Chapter 70
The VARIOGRAM Procedure

Chapter Table of Contents

Chapter 70
The VARIOGRAM Procedure

Overview

The VARIOGRAM procedure computes sample or empirical measures of spatial continuity for two-dimensional spatial data. These continuity measures are the regular semivariogram, a robust version of the semivariogram, and the covariance. The continuity measures are written to an output data set, allowing plotting or parameter estimation for theoretical semivariogram or covariance models. Both isotropic and anisotropic measures are available.

The VARIOGRAM procedure produces two additional output data sets that are useful in the analysis of pairwise distances in the original data. The OUTPAIR= data set contains one observation for each pair of points. The coordinates, distance, angle, and values of the analysis variables are written to this data set. The OUTDISTANCE= data set contains histogram information on the count of pairs within distance intervals, which is useful for determining unit lag distances.

Introduction to Spatial Prediction

Spatial prediction, in general, is any prediction method that incorporates spatial dependence. A simple and popular spatial prediction method is ordinary kriging.

Ordinary kriging requires a model of the spatial continuity, or dependence. This is typically in the form of a covariance or semivariogram.

Spatial prediction, then, involves two steps. First, you model the covariance or semivariogram of the spatial process. This involves choosing both a mathematical form and the values of the associated parameters. Second, you use this dependence model in solving the kriging system at a specified set of spatial points, resulting in predicted values and associated standard errors.

SAS/STAT software has two procedures corresponding to these steps for spatial prediction of two-dimensional data. The VARIOGRAM procedure is used in the first step. By computing a sample estimate of the variogram or covariance, you can choose a theoretical model based on graphical or other means.

Getting Started

In activities such as reservoir estimation in mining, petroleum exploration, and environmental modeling of air and water pollution, it often happens that data on one or more quantities are available at given spatial locations, and the goal is to predict the measured quantities at unsampled locations. Often, these unsampled locations are on a regular grid, and the predictions are used to produce surface plots or contour maps.

A popular method of spatial prediction is ordinary kriging, which produces both predicted values and associated standard errors. Ordinary kriging requires the complete specification (the form and parameter values) of the spatial dependence of the spatial process in terms of a covariance or semivariogram model.

Typically the semivariogram model is not known in advance and must be estimated, either visually or by some estimation method.

PROC VARIOGRAM computes the sample semivariogram, from which you can find a suitable theoretical semivariogram by visual methods.

The following example goes through a typical problem to show how you can compute a sample variogram and determine an appropriate theoretical model.

Preliminary Spatial Data Analysis

The simulated 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.

First, the data are input.

```
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
   ;
```

It is instructive to see the locations of the measured points in the area where you want to perform spatial prediction. It is desirable to have these locations scattered evenly around the prediction area. If this is not the case, the prediction error might be unacceptably large where measurements are sparse. The following GPLOT procedure is useful in determining potential problems:

```
proc gplot data=thick;
   title 'Scatter Plot of Measurement Locations';
   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;
```

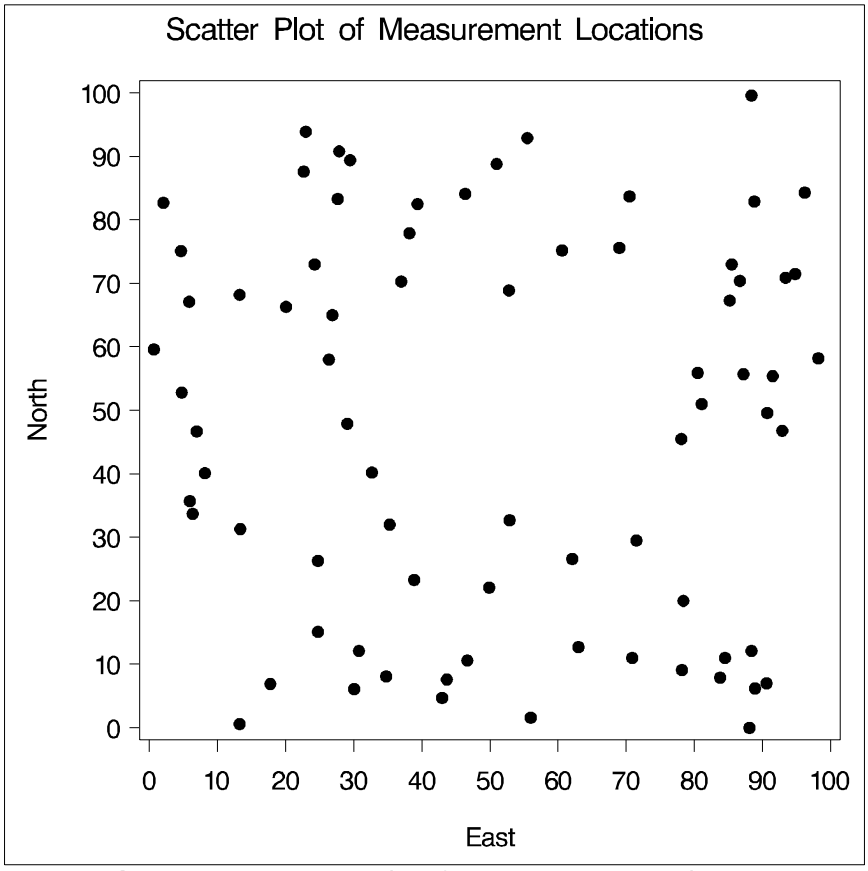


Figure 70.1. Scatter Plot of Measurement Locations

As Figure 70.1 indicates, while the locations are not ideally spread around the prediction area, there are not any large areas lacking measurements. You now can look at a surface plot of the measured variable, the thickness of coal seam, using the G3D procedure. This is a crucial step. Any obvious surface trend has to be removed before you compute and estimate the model of spatial dependence (the semivariogram model).

```
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;
```

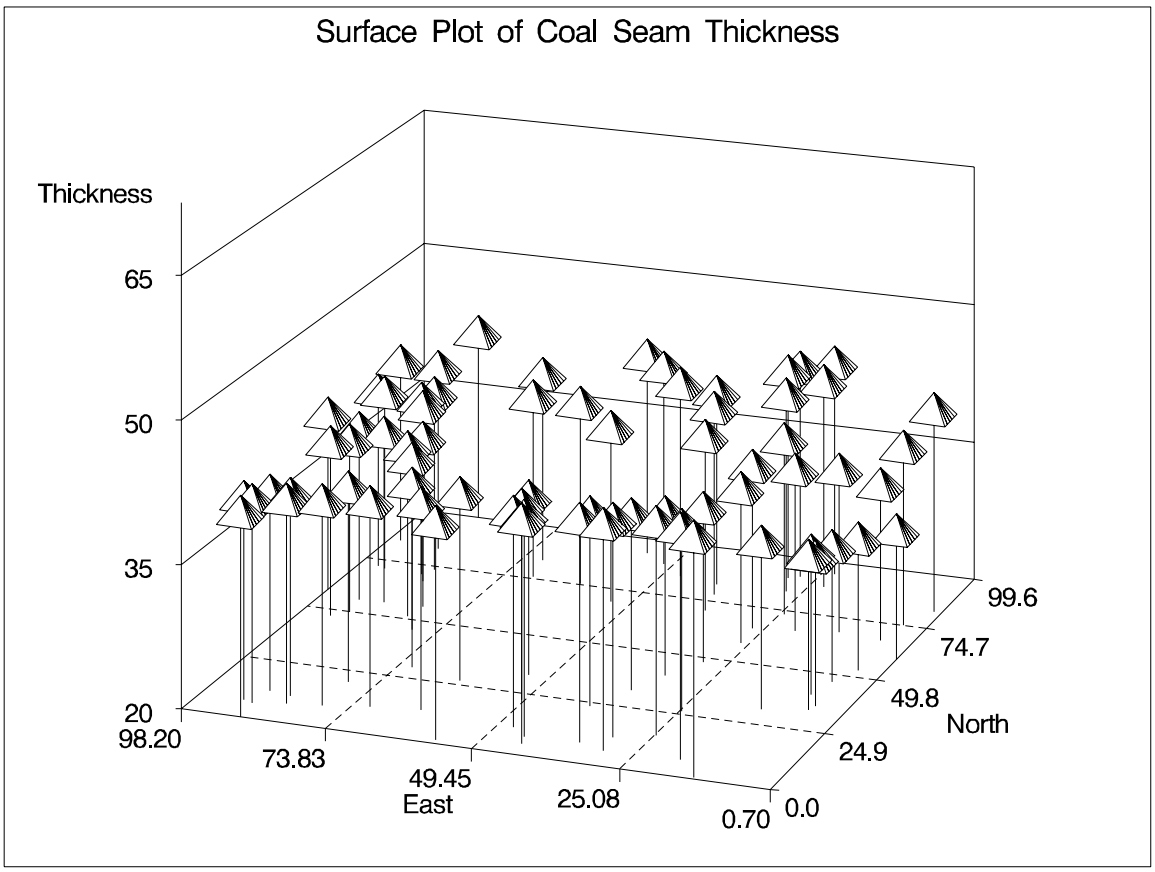


Figure 70.2. Surface Plot of Coal Seam Thickness

Figure 70.2 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.

Preliminary Variogram Analysis

Recall that the goal of this example is spatial prediction. In particular, you would like to produce a contour map or surface plot on a regular grid of predicted values based on ordinary kriging. Ordinary kriging requires the complete specification of the spatial covariance or semivariogram.

You can use PROC VARIOGRAM, along with a DATA step and PROC GPLOT, to estimate visually a reasonable semivariogram model (both the form and associated parameters) for the thickness data.

Before proceeding with this estimation, consider the formula for the empirical or experimental semivariogram $\gamma_z(h)$. Denote the coal seam thickness process by $\{Z(r), r \in D \subset \mathcal{R}^2\}$. You have measurements $(Z(r_i), i = 1, \ldots, 75)$. The standard formula for $\gamma_z(h)$ (isotropic case) is

$$2\gamma_z(h) = \frac{1}{\mid N(h) \mid} \sum_{N(h)} (Z(r_i) - Z(r_j))^2$$

where $N(h)$ is given by

$$N(h) = \{i, j :\mid r_i - r_j \mid = h\}$$

and $\mid N(h) \mid$ is the number of such pairs (i, j).

