THE POWER TO KNOW.

SAS/STAT® 9.2 User's Guide

The SIM2D Procedure (Book Excerpt)

SAS® Documentation

This document is an individual chapter from *SAS/STAT® 9.2 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2008. *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st electronic book, March 2008
2nd electronic book, February 2009

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at **support.sas.com/publishing** or call 1-800-727-3228.

Chapter 79

The SIM2D Procedure

Contents

Overview: SIM2D Procedure

The SIM2D procedure produces a spatial simulation for a Gaussian random field with a specified mean and covariance structure in two dimensions by using an LU decomposition technique.

The simulation can be conditional or unconditional. If it is conditional, a set of coordinates and associated field values are read from a SAS data set. The resulting simulation honors these data values.

You can specify the mean structure as a quadratic function in the coordinates. You can specify the covariance by naming the form and supplying the associated parameters.

PROC SIM2D can handle anisotropic and nested semivariogram models. Three covariance models are supported: Gaussian, exponential, and spherical. A single nugget effect is also supported.

You can specify the locations of simulation points in a GRID statement, or they can be read from a SAS data set. The grid specification is most suitable for a regular grid; the data set specification can handle any irregular pattern of points.

The SIM2D procedure writes the simulated values for each grid point to an output data set. The SIM2D procedure now uses ODS Graphics to create graphs as part of its output. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." For more information about the graphics available in PROC SIM2D, see the section "ODS Graphics" on page 6173.

Introduction to Spatial Simulation

The purpose of spatial simulation is to produce a set of partial realizations of a spatial random field (SRF) $Z(\boldsymbol{s}), \boldsymbol{s} \in D \subset \mathcal{R}^2$ in a way that preserves a specified mean $\mu(\boldsymbol{s}) = E[Z(\boldsymbol{s})]$ and covariance structure $C_z(\boldsymbol{s}_1 - \boldsymbol{s}_2) = \text{Cov}(Z(\boldsymbol{s}_1), Z(\boldsymbol{s}_2))$. The realizations are partial in the sense that they occur only at a finite set of locations $(\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n)$. These locations are typically on a regular grid, but they can be arbitrary locations in the plane.

PROC SIM2D produces simulations for continuous processes in two dimensions by using the lower-upper (LU) decomposition method. In these simulations the possible values of the measured quantity $Z(\boldsymbol{s}_0)$ at location $\boldsymbol{s}_0 = (x_0, y_0)$ can vary continuously over a certain range. An additional assumption, needed for computational purposes, is that the spatial random field $Z(\boldsymbol{s})$ is Gaussian. The section "Details: SIM2D Procedure" on page 6168 provides more information about different types of spatial simulation and associated computational methods.

Spatial simulation is different from spatial prediction, where the emphasis is on predicting a point value at a given grid location. In this sense, spatial prediction is local. In contrast, spatial simulation is global; the emphasis is on the entire realization $(Z(\boldsymbol{s}_1), Z(\boldsymbol{s}_2), \cdots, Z(\boldsymbol{s}_n))$.

Given the correct mean $\mu(\boldsymbol{s})$ and covariance structure $C_z(\boldsymbol{s}_1 - \boldsymbol{s}_2)$, SRF quantities that are difficult or impossible to calculate in a spatial prediction context can easily be approximated by functions of multiple simulations.

Getting Started: SIM2D Procedure

Spatial simulation, just like spatial prediction, requires a model of spatial dependence, usually in terms of the covariance $C_z(\boldsymbol{h})$. For a given set of spatial data $Z(\boldsymbol{s}_i), i = 1, \cdots, n$, the covariance structure (both the form and parameter values) can be found by the VARIOGRAM procedure. This example uses the coal seam thickness data that are also used in the section "Getting Started: VARIOGRAM Procedure" on page 7511.

In this example, the data consist of coal seam thickness measurements (in feet) taken over an area of 100×100 (10^6 ft^2). The coordinates are offsets from a point in the southwest corner of the measurement area, with the north and east distances in units of thousands of feet.

Preliminary Spatial Data Analysis

The semivariogram analysis of the thick data set in "Example 95.1: Theoretical Semivariogram Model Fitting" on page 7562 of the VARIOGRAM procedure considered the spatial random field (SRF) $Z(\boldsymbol{s})$ of the thickness values to be free of surface trends. The expected value $E[Z(\boldsymbol{s})]$ is then a constant $\mu(\boldsymbol{s}) = \mu$, which suggests that you can work with the original thickness data rather than residuals from a trend surface fit. In fact, a reasonable approximation of the spatial process generating the coal seam data is given by

$$Z(\boldsymbol{s}) = \mu + \varepsilon(\boldsymbol{s})$$

where $\varepsilon(\boldsymbol{s})$ is a Gaussian SRF with Gaussian covariance structure

$$C_z(\boldsymbol{h}) = c_0 \exp\left(-\frac{h^2}{a_0^2}\right)$$

Note that the term "Gaussian" is used in two ways in this description. For a set of locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n$, the random vector

$$\boldsymbol{Z}(\boldsymbol{s}) = \begin{bmatrix} Z(\boldsymbol{s}_1) \\ Z(\boldsymbol{s}_2) \\ \vdots \\ Z(\boldsymbol{s}_n) \end{bmatrix}$$

has a multivariate Gaussian or normal distribution $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The (i, j)th element of $\boldsymbol{\Sigma}$ is computed by $C_z(\boldsymbol{s}_i - \boldsymbol{s}_j)$, which happens to be a Gaussian functional form.

Any functional form for $C_z(\boldsymbol{h})$ that yields a valid covariance matrix $\boldsymbol{\Sigma}$ can be used. Both the functional form of $C_z(\boldsymbol{h})$ and the parameter values

$$\mu = 40.1173$$

$$c_0 = 7.2881$$

$$a_0 = 30.6239$$

are estimated by using PROC VARIOGRAM and PROC NLIN in "Example 95.1: Theoretical Semivariogram Model Fitting" on page 7562 in the VARIOGRAM procedure. Specifically, the expected value μ is reported in the VARIOGRAM procedure OUTV output data set, and the parameters c_0 and a_0 are estimates derived from a weighted least squares fit.

The choice of a Gaussian functional form for $C_z(\boldsymbol{h})$ is simply based on the data, and it is not at all crucial to the simulation. However, it *is* crucial to the simulation method used in PROC SIM2D that $Z(\boldsymbol{s})$ be a Gaussian SRF. For details, see the section "Computational and Theoretical Details of Spatial Simulation" on page 6168.

Investigating Variability by Simulation

The variability of $Z(\boldsymbol{s})$, as modeled by

$$Z(\boldsymbol{s}) = \mu + \varepsilon(\boldsymbol{s})$$

with the Gaussian covariance structure $C_z(\boldsymbol{h})$ found previously, is not obvious from the covariance model form and parameters. The variation around the mean of the surface is relatively small, making it difficult visually to pick up differences in surface plots of simulated realizations.

Instead, you can compute the mean for each location on a grid from a series of realizations in a simulation. Then, the standard deviation of all the simulated values at each grid location will provide you with a measure of the variability of $Z(\boldsymbol{s})$ for the given covariance structure. You can also investigate variations at selected grid points in more detail, as shown in the "Example 79.2: Variability at Selected Locations" on page 6179.

In the present example you will see how to use ODS Graphics with PROC SIM2D to investigate the mean and standard deviation of simulated values. In particular, you produce 5,000 realizations of a simulation with PROC SIM2D, where you specify the Gaussian model with the parameters found previously. You want the simulated data to pass through the simulated values, so first you define the data with the following DATA step:

```
data thick;
   input East North Thick @@;
   label Thick='Coal Seam Thickness';
   datalines;
    0.7  59.6  34.1   2.1  82.7  42.2   4.7  75.1  39.5
    4.8  52.8  34.3   5.9  67.1  37.0   6.0  35.7  35.9
    6.4  33.7  36.4   7.0  46.7  34.6   8.2  40.1  35.4
   13.3   0.6  44.7  13.3  68.2  37.8  13.4  31.3  37.8
   17.8   6.9  43.9  20.1  66.3  37.7  22.7  87.6  42.8
   23.0  93.9  43.6  24.3  73.0  39.3  24.8  15.1  42.3
   24.8  26.3  39.7  26.4  58.0  36.9  26.9  65.0  37.8
   27.7  83.3  41.8  27.9  90.8  43.3  29.1  47.9  36.7
   29.5  89.4  43.0  30.1   6.1  43.6  30.8  12.1  42.8
   32.7  40.2  37.5  34.8   8.1  43.3  35.3  32.0  38.8
   37.0  70.3  39.2  38.2  77.9  40.7  38.9  23.3  40.5
   39.4  82.5  41.4  43.0   4.7  43.3  43.7   7.6  43.1
   46.4  84.1  41.5  46.7  10.6  42.6  49.9  22.1  40.7
   51.0  88.8  42.0  52.8  68.9  39.3  52.9  32.7  39.2
   55.5  92.9  42.2  56.0   1.6  42.7  60.6  75.2  40.1
   62.1  26.6  40.1  63.0  12.7  41.8  69.0  75.6  40.1
   70.5  83.7  40.9  70.9  11.0  41.7  71.5  29.5  39.8
   78.1  45.5  38.7  78.2   9.1  41.7  78.4  20.0  40.8
   80.5  55.9  38.7  81.1  51.0  38.6  83.8   7.9  41.6
   84.5  11.0  41.5  85.2  67.3  39.4  85.5  73.0  39.8
   86.7  70.4  39.6  87.2  55.7  38.8  88.1   0.0  41.6
   88.4  12.1  41.3  88.4  99.6  41.2  88.8  82.9  40.5
   88.9   6.2  41.5  90.6   7.0  41.5  90.7  49.6  38.9
   91.5  55.4  39.0  92.9  46.8  39.1  93.4  70.9  39.7
   55.8  50.5  38.1  96.2  84.3  40.3  98.2  58.2  39.5
   ;
```

Since this is a conditional simulation, you can specify the OBSERV option in the PLOT option in PROC SIM2D to see the locations and values of the measured points in the area where you want to perform spatial simulations.

