Paper 279

Geostatistics using SAS® Software

A. Katherine Ricci, Owen Analytics Inc., Deep River, CT

ABSTRACT

Two experimental procedures in SAS/STAT® Software Release 6.12, PROC VARIOGRAM and PROC KRIGE2D, allow two dimensional geostatistical modeling and estimation. A brief background in theory precedes a full geostatisticat analysis of spatially correlated data. The data are bathymetric soundings of depth in Lake Huron, presented in the NOAA Report ERL GLERL-16 Computerized *Bathymetry* and *Shorelines* of *the* Great *Lake* (Schwab and Sellers, 1996).

INTRODUCTION

Geostatistics was defined as "... the application of the formalism of random functions to the reconnaissance and estimation of natural phenomena" by G. Matheron, who introduced the theory of regionalized variables for mining applications. The technique of geostatistics is used to create of model of spatially correlated random variables based on samples, and then estimate values at unsampled locations using the model.

WHY GEOSTATISTICS?

Many classical statistical methods rely upon the independence of data samples for inference. In practice the independence is often counter-intuitive. Many natural processes exhibit a correlation in the spatial dimension. Samples at closer distances are more alike than those at further distances. In some cases, the direction of the points affects the relationship. Finally, there may be a distance beyond which samples are effectivelyindependent.

Geostatistical methods model the covariance structure in a process. The predictive model for covarlance is selected from a family of functions, variograms, and fit to empirical data to obtain parameter estimates. The resulting model is used to predict values at unsampled locations. Sampled points need not be gridded, or evenly spaced for estimations, so the techniques are well suited to applications where precise sampling locations cannot be specified.

REGIONALIZED VARIABLES

A regionalized variable Z(s) is the value of Z at point s in $S \subseteq R^d$ with each s continuous in S. Any two variables Z(s) and $Z(s+h)$ are autocorrelated and depend at least partially on vector h in magnitude and direction. The statistics of interest is the variance of the difference of the Z values at locations s and $s+h$, $Var[Z(s)-Z(s+h)]$. This statistic is also referred to as a mean square difference in time-series analysis, and a structure function inprobability.

If $E[Z(s)]=m$, and for each set of random variables Z(s) and $Z(s+h)$ the covariance exists and depends only on the vector h, and $Cov[Z(s),Z(s+h)]=C(h)$ for every s and h, then Z(s) is said to be second order stationary. All Gaussian processes are second order stationary. If Z(s) is second order stationary then $E[Z(s)-Z(s+h)]=0$ and $Var[Z(s)-Z(s+h)]=E[\{Z(s)-Z(s+h)\}^2]$. The *variogram* function is the function $2\gamma(s_1-s2)=Var[Z(s_1)-Z(s_2)]$. $2\gamma()$ is a function of the random process Z(). The function $\gamma(h)=(2\gamma(h))/2$ is the *semivariogram*.

VARIOGRAMS

The variogram is usually expressed in terms of vector $h=s_1-s_2$. This vector can also be expressed in terms of magnitude and direction angle, $h=(L,\theta)$, where L is often referred to as the lag. If $2\gamma()$ is a function only of the lag of h, then the variogram is called isotropic.

If a sample $z(s_1), z(s_2), \ldots . z(s_n)$ is taken from a population of regionalized variables in R^2, every possible pair of points is classified by direction and magnitude. The points $s_1=(5,4)$ and $s_2=(2,4)$ have a difference of h $(L=3,\theta=90°)$. This pair of points is in the same classification as (5,4) and (8,4), but not (4,5) and (4,8), which have a different direction parameter.

The variogram can also be expressed in terms of the covariogram $2\gamma(h)=E[\{Z(s)-Z(s+h)\}^2]=2[C(0)-C(h)]$. The function C() is called a covariogram (or autocovariance function in timeseries analysis). It follows that $C(0)=Cov[Z(s),Z(s)]=Var[Z(s)]$. If $C(0)>0$ then $r(h)=C(h)/C(0)$ is called a correlogram (or autocovariance function in times-series analysis).

The quantity $2C(0)$ (or $2Var[Z(s)]$) is called the sill of the variogram. This sill is the limit of the variogram as the lag increases. The smallest vector r for which $2\gamma(r)=2C(0)$ is the range of the variogram in direction r, and the lag at which the variogram approaches its limit. All pairs of points whose distance is beyond the range are assumed to be independent.