For actual data, it is unlikely that any pair (i, j) would exactly satisfy $\mid r_i - r_j \mid = h$, so typically a range of pairwise distances, $\mid r_i - r_j \mid \in [h - \delta h, h + \delta h)$, is used to group pairs (r_i, r_j) for a single term in the expression for $\gamma_z(h)$. Using this range, $N(h)$ is modified by

$$N(h, \delta h) = \{i, j :\mid r_i - r_j \mid \in [h - \delta h, h + \delta h)\}$$

PROC VARIOGRAM performs this grouping with two required options for variogram computation: the LAGDISTANCE= and MAXLAGS= options.

The meaning of the required LAGDISTANCE= option is as follows. Classify all pairs of points into intervals according to their pairwise distance. The width of the distance interval is the LAGDISTANCE= value. The meaning of the required MAXLAGS= option is simply the number of intervals.

The problem is that a surface plot of the original data, or the scatter plot of the measurement locations, is not very helpful in determining the distribution of these pairwise distances; it is not clear what values to give to the LAGDISTANCE= and MAXLAGS= options.

You use PROC VARIOGRAM with the OUTDISTANCE= option to produce a modified histogram of the pairwise distances in order to find reasonable values for the LAGDISTANCE= and MAXLAGS= options. In the following analysis, you use the NOVARIOGRAM option in the COMPUTE statement and the OUTDISTANCE=

option in the PROC VARIOGRAM statement. You need the NOVARIOGRAM option to keep an error message from being issued due to the absence of the LAGDISTANCE= and MAXLAGS= options.

The DATA step after the PROC VARIOGRAM statement computes the midpoint of each distance interval. This midpoint is then used in the GCHART procedure. Since the number of distance intervals is not specified by using the NHCLASSES= option in the COMPUTE statement, the default of 10 is used.

```
proc variogram data=thick outdistance=outd;
   compute novariogram;
   coordinates xc=east yc=north;
   var thick;
run;

title 'OUTDISTANCE= Data Set Showing Distance Intervals';
proc print data=outd;
run;

data outd; set outd;
   mdpt=round((lb+ub)/2,.1);
   label mdpt = 'Midpoint of Interval';
run;

axis1 minor=none;
axis2 minor=none label=(angle=90 rotate=0);
title 'Distribution of Pairwise Distances';
proc gchart data=outd;
   vbar mdpt / type=sum sumvar=count discrete frame
               cframe=ligr gaxis=axis1 raxis=axis2 nolegend;
run;
```

OUTDISTANCE= Data Set Showing Distance Intervals

Obs	VARNAME	LAG	LB	UB	COUNT	PER
1	thick	0	0.000	6.969	45	0.01622
2	thick	1	6.969	20.907	263	0.09477
3	thick	2	20.907	34.845	383	0.13802
4	thick	3	34.845	48.783	436	0.15712
5	thick	4	48.783	62.720	495	0.17838
6	thick	5	62.720	76.658	525	0.18919
7	thick	6	76.658	90.596	412	0.14847
8	thick	7	90.596	104.534	179	0.06450
9	thick	8	104.534	118.472	35	0.01261
10	thick	9	118.472	132.410	2	0.00072
11	thick	10	132.410	146.348	0	0.00000

Figure 70.3. OUTDISTANCE= Data Set Showing Distance Intervals

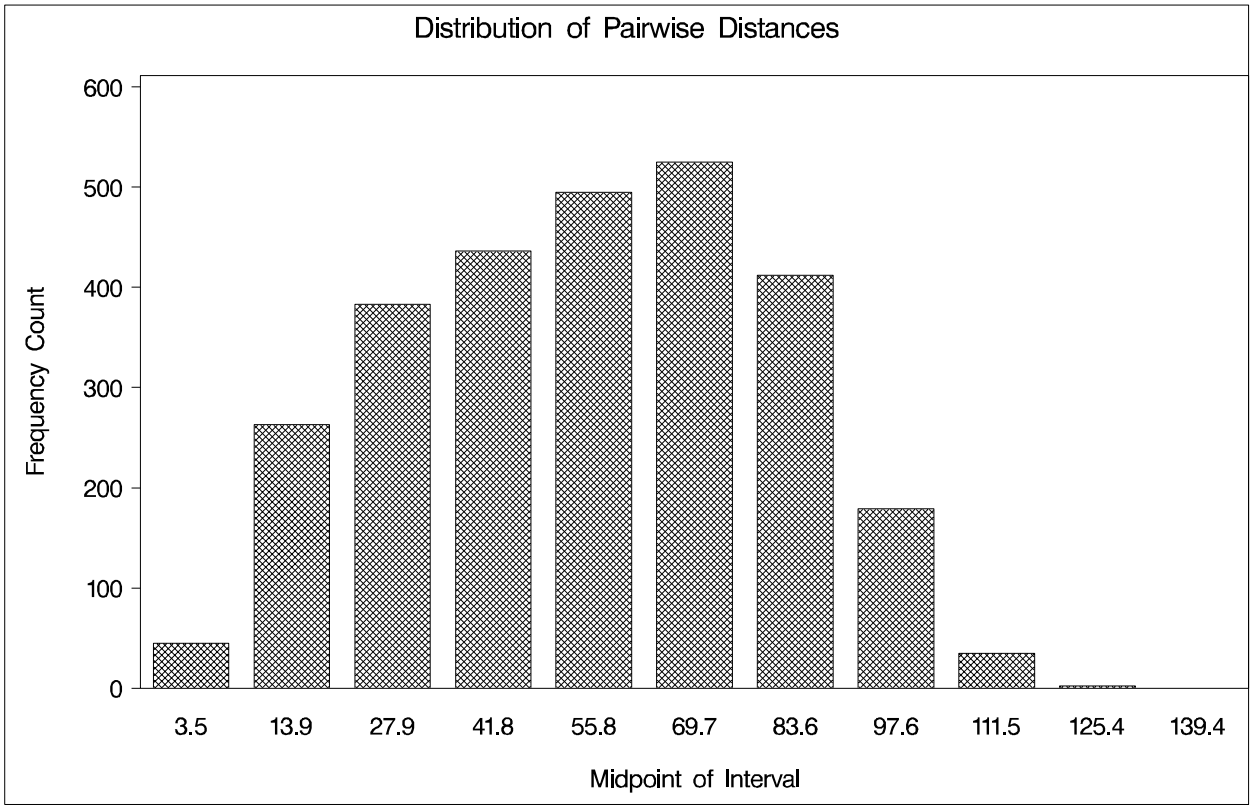


Figure 70.4. Distribution of Pairwise Distances

For plotting and estimations purposes, it is desirable to have as many points as possible for the plot of $\gamma_z(h)$ against h. This corresponds to having as many distance intervals as possible, that is, having a small value for the LAGDISTANCE= option.

However, a rule of thumb used in computing sample semivariograms is to use at least 30 point pairs in computing a single value of the empirical or experimental semivariogram.

If the LAGDISTANCE= value is set too small, there may be too few points in one or more of the intervals. On the other hand, if the LAGDISTANCE= value is set to a large value, the number of point pairs in the distance intervals may be much greater than that needed for estimation precision, thereby "wasting" point pairs at the expense of variogram points.

Hence, there is a tradeoff between the number of distance intervals and the number of point pairs within each interval.

As discussed in the section "OUTDIST=*SAS-data-set* " on page 3670 the first few distance intervals, corresponding to lag 0 and lag 1, are typically the limiting intervals. This is particularly true for lag 0 since it is half the width of the remaining intervals. For the default of NHCLASSES=10, the lag 0 class contains 45 points, which is reasonably close to 30, but the lag 1 class contains 263 points.

If you rerun PROC VARIOGRAM with NHCLASSES=20, these numbers become 8 and 83 for lags 0 and 1, respectively. Because of the asymmetrical nature of lag 0, you are willing to violate the rule of thumb for the 0th lag. You will, however, have sufficient numbers in lag 1 and above.

```
proc variogram data=thick outdistance=outd;
   compute nhc=20 novariogram;
   coordinates xc=east yc=north;
   var thick;
run;

title 'OUTDISTANCE= Data Set Showing Distance Intervals';
proc print data=outd;
run;

data outd; set outd;
   mdpt=round((lb+ub)/2,.1);
   label mdpt = 'Midpoint of Interval';
run;

axis1 minor=none;
axis2 minor=none label=(angle=90 rotate=0);
title 'Distribution of Pairwise Distances';
proc gchart data=outd;
   vbar mdpt / type=sum sumvar=count discrete frame
               cframe=ligr gaxis=axis1 raxis=axis2 nolegend;
run;
```

OUTDISTANCE= Data Set Showing Distance Intervals

Obs	VARNAME	LAG	LB	UB	COUNT	PER
1	thick	0	0.000	3.484	8	0.00288
2	thick	1	3.484	10.453	83	0.02991
3	thick	2	10.453	17.422	143	0.05153
4	thick	3	17.422	24.391	167	0.06018
5	thick	4	24.391	31.360	198	0.07135
6	thick	5	31.360	38.329	197	0.07099
7	thick	6	38.329	45.298	203	0.07315
8	thick	7	45.298	52.267	235	0.08468
9	thick	8	52.267	59.236	234	0.08432
10	thick	9	59.236	66.205	284	0.10234
11	thick	10	66.205	73.174	264	0.09514
12	thick	11	73.174	80.143	236	0.08505
13	thick	12	80.143	87.112	221	0.07964
14	thick	13	87.112	94.081	165	0.05946
15	thick	14	94.081	101.050	75	0.02703
16	thick	15	101.050	108.018	41	0.01477
17	thick	16	108.018	114.987	15	0.00541
18	thick	17	114.987	121.956	5	0.00180
19	thick	18	121.956	128.925	1	0.00036
20	thick	19	128.925	135.894	0	0.00000
21	thick	20	135.894	142.863	0	0.00000

