# TOWARDS A METHODOLOGY FOR SELECTING A 'CHARACTERISTIC' SAMPLE FROM AN EXISTING DATABASE: AN EVOLUTIONARY APPROACH

Jarvis, C.H.*, Stuart, N.*, Kelsey, J., Baker, R.H.A.

## ABSTRACT

The performance of an environmental model depends highly on the 'representativeness' of the sample data with which it has been developed in relation to the overall data set. Additionally, the adequacy of the sample in relation to the modelling technique itself is important. While advances have been made towards achieving both goals individually, in few cases are these combined. Additionally, the situation with regard to pre-existing data at fixed points remains 'ad hoc'. Given that increasing volumes of pre-sampled point data now exist in electronic form, these omissions are an issue of currency within the wide range of topics embraced by the term 'environmental modelling'. That this increase in data richness has frequently been matched by the rising cost of data means the ability to select appropriate data is paramount to the economic feasibility of the overall project itself. A new sampling methodology is presented to deal with these issues, framed within the context of the choice of meteorological station data for the development of interpolated weather surfaces but of wider applicability.

Fundamental to the new strategy is the representation of sampling requirements by means of a multi-criteria function. Criteria are developed to deal with each individual variable, for example a desired height function/range or the maximisation of record length. Data selection may also be tailored to the model types for which the data is to be used, for example by ensuring that an adequate number of nearby points are selected for interpolation by kriging. This combination of possibly contradictory criteria are then optimised together using a genetic algorithm in conjunction with the geographical information system in which the data are stored. While the solution of multi-criteria problems can be achieved using a variety of more traditional techniques, the GA approach was chosen both for its ability to manage large volumes of data and for its management of compromise. As a working progress report, discussion focuses on the reasons for the choice of an evolutionary approach and the representation of the sampling problem within this framework.

## INTRODUCTION

However sophisticated a model or procedure, the results will only be as good as the data and knowledge underpinning its conception. This statement in itself is not new or radical and a number of sampling methodologies are very familiar in a wide range of disciplines. Such techniques include for example the random, stratified or systematic families of sampling. More recently these familiar methods have been augmented by those such as gradsect sampling (Austin & Heyligers, 1989) which deliberately sample across steep gradients of environmental change. Such moves highlight an increased awareness of the

need to maximise the possible environmental range within a sample. The importance of work by biologists Margules & Stein (1989) which stresses the limitation of sampling within one dimension only has also been noted within a geographical framework (Lees 1994, Aspinall & Lees 1994). The latter argue that not only should sampling be carried out with respect to full range of environmental criteria, but also that each environmental space should be sampled individually. Further dimensions to the sampling problem that are difficult to incorporate within the context of traditional sampling methodologies include the purpose for which the data is to be used (e.g. Pettitt & McBratney 1993, variogram production), together with strategies which formally take into account the cost of sampling within the overall sampling objective (e.g. Bras and Rodriguez-Iturbe, 1976). What may be summarised from the current literature therefore are the concepts of *representative* data, the need to look at *multidimensionality* and the possibility of tailoring our sample according to future *analytical requirements*.

In the case of choosing the most characteristic *pre-located* site data from a wider set, as distinct from a free choice of sample sites over a particular landscape, the means by which *all* of the principles discussed above may be applied within the sampling process is somewhat ad hoc. Such a situation is however frequently to be found with the increasing volumes of pre-sampled data available in electronic form, the task of choosing relevant data becoming commensurably awesome. Even when a restricted and relevant data set exists, its full cost may deem a project economically not viable and therefore a means of partial selection critical. Applying the definition of Eastman (1995) which classifies problems with a single objective (making an 'appropriate' sample choice) subject to a number of possibly conflicting criteria (representative data, multidimensionality, analytical requirements, cost) as multi-criteria issues, established techniques for solving such tasks are used as a base for further methodological discussion. It should be noted that some ambiguity in the use of such terminology arises even within a GIS environment, Jankowski (1995) referring to *both* multi-criteria and multi-objective techniques as 'multiple criteria decision making methods'.