Experimental variograms are estimated from a random sample. If N(h) is the number of sample pairs with classification h, the isotropic method of moments estimate (or classic variogram estimator) is

$$2\gamma(h)=\frac{1}{N(h)}\sum_{N(h)}(z(s)-z(s+h))^2$$

The associated covariogram is

$$C(h)=\frac{1}{N(h)}\sum_{N(h)}(z(s_i)-\bar{z})(z(s_j)-\bar{z})$$

Cressie (1993) presented an alternate robust variogram estimator, which is stable in the presence of outliers:

$$2\gamma(h)=\frac{\left[\frac{1}{N(h)}\sum_{N(h)}\sqrt{(z(s)-z(s+h))}\right]^4}{\left(0.457+0.494/N(h)_1\right.}$$

Both estimators can be calculated with PROC VARIOGRAM.

THEORETICAL VARIOGRAM MODELS

The experimental models are not necessarily suitable for estimation. For kriging, the variogram function must possess certain mathematical properties, and the experimental data must be fitted to theoretical modds. Many valid vartograms have been documented, and all models are expressed in terms of the semivarlogram, and assume that $\gamma(0)=0$. PROC KRIGE2D accepts four theoretical isotropic semivartogram models: Spherical, Gaussian, Exponential, and Power.

All theoretical variogram models are isotropic. An isotropic model assumes the direction angle θ has no influence on the corretation structure, and only the lag parameter is considered. Actual data can have a directional trend, and these spatial processes are called anisotropic. Anisotropic processes can differ in model form, sill, or range, depending on direction. Multiple isotropic variogram modds are used to reflect the anisotropy.

A process with the same sill and form in all directions, but different ranges exhibit geometric anisotropy. The ratio between the lowest range and highest range, and the difference between their two angles can be used to transform the model to an isotropic model suitable for kriging.

A more common type of anisotropy is called zonal isotropy, in which the sill or the form may differ by direction. Geologic processes often are zonal anisotropic. Multiple variograms are used for estimation in the directions indicated, and theoretical variograms are fitted as if each were isotropic.

PROC KRIGE2D can compute estimations for both types of anisotropic processes.

Spatial data may also have a discontinuity close to vector 0. This is called the nugget effect. In mining applications, the presence of pockets of minerals, or nuggets, resulted in high local variation. In a theoretical variogram model, the nugget effect can be contrdled by an additive parameter c_n which effectively shifts the variogram model.

PROC KRIGE2D also allows for a nugget effect in spatial data.

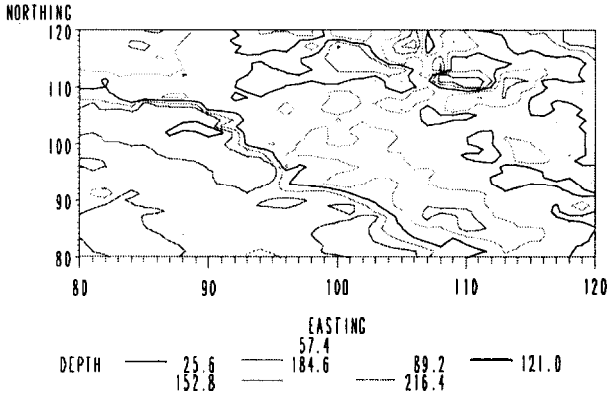
ANALYSIS

EXPLORATORY DATA ANALYSIS

The bathymetric data set for Lake Huron is the water depth in meters sampled on a grid with 2 km spacing with 188 north-south levels and 209 east-west levels for a total of 39,292 points covering 157,168 square kilometers.

For this analysis a 82 km by 82 km section containing 1,681 data points was selected. A contour plot clearly shows a depth pattern in the Northwest-Southeast direction.

Contour Plot of Sampled Points

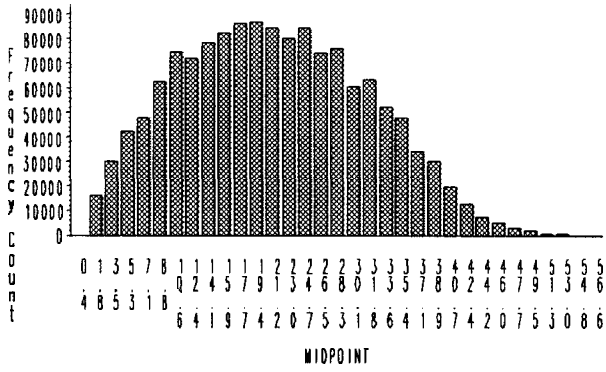


The mean depth for the sample is 112.47, with a variance of 1859.07. The isotropic experimental sill, 2C(0), is 3718.14.

