SAS/STAT® 9.2 User's Guide

The VARIOGRAM Procedure

(Book Excerpt)

SAS® Documentation

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

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

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

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

Chapter 95

The VARIOGRAM Procedure

Contents

Overview: VARIOGRAM Procedure

The VARIOGRAM procedure computes empirical measures of spatial continuity for two-dimensional spatial data. These measures are a function of the distances between the sample data pairs. When the data are free of nonrandom (or systematic) surface trends, the estimated continuity measures are the empirical semivariance and covariance. These measures can be used in subsequent analysis to perform spatial prediction. The procedure plots empirical semivariograms, which can also be saved to an output data set to enable parameter estimation for theoretical semivariogram or covariance models. Both isotropic and anisotropic measures are available.

In addition, PROC VARIOGRAM provides the Moran I and Geary c spatial autocorrelation statistics. The procedure also produces the OUTPAIR= and OUTDISTANCE= data sets that contain information about the semivariogram analysis.

The VARIOGRAM procedure now uses ODS Graphics to create graphs as part of its output. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." For more information about the graphics available in PROC VARIOGRAM, see the section "ODS Graphics" on page 7561.

Introduction to Spatial Prediction

Many activities in science and technology involve measurements of one or more quantities at given spatial locations, with the goal of predicting the measured quantities at unsampled locations. Application areas include reservoir prediction in mining and petroleum exploration, as well as modeling in a broad spectrum of fields (for example, environmental health, environmental pollution, natural resources and energy, hydrology, risk analysis). Often, the unsampled locations are on a regular grid, and the predictions are used to produce surface plots or contour maps.

The preceding tasks fall within the scope of *spatial prediction*, which, in general, is any prediction method that incorporates spatial dependence. The study of these tasks involves naturally occurring uncertainties that cannot be ignored. Stochastic analysis frameworks and methods are used to account for these uncertainties. Hence, the terms *stochastic spatial prediction* and *stochastic modeling* are also used to characterize this type of analysis.

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 characterizing the spatial process. For this purpose, models for the spatial dependence are expressed in terms of the distance between any two locations in the spatial domain of interest. These models take the form of a covariance or semivariance function.

Spatial prediction, then, involves two steps. First, you model the covariance or semivariance of the spatial process. These measures are typically not known in advance. This step involves computing an empirical estimate, as well as determining both the mathematical form and the values of any parameters for a theoretical form of the dependence model. Second, you use this dependence model to solve 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 (that is, calculating and modeling the dependence model), and the KRIGE2D procedure performs the kriging operations to produce the final predictions.

This introduction concludes with a note on terminology. You might commonly encounter the terms *estimation* and *prediction* used interchangeably by experts in different fields; this could be a source of confusion. A precise statistical vernacular uses the term *estimation* to refer to inferences about the value of fixed but unknown parameters, whereas *prediction* concerns inferences about the value of random variables—see, for example, Cressie (1993, p. 106). In light of these definitions, kriging methods are clearly predictive techniques, since they are concerned with making inferences about the value of a spatial random field at observed or unobserved locations. The SAS/STAT suite of procedures for spatial analysis and prediction (VARIOGRAM, KRIGE2D, and SIM2D) follows the statistical vernacular in the use of the terms *estimation* and *prediction*.

Getting Started: VARIOGRAM Procedure

PROC VARIOGRAM uses your data to compute the empirical semivariogram. This computation refers to the steps you take to derive the empirical semivariance from the data, and then to produce the corresponding semivariogram plot.

You can proceed further with the semivariogram analysis if the data are free of systematic trends. In that case, you can use the empirical outcome to determine a theoretical semivariogram model by using automated or visual methods. The model characterizes the type of theoretical semivariance function you will use to describe spatial dependence in your data set.

Some of the following graphical displays are requested by using the ODS GRAPHICS statement. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." For specific information about the graphics available in the VARIOGRAM procedure, see the section "ODS Graphics" on page 7561.

Preliminary Spatial Data Analysis

The following data set simulates measurements of coal seam thickness (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.

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

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 the sampling locations scattered evenly throughout the prediction area. If the locations are not scattered evenly, the prediction error might be unacceptably large where measurements are sparse.

You can run PROC VARIOGRAM in this preliminary analysis to determine potential problems. In the following statements, the NOVARIOGRAM option in the COMPUTE statement specifies that only the descriptive summaries and a plot of the raw data should be produced.

```
ods graphics on;

proc variogram data=thick plots=pairs(thr=30);
   compute novariogram nhc=20;
   coordinates xc=East yc=North;
   var Thick;
run;
```

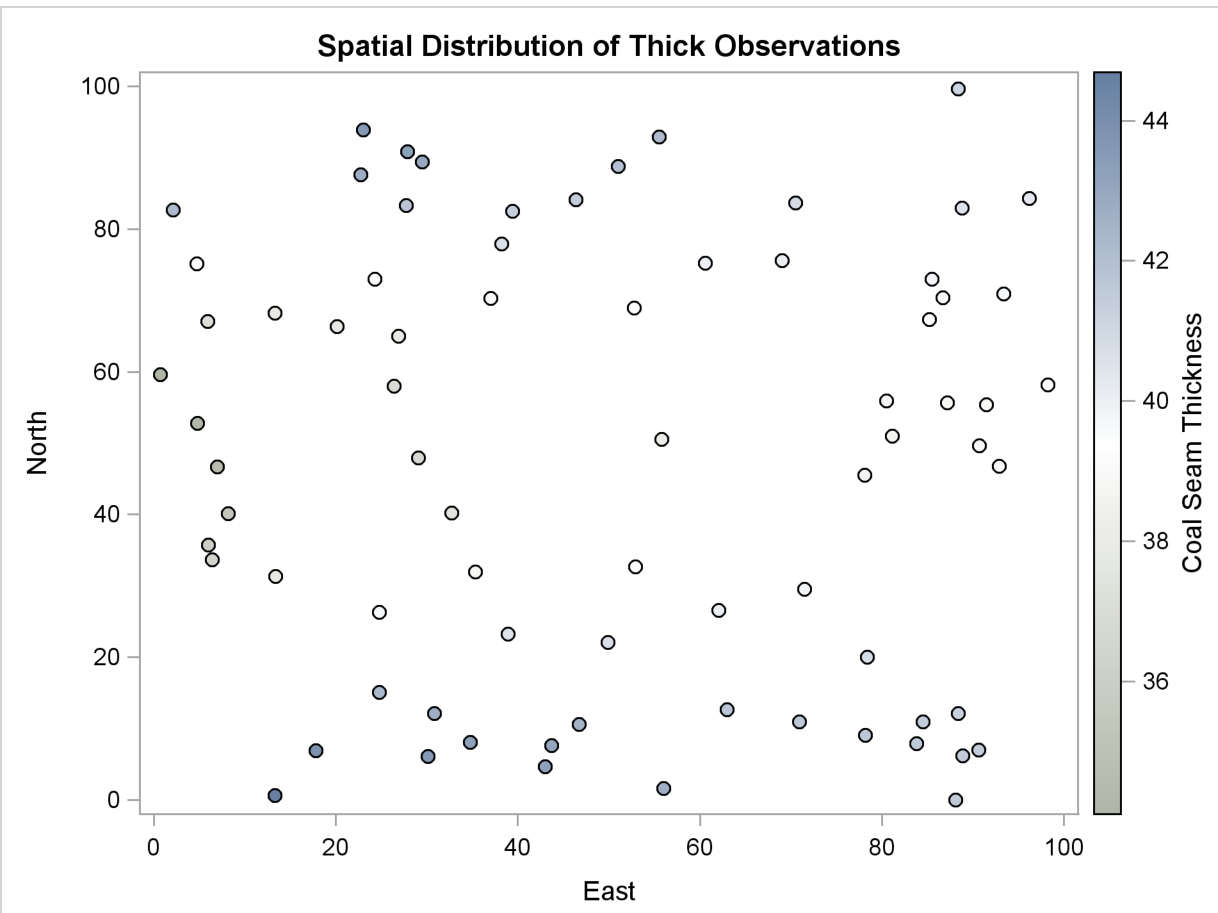
PROC VARIOGRAM produces the table in Figure 95.1 that shows the number of Thick observations read and used. This table provides you with useful information in case you have missing values in the input data.