A wide variety of traditional optimisation and search techniques exist which have been drawn upon in the development of 'multiple criteria decision making methods' (Table 1). As Schwefel (op cit, p165) notes, “the question of which is the best strategy is itself a kind of optimisation problem” ! Just as multi-objective analysis has tended to be viewed as a 'natural extension of mathematical programming' (Jankowski 1995), so have the overlay and hierarchical methods of multi-criteria analysis found favour within the context of a geographical information systems (e.g. Carver 1991, Eastman et al 1995).

- Gradient analysis
- Enumerative
  - Mixed Integer Programming
  - Dynamic Programming
  - Overlay methods
  - Hierarchical methods
- Random
  - Random walks

Random search & save

**Table 1, Conventional search and optimisation methods**
(After Goldberg, 1989, p3)

While a straightforward solution to the multi-criteria sampling problem was sought, a number of disadvantages in using these conventional methods were identified. These are summarised in Table 2 below.

**Gradient analysis:**

- Search spaces are often calculus unfriendly i.e. non-differentiable and are difficult to define.
- Local in scope (e.g. calculus (or gradient) methods) and may therefore converge prematurely at a non global solution
- Not designed with multiple solutions in mind, which occur with competing objectives, and are therefore devalued as a decision support tool
- Often highly tailored to solve a particular problem, so both may not be easily adapted and fail to enhance the GIS in other potential application areas
- Not adaptable to situations with multiple solutions without many runs (Fonseca & Fleming, 1995)

**Enumerative methods:**

- Solutions problematic to understand with increasing number of criteria (e.g. overlay analysis, Janssen & Rietveld 1990)
- Loss of information using thresholds with continuous data & problems in defining thresholds in overlay analysis (Janssen & Rietveld op cit.)
- Unable to compromise between conflicting criteria (e.g. hierarchical methods)
- Not guaranteed to find a solution (e.g. Overlay/weighted linear combination methods or hierarchical ranking methods may render the entire search space 'unsuitable' )
- May not address combinatorial side of a problem (e.g. overlay/hierarchical methods ignore geographical relationship *between* possible sites, important in sampling context)
- Unable to cope with large volumes of data (incl. dynamic programming, Schwefel 1995 p12)
- Not necessarily adaptable to situations with multiple solutions in mind (e.g. simplex, Fonseca & Fleming 1995)

**Random search:**

- Unable to cope with large volumes of data (choosing 200 sites from a total of 985 has 985!/((985-200)!*200!) possible combinations)

**Table 2, Restrictions of conventional algorithms**

In addition to conventional search and optimisation algorithms, a newer class of methods known as evolutionary algorithms (incl. genetic algorithms) has also been used in multi-criteria optimisation (Goldberg, op cit, p197). Because of the major obstacles emerging should traditional methods be used in the case of the sampling problem (search space size, the objective as a collective of suitable sites, and conflicting criteria selection) these more

recent techniques are evaluated for their potential use in the sampling application. The results of this review are shown in Table 3, from which the decision to develop a genetic algorithm approach to the extraction of a characteristic sub-sample of fixed data points from an established database was taken.

- Makes use of guided but randomised search to give wider coverage of data space than that provided by deterministic methods (Schwefel op cit p109)
- Work from a large number of points simultaneously, so climbing many 'local' peaks in parallel and therefore improving upon simulated annealing in addition to traditional hill climbing methods (Michalewicz, 1992, p29)
- No auxiliary information in the form of derivatives is required, although the prior tabulation of distances between sites is needed for efficiency.
- Use of codings of parameter set, rather than parameters themselves, means that GAs are largely unconstrained by previously mentioned limitations such as continuity, derivative existence etc. (Goldberg, op cit, p7)
- General, rather than specific, technique allows code and methodology to be used for a wide variety of tasks when coupled with a GIS subject to alteration of codings/objective functions (travelling salesman e.g. Goldberg op cit, p170; groundwater pollution containment, Ritzel et al 1994; facilities siting, Pereira et al 1993)
- But, theoretical study of multi-criterion evolutionary algorithms is lacking (Michalewicz, op cit, p9)