Figure 70.5. OUTDISTANCE= Data Set Showing Distance Intervals

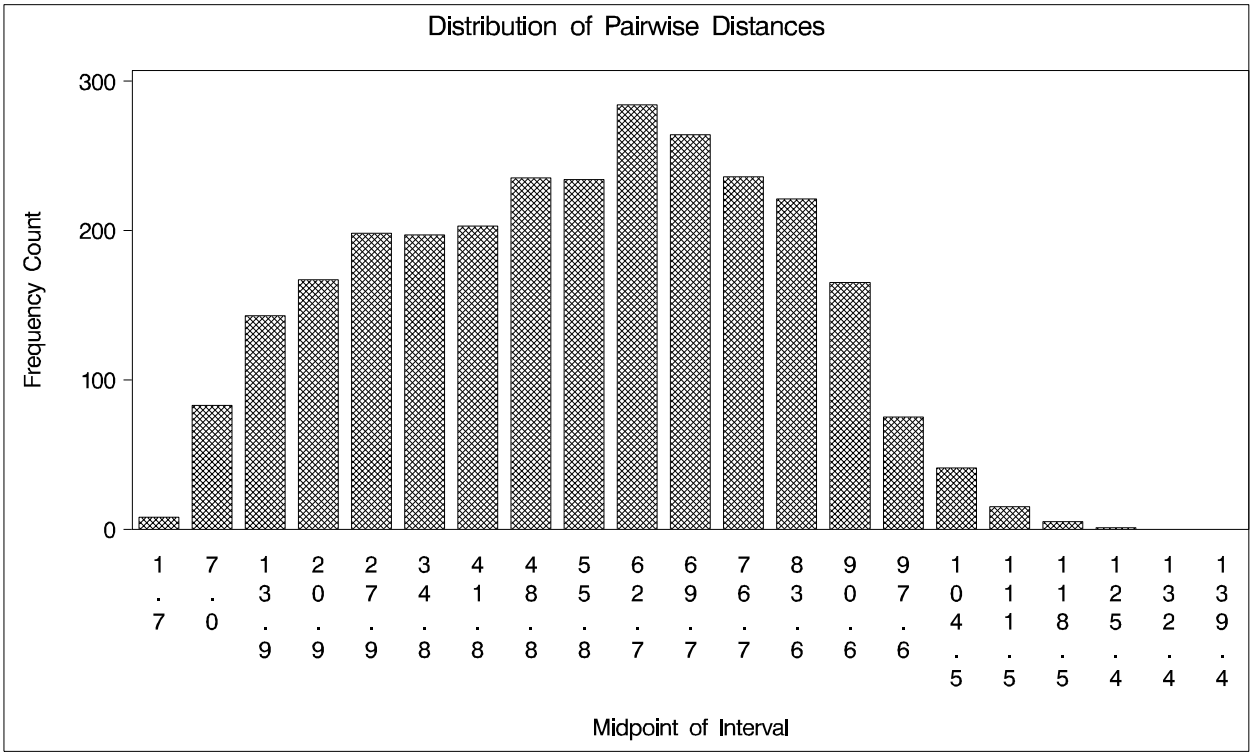


Figure 70.6. Distribution of Pairwise Distances

The length of the lag 1 class $(3.484, 10.453)$ is 6.969. You round off and use LAGDISTANCE=7.0 in the COMPUTE statement.

The use of the MAXLAGS= option is more difficult. From Figure 70.5, note that up to a pairwise distance of 101, you have a sufficient number of pairs. With your choice of LAGDISTANCE=7.0, this yields a maximum number of lags $\frac{101}{7} \approx 14$.

The problem with using the maximum lag value is that it includes pairs of points so far apart that they are likely to be independent. Using pairs of points that are independent adds nothing to the empirical semivariogram plot; they are essentially added noise.

If there is an estimate of correlation length, perhaps from a prior geologic study of a similar site, you can specify the MAXLAGS= value so that the maximum pairwise distance does not exceed two or three correlation lengths. If there is no estimate of correlation length, you can use the following rule of thumb: use $\frac{1}{2}$ to $\frac{3}{4}$ of the "diameter" of the region containing the data. A MAXLAGS= value of 10 is within this range.

You now rerun PROC VARIOGRAM with these values.

Sample Variogram Computation and Plots

Using the values of LAGDISTANCE=7.0 and MAXLAGS=10 computed previously, rerun PROC VARIOGRAM without the NOVARIOGRAM option. Also, request a robust version of the semivariogram; then, plot both results against the pairwise distance of each class.

```
proc variogram data=thick outv=outv;
   compute lagd=7 maxlag=10 robust;
   coordinates xc=east yc=north;
   var thick;
run;

title 'OUTVAR= Data Set Showing Sample Variogram Results';
proc print data=outv label;
   var lag count distance variog rvario;
run;

data outv2; set outv;
   vari=variog; type = 'regular'; output;
   vari=rvario; type = 'robust'; output;
run;

title 'Standard and Robust Semivariogram for Coal Seam
       Thickness Data';
proc gplot data=outv2;
   plot vari*distance=type / frame cframe=ligr vaxis=axis2
                             haxis=axis1;
   symbol1 i=join l=1 c=blue   /* v=star   */;
   symbol2 i=join l=1 c=yellow /* v=square */;
   axis1 minor=none
         label=(c=black 'Lag Distance') /* offset=(3,3) */;
   axis2 order=(0 to 9 by 1) minor=none
         label=(angle=90 rotate=0 c=black 'Variogram')
         /* offset=(3,3) */;
run;
```

OUTVAR= Data Set Showing Sample Variogram Results

Obs	Lag Class Value (in LAGDIST= units)	Number of Pairs in Class	Average Lag Distance for Class	Variogram	Robust Variogram
1	-1	75	.	.	.
2	0	8	2.5045	0.02937	0.01694
3	1	85	7.3625	0.38047	0.19807
4	2	142	14.1547	1.15158	0.98029
5	3	169	21.0913	2.79719	3.01412
6	4	199	27.9691	4.68769	4.86998
7	5	199	35.1591	6.16018	6.15639
8	6	205	42.2547	7.58912	8.05072
9	7	232	48.7775	7.12506	7.07155
10	8	244	56.1824	7.04832	7.62851
11	9	285	62.9121	6.66298	8.02993
12	10	262	69.8925	6.18775	7.92206

Figure 70.7. OUTVAR= Data Set Showing Sample Variogram Results

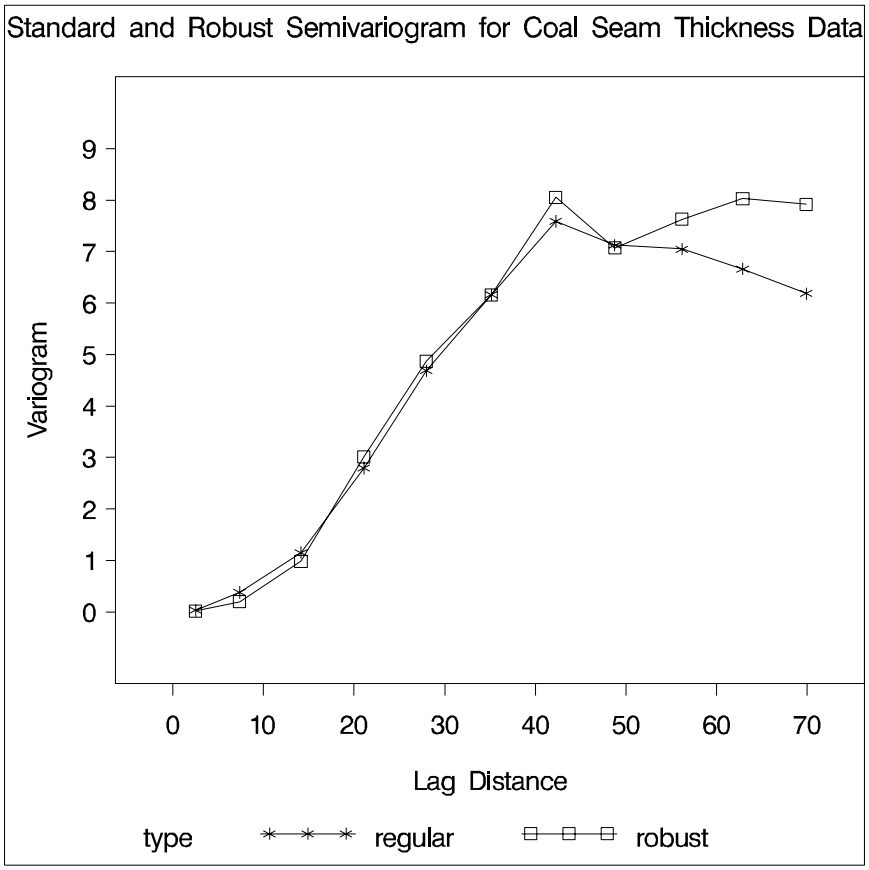


Figure 70.8. Standard and Robust Semivariogram for Coal Seam Thickness Data

Figure 70.8 shows first a slow, then a rapid rise from the origin, suggesting a Gaussian type form:

$$\gamma_z(h) = c_0 \left[1 - \exp\left(-\frac{h^2}{a_0^2}\right)\right]$$

See the section "Theoretical and Computational Details of the Semivariogram" on page 3664 for graphs of the standard semivariogram forms.

By experimentation, you find that a scale of $c_0 = 7.5$ and a range of $a_0 = 30$ fits reasonably well for both the robust and standard semivariogram

The following statements plot the sample and theoretical variograms:

```
data outv3; set outv;
   c0=7.5; a0=30;
   vari = c0*(1-exp(-distance*distance/(a0*a0)));
   type = 'Gaussian'; output;
   vari = variog; type = 'regular'; output;
   vari = rvario; type = 'robust'; output;
run;

title 'Theoretical and Sample Semivariogram for Coal Seam
       Thickness Data';
proc gplot data=outv3;
   plot vari*distance=type / frame cframe=ligr vaxis=axis2
                             haxis=axis1;
   symbol1 i=join l=1 c=blue    /* v=star    */;
   symbol2 i=join l=1 c=yellow  /* v=square  */;
   symbol3 i=join l=1 c=cyan    /* v=diamond */;
   axis1 minor=none
         label=(c=black 'Lag Distance') /* offset=(3,3) */;
   axis2 order=(0 to 9 by 1) minor=none
         label=(angle=90 rotate=0 c=black 'Variogram')
         /* offset=(3,3) */;
run;
```

