Civil Engineering Seminar
13 October 2006

GEOSTATISTICS:WHAT IS IT AND WHAT'S IT GOOD FOR?

Donald E. Myers
Emeritus Professor of Mathematics
Emeritus Joint Professor of Hydrology
Member of Committee on Remote Sensing and Spatial Analysis

NOTE: you may need to copy/paste some of these hyperlinks as opposed to merely clicking on them, in particular if the URL is longer than one line
http://www.u.arizona.edu/~donaldm

COMMERCIAL
I will be teaching Geostatistics next semester, Mathematics/Geography 574, TTH 12:30-1:45

WHAT IS GEOSTATISTICS?
Geostatistics is concerned with spatial data and uncertainty

> **Each data value is associated with a location**
> **Each data value/location has a *support***
> **Changing the support can be useful but also cause problems**

> **The data locations need not be on a regular pattern.**

Spatial data is often expensive to collect and thus there usually isn't enough of it.

Collecting it may alter the environment, e.g drilling wells or boreholes

Uncertainty can occur or appear in a problem/application in at least two ways: (1) something has not yet occurred and might occur in different ways, (2) lack of knowledge, i.e., data. Geostatistics is more concerned with the latter than with the former.

Some examples of spatial data
    Geohydrological parameters
        Hydraulic conductivity
        Porosity

    Water table/aquifer properties
        Depth to groundwater (head)

Wolfcamp data
www.uni-muenster.de/ZIV/Mitarbeiter/BennoSueselbeck/s-html/helpfiles/aquifer.html

http://www.ic.arizona.edu/ic/math574/analyses/wolfcamp/wo_plt.gif

    Concentrations of various natural substances in the subsurface
        Ore grades
        Soil nutrients

    Contaminant concentrations
        In an aquifer
        In the soil

Weather/climate characteristics (could be spatial-temporal as well)
        Humidity
        Temperature

Wind velocity
Components of wind direction

An Important Characteristic Commonly Observed for Spatial Data

If two locations are close together the values are more "similar" than if the locations are far apart.

A key question of course is, what do we mean by "similar"?

Another way of saying it: If you want to predict/estimate the value at a non-data location, then the data at nearby locations is more useful than the data at locations far away.

EXAMPLES OF (Generic) PROBLEMS TO SOLVE, GIVEN A SPATIAL DATA SET

1. Estimate/predict the values at non-data locations
   "Contour"/ interpolate the data

2. Estimate the average value over a region (2-space or 3-space), e.g., rectangle or "block"

3 Generate alternative "contours" fitting the data and preserving some statistical properties
   Shape of frequency distribution
   Measure of spatial "similarity"
   Contoured data may be too "smooth" and hence unrealistic

4. Provide some measure of the quality/reliability of the estimates

This not an exhaustive list, e.g. Bayesian methods

A LITTLE BIT OF HISTORY

What is now known as geostatistics largely traces back to the work of Georges Matheron at the Ecole des Mines, Paris, France beginning in 1965, focusing on applications in mining, hydrology and petroleum. This led to the establishment of a research center at Fontainebleau (about 60 km south of Paris). At least two other persons were doing somewhat similar work at about the same time, B. Matérn in Sweden (forestry applications) and Y. Ghandin in the Soviet Union (meterology and atmospheric physics). Closely related work was done by R. Hardy at Iowa State in 1971 (Radial Basis functions). Geostatistics developed largely outside of the statistics discipline. This is still true today although it receives much more attention in the statistics literature than it used to. .

See also http://www.u.arizona.edu/~donaldm/homepage/whatis.html
www.u.arizona.edu/~donaldm/homepage/ua-geostatistics.html

A BIT OF MATHEMATICS/STATISTICS

Geostatistics is a model based method/technique as opposed to a design based method

The data is viewed as a non-random sample from one *realization* of a random function.

