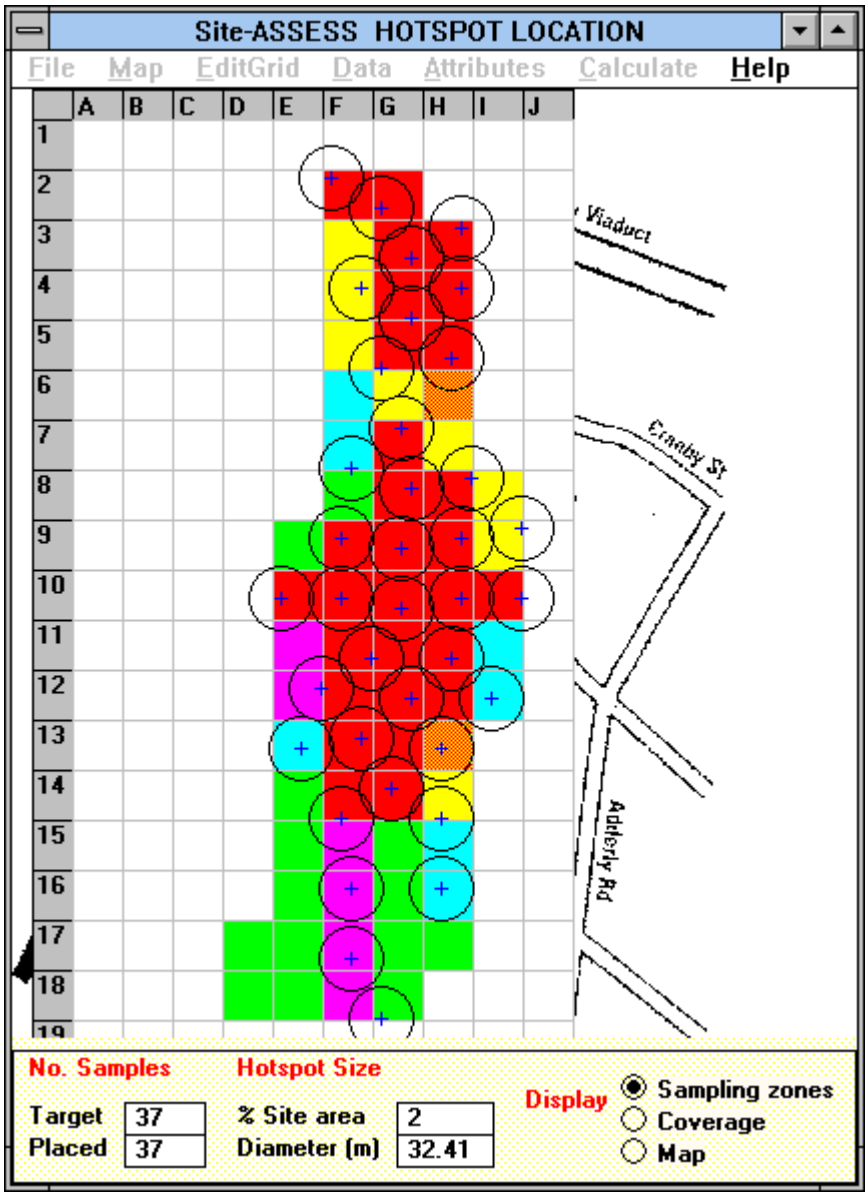


Fig. 7: First stage sampling strategy with clustering in central area.

Figure 5 shows two ways of allocating attributes to information cells and working out their influence on the final score. Figure 5A shows that user judgement on the importance of an attribute can be set using a sliding scale to specify a score in the nominal range 0 - 5. Alternatively, Figure 5B shows a menu which allows site assessors to respond to simple questions; the answers are then used to work out a score for the designated information cell. The user can overrule the knowledge base if specific data on items such as quantity handled, years in use, years since last use and leaching potential, are available.

