Chapter 58
The SIM2D Procedure

Chapter Table of Contents

Chapter 58
The SIM2D Procedure

Overview

The SIM2D procedure produces a spatial simulation for a Gaussian random field with a specified mean and covariance structure in two dimensions 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 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 does not produce any displayed output.

Introduction to Spatial Simulation

The purpose of spatial simulation is to produce a set of partial realizations of a spatial random field (SRF) $Z(\mathbf{s}), \mathbf{s} \in D \subset \mathcal{R}^2$ in a way that preserves a specified mean $\mu(\mathbf{s}) = E[Z(\mathbf{s})]$ and covariance structure $C_z(\mathbf{s}_1 - \mathbf{s}_2) = cov(Z(\mathbf{s}_1), Z(\mathbf{s}_2))$.

The realizations are partial in the sense that they occur only at a finite set of locations $(\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n)$. These locations are typically on a regular grid, but they can be arbitrary locations in the plane.

There are a number of different types of spatial simulation and associated computational methods. PROC SIM2D produces simulations for continuous processes in two dimensions. This means that the possible values of the measured quantity $Z(\mathbf{s}_0)$ at location $\mathbf{s}_0 = (x_0, y_0)$ can vary continuously over a certain range.

An additional assumption, needed for computational purposes, is that the spatial random field $Z(\mathbf{s})$ is Gaussian.

Spatial simulation is different from spatial prediction, where the emphasis is on producing a point estimate 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 $(Z(\mathbf{s}_1), Z(\mathbf{s}_2), \cdots, Z(\mathbf{s}_n))$.

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

Getting Started

Spatial simulation, just like spatial prediction, requires a model of spatial dependence, usually in terms of the covariance $C_z(h)$. For a given set of spatial data $Z(\mathbf{s}_i), i = 1, \cdots, n$, 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 is also used in the "Getting Started" section of Chapter 70, "The VARIOGRAM Procedure."

Preliminary Spatial Data Analysis

In this example, the data consist of coal seam thickness measurements (in feet) taken over an approximately square area. 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.

It is instructive to see the locations of the measured points in the area where you want to perform spatial simulations. It is generally desirable to have these locations scattered evenly around the simulation area.

First, the data are input and the sample locations plotted.

```
data thick;
   input east north thick @@;
   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
      94.8    71.5    39.7    96.2    84.3    40.3    98.2    58.2    39.5
      ;

  proc gplot data=thick;
     title 'Locations of Measured Samples';
     plot north*east / frame cframe=ligr haxis=axis1
                       vaxis=axis2;
     symbol1 v=dot color=blue;
     axis1 minor=none;
     axis2 minor=none label=(angle=90 rotate=0);
     label east   = 'East'
           north  = 'North'
        ;
  run;
```
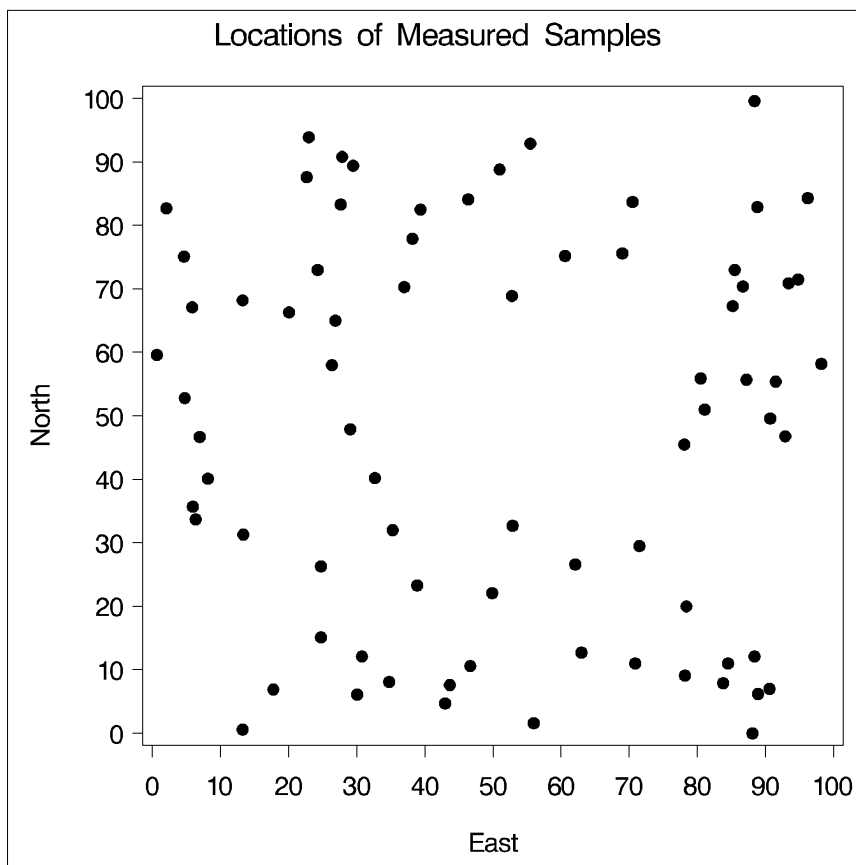