**Table 3, Advantages of the evolutionary approach to sampling**

## METHODOLOGY

In developing a new sampling methodology, the first question to be asked is 'what is required of the sample data?'. As established within the literature review, the two main goals should be the 'representativeness' of the data with regard to the overall problem space, and the tailoring of data to suit the requirements of further analytical techniques. Specific criteria relevant to the search may then be introduced within the framework of more general evolutionary code. The local context of the sampling problem therefore requires close analysis. In this case the sample is to be used in the interpolation of point type meteorological data for England & Wales. The problem task is to choose a total of 200 sites from a possible 985, the number of sites restricted for economic reasons. The available meta-data for each site, derived data and its source are detailed below (Table 4) and in the first instance exclude partial or subjective data such as quality statements.

| MET. OFFICE | DERIVED |
|---|---|
| Location | Location |
| Start of recording (year) | Record length |
| End of recording (year) | Age of information |
| | Currency of data |

| | |
|---|---|
| GIS | |
| Ordnance Survey height data | Height |
| (1: 50,000 raster) | Second derivative of slope |
| | Aspect |

**Table 4, Sources of meta data for each site**

Derivations have been developed within the GIS GRASS or by spreadsheet, and move the initial data closer to a form useful in meeting the sampling criteria which are outlined below.

A. Goal 1: 'Representativeness' of sample data relative to full data set

Adequate coverage of sites throughout country
Distribution of aspects representative of those of UK as a whole
Distribution of heights representative of total range within UK

B. Goal 2: Sampling requirements related to interpolation tasks

Nearby sites required for successful production of variogram
Long records required for improved temporal predictions
Currency of data important for management of future infestations
Where data recording has stopped but site important for other characteristics, time since recording discontinued

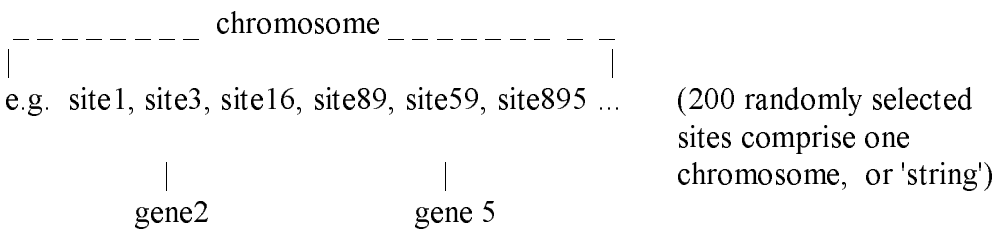
**Table 5, Development of sampling criteria**

Straightforward formulation of the criteria does not necessarily imply ease of optimisation however, as hinted at by the ease with which more traditional approaches may be discounted in this problem case. A number of methods of multi-criterion optimisation using evolutionary techniques exist, and this subject in itself forms an area of active research. Choice of method incorporates the two main stages involved within the evolutionary approach, those of fitness assignment (or suitability of the site) and search strategies (the probabilistic search through discrete space). In the main, attention has been given to the former (Jankowski 1995) and thus this is where most alternatives lie. Given the apparent lack of comparative material and a desire to maintain simplicity where possible, the most straightforward of approaches is chosen firstly: the use of the popular aggregation (weighted sum) approach to fitness familiar from other techniques together with standard (Michalewicz, 1992) search operations.

The basic implementation of the evolutionary approach is summarised below (After Davis, 1991):

- Initialise a population of chromosomes

(Where each chromosome is composed of 200 randomly selected sites (genes), selected from the possible 985 sites)




A floating point rather than the more conventional binary representation is used to prevent prohibitively long strings. This also has other advantages: it is intuitively closer to the problem space, facilitates the design of problem specific operators, and is faster and more consistent (Michalewicz, op cit, Chap. 5). This means that the approach used is not strictly that of a 'pure' (Goldberg 1989) genetic algorithm, hence the use of the broader term evolutionary algorithm.

- Evaluate each chromosome in the population according to its fitness (Fitness assignment)

(Where fitness equals the weighted sum of each of the eight (normalised) individual criterion fitnesses, akin to approaches developed for use with conventional optimisers)

e.g. Height fitness for chromosome 1 within the population (population = 200)

f(height)= (max. height - min height )/max. range possible within all 985 sites

e.g. Nearest neighbour calculation

$$f(\text{nearness}) = \left( \sum_{\text{gene 1}}^{\text{gene 200}} \text{Distance to nearest neighbour within string} \right) / 2 * \text{lowest 100 distances for complete set of sites}$$

- Create new chromosomes by mating current chromosomes, using procedures of crossover and mutation (search strategy) and where the probability of breeding relates to fitness.

e.g. simple, single point crossover

Initial two strings: s1, s4, s6 | s768, s500, s345 ...
with crossover points s98, s687 | s27, s879, s56 ...