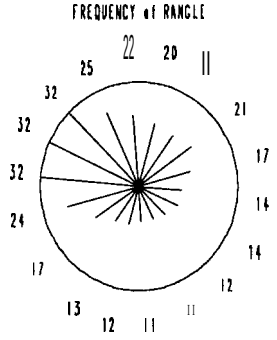
The first step in variogram estimation is to determine the optimal distance unit for each lag. A minimum of 30 pairs of points is needed in each lag. A histogram of distances can help with this process. A rose diagram, or polar plot, of range values by angle class is another good visual toot.

These calculations and graphs are produced by the %GEOEAS macro.

Histogram of Intervals



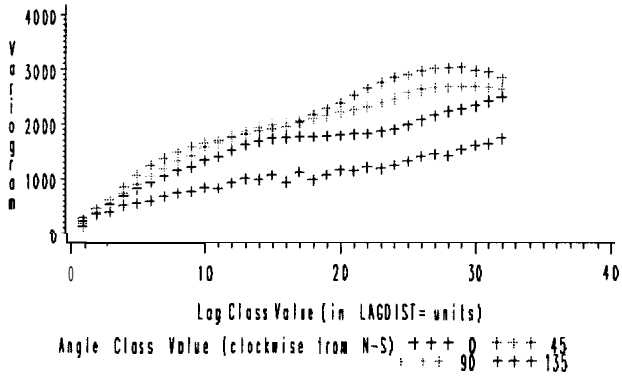
Rose Variogram Plot



The Lake Huron data clearly is an anisotropic process, with maximum range of 35 and minimum range of 11, with a range ratio of 3.2. The angle classes are 0, 45, 90 and 135. The optimum lag distance is 1.6, which results in a maximum of 35 lags.

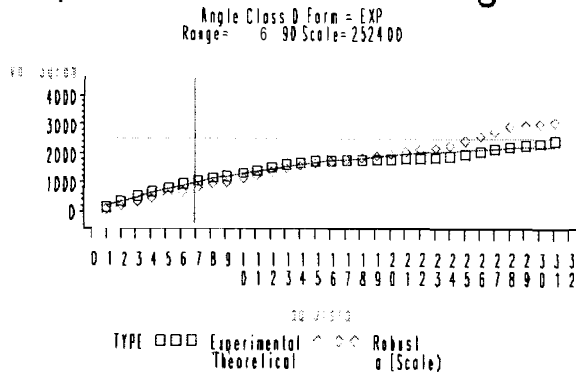
With this information, the experimental variogram can be calculated and plotted.

Experimental Variogram

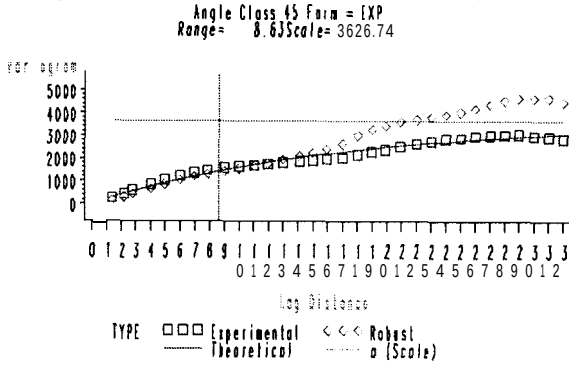


All four experimental variograms have the form of an exponential variogram. The scale and range parameters are obtained from a weighted nonlinear regression of the exponential variogram function using PROC NLIN (Cressie, 1993). Each observation is weighted by $N(h)/\gamma(h)^2$. The macro %FITVARIO simplifies variogram fitting for anisotropic processes, producing the MDATA= data set required by PROC KRIGE2D.