Figure 58.1.   Locations of Measured Samples

```
proc g3d data=thick;
   title 'Surface Plot of Coal Seam Thickness';
   scatter east*north=thick / xticknum=5 yticknum=5
      grid zmin=20 zmax=65;
   label east  = 'East'
         north = 'North'
         thick = 'Thickness'
      ;
run;
```
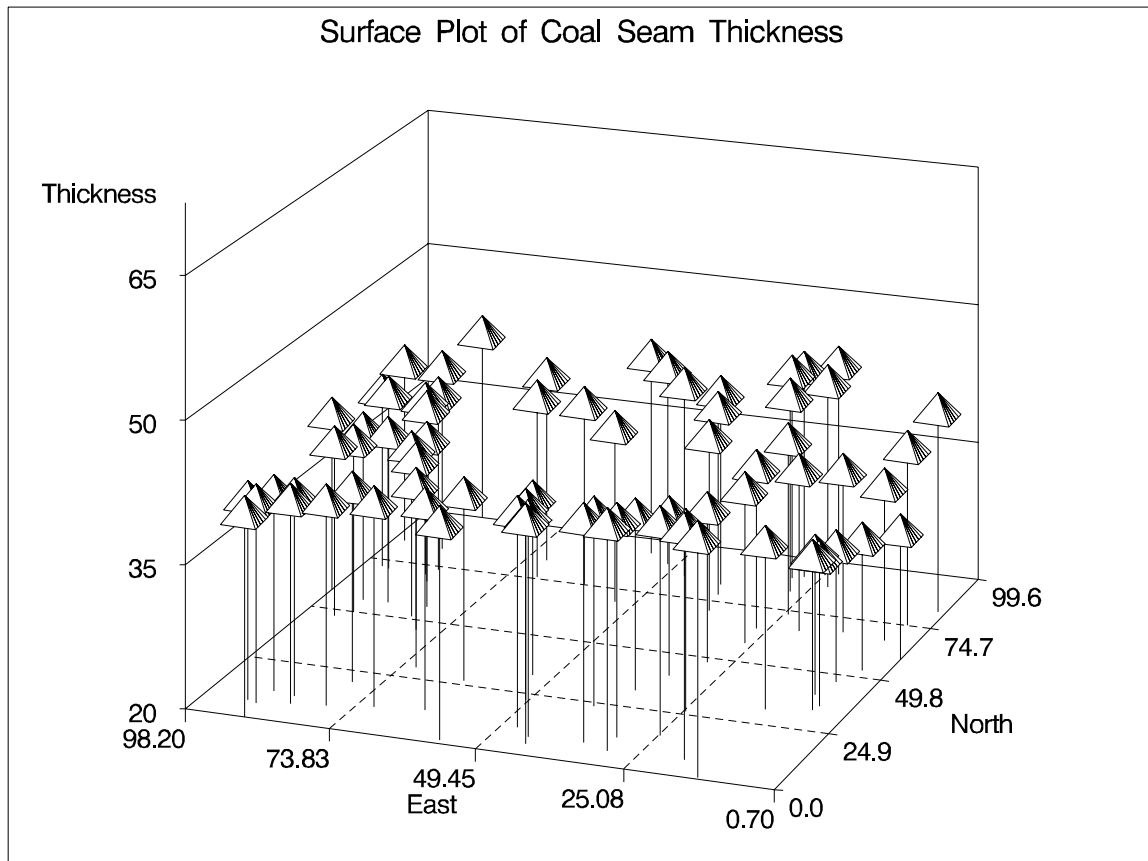