Figure 95.1 Number of Observations for the thick Data Set

```
                    The VARIOGRAM Procedure
                  Dependent Variable: Thick

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

Then, the scatter plot of the observed data is produced as shown in Figure 95.2. According to the figure, while the locations are not ideally spread around the prediction area, there are not any extended areas lacking measurements. The same graph also provides the values of the measured variable by using colored markers.

Figure 95.2 Scatter Plot of the Observations Spatial Distribution



The following is a crucial step. Any obvious surface trend must be removed before you compute the empirical semivariogram and proceed to estimate a model of spatial dependence (the theoretical semivariogram model). You can observe in Figure 95.2 the small-scale variation typical of spatial data, but a first inspection indicates no obvious major systematic trend.

Assuming, therefore, that the data are free of surface trends, you can work with the original thickness rather than residuals obtained from a trend removal process. The following analysis also assumes that the spatial characterization is independent of the direction of the line connecting any two equidistant pairs of data; this is a property known as isotropy. See "Example 95.2: An Anisotropic Case Study with Surface Trend in the Data" on page 7566 for a more detailed approach to trend analysis and the issue of anisotropy.

Following the previous exploratory analysis, you then need to classify each data pair as a member of a distance interval (lag). PROC VARIOGRAM performs this grouping with two required options for semivariogram computation: the LAGDISTANCE= and MAXLAGS= options. These options are based on your assessment of how to group the data pairs within distance classes.

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 each distance interval is the LAGDISTANCE= value. The meaning of the required MAXLAGS= option is simply the number of intervals you consider. The problem is that given only the scatter plot of the measurement locations, it is not clear what values to give to the LAGDISTANCE= and MAXLAGS= options.

Ideally, you would like a sufficient number of distance classes that capture the extent to which your data are correlated, and that each contain a minimum of data pairs to increase the accuracy in your computations. A rule of thumb used in semivariogram computations is that you should have at least 30 pairs per lag class. This is an empirical arbitrary threshold; see the section "Choosing the Size of Classes" on page 7547 for further details.

In the preliminary analysis, you use the option NHC= in the COMPUTE statement to help you experiment with these numbers and choose values for the LAGDISTANCE= and MAXLAGS= options. Here, in particular, you requested NHC=20 to preview a classification that uses 20 distance classes across your spatial domain. A zero lag class is always considered; therefore the output shows the number of distance classes to be one more than the number you specified.

Based on your selection of the NHC= option, the NOVARIOGRAM option produces a pairwise distances table from your observations shown in Figure 95.3, and the corresponding histogram in Figure 95.4. For illustration purposes, you also specified in the code a threshold of minimum data pairs per distance class in the PAIRS option as THR=30. As a result, a reference line appears in the histogram so that you can visually identify any lag classes with pairs that fall below your specified threshold.

Figure 95.3 Pairwise Distance Intervals Table

Pairwise Distance Intervals

Lag Class	Bounds		Number of Pairs	Percentage of Pairs
0	0.00	3.48	7	0.25%
1	3.48	10.45	81	2.92%
2	10.45	17.42	138	4.97%
3	17.42	24.39	167	6.02%
4	24.39	31.36	204	7.35%
5	31.36	38.33	210	7.57%
6	38.33	45.30	213	7.68%
7	45.30	52.27	253	9.12%
8	52.27	59.24	237	8.54%
9	59.24	66.20	280	10.09%
10	66.20	73.17	252	9.08%
11	73.17	80.14	230	8.29%
12	80.14	87.11	217	7.82%