Furthermore, the MEANS suboption in the PLOT option specifies that you want to create a plot that shows the means of the simulated values across the region. The MEANS suboption with no other arguments produces a plot that shows the contours of the simulated means in the foreground and the gradient of the simulated standard deviations in the background.

You obtain these PROC SIM2D results at the nodes of an output grid that you specify according to your application needs. In the present analysis, a convenient area that encompasses all the Thick data points is a square with a side length of 100,000 feet. You define a regular grid for your simulation in this area. Assume a distance of 2,500 feet between grid nodes in both directions for a smooth contour plot. Based on this choice, your square grid has 41 nodes on each side. This means that PROC SIM2D computes the simulated values at a total of 1,681 grid points. You use the GRID statement of the PROC SIM2D to specify this grid.

The SIMULATE statement specifies the parameters of your simulation across the output grid. In particular, the VAR= option specifies the conditional simulation variable. The number of realizations in the simulation is specified with the NUMREAL option. The SEED= option specifies the seed for the simulation random number generator.

The spatial correlation model for the simulation is also specified in the SIMULATE statement. You specify the model type by using the FORM= option. The options SCALE= and RANGE= specify the covariance structure sill c_0 and range a_0 parameters, respectively, as discussed in the previous section.

Although it is not included in the original spatial structure, note that a minimal nugget effect is specified with the NUGGET= option to avoid singularity issues. Singularity can appear in the present example as a result of the combined use of the Gaussian covariance model and relatively short distances between nodes, data, or nodes and data in the simulation area.

These steps are implemented using the following DATA step and statements:

```
ods graphics on;

proc sim2d data=thick outsim=sim plot=(observ means);
   coordinates xc=East yc=North;
   simulate var=Thick numreal=5000 seed=79931
      scale=7.2881 range=30.6239 nugget=1e-8 form=gauss;
   mean 40.1173;
   grid x=0 to 100 by 2.5 y=0 to 100 by 2.5;
run;
```

The table in Figure 79.1 shows the number of observations read and used in the conditional simulation. This table can provide you with useful information in case you have missing values in the input data.

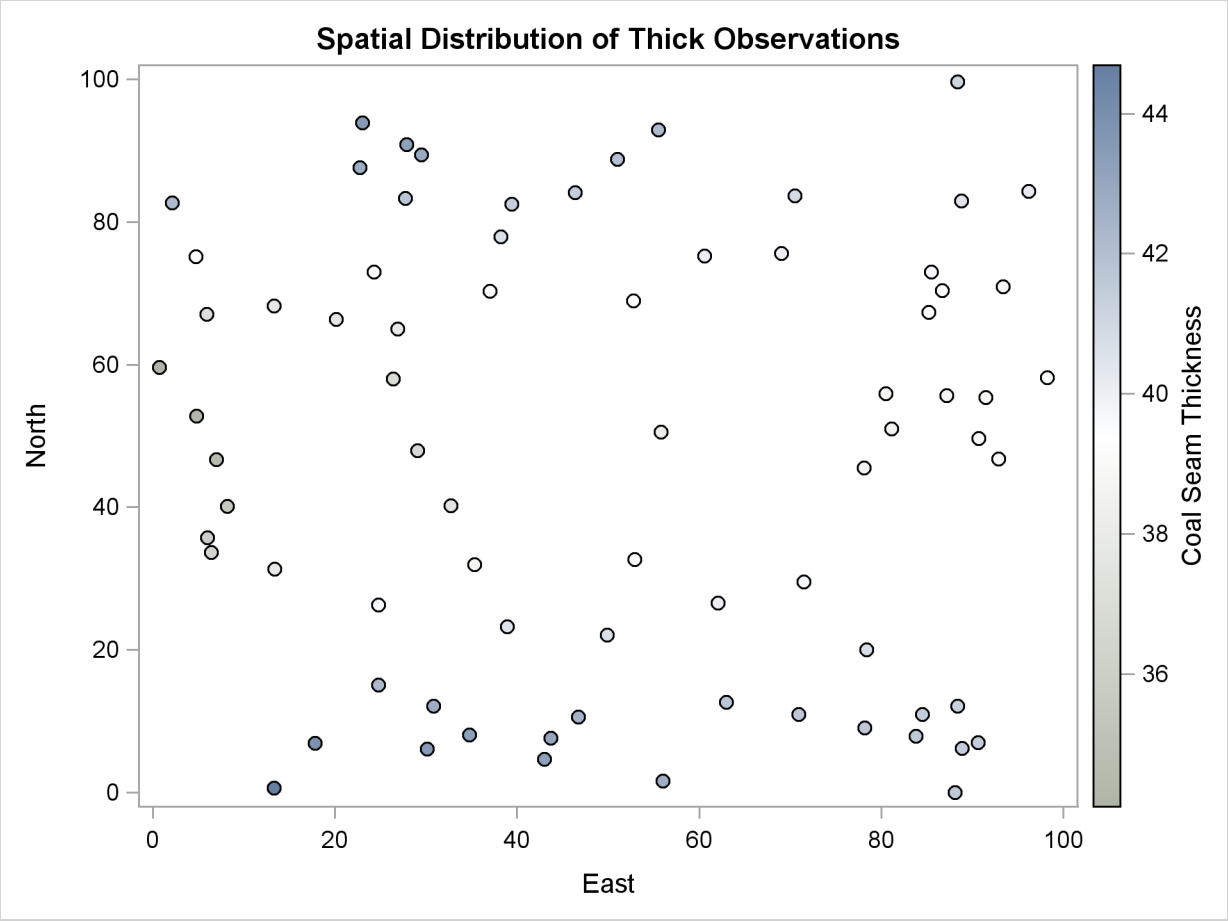
Figure 79.1 Number of Observations for the thick Data Set

```
                    The SIM2D Procedure
          Simulation: Sim1, Dependent Variable: Thick

            Number of Observations Read          75
            Number of Observations Used          75
```

The sample locations are then plotted in Figure 79.2. The figure clearly shows some small-scale variation that is typical of spatial data.

Figure 79.2 Scatter Plot of the Observations Spatial Distribution



PROC SIM2D also produces the table shown in Figure 79.3, which contains information about the type of simulation you run and the number of realizations requested.

Figure 79.3 Simulation Analysis Information

Simulation Information

Type	Conditional
Number of Realizations	5000

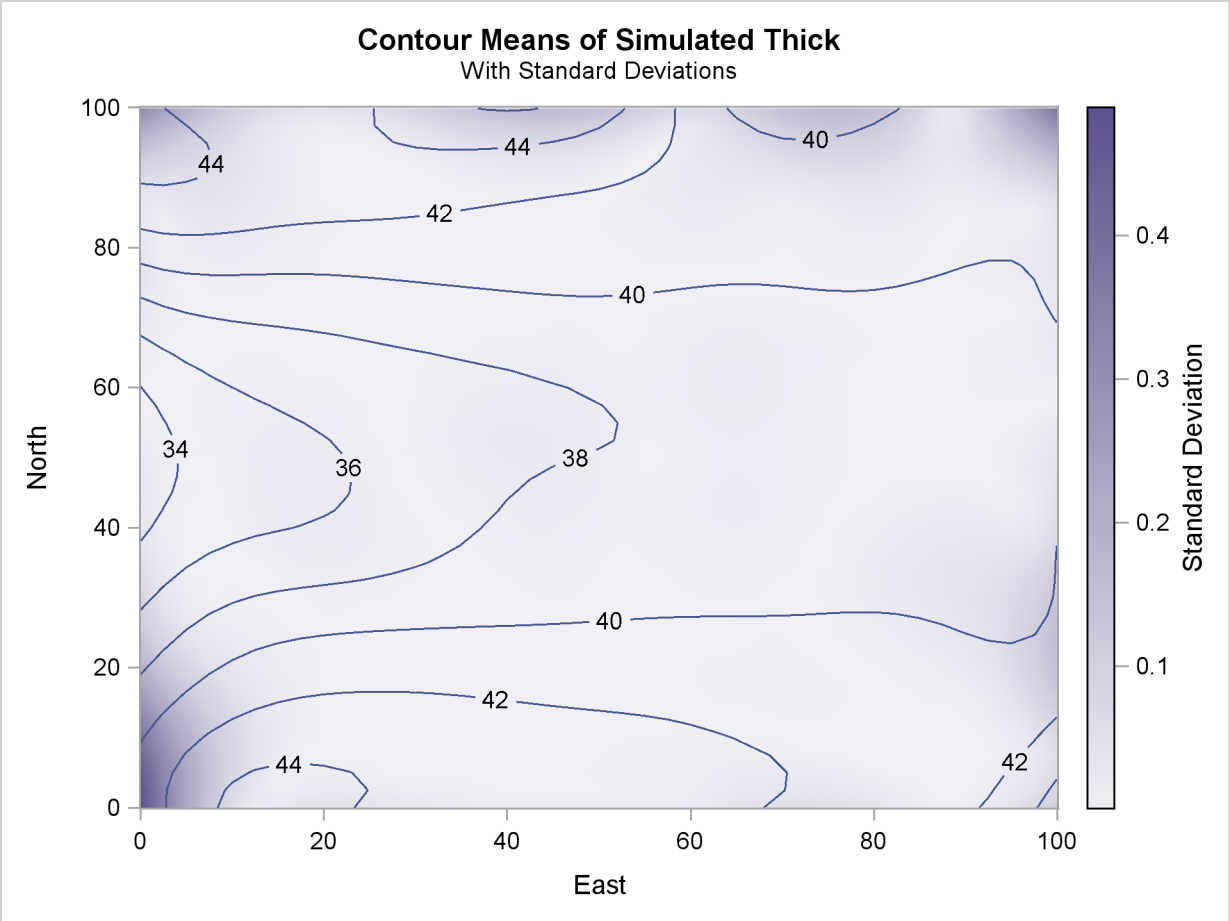
The table in Figure 79.4 displays the spatial correlation model information that will be used by PROC SIM2D for the current simulation. If applicable, the table provides the effective range, as well. This is the distance r_ϵ at which the covariance is 5% of its value at zero. Here you specified the Gaussian model, for which the effective range r_ϵ is $\sqrt{3}a_0$.