Figure 58.2.   Surface Plot of Coal Seam Thickness

Figure 58.2 shows the small scale variation typical of spatial data, but there does not appear to be any surface trend. Hence, 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

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

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

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

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

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

has a multivariate Gaussian or normal distribution $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The (i,j)th element of $\boldsymbol{\Sigma}$ is computed by $C_z(\mathbf{s}_i - \mathbf{s}_j)$, which happens to be a Gaussian functional form. Any functional form for $C_z(\mathbf{h})$ yielding a valid covariance matrix $\boldsymbol{\Sigma}$ can be used. Both the functional form of $C_z(\mathbf{h})$ and the parameter values

$$\mu = 40.14$$
$$c_0 = 7.5$$
$$a_0 = 30.0$$

are visually estimated using PROC VARIOGRAM, a DATA step, and the GPLOT procedure. Refer to the "Getting Started" section beginning on page 3644 in the chapter on the VARIOGRAM procedure for details on how these parameter values are obtained.

The choice of a Gaussian functional form for $C_z(\mathbf{h})$ 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 $Z(\mathbf{s})$ be a Gaussian SRF. For details, see the section "Computational and Theoretical Details of Spatial Simulation" beginning on page 3103.

Investigating Variability by Simulation

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

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

with the Gaussian covariance structure $C_z(\mathbf{h})$ 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 investigate variations at selected grid points.

To do this investigation, this example uses PROC SIM2D and specifies the Gaussian model with the parameters found previously. Five thousand simulations (iterations) are performed on two points: the extreme south-west point of the region and a point towards the north-east corner of the region. Because of the irregular nature of these points, a GDATA= data set is produced with the coordinates of the selected points.

Summary statistics are computed for each of these grid points by using a BY statement in PROC UNIVARIATE.

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

proc sim2d data=thick outsim=sim1;
   simulate var=thick numreal=5000 seed=79931
      scale=7.5 range=30.0 form=gauss;
   mean 40.14;
   coordinates xc=east yc=north;
   grid gdata=grid xc=xc yc=yc;
run;