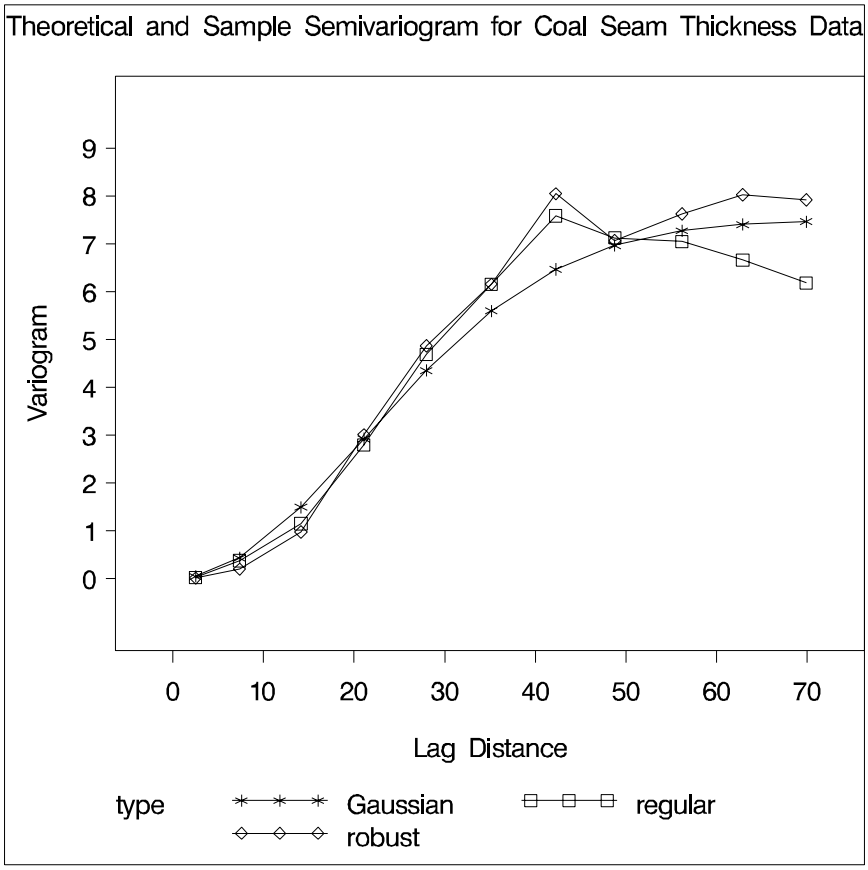


Figure 70.9. Theoretical and Sample Semivariogram for Coal Seam Thickness Data

Figure 70.9 shows that the choice of a semivariogram model is adequate. You can use this Gaussian form and these particular parameters in PROC KRIGE2D to produce a contour plot of the kriging estimates and the associated standard errors.

Syntax

The following statements are available in PROC VARIOGRAM.

PROC VARIOGRAM *options* **;**
COMPUTE *computation-options* **;**
COORDINATES *coordinate-variables* **;**
DIRECTIONS *directions-list* **;**
VAR *analysis-variables-list* **;**

The COMPUTE and COORDINATES statements are required.

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

Table 70.1. Options Available in the VARIOGRAM Procedure

Task	Statement	Option
Data Set Options		
specify input data set	PROC VARIOGRAM	DATA=
write spatial continuity measures	PROC VARIOGRAM	OUTVAR=
write distance histogram information	PROC VARIOGRAM	OUTDISTANCE=
write pairwise point information	PROC VARIOGRAM	OUTPAIR=
Declaring the Role of Variables		
specify the analysis variables	VAR	
specify the x, y coordinates in the DATA= data set	COORDINATES	XCOORD= YCOORD=
Controlling Continuity Measure Computations		
specify the basic lag distance	COMPUTE	LAGDISTANCE=
specify the tolerance around the lag distance	COMPUTE	LAGDISTANCE=
specify the maximum number of lags in computations	COMPUTE	MAXLAGS=
specify the number of angle classes	COMPUTE	NDIRECTIONS=
specify the angle tolerances for angle classes	COMPUTE	ANGLETOL=
specify the bandwidths for angle classes	COMPUTE	BANDWIDTH=
compute robust semivariogram	COMPUTE	ROBUST
suppress computation of all continuity measures	COMPUTE	NOVARIOGRAM
Controlling Distance Histogram Data Set		
specify the distance histogram data set	PROC VARIOGRAM	OUTDISTANCE=
specify the number of histogram classes	COMPUTE	NHCLASSES=
Controlling Pairwise Information Data Set		
specify the pairwise data set	PROC VARIOGRAM	OUTPAIR=
specify the maximum distance for the pairwise data set	COMPUTE	OUTPDISTANCE=

PROC VARIOGRAM Statement

PROC VARIOGRAM *options* **;**

You can specify the following options in the PROC VARIOGRAM statement.

DATA=*SAS-data-set*
specifies a SAS data set containing the x and y coordinate variables and the VAR statement variables.

OUTDISTANCE=*SAS-data-set*
OUTDIST=*SAS-data-set*
OUTD=*SAS-data-set*
specifies a SAS data set in which to store summary distance information. This data set contains a count of all pairs of data points within a given distance interval. The number of distance intervals is controlled by the NHCLASSES= option in the COM-

PUTE statement. The OUTDISTANCE= data set is useful for plotting modified histograms of the count data for determining appropriate lag distances. See the section "OUTDIST=*SAS-data-set*" on page 3670 for details.

OUTPAIR=*SAS-data-set*
OUTP=*SAS-data-set*

specifies a SAS data set in which to store distance and angle information for each pair of points in the DATA= data set. This option should be used with caution when the DATA= data set is large. If n denotes the number of observations in the DATA= data set, the OUTPAIR= data set contains $\frac{n(n-1)}{2}$ observations unless you restrict it with the OUTPDISTANCE= option in the COMPUTE statement. The OUTPDISTANCE= option in the COMPUTE statement excludes pairs of points when the distance between the pairs exceeds the OUTPDISTANCE= value. See the section "OUTPAIR=*SAS-data-set*" on page 3673 for details.

OUTVAR=*SAS-data-set*
OUTVR=*SAS-data-set*

specifies a SAS data set in which to store the continuity measures. See the section "OUTVAR=*SAS-data-set*" on page 3669 for details.

COMPUTE Statement

COMPUTE *computation-options* **;**

The COMPUTE statement provides a number of options that control the computation of the semivariogram, the robust semivariogram, and the covariance.

ANGLETOLERANCE=*angle tolerance*
ANGLETOL=*angle tolerance*
ATOL=*angle tolerance*

specifies the tolerance, in degrees, around the angles determined by the NDIRECTIONS= specification. The default is $\frac{180^o}{2\times n_d}$, where n_d is the NDIRECTIONS= specification.

See the section "Theoretical and Computational Details of the Semivariogram" on page 3664 for more detailed information.

BANDWIDTH=*bandwidth distance*
BANDW=*bandwidth distance*

specifies the bandwidth, or perpendicular distance cutoff for determining the angle class for a given pair of points. The distance classes define a series of cylindrically shaped areas, while the angle classes radially cut these cylindrically shaped areas. For a given angle class $(\theta_1 - \delta\theta_1, \theta_1 + \delta\theta_1)$, as you proceed out radially, the area encompassed by this angle class becomes larger. The BANDWIDTH= option restricts this area by excluding all points with a perpendicular distance from the line $\theta = \theta_1$ that is greater than the BANDWIDTH= value.

If you do not specify the BANDWIDTH= option, no restriction occurs. See Figure 70.15 on page 3668 for more detailed information.

DEPSILON=*distance value*
DEPS=*distance value*

specifies the distance value for declaring that two distinct points are zero distance apart. Such pairs, if they occur, cause numeric problems. If you specify DEPSILON=ε, then pairs of points P_1 and P_2 for which the distance between them $\mid P_1P_2 \mid < \varepsilon$ are excluded from the continuity measure calculations. The default value of the DEPSILON= option is 100 times machine epsilon; this product is approximately 1E-10 on most computers.

LAGDISTANCE=*distance unit*
LAGDIST=*distance unit*
LAGD=*distance unit*

specifies the basic distance unit defining the lags. For example, a specification of LAGDISTANCE=x results in lag distance classes that are multiples of x. For a given pair of points P_1 and P_2, the distance between them, denoted $\mid P_1P_2 \mid$, is calculated. If $\mid P_1P_2 \mid = x$, then this pair is in the first lag class. If $\mid P_1P_2 \mid = 2x$, then this pair is in the second lag class, and so on.

For irregularly spaced data, the pairwise distances are unlikely to fall exactly on multiples of the LAGDISTANCE= value. A distance tolerance of δx is used to accommodate a spread of distances around multiples of x (the LAGTOLERANCE= option specifies the distance tolerance). For example, if $\mid P_1P_2 \mid$ is within $x \pm \delta x$, you would place this pair in the first lag class; if $\mid P_1P_2 \mid$ is within $2x \pm \delta x$, you would place this pair in the second lag class, and so on.

You can determine the candidate values for the LAGDISTANCE= option by plotting or displaying the OUTDISTANCE= data set.

A LAGDISTANCE= value is required unless you specify the NOVARIOGRAM option.

See the section "Theoretical and Computational Details of the Semivariogram" on page 3664 for more details.

LAGTOLERANCE=*tolerance number*
LAGTOL=*tolerance number*
LAGT=*tolerance number*

specifies the tolerance around the LAGDISTANCE= value for grouping distance pairs into lag classes. See the preceding description of the LAGDISTANCE= option for information on the use of the LAGTOLERANCE= option, and see the section "Theoretical and Computational Details of the Semivariogram" on page 3664 for more details.

If you do not specify the LAGTOLERANCE= option, a default value of $(1/2)$ times the LAGDISTANCE= value is used.

MAXLAGS=*number of lags*
MAXLAG=*number of lags*
MAXL=*number of lags*

specifies the maximum number of lag classes used in constructing the continuity measures. This option excludes any pair of points P_1 and P_2 for which the distance

between them, $\mid P_1P_2 \mid$, exceeds the MAXLAGS= value times the LAGDISTANCE= value.

You can determine candidate values for the MAXLAGS= option by plotting or displaying the OUTDISTANCE= data set.

A MAXLAGS= value is required unless you specify the NOVARIOGRAM option.

NDIRECTIONS=*number of directions*
NDIR=*number of directions*
ND=*number of directions*

specifies the number of angle classes to use in computing the continuity measures. This option is useful when there is potential anisotropy in the spatial continuity measures. Anisotropy occurs when the spatial continuity or dependence between a pair of points depends on the orientation or angle between the pair. Isotropy is the absence of this effect: the spatial continuity or dependence between a pair of points depends only on the distance between the points, not the angle.