Figure 79.4 Simulation Covariance Model Information

Covariance Model Information	
Type	Gaussian
Sill	7.2881
Range	30.6239
Effective Range	53.042151
Nugget Effect	1E-8

Eventually, the SIM2D procedure produces the requested means plot shown in Figure 79.5. The contours of the mean of the simulated values show the average of the simulated realizations at each grid node, which is based on the given spatial structure characteristics. In this case, these means are also conditioned by the Thick observations across the region.

Figure 79.5 Contour Plot of Conditionally Simulated Coal Seam Thickness



Note also that the gradient showing the standard deviation of the simulated values at each grid node is generally small throughout the region. A few exceptions are evident close to the region borders. In these areas the simulated realizations depend on a limited amount of neighboring data. The simulation at these locations relies mainly on the underlying spatial structure.

In addition to the simulation analysis, you can use the PROC SIM2D output to obtain statistical information about the simulated values at selected locations. Let us assume that you would like some basic statistics for the extreme southwest point at (East=0, North=0) and the point (East=75, North=75) toward the northeast corner of the region. You use the following DATA step to select the realizations for these points from the OUTSIM= output data set:

```
data selected;
   set sim(where=((gxc=0 and gyc=0) or (gxc=75 and gyc=75)));
   label gxc = "X-coord";
   label gyc = "Y-coord";
run;
```

Then, you use PROC SORT to sort the selected data set entries and PROC MEANS to produce the simulation statistics for the selected points. The following statements yield the mean, standard deviation, and maximum values of the 5,000 realizations of the Thick values at each one of the selected locations:

```
proc sort data=selected;
   by gxc gyc;
run;

proc means data=selected Mean Std Max;
   class gxc gyc;
   ways 2;
   where (   ((gxc =  0) & (gyc =  0))
           | ((gxc = 75) & (gyc = 75)));
   var SValue;
run;

ods graphics off;
```

The requested statistics for the grid points (East=0, North=0) and (East=75,North=75) are shown in Figure 79.6.

Figure 79.6 Simulation Statistics at Grid Points (East=0, North=0) and (East=75, North=75)

The MEANS Procedure

Analysis Variable : SVALUE Simulated Value at Grid Point

X-coord	Y-coord	N Obs	Mean	Std Dev	Maximum
0	0	5000	40.6064093	0.4900041	42.4131782
75	75	5000	40.1099449	0.0020988	40.1187941

"Example 79.2: Variability at Selected Locations" on page 6179 shows you how to perform a simulation at a set of selected locations rather than a domain-wide grid, and how to obtain more detailed statistics from the simulation.

Syntax: SIM2D Procedure

The following statements are available in PROC SIM2D:

PROC SIM2D *options* ;
  BY *variables* ;
  COORDINATES *coordinate-variables* ;
  GRID *grid-options* ;
  SIMULATE *simulate-options* ;
  MEAN *mean-options* ;

The SIMULATE and MEAN statements are hierarchical; you can specify any number of SIMULATE statements, but you must specify at least one. If you specify a MEAN statement, it refers to the preceding SIMULATE statement. If you do not specify a MEAN statement, a zero-mean model is simulated.

You must specify a single COORDINATES statement to identify the x and y coordinate variables in the input data set when you perform a conditional simulation. You must also specify a single GRID statement to specify the grid information.

The following table outlines the options available in PROC SIM2D classified by function.

Table 79.1 Options Available in the SIM2D Procedure

Task	Statement	Option
Data Set Options		
Specify input data set	PROC SIM2D	DATA=
Specify grid data set	GRID	GDATA=
Write simulated values	PROC SIM2D	OUTSIM=
Specify plot display and options	PROC SIM2D	PLOTS=
Specify quadratic form data set	MEAN	QDATA=
Specify plot display and options	PROC SIM2D	PLOTS=
Declaring the Role of Variables		
Specify variables to define analysis subgroups	BY	
Specify the conditioning variable	SIMULATE	VAR=
Specify the x and y coordinate variables in the data set	COORDINATES	XC= YC=
Specify the x and y coordinate variables in the GDATA= data set	GRID	XC= YC=
Specify the constant coefficient variable in the QDATA= data set	MEAN	CONST=
Specify the linear x coefficient variable in the QDATA= data set	MEAN	CX=
Specify the linear y coefficient variable in the QDATA= data set	MEAN	CY=

Table 79.1 *continued*

Task	Statement	Option
Specify the quadratic x coefficient variable in the QDATA= data set	MEAN	CXX=
Specify the quadratic y coefficient variable in the QDATA= data set	MEAN	CYY=
Specify the quadratic xy coefficient variable in the QDATA= data set	MEAN	CXY=
Controlling the Simulation		
Specify the number of realizations	SIMULATE	NUMREAL=
Specify the seed value for the random generator	SIMULATE	SEED=
Controlling the Mean Quadratic Surface		
Specify the CONST term	MEAN	CONST=
Specify the linear x term	MEAN	CX=
Specify the linear y term	MEAN	CY=
Specify the quadratic x term	MEAN	CXX=
Specify the quadratic y term	MEAN	CYY=
Specify the quadratic cross term	MEAN	CXY=
Controlling the Semivariogram Model		
Specify a nugget effect	SIMULATE	NUGGET=
Specify a functional form	SIMULATE	FORM=
Specify nested functional forms	SIMULATE	FORM=($f_1, \cdots, f_k$)
Specify a range parameter	SIMULATE	RANGE=
Specify nested range parameters	SIMULATE	RANGE=($r_1, \cdots, r_k$)
Specify a scale parameter	SIMULATE	SCALE=
Specify nested scale parameters	SIMULATE	SCALE=($s_1, \cdots, s_k$)
Specify an angle for an anisotropic model	SIMULATE	ANGLE=
Specify nested angles	SIMULATE	ANGLE=($a_1, \cdots, a_k$)
Specify a minor-major axis ratio for an anisotropic model	SIMULATE	RATIO=
Specify nested minor-major axis ratios	SIMULATE	RATIO=($ra_1, \cdots, ra_k$)

PROC SIM2D Statement

PROC SIM2D *options* **;**

You can specify the following options with the PROC SIM2D statement.

DATA=*SAS-data-set*
specifies a SAS data set that contains the x and y coordinate variables and the VAR= variables that are used in the SIMULATE statements. This data set is required if you specify the BY

statement or if any of the SIMULATE statements are conditional—that is, if you specify the VAR= option in any of those. Otherwise, you do not need the DATA= option and this option is ignored if you specify it.

NARROW
restricts the variables included in the OUTSIM= data set. When you specify the NARROW option, only four variables are included. This option is useful when a large number of simulations are produced. Including only four variables reduces the memory required for the OUTSIM= data set. For details on the variables that are excluded with the NARROW option, see the section "Output Data Set" on page 6172.

NOPRINT
suppresses the normal display of results. The NOPRINT option is useful when you want only to create one or more output data sets with the procedure. Note that this option temporarily disables the Output Delivery System (ODS); see the section "ODS Graphics" on page 6173 for more information.

OUTSIM=*SAS-data-set*
specifies a SAS data set to store the simulation values, iteration number, simulate statement label, variable name, and grid location. For details, see the section "Output Data Set" on page 6172.

PLOTS <(*global-plot-option***)> <= *plot-request* <(***options***)>>**

PLOTS <(*global-plot-option***)> <= (***plot-request* <(***options***)>** **<... *plot-request* <(***options***)>>)>**
controls the plots produced through ODS Graphics. When you specify only one plot request, you can omit the parentheses around the plot request. Here are some examples:

```
plots=none
plots=observ
plots=(observ(outl) means)
plots=(means(fill=mean line=sd obs=grad) means(fill=sd))
```

You must enable ODS Graphics before requesting plots, for example, like this:

```
ods graphics on;

proc sim2d data=thick outsim=sim;
   coordinates xc=East yc=North;
   simulate var=Thick numreal=5000 seed=79931
      scale=7.2881 range=30.6239 form=gauss;
   mean 40.1173;
   grid x=0 to 100 by 2.5 y=0 to 100 by 2.5;
run;

ods graphics off;
```

For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." PROC SIM2D does not have default plots. This means that unless you have enabled ODS Graphics, specified the PLOTS= option, or specified PLOTS=ALL, then PROC SIM2D will produce no plots.

The following *global-plot-option* is available:

ONLY
produces only plots specifically requested.

The following individual *plot-requests* and plot *options* are available:

ALL
produces all appropriate plots. You can specify other options with ALL. For example, to request all appropriate plots and an additional plot of the simulation means, specify PLOTS=(ALL MEANS).

EQUATE
specifies that all appropriate plots be produced in a way that the axes coordinates have equal size units.

NONE
supresses all plots.

OBSERVATIONS < (*observations-plot-options*) >

OBSERV < (*observations-plot-options*) >

OBS < (*observations-plot-options*) >
produces the observed data plot. Only one observations plot will be created if you specify the OBSERVATIONS option more than once within a PLOTS option.