> **A random function might be thought of as a collection of random variables, one associated with each location in the region of interest. Alternatively as a random variable whose possible values are functions.**

Let *s* denote a location in space (usually 2 or 3 dimensional), Z(*s*) a random function

Z(s) is assumed to satisfy a form of *stationarity* (which is not testable from data)

Second order or intrinsic

Both conditions (essentially) imply that E[Z(*s*)] is a constant

"E" denotes expected value

Quantifiying "similarity"
(Auto)Covariance function (requires second order stationarity)

$$C(h) = \text{Cov}\{Z(s+h), Z(s)\}, h \text{ a vector}$$

"Cov" denotes covariance

Must be a positive definite function

Variogram (requires intrinsic stationarity, a weaker condition)

$$\gamma(h) = 0.5\text{Var}[Z(s+h) - Z(s)], h \text{ a vector}$$

"Var" denotes variance

Must be a conditionally negative definite function

IF Z(*s*) is second order stationary then $\gamma(h) = C(0) - C(h)$

These quantify spatial correlation, i.e., spatial dependence

$C(0) \neq 0$ and $C(h) = 0$ for $h \neq 0$ corresponds to no spatial dependence. The corresponding variogram is called a pure nugget model

A crucial problem is how to use data to determine the choice of the covariance function or variogram.

The practical solution is to fit the data to one of a class of known valid covariance functions or variograms, testing for positive definiteness or conditional negative definiteness is very non-trivial.

Suppose given data $Z(s_1), ..., Z(s_n)$ one objective is to estimate/predict $Z(s)$, s a non-data location.

IF there were no spatial dependence then the best estimator would be

$$Z^*(s) = E[Z(s)], \quad \text{"E" denotes expected value}$$

and this could be estimated by

$$Z^*(s) = \sum_{[i=1,..n]} (1/n)Z(s_i) \quad \text{(Arithmetic average)}$$

This estimator is "best" in two senses

(1) It is unbiased, i.e. $E[Z^*(s) - Z(s)] = 0$

(2) It has minimal estimation variance ,
        $Var[Z^*(s) - Z(s)]$ is minimal

but it is "non-local".

This suggests using

$$Z^*(s) = \sum_{[i=1,..n]} \lambda_i Z(s_i) \qquad \text{(Weighted arithmetic average)}$$

where the weights are chose to satisfy conditions (1) and (2). This is known as the "kriging" estimator. The weights are obtained as the solution of a system of linear equations whose coefficients are determined by the covariance function or variogram. One of the equations is simply

$$\sum_{[i=1,...n]} \lambda_i = 1$$

but this does not imply that the weights are between 0 and 1.

The weights on near data locations will be the largest and far away location weights will be zero or near zero.

> This form of the estimator is motivated by another important statistical property. If $Z(s)$, $Z(s_1)$,..., $Z(s_n)$ are considered as random variables (each with mean zero) that have a multivariate Normal (Gaussian) distribution then the conditional expectation of $Z(s)$ given $Z(s_1)$,..., $Z(s_n)$ is exactly the weighted arithmetic average given above

The minimized estimation variance is called the "kriging variance", it is not quite a true variance in the sense that one might use it to construct confidence intervals but rather is a relative measure of reliability.

The kriging estimator has one more important property, it is "exact". If we estimate at a data location, the estimated value will be the data value. Note that there are no distributional assumptions.

Non-Linear Geostatistics

(Non-linear) Transformations are very common in other parts of statistics, sometimes to ensure that a necessary statistical assumption is satisfied. Ordinarily it is not necessary to re-transform but in most geostatistical applications re-transformation would be necessary. Consequently only two transformations receive much attention. First the logarithmic transform:

$$W(s) = Ln\ (Z(s))$$

This is best used when the random function is multivariate lognormally distributed (in order to compute the bias correction when re-transforming)

$$I_Z(s;\ a)\ =\ 1\ if\ \ Z(s) \le a$$

$$=\ 0\ if\ \ Z(s) > a$$

This is called the indicator transform. In this case we will be estimating the conditional probability of $Z(s) \le a$ (Conditional on the data)

Simulation (Alternative scenarios)

As noted above, the data is viewed as a non-random sample from one *realization* of a random function. This does not imply that the realization is unique, hence the question is can we generate other realizations? Note that simulation as used herein does not mean the same thing as in simulating the solution of a differential equation.

The problem is to preserve certain properties; the data, the spatial correlation and perhaps the marginal probability distribution.

Recall that a random variable can be simulated (multiple realizations) by using a random number generator. We want to simulate one realization for each of multiple random variables (one at each of multiple locations) but not independently.

Usually we will want to do this whole process multiple times, i.e. multiple realizations

(1) The simulated value at each data location will be the original data value

(2) Preserve the "spatial correlation"
This can occur in at least two ways
(a) On average
(b) separately for each realization.