The angle classes formed from the NDIRECTIONS= option start from N–S and proceed clockwise. For example, NDIRECTIONS=3 produces three angle classes. In terms of compass points, these classes are centered at 0^o (or its reciprocal 180^o), 60^o (or its reciprocal 240^o), and 120^o (or its reciprocal 300^o). For irregularly spaced data, the angles between pairs are unlikely to fall exactly in these directions, so an angle tolerance of $\delta\theta$ is used (the ANGLETOLERANCE= option specifies the angle tolerance). If NDIRECTIONS=n_d, the base angle is $\theta = \frac{180^o}{n_d}$, and the angle classes are

$$(k\theta - \delta\theta, k\theta + \delta\theta) \qquad k = 0, \ldots, n_d - 1$$

If you do not specify the NDIRECTIONS= option, no angles are formed, and the spatial continuity measures are assumed to be isotropic.

The NDIRECTIONS= option is useful for exploring possible anisotropy. The DIRECTIONS statement, described in the "DIRECTIONS Statement" section on page 3662, provides greater control over the angle classes. See the section "Theoretical and Computational Details of the Semivariogram" on page 3664 for more detailed information.

NHCLASSES=*number of histogram classes*
NHCLASS=*number of histogram classes*
NHC=*number of histogram classes*

specifies the number of distance or histogram classes to write to the OUTDISTANCE= data set. The actual number of classes is one more than the NHCLASSES= value since a special lag 0 class is also computed. See the OUTDISTANCE= option on page 3657 and the section "OUTDIST=*SAS-data-set* " on page 3670 for details.

The default value of the NHCLASSES= option is 10. This option is ignored if you do not specify an OUTDISTANCE= data set.

NOVARIOGRAM
prevents the computation of the continuity measures. This option is useful for preliminary analysis when you require only the OUTDISTANCE= or OUTPAIR= data sets.

OUTPDISTANCE=*distance limit*
OUTPDIST=*distance limit*
OUTPD=*distance limit*
specifies the cutoff distance for writing observations to the OUTPAIR= data set. If you specify OUTPDISTANCE=d_{max}, the distance $\mid P_1P_2 \mid$ between each pair of points P_1 and P_2 is checked against d_{max}. If $\mid P_1P_2 \mid > d_{max}$, the observation for this pair is not written to the OUTPAIR= data set. If you do not specify the OUTPDISTANCE= option, all distinct pairs are written. This option is ignored if you do not specify an OUTPAIR= data set.

ROBUST
requests that a robust version of the semivariogram be calculated in addition to the regular semivariogram and covariance.

COORDINATES Statement

COORDINATES *coordinate-variables* **;**

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

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

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 in the DATA= data set.

DIRECTIONS Statement

DIRECTIONS *directions-list* **;**

The DIRECTIONS statement enables detailed control for defining the angle classes. It is a list of angles, separated by commas, with optional angle tolerances and bandwidths within parentheses following the angle. One or more angles are required. If you do not specify the optional angle tolerance, the default value of 45^o is used. If you do not specify the optional bandwidth, no bandwidth is checked.

For example, suppose you want to compute three separate semivariograms at angles $\theta_1 = 0^o, \theta_2 = 60^o$, and $\theta_3 = 120^o$, with corresponding angle tolerance $\delta\theta_1 = 22.5^o, \delta\theta_2 = 12.5^o$, and $\delta\theta_3 = 22.5^o$, with bandwidths 50 and 40 distance units on the first two angle classes and no bandwidth check on the last angle class.

The appropriate DIRECTIONS statement is

```
directions 0.0(22.5,50), 60.0(12.5,40),120,235(22.5);
```

VAR Statement

VAR *analysis-variables-list* **;**

Use the VAR statement to specify the analysis variables. You can specify only numeric variables. If you do not specify a VAR statement, all numeric variables in the DATA= data set that are not in the COORDINATES statement are used.

Details

Theoretical Semivariogram Models

The VARIOGRAM procedure computes the sample, or experimental semivariogram. Prediction of the spatial process at unsampled locations by techniques such as ordinary kriging requires a theoretical semivariogram or covariance.

It is necessary, then, to decide on a theoretical variogram based on the sample variogram. While there are methods of fitting semivariogram models, such as least squares, maximum likelihood, and robust methods (Cressie 1993, section 2.6), these techniques are not appropriate for data sets resulting in a small number of variogram points. Instead, a visual fit of the variogram points to a few standard models is often satisfactory. Even when there are sufficient variogram points, a visual check against a fitted theoretical model is appropriate (Hohn 1988, p. 25ff).

In some cases, a plot of the experimental semivariogram suggests that a single theoretical model is inadequate. Nested models, anisotropic models, and the nugget effect increase the scope of theoretical models available. All of these concepts are discussed in this section. The specification of the final theoretical model is provided by the syntax of PROC KRIGE2D.

Note the general flow of investigation. After a suitable choice is made of the LAGDIST= and MAXLAGS= options and, possibly, the NDIR= option (or a DIRECTIONS statement), the experimental semivariogram is computed. Potential theoretical models, possibly incorporating nesting, anisotropy, and the nugget effect, are computed by a DATA step, then they are plotted against the experimental semivariogram and evaluated. A suitable theoretical model is thus found visually, and the specification of the model is used in PROC KRIGE2D. This flow is illustrated in Figure 70.10; also see the "Getting Started" section on page 3644 for an illustration in a simple case.

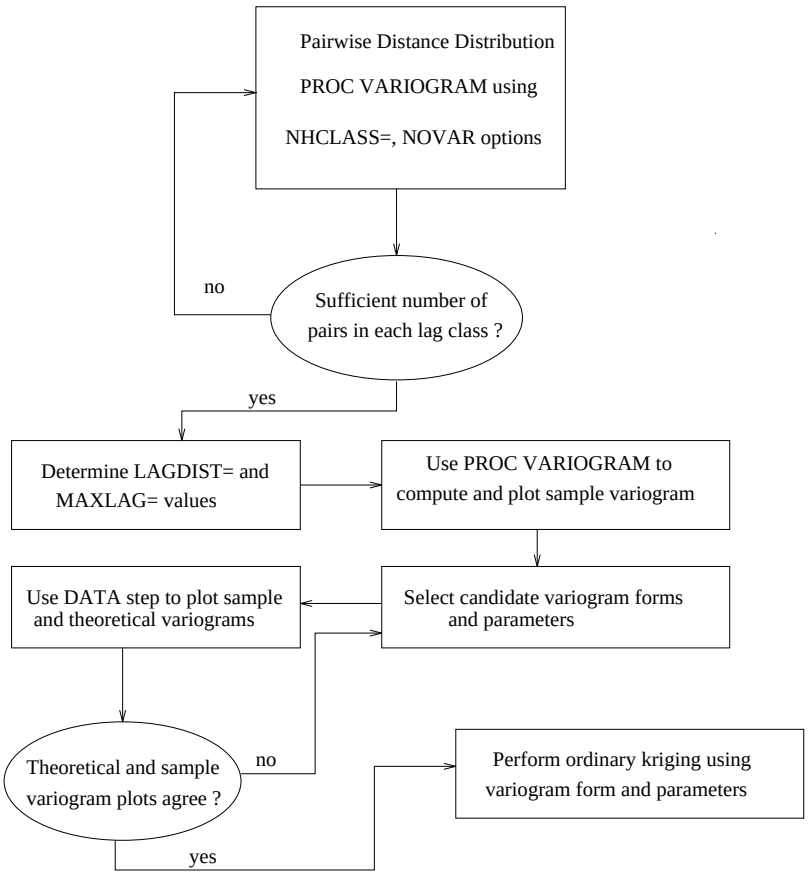


Figure 70.10. Flowchart for Variogram Selection

Theoretical and Computational Details of the Semivariogram

The basic starting point in computing the semivariogram is the enumeration of pairs of points for the spatial data. Figure 70.11 shows a spatial domain in which a set of measurements are made at the indicated locations. Two points P_1 and P_2, with coordinates $(x_1, y_1), (x_2, y_2)$, are selected for illustration. A vector, or directed line segment, is drawn between these points. This pair is then categorized first by orientation of this directed line segment and then by its length. That is, the pair P_1P_2 is placed into an angle and distance class.

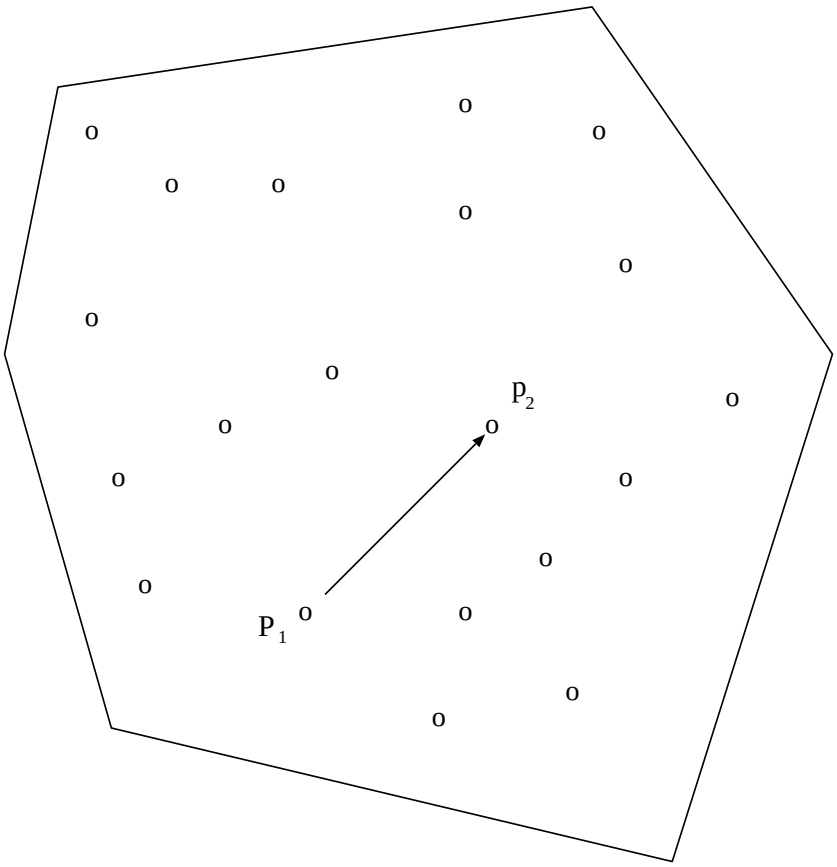


Figure 70.11. Selection of Points P_1 and P_2 in Spatial Domain

Angle Classification

Suppose you specify NDIR=3 in the COMPUTE statement in PROC VARIOGRAM. This results in three angle classes defined by midpoint angles between 0^o and 180^o: $0^o \pm \delta\theta$, $60^o \pm \delta\theta$, and $120^o \pm \delta\theta$, where $\delta\theta$ is the angle tolerance. If you do not specify an angle tolerance using the ATOL= option in the COMPUTE statement, the following default value is used.

$$\delta\theta = \frac{180^o}{2 \times NDIR}$$

For three classes, $\delta\theta = 30^o$. When the example directed line segment P_1P_2 is superimposed on the coordinate system showing the angle classes, its angle, measured clockwise from north, is approximately 45^o. In particular, it falls within $[60^o - \delta\theta, 60^o + \delta\theta) = [30^o, 90^o)$, the second angle class. See Figure 70.12.

