Paper 270-25

A primer in spatial simulation with PROC SIM2D

Robert G. Downer, Louisiana State University, Baton Rouge, LA

ABSTRACT

Spatial simulation is a valuable tool for investigating the behavior of spatial random variables. The new SIM2D procedure in SAS simulates normal spatial data with a variety of possible covariance structures. The simulations can be unconditional or conditional on the observed sample data. This general paper motivates spatial simulation, describes and illustrates the SIM2D procedure and also discusses its possible applications.

1. MOTIVATION

Spatially referenced data comes in many forms and arises in many subject areas. Knowledge of the long-run behavior of the variable(s) can be very beneficial with respect to understanding the underlying process and the context of a recently obtained set of observations. With a reasonable assumed model for the data, spatial simulation can give us valuable insight for the entire region which isn't possible through the limited spatial coverage of the sample.

Consider the following example where we have phosphorous concentrations in a field with the values and locations as displayed below by PROC G3D.

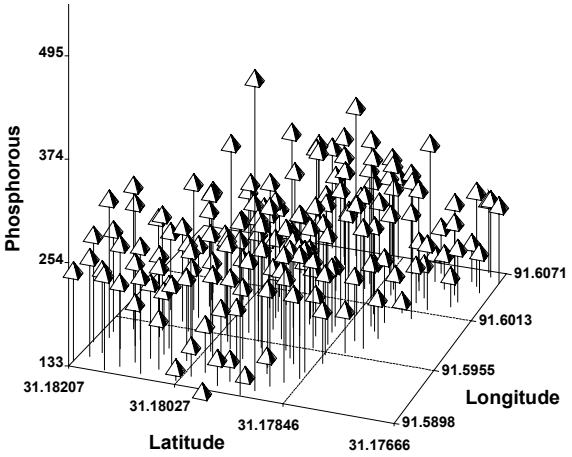

Figure 1: Scatterplot of Phosphorous Concentrations

We're interested in knowing whether the observations imply that the phosphorous level for the entire field is high this year relative to past years. If it is indeed high, expensive fertilizer may not be uniformly applied as it is not required to produce the desired yield.

A simplistic approach assumes the concentrations are independent normal observations from the historical distribution with known mean and standard deviation. We obtain the sample average of the recently sampled points and state that we have evidence that the mean concentration is now higher based on a simple Z test. One problem with this approach is that it ignores the spatial dependence of the observed data. Similar values of spatial random variables tend to occur close together geographically (i.e. exhibit positive spatial correlation). The sampling error in the probability calculation is based on the sampling distribution of the sample mean. Spatial sampling error (how representative is the sample of the entire region) is not considered.

An improved approach would be to use spatial prediction methods (e.g. kriging) to estimate the concentration at other points in the region. The spatial covariance structure is involved but the standard errors of all the predicted values would have to be incorporated into some aggregate assessment of whether the region as a whole has a high nutrient level.

A practical way of evaluating the observed sample is through simulation. A conditional simulation would assume that the observed data (which has an estimated mean and covariance structure) is a part of a realization of a multivariate normal random variable. The distribution which generates the random variable at other points on the grid is conditional on the observed data. Unconditionally, one could simulate using a historical mean and covariance matrix and consider the observed sample within the many realizations possible through simulation.

2. SPATIAL VARIABILITY

Spatial variability, the variation among observations in space is a focus of study in subject areas such as the agricultural, earth, health and social sciences. Representative examples of spatial patterns which are of interest to researchers and policy makers are: wheat yields, groundwater contamination and electricity usage.

Positive spatial dependence or positive spatial autocorrelation occurs when neighboring observations tend to be more similar than observations further apart. A common measure of spatial similarity is the variogram, defined as

$$2\gamma(h) = \mathrm{Var}(\mathbf{s}+h) = [\, Z(\mathbf{s}+h) - Z(\mathbf{s}) \,]$$

and estimated as

$$(1/n(h))\, \Sigma_h \, (z_i - z_j)^2$$

where Z is the random variable recorded at spatial co-ordinate **s** and h is their lag distance apart. The estimated variogram at lag h is the average squared deviation of pairs of values that are within h units apart. It is a measure of similarity between pairs of observations as a function of distance.

