JOURNAL OF
ENVIRONMENTAL HYDROLOGY
The Electronic Journal of the International Association for Environmental Hydrology
On the World Wide Web at http://www.hydroweb.com
VOLUME 11 2003

LINEAR GEOSTATISTICAL MODELS FOR ESTIMATING GROUNDWATER QUALITY VARIABLES

G.L. Verma[1]
H.S. Bhatia[2]
D. Katyal[1]

[1]Department of Applied Chemistry
Delhi College of Engineering, New Campus, Delhi, India
[2]Department of Environmental Engineering,
Delhi College of Engineering, Delhi, India

Geostatistics provides set of probabilistic techniques, which are useful to detect and find the mode of patterns of spatial dependence of attribute values in space, and further use these models for the assessment of uncertainty about unknown values at locations not sampled. In this paper, an analysis of geostatistical methods is presented, including a discussion of possible application and limitations for regional groundwater quality, and its application to mapping problems.

INTRODUCTION

Geostatistics is the discipline that provides a set of models and tools for the estimation of block averages or local averages from sample observations, taking both large scale variation (the trend) and small scale variation (spatial correlation) into account. The mathematical formulation of the set of relations connecting independent condition defining variables to the measurement is the model. Which independent variables are used, and how the measurement variable depends on them is the model structure. Theory and data should be balanced when choosing the model structure. The simplest adequate model that is parsimonious in its parameters is usually chosen. Given a set of relations that define the model structure, we have to choose a statistical procedure to estimate the characteristic of interest. Statistics is the discipline that provides the theory and tools for deciding on the data, theory complexity, trade offs inferring population characteristics from limited sample information, and for assigning accuracy measures to the inference.

Linear Models

For estimating groundwater quality variables at locations where they are not measured, we choose statistical models, and the family of linear models provides a comprehensive framework for most commonly used statistical models .A limitation of linear models is that they only allow additive effects. For groundwater quality variables (being nonnegative and highly skewed) additivity of effects is often a reasonable assumption after log transforming the measurements. Wider classes of problems can be formed, e.g. by nonlinear transformation of variables (as in generalized linear models, McCullagh and Nelder, 1989) or by defining linear relations locally (approximating more complex relations with local linear models as in generalized additive models, Hastie and Tibshirani, 1990). Comprehensive treatment of linear models is found in Searle (1971), Rao (1973), Christensen (1987) and in Draper and Smith (1981). Cressie (1992) and Christensen (1991) extended these treatments to mapping problems.

METHODOLOGY

Linear models provide a flexible way of expressing a wide range of problems in a compact notation. Consider the two following problems.

1) n groundwater quality measurements are collected randomly from a homogenous spatial units, and we want to estimate the mean value of the measured variable in this unit and its estimation variance.

2) n groundwater quality measurements are collected in two spatial units, q in the first, n-q in the second and we want to estimate the mean of each group.

The observations $z(x_i)$ at location x_i, $i=1...n$ can be written as

1) $z(x_i) = m + d(x_i),\ \ i=1...n$

2) $z(x_i) = m_1 + d(x_i),\ \ i=1...q$

$z(x_i) = m_2 + d(x_i),\ \ i=q+1...n$

Here conceptually the measurements are taken as the sum of a structural, systematic part and a residual, unsystematic part. The structural part consists of (i) m, the mean value of the spatial unit; (ii) m_1 and m_2, the mean values of units 1 and 2. The unsystematic part in all the three problems is $d(x_i)$: the deviation of the i-th measurement from the structural part.

In a linear model the observation $z(x_i)$ is represented by a random variable $Z(x_i)$, and *Z(xi) is* modeled as a sum of its expected value E $(Z(x_i))= m(x_i)$ and a random deviation from $m(x_i)$, $e(x_i)$

$Z(x_i)=m(x_i)+e(x_i)$, $E(e(x_i))=0$,

Note that *e* refers to residual variation that is not accounted for by $m(x)$.

The expected value $m(x)$ is modeled as a linear function of p unknown, independent variables that have a causal influence upon $Z(x)$, and p unknown coefficients β_j that relate these independent variables to the observations

$$m(x_i) = \sum_{j=1}^{p} f_j(x_i)\beta_j,$$

which gives, using the vector notation

$$Z(x_i) = f(x_i)\beta + e(x_i)$$

With $f(x_i)=(f_1(x_i), f_2(x_i), \ldots\ldots, f_p(x_i))$ the row vector with the values of the independent variables at location x_i, and $\beta = (\beta_1, \ldots\ldots \beta_p)'$, the column vector with the unknown coefficients.