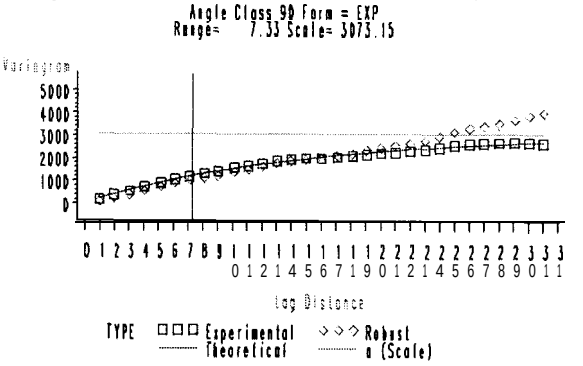
Experimental and Fitted Varioarams



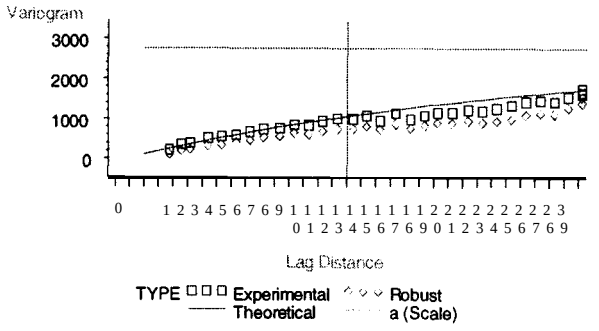
Experimental and Fitted Variograms



Experimental and Fitted Variograms



Experimental and Fiied Variograms



LOCAL ORDINARY **KRIGING**

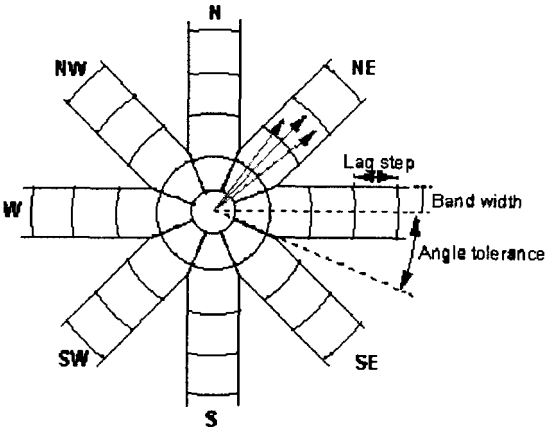
Kriging refers to estimation techniques using the theoretical variogram and covariograms. This term was coined referring to G.H. Krige, a South African mining engineer who used similar methods in the early 1950's. Kriging uses weighted linear combinations of the sample data to estimate block or point data.

Ordinary kriging finds the best linear unbiased estimator for the point to be estimated.

$$\hat{Z}(s_0) = \sum_{j=1}^{N(h)} \lambda_j \cdot Z(s_j), \sum_{j=1}^{N(h)} \lambda_j = 1$$

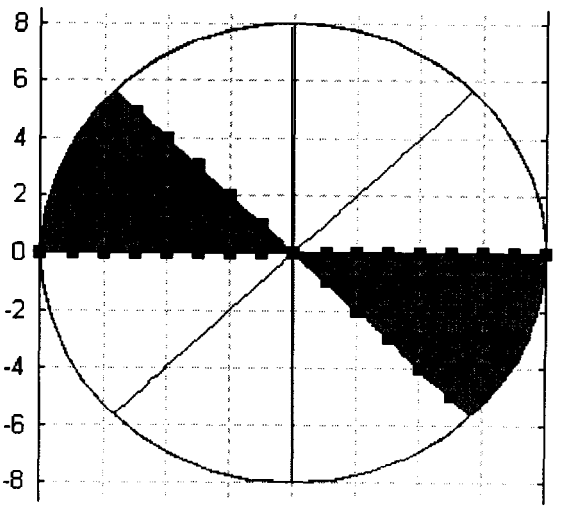
The estimator must also minimize the mean-square error, which results in the equation $C\lambda_0 = C_0$, C is a matrix of known covariograms, and C_0 is a vector of covariograms with the unsampled point. Solving for λ provides the solution.

The specific technique of local ordinary kriging limits the sample points used in the estimation matrix to a predetermined distance radius around the estimated points. The kriging radius should contain at least 30 points for a "good" estimate.