Result: s98, s687,s768, s500, s345 ...

s1, s4, s6 s27, s879, s56 ...

e.g. mutation.

A value selected for mutation is replaced by a randomly generated number between that variable's lower and upper bounds.

In both cases, the possibility exists to create illegal combinations, that is strings in which a particular site appears more than once. For the promotion of diversity, no repair algorithm was applied to the crossover operation on the basis that the fitness function rewards a spread of points. The basic algorithm was however altered to prevent any illegal mutations from occurring, and additionally f(nearness) was amended to avoid rewarding duplicate sites.

A further issue arises when the contribution of one gene to fitness depends on the value of another gene (Epistasis). This can lure the evolutionary algorithm into suboptimal convergence, as explained by Michalewicz (op cit, p52). Because this situation arises in the case of the height fitness for example, an inversion operator was incorporated within the methodology which aims to group the maximum and minimum height genes together within a string to form an important building block.

e.g. inversion

| | | | |
|---|---|---|---|
| Initial string: | s1, s4, s78, s627, s3, s783, s876 ... | | |
| | | \| | \| |
| After fitness calculations: | max. height | | min height |

After inversion: s1, s4, s783, s78, s627, s3, s876 ...

- Delete members of the population to make room for new chromosomes
- Evaluate new chromosomes and insert them into the population
- If the set number of generations is up, terminate and return the best chromosome
- Repeat experiment five times and average results

## RESULTS AND DISCUSSION

In order for the evolutionary approach to sampling to be evaluated, its advantages and improvements compared to older strategies must be displayed. A major problem facing the study in this regard is that the performance of the evolutionary algorithm is a vector.

Because of conflicting criteria, such as the requirement for sites to be spread throughout England and Wales in contrast with the need for nearby sites in order to be able produce an adequate variogram for interpolation, a number of non dominated solutions are likely to exist. How therefore can a set of runs be evaluated? Perhaps only by the relative performance of the sample data in producing an interpolated data surface. However, because this experiment is set within the context of a real requirement for meteorological data, only meta-data is as yet available.

Instead, average initial results of the overall fitness of the solution were compared to that from a simple deterministic methodology in which sites were 'weeded' in a sequential fashion according to each criterion in turn, ranked according to perceived importance. While using this comparative technique selections could be made relatively straightforwardly according to second derivative of slope, age and continuance, problems arose in incorporating the full range of environmental parameters such as height and in balancing a good spread of sites with adequate nearest neighbours in geographically varied areas. The overall fitness value achieved by the ad hoc technique compared well with the 'best' value of the initial random populations of the evolutionary algorithm but unsurprisingly was not well balanced amongst the scores for individual criteria.

One of the obstacles to the wider application of evolutionary algorithms is the need to experiment using a number of parameters such as the mutation rate, the probability of crossover between populations, population size used and the form of the operators themselves before obtaining a satisfactory result. The success of the evolutionary procedure may be shown by reference to the average *convergence curve* of fitness per generation over several runs using identical parameters from different random start points and additionally the use of different strategies from similar start points. Work in progress currently shows a steady increase in the average fitness of the total population per generation, with slower gains in 'best' fitness. This steady increase in average population value indicates a performance better than purely random and that the methodology holds promise. Nevertheless, it is the case that further tuning of the crossover and mutation rates is required, and alternative evolutionary operators such as expected crossover and mutation and ranked crossover are currently being explored. Also, the weighted sum method which aggregates all individual fitnesses within one overall, compensatory function is acknowledged as being sub-optimal in comparison to more sophisticated algorithms such as Shaffer's (1985, in Michalawicz 1992) VEGA system which ranks populations according to pareto dominance (the ranked position according to each individual criterion, not accumulated fitness).