(3) Preserve the marginal distribution
Intuitively this means that the shape of the histogram of the simulated values should be the same as the shape of the histogram of the original data

There are a number of different algorithms for simulating spatial data, these include

(i) Turning Bands
Actually a method to generate realizations in 3-space from realizations in 1-space

(ii) LU or Cholesky decomposition
Uses the LU or Cholesky decomposition of a covariance matrix.

(iii) Sequential Gaussian
Based on the form of the conditional expectation in the case of multivariate gaussian distribution

(iv) Simulated annealing
Uses the optimization method known as simulated annealing

Simulation for porosity modeling.
www.statios.com/Resources/08-sgsim.pdf#search=%22simulation%2C%20geostatistics%22

The above is not a listing of all the ideas and techniques/methods used in geostatistics

Connections with Other Methods/ideas

Trend surfaces

In this case the interpolator appears as a regression on powers of the position coordinates or products of these powers. It is essentially like kriging with a pure nugget effect variogram. The trend surface will excessively smooth the data and is not "exact"

Radial Basis Functions (including thin plate splines)

In this one generates an interpolating function as a linear combination of translates of a positive definite function (covariance function) or conditionally positive definite function (e.g. variogram).The thin plate spline is a special case. In turn each of these can be re-written as a kriging estimator (perhaps using a generalized covariance function).

Inverse Distance Weighting

This estimator is again a weighted average but the coefficients/weights are determined solely by distances between the interpolation location and each of the data locations. There is no theory and it is really "adaptable to the data"

GEOSTATISTICS SOFTWARE
Geostatistics is computational intensive and hence software is very important. For a more complete listing see
http://www.ai-geostats.org/software/Geostatistics_Softfaq.htm

Free software

R, extensible software, source code available and binaries for Linnux and Windows
http://www.r-project.org/

See the gstat and geoR packages

Short Introduction to Geostatistical and Spatial Data Analysis with GRASS and R statistical data language
http://grass.itc.it/statsgrass/grass_geostats.html

GRASS: Geostatistics and spatial data analysis
http://grass.itc.it/statsgrass/index.php

Commercial software
Using ArcGIS Geostatistical Analyst
www.ci.uri.edu/projects/geostats/Using_ArcGIS_Geostat_Anal_Tutor.pdf

ArcGIS Geostatistical Analyst
www.esri.com/library/whitepapers/pdfs/geostat.pdf

ArcGIS Geostatistical Analyst tutorial
webhelp.esri.com/arcgisdesktop/9.2/pdf/Geostatistical_Analyst_Tutorial.pdf

Developing a Three-Dimensional Surface-Analytical and Geostatistic Tool Using ArcObjects
http://gis.esri.com/library/userconf/proc04/abstracts/a2096.html

Petroleum geostatistics using SYSTAT
http://www.ritme.com/tech/systat/resources/pdf/Petroleum.pdf

Simulation software using fractals, applications to petroleum
http://www.dpr.csiro.au/ourcapabilities/petroleumgeoengineering/geostatistics/projects/geostatisticslevysim/

There are also libraries of routines available to use with MATLAB (or the free version OCTAVE)

WHAT IS IT GOOD FOR?

Geostatistics is used in a variety of disciplines related to civil engineering but also many not at all seemingly related. One of best ways to see the usefulness of geostatistics is to examine the literature.

Doing a search on Google for "civil engineering, geostatistics" shows at least the following civil engineering departments that teach geostatistics or which recommend it to their students

> **University of Minnesota**
> **University of Notre Dame**
> **University of California at Davis**
> **New Mexico Tech**
> **University of California at Berkeley**
> **Oregon State University**
> **Michigan Tech**
> **University of Washington**
> **Stanford University**
> **University of Southern California**
> **Kansas State University**

Water Resources Research, J. Hydrology, Groundwater all have lots of papers with applications of geostatistics to problems in hydrology. (The Department of Hydrology has always been a significant source for students in my geostatistics class)

Soil Science Society of America J., Geoderma, European J. of Soil Science have lots of papers with applications to problems in soils

Mathematical Geology is one of the premier journals for papers on geostatistics (both applications and theory).

The book, "Reliability and statistics in geotechnical engineering", G.B. Baecher and J. T. Christian, J. Wiley devotes parts of multiple chapters to geostatistics and applications of geostatistics"

The J. of Geophysics often has papers using geostatistics.

Petroleum was and continues to be a discipline where geostatistics has wide application.