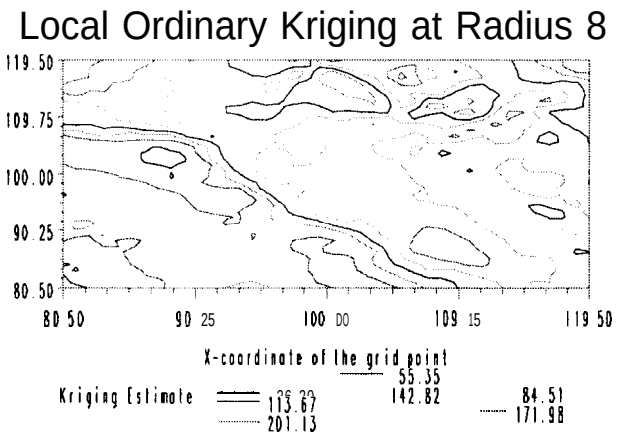


(Image from Sharov)

The take Huron data set is evenly spaced a 1 unit (2 km), and each variogram covers 45°, so a minimum kriging radius of 6 units is required. A krtging radius of 8 units, or 16 km is sufficient for this application.



The final data should have depth estimates for every 1 km, so the grid for estimation is 81 to 119 by 2 in both north and east directions. The resulting contour plot shows the new patterns.



CONCLUSION

Geostatistics is a collection of techniques which model processes using inherent spatial relationships between the data. The potential areas of application for geostatistics include environmental monitoring, pollution control, mining and petroleum engineering, agricultural experiments, or any area where spatial dependence affects a process.

SAS/STAT Version 6.12 provides two very powerful procedures, PROC VARIOGRAM and PROC KRIGE2D. These procedures, in combination with other SAS toots, make a versatile modelling environment for any project with data spatially dependent in two dimensions.

CONTACT INFORMATION

Contact the author at:

Kate Ricci
Owen Analytics, Inc.
500 Main Street
Deep River, CT 06417
Work Phone: 860-526-2222
Email: akricci @owenanalytics.com

