

Geostatistical Tools in the SAS® System

E. Barry Moser, Louisiana State University

ABSTRACT

The SAS System provides several procedures for the development of geostatistical spatial prediction models including PROC VARIOGRAM followed by PROC KRIGE2D, PROC MIXED using spatial covariance structures, and PROC MIXED or PROC REG using a trend surface. One can also reproduce and extend these methods using PROC IML. The use of these procedures for geostatistical spatial prediction will be illustrated, along with a variety of supporting graphics that are useful diagnostics in conjunction with these methods.

INTRODUCTION

Geostatistics has focused largely upon the prediction of a continuous random variable at unsampled locations in a continuous field. More recently, the techniques of geostatistics have been used to model spatial dependencies in an explanatory model of a random variable. Extensions of the techniques have also been used to model discrete random variables (see Gotway and Stroup, 1997). Let's begin by specifying a general (linear) model for the response that includes terms for a trend and spatial

$$Z(s) = \mu(s) + \kappa(s) + \eta(s) + \varepsilon(s)$$

associations. Following Cressie (1991), where s are the spatial coordinates of the measurement, $Z(s)$ is the response random variable spatial process, $\mu(s)$ is the non-random trend across the surface, $\kappa(s)$ is a large-scale, zero-mean, stationary spatial process, $\eta(s)$ is a small-scale, zero-mean, stationary spatial process, and $\varepsilon(s)$ is a zero-mean, independent measurement error process. We differentiate $\kappa(s)$ from $\eta(s)$ in that measurements are made near enough together to observe information on $\kappa(s)$, while we are not able to observe information on $\eta(s)$. Thus, $\eta(s)$ shows up as unexplained variation combined with any measurement error. The focus of the spatial modeling is on $\mu(s)$ and $\kappa(s)$. Thus, a simplified version of the model would be

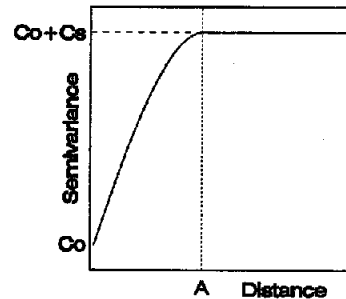
$$Z(s) = \mu(s) + \kappa(s).$$

There are a number of spatial prediction tools that have been developed for this model depending upon what terms are included in the model, and what terms are treated as known and unknown. Simple kriging, for example, would treat $\mu(s)$ as known, a trend surface model would exclude $\kappa(s)$, ordinary kriging fixes $\mu(s)$ as a constant μ , while universal kriging would use all terms. We will treat this model as a mixed model since it has components of both fixed and random effects. In most situations, only a single measurement is made at a given location, and therefore, assumptions are required in order to obtain estimates of errors of estimates and predictions. One assumption is stationarity, and it states that the spatial process doesn't depend upon the location of measurements, but rather upon the distance and direction separating each pair of points. Isotropy further requires that the process be only a function of distance and not direction. Otherwise the process is anisotropic.

THE VARIOGRAM

The variogram is a model for the spatial dependencies among the spatially referenced data. It is based upon the assumption that measurements made near in space will tend to be much more alike than measurements made further apart. In the traditional sense, the ideal variogram is a monotonically non-decreasing function of distance h that is assumed to reach a plateau or

asymptote c_s called the sill at a distance A called the range, beyond which no spatial association occurs. Due to measurement error and small-scale processes, the variogram may be non-zero at a distance of zero. When this occurs, this non-zero value c_0 is called the nugget. The sill may then be redefined as the sum of the nugget c_0 and the relative sill c_s .



A number of functions have been proposed as models for the variogram. The exponential variogram is a simple, yet commonly used function for the variogram that captures the qualities of the ideal variogram. Its functional form is:

$$\gamma(h) = c_0 + c_s [1 - \exp(-h/A)],$$

where $\gamma(h)$ is the semivariance. Additional models can be found in Moser and Macchiavelli (1996) and Cressie (1991). The empirical semivariogram can be computed using the formula of Matheron (see Cressie, 1991:69):

$$\hat{\gamma}(h) = \frac{1}{n(h)} \sum_{i,j \in n(h)} [Z(s_i) - Z(s_j)]^2$$

where $n(h)$ is the number of points separated by a distance h , and summation is over this set of points. In practice, the squared differences are placed into distance categories and summed up over each category. PROC VARIOGRAM can be used to estimate the semivariogram and can be tuned to the number and size of the distance categories used. It will be illustrated later. Next, the theoretical model, e.g., the exponential semivariogram, can be fit using non-linear regression to the empirical semivariogram information. This could be done using PROC NLIN.