The OBSERVATIONS option has the following suboptions:

GRADIENT
specifies that observations be displayed as circles colored by the observed measurement.

OUTLINE
specifies that observations be displayed as circles with a border but with a completely transparent fill.

OUTLINEGRADIENT
is the same as OBSERVATIONS(GRADIENT) except that a border is shown around each observation.

SHOWMISSING
specifies that observations with missing values be displayed in addition to the observations with nonmissing values. By default, missing values locations are not shown on the plot. If you specify multiple instances of the OBSERVATIONS option, and you specify the SHOWMISSING suboption in any of those, then the resulting observations plot will display the observations with missing values.

For the GRADIENT, OUTLINE, and OUTLINEGRADIENT suboptions: The OUTLINEGRADIENT is the default suboption if you do not specify any of those three. If you specify multiple instances of the OBSERVATIONS option or multiple suboptions for OBSERVATIONS, then the resulting observations plot will honor the last specified GRADIENT, OUTLINE, or OUTLINEGRADIENT suboption.

MEANS < (*means-plot-options*) >

MEA < (*means-plot-options*) >

specifies that a plot of the means and standard deviations of the simulation realizations at each grid point be produced. You can specify the MEANS option multiple times in the same PLOTS option to request instances of plots with the following *means-plot-options*:

FILL=NONE | MEAN | SD

produces a surface plot for either the values of the means or the standard deviations. FILL=SD is the default; however, if you do not specify the FILL suboption the behavior depends on the LINE suboption as follows: If you specify LINE=NONE or do not specify the LINE suboption at all, then the FILL suboption will be set to its default value. If LINE=PRED or LINE=SE, then the FILL suboption will be set to the same value as the LINE suboption.

LINE=NONE | MEAN | SD

produces a contour line plot for either the values of the means or the standard deviations. LINE=MEAN is the default; however, if you do not specify the LINE suboption the behavior depends on the FILL suboption as follows: If you specify FILL=NONE or do not specify the FILL suboption at all, then the LINE suboption will be set to its default value. If FILL=PRED or FILL=SE, then the LINE suboption will be set to the same value as the FILL suboption.

OBS=*obs-options*

produces an overlaid scatter plot of the observations in addition to the specified contour plots. The following *obs-options* are available:

GRAD

specifies that observations be displayed as circles colored by the observed measurement. The same color gradient is used to display the means surface and the observations. The conditional simulation honors the observed values, so the means surface at the observation locations has the same color as the corresponding observations.

LINEGRAD

is the same as OBS=GRAD except that a border is shown around each observation.

NONE

specifies that no observations be displayed.

OUTL

specifies that observations be displayed as circles with a border but with a completely transparent fill.

OBS=NONE is the default.

If you specify multiple instances of the FILL, LINE, or OBS suboptions in the same MEANS option, then the resulting means plot will honor the last value specified for any of the suboptions.

Any combination where you specify FILL=NONE and LINE=NONE is not available.

The MEANS option requires that simulations be requested at more than one location on a rectangular grid. The plot is not available when you use the GRIDDATA= data set option in the GRID statement, because the GRIDDATA= option allows you to specify any coordinates for the simulation locations.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC SIM2D to obtain separate analyses on observations in groups defined by the BY variables. Hence, in PROC SIM2D it makes sense to use the BY statement when you request conditional simulations where observations are involved. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the BY statement options NOTSORTED or DESCENDING in the BY statement for the SIM2D procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

The use of the BY statement assumes that you have specified an input data set with the DATA= option in the PROC SIM2D statement. Note that in PROC SIM2D if you do not specify the VAR= option in the SIMULATE statement, then this is a request for unconditional simulation even if you have specified the DATA= option in the PROC SIM2D statement. Therefore, it is possible to specify the BY statement and request mixed types of simulations by specifying multiple SIMULATE statements in the same PROC SIM2D step.

For more information about the BY statement, see *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the *Base SAS Procedures Guide*.

COORDINATES Statement

COORDINATES *coordinate-variables* **;**

The following two options give the name of the variables in the DATA= data set containing the values of the x and y coordinates of the conditioning data.

Only one COORDINATES statement is allowed, and it is applied to all SIMULATE statements that have the VAR= specification. In other words, it is assumed that all the VAR= variables in all SIMULATE statements have the same x and y coordinates.

You can abbreviate the COORDINATES statement as COORD.

XCOORD=*(variable-name)*

XC=*(variable-name)*

gives the name of the variable containing the x coordinate of the data in the DATA= data set.

YCOORD=*(variable-name)*

YC=*(variable-name)*

gives the name of the variable containing the y coordinate of the data locations in the DATA= data set.

GRID Statement

GRID *grid-options* **;**

The following options can be used to specify the grid of spatial locations at which to perform the simulations. A single GRID statement is required and is applied to all SIMULATE statements.

There are two basic methods for specifying the grid. You can specify the x and y coordinates explicitly, or they can be read from a SAS data set.

The options for the explicit specification of grid locations are as follows.

X=*number*

X=$x_1, \ldots, x_m$

X=x_1 *to* x_m

X=x_1 *to* x_m *by* δx

specifies the x coordinate of the grid locations.

Y=*number*

Y=$y_1, \ldots, y_m$

Y=y_1 *to* y_m

Y=y_1 *to* y_m *by* δy

specifies the y coordinate of the grid locations.

For example, the following two GRID statements are equivalent:

```
grid x=1,2,3,4,5  y=0,2,4,6,8,10;

grid x=1 to 5 y=0 to 10 by 2;
```

To specify grid locations from a SAS data set, you must provide the name of the data set and the variables containing the values of the x and y coordinates.

GRIDDATA=*SAS-data-set*

GDATA=*SAS-data-set*

specifies a SAS data set containing the x and y grid coordinates.

XCOORD=*(variable-name)*

XC=*(variable-name)*

gives the name of the variable containing the x coordinate of the grid locations in the GRIDDATA= data set.

YCOORD=*(variable-name)*

YC=*(variable-name)*

gives the name of the variable containing the y coordinate of the grid locations in the GRIDDATA= data set.

SIMULATE Statement

SIMULATE *simulate-options* **;**

The SIMULATE statement specifies details on the simulation and the covariance model used in the simulation. You can specify the following options with a SIMULATE statement, which can be abbreviated by SIM.

NUMREAL=*number*

NUMR=*number*

NR=*number*

specifies the number of realizations to produce for the spatial process specified by the covariance model. Note that the number of observations in the OUTSIM= data set contributed by a given SIMULATE statement is the product of the NUMREAL= value with the number of grid points. This can cause the OUTSIM= data set to become large even for moderate values of the NUMREAL= option.

VAR=*(variable-name)*

specifies the single numeric variable used as the conditioning variable in the simulation. In other words, the simulation is conditional on the values of the VAR= variable found in the DATA= data set. If you omit the VAR= option or if all observations of the VAR= variable are missing values, then the simulation is *unconditional*. Since multiple SIMULATE statements are allowed, you can perform both unconditional and conditional simulations with a single PROC SIM2D statement.

Covariance Model Specification

There are two ways to specify a semivariogram or covariance model. In the first method, you can specify the required parameters SCALE, RANGE, and FORM, and possibly the optional parameters NUGGET, ANGLE, and RATIO, explicitly in the SIMULATE statement.

In the second method, you can specify an MDATA= data set. This data set contains variables corresponding to the required SCALE, RANGE, and FORM parameters, and, optionally, variables for the NUGGET, ANGLE, and RATIO parameters.

The two methods are exclusive; either you specify all parameters explicitly, or they are all are read from the MDATA= data set.

ANGLE=*angle*

ANGLE=*(angle*$_1$*,...,angle*$_k$*)*

specifies the angle of the major axis for anisotropic models, measured in degrees clockwise from the N–S axis. In the case of a nested semivariogram model, you can specify an angle for each nesting. The default is ANGLE=0.

FORM=*form_spec*

FORM=*(form_spec*$_1$*, form_spec*$_2$*,...,form_spec*$_k$*)*

specifies the functional form or forms of the semivariogram model, where *form_spec* can take only the values SPHERICAL, EXPONENTIAL, and GAUSSIAN. The two ways of specifying the FORM= parameter allow specification of both nested and non-nested models.

The following abbreviations are permitted. For the spherical model, you can specify the *form_spec* as FORM=SPHERICAL, FORM=SPH, or FORM=S. For the exponential model, you can specify the *form_spec* as FORM=EXPONENTIAL, FORM=EXP, or FORM=E. For the Gaussian model, you can specify the *form_spec* as FORM=GAUSSIAN, FORM=GAUSS, or FORM=G.

MDATA=*SAS-data-set*

specifies the input data set that contains parameter values for the covariance or semivariogram model. The MDATA= data set must contain variables named SCALE, RANGE, and FORM, and it can optionally contain the variables NUGGET, ANGLE, and RATIO.

The FORM variables must be character, and they can assume the same values allowed in the explicit FORM= syntax described previously. The RANGE and SCALE variables must be numeric. The optional variables ANGLE, RATIO, and NUGGET must also be numeric if present.