MACROS

```
%macro geoeda(
dsn=_last_,xc=xc,yc=yc,geovar=,gcat=gseg);
%*************************************
% Geostatistical Exploratory Data Analysis
%*************************************
% Parameters:
%       DSN = Input Data Set (default=-last-)
%       XC = X Coordinate
%       YC = Y Coordinate
%       GEOVAR = geostatistic variable
%       GCAT = Catalog for graphs
(default=WORK.GSEG)
%*************************************
%Original Author: A. K. Ricci, May 1997
a*************************************.
proc variogram data=&dsn outdistance=outd;
  compute novariogram nhclasses=&classes.;
  coordinates xc=&xc. yc=&yc.;
  var &geovar.;
run;

data outd;
  set outd;
  midpoint=round((lb+ub)/2,.1);
  range=ub - lb ;
run ;

title "Variogram Interval Estimation of
&classes. Classes";
proc print noobs;
run;

title "Histogram of Intervals";
proc gchart gout=&gcat;
  vbar midpoint /type=sum sumvar=count
discrete name="Histogram"
description="Histogram of &Classes.
Intervals";
run ;

proc variogram data=&dsn outvar=outv ;
  compute lagd=&lagdist.maxlag=&classes
ndir=18 robust;
  coordinates xc=&xc. yc=&yc.;
  var &geovar.;
run;

data star;
 retain range c0 0;
 set outv;
   by angle;
  if angle=. then c0=covar;
  if angle>=0;
  if first.angle then range=0;
  if variog<=c0;
  if count>0;
 run;

%* Make a complete data set. Must reverse
angle because Star plots are counter
clockwise ;
data star;
  set star; by angle;
  if last.angle;
  output;
  angle+180;
  rangle=360-angle;
  label rangle="Angle";
  output;
```

```
run ;

proc sort data=star; by angle; run;

title "Rose Variogram Plot";
proc print data=star(where=(angle<180)); run;

proc gchart data=star gout=&Gcat.;
   star rangle/angle=85 freq=lag discrete
   starmax=&classes. noconnect slice=none
   value=outsize name="Rose"
   Description="Rose Diagram of &classes.
intervals";
run ;
%mend geoeda;

%MACRO FitVario
(DSN=_LAST_,MODEL=model,  VARIOCTL=varioctl,Angle=,
Form=,gcat=gseg);
%*********************************************
%* Fit a variogram for an angle and Form
%*********************************************
% Parameters:
%       DSN =  Input data set (default=-last-)
%       MODEL= Output Model Data Set
%       VARIOCTL = Variogram Control Data Set
%       ANGLE = Angle of Variogram
%       FORM = Variogram Form (SPH,PW,EXP,GAUSS)
%       GCAT = graphics output catalog
(default=work.gseg)
%*********************************************
% Original Author: A. Katherine Ricci, May 1997
%*********************************************
Title "Nonlinear Regression for Angle Class &angle.
Form &FORM.";

%* varioctl is the control data set for the entire
experimental variogram, and covar is the sample
covariance ;

data _null_ ;
  set &varioctl;
  c0=covar*2;
  format c0 covar comma7.2;
  call symput('C0',c0);
  call symput('semic0',covar);
run;

%* Find the starting range and scale;
data _null_;
  set &DSN(where=(variog<=&semic0.));
    format lag variog comma7.2;
    if variog^=. ;
    call symput('a0',lag);
    call symput('c1',variog);
run;

%* Fit the variogram, and save the results;
proc nlin data=&dsn(where=(distance>0))
  method=DUD best=3 maxiter=200 save nohalve
  outest=est&angle.(where=(_TYPE_="FINAL"));
  parms c0=&c1. a0=&a0. ;

  %IF &FORM^=PW %THEN %do;
  bounds 1<=c0<=&c0. , 1<=a0<=&classes. ;
   %end;
  %else %do;
  bound 1<=c0 ,  1 <=a0<=&classes;
  %end;
  %IF &FORM=EXP %THEN
    %DO;
  a1=1/a0;
  expon=exp(-distance*a1);
  model variog=c0*(1-expon);
  _weight_ = count/((c0*(1-expon))**2) ;
    %END;

  %IF &FORM=GAUSS %THEN
    %DO;
  a1=1/a0;
  expon=exp(-1*(distance*a1)**2);
  model variog=c0*(1-expon);
  _weight_ = count/((c0*(1-expon))**2) ;
    %END;

  %IF &FORM=PW %THEN
    %DO;
  model variog=c0*distance**a0; _weight_ =
count/((c0*distance**a0)**2)
    %END;

  %IF &FORM=SPH %THEN
    %DO;
  if distance<A0 then do;
     model variog=c0*((3/2)*(distance/a0)-
(1/2)*(distance/a0)**3);
_weight_=count/(c0*((3/2)*(distance/a0)-
(1/2)*(distance/a0)**3))**2;
     end;
  else do;
     model variog=c0;
     _weight_ = count/(c0**2);
     end;
  if (_OBS_=1 and _MODEL_=1) then Do;
    sill = c0; put a0=sill;
  end;
    %END;
run ;

%* Create the MDATA= model data set for proc
krige2d ;
data &model(keep=scale range angle ratio
form);
  set est&angle.;
  format scale range comma8.2 ;
  scale=c0;
  call symput("SCALE&angle",put(scale,8.2));
  range=a0;
  hrange=a0/2;
  call symput("RANGE&angle",put(hrange,8.2));
  angle=&angle.;
  ratio=1E8;
  form="&FORM";
run ;

%* Create a hold dataset for the variogram
with fitted values;
data &DSN ;
  merge &DSN &MODEL(keep=angle scale range);
  by angle;
  %IF &FORM=EXP %THEN %do;
     fvariog=scale*(1-exp(-distance/range));
  %END;
  %IF &FORM=GAUSS %THEN %do;
     fvariog=scale*(1-exp(-
(distance/range)**2));
  %END;
  %IF &FORM=PW %THEN %do;
     fvariog=scale*(distance**range);
  %END;
  %IF &FORM=SPH %THEN %do;
     if distance<=range then
fvariog=scale*((3/2)*(distance/range)-
```

```
(1/2)*(distance/range)**3);
     else
       fvariog=scale;
  %END;
run:

%* For the graph, find the highest lag and
variogram;
proc summary data=&DSN noprint nway;
  var variog fvariog rvario;
  output out=maxvarimax=;
run ;

data -null-;
  set maxvari;
  format vari cormnall.4 ;
  vari=&&scale&angle. ;
  if variog>vari then vari=variog;
  if fvariog>vari then vari=fvariog;
  if rvario>vario then vari=rvario;
  call symput('maxvari',vari);
  varunit=floor(vari/20);
  call symput('varunit',varunit);
run ;

%* Set the Graphing parameters ;
axis1 minor=none label=(c=green 'Lag Distance')
offset=(1,1)
      order=(0 to &classes. by 1) ;
*axis2 minor=(number=1) major=(number=2) order=(0 to
&maxvari. by
&varunit.)
      label=(c=green 'Variogram') offset=(1,1) ;
axis2  label=(c=green 'Variogram') offset=(1,1) ;

data plotdata;
  set &DSN;
  vari=variog ;  type='Experimental'; output;
  vari=fvariog ;  type='Theoretical'; output;
  vari=rvario ; type='Robust'     ; output;
  vari=scale ;  type='a (Scale) '; output;
run;

symbol1 i=none l=1 v=square c=blue ;
symbol2 i=none l=1 v=diamond c=green ;
symbol3 i=join l=1 v=none c=red ;
symbol4 i=join l=1 v=none c=green ;

Title "Experimental and Fitted Variograms";
Title2 "Angle Class &angle. Form = &FORM.";
Title3 "Range=&&RANGE&ANGLE. Scale=&&SCALE&ANGLE.";
run;

proc gplot data=plotdata gout=&gcat.;
    plot vari*distance=type /
        vaxis=axis2 haxis=axis1
       HREF=&&RANGE&ANGLE CHREF=red
        name="Vario&angle" description="Experimental
Variogram for &angle.";
run ;
%mend fitvario;

%macro krige(DSN=_LAST_,EST=,MODEL=,XC=,YC=
       ,GEOVAR=,GRID=,RADIUS=,MINPOINT=8,GCAT=gseg);

Title "Local Ordinary Kriging at Radius &RADIUS.";
proc krige2d data=&DSN outest=&EST.;
  pred var=&GEOVAR. radius=&radius.
minpoints=&minpoint. ;
  model mdata=&MODEL. ;
  coord xcoord=&XC. ycoord=&YC. ;
  GRID &grid;
run;

data valid;
  set &DSN. &EST.(rename=(gxc=&XC.
gyc=&YC.));
run;

proc gcontour data=&est gout=&GCAT.;
  plot  gyc*gxc=estimate /
        name="Krige"
        Description="Contour Plot of Krige
Estimates";
run:

Title "Contour Plot of Sampled Points";
proc gcontour data=&dsn gout=&GCAT.;
  plot &yc*&xc=&geovar /
        name="Sample"
        Description="Contour Plot of Sample
Data";
run;
%mend krige;
```