Representing all observations, the model can be written as

$$Z(x) = F_x\beta + e(x)$$

With $Z(x)=(Z(x_1),...Z(x_n))'$, $F_x=[f_j(x_i)]_{nxp}=(f_1(x),\ldots\ldots..,f_p(x))$ with $f_j(x)=(f_j(x_1),\ldots\ldots,f_j((x_n))'$, and $e(x)=(e(x_1),\ldots,e(x_n))'$. When we define F_x and β accordingly, this model covers the two simple problems discussed above:

5) $p = 1, f(x_i) = 1...n,$ and $\beta_1 = m$

6) $p = 2,$ if $i \le q$ then $f(x_i) = (1,0)$ else $f(x_i) = (0,1)$ and $\beta = (m_1, m_2)'$

Thus we can structure the problem by choosing F_x (and thus defining the size of β). The *j*th column in F_x, *fj(x) defines* the structure of the relation of the measurement variable *z(x) of* the *j*th parameter β_j. If β_j is the overall mean as *m* in problem (1), then $f_j(x)$ is a column of ones. If the model contains categories as in problem (2) and β_j is the mean of the *j*th- category, then $f_j(x)$ is a binary variable that for every observation denotes whether it belongs to the *j*-th category with a one, or not , with a zero.

1. Linear models with independent, identically distributed errors:

In the simplest case we assume that the errors *e(x)* are independent and identically distributed, resulting in the model

$$Z(x) = F_x\beta + e(x), E(e(x) = 0, Cov(e(x)) = \sigma^2 I, \quad (1)$$

which leads to the ordinary least squares estimates (provided that F_x has full rank.)

$$\hat{\beta} = \left(F_x' F_x\right)^{-1} F_x' z(x) \quad (2a)$$

and estimation variances and covariances for $\left(\beta-\hat{\beta}\right)$

$$Cov\left(\beta-\hat{\beta}\right)=\left(F_x{}'F_x\right)^{-1}\sigma^2 \tag{2b}$$

Where σ^2 is estimated by

$$s^2=z(x)'(I-F_x(F_x{}'F_x)^{-1}F_x{}')z(x)/(n-R) \tag{3}$$

with R the rank (the number of columns)of F_x.

At an unsampled location x_0, given this estimate $\hat{\beta}$, the value of $z(x_0)$ (and the mean value of r independent replications of $z(x_0)$ is estimated by

$$\hat{z}(x_0)=f(x_0)\hat{\beta} \tag{4}$$

with $f(x_0)$ the value of the independent variables at location x_0. The estimation variance of the estimator of $z(x_0)$ (r=1) or the mean of r independent replications of $z(x_0)$ is given by

$$\sigma_r^2(x_0)=(\frac{1}{r}+f(x_0)\ (F_x^{'}F_x)^{-1}\ f(x_0)^{'})\sigma^2 \tag{5}$$

2. Linear models with dependent errors

A wider class of problems than the one with independent identically distributed errors is obtained when the errors are allowed to be dependent.

$$Z(x)=F_x\beta+e(x), E\left(e(x)\right)-0, Cov\left(e(x)\right)=V \tag{6}$$

With $V=\left[Cov\left(e(x_i),e\left(x_j\right)\right)\right]_{nxn}$ .This leads to weighed least square (WLS) estimates of β

$$\hat{\beta}*=\left(F_x{}'V^{-1}F_x\right)^{-1}F_x{}'V^{-1}z(x) \tag{7a}$$

with estimation covariances

$$Cov\left(\beta-\hat{\beta}*\right)=\left(F_x{}'V^{-1}F_x\right)^{-1} \tag{7b}$$

Under this model, given $\hat{\beta}*$ by 7(a and b), the estimate of $z(x_0)$ is

$$\hat{z}_{wls}\left(x_0\right)=f\left(x_0\right)\hat{\beta}*+v_0'V^{-1}\left(z(x)-F_x\hat{\beta}*\right) \tag{8}$$

where $v_0'=\left(Cov\left(e(x_1),e(x_0)\right),\ldots Cov\left(e(x_n),e(x_0)\right)\right)$ and has estimation variance:

$$\sigma_{wls}^2(x_0)=C(0)-v_0^{'}V^{-1}v_0+(f(x_0)-v_0^{'}V^{-1}F_x)(F_x^{'}V^{-1}F_x)^{-1}(f(x_0)-v_0^{'}V^{-1}F_x)^{'} \tag{9}$$

with $C(0) = Var(e(x_0))$. In statistical terms, the value estimated in (8) is expressed as the sum of the best linear unbiased estimate of $m(x_0), \hat{m}(x_0) = f(x_0)\hat{\beta}$ and the best linear unbiased predictor of the (correlated) error, $\hat{e}(x_0) = v_0' V^{-1}\left(z(x) - F_x\hat{\beta}*\right)$.