13	87.11	94.08	154	5.55%
14	94.08	101.05	71	2.56%
15	101.05	108.02	41	1.48%
16	108.02	114.99	14	0.50%
17	114.99	121.96	5	0.18%
18	121.96	128.93	1	0.04%
19	128.93	135.89	0	0.00%
20	135.89	142.86	0	0.00%

The NOVARIOGRAM option also produces a table with useful facts about the pairs and the distances between the most remote data in selected directions, shown in Figure 95.5. In particular, the lag distance value is calculated based on your selection of the NHC= option. The last three table entries report the overall maximum distance among your data pairs, as well as the maximum distances in the main axes directions (that is, the vertical or N–S axis, and the horizontal or E–W axis). This information is also provided in the inset of Figure 95.4. When you specify a threshold in the PAIRS suboption of the PLOTS= option, as in this example, the threshold also appears in the table. Then, the line that follows indicates the highest lag class with the following property: Each one of the distance classes that lie farther away from this lag features a pairs population below the specified threshold.

With the preceding information you can determine appropriate values for the LAGDISTANCE= and MAXLAGS= options in the COMPUTE statement. In particular, the classification that uses 20 distance classes is satisfactory, and you can choose LAGDISTANCE=7 after following the suggestion in Figure 95.5.

Figure 95.4 Distribution of Pairwise Distances

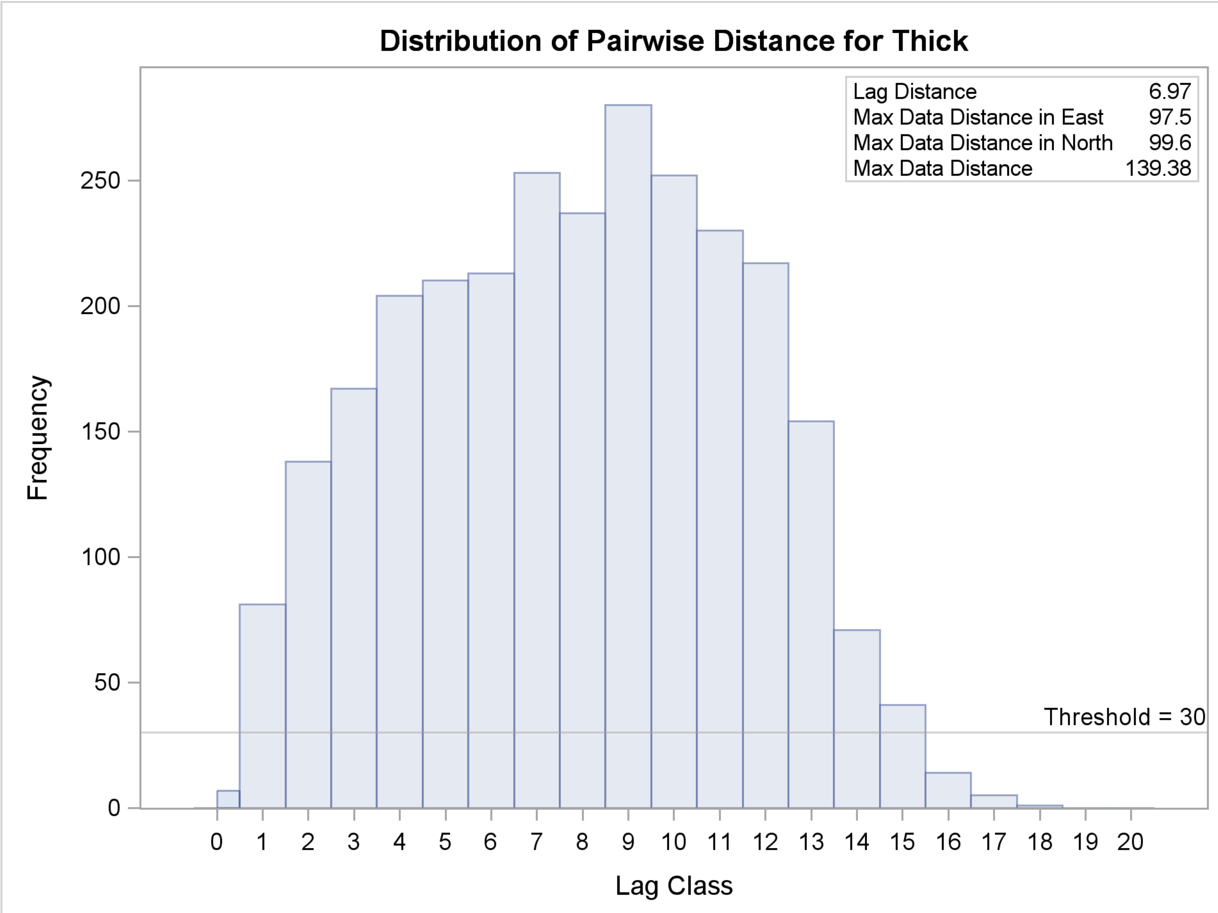


Figure 95.5 Pairs Information Table

Pairs Information

Number of Lags	21
Lag Distance	6.97
Minimum Pairs Threshold	30
Highest Lag With Pairs > Threshold	15
Maximum Data Distance in East	97.50
Maximum Data Distance in North	99.60
Maximum Data Distance	139.38

The MAXLAGS= option needs to be specified based on the spatial extent to which your data are correlated. Unless you know this size, in the present omnidirectional case you can assume the correlation extent to be roughly equal to half the overall maximum distance between data points.

The table in Figure 95.5 suggests that this number corresponds to 139,380 feet, which is most likely on or close to a diagonal direction (that is, the northeast–southwest or northwest–southeast direction). Hence, you can expect the correlation extent in this scale to be around $139.4/2 = 69,700$ feet. Consequently, this is the distance up to which you should consider lag classes for the empirical semivariogram computations. Given your lag size selection, Figure 95.3 indicates that this distance corresponds to about 10 lags; hence you can set MAXLAGS=10.

Overall, for a specific NHC= choice of class count, you can expect your choice of MAXLAGS= to be approximately half the number of the lag classes (see the section "Spatial Extent of the Empirical Semivariogram" on page 7548 for more details).

Once you have starting values for the LAGDISTANCE= and MAXLAGS= options, you can run the VARIOGRAM procedure multiple times to inspect and compare the results you get by specifying different values for these options.

Empirical Semivariogram Computation

Using the values of LAGDISTANCE=7 and MAXLAGS=10 computed previously, rerun PROC VARIOGRAM without the NOVARIOGRAM option in order to compute the empirical semivariogram. You specify the CL option in the COMPUTE statement to calculate the 95% confidence limits for the classical semivariance. The section "COMPUTE Statement" on page 7525 describes how to use the ALPHA= option to specify a different confidence level.

Also, you can request a robust version of the semivariance with the ROBUST option in the COMPUTE statement. PROC VARIOGRAM produces a plot showing both the classical and the robust empirical semivariograms. See the details of the PLOT option to specify different instances of plots of the empirical semivariogram. In addition, ask for the autocorrelation Moran's I and Geary's c statistics under the assumption of randomization using binary weights. The following statements implement all of the preceding requests:

```
proc variogram data=thick outv=outv;
   compute lagd=7 maxlag=10 cl robust
           autocorr(assum=random);
   coordinates xc=East yc=North;
   var Thick;
run;

ods graphics off;
```

Figure 95.6 displays the PROC VARIOGRAM output empirical semivariogram table for the preceding code.

Figure 95.6 Output Table for the Empirical Semivariogram Analysis

The VARIOGRAM Procedure
Dependent Variable: Thick

Empirical Semivariogram

Lag Class	Pair Count	Average Distance	Semivariance Robust	Semivariance Classical	Semivariance Standard Error	95% Confidence Limits	
0	7	2.64	0.0284	0.0336	0.0179	0.000	0.0687
1	82	7.29	0.2098	0.3937	0.0615	0.273	0.5142
2	138	14.16	1.0079	1.1794	0.1420	0.901	1.4577
3	169	21.08	3.0183	2.7988	0.3045	2.202	3.3956
4	205	27.93	4.8107	4.6024	0.4546	3.711	5.4934
5	213	35.17	5.9904	5.9278	0.5744	4.802	7.0536
6	214	42.20	8.1040	7.5181	0.7268	6.094	8.9426
7	250	48.78	7.5326	7.2210	0.6459	5.955	8.4869
8	247	56.16	8.0662	7.1952	0.6475	5.926	8.4642
9	281	62.89	8.2792	6.8445	0.5774	5.713	7.9763
10	250	69.93	8.1440	6.3577	0.5686	5.243	7.4722

Figure 95.7 shows the output from the requested autocorrelation analysis. This includes the observed (computed) Moran's I and Geary's c coefficients, the expected value and standard deviation for each coefficient, the corresponding Z score, and the p-value in the Pr $>| \, Z \, |$ column. The low p-values suggest strong autocorrelation for both statistics types. Note that a two-sided p-value is reported, which is the probability that the observed coefficient lies farther away from $| \, Z \, |$ on either side of the coefficient's expected value—that is, lower than $-Z$ or higher than Z. The sign of Z for both Moran's I and Geary's c coefficients indicates positive autocorrelation in the Thick data values; see the section "Interpretation" on page 7556 for more details.

Figure 95.7 Output Table for the Autocorrelation Statistics

Autocorrelation Statistics

Assumption	Coefficient	Observed	Expected	Std Dev	Z	Pr > |Z|
Randomization	Moran's I	1.0833	-0.0135	0.147	7.47	<.0001
Randomization	Geary's c	0.0185	1.0000	0.228	-4.30	<.0001

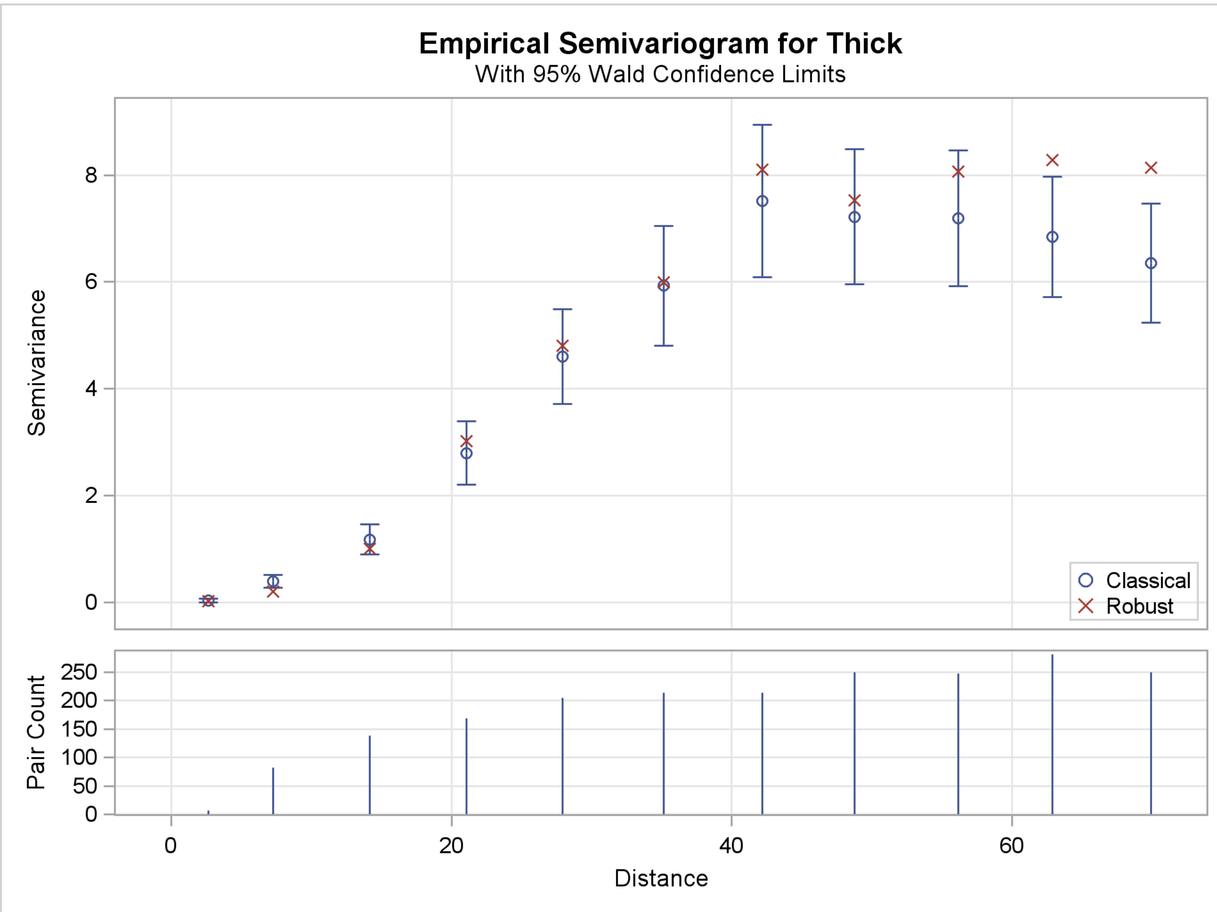
Figure 95.8 shows both the classical and robust empirical semivariograms. In addition, the plot features the approximate 95% confidence limits for the classical semivariance. The figure exhibits a typical behavior of the computed semivariance uncertainty, where in general the variance increases with distance from the origin at Distance=0.

The needle plot in the lower part of the Figure 95.8 provides the number of pairs that were used in the computation of the empirical semivariance for each lag class shown. Note that in general this is a pairwise distribution that is different from the distribution depicted in Figure 95.4. First, the number of pairs shown in the needle plot depends on the particular criteria you specify in the

COMPUTE statement of PROC VARIOGRAM. Second, the distances shown for each lag on the Distance axis are not the midpoints of the lag classes as in the pairwise distances plot, but rather the average distance from the origin Distance=0 of all pairs in a given lag class.

"Example 95.1: Theoretical Semivariogram Model Fitting" on page 7562 continues from this point to show how to choose a theoretical model for the thickness spatial dependence.

Figure 95.8 Classical and Robust Empirical Semivariograms for Coal Seam Thickness Data



Syntax: VARIOGRAM Procedure

The following statements are available in PROC VARIOGRAM:

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

The COMPUTE and COORDINATES statements are required.

Table 95.1 outlines the options available in PROC VARIOGRAM classified by function.

Table 95.1 Options Available in the VARIOGRAM Procedure

Task	Statement	Option
Data Set Options		
Specify input data set	PROC VARIOGRAM	DATA=
Suppress normal display of results	PROC VARIOGRAM	NOPRINT
Write autocorrelation weights information	PROC VARIOGRAM	OUTACWEIGHTS=
Write distance histogram information	PROC VARIOGRAM	OUTDISTANCE=
Write pairwise point information	PROC VARIOGRAM	OUTPAIR=
Write spatial continuity measures	PROC VARIOGRAM	OUTVAR=
Specify plot display and options	PROC VARIOGRAM	PLOTS
Declaring the Role of Variables		
Specify variables to define analysis subgroups	BY	
Specify the analysis variables	VAR	
Specify the x, y coordinates in the DATA= data set	COORDINATES	XCOORD= YCOORD=
Controlling Continuity Measure Computations		
Specify the confidence level	COMPUTE	ALPHA=
Specify the angle tolerances for angle classes	COMPUTE	ANGLETOL=
Compute autocorrelation statistics	COMPUTE	AUTOCORRELATION
Specify the bandwidths for angle classes	COMPUTE	BANDWIDTH=
Compute semivariance estimate variance	COMPUTE	CL
Specify the minimum distance that indicates any two distinct points are not collocated	COMPUTE	DEPSILON=
Specify the basic lag distance	COMPUTE	LAGDISTANCE=
Specify the tolerance around the lag distance	COMPUTE	LAGTOLERANCE=
Specify the maximum number of lags in computations	COMPUTE	MAXLAGS=
Specify the number of angle classes	COMPUTE	NDIRECTIONS=
Suppress computation of all continuity measures	COMPUTE	NOVARIOGRAM
Compute robust semivariance	COMPUTE	ROBUST
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.

NOPRINT
suppresses the normal display of results. The NOPRINT option is useful when you want only to create one or more output data sets with the procedure. Note that this option temporarily disables the Output Delivery System (ODS); see the section "ODS Graphics" on page 7561 for more information.

OUTACWEIGHTS=*SAS-data-set*

OUTACW=*SAS-data-set*

OUTA=*SAS-data-set*
specifies a SAS data set in which to store the autocorrelation weights 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 OUTACWEIGHTS= data set contains $[n(n-1)]/2$ observations.

See the section "OUTACWEIGHTS=*SAS-data-set*" on page 7557 for details.

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 COMPUTE 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 7558 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 $[n(n-1)]/2$ 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 7558 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 7559 for details.

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

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

controls the plots produced through ODS Graphics. When you specify only one plot request, you can omit the parentheses around the plot request. Here are some examples:

```
plots=none
plots=observ
plots=(observ semivar)
plots(unpack)=semivar
plots=(semivar(cla unpack) semivar semivar(rob))
```

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

```
ods graphics on;

proc variogram data=thick;
   compute novariogram;
   coordinates xc=East yc=North;
   var Thick;
run;

ods graphics off;
```

For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." If you have enabled ODS Graphics but do not specify the PLOTS= option or have specified PLOTS=ALL, then PROC VARIOGRAM produces a default set of plots, which might be different for different COMPUTE statement options, as discussed in the following.

- If you specify NOVARIOGRAM in the COMPUTE statement, the VARIOGRAM procedure produces a scatter plot of your observations spatial distribution, as well as the histogram of the pairwise distances of your data. For an example of the observations plot, see Figure 95.2. For an example of the pairwise distances plot, see Figure 95.4.
- If you do not specify NOVARIOGRAM in the COMPUTE statement, the VARIOGRAM procedure computes the empirical semivariogram for the specified LAGDISTANCE= and MAXLAGS= options. The observations plot appears by default in this case, too. The VARIOGRAM procedure also produces a plot of the classical empirical semivariogram. If you further specify ROBUST in the COMPUTE statement, then the VARIOGRAM procedure instead produces a plot of both the classical and robust empirical semivariograms, in addition to the observations plot. For an example of the empirical semivariograms plot, see Output 95.3.4.

The *global-plot-options* include the following:

ONLY

suppresses the default plots. Only plots specifically requested are displayed.

UNPACKPANEL

UNPACK

suppresses paneling. By default, multiple plots can appear in some output panels. Specify UNPACKPANEL to get each plot in a separate panel. You can specify PLOTS(UNPACKPANEL) to unpack the default plots. You can also specify UNPACKPANEL as a suboption with the SEMIVAR option.

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

ALL

produces all appropriate plots. You can specify other options with ALL. For example, to request all default plots and an additional classical empirical semivariogram, specify PLOTS=(ALL SEMIVAR(CLA)).

EQUATE

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

NONE

suppresses all plots.

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

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

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

produces the observed data plot. Only one observations plot will be created if you specify the OBSERVATIONS option more than once within a PLOTS option.

The OBSERVATIONS option has the following suboptions:

GRADIENT

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

OUTLINE

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

OUTLINEGRADIENT

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

SHOWMISSING

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

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

PAIRS < **(***pairs-plot-options***)** >

specifies that the pairwise distances histogram be produced. By default, the horizontal axis displays the lag class number. The vertical axis shows the frequency (count) of pairs in the lag classes. Notice that the zero lag class width is half the width of the other classes.

The PAIRS option has the following suboptions:

MIDPOINT

MID

specifies that the plot created with the PAIRS option display the lag class midpoint value on the horizontal axis, rather than the default lag class number. The midpoint value is the actual distance of a lag class center from the assumed origin point at distance zero (see also the illustration in Figure 95.14).

NOINSET

NOI

specifies that the plot created with the PAIRS option be produced without the default inset that provides additional information about the pairs distribution.

THRESHOLD=*minimum pairs*

THR=*minimum pairs*

specifies that a reference line appear in the plot created with the PAIRS option to indicate the *minimum pairs* frequency of data pairs. You can use this line as an exploratory tool when you want to select lag classes that contain at least THRESHOLD point pairs. The option helps you to identify visually any portion of the PAIRS distribution that lies below the specified THRESHOLD value.

Only one pairwise distances histogram will be created if you specify the PAIRS option within a PLOTS option. If you specify multiple instances of the PAIRS option, the resulting plot will have the following features:

- If the MIDPOINT or NOINSET suboption has been specified in any of the instances, it will be activated in the resulting plot.
- If you have specified the THRESHOLD= suboption more than once, then the THRESHOLD= value specified last will prevail.

SEMIVARIOGRAM < **(***semivar-plot-options***)** >

SEMIVAR < **(***semivar-plot-options***)** >

specifies that the empirical semivariogram plot be produced. You can specify the SEMIVAR option multiple times in the same PLOTS option to request instances of plots with the following *semivar-plot-options*:

ALL | CLASSICAL | ROBUST

ALL | CLA | ROB

specifies a single type of empirical semivariogram (classical or robust) to plot, or specifies that all the available types be included in the same plot. The default is ALL.

UNPACKPANEL

UNPACK

specifies that paneled semivariogram plots be displayed separately. By default, plots appear in a panel, when appropriate.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC VARIOGRAM to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

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

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

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

COMPUTE Statement

COMPUTE *computation-options* **;**

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

ALPHA=*number*

specifies a parameter to obtain the confidence level for constructing confidence limits in the classical empirical semivariance estimation. The value of *number* must be between 0 and 1, and the confidence level is $1-number$. The default is ALPHA=0.05, which corresponds to the default confidence level of 95%. If the CL option is not specified, ALPHA= is ignored.

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 $180°/(2n_d)$, where n_d is the NDIRECTIONS= specification. If you do not specify the NDIRECTIONS= option or the DIRECTIONS statement, ANGLE-TOLERANCE= is ignored.

See the section "Theoretical and Computational Details of the Semivariogram" on page 7536 for further information.

Experimental

AUTOCORRELATION < **(***autocorrelation-options***)** >

AUTOCORR < **(***autocorrelation-options***)** >

AUTOC < **(***autocorrelation-options***)** >

specifies that autocorrelation statistics be calculated. You can further specify the following *autocorrelation-options* in parentheses following the experimental AUTOCORRELATION option.

ASSUMPTION < **=** *assumption-options* >

ASSUM < **=** *assumption-options* >

specifies the type of autocorrelation assumption to use. The *assumption-options* can be one of the following:

NORMALITY | NORMAL | NOR

specifies use of the normality assumption.

RANDOMIZATION | RANDOM | RAN

specifies use of the randomization assumption.

The default is ASSUMPTION=NORMALITY.

STATISTICS < **=** **(***stats-options***)** >

STATS < **=** **(***stats-options***)** >

specifies the autocorrelation statistics in detail. The *stats-options* can be one or more of the following:

ALL

applies all available types of autoregression statistics.

GEARY | GEA

specifies use of the Geary c statistics.

MORAN | MOR

specifies use of the Moran I statistics.

The default is STATISTICS=ALL.

WEIGHTS < = *weights-options* **>**

WEI < = *weights-options* **>**