The number of observations present in the MDATA= data set corresponds to the level of nesting of the covariance or semivariogram model. For example, to specify a non-nested model that uses a spherical covariance, an MDATA= data set might contain the following statements:

```
data md1;
   input scale range form $;
   datalines;
   25 10 SPH
run;
```

The PROC SIM2D statement to use the MDATA= specification is of the form shown in the following:

```
proc sim2d data=...;
   sim var=.... mdata=md1;
run;
```

This is equivalent to the following explicit specification of the covariance model parameters:

```
proc sim2d data=...;
   sim var=.... scale=25 range=10 form=sph;
run;
```

The following MDATA= data set is an example of an anisotropic nested model:

```
data md2;
   input scale range form $ nugget angle ratio;
   datalines;
   20 8 S  5 35 .7
   12 3 G  5 0  .8
   4  1 G  5 45 .5
   ;

proc sim2d data=...;
   sim var=.... mdata=md2;
run;
```

This is equivalent to the following explicit specification of the covariance model parameters:

```
proc sim2d data=...;
   sim var=.... scale=(20,12,4) range=(8,3,1) form=(S,G,G)
                angle=(35,0,45) ratio=(.7,.8,.5) nugget=5;
run;
```

This example is somewhat artificial in that it is usually hard to detect different anisotropy directions and ratios for different nestings by using an experimental semivariogram. Note that the NUGGET value is the same for all nestings. This is always the case; the nugget effect is a single additive term for all models. For further details, refer to the section "The Nugget Effect" on page 2952 in the KRIGE2D procedure.

The SIMULATE statement can be given a label. This is useful for identification in the OUTSIM= data set when multiple SIMULATE statements are specified. For example:

```
proc sim2d data=...;
   gauss1: sim var=.... form=gauss;
   mean ....;
   gauss2: sim var=.... form gauss;
   mean ....;
   exp1: sim var=.... form=exp;
   mean ....;
   exp2: sim var=.... form=exp;
   mean ....;
run;
```

In the OUTSIM= data set, the values "GAUSS1," "GAUSS2," "EXP1," and "EXP2" for the LABEL variable help to identify the realizations corresponding to the four SIMULATE statements. If you do not provide a label for a SIMULATE statement, a default label of SIM*n* is given, where *n* is the number of unlabeled SIMULATE statements seen so far.

NUGGET=*number*

specifies the nugget effect for the model. This effect is due to a discontinuity in the semivariogram as determined by plotting the sample semivariogram (refer to the section "The Nugget Effect" on page 2952 in the KRIGE2D procedure for details). For models without any nugget effect, the NUGGET= option is left out. The default is NUGGET=0.

RANGE=*range*

RANGE=*(range*$_1$*,. . . ,range*$_k$*)*

specifies the range parameter in the semivariogram models. In the case of a nested semivariogram model, you must specify a range for each nesting.

The range parameter is the divisor in the exponent in all supported models. It has the units of distance or distance squared for these models, and it is related to the correlation scale for the underlying spatial process.

Refer to the section "Theoretical Semivariogram Models" on page 2945 in the KRIGE2D procedure for details on how the RANGE= values are determined.

RATIO=*ratio*

RATIO=*(ratio*$_1$*,. . . ,ratio*$_k$*)*

specifies the ratio of the length of the minor axis to the length of the major axis for anisotropic models. The value of the RATIO= option must be between 0 and 1. In the case of a nested semivariogram model, you can specify a ratio for each nesting. The default is RATIO=1.

SCALE=*scale*

SCALE=*(scale*$_1$*,. . . ,scale*$_k$*)*

specifies the scale (or *sill*) parameter in semivariogram models. In the case of a nested semivariogram model, you must specify a scale for each nesting. The scale parameter is the multiplicative factor in all supported models; it has the same units as the variance of the VAR= variable.

Refer to the section "Theoretical Semivariogram Models" on page 2945 in the KRIGE2D procedure for details on how the SCALE= values are determined.

SEED=*seed value*

specifies the seed to use for the random number generator. The SEED= value has to be an integer.

SINGULAR=*number*

gives the singularity criteria for solving the set of linear equations involved in the computation of the mean and covariance of the conditional distribution associated with a given SIMULATE statement. The larger the value of the SINGULAR= option, the easier it is for the covariance matrix system to be declared singular. The default is SINGULAR=1E–8.

For more details on the use of the SINGULAR= option, see the section "Computational and Theoretical Details of Spatial Simulation" on page 6168.

MEAN Statement

MEAN *spec*$_1$,. . . ,*spec*$_6$ **;**

MEAN QDATA= *SAS-data-set CONST=var*$_1$ *CX=var*$_2$ *CY=var*$_3$
CXX=var$_4$ *CYY=var*$_5$ *CXY=var*$_6$ **;**

MEAN QDATA= *SAS-data-set* **;**

A mean function $\mu(s)$ that is a quadratic in the coordinates can be written

$$\mu(s) = \mu(x, y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy$$

The MEAN statement is used to specify the quadratic surface to use as the mean function for the simulated SRF. There are three ways to specify the MEAN statement. The MEAN statement allows the specification of the coefficients $\beta_0, \cdots, \beta_5$ either explicitly or through a QDATA= data set.

An example of an explicit specification is the following:

```
mean 1.4 + 2.5*x + 3.6*y + 0.47*x*x + 0.58*y*y + 0.69*x*y;
```

In this example, all terms have a nonzero coefficient. Any term with a zero coefficient is simply left out of the specification. For example,

```
mean 1.4;
```

is a valid quadratic form with all terms having zero coefficients except the constant term.

An equivalent way of specifying the mean function is through the QDATA= data set. For example, the MEAN statement

```
mean 1.4 + 2.5*x + 3.6*y + 0.47*x*x + 0.58*y*y + 0.69*x*y;
```

can be alternatively specified by the following DATA step and MEAN statement:

```
data q1;
   input c1 c2 c3 c4 c5 c6;
   datalines;
   1.4 2.5 3.6 0.47 0.58 0.69
run;

proc sim2d data=....;
   simulate ...;
   mean qdata=q1 const=c1 cx=c2 cy=c3 cxx=c4
            cyy=c5 cxy=c6;
run;
```

The QDATA= data set specifies the data set containing the coefficients. The parameters CONST=, CX=, CY=, CXX=, CYY=, and CYX= specify the variables in the QDATA= data set that correspond to the constant, linear x, linear y, and so on. For any coefficient not specified in this list, the

QDATA= data set is checked for the presence of variables with default names of CONST, CX, CY, CXX, CYY, and CXY. If these variables are present, their values are taken as the corresponding coefficients. Hence, you can rewrite the previous example as follows:

```
data q1;
   input const cx cy cxx cyy cxy;
   datalines;
   1.4 2.5 3.6 0.47 0.58 0.69
   ;

proc sim2d data=....;
   simulate ...;
   mean qdata=q1;
run;
```

If a given coefficient does not appear in the list or in the data set with the default name, a value of zero is assumed.

Details: SIM2D Procedure

Computational and Theoretical Details of Spatial Simulation

Introduction

There are a number of approaches to simulating spatial random fields or, more generally, simulating sets of dependent random variables. This includes sequential indicator methods, turning bands, and the Karhunen-Loeve Expansion. Refer to Christakos (1992, Chapter 8) and Deutsch and Journel (1992, Chapter V) for details.

A particularly simple method available for Gaussian spatial random fields is the LU decomposition method. This method is computationally efficient. For a given covariance matrix, the $LU = \mathbf{L}\mathbf{L}^T$ decomposition is computed once, and the simulation proceeds by repeatedly generating a vector of independent $N(0, 1)$ random variables and multiplying by the $\mathbf{L}$ matrix.

One problem with this technique is memory requirements; memory is required to hold the full data and grid covariance matrix in core. While this is especially limiting in the three-dimensional case, you can use PROC SIM2D, which handles only two-dimensional data, for moderately sized simulation problems.

Theoretical Development

It is a simple matter to produce an $N(0, 1)$ random number, and by stacking k $N(0, 1)$ random numbers in a column vector, you can obtain a vector with independent standard normal components

$\mathbf{W} \sim N_k(\mathbf{0}, \mathbf{I})$. The meaning of the terms *independence* and *randomness* in the context of a deterministic algorithm required for the generation of these numbers is subtle; refer to Knuth (1981, Chapter 3) for details.

Rather than $\mathbf{W} \sim N_k(\mathbf{0}, \mathbf{I})$, what is required is the generation of a vector $\mathbf{Z} \sim N_k(\mathbf{0}, \mathbf{C})$—that is,

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_k \end{bmatrix}$$

with covariance matrix

$$\mathbf{C} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1k} \\ C_{21} & C_{22} & \cdots & C_{2k} \\ & \ddots & & \\ C_{k1} & C_{k2} & \cdots & C_{kk} \end{pmatrix}$$

If the covariance matrix is symmetric and positive definite, it has a Cholesky root $\mathbf{L}$ such that $\mathbf{C}$ can be factored as

$$\mathbf{C} = \mathbf{L}\mathbf{L}^T$$

where $\mathbf{L}$ is lower triangular. Refer to Ralston and Rabinowitz (1978, Chapter 9, Section 3-3) for details. This vector $\mathbf{Z}$ can be generated by the transformation $\mathbf{Z} = \mathbf{L}\mathbf{W}$. Note that this is where the assumption of a Gaussian SRF is crucial. When $\mathbf{W} \sim N_k(\mathbf{0}, \mathbf{I})$, then $\mathbf{Z} = \mathbf{L}\mathbf{W}$ is also Gaussian. The mean of $\mathbf{Z}$ is

$$E(\mathbf{Z}) = \mathbf{L}(E(\mathbf{W})) = \mathbf{0}$$

and the variance is

$$\text{Var}(\mathbf{Z}) = \text{Var}(\mathbf{L}\mathbf{W}) = E(\mathbf{L}\mathbf{W}\mathbf{W}^T\mathbf{L}^T) = \mathbf{L}E(\mathbf{W}\mathbf{W}^T)\mathbf{L}^T = \mathbf{L}\mathbf{L}^T = \mathbf{C}$$