REFERENCES

Cressie, N.A.C., Statistics for ***Spatial Data,*** New York: John Wiley **&** Sons, Inc., 1993

Gill, A., ***Geostatistics,*** Groundwater Group, Adelaide University, Australia, http://www.maths.adelaide.edu.au/Applied/UA_DAM_FLUIDS/GROUNDWATER/GEOSTATS/geostats.html

Ingram, P., ***Introduction*** to ***Geostatistics,*** Macquarie University, Sydney Australia, http://atlas.es.mq.edu.au/users/pingram/geostat.html

Isaaks, E.H. and **R.M. Srivastava,** *An **Introduction** to Applied **Geostatistics,*** New York: Oxford University Press, 1989

Journel, A.G. and Huijbregts, Ch.J., *Mining Geostatistics, New* York: Academic Press, 1978

Lang, C., Kriging Interpolation, New York: Department of Computer Science, Cornell University, 1995, http://www.tc.cornell.edu/Visualization/contrib/cs490-94to95/clang/kriging.html

SAS Institute, Inc., **SAS/STAT®** Technical *Report: Spatial Prediction using **the SAS®** System,* Cary NC: SAS Institute, Inc., 1996.80 pp.

Schwab, David J. and Sellers, Diana L., NOAA Report ERL GLERL-16 *Computerized Bathymetty and Shorelines of the Great Lakes,* 1996

Sharov, A., *Elements of **Geostatistics,*** Blacksburg, Virginia: Department of Entomology, Quantitative Population Ecology, Virginia Tech, 1996, http://www.gypsymoth.ento.vt.edu/~sharov/PopEcol/lec2/geostat.html

Shibli, S.A.R., *The AI-GEOSTATS Frequently Asked Questions **(FAQ),*** Ispra, **Italy:** Environmental Monitoring Unit of the Joint Research Centre, 1997, http://java.ei.jrc.it/rem/gregoire/ai-geostats_faq.html