KRIGING

Once we have a fitted semivariogram model, we can use it to perform optimal spatial prediction of the random variable at locations that were not sampled. This process of optimal spatial prediction is called kriging. What is interesting about the ordinary kriging estimator is that it is a perfect interpolator because its predictions at observed locations are exactly the value measured at the observed location – this is only false when measurement error is present and multiple measurements are made at a given location or locations. This predictor is also linear and unbiased. The SAS procedure KRIGE2D can be used to produce kriging predictions, as can PROC MIXED. The kriging equations and estimators can be found in Cressie (1991) and Moser and Macchiavelli (1996).

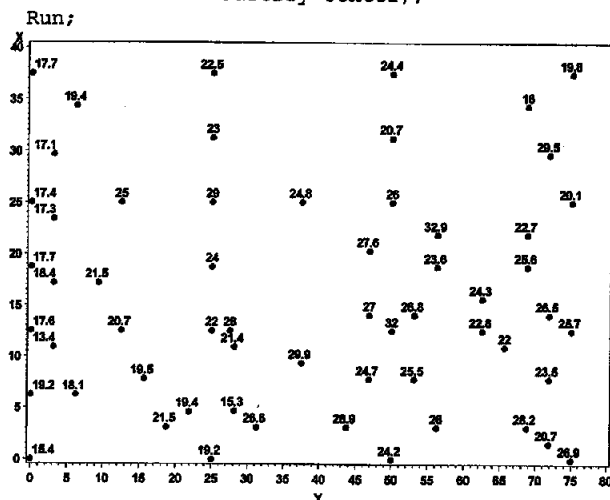
It is important to note that the variogram is an integral aspect of geostatistical prediction. Changes in the range of the semivariogram affect the correlatedness of points a specified distance apart. A large range, relative to the observed distances, implies that measurements tend to be correlated and that predictions will be more like the surrounding points (though the prediction could be quite different from the overall mean) and will have smaller prediction variance; while if the range is small,

measurements tend to be uncorrelated and so predictions tend to be more like the overall mean with a larger prediction variance. Changes in the sill do not affect the predicted values, but do affect the prediction variance.

GEOSTATISTICAL EXAMPLE

Water and Clay content were recorded at 60 locations in a rectangular field of 50m X 100m (Clarke and Dane, 1991). The clay data are presented on the map of sampling locations below using the SAS/GRAPH statements:

```
Proc GPlot Data=WC;
  Plot X*Y=1/Overlay VAxis=Axis1 HAxis=Axis2;
  Axis1 Order=(0 To 40 By 5);
  Axis2 Order=(0 To 80 By 5);
  Symbol1 C=Black V=Dot I=None
  PointLabel=("#Clay" Position=Top
    Justify=Center);
```

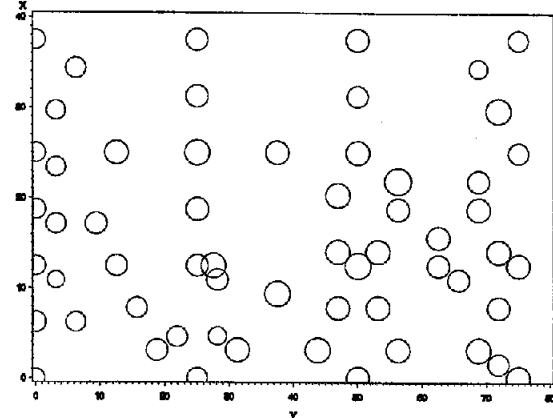


EXPLORATORY ANALYSIS

Before fitting variogram models to the data, it is a good idea to examine the data using an exploratory analysis. A bubble plot of the response can give a quick view of the data without the need for interpolation.

```
Proc GPlot Data=Spatial;
  Bubble X*Y=Clay / BSize=20 BScale=Area
    VAxis=Axis1 HAxis=Axis2;
```