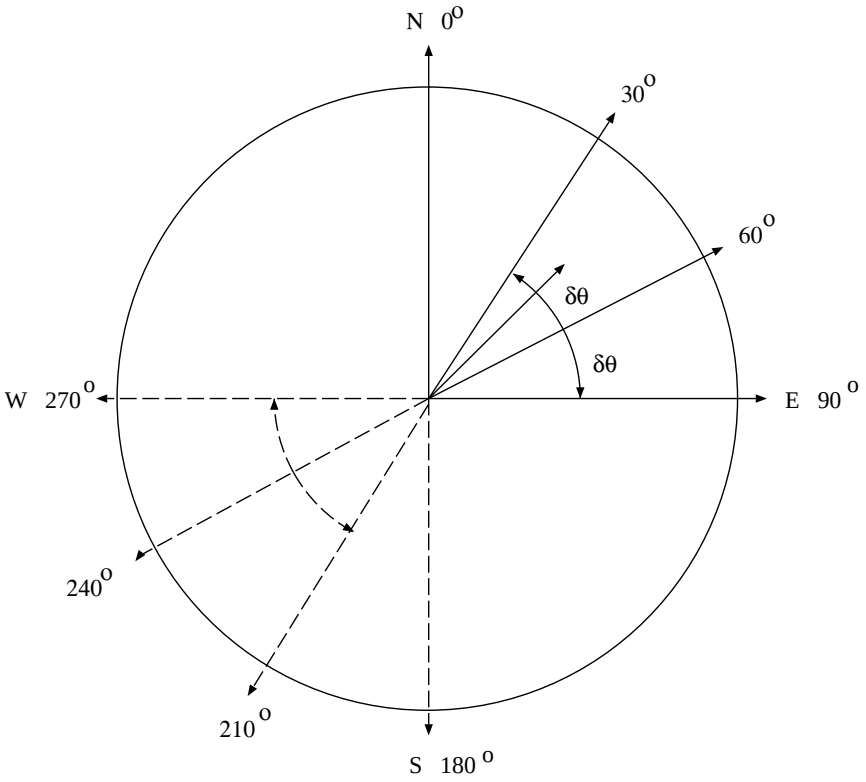


Figure 70.12. Selected Pair P_1P_2 Falls within the Second Angle Class

Note that if the designated points P_1 and P_2 are labeled in the opposite order, the orientation is in a reciprocal direction, that is, approximately 225^o for the point pair instead of approximately 45^o. This does not affect angle class selection; the angle classes $[60^o - \delta\theta, 60^o + \delta\theta)$ and $[240^o - \delta\theta, 240^o + \delta\theta)$ are the same.

If you specify an angle tolerance less than the default, for example, $ATOL = 15^o$, some point pairs might be excluded. For example, the selected point pair P_1P_2 in Figure 70.12, while closest to the 60^o axis, might lie outside $[60 - \delta\theta, 60 + \delta\theta) = [45^o, 75^o)$. In this case, the point pair P_1P_2 would be excluded from the variogram computation.

On the other hand, you can specify an angle tolerance *greater* than the default. This can result in a point pair being counted in more than one angle class. This has a smoothing effect on the variogram and is useful when there is a small amount of data available.

An alternative way to specify angle classes and angle tolerances is with the DIRECTIONS statement. The DIRECTIONS statement is useful when angle classes are not equally spaced. When you specify the DIRECTIONS statement, you should also

specify the angle tolerance. The default value of the angle tolerance is 45^o when a DIRECTIONS statement is used instead of the NDIRECTIONS= option in the COMPUTE statement. This may not be appropriate for a particular set of angle classes. See the "DIRECTIONS Statement" section on page 3662 for more details on the DIRECTIONS statement.

Distance Classification

Next, the distance class for the point pair P_1P_2 is determined. The directed line segment P_1P_2 is superimposed on the coordinate system showing the distance or lag classes. These classes are determined by the LAGD= specification in the COMPUTE statement. Denoting the length of the line segment by $\mid P_1P_2 \mid$ and the LAGD value by Δ, the lag class L is determined by

$$L(P_1P_2) = \left\lfloor \frac{\mid P_1P_2 \mid +.5}{\Delta} \right\rfloor$$

where $\lfloor x \rfloor$ denotes the largest integer $\leq x$.

When the directed line segment P_1P_2 is superimposed on the coordinate system showing the distance classes, it is seen to fall in the first lag class; see Figure 70.13 for an illustration for $\Delta = 1$.

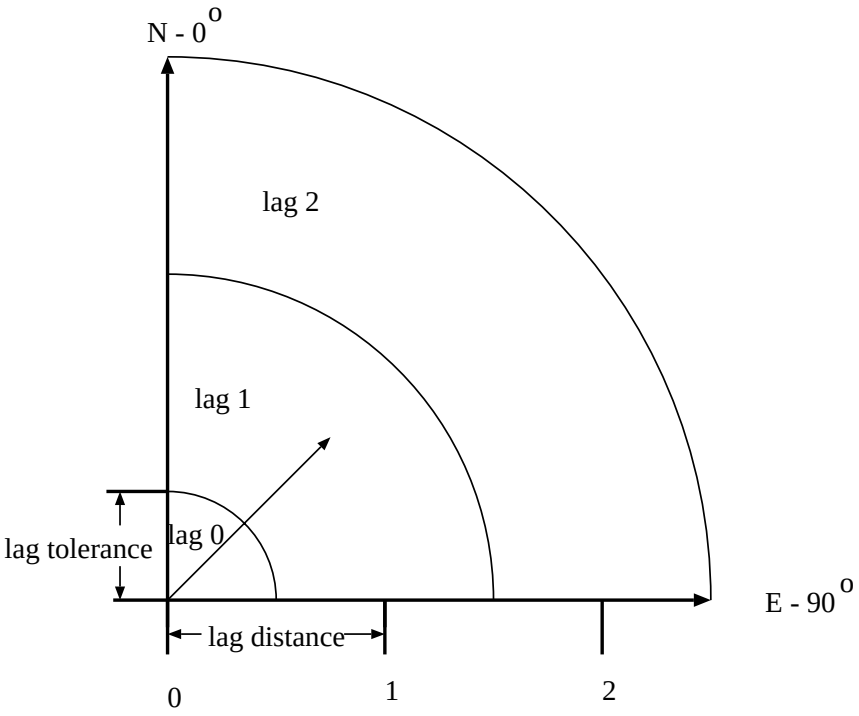


Figure 70.13. Selected Pair P_1P_2 Falls within the First Lag Class

Because pairwise distances are positive, lag class zero is smaller than lag classes $1, \cdots, MAXLAG - 1$. For example, if you specify LAGD=1.0 and MAXLAG=10,

and you do not specify a LAGTOL= value in the COMPUTE statement in PROC VARIOGRAM, the ten lag classes generated by the preceding equation are

$$[0, .5), [.5, 1.5), [1.5, 2.5), \cdots, [8.5, 9.5)$$

This is because the default lag tolerance is one-half the LAGD= value, resulting in no gaps between the distance class intervals. This is shown in Figure 70.14.

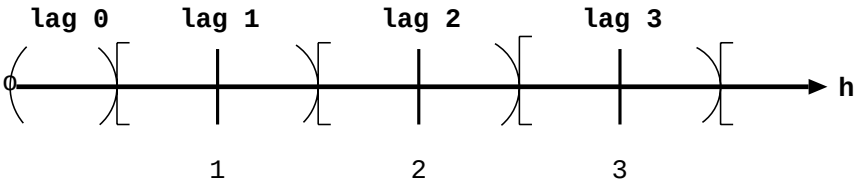


Figure 70.14. Lag Distance Axis Showing Lag Classes

On the other hand, if you do specify a distance tolerance with the DTOL= option in the COMPUTE statement, a further check is performed to see if the point pair falls within this tolerance of the nearest lag. In the preceding example, if you specify LAGD=1.0 and MAXLAG=10 (as before) and also specify LAGTOL=0.25, the intervals become

$$[0, 0.25), [0.75, 1.25), [1.75, 2.25), \cdots, [8.75, 9.25)$$

Note that this specification results in gaps in the lag classes; a point pair P_1P_2 might fall, for example, in the interval

$$\mid P_1P_2 \mid \in [1.25, 1.75)$$

and hence be excluded from the semivariogram calculation. The maximum LAGTOL= value allowed is half the LAGD= value; no overlap of the distance classes is allowed.

Bandwidth Restriction

Because the areal segments generated from the angle and distance classes increase in area as the lag distance increases, it is sometimes desirable to restrict this area (Duetsch and Journel 1992, p. 45). If you specify the BANDW= option in the COMPUTE statement, the lateral, or perpendicular, distance from the axis defining the angle classes is fixed.

For example, suppose two points P_3, P_4 are picked from the domain in Figure 70.11 and are superimposed on the grid defining distance and angle classes, as shown in Figure 70.15.