The nominal hotspot size can be varied to study its impact on the number of sample locations required. The hotspot size, assumed of circular shape, is expressed as a percentage of the total site area. The scores shown in Figure 6 are then reviewed and cross-checked with the assessor's strength of belief on levels of contamination in differents part of the site. The central area is most suspected of being contaminated, and was the location of a gas works for many years (see Figure 2). The far north-east part of the site is also of high *a priori* probability; it is suspected to be a site where waste material was deposited.

The spatial distribution of the samples is characterised by higher sampling density in the central areas and the north-east part of the site (Figure 7) reflecting the higher *a priori* probabilities as described above. In broad terms *Site-ASSESS* therefore outputs the sorts of sampling pattern that an experienced sampler would produce. But it does so in a more consistent and reproducible manner that is underpinned by a statistically defensible methodology.

Using *Site-ASSESS* the assessor is able to compare sampling strategies designed on the basis of prior information with strategies based on the equiprobable assumption (Ferguson 1992). The reduction in total number of sample locations is apparent especially at smaller hotspot sizes (Figure 8); for this site roughly a 30% reduction in total number of sample locations is achieved. Of course this reduction is obtained at the expense of relaxing the hit probability in some areas of the site. Site investigators should make their own judgement as to whether prior information gathered on a site is robust enough to give confidence in the sampling strategy adopted.

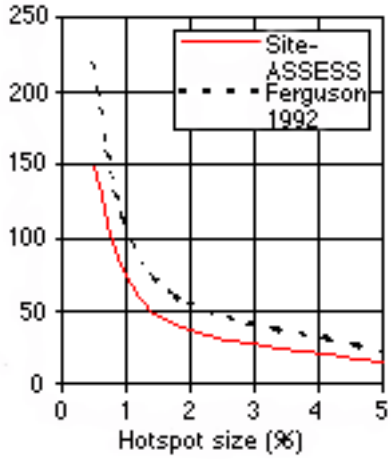


Fig.8: Number of sample locations as a function of hotspot size, (expressed as a percentage of the total site area) to achieve a 0.95 probability of hitting a single circular hotspot if it exists.

Conclusion

The system developed here provides a standardised framework for spatial data handling and qualitative reasoning for sampling strategies design. It uses readily available prior information acquired during a desk study and walkover survey. It should be noted that many practitioners fail to use these data in sampling design (although the data have been collected at some expense) and recourse to a regular grid pattern is the norm for site sampling. Sampling and analysis costs may amount to hundred of pounds for each sample taken, so minimisation of sample numbers is important. The subsequent costs of missing a contaminant hotspot through insufficient sampling could, however, be orders of magnitude higher.

The system, based on GIS techniques, statistical methods and expert knowledge, provides a pioneering approach for handling and analysing spatial data to design cost-effective sampling strategies. The purpose of such strategies is to test the conceptual models developed on the basis of phase 1 information (desk study and preliminary walkover survey). Of course, decisions on the need for further investigation or remedial actions are made after analytical results are available and are used, with other information, in an appropriate risk assessment.

Acknowledgement

This work was funded by the UK Department of the Environment but the views expressed are those of the authors and do not necessarily represent those of the Department. We are grateful to British Gas Properties for permission to publish the case study.

References

Department of the Environment (1994a) Sampling Strategies for Contaminated Land. CLR Report No 4., Department of the Environment, London

Department of the Environment (1994b) Guidance on Preliminary Site Inspection of Contaminated Land. CLR Report No 2., Department of the Environment, London

Ferguson, C.C. (1992) The statistical basis for spatial sampling of contaminated land. *Ground Engineering 25/5*, 34-38

Ferguson, C.C. and Abbachi, A (1993) Incorporating expert judgement into statistical sampling designs for contaminated sites. *Land Contamination & Reclamation 1*, 135-142

Fletcher, R.O. (1987) *Practical methods of optimisation*. Academic Press, London

ICC (1995) Search for Contaminative Uses Report at EXPO'95 Site. Available from ICC Information Group Ltd., 16-26 Banner Street, London EC1Y 8QE.

Johnson, R. A. (1994) *Miller and Freund's Probability and Statistics for Engineers*, Fifth Edition. Prentice-Hall, Englewood Cliffs, New Jersey.