ELSEVIER

Computers & Geosciences 32 (2006) 1320–1333

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

# Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood ☆

R. Webster[a,*], S.J. Welham[a], J.M. Potts[b], M.A. Oliver[c]

[a]*Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK*
[b]*BioSS, The Macaulay Institute, Craigiebuckler, Aberdeen AB15 8QJ, UK*
[c]*Department of Soil Science, The University, P.O. Box 233, Reading RG6 6DW, UK*

Received 25 July 2005; received in revised form 1 December 2005; accepted 2 December 2005

**Abstract**

The variogram is essential for local estimation and mapping of any variable by kriging. The variogram itself must usually be estimated from sample data. The sampling density is a compromise between precision and cost, but it must be sufficiently dense to encompass the principal spatial sources of variance. A nested, multi-stage, sampling with separating distances increasing in geometric progression from stage to stage will do that. The data may then be analyzed by a hierarchical analysis of variance to estimate the components of variance for every stage, and hence lag. By accumulating the components starting from the shortest lag one obtains a rough variogram for modest effort. For balanced designs the analysis of variance is optimal; for unbalanced ones, however, these estimators are not necessarily the best, and the analysis by residual maximum likelihood (REML) will usually be preferable.

The paper summarizes the underlying theory and illustrates its application with data from three surveys, one in which the design had four stages and was balanced and two implemented with unbalanced designs to economize when there were more stages. A Fortran program is available for the analysis of variance, and code for the REML analysis is listed in the paper.


*Keywords:* Nested sampling; Analysis of variance; Variance components; Variogram; Balance; REML

## 1. Introduction

Designing efficient sampling schemes for estimating such quantities as the amounts of metals in ore bodies, the concentrations of trace elements in soil, and of pollutants in wastes has taxed earth scientists for many years and continues to do so. Regional totals and means and their associated variances can be estimated from random samples; the underlying theory was established in the 1930s. The main uncertainty at the outset of a survey is the population variance, but once an investigator has a rough prior estimate of it he or she can design a reasonably efficient scheme that will provide the desired confidence for a modest sample size.

☆Code on server at http://www.iamg.org/CGEditor/index.htm. This paper is a contribution to the 'Studies for Students' series of the International Association of Mathematical Geology.

*Corresponding author.
*E-mail address:* richard.webster@bbsrc.ac.uk (R. Webster).

0098-3004/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2005.12.002

In many instances, however, single estimates of either totals or means are not what are wanted; rather investigators want to know *where* there are exploitable ores, deficiencies in trace elements, or excessive concentrations of pollutants and potential toxic chemicals. They want *local* estimates, and they may well want to map the variation also. For this the total size of the sample per se, i.e. the number of sampling points, is of little consequence. What matters is the sampling density in relation to the spatial scale(s) of variation. Any region of interest, whether a field, a catchment, or an administrative district, has its particular spatial scale or scales of variation, and sampling must be sufficiently intense to resolve the variation from place to place at one or more of those scales.

Most quantitative spatial prediction and much interpolation are now done by kriging, for which the estimation of the variogram is an essential intermediate step. Again, sampling must be dense enough to estimate the variogram over the lag distances that embrace a large proportion of the variance. Too sparse sampling can lead to 'flat' sample variograms, i.e. ones that appear to be all nugget. The consequence is that we learn nothing of the underlying spatial structure, and it makes no sense to attempt interpolation from the data in those circumstances. It is a recurrent shortcoming in resource survey, and it is one that we have faced ourselves (Oliver and Webster, 1987). We might attempt to sample at intervals so short as to ensure that the resulting data are spatially dependent. However, we must recognize that measurements, whether in the field or in the laboratory on material collected, can be expensive. If the sampling interval were less than, say, one-tenth of the range of the variogram then we should be wasting effort and spending more than necessary to achieve our objectives. There must be some intermediate sampling density that would enable surveyors to estimate the variogram economically. The question is how to discover that density. It is one that we encounter time and time again in our consulting work.

Just as investigators want rough estimates of population variances to decide how large samples should be for estimating regional totals or means, so we want rough prior estimates of variograms. The principles for obtaining such estimates with modest resources were worked out long ago by Youden and Mehlich (1937) in their adaptation of analysis of variance. Unfortunately, they published their innovation in the house journal of their institute where it lay unrecognized for many years. The technique was rediscovered several times by, for example, Olson and Potter (1954) in geology, Hammond et al. (1958) in agronomy, and Webster and Butler (1976) in soil survey, though it was the last who finally turned the results into variograms.

Despite the publicity that we (Webster and Butler, 1976; Webster, 1984; Oliver and Webster, 1986; Webster and Oliver, 1990, 2001) and others such as Miesch (1975) have given to the subject, the technique still seems poorly understood and little used, partly perhaps because of the lack of readily available software. Our purpose here is to rectify the first and to provide code on the Web site of the Journal (http://www.iamg.org/CGEditor/index.htm) so that readers should be able to do the analysis painlessly.

## 2. Theory

The conventional variogram of geostatistics, defined formally in Eq. (4), can distinguish the spatial variation in a region at one or two spatial scales within an order of magnitude simultaneously, but rarely more. Until we know approximately what those scales are, however, we cannot design a sampling scheme that will ensure that we can estimate the variogram sensibly. Further, the variation might derive from two or more sources with spatial scales spanning several orders of magnitude. So we want to be able to estimate the contributions to the variance from all distances from the smallest likely to be of interest to the largest. It can be done by spatially nested sampling followed by a hierarchical analysis of variance.

The idea underlying the technique is that a population of units in the field can be divided hierarchically into classes at two or more categoric levels or stages. At each stage the population is subdivided into classes consisting of units that are geographically close to one another. The population is first divided into classes at the highest level (stage 1), each of those classes is subdivided into subclasses at the level below (stage 2), and so on until the subdivisions are the smallest of interest at stage $m-1$. Points are then selected at random within these finest subdivisions, and it is at these that the material of interest is measured to give data at the lowest level (stage $m$).

As an example, think of an agricultural region comprising administrative districts (stage 1), each