Run;
Quit;



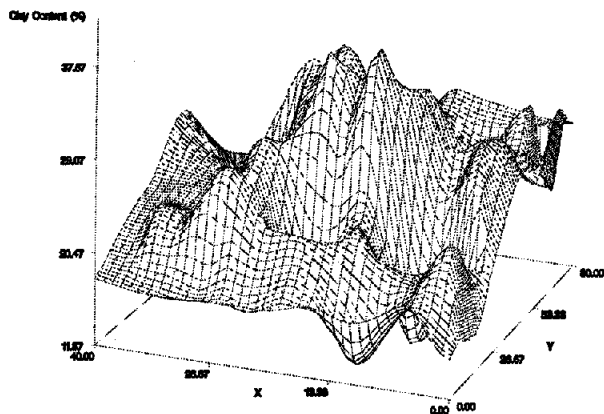
In the bubble plot, the area of the circles is proportional to the value of clay at that specific location. Very strong outliers should show up as unusually large or small bubbles. A polynomial interpolated surface plot might also be useful to get a basic idea of trends and unusual values. The G3GRID procedure provides

the interpolation to a dense grid of points from which the G3D procedure produces the surface plot.

```
Proc G3Grid Data=Spatial Out=Clay;
  Grid X*Y=Clay / Axis1=0 To 40 By 1
    Axis2=0 To 80 By 1;
```

Run;

```
Proc G3D Data=Clay;
  Plot X*Y=Clay;
  Run;
```

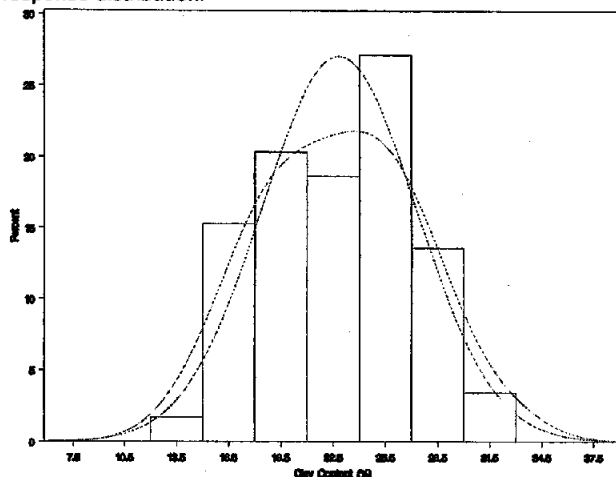


There appears to be some trend along the y-axis direction. One of the decisions that we will have to make is whether to consider this apparent trend as a trend in the mean, $\mu(s)$ from our model, or whether this is simply a piece of a much larger stochastic process, $\kappa(s)$ from our model. If we had additional information about the field, such as its past history, or known larger trends in soil types, the decision might be simplified.

Since we will apply kriging methods to these data, we might like to examine how closely the data conform to our idealized normal distribution model, under which we know that our inferences will be optimal.

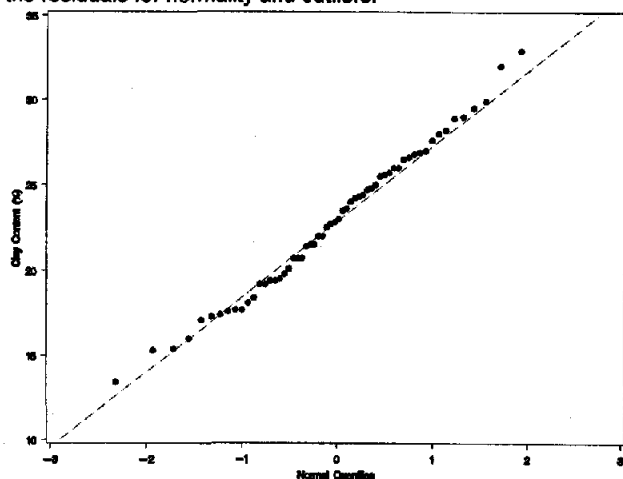
```
Proc Univariate Data=Spatial;
  Var Clay;
  Histogram / Normal(mu=est sigma=est) Kernel;
  QQPLOT / Normal(mu=est sigma=est);
  Run;
```

These statements will produce summary statistics for the data, including lists of extreme values, but will also produce histograms and quantile-quantile plots to help us visually explore the response distribution.