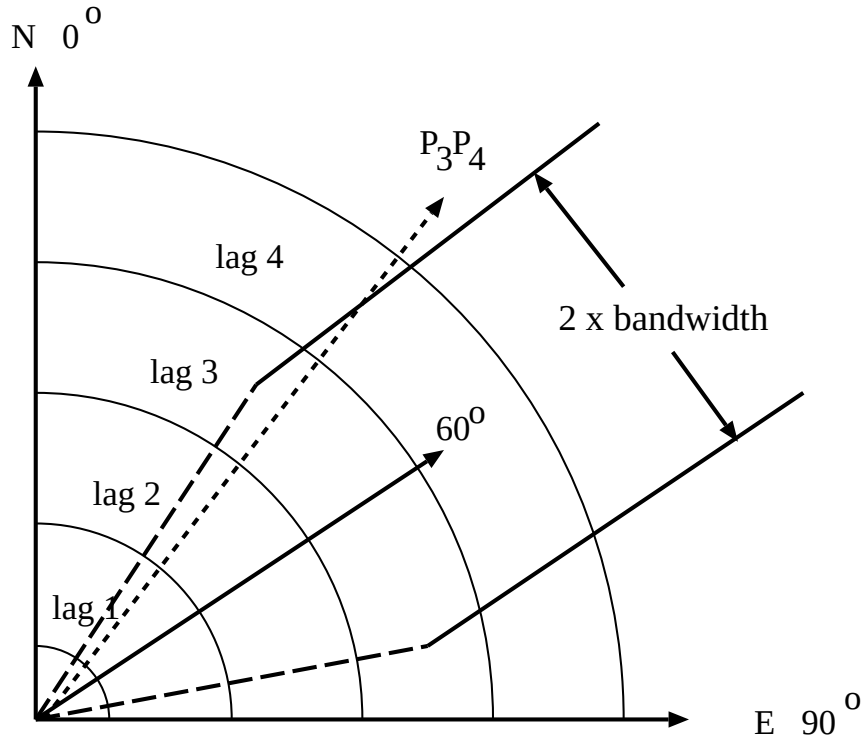


Figure 70.15. Selected Pair P_3P_4 Falls Outside Bandwidth Limit

The endpoint of vector P_3P_4 falls within the angle class around 60^o and the 5th lag class; however, it falls outside the restricted area defined by the bandwidth. Hence, it is excluded from the semivariogram calculation.

Finally, a pair P_iP_j that falls in a lag class larger than the value of the MAXLAG= option is excluded from the semivariogram calculation.

From this description, it is clear that the number of pairs within each angle/distance class is strongly affected by the angle and lag tolerances. Since it is desirable to have the maximum number of point pairs within each class, the angle tolerance and the distance tolerance should usually be the default values.

Semivariogram Computation

With the classification of a point pair P_iP_j into an angle/distance class, as shown in the preceding section, the semivariogram computation proceeds as follows.

Denote all pairs P_iP_j belonging to angle class $[\theta_k - \delta\theta_k, \theta_k + \delta\theta_k)$ and distance class $L = L(P_iP_j)$ by $N(\theta_k, L)$. For example, in the preceding illustration, P_1P_2 belongs to $N(60^o, 1)$.

Let $\mid N(\theta_k, L) \mid$ denote the *number* of such pairs. Let V_i, V_j be the measured values at points P_i, P_j. The component of the standard (or method of moments) semivariogram corresponding to angle/distance class $N(\theta_k, L)$ is given by

$$2\gamma(h_k) = \frac{1}{\mid N(\theta_k, L) \mid} \sum_{P_i P_j \in N(\theta_k, L)} (V_i - V_j)^2$$

where h_k is the average distance in class $N(\theta_k, L)$; that is,

$$h_k = \frac{1}{\mid N(\theta_k, L) \mid} \sum_{P_i P_j \in N(\theta_k, L)} \mid P_i P_j \mid$$

The robust version of the semivariogram, as suggested by Cressie (1993), is given by

$$2\bar{\gamma}(h_k) = \frac{\Psi^4(h_k)}{0.457 + 0.494/N(\theta_k, L)}$$

where

$$\Psi(h_k) = \frac{1}{N(\theta_k, L)} \sum_{P_i P_j \in N(\theta_k, L)} (V_i - V_j)^{\frac{1}{2}}$$

This robust version of the semivariogram is computed when you specify the ROBUST option in the COMPUTE statement in PROC VARIOGRAM.

PROC VARIOGRAM computes and writes to the OUTVAR= data set the quantities $h_k, \theta_k, L, N(\theta_k, L), \gamma(h)$, and $\bar{\gamma}(h)$.

Output Data Sets

The VARIOGRAM procedure produces three data sets: the OUTVAR=*SAS-data-set*, the OUTPAIR=*SAS-data-set*, and the OUTDIST=*SAS-data-set*. These data sets are described in the following sections.

OUTVAR=*SAS-data-set*

The OUTVAR= data set contains the standard and robust versions of the sample semivariogram, the covariance, and other information at each lag class.

The details of the computation of the variogram, the robust variogram, and the covariance is described in the section "Theoretical and Computational Details of the Semivariogram" on page 3664.

The OUTVAR= data set contains the following variables:

- ANGLE, which is the angle class value (clockwise from N–S)
- ATOL, which is the angle tolerance for the lag/angle class
- AVERAGE, which is the average variable value for the lag/angle class
- BANDW, which is the band width for the lag/angle class
- COUNT, which is the number of pairs in the lag/angle class

- COVAR, which is the covariance value for the lag/angle class
- DISTANCE, which is the average lag distance for the lag/angle class
- LAG, which is lag class value (in LAGDISTANCE= units)
- RVARIO, which is the sample robust variogram value for the lag/angle class
- VARIOG, which is the sample variogram value for the lag/angle class
- VARNAME, which is the name of the current VAR= variable

The bandwidth variable, BANDW, is not included in the data set if no bandwidth specification is given in the COMPUTE statement or in a DIRECTIONS statement.

OUTDIST=*SAS-data-set*

The OUTDIST= data set contains counts for a modified histogram showing the distribution of pairwise distances. The purpose of this data set is to enable you to make choices for the value of the LAGDISTANCE= option in the COMPUTE statement in subsequent runs of PROC VARIOGRAM.

For plotting and estimation purposes, it is desirable to have as many points as possible for a variogram plot. However, a rule of thumb used in computing sample semivariograms is to use at least 30 points in each interval whenever possible. Hence, there is a lower limit to the value of the LAGDISTANCE= option.

Since the distribution of pairwise distances is seldom known in advance, the information contained in the OUTDIST= data set enables you to choose, in an iterative fashion, a value for the LAGDISTANCE= parameter. The value you choose is a compromise between the number of pairs making up each variogram point and the number of variogram points.

In some cases, the pattern of measured points may result in some lag or distance classes having a small number of pairs, while the remaining classes have a large number of pairs. By adjusting the value of the LAGDISTANCE= option to honor the rule of thumb (at least 30 pairs), you are "wasting" pairs in the other distance classes.

One strategy for solving this problem is to use less than 30 pairs for these distance classes. Then, either delete the corresponding variogram points or use them and accept the increased uncertainty. Unfortunately, the deficient distance classes are usually those close to the origin ($h = 0$). This is the crucial portion of the experimental variogram curve for determining the form of the theoretical variogram and for detecting the presence of a nugget effect.

Another alternative is to force distance classes to contain approximately the same number of pairs. This results in distance classes of unequal widths.

While PROC VARIOGRAM does not produce such distance classes directly, the OUTPAIR= data set, described in the section "OUTPAIR=*SAS-data-set*" on page 3673, contains information on all distinct pairs of points. You can use this data set, along with the RANK procedure, to produce experimental variogram-based equal numbers of pairs in each distance class.

To request an OUTDIST= data set, you specify the OUTDIST= data set in the PROC VARIOGRAM statement and the NOVARIOGRAM option in the COMPUTE state-

ment. The NOVARIOGRAM option prevents any variogram or covariance computation from being performed.

Computation of the Distribution Distance Classes

The simplest way of determining the distribution of pairwise distances is to determine the maximum distance h_{max} between pairs and divide this distance by some number N of intervals to produce distance classes of length $\delta = \frac{h_{max}}{N}$. The distance between each pair of points P_1, P_2, denoted $\mid P_1P_2 \mid$, is computed, and the pair P_1P_2 is counted in the kth distance class if $\mid P_1P_2 \mid \in [(k-1)\delta, k\delta)$ for $k = 1, \cdots, N$.

The actual computation is a slight variation of this. A bound, rather than the actual maximum distance, is computed. This bound is the length of the diagonal of a bounding rectangle for the data points. This bounding rectangle is found by using the maximum and minimum x and y coordinates, $x_{max}, x_{min}, y_{max}, y_{min}$, and forming the rectangle determined by the points

$(x_{max}, y_{max}), (x_{max}, y_{min}), (x_{min}, y_{min}), (x_{min}, y_{max})$

See Figure 70.16 for an illustration of the bounding rectangle.

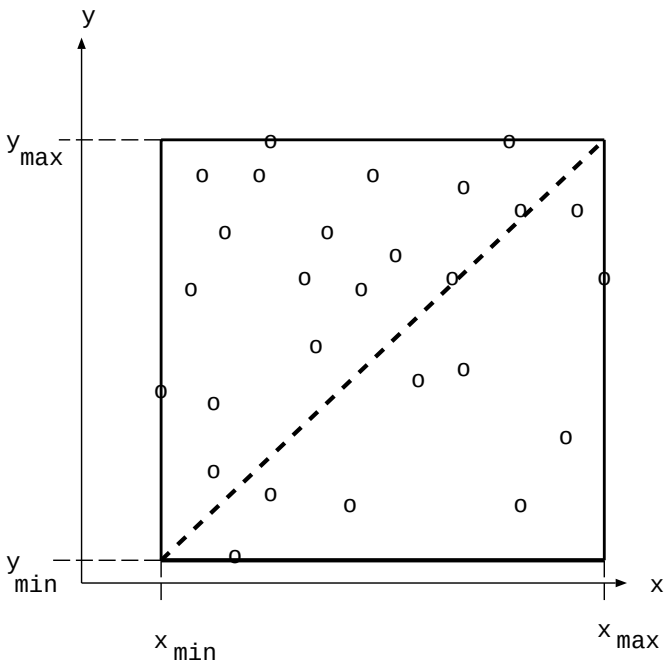


Figure 70.16. Bounding Rectangle to Determine Maximum Pairwise Distance

The pairwise distance bound, denoted by h_b, is given by

$$h_b^2 = (x_{max} - x_{min})^2 + (y_{max} - y_{min})^2$$

Using h_b, the interval $(0, h_b]$ is divided into $N+1$ subintervals, where N is the value of the NHCLASSES= option specified in the COMPUTE statement, or $N = 10$ if the NHCLASSES= option is not specified. The basic distance unit is $h_0 = \frac{h_b}{N}$; the distance intervals are centered on $h_0, 2h_0, \cdots, Nh_0$, with a distance tolerance of $\pm\frac{h_0}{2}$. The extra subinterval is $(0, h_0/2)$, corresponding to the 0th lag. It is half the length of the remaining subintervals, and it often contains the smallest number of pairs.