Consider now an SRF $Z(\boldsymbol{s}), \boldsymbol{s} \in D \subset \mathcal{R}^2$, with spatial covariance function $C(\boldsymbol{h})$. Fix locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_k$, and let $\mathbf{Z}$ denote the random vector

$$\mathbf{Z} = \begin{bmatrix} Z(\boldsymbol{s}_1) \\ Z(\boldsymbol{s}_2) \\ \vdots \\ Z(\boldsymbol{s}_k) \end{bmatrix}$$

with corresponding covariance matrix

$$\mathbf{C}_z = \begin{pmatrix} C(\mathbf{0}) & C(\boldsymbol{s}_1 - \boldsymbol{s}_2) & \cdots & C(\boldsymbol{s}_1 - \boldsymbol{s}_k) \\ C(\boldsymbol{s}_2 - \boldsymbol{s}_1) & C(\mathbf{0}) & \cdots & C(\boldsymbol{s}_2 - \boldsymbol{s}_k) \\ & \ddots & & \\ C(\boldsymbol{s}_k - \boldsymbol{s}_1) & C(\boldsymbol{s}_k - \boldsymbol{s}_2) & \cdots & C(\mathbf{0}) \end{pmatrix}$$

Since this covariance matrix is symmetric and positive definite, it has a Cholesky root, and the $Z(\boldsymbol{s}_i), i = 1, \cdots, k$, can be simulated as described previously. This is how the SIM2D procedure implements unconditional simulation in the zero-mean case. More generally,

$$Z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$$

with $\mu(\boldsymbol{s})$ being a quadratic form in the coordinates $\boldsymbol{s} = (x, y)$, and the $\varepsilon(\boldsymbol{s})$ being an SRF having the same covariance matrix $\mathbf{C}_z$ as previously. In this case, the $\mu(\boldsymbol{s}_i), i = 1, \cdots, k$, is computed once and added to the simulated vector $\varepsilon(\boldsymbol{s}_i), i = 1, \cdots, k$, for each realization.

For a conditional simulation, this distribution of

$$\mathbf{Z} = \begin{bmatrix} Z(\boldsymbol{s}_1) \\ Z(\boldsymbol{s}_2) \\ \vdots \\ Z(\boldsymbol{s}_k) \end{bmatrix}$$

must be conditioned on the observed data. The relevant general result concerning conditional distributions of multivariate normal random variables is the following. Let $\mathbf{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

The subvector $\mathbf{X}_1$ is $k \times 1$, $\mathbf{X}_2$ is $n \times 1$, $\boldsymbol{\Sigma}_{11}$ is $k \times k$, $\boldsymbol{\Sigma}_{22}$ is $n \times n$, and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^T$ is $k \times n$, with $k + n = m$. The full vector $\mathbf{X}$ is partitioned into two subvectors, $\mathbf{X}_1$ and $\mathbf{X}_2$, and $\boldsymbol{\Sigma}$ is similarly partitioned into covariances and cross covariances.

With this notation, the distribution of $\mathbf{X}_1$ conditioned on $\mathbf{X}_2 = \boldsymbol{x}_2$ is $N_k(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, with

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$$

and

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Refer to Searle (1971, pp. 46–47) for details. The correspondence with the conditional spatial simulation problem is as follows. Let the coordinates of the observed data points be denoted $\tilde{\boldsymbol{s}}_1, \tilde{\boldsymbol{s}}_2, \cdots, \tilde{\boldsymbol{s}}_n$, with values $\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_n$. Let $\tilde{\mathbf{Z}}$ denote the random vector

$$\tilde{\mathbf{Z}} = \begin{bmatrix} Z(\tilde{\boldsymbol{s}}_1) \\ Z(\tilde{\boldsymbol{s}}_2) \\ \vdots \\ Z(\tilde{\boldsymbol{s}}_n) \end{bmatrix}$$

The random vector $\tilde{\mathbf{Z}}$ corresponds to $\mathbf{X}_2$, while $\mathbf{Z}$ corresponds to $\mathbf{X}_1$. Then $\left(\mathbf{Z} \mid \tilde{\mathbf{Z}} = \tilde{\mathbf{z}}\right) \sim N_k(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{C}})$ as in the previous distribution. The matrix

$$\tilde{\mathbf{C}} = \mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}$$

is again positive definite, so a Cholesky factorization can be performed.

The dimension n for $\tilde{\mathbf{Z}}$ is simply the number of nonmissing observations for the VAR= variable; the values $\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_n$ are the values of this variable. The coordinates $\tilde{\boldsymbol{s}}_1, \tilde{\boldsymbol{s}}_2, \cdots, \tilde{\boldsymbol{s}}_n$ are also found in the DATA= data set, with the variables corresponding to the x and y coordinates identified in the COORDINATES statement. Note that all VAR= variables use the same set of conditioning coordinates; this fixes the matrix $\mathbf{C}_{22}$ for all simulations.

The dimension k for $\mathbf{Z}$ is the number of grid points specified in the GRID statement. Since there is a single GRID statement, this fixes the matrix $\mathbf{C}_{11}$ for all simulations. Similarly, $\mathbf{C}_{12}$ is fixed.

The Cholesky factorization $\tilde{\mathbf{C}} = \mathbf{L}\mathbf{L}^T$ is computed once, as is the mean correction

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \mathbf{C}_{12}\mathbf{C}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$$

Note that the means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are computed using the grid coordinates $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_k$, the data coordinates $\tilde{\boldsymbol{s}}_1, \tilde{\boldsymbol{s}}_2, \cdots, \tilde{\boldsymbol{s}}_n$, and the quadratic form specification from the MEAN statement. The simulation is now performed exactly as in the unconditional case. A $k \times 1$ vector of independent standard $N(0, 1)$ random variables is generated and multiplied by $\mathbf{L}$, and $\tilde{\boldsymbol{\mu}}$ is added to the transformed vector. This is repeated N times, where N is the value specified for the NR= option.

Computational Details

In the computation of $\tilde{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}$ described in the previous section, the inverse $\boldsymbol{\Sigma}_{22}^{-1}$ is never actually computed; an equation of the form

$$\boldsymbol{\Sigma}_{22}\mathbf{A} = \mathbf{B}$$

is solved for $\mathbf{A}$ by using a modified Gaussian elimination algorithm that takes advantage of the fact that $\boldsymbol{\Sigma}_{22}$ is symmetric with constant diagonal $C_z(0)$ that is larger than all off-diagonal elements. The SINGULAR= option pertains to this algorithm. The value specified for the SINGULAR= option is scaled by $C_z(0)$ before comparison with the pivot element.

Memory Usage

For conditional simulations, the largest matrix held in core at any one time depends on the number of grid points and data points. Using the previous notation, the data-data covariance matrix $\mathbf{C}_{22}$ is $n \times n$, where n is the number of nonmissing observations for the VAR= variable in the DATA= data set. The grid-data cross covariance $\mathbf{C}_{12}$ is $n \times k$, where k is the number of grid points. The grid-grid covariance $\mathbf{C}_{11}$ is $k \times k$. The maximum memory required at any one time for storing these matrices is

$$\max\left(k(k+1), n(n+1) + 2(n \times k)\right) \times \text{sizeof(double)}$$

There are additional memory requirements that add to the total memory usage, but usually these matrix calculations dominate, especially when the number of grid points is large.

Output Data Set

The SIM2D procedure produces a single output data set: the OUTSIM=*SAS-data-set*. The OUTSIM= data set contains all the needed information to uniquely identify the simulated values.

The OUTSIM= data set contains the following variables:

- LABEL, which is the label for the current SIMULATE statement
- VARNAME, which is the name of the conditioning variable for the current SIMULATE statement
- _ITER_, which is the iteration number within the current SIMULATE statement
- GXC, which is the x coordinate for the current grid point
- GYC, which is the y coordinate for the current grid point
- SVALUE, which is the value of the simulated variable

If you specify the NARROW option in the PROC SIM2D statement, the LABEL and VARNAME variables are not included in the OUTSIM= data set. This option is useful in the case where the number of data points, grid points, and realizations are such that they generate a very large OUTSIM= data set. The size of the OUTSIM= data set is reduced when these variables are not included.

In the case of an unconditional simulation, the VARNAME variable is not included. In the case of mixed conditional and unconditional simulations (that is, when multiple SIMULATE statements are specified, and one or more contain a VAR= specification and one or more do *not* contain a VAR= specification), the VARNAME variable is included but is given a missing value for those observations corresponding to an unconditional simulation.

Displayed Output

In addition to the output data set, the SIM2D procedure produces output objects as well. The SIM2D procedure output objects are the following:

- a default "Number of Observations" table that displays the number of observations read from the input data set and the number of observations used in the analysis
- a map showing the spatial distribution of the observations of the current VAR variable in the SIMULATE statement, in the case of conditional simulations. The observations are displayed by default with circled markers whose color indicates the VAR value at the corresponding location.
- a default table for each SIMULATE statement that summarizes the simulation specifications
- a default table for each SIMULATE statement that shows the covariance model parameters for the corresponding simulation
- plots of the mean and standard deviation of the simulation realizations at each point of the specified output grid. You can produce more than one of these plots for every SIMULATE statement with styles that you can specify by using the available suboptions of the PLOTS=MEANS option.

ODS Table Names

Each table created by PROC SIM2D has a name associated with it, and you must use this name to reference the table when using ODS Graphics. These names are listed in Table 79.2.

Table 79.2 ODS Tables Produced by PROC SIM2D

ODS Table Name	Description	Statement	Option
ModelInfo	Parameters of the covariance model used in current simulation	PROC	default output
NObs	Number of observations read and used	PROC	default output
SimuInfo	General information about the simulation	PROC	default output

ODS Graphics

This section describes the use of the Output Delivery System (ODS) for creating graphics with the SIM2D procedure.

To request these graphs, you must specify the ODS GRAPHICS statement. For additional control of the graphics that are displayed, see the PLOTS= option in the section "PROC SIM2D Statement" on page 6157. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

ODS Graph Names

PROC SIM2D assigns a name to each graph it creates by using ODS Graphics. You can use these names to reference the graphs when using ODS Graphics. The names are listed in Table 79.3.