The nonparametric kernel density estimate is not too different from the theoretical curve, while the following Q-Q plot suggests

that normality is not an unreasonable model for these data. It is important to note here that if we suspect a trend in the mean, then we should first remove the trend from the data and examine the residuals for normality and outliers.



BUILDING THE VARIOGRAM MODEL

Once we have examined the data and believe that our data and methods are reasonably compliant, then we can develop a variogram model for the data. Again, if a trend is present, we should first remove it, then proceed with the examination and estimation of the variogram. Before estimating the empirical variogram, we will need to make decisions concerning the number and interval-widths of the distance categories. In addition, this would be an appropriate time to examine the distribution of squared differences between measurements within each distance category to check for outliers that might unduly influence the variogram estimate. Begin by building the distance pairs data set:

```
Proc Variogram Data=SpatialData OutPair=Pairs;
  Coordinates XC=X YC=Y;
  Var Clay;
  Compute NoVariogram NhClasses=20;
Run;
```

Note that the computation of the variogram has been suppressed. The OUTPAIR data set contains the following information:

SAS Data Set: OUTPAIR				
Variable				
1	COS	Num	1	Cosine of Angle Between Pairs
2	DISTANCE	Num	5	Distance between Pairs
3	V1	Num	5	Variable Value for First Point in Pair
4	V2	Num	5	Variable Value for Second Point in Pair
5	VARIABLE	Char	4	Variable Name
6	X1	Num	5	X-Coordinate of First Point in Pair
7	X2	Num	5	X-Coordinate of Second Point in Pair
8	Y1	Num	5	Y-Coordinate of First Point in Pair
9	Y2	Num	5	Y-Coordinate of Second Point in Pair

From this data set, one could use a number of tools, including SAS/INSIGHT to examine histograms of the distances to help select a reasonable set of distance categories. From this data set a square-root plot of the differences (Cressie, 1991:41) can be constructed to check for outliers:

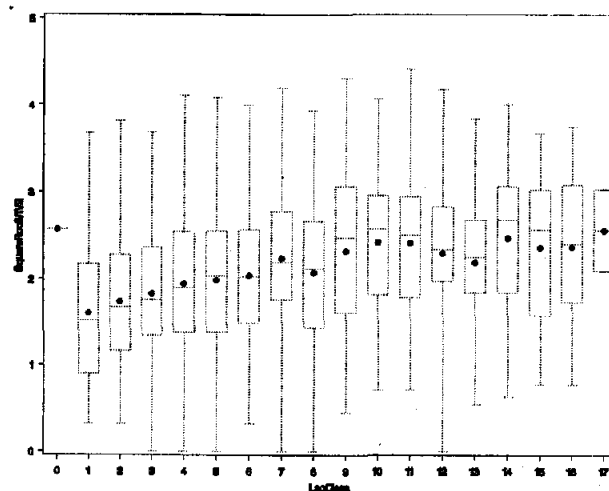
```
Data SquareRoot; Set Pairs;
  LagClass=Int (Distance/5+0.50001);
  SquareRootV1V2=Sqrt (Abs (V1-V2));
Run;
```

The results are plotted as boxplots

```
Proc Sort Data=SquareRoot;
  By LagClass;
```

Run;

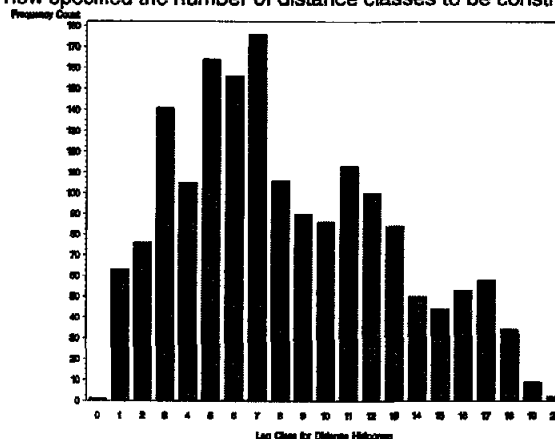
```
Proc BoxPlot Data=SquareRoot;
  Plot SquareRootV1V2*LagClass;
Run;
```



Note the large range of differences over the distance lag classes. To explicitly examine the distribution of numbers of differences within each class the following code can be used:

```
Proc Variogram Data=SpatialData
  OutDistance=Distance;
  Coordinates XC=X YC=Y;
  Var Clay;
  Compute NoVariogram NhClasses=20;
Run;
Proc GChart Data=Distance;
  VBar Lag / SumVar=Count Discrete;
Run; Quit;
```

Again, no variogram estimates are requested, though we have now specified the number of distance classes to be constructed.