3. Multivarable estimation

When s variables $Z_k(x)$, k=1……s each follow a linear model $Z_k(x) = F_{k,x}\beta_k + e_k(x)$, and the $e_k(x)$ are correlated, then it makes sense to extend the weighed least squares model to allow multivariable estimation. Without loss of generality, assume $s = 2$. When $z(x) = \left(z_1(x), z_2(x)\right)'$ and $B = (\beta 1, \beta 2)'$ are substituted for $z(x)$ and β, and when

$$\mathbf{f}(x_0) = \begin{bmatrix} f_1(x_0) & 0 \\ 0 & f_2(x_0) \end{bmatrix}, \mathbf{F}_x = \begin{bmatrix} F_{1,x} & 0 \\ 0 & F_{2,x} \end{bmatrix}, \mathbf{V} = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \mathbf{v}_0 = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$$

with $V_{21} = [Cov(e_2(x_i), e_1(x_j))]$, $v_{21} = (Cov(e_2(x_1), e_1(x_0)), \ldots, Cov(e_2(x_n), e_1(x_0)))'$ and 0 a conforming zero matrix or vector, are substituted for $f(x_0)$, F_x, V and v_0, then the left hand side of both (8) and (9) yield the multi variable estimates: the left hand side of (8) then becomes the estimate vector $\hat{z}(x_0) = (\hat{z}_1(x_0), \hat{z}_2(x_0))'$, and the left hand side of (9) becomes the (2x2) matrix with estimation covariance.

Confidence intervals

When the estimation error $z(x_0) - \hat{z}(x_0)$ is normally distributed with zero mean and variance $\sigma^2(x_0)$ confidence intervals can be constructed for $z(x_0)$ (depending on the model used, (4) and (5), or (8) and (9)), and the interval

$$[\hat{z}(x_0) - 2\sigma(x_0), \hat{z}(x_0) + 2\sigma(x_0)]$$

is a 95% confidence interval for $z(x_0)$

RESULTS AND CONCLUSIONS

For mapping groundwater quality, variables for units of size of the measurements, estimating value that would actually be measured at an unsampled location is not possible because the variation in the measurements is frequently too large, and consequently we cannot estimate groundwater quality with a reasonable accuracy at that scale. At a lower spatial resolution, it is possible to estimate groundwater quality variables because the pattern of local average groundwater quality is smoother than the pattern of measurements. For the estimation of this smooth pattern of local averages we need a model that allows spatial local average values.

In such case, it would be convenient to use the linear model with independent (IID) errors presented in the above section, because it is simple and it is supported by a rich body of research in classical statistics . However if we want to attain independence from a design based argument (e.g. Hansel et al.; De gruijter ter braak, 1990), then this model is only adequate for estimating the pattern of local values when (i) the observations are collected randomly from the areas with constant values for the $f(x_i)$ (i.e. from the categories distinguished or from an area with a specific value for the

regressors) (ii) the areas for which we want estimates of a spatial average coincide with these areas of constant $f(x_i)$. These conditions limit the suitability of the above model for estimating local averages severely.

Thus for mapping of groundwater quality variables, one should continue with models that allow errors to be substantially dependent. Moreover, although part of the residual errors can usually be attributed to measurement error, it is very likely that in a linear model intended for spatial estimation, a large proportion of the residual, unexplained variation is caused by unknown spatially smooth factors resulting in a spatially dependent error. For mapping purposes, we can use this spatial dependence to capture the spatial structure present in the measurements beyond the part explained by the independent variables. Average values of arbitrarily shaped elements can be estimated efficiently if we are willing to assume a model with a spatially dependent error structure.

REFERENCES

Cressie, N.; Statistics for spatial data ,Wiley, New York.

Cressie, N. and D.L. Zimmerman; Stability of the geostastical method , Mathematical Geology, Vol .24, No.1, pp 45-49.

Christensen, R.; Plane answers to complex question: the theory of linear models , Springer, NewYork.

Christensen, R.; Linear models for multivariate time series and spatial data. Springer, New York.

De Gruijter, J.J. and C.J.F. terBraak; Model free estimation from spatial samples: A reappraisal of classical sampling theory, Mathematical Geology, Vol. 22, No.(4), pp 407-415.

Draper, N. and H. Smith; Applied regression analysis, Second edition. Wiley, New York.

Hansen, M.H., W.H. Madow, and B.J. Tepping; An evaluation of model dependent probability–Sampling inferences in sample surveys, Journal of the American statistical Association ,Vol 78, pp 776-793.

Hastie, T. and R. Tibshirani; Generalized additive models, Chapman and Hall, London

McCullagh, P. and J.A. Nelder; Generalized Linear models, Second edition, Chapman and Hall, London.

Rao, C.D; Linear statistical inference and its applications, Second edition, Wiley, New York.

Searle, S.R; Linear models, Wiley, New York.

ADDRESS FOR CORRESPONDENCE
Deeksha Katyal
Department of Applied Chemistry
Delhi College of Engineering New Campus
Bawana Road
Delhi
India

E-mail: deeksha_dce@rediffmail.com