specifies the scheme used for the computation of the autocorrelation weights. You can choose one of the following *weights-options*:

BINARY < (*binary-option***) >**

specifies that binary weights be used. You also have the following *binary-option*:

ROWAVERAGING | ROWAVG | ROW

specifies that asymmetric autocorrelation weights be assigned to data pairs. For each observation, if there are nonzero weights, the ROWAVG option standardizes those weights so that they sum to 1. No row averaging is performed by default.

DISTANCE < (*distance-options***) >**

specifies that autocorrelation weights be assigned based on the point pair distances. You also have the following *distance-options*:

NORMALIZE | NORMAL | NOR

specifies that normalized pair distances be used in the distance-based weights expression. The distances are normalized with respect to the maximum pairwise distance h_b, as it is defined in the section "Computation of the Distribution Distance Classes" on page 7545. By default, nonnormalized values are used in the computations.

POWER | POW

specifies the power to which the pair distance is raised in the distance-based weights expression. POWER is a nonnegative number, and its default value is POWER=1.

ROWAVERAGING | ROWAVG | ROW

specifies that asymmetric autocorrelation weights be assigned to data pairs. For each observation, if there are nonzero weights, the ROWAVG option standardizes those weights so that they sum to 1. No row averaging is performed by default.

SCALE | SCA

specifies the scaling factor in the distance-based weights expression. SCALE is a nonnegative number, and its default value is SCALE=1.

The default is WEIGHTS=BINARY. See the section "Autocorrelation Statistics (Experimental)" on page 7552 for further details about the autocorrelation weights.

When you specify the AUTOCORRELATION option with no *autocorrelation-options*, PROC VARIOGRAM computes by default both the Moran I and Geary c statistics with p-values computed under the normality assumption with binary weights.

If you specify more than one ASSUMPTION in the *autocorrelation-options*, all but the last specified ASSUMPTION will be ignored. The same holds if you specify more than one POWER= or SCALE= parameter in the WEIGHT=DISTANCE *distance-options*.

If you specify the WEIGHT=BINARY option in the AUTOCORRELATION option and the NOVARIOGRAM option at the same time, then you must also specify the LAGDISTANCE= option in the COMPUTE statement. See the section "Autocorrelation Weights" on page 7552 for more 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. See Figure 95.15 for a visual representation of the bandwidth.

If you do not specify the BANDWIDTH= option, no restriction occurs. If you do not specify the NDIRECTIONS= option or the DIRECTIONS statement, BANDWIDTH= is ignored.

CL

requests confidence limits for the classical semivariance estimate. The confidence limits' lower bound is always nonnegative, adhering to the behavior of the theoretical semivariance. You can control the confidence level with the ALPHA= option.

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=$\Delta\varepsilon$, then pairs of points P_1 and P_2 for which the distance between them $| P_1 P_2 | < \Delta\varepsilon$ are excluded from the continuity measure calculations. The default value of the DEPSILON= option is 100 times the machine precision; 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 $| P_1 P_2 |$, is calculated. If $| P_1 P_2 | = x$, then this pair is in the first lag class. If $| P_1 P_2 | = 2x$, 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 $| P_1 P_2 |$ is within $x \pm \delta x$, you would place this pair in the first lag class; if $| P_1 P_2 |$ is within $2x \pm \delta x$, you would place this pair in the second lag class, and so on.

You can experiment and determine the candidate values for the LAGDISTANCE= option by plotting the pairwise distance histogram for different numbers of histogram classes, using the NHC= option.

A LAGDISTANCE= value is required for the semivariance and the autocorrelation computations. You need not specify LAGDISTANCE= only when you specify the NOVARIOGRAM option without the AUTOCORRELATION option.

See the section "Theoretical and Computational Details of the Semivariogram" on page 7536 for more information.

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 description of the LAGDISTANCE= option for information about the use of the LAGTOLERANCE= option, and the section "Theoretical and Computational Details of the Semivariogram" on page 7536 for more details.

If you do not specify the LAGTOLERANCE= option, a default value of $\frac{1}{2}$ 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, $\mid P_1 P_2 \mid$, 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 is a field property where the characterization of spatial continuity depends on the data pair orientation (or angle between the N–S direction and the axis defined by the data pair). Isotropy is the absence of this effect; that is, the description of spatial continuity 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° (or its reciprocal, 180°), 60° (or its reciprocal, 240°), and 120° (or its reciprocal, 300°). 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 $\theta = 180° / n_d$, 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. This is the omnidirectional case where the spatial continuity measures are assumed to be isotropic.

The NDIRECTIONS= option is useful for exploring possible anisotropy. The DIRECTIONS statement, described in the section "DIRECTIONS Statement" on page 7531, provides greater control over the angle classes.

See the section "Theoretical and Computational Details of the Semivariogram" on page 7536 for more information.

NHCLASSES=*number of histogram classes*

NHCLASS=*number of histogram classes*

NHC=*number of histogram classes*

specifies the number of distance classes to consider in the spatial domain in the exploratory stage of the empirical semivariogram computation. The actual number of classes is one more than the NHCLASSES= value, since a special lag zero class is also computed. The NHC= option is used to produce the distance intervals table, the histogram of pairwise distances, and the OUTDISTANCE= data set. See the OUTDISTANCE= option as well as the section "OUTDIST=*SAS-data-set*" on page 7558 and the section "Theoretical and Computational Details of the Semivariogram" on page 7536 for more information.

The default value of the NHCLASSES= option is 10.

NOVARIOGRAM

prevents the computation of the continuity measures. This option is useful for preliminary analysis, or 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 $| P_1 P_2 |$ between each pair of points P_1 and P_2 is checked against d_{max}. If $| P_1 P_2 | > d_{max}$, 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 semivariance be calculated in addition to the classical semivariance.

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)*

X=*(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)*

Y=*(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* **;**

You use the DIRECTIONS statement to define angle classes. You can specify angle classes as a list of angles, separated by commas, with optional angle tolerances and bandwidths within parentheses following the angle. You must specify at least one angle.

If you do not specify the optional angle tolerance, the default value of 45° is used. If you do not specify the optional bandwidth, no bandwidth is checked. If you specify a bandwidth, you must also specify an angle tolerance.

For example, suppose you want to compute three separate semivariograms at angles $\theta_1 = 0°$, $\theta_2 = 60°$, and $\theta_3 = 120°$, with corresponding angle tolerances $\delta\theta_1 = 22.5°$, $\delta\theta_2 = 12.5°$, and $\delta\theta_3 = 22.5°$, 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 as follows:

```
directions 0.0(22.5,50), 60.0(12.5,40),120(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: VARIOGRAM Procedure

Theoretical Semivariogram Models

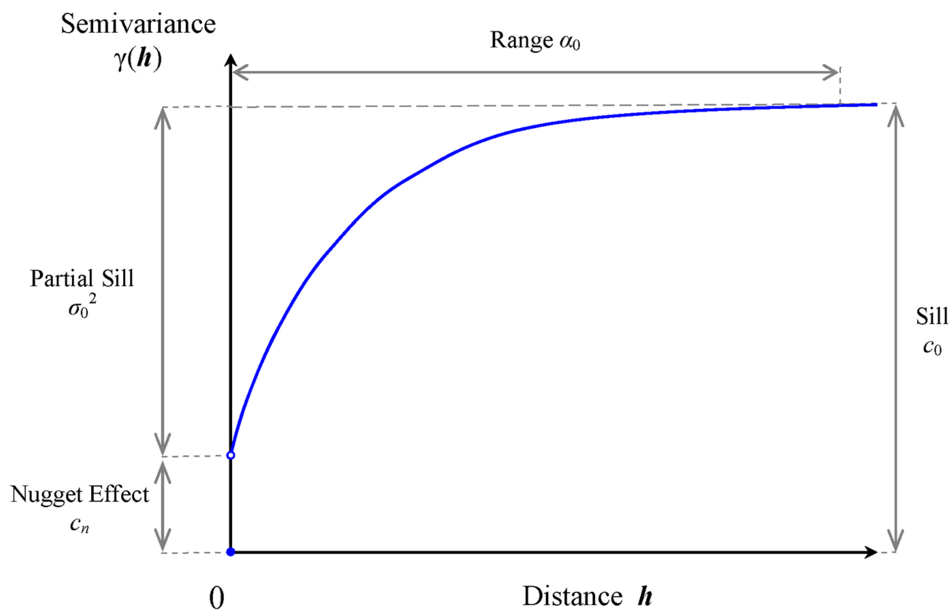
The VARIOGRAM procedure computes the empirical (also known as *sample* or *experimental*) semivariogram from a set of point measurements. Semivariograms are used in the first steps of spatial prediction as tools that provide insight into the spatial continuity and structure of a random process. Naturally occurring randomness is accounted for by describing a process in terms of the *spatial random field* (SRF) concept (Christakos 1992). An SRF is a collection of random variables throughout your spatial domain of prediction. For some of them you already have measurements, and your data set constitutes part of a single realization of this SRF. Based on your sample, spatial prediction aims to provide you with values of the SRF at locations where no measurements are available.

Prediction of the SRF values at unsampled locations by techniques such as ordinary kriging requires the use of a theoretical semivariogram or covariance model. Due to the randomness involved in stochastic processes, the theoretical semivariance cannot be computed. Instead, it is possible that the empirical semivariance can provide an estimate of the theoretical semivariance, which can then be used to characterize the spatial structure of the process.

It is critical to note that the empirical semivariance provides an estimate of its theoretical counterpart only when the SRF satisfies stationarity conditions. These conditions imply that the SRF has a constant (or zero) expected value. Consequently, your data need to be sampled from a trend-free random field and need to have a constant mean, as assumed in "Example 95.1: Theoretical Semivariogram Model Fitting" on page 7562. Equivalently, your data could be residuals of an initial sample that has had a surface trend removed, as portrayed in "Example 95.2: An Anisotropic Case Study with Surface Trend in the Data" on page 7566. For a closer look at stationarity, see the section "Stationarity" on page 7538. For details about different stationarity types and conditions see, for example, Chilès and Delfiner (1999, Section 1.1.4).