What this shows is that the empirical variogram estimates obtained for each distance lag class will be based on very different numbers of distance pairs, and so the quality of each estimate will vary over the distances. We will use this information later in our estimate of the variogram.

Once we have settled upon the number and size of the distance lag classes, we can request computation of the empirical semivariogram.

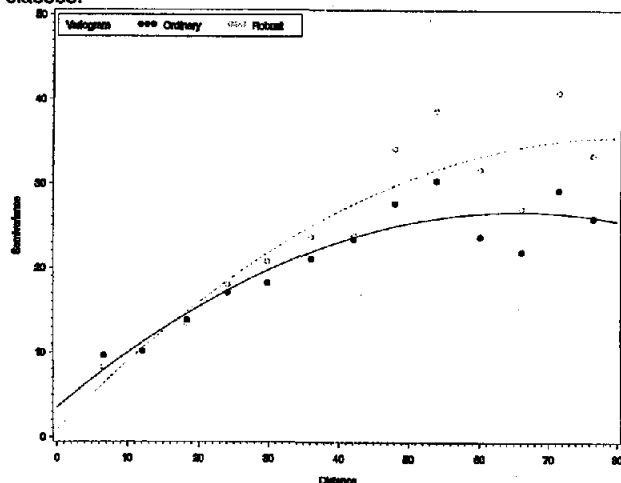
```
Proc Variogram Data=SpatialData
  OutVar=Variogram;
  Compute LagDistance=8 MaxLags=10
  NDirections=0 Robust;
  Coordinates XCoord=X YCoord=Y;
```

```
Var Clay;
Run;
```

Here, both the standard and robust estimators of the semivariogram will be computed. In addition, isotropy has been required since the number of directions classes is set to zero. A listing of the OUTVAR data set contains columns for the number of distances, the average distance, and the semivariance, covariance, and robust semivariance estimates for each lag class.

OUTVAR	VARNAME	CLASS	COUNT	AVERAGE DISTANCE	SEMIVARIANCE	COVARIANCE	ROBUST SEMIVARIANCE
1	Clay	1	34	21.8227	17.5793		
2	Clay	2	17	13.8891	23.2523	12.5890	17.5623
3	Clay	3	108	8.2734	22.0298	8.0273	17.5498
4	Clay	4	269	5.0173	23.1442	1.9623	17.5391
5	Clay	5	328	3.3525	23.5288	1.1098	17.5281
6	Clay	6	245	2.0225	23.2798	1.4787	17.5158
7	Clay	7	179	28.8387	23.1443	1.2284	17.5043
8	Clay	8	201	17.8318	23.7224	20.1251	17.4922
9	Clay	9	124	25.7225	23.1443	18.8451	17.4803
10	Clay	10	85	35.9953	21.2348	16.2284	17.4684
11	Clay	11	41	71.9906	18.8451	14.1812	17.4569
12	Clay	12	22	10.9808	24.1543	14.8080	17.4452

This information can then be used as input to PROC NLIN to obtain parameter estimates for specified variogram models. To get an idea as to the model or models that we might try, we should first plot the empirical semivariograms. The plot below is based upon 15 lag classes.



It appears that these data conform reasonably well to the ideal variogram, though it is common for the empirical semivariogram not to be monotonically increasing.