There are applications in ecology, entomology, epidemiology, geography, agronomy and a number of journals devoted to environmental assessment and modeling, atmospheric sciences and climatology (including "tree rings")

A Few Examples
An example taken from a dissertation in this department
1990, M.Ali, E. Nowatzki and D.E. Myers, Probabilistic Analysis of Collapsing soil by Indicator Kriging Math. Geology, 22, 15-38

1989, M.Ali ,E. Nowatzki and ,Myers,D.E., Geostatistical Methods to Predict Collapsing Soils. in Proceedings of the XIIth International Conference on Soil Mechanics and Foundation Engineering, August 1989, Rio de Janiero, Brazil, A.A. Balkema, Rotterdam 567-570

1990, M. Ali ,E. Nowatzki and ,Myers,D.E., Use of Geostatistics to Estimate the Probability of Occurrence of Collapse Susceptible Soils

in Tucson,AZ. in Proceedings of the 1989 Foundation Engineering Congress, 25-29 June 1989, Evanston, Illinois, 176-190

1989, M. Ali, E. Nowatzki and ,Myers,D.E., Geostatistical techniques to estimate collapse-related soil parameters. in Engineering Geology and Geotechnical Engineering, Watters (ed), Balkema Publishing, Rotterdam, 289-296

General papers on geostatistics and civil engineering
Blanchin, R., Chilès, J. P., and Deverly, F., 1989, Some Applications of Geostatistics to Civil Engineering,in M. Armstrong (ed.),Geostatistics (Vol. 2): Kluwer, Dordrecht, Netherlands, p. 785–795.

Geostatistics in geotechnical engineering: A fad or an empowering approach
www.rocscience.com/library/rocnews/Spring2003/GeostatisticsArticle.pdf

Geostatistics for Environmental and Geotechnical Applications, ASTM STP 1283, eds. S. Rouhani, R.M. Srivastava, A.J. Desbarats, M.V. Cromer, and A.I. Johnson. 1996. Description: This Special Technical Publication of the ASTM contains an excellent collection of papers that provide an overview of geostatistics and describe applications to environmental and geotechnical engineering.

A very visible(?) application
The Channel Tunnel: Geostatistical prediction of the geological conditions and its validation by the reality. Raymonde Blanchin and Jean-Paul Chilès, Mathematical Geology 25 (1993) 963-974

Rock mechanics

La Pointe, P.R. "Analysis of the spatial variation in rock mass properties through geostatistics," in Proceedings of the 21st Symposium on Rock Mechanics: A State 0f the Art, University of Missouri, Rolla, May 28 – 30, 1980, pp. 570-580. Description: This is a well-written paper that describes the specific application of geostatistics to rock engineering.

1997, Roko, R., Kim, Y.C. and Myers, D.E.,   Variogram characterization of joint surface morphology and asperity deformation during shearing. International J. of Rock Mechanics 34, 71-84

A cross disciplinary application in Petroleum and Hydrology
DOE report: Integrated Approach Towards the Application of Horizontal Wells to Improve Waterflooding Performance
http://www.osti.gov/bridge/product.biblio.jsp?osti_id=13821

Application of seismic stratigraphy and sedimentology to regional hydrogeological investigations: an example from Oak Ridges Moraine, southern Ontario, Canada. Sharpe, D R.; Pugin, A; Pullan, S E.; Gorrell, G. Canadian Geotechnical Journal, 40, (2003), 711-730

Plant Pathology
1994, M. Nelson, R.Felix-Gastelum, T. V. Orum, L.J. Stowell and Myers, D.E., Geographic Information Systems and Geostatistics in the Design and Validation of Regional Plant Virus Management Systems. Phytopathology 84, 898-905

Petroleum
Damsleth,E. and Omre,H.;1997: Geostatistical Approaches in Reservoir Evaluation; Invited Technology Today Series in Journal of

Petroleum Technology; May 1997, pp 498-502.

Omre,H. and Tjelmeland,H.;1997: Petroleum Geostatistics; in Baafi and Schofield (eds) Geostatistics Wollongong '96, Vol I; Kluwer Acad. Publ.; pp 41-52.

Books and Reference Materials
An extensive listing
http://www.ai-geostats.org/books/

Petroleum geostatistics for nongeostatisticians
http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=LEEDFF000019000005000474000001&idtype=cvips&gifs=yes

Achievements and Challenges in Petroleum Geostatistics
http://www.ingentaconnect.com/content/geol/pg/2001/00000007/A00