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

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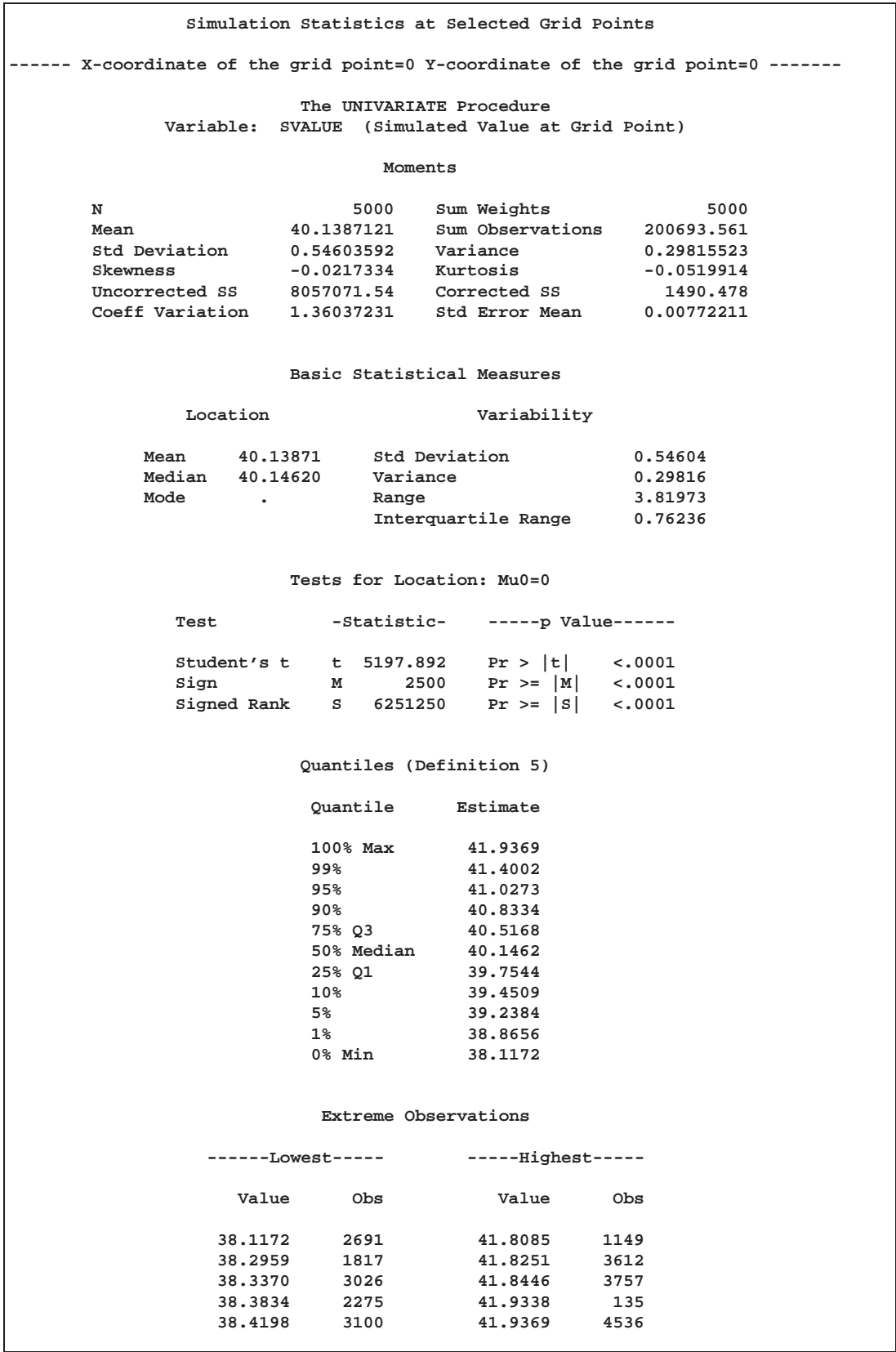
```
                    Simulation Statistics at Selected Grid Points

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

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

                                   Moments

         N                         5000    Sum Weights                    5000
         Mean                 40.1387121   Sum Observations         200693.561
         Std Deviation         0.54603592  Variance                 0.29815523
         Skewness             -0.0217334   Kurtosis                -0.0519914
         Uncorrected SS       8057071.54   Corrected SS              1490.478
         Coeff Variation       1.36037231  Std Error Mean           0.00772211


                            Basic Statistical Measures

              Location                          Variability

          Mean      40.13871      Std Deviation               0.54604
          Median    40.14620      Variance                    0.29816
          Mode         .          Range                       3.81973
                                  Interquartile Range         0.76236


                            Tests for Location: Mu0=0

             Test              -Statistic-      -----p Value------

             Student's t    t   5197.892      Pr > |t|     <.0001
             Sign           M      2500        Pr >= |M|    <.0001
             Signed Rank    S   6251250        Pr >= |S|    <.0001


                            Quantiles (Definition 5)

                            Quantile      Estimate

                            100% Max       41.9369
                            99%            41.4002
                            95%            41.0273
                            90%            40.8334
                            75% Q3         40.5168
                            50% Median     40.1462
                            25% Q1         39.7544
                            10%            39.4509
                            5%             39.2384
                            1%             38.8656
                            0% Min         38.1172


                            Extreme Observations

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

                    Value     Obs              Value     Obs

                   38.1172    2691            41.8085    1149
                   38.2959    1817            41.8251    3612
                   38.3370    3026            41.8446    3757
                   38.3834    2275            41.9338     135
                   38.4198    3100            41.9369    4536
```