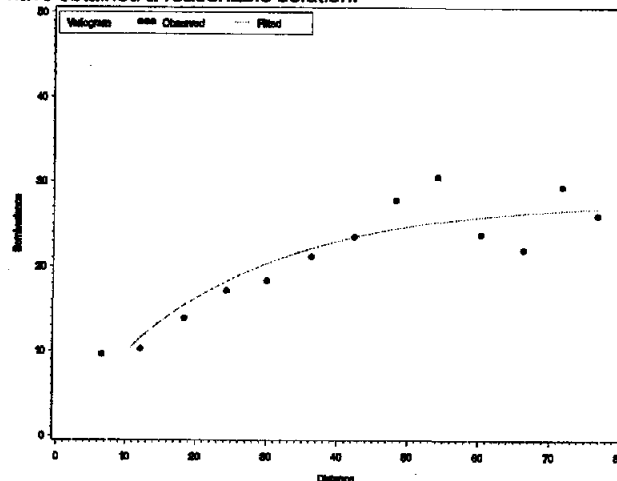
The NLIN procedure can be used to fit, for example, the exponential variogram model to the empirical semivariogram. In addition, weighted least-squares, using the WEIGHT statement, can be used to deal with the unequal class sizes and different variances among classes – though it doesn't account for the dependence among the squared differences.

```
Proc NLIN Data=EmpiricalVariogram;
  Params A=20,30,40 C=12,15,18;
  Semivariogram=C*(1-Exp(-Distance/A));
  Model Variog=Semivariogram;
  _Weight_ = Count/Semivariogram**2;
  Der.A=-(C*Distance*Exp(-Distance/A))
    /(A*A);
  Der.C=1-Exp(-Distance/A);
Run;
```

The resulting estimates are:

Estimator	Estimate	Std Error	Approx. 95% Confidence Limits
A	22.8302	4.5378	12.8424 32.8178
C	27.7858	2.8443	21.4450 33.9663

Thus, we estimate the range to be 23 meters, and the sill to be 28. A graphical assessment of the results can help verify that we have obtained a reasonable solution.



MAKING THE KRIGING PREDICTIONS

In order to use the kriging methods, we will need a point or set of points at which predictions are to be made. Recall that predicting to locations at which measurements were made does not yield new insight because kriging is a perfect interpolator. To illustrate this result and to show how to make predictions at desired locations, a data set containing a new point and an old point, a location for which we have a measurement, is built:

```
Data ToBePredicted;
  Input X Y Water Clay;
Datalines;
16.25 6.25 . . Old Point
12.00 17.00 . . New Point
;
```

Next, KRIGE2D is used to perform the kriging computations:

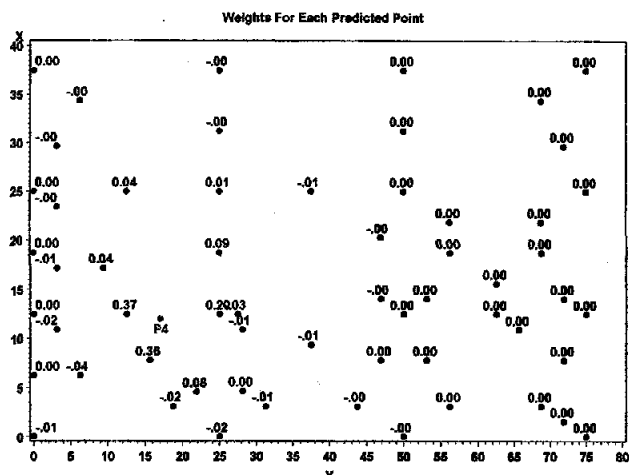
```
Proc Krige2d Data=Simple
  OutEst=Results(Rename=(GYC=Y GX=C));
  Pred Var=Clay;
  Model Scale=29.2 Range=25.8 Nugget=0 Form=Exp;
  Coord XC=X YC=Y;
  Grid GridData=ToBePredicted XC=X YC=Y;
Run;
```

Note that specification of a semivariogram model, along with parameters for the semivariogram is required. The GRID statement is used to identify a data set containing locations at which predictions will be made. The results are:

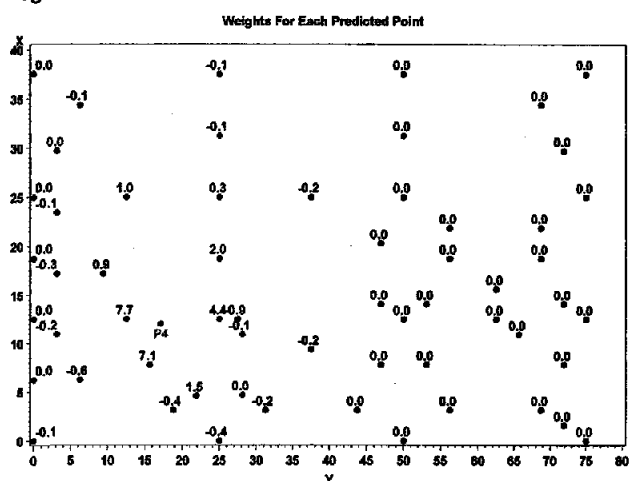
X	Y	Kriging Estimate	Standard Error Kriging Estimate
16.25	6.25	18.1000	0.00000
12.00	17.00	19.2527	2.35809