The semivariogram $\gamma(h)$ is defined as one-half the variogram. The covariance C(h) between two observations h units apart is estimated as

$$1/(n(h))\sum_{i,j}(z_i-\bar{z})(z_j-\bar{z})$$

and can be defined in terms of the semivariogram as C(0)[1 – f(h)] where f(h) is the functional form of the semivariogram model and C(0) = σ^2 is variance of the spatial process and the sill of the variogram. The range r is the distance at which the sill or asymptote of the variogram occurs. Three common covariance functional forms are the exponential, spherical and gaussian which are defined below:

Spherical:

$$\gamma\ (h) = \sigma^2[\ (3h/2r) – (h/2r)^3\ ]$$

Exponential:

$$\gamma\ (h) = \sigma^2[\ 1 – \exp(-3h/r)\ ]$$

Gaussian:

$$\gamma\ (h) = \sigma^2[\ 1 – \exp(-3h^2/r^2)\ ]$$

These semivariogram models are isotropic in that the covariance functional form for pairs of observations h units apart is the same regardless of direction.

3. METHODOLOGY

PROC SIM2D simulates continuous normal data in two dimensions. A continuous random variable Z is geographically referenced with locations ($\mathbf{s}_1,\ldots,\mathbf{s}_n$) each of which consists of an (x,y) co-ordinate. The n locations can be the elements of a regular grid or specific locations selected from the region. Deviates Z(**s**) are simulated from a multivariate normal distribution with mean u(**s**) and covariance matrix Σ. The mean u(**s**) can be specified as a constant (i.e. as a horizontal plane) or as a quadratic surface (as a function of the (x,y) co-ordinates). Each assumes the covariance C(h) of observations h units apart follows a specific form (e.g spherical, exponential, Gaussian) which is a function of distance. Conditional simulations consider the normal random vector **X** to be partitioned as **X** = ($\mathbf{X}_1$,$\mathbf{X}_2$) and is viewed as having mean vector (μ_1,μ_2) and covariance matrix Σ with submatrices Σ_1,Σ_2 diagonal within Σ. The simulated deviates in $\mathbf{X}_1$ are considered conditional on the observed $\mathbf{X}_2$.

The basic syntax for the SIM2D procedure is as follows:

```
PROC SIM2D options
  COORDINATES options
  GRID options
  SIMULATE options
  MEAN options
```

A SAS data set containing (x,y) co-ordinates is specified through the DATA= option. It is required for conditional simulations. The data set which will contain the simulation values, the iteration, variable name and location are specified in an OUTSIM= option.

The COORDINATES statement specifies the x and y coordinates (as XC = variable name, YC= variable name) of the conditioning data set.

The GRID statement allows one to specify the grid spatial locations for the generated simulation values. It can be done explicitly as a SAS data set through an explicit GDATA=option or by specifying grid steps or increments (e.g. X = x_1, x_2,…, x_m or , X= x_1 to x_m by d_x) as in the example below.

The SIMULATE statement is rather comprehensive. It specifies several simulation details including the number of simulations (NUMR=option), the variable in a SAS data set to be used for conditioning (VAR= option) and the covariance model specification. The covariance or semivariogram model specification requires a SCALE, RANGE and FORM parameter. These structures can be abbreviated as SPH, EXP, and GAU in the FORM= option. The scale or sill is the value at which the covariance between pairs of observations levels off and the range is distance at which the covariance function levels off. An optional parameter is the NUGGET, the value of the assumed semivariogram function at lag distance zero). These parameters may be specified explicitly or in a SAS data set through an MDATA= option.

The MEAN statement allows one to specify the mean as a quadratic function of the co-ordinates or simply as a constant. The coefficients β_o through β_5 of the surface (for the intercept, first and second order terms in x and y as well as the cross term xy) can be listed explicitly or through the QDATA=option.

4. EXAMPLE

The phosphorous concentrations shown in Figure 1 show some local variation and similar values do tend to occur together spatially. The total area of the field is only 163.5 acres with latitude ranging from –92.392 to –92.411 and longitude from 31.176 to 31.183. It is not unreasonable to assume the observed variability in concentration is simply variation about a constant mean (over a larger area) and assume strict stationarity (constant mean and variance). For this data set, removing any weak linear trend via ordinary least squares regression leads to a very similar model for the covariance structure. The implied covariance between any two points is assumed to be a function of distance only (isotropy)

A preliminary run of PROC VARIOGRAM with the NODISTANCE option reveals the distribution of point pairs in 10 distance classes. A revised run with 8 classes and a lag distance of .001 gives the following output:

LAG	COUNT	DISTANCE	VARIOG	RVARIO
0	32	.000426955	1388.97	1387.76
1	1154	.001053656	1650.88	1206.83
2	1585	.002020681	1722.31	1487.91
3	1729	.002985041	1728.06	1634.80
4	1432	.003997599	1768.44	1628.76
5	1198	.004995086	1742.49	1577.00
6	1106	.005988691	1797.98	1591.50
7	987	.006987030	1770.30	1509.94
8	900	.007986653	1796.15	1688.63

Both the traditional variogram and robust variogram appear to be levelling off at a distance near .003. An exponential variogram model was fit to the oberved empirical variogram. Each are shown in Figure 2 below.