This method of partitioning the interval $(0, h_b]$ is identical to what is done when you actually compute the sample variogram.

The lag classes corresponding to h_0=1 are shown in Figure 70.17.

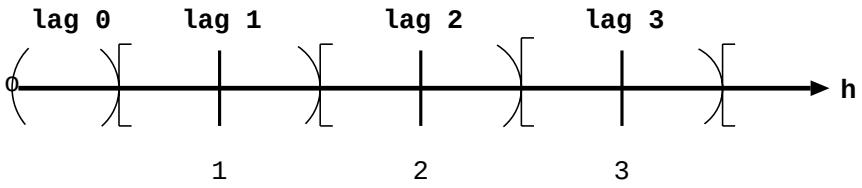


Figure 70.17. Lag Classes Corresponding to $h_0 = 1$

By increasing or decreasing the value of the NHCLASSES= option, you can adjust the lag or distance class with the smallest count so that this count is around 30 or some other value that you judge appropriate.

Once you determine an appropriate value for the NHCLASSES= option, you can use the width of the lag classes as a candidate value for the LAGDIST= option in the COMPUTE statement. The width of the lag classes is determined by the upper bound (UB) and lower bound (LB) variables.

For example, read the observation from the OUTDIST= data set corresponding to lag 1 and compute the quantity UB-LB. Use this value for the LAGDIST= option in the COMPUTE statement.

Note: Do not use the 0th lag class; it is half the length of the other intervals. Use lag 1 instead.

Variables in the OUTDIST= data set

The following variables are written to the OUTDIST= data set:

- COUNT, which is the number of pairs falling into this lag class
- LAG, which is the lag class value
- LB, which is the lower bound of the lag class interval
- UB, which is the upper bound of the lag class interval
- PER, which is the percent of all pairs falling in this lag class
- VARNAME, which is the name of the current VAR= variable

OUTPAIR=*SAS-data-set*

The OUTPAIR= data set contains one observation for each distinct pair of points P_1, P_2 in the original data set, unless you specify the OUTPDISTANCE= option in the COMPUTE statement.

If you specify OUTPDISTANCE=D_{max} in the COMPUTE statement, all pairs P_1, P_2 in the original data set that satisfy the relation $\mid P_1P_2 \mid \leq D_{max}$ are written to the OUTPAIR= data set.

Note that the OUTPAIR= data set can be very large even for a moderately sized DATA= data set. For example, if the DATA= data set has NOBS=500, the OUTPAIR= data set has NOBS$(\text{NOBS} - 1)/2$ =124,750 if no OUTPDISTANCE= restriction is given in the COMPUTE statement.

The OUTPAIR= data set contains information on the distance and orientation for each point pair, and you can use it for specialized continuity measure calculations.

The OUTPAIR= data set contains the following variables:

- AC, which is the angle class value
- COS, which is the cosine of the angle between pairs
- DC, which is the distance (lag) class
- DISTANCE, which is the distance between pairs
- V1, which is the variable value for the first point in the pair
- V2, which is the variable value for the second point in the pair
- VARNAME, which is the variable name for the current VAR variable
- X1, which is the x coordinate of the first point in the pair
- X2, which is the x coordinate of the second point in the pair
- Y1, which is the y coordinate of the first point in the pair
- Y2, which is the y coordinate of the second point in the pair

Computational Resources

The computations of the VARIOGRAM procedure are basically binning: for each pair of observations in the input data set, a distance and angle class is determined and recorded. Let N_d denote the number of distance classes, N_a denote the number of angle classes, and N_v denote the number of VAR variables. The memory requirements for these operations are proportional to $N_d \times N_a \times N_v$. This is typically small.

The CPU time required for the computations is proportional to the number of pairs of observations, or to $N^2 \times N_v$, where N is the number of observations in the input data set.

Example

Example 70.1. A Box Plot of the Square Root Difference Cloud

The Gaussian form chosen for the variogram in the "Getting Started" section on page 3644 is based on the consideration of the plots of the sample variogram. For the coal thickness data, the Gaussian form appears to be a reasonable choice.

It can often happen, however, that a plot of the sample variogram shows so much scatter that no particular form is evident. The cause of this scatter can be one or more outliers in the pairwise differences of the measured quantities.

A method of identifying potential outliers is discussed in Cressie (1993, section 2.2.2). This example illustrates how to use the OUTPAIR= data set from PROC VARIOGRAM to produce a square root difference cloud, which is useful in detecting outliers.

For the spatial process $Z(s), s \in \mathcal{R}^2$, the square root difference cloud for a particular direction $\mathbf{e}$ is given by

$$\mid Z(s_i + he) - Z(s_i) \mid^{\frac{1}{2}}$$

for a given lag distance h. In the actual computation, all pairs of points P_1, P_2 within a distance tolerance around h and an angle tolerance around the direction $\mathbf{e}$ are used. This generates a number of point pairs for each lag class h. The spread of these values gives an indication of outliers.

Following the example in the "Getting Started" section on page 3644, this example uses a basic lag distance of 7 units, with a distance tolerance of 3.5, and a direction of N–S, with a 30^o angle tolerance.

First, input the data, then use PROC VARIOGRAM to produce an OUTPAIR= data set. Then use a DATA step to subset this data by choosing pairs within 30^o of N–S. In addition, compute lag class and square root difference variables. Next, summarize the results using the MEANS procedure and present them in a box plot using the SHEWHART procedure. The box plot facilitates the detection of outliers.

You can conclude from this example that there does not appear to be any outliers in the N–S direction for the coal seam thickness data.

```
title 'Square Root Difference Cloud Example';
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 variogram data=thick outp=outp;
coordinates xc=east yc=north;
var thick;
compute novar;
run;

data sqroot;
   set outp;

/*- Include only points +/- 30 degrees of N-S -------*/
   where abs(cos) < .5;

/*- Unit lag of 7, distance tolerance of 3.5  -------*/
   lag_class=int(distance/7 + .5000001);
   sqr_diff=sqrt(abs(v1-v2));
run;

proc sort data=sqroot;
   by lag_class;
run;

proc means data=sqroot noprint n mean std;
   var sqr_diff;
   by lag_class;
   output out=msqrt n=n mean=mean std=std;
run;

title2 'Summary of Results';
proc print data=msqrt;
   id lag_class;
   var n mean std;
run;
```

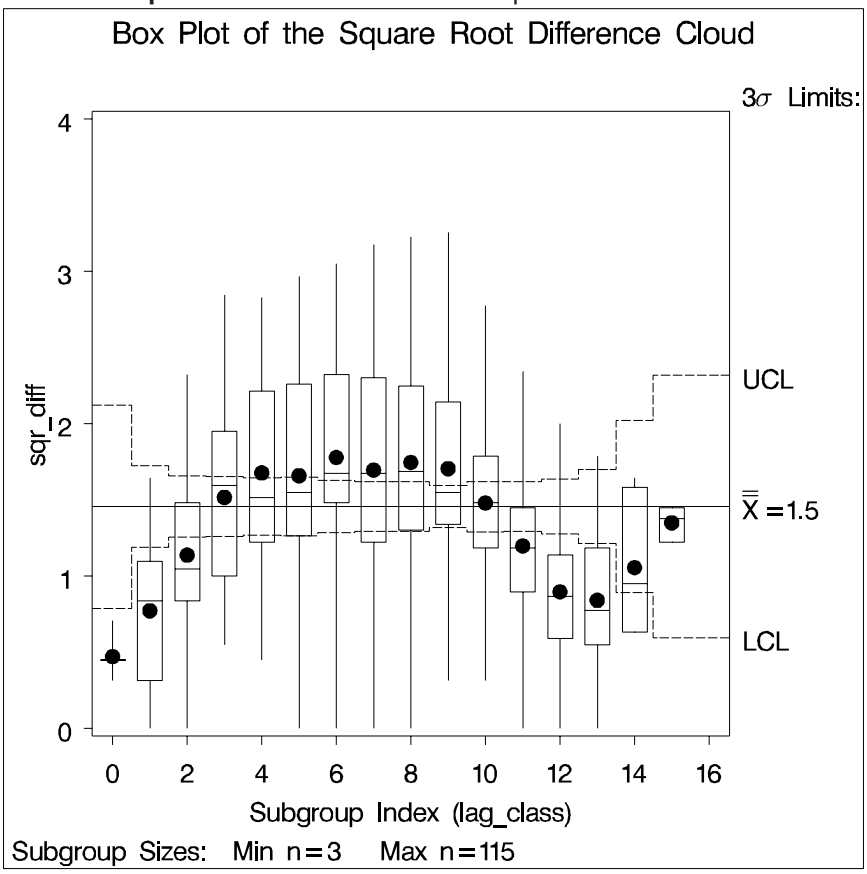
```
title 'Box Plot of the Square Root Difference Cloud';
proc shewhart data=sqroot;
   boxchart sqr_diff*lag_class / cframe=ligr haxis=axis1
                                 vaxis=axis2;
   symbol1 v=dot c=blue height=3.5pct;
   axis1 minor=none;
   axis2 minor=none label=(angle=90 rotate=0);
run;
```

Output 70.1.1. Summary of Results

Square Root Difference Cloud Example
Summary of Results

lag_class	n	mean	std
0	5	0.47300	0.14263
1	31	0.77338	0.41467
2	55	1.13908	0.47604
3	58	1.51768	0.51989
4	63	1.67858	0.60494
5	61	1.66014	0.70687
6	75	1.77999	0.64590
7	85	1.69703	0.75362
8	84	1.74687	0.68785
9	115	1.70635	0.57173
10	82	1.48100	0.48105
11	85	1.19877	0.47121
12	68	0.89765	0.42510
13	38	0.84223	0.44249
14	7	1.05653	0.42548
15	3	1.35076	0.11472

Output 70.1.2. Box Plot of the Square Root Difference Cloud



References

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.

Hohn, M.E. (1988), *Geostatistics and Petroleum Geology*, New York: Van Nostrand Reinhold.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

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

1st printing, October 1999