Figure 58.3. Simulation Statistics at Grid Point (XC=0, YC=0)

```
                   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)

                                    Moments

        N                         5000     Sum Weights                      5000
        Mean                40.1386868     Sum Observations           200693.434
        Std Deviation       0.00250643     Variance                    6.2822E-6
        Skewness            0.00937779     Kurtosis                   -0.0088601
        Uncorrected SS      8055570.91     Corrected SS               0.03140472
        Coeff Variation     0.00624443     Std Error Mean             0.00003545


                            Basic Statistical Measures

                  Location                        Variability

               Mean     40.13869     Std Deviation           0.00251
               Median   40.13870     Variance              6.2822E-6
               Mode        .         Range                   0.01756
                                     Interquartile Range     0.00346


                          Tests for Location: Mu0=0

                 Test              -Statistic-      -----p Value------

                 Student's t    t   1132380     Pr > |t|    <.0001
                 Sign           M      2500     Pr >= |M|   <.0001
                 Signed Rank    S   6251250     Pr >= |S|   <.0001


                             Quantiles (Definition 5)

                            Quantile      Estimate

                            100% Max       40.1468
                            99%            40.1445
                            95%            40.1428
                            90%            40.1419
                            75% Q3         40.1404
                            50% Median     40.1387
                            25% Q1         40.1369
                            10%            40.1355
                            5%             40.1346
                            1%             40.1328
                            0% Min         40.1293


                               Extreme Observations

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

                      Value      Obs           Value      Obs

                     40.1293     2176         40.1465     1278
                     40.1299     1262         40.1465     3980
                     40.1302     2383         40.1468      676
                     40.1306     2156         40.1468     1514
                     40.1308      643         40.1468      329
```

Figure 58.4. Simulation Statistics at Grid Point (XC=75, YC=75)

Syntax

The following statements are available in PROC SIM2D.

> **PROC SIM2D** *options* ;
>   **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 58.1. 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=
specify quadratic form data set	MEAN	QDATA=
write simulated values	PROC SIM2D	OUTSIM=
Declaring the Role of Variables		
specify the conditioning variable	SIMULATE	VAR=
specify the x and y coordinate variables in the DATA= 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=
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=

Table 58.1. (continued)

Task	Statement	Option
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 containing the x and y coordinate variables and the SIMU-LATE VAR= variables. This data set is required if any of the SIMULATE statements are conditional, that is, if you specify the VAR= option. If none of the SIMULATE statements are conditional then 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 3107.

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 3107.

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 a 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, the simulation is *uncon-*

ditional. 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 allows specification of both nested and nonnested 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 semi-variogram 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 using a spherical covariance, an MDATA= data set might look like the following.

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

The PROC SIM2D statement to use the MDATA= specification is of the form

```
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 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 1727 in Chapter 34, "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 SIMn 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 1727 in the chapter on 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" beginning on page 1721 in the chapter on 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 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" beginning on page 1721 in the chapter on 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. If you omit the SEED= value, the system clock is used.

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" beginning on page 3103.

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

```
mean 1.4 + 2.5*x + 3.6*y + .47*x*x + .58*y*y + .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 following MEAN statement

```
mean 1.4 + 2.5*x + 3.6*y + .47*x*x + .58*y*y + .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 .47 .58 .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

```
data q1;
   input const cx cy cxx cyy cxy;
   datalines;
   1.4 2.5 3.6 .47 .58 .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

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 Duetsch 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 $LU = \mathbf{LL}^T$ 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 a little subtle; refer to Knuth (1981, Vol. 2, 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

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

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 $\mathbf{Z} = \mathbf{LW}$. Note that this is where the assumption of a Gaussian SRF is crucial. When $\mathbf{W} \sim N_k(\mathbf{0}, \mathbf{I})$, then $\mathbf{Z} = \mathbf{LW}$ is also Gaussian. The mean of $\mathbf{Z}$ is