Table 79.3 ODS Graphics Produced by PROC SIM2D

ODS Graph Name	Plot Description	Statement	Option
MeansPlot	Contour plot of the simulated means, surface of the standard deviation, and outlines of the observation locations	PROC	PLOTS=MEANS
ObservationsPlot	Scatter plot of observed data and colored markers indicating observed values	PROC	PLOTS=OBSERV

To request this graph, you must specify the ODS GRAPHICS statement in addition to the statements indicated in Table 79.3. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

Examples: SIM2D Procedure

Example 79.1: Simulation

Continuing with the coal seam thickness example from the section "Getting Started: SIM2D Procedure" on page 6149, this example asks a more complicated question. This question is economic in nature, and the (approximate) answer requires the use of simulation.

Simulating a Subregion for Economic Feasibility

The coal seam must be of a minimum thickness, called a *cutoff value*, for a mining operation to be profitable. Suppose that, for a subregion of the measured area, the cost of mining is higher than in the remaining areas due to the geology of the overburden. This higher cost results in a higher thickness cutoff value for the subregion. Suppose also that it is determined from a detailed cost analysis that at least 60% of the subregion must exceed a seam thickness of 39.7 feet for profitability.

How can you use the SRF model (μ and $C_z(\boldsymbol{s})$) and the measured seam thickness values $Z(\boldsymbol{s}_i), i = 1, \cdots, 75$, to determine, in some approximate way, if at least 60% of the subregion exceeds this minimum?

Spatial prediction does not appear to be helpful in answering this question. While it is easy to determine whether a predicted value at a location in the subregion is above the 39.7-feet cutoff value, it is not clear how to incorporate the standard error associated with the predicted value. The standard error is what characterizes the stochastic nature of the prediction (and the underlying SRF). It is clear that it must be included in any realistic approach to the problem.

A conditional simulation, on the other hand, seems to be a natural way of obtaining an approximate answer. By simulating the SRF on a sufficiently fine grid in the subregion, you can determine the proportion of grid points in which the mean value over realizations exceeds the 39.7-feet cutoff and compare it with the 60% value needed for profitability.

It is desirable in any simulation study that the quantity being estimated (in this case, the proportion exceeding the 39.7-feet cutoff) not depend on the number of simulations performed. For example, suppose that the maximum seam thickness is simulated. It is likely that the maximum value increases as the number of simulations performed increases. Hence, a simulation is not useful for such an estimate. A simulation is useful for determining the *distribution* of the maximum, but there are general theoretical results for such distributions, making such a simulation unnecessary. Refer to Leadbetter, Lindgren, and Rootzen (1983) for details.

In the case of simulating the proportion exceeding the 39.7-feet cutoff, it is expected that this quantity will settle down to a fixed value as the number of realizations increases. At a fixed grid point, the quantity being compared with the cutoff value is the mean over all simulated realizations; this mean value settles down to a fixed number as the number of realizations increases. In the same manner, the proportion of the grid where the mean values exceed the cutoff also becomes constant. This can be tested using PROC SIM2D.

A crucial, nonprovable assumption in applying SRF theory to the coal seam thickness data is that the values $Z(\boldsymbol{s}_i), i = 1, \cdots, 75$, represent a *single* realization from the set of all possible realizations consistent with the SRF model (μ and $C_z(\boldsymbol{h})$). A conditional simulation repeatedly produces other possible simulated realizations consistent with the model and data. However, the only concern of the mining company is this single unique realization. It is not concerned about similar coal fields to be mined sometime in the future; it might never see another coal field remotely similar to this one, or it might not be in business in the future.

Hence the proportion found by generating repeated simulated realizations must somehow relate back to the unique realization that is the coal field (seam thickness). This is done by interpreting the proportion found from a simulation to the spatial mean proportion for the unique realization. The term "spatial mean" is simply an appropriate integral over the fixed (but unknown) spatial function $z(\boldsymbol{s})$. (The SRF is denoted $Z(\boldsymbol{s})$; a particular realization, a deterministic function of the spatial coordinates, is denoted $z(\boldsymbol{s})$.)

This interpretation requires an ergodic assumption, which is also needed in the original estimation of $C_z(\boldsymbol{s})$. Refer to Cressie (1993, pp. 53–58) for a discussion of ergodicity and Gaussian SRFs.

Implementation Using PROC SIM2D

The subregion to be considered is the southeast corner of the field, which is a square region with a length of 40 distance units (in thousands of feet). First, you input the thickness data as the following DATA step shows:

```
data thick;
   input East North Thick @@;
   label Thick='Coal Seam Thickness';
   datalines;
 0.7  59.6  34.1   2.1  82.7  42.2   4.7  75.1  39.5
 4.8  52.8  34.3   5.9  67.1  37.0   6.0  35.7  35.9
 6.4  33.7  36.4   7.0  46.7  34.6   8.2  40.1  35.4
13.3   0.6  44.7  13.3  68.2  37.8  13.4  31.3  37.8
17.8   6.9  43.9  20.1  66.3  37.7  22.7  87.6  42.8
23.0  93.9  43.6  24.3  73.0  39.3  24.8  15.1  42.3
24.8  26.3  39.7  26.4  58.0  36.9  26.9  65.0  37.8
27.7  83.3  41.8  27.9  90.8  43.3  29.1  47.9  36.7
29.5  89.4  43.0  30.1   6.1  43.6  30.8  12.1  42.8
32.7  40.2  37.5  34.8   8.1  43.3  35.3  32.0  38.8
37.0  70.3  39.2  38.2  77.9  40.7  38.9  23.3  40.5
39.4  82.5  41.4  43.0   4.7  43.3  43.7   7.6  43.1
46.4  84.1  41.5  46.7  10.6  42.6  49.9  22.1  40.7
51.0  88.8  42.0  52.8  68.9  39.3  52.9  32.7  39.2
55.5  92.9  42.2  56.0   1.6  42.7  60.6  75.2  40.1
62.1  26.6  40.1  63.0  12.7  41.8  69.0  75.6  40.1
70.5  83.7  40.9  70.9  11.0  41.7  71.5  29.5  39.8
78.1  45.5  38.7  78.2   9.1  41.7  78.4  20.0  40.8
80.5  55.9  38.7  81.1  51.0  38.6  83.8   7.9  41.6
84.5  11.0  41.5  85.2  67.3  39.4  85.5  73.0  39.8
86.7  70.4  39.6  87.2  55.7  38.8  88.1   0.0  41.6
88.4  12.1  41.3  88.4  99.6  41.2  88.8  82.9  40.5
88.9   6.2  41.5  90.6   7.0  41.5  90.7  49.6  38.9
91.5  55.4  39.0  92.9  46.8  39.1  93.4  70.9  39.7
55.8  50.5  38.1  96.2  84.3  40.3  98.2  58.2  39.5
;
```

PROC SIM2D is run on the entire data set for conditioning, while the simulation grid covers only this subregion. It is convenient to be able to vary the seed, the grid increment, and the number of simulations performed. The following macro implements the computation of the percent area exceeding the cutoff value by using the seed, the grid increment, and the number of simulated realizations as macro arguments.

Within the macro, the data set produced by PROC SIM2D is transposed with PROC TRANSPOSE so that each grid location is a separate variable. The MEANS procedure is then used to average the simulated value at each grid point over all realizations. It is this average that is compared to the cutoff value. The last DATA step does the comparison, uses an appropriate loop to determine the percent of the grid locations that exceed this cutoff value, and writes the results to the listing file in the form of a report. This sequence is implemented with the following statements:

```
/* Construct macro for conditional simulation ------------------*/

%let cc0=7.2881;
%let aa0=30.6239;
%let ngt=1e-8;
%let form=gauss;
%let cut=39.7;

%macro area_sim(seed=,nr=,ginc=);

   %let ngrid=%eval(40/&ginc+1);
   %let tgrid=%eval(&ngrid*&ngrid);

   proc sim2d data=thick outsim=sim1;
      coordinates xc=east yc=north;
      simulate var=thick numreal=&nr seed=&seed
               scale=&cc0 range=&aa0 nugget=&ngt form=&form;
      mean 40.1173;
      grid x=60 to 100 by &ginc
           y= 0 to  40 by &ginc;
   run;

   proc transpose data=sim1 out=sim2 prefix=sims;
      by _iter_;
      var svalue;
   run;

   proc means data=sim2 noprint n mean;
      var sims1-sims&tgrid;
      output out=msim n=numsim mean=ms1-ms&tgrid;
   run;

   data _null_;
      file print;
      array simss ms1-ms&tgrid;
      set msim;

      cflag=0;
      do ss=1 to &tgrid;
         tempv=simss[ss];
         if simss[ss] > &cut then do;
            cflag + 1;
         end;
      end;

      area_per=100*(cflag/&tgrid);
      put // +5 'Conditional Simulation of Coal Seam'
                ' Thickness for Subregion';
      put / +5 'Subregion is South-East Corner 40 by 40 distance units';
      put / +5 "Seed:&seed" +2 "Grid Increment:&ginc";
      put / +5 "Total Number of Grid Points:&tgrid" +2
               "Number of Simulations:&nr";
```

```
      put / +5 "Percent of Subregion Exceeding Cutoff of %left(&cut) ft.:"
                +2 area_per 5.2;
   run;
%mend area_sim;
```

In the following statement, you invoke the macro three times. Each time, the macro is invoked with a different seed and combination of the grid increment and number of simulations. The macro is first invoked with a relatively coarse grid (grid increment of 10 distance units) and a small number of realizations (5). The output of this conditional simulation is shown in Output 79.1.1.

```
/* Execute macro for coarse grid ------------------------------*/
%area_sim(seed=12345,nr=5,ginc=10);
```