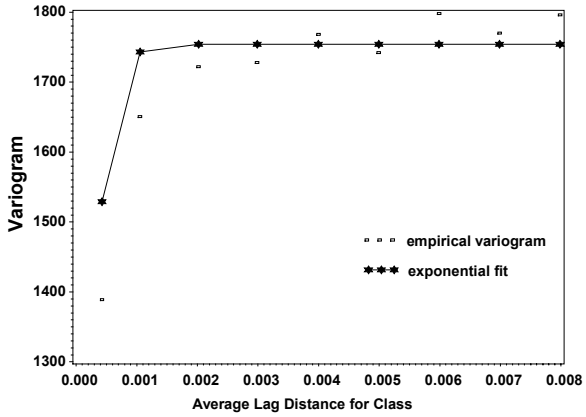

Figure 2: Empirical Variogram and Exponential Fit

The range (of covariance) is estimated at .00208 in cartesian distance which corresponds to approximately 700 feet for this field. This estimated exponential variogram is input for a conditional simulation using PROC SIM2D. The observed phosphorous concentrations are approximately normal with a mean of 235.95. This is included with the input data set to complete the required simulation parameters via the MEAN statement. The latitude and longitude have been transformed by subtracting 30 and adding 92 respectively. The following code performed a simulation of 1000 repetitions.

```
proc sim2d data = oneb outsim = sim1;
coordinates xc = lon yc = lat ;
simulate numr = 1000 form = exp range =
.000208 scale = 1754.2 ;
mean qdata = oneb;
grid x = 1.1765 to 1.1820 by .000174 y =
.17666 to .18207 by .000174 ;
run;
```

The chosen grid specifications resulted in 1024 observations per repetition and hence over 1 million observations in total..

The observed sample had a maximum of 495 . The 99^{th} , 95^{th} ,75^{th} and 5^{th} percentiles were 343, 300 , 283 and 171 respectively. The average maximum value of the 1000 simulations was 370.4 with a simulation standard deviation of 16.The maximum (of maximums) was 441.6 so the outlier in the observed sample is indeed extreme. The average 95^{th} percentile was 304.2 with a simulation standard deviation of 4.7. The observed 95^{th} percentile of 300 was not too unusual. The average third quartile was 264 with a simulation standard deviation of 3.9 so the observed counterpart was indeed high. The average fifth percentile (167.4, standard deviation 4.9) was much closer to the observed. Assuming the estimated covariance structure is reasonable, the phosphorous concentrations for the current year appear to be rather high and less fertilizer will be needed. Variable rate application within the field based on spatial prediction (e.g. via kriging) may be economical.

5. DISCUSSION

Applications of PROC SIM2D will arise in many other subject areas. Simulation of mineral deposits based on observed data would be valuable to geologists, the simulation of pollutant levels would have similar importance to environmental administrators and sales simulations would be just another tool for market analysts.

There are many other possible extensions of the SIM2D procedure which have not been summarized here. One may want to estimate a covariance structure for an entire region and then simulate only for a subregion or vise versa. It may also be appropriate to simulate with different intensity within subregions or only at the observed sample locations.

The specified covariance structure is likely to be from an estimated semivariogram model. The functional form itself may be tentative. The sill, range, and/or nugget will generally have been estimated through the use of other software (PROC VARIOGRAM used above) and their true values may be in a wide interval of possible values. Hence a range of estimated parameters and/or functional forms can be attempted to give a better understanding of possible long term behavior of the spatial process. The mean and covariance structure may be more complex so that anisotropic and/or nested models are required. Anisotropic models involve specified covariance functions which are not global to the region but along certain angular axes while nested models allow subregional covariance structures.

PROC SIM2D can be also used to empirically investigate statistical methodology. For example, the simulation data may be used to consider the distribution of a test statistic such as Moran's spatial correlation coefficient under various covariance structures. The procedure may also be used to test other SAS procedures. SIM2D output data can be used to test linear models which adjust for spatial correlation. The performance of the KRIGE2D procedure and the properties of the kriged estimates may be investigated by using simulations from the SIM2D procedure.

ACKNOWLEDGMENTS

The author would like to thank Steven Moore and Maurice Wolcott at the LSU Agricultural Center's Dean Lee Research Station in Alexandria, Louisiana for providing the data of the example.

CONTACT INFORMATION

Robert G. Downer
Department of Experimental Statistics
161 Ag. Administration Building
Louisiana State University
Baton Rouge, Louisiana
(225) 388-8303
Fax: (225) 388-8344
Email: rdowner@lsu.edu