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

and the variance is

$$\mathrm{Var}(\mathbf{Z}) = \mathrm{Var}(\mathbf{LW}) = E(\mathbf{LWW}^T\mathbf{L}^T) = \mathbf{L}E(\mathbf{WW}^T)\mathbf{L}^T = \mathbf{LL}^T = \mathbf{C}$$

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

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

with corresponding covariance matrix

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

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

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

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

For a conditional simulation, this distribution of

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

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

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

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

and

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

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 $\mathbf{X}_2 = \mathbf{x}_2$ is $N_k(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$, with

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

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 $\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2, \cdots, \tilde{\mathbf{s}}_n$, with values $\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_n$. Let $\tilde{\mathbf{Z}}$ denote the random vector

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

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 $\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2, \cdots, \tilde{\mathbf{s}}_n$ 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

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

Note that the means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are computed using the grid coordinates $\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_k$, the data coordinates $\tilde{\mathbf{s}}_1, \tilde{\mathbf{s}}_2, \cdots, \tilde{\mathbf{s}}_n$, 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 $\mathbf{A}$ 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

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

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.

Example

Example 58.1. Simulation

Continuing with the coal seam thickness example from the "Getting Started" section beginning on page 3088, 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 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 percent of the subregion must exceed a seam thickness of 39.7 feet for profitability.

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

Spatial prediction does not appear to be helpful in answering this question. While it is easy to determine if 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 percent value needed for profitability.

Example 58.1. Simulation ⬥ 3109

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 $Z(\mathbf{s}_i), i = 1, \cdots, 75$ represent a *single* realization from the set of all possible realizations consistent with the SRF model (μ and $C_z(\mathbf{h})$). A conditional simulation repeatedly produces other possible simulated realizations consistent with the model and data. However, the only concern of the mining company is with this single unique realization. It is not concerned with similar coal fields to be mined sometime in the future; it may never see another coal field remotely similar to this one, or it may 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 $z(\mathbf{s})$. (The SRF is denoted $Z(\mathbf{s})$; a particular realization, a deterministic function of the spatial coordinates, is denoted $z(\mathbf{s})$.)

This interpretation requires an ergodic assumption, which is also needed in the original estimation of $C_z(\mathbf{s})$. 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 length 40 distance units (in thousands of feet). 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.

The data set produced by PROC SIM2D is transposed 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 and determines 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.

The macro is first invoked with a relatively coarse grid (grid increment of 10 distance units) and a small number of realizations (5). The next invocation uses a finer grid and 50 realizations, and the final invocation uses the same grid increment and 500 realizations. Each time, the macro is invoked with a different seed. The simulations indicate that around 87 percent of the subregion exceeds the cutoff value.

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 more realizations and experimentation with different seeds.

```
/*- Set the covariance model parameters and cutoff value -*/
%let cc0=7.5;
%let aa0=30.0;
%let form=gauss;
%let cut=39.7;

data thick;
   input east north thick @@;
   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
   94.8  71.5  39.7   96.2  84.3  40.3   98.2  58.2  39.5
   ;


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

Example 58.1.   Simulation   ⬧   3111

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

   proc sim2d data=thick outsim=sim1;
      simulate var=thick numreal=&nr seed=&seed
         scale=&cc0 range=&aa0 form=&form;
      mean 40.14;
      coordinates xc=east yc=north;
      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;

   /*- Determine the percentage of sites exceeding cutoff -*/
   data _null_;
      file print;
      array simss ms1-ms&tgrid;
      set msim;

   /*- Loop over the grid sites to test cutoff  -*/
      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
               &cut ft.:"
               +2 area_per 5.2;
   run;
%mend area_sim;
```

```
%area_sim(seed=12345,nr=5,ginc=10);
%area_sim(seed=54321,nr=50,ginc=1);
%area_sim(seed=655311,nr=500,ginc=1);
```

Output 58.1.1.   Conditional Simulation of Coal Seam Thickness

```
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.:  80.00
```

```
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.:  87.33
```

```
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.:  87.57
```

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, Inc.

Duetsch, 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*, Second Edition, New York: McGraw-Hill, Inc.

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