Additional insight into how the observed spatial data contribute to the predictions can be obtained by examining the kriging weights associated with each observed point. For the first point in the table, the kriging weight is 1 for the observed value of 18.1, and 0 for all other points. For the second point, however, the graph illustrates how the range of the semivariogram modifies the contributions of the surrounding points. Our new point is labeled as P4 in the graph and the kriging weights are printed alongside

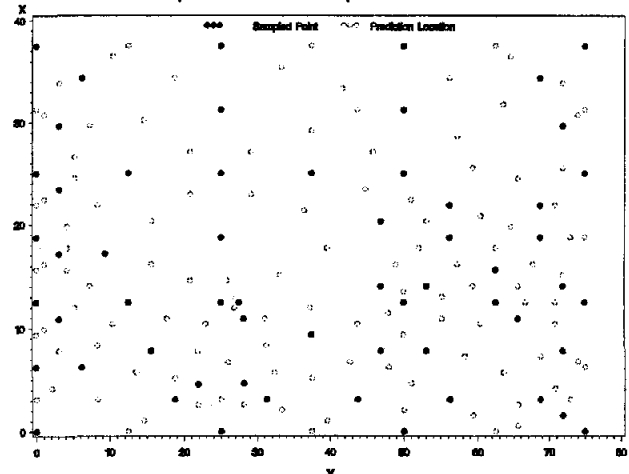
the observed points. Note that the prediction is largely a function of 3 nearby points.



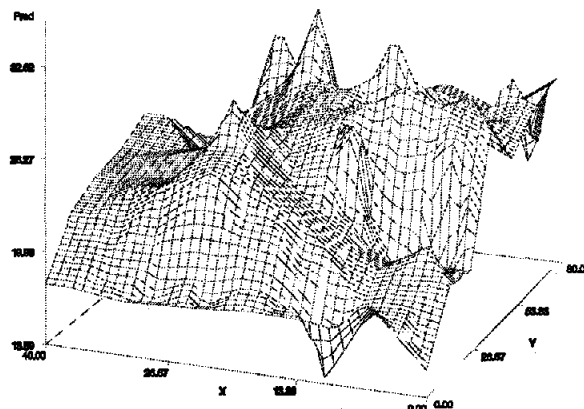
The actual contributions to the prediction can also be graphed. It is important to recognize that weights and contributions can be negative.



For generating the surface, a number of new points were located throughout the field at which predictions are to be made. These are indicated as open circles in the plot below.



These data are then operated on using the KRIGE2D procedure to produce a data set of predicted values. These can be combined with the original data and used to produce a surface map of clay content.



The positive and negative spikes might be of interest to the agronomist and could be explored with additional field measurements. You should also note that this surface is a bit different from the polynomial interpolated surface produced earlier to get an initial view of the spatial data.

SPATIAL PREDICTION WITH PROC MIXED

In addition to KRIGE2D, the above spatial predictions can be produced using PROC MIXED. However, MIXED can be used to estimate the semivariogram and compute the predictions in one step. Note, however, that you will lose a lot of the diagnostic information on the semivariogram using this approach. To predict to new points, create a data set that contains the original data, plus contains observations corresponding to new points to be predicted. For the new points, the response variable is set to missing. Next, MIXED is called with an intercept-only model, and a spatial structure is specified for the covariance matrix of errors. Note that the entire field is treated as a single subject (SUBJECT=INTERCEPT).

```
Proc Mixed Data=Spatial2 Method=REML;
  Model Clay = / OutP=Estimates;
  Repeated / Subject=Intercept
    Type=SP(EXP) (X Y);
  Parm (5 To 35 By 5) (10 To 40 By 5);
Run;
```