Output 79.1.1 Conditional Simulation of Coal Seam Thickness on a Coarse Grid

```
Conditional Simulation of Coal Seam Thickness for Subregion

Subregion is South-East Corner 40 by 40 distance units

Seed:12345  Grid Increment:10

Total Number of Grid Points:25  Number of Simulations:5

Percent of Subregion Exceeding Cutoff of 39.7 ft.:  76.00
```

The next invocation, in the following statement, uses a finer grid and 50 realizations. The output of the second conditional simulation is shown in Output 79.1.2.

```
/* Execute macro for fine grid and fewer simulations -----------*/
%area_sim(seed=54321,nr=50,ginc=1);
```

Output 79.1.2 Conditional Simulation of Coal Seam Thickness on a Fine Grid

```
Conditional Simulation of Coal Seam Thickness for Subregion

Subregion is South-East Corner 40 by 40 distance units

Seed:54321  Grid Increment:1

Total Number of Grid Points:1681  Number of Simulations:50

Percent of Subregion Exceeding Cutoff of 39.7 ft.:  76.74
```

The final invocation, in the following statement, uses the same grid increment and 500 realizations. The output of this conditional simulation is shown in Output 79.1.3.

```
/* Execute macro for fine grid and more simulations ------------*/
%area_sim(seed=655311,nr=500,ginc=1);
```

Output 79.1.3 Conditional Simulation of Coal Seam Thickness on a Fine Grid

```
Conditional Simulation of Coal Seam Thickness for Subregion

Subregion is South-East Corner 40 by 40 distance units

Seed:655311  Grid Increment:1

Total Number of Grid Points:1681  Number of Simulations:500

Percent of Subregion Exceeding Cutoff of 39.7 ft.:  76.68
```

The results from the preceding simulations indicate that 76% to 77% of the subregion exceeds the cutoff value.

Note also that the number of grid points in the simulation increases with the square of the decrease in the grid increment, leading to long CPU processing times. Increasing the number of realizations results in a linear increase in processing times. Hence, using as coarse a grid as possible allows for more realizations and experimentation with different seeds.

Example 79.2: Variability at Selected Locations

This example exhibits a more detailed investigation of the variation of simulated Thick values. You use the same thick data set from the section "Getting Started: SIM2D Procedure" on page 6149, and you are interested in the simulated values statistics at two selected grid points.

Specifically, you perform a simulation asking for 5,000 realizations (iterations) at two points of the region defined in the section "Preliminary Spatial Data Analysis" on page 6149. These are the extreme southwest point and a point toward the northeast corner of the region. Since you do not want to perform the simulation across the whole region, you need to produce a GDATA= data set to specify the coordinates of the selected points. These steps are implemented using the following DATA step and statements:

```
data thick;
   input East North Thick @@;
   label Thick='Coal Seam Thickness';
   datalines;
    0.7  59.6  34.1   2.1  82.7  42.2   4.7  75.1  39.5
    4.8  52.8  34.3   5.9  67.1  37.0   6.0  35.7  35.9
    6.4  33.7  36.4   7.0  46.7  34.6   8.2  40.1  35.4
   13.3   0.6  44.7  13.3  68.2  37.8  13.4  31.3  37.8
   17.8   6.9  43.9  20.1  66.3  37.7  22.7  87.6  42.8
   23.0  93.9  43.6  24.3  73.0  39.3  24.8  15.1  42.3
   24.8  26.3  39.7  26.4  58.0  36.9  26.9  65.0  37.8
   27.7  83.3  41.8  27.9  90.8  43.3  29.1  47.9  36.7
   29.5  89.4  43.0  30.1   6.1  43.6  30.8  12.1  42.8
   32.7  40.2  37.5  34.8   8.1  43.3  35.3  32.0  38.8
```

```
   37.0  70.3  39.2  38.2  77.9  40.7  38.9  23.3  40.5
   39.4  82.5  41.4  43.0   4.7  43.3  43.7   7.6  43.1
   46.4  84.1  41.5  46.7  10.6  42.6  49.9  22.1  40.7
   51.0  88.8  42.0  52.8  68.9  39.3  52.9  32.7  39.2
   55.5  92.9  42.2  56.0   1.6  42.7  60.6  75.2  40.1
   62.1  26.6  40.1  63.0  12.7  41.8  69.0  75.6  40.1
   70.5  83.7  40.9  70.9  11.0  41.7  71.5  29.5  39.8
   78.1  45.5  38.7  78.2   9.1  41.7  78.4  20.0  40.8
   80.5  55.9  38.7  81.1  51.0  38.6  83.8   7.9  41.6
   84.5  11.0  41.5  85.2  67.3  39.4  85.5  73.0  39.8
   86.7  70.4  39.6  87.2  55.7  38.8  88.1   0.0  41.6
   88.4  12.1  41.3  88.4  99.6  41.2  88.8  82.9  40.5
   88.9   6.2  41.5  90.6   7.0  41.5  90.7  49.6  38.9
   91.5  55.4  39.0  92.9  46.8  39.1  93.4  70.9  39.7
   55.8  50.5  38.1  96.2  84.3  40.3  98.2  58.2  39.5
   ;

data grid;
   input xc yc;
   datalines;
   0   0
   75  75
run;
```

Then, you run PROC SIM2D with the same parameters and characteristics as those in the section "Preliminary Spatial Data Analysis" on page 6149. This time, however, you ask for simulated values only at the two locations you specified in the previous DATA step. The following statements execute the requested simulation:

```
proc sim2d data=thick outsim=sim1;
   coordinates xc=East yc=North;
   simulate var=Thick numreal=5000 seed=79931
            scale=7.2881 range=30.6239 form=gauss;
   mean 40.1173;
   grid gdata=grid xc=xc yc=yc;
run;
```

Once the simulation is performed, summary statistics are computed for each of the specified grid points. In particular, you use PROC UNIVARIATE and a BY statement to request at these locations the quantiles and the extreme observations, as the following statements show:

```
proc sort data=sim1;
   by gxc gyc;
run;

proc univariate data=sim1;
   var svalue;
   by gxc gyc;
   ods select Quantiles ExtremeObs;
   title 'Simulation Statistics at Selected Grid Points';
run;
```

The summary statistics for the first grid point (East=0, North=0) are presented in Output 79.2.1.

Output 79.2.1 Simulation Statistics at Grid Point (East=0, North=0)

```
                Simulation Statistics at Selected Grid Points

------ X-coordinate of the grid point=0 Y-coordinate of the grid point=0 -------

                           The UNIVARIATE Procedure
                Variable:  SVALUE  (Simulated Value at Grid Point)

                           Quantiles (Definition 5)

                            Quantile      Estimate

                            100% Max       42.1810
                            99%            41.6986
                            95%            41.3633
                            90%            41.1890
                            75% Q3         40.9044
                            50% Median     40.5713
                            25% Q1         40.2191
                            10%            39.9463
                            5%             39.7552
                            1%             39.4201
                            0% Min         38.7473

------ X-coordinate of the grid point=0 Y-coordinate of the grid point=0 -------

                              Extreme Observations

                    ------Lowest-----        -----Highest-----

                       Value      Obs           Value      Obs

                     38.7473     2691         42.0656     1149
                     38.9080     1817         42.0805     3612
                     38.9449     3026         42.0980     3757
                     38.9867     2275         42.1782      135
                     39.0194     3100         42.1810     4536
```

Finally, Output 79.2.2 displays the summary statistics for the second grid point (East=75, North=75).

Output 79.2.2 Simulation Statistics at Grid Point (East=75, North=75)

```
                Simulation Statistics at Selected Grid Points

----- X-coordinate of the grid point=75 Y-coordinate of the grid point=75 ------

                           The UNIVARIATE Procedure
              Variable:  SVALUE  (Simulated Value at Grid Point)

                           Quantiles (Definition 5)

                            Quantile      Estimate

                            100% Max       40.1168
                            99%            40.1148
                            95%            40.1134
                            90%            40.1126
                            75% Q3         40.1114
                            50% Median     40.1101
                            25% Q1         40.1086
                            10%            40.1075
                            5%             40.1068
                            1%             40.1053
                            0% Min         40.1024

----- X-coordinate of the grid point=75 Y-coordinate of the grid point=75 ------

                            Extreme Observations

                  ------Lowest-----        -----Highest-----

                     Value      Obs           Value      Obs

                   40.1024     7176         40.1164     8980
                   40.1029     6262         40.1165     9272
                   40.1032     7383         40.1167     5676
                   40.1037     7156         40.1167     6514
                   40.1037     5643         40.1168     5329
```

References

Christakos, G. (1992), *Random Field Models in Earth Sciences*, New York: Academic Press.

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

Deutsch, C. V. and Journel, A. G. (1992), *GSLIB: Geostatistical Software Library and User's Guide*, New York: Oxford University Press.

Knuth, D. E. (1981), *The Art of Computer Programming: Seminumerical Algorithms*, Vol. 2, Second Edition, Reading, MA: Addison-Wesley.

Leadbetter, M. R., Lindgren, G., and Rootzen, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, New York: Springer-Verlag.

Ralston, A. and Rabinowitz, P. (1978), *A First Course in Numerical Analysis*, New York: McGraw-Hill.

Searle, S. R. (1971), *Linear Models*, New York: John Wiley & Sons.

Subject Index

Syntax Index

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **suggest@sas.com**.

SAS® Publishing Delivers!

Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.

SAS® Press

Need to learn the basics? Struggling with a programming problem? You'll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

support.sas.com/saspress

SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – **free** on the Web.
- Hard-copy books.

support.sas.com/publishing

SAS® Publishing News

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

support.sas.com/spn

sas | THE POWER TO KNOW®