## CONCLUSIONS AND FURTHER DIRECTIONS

While simplistic in evolutionary computing terms, in comparison with traditional sampling methodologies the technique outlined appears rather complex. However, the approach outlined does allow for the sampling of fixed point sites of data both characteristic of the underlying landscape spaces and tailored for analysis. To date, only ad hoc alternatives

exist. Additionally, with the tailoring of the criteria and therefore the fitness function, the methodology developed is equally applicable to the selection of other geographically referenced point data, such as for example soil survey results and indeed to very different geographical problems.

In order to facilitate the exploration of a wider variety of genetic operators and parameters, work is undergoing to visualise the problem search space according to the defined function as a means of guidance. The generation of equally representative training and testing site combinations in order to avoid over reliance on cross-validation techniques is also underway.

## REFERENCES


Aspinall, R.J., Lees B.G. (1994) Sampling and analysis of spatial environmental data, In Waugh, T.C., Healey (Eds) (1994) Advances in GIS Research, *Proceedings of Sixth International Symposium on Spatial Data Handling,* Waugh: Edinburgh, p1086-1098.

Austin, M.P. , Heyligers, P.C.(1989) Vegetation survey design for conservation: Gradsect sampling of forests in north-eastern New South Wales, *Biological Conservation*, **50**, 13-32.

Bras, R.L., Rodriguez-Iturbe, I. (1976) Network design for the estimation of areal mean of rainfall events, *Water Resources Research*, **12,** 1185-1195.

Carver, S. J. (1991) Integrating multicriteria evaluation with GIS, *International Journal of Geographical Information Systems*, **5**, 321-339.

Davis, L. (1991) *Handbook of Genetic Algorithms*, Van Nostrand Reinhold: New York.

Eastman, J., Jin, W., Kyem, P.A.K., Toledano, J. (1995) Raster procedures for multi-criteria / multi-objective decisions, *Photogrammetric Engineering & Remote Sensing*, **61(5),** 539-547.

Fonseca, C.M., Fleming, P.J. (1995) An overview of evolutionary algorithms in multiobjective optimization, *Evolutionary Computing*, **3(1),** 1-16.

Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning, Addison*-Wesley: Reading, Mass.

Jankowski, P. (1995) Integrating geographical information systems and multiple criteria decision-making methods, *International Journal of Geographical Information Systems*, **9(3),** 251-273.

Janssen, R., Rietveld, P. (1990) Multicriteria analysis and GIS: an application to agricultural land use in the Netherlands, In: *Geographical Information Systems for Urban and Regional Planning*, Eds. Scholten, H.J., Stillwell, J.C.H., Kluwer: Dordrecht.

Lees B. (1993) Sampling strategies for machine learning using GIS, In *Proceedings of the Second International Conference on Environmental Modelling,* Breckenridge, NCGIA: CA.

Margules, C.R., Stein, J.L. (1989) Pattern in the distributions of species and the selection of nature reserves: an example from Eucalyptus forests in south-eastern New South Wales, *Biological Conservation*, **50(4),** 219-38.

Michalewicz, M. (1992) *Genetic Algorithms + Data Structures = Evolution Programs*, Springer- Verlag: Berlin.

Pereira, A.G., Peckham, R.J., Antunus, M.P. (1993) GENET: A method to generate alternatives for facilities siting using genetic algorithms, *In European Conference on Geographical Information Systems*, 1993, 973-982.

Pettitt, A.N., McBratney, A.B. (1993) Sampling designs for estimating spatial variance components, *Applied Statistics*, **42(1),** 185-209.

Ritzel, B.J., Eheart, J.W., Ranjithan, S. (1994) Using genetic algorithms to solve a multiple objective groundwater pollution containment problem, *Water Resources Research*, **30(5),** 1589-1603.

Schwefel, H.-P., (1995) *Evolution and optimum seeking*, Wiley: New York.

Shaffer, J.D., Grefenstette, J.J. (1985) Multi-objective learning via genetic algorithms, In Joshi, A.(Ed), *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann: San Mateo, CA, p593-595.

## AFFILIATIONS

* University of Edinburgh
Department of Geography
Drummond Street
EDINBURGH
chj@geo.ed.ac.uk

Central Science Laboratory,
Ministry for Agriculture, Food & Fisheries
Hatching Green
Harpenden
Herts.