It is very important to specify starting values (PARMS) for the semivariogram parameters as the non-linear estimation method may appear to converge, but far away from the solution. Thus, you want the algorithm to search in a close neighborhood of the solution. Above, a grid of starting values is specified in the hopes of covering the solution somewhere inside the grid. MIXED will pick a starting point from this grid based upon the point with best fit. Results of this fit are:

Covariance Parameter Estimates		
Cov. Parm	Subject	Estimate
SP(EXP)	Intercept	0.0302
Residual		0.0084

The range and sill, as estimated by MIXED and NLIN will differ as they do here. With NLIN we used a weighting scheme, distance classes, and excluded extreme distance classes with few points. MIXED uses all of the pairs of distances and does not construct distance classes. One also has to insure that the same parameterizations of the models are being used in both methods for direct comparisons.

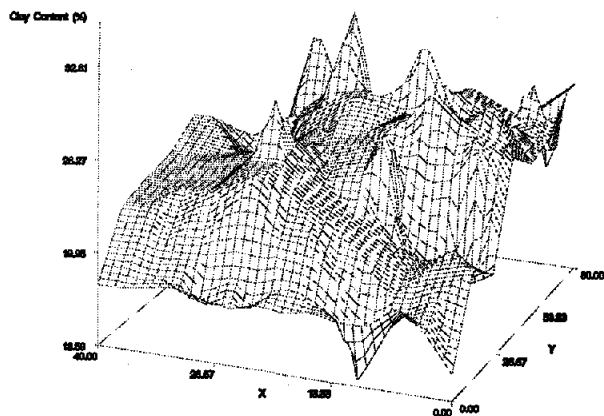
A major advantage that MIXED has over KRIGE2D is the ability to include additional factors into the model, especially with respect to the mean. An extension of the above mixed model would be to add a simple trend surface.

```
Proc Mixed Data=Spatial;
  Model Clay = X Y X*Y / Solution OutP=Clay;
  Repeated / Subject=Intercept
    Type=SP(EXP) (X Y);
  Farms (5 To 35 By 5) (10 To 40 By 5);
Run;
```

Source	Sum of Squares	Mean Square	DF	Pr > F
Corrected Total	15.0822		100	
Model	15.0822		3	
Error	0.0000		97	
Total	15.0822		100	

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	18.1015	2.1043	97	8.60	<.0001
X	0.08402	0.09632	97	0.87	0.3869
Y	0.1415	0.04613	97	3.06	0.0034
X*Y	-0.00318	0.002061	97	-1.52	0.1346

The prediction surface resulting from this model is practically the same as for the ordinary kriging predictions. The standard errors of prediction, however, are a bit different (not shown). Notice that the trend is found only in the y-direction.



CONCLUSION

The SAS System provides a rich environment for modelling spatial data. The relatively new procedures VARIOGRAM and KRIGE2D specifically address spatial prediction, but enhancements to MIXED and its related macros for generalized linear modeling (%GLIMMIX) and non-linear modeling (%NLINMIX) provide considerable flexibility in modelling spatial data. For algorithms not contained in these procedures, such as cokriging, IML provides a programming environment within which these algorithms can be written. The associated diagnostic and graphical procedures permit extensive checks on data integrity and model appropriateness, and provide excellent means of displaying results.

REFERENCES

- Clarke, G.P.Y. and J.H. Dane. 1991. *A simplified theory of point kriging and its extension to co-kriging and sampling optimization*. Bulletin 609 of the Alabama Agricultural Experiment Station, Auburn Univ., Auburn, AL, 44pp.
- Cressie, N. 1991. *Statistics For Spatial Data*. John Wiley & Sons, Inc., New York, 900p.
- Gotway, C. A. , and W. W. Stroup. 1997. A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 157-178.
- Moser, E.B., and R. E. Macchiavelli. 1996. Methods for Spatial Analysis. Pages 91-113 in V. P. Singh and M. Fiorentino (eds). *Geographical Information Systems in Hydrology*. Kluwer Academic Publishers, Netherlands.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

E. Barry Moser
Louisiana State University
Baton Rouge, LA 70803-5606
Work Phone: 225-578-8376
Fax: 225-578-8344
Email: bmoser@lsu.edu
Web: <http://www.stat.lsu.edu/faculty/moser/>