When you obtain a valid empirical estimate of the theoretical semivariance, it is then necessary to choose a type of theoretical semivariogram model based on that estimate. Commonly used theoretical semivariogram shapes rise monotonically as a function of distance. The shape is typically characterized in terms of particular parameters; these are the *range* a_0, the *sill* (or *scale*) c_0, and the *nugget effect* c_n. Figure 95.9 displays a theoretical semivariogram of a spherical semivariance model, and points out the semivariogram characteristics.

Figure 95.9 A Theoretical Semivariogram of Spherical Type and Its Characteristics



Specifically, the sill is the semivariogram upper bound. The range a_0 denotes the distance at which the semivariogram reaches the sill. When the semivariogram increases asymptotically toward its sill value, as occurs in the exponential and Gaussian semivariogram models, the term *effective* (or *practical*) range is also used. The effective range r_ϵ is defined as the distance at which the semivariance value achieves 95% of the sill. In particular, for these models the relationship between the range and effective range is $r_\epsilon = 3a_0$ (exponential model) and $r_\epsilon = \sqrt{3}a_0$ (Gaussian model).

The nugget effect c_n represents a discontinuity of the semivariogram that can be present at the origin. It is typically attributed to microscale effects or measurement errors. The semivariance is always 0 at distance $\boldsymbol{h} = 0$; hence, the nugget effect demonstrates itself as a jump in the semivariance as soon as $\boldsymbol{h} > 0$ (note in Figure 95.9 the discontinuity of the function at $\boldsymbol{h} = 0$ in the presence of a nugget effect).

The sill c_0 comprises the nugget effect, if present, and the *partial sill* ${\sigma_0}^2$; that is, $c_0 = c_n + {\sigma_0}^2$. If the SRF $Z(\boldsymbol{s})$ is second-order stationary (see the section "Stationarity" on page 7538), the estimate of the sill is an estimate of the constant variance $\text{Var}[Z(\boldsymbol{s})]$ of the field. Nonstationary processes have variances that depend on the location $\boldsymbol{s}$. Their semivariance increases with distance, hence their semivariograms do not have a sill.

Not every function is a suitable candidate for a theoretical semivariogram model. The semivariance function $\gamma_z(\boldsymbol{h})$, as defined in the following section, is a so-called *conditionally negative-definite* function that satisfies (Cressie 1993, p. 60)

$$\sum_{i=1}^{m}\sum_{j=i}^{m} q_i q_j \gamma_z(\boldsymbol{s}_i - \boldsymbol{s}_j) \le 0$$

for any number m of locations $\boldsymbol{s}_i, \boldsymbol{s}_j$ in $\mathcal{R}^2$ with $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$, and any real numbers q_i such that $\sum_{i=1}^{m} q_i = 0$. Permissible, commonly used theoretical semivariogram models include the ones shown in Table 95.2.

Table 95.2 Some Permissible Theoretical Semivariogram Models ($a_0 > 0$)

Model Type	Semivariance
Exponential	$\gamma_z(\boldsymbol{h}) = \begin{cases} 0 & , \text{if } \mid \boldsymbol{h} \mid = 0 \\ c_n + \sigma_0{}^2 \left[1 - \exp\left(-\frac{\lvert \boldsymbol{h} \rvert}{a_0}\right)\right] & , \text{if } 0 < \mid \boldsymbol{h} \mid \end{cases}$
Gaussian	$\gamma_z(\boldsymbol{h}) = \begin{cases} 0 & , \text{if } \mid \boldsymbol{h} \mid = 0 \\ c_n + \sigma_0{}^2 \left[1 - \exp\left(-\frac{\lvert \boldsymbol{h} \rvert^2}{a_0^2}\right)\right] & , \text{if } 0 < \mid \boldsymbol{h} \mid \end{cases}$
Power	$\gamma_z(\boldsymbol{h}) = \begin{cases} 0 & , \text{if } \mid \boldsymbol{h} \mid = 0 \\ c_n + \sigma_0{}^2 \boldsymbol{h}^{a_0} & , \text{if } 0 < \mid \boldsymbol{h} \mid \end{cases}$
Spherical	$\gamma_z(\boldsymbol{h}) = \begin{cases} 0 & , \text{if } \mid \boldsymbol{h} \mid = 0 \\ c_n + \sigma_0{}^2 \left[\frac{3}{2}\frac{\lvert \boldsymbol{h} \rvert}{a_0} - \frac{1}{2}(\frac{\lvert \boldsymbol{h} \rvert}{a_0})^3\right] & , \text{if } 0 < \mid \boldsymbol{h} \mid \le a_0 \\ c_0, & , \text{if } a_0 < \mid \boldsymbol{h} \mid \end{cases}$

You can review these models in further detail in the section "Theoretical Semivariogram Models" on page 2945 in the KRIGE2D procedure documentation.

The theoretical semivariogram models are used to describe the spatial structure of random processes. Based on their shape and characteristics, the semivariograms of these models can provide a plethora of information (Christakos 1992, Section 7.3):

- Examination of the semivariogram variation in different directions provides information about the isotropy of the random process (see also the discussion about isotropy in the following section).
- The semivariogram range determines the zone of influence extending from any given location. Values at surrounding locations within this zone are correlated with the value at the specific location by means of the particular semivariogram.
- The semivariogram behavior at large distances indicates the degree of stationarity of the process. In particular, an asymptotic behavior suggests a stationary process, whereas either a linear increase and slow convergence to the sill or a fast increase is an indicator of nonstationarity.
- The semivariogram behavior close to the origin indicates the degree of regularity of the process variation. Specifically, a parabolic behavior at the origin implies a very regular spatial variation, whereas a linear behavior characterizes a nonsmooth process. The presence of a nugget effect is additional evidence of irregularity in the process.
- The semivariogram behavior within the range provides description of potential periodicities or anomalies in the spatial process.

A brief note on terminology: In some fields (for example, geostatistics) the term homogeneity is sometimes used instead of stationarity in spatial analysis, whereas in statistics homogeneity is defined differently (Banerjee, Carlin, and Gelfand 2004, Section 2.1.3). In particular, the alternative terminology characterizes as homogeneous the stationary SRF in $\mathcal{R}^n, n > 1$, whereas it retains the term stationary for such SRF in $\mathcal{R}^1$ (SRF in $\mathcal{R}^1$ are also known as *random processes*). Often, studies in a single dimension refer to temporal processes; hence, you might see time-stationary random processes called "temporally stationary" or simply stationary, and stationary SRF in $\mathcal{R}^n, n > 1$, characterized as "spatially homogeneous" or simply homogeneous. This distinction made by the alternative nomenclature is more evident in spatiotemporal random fields (S/TRF), where the different terms clarify whether stationarity applies in the spatial or the temporal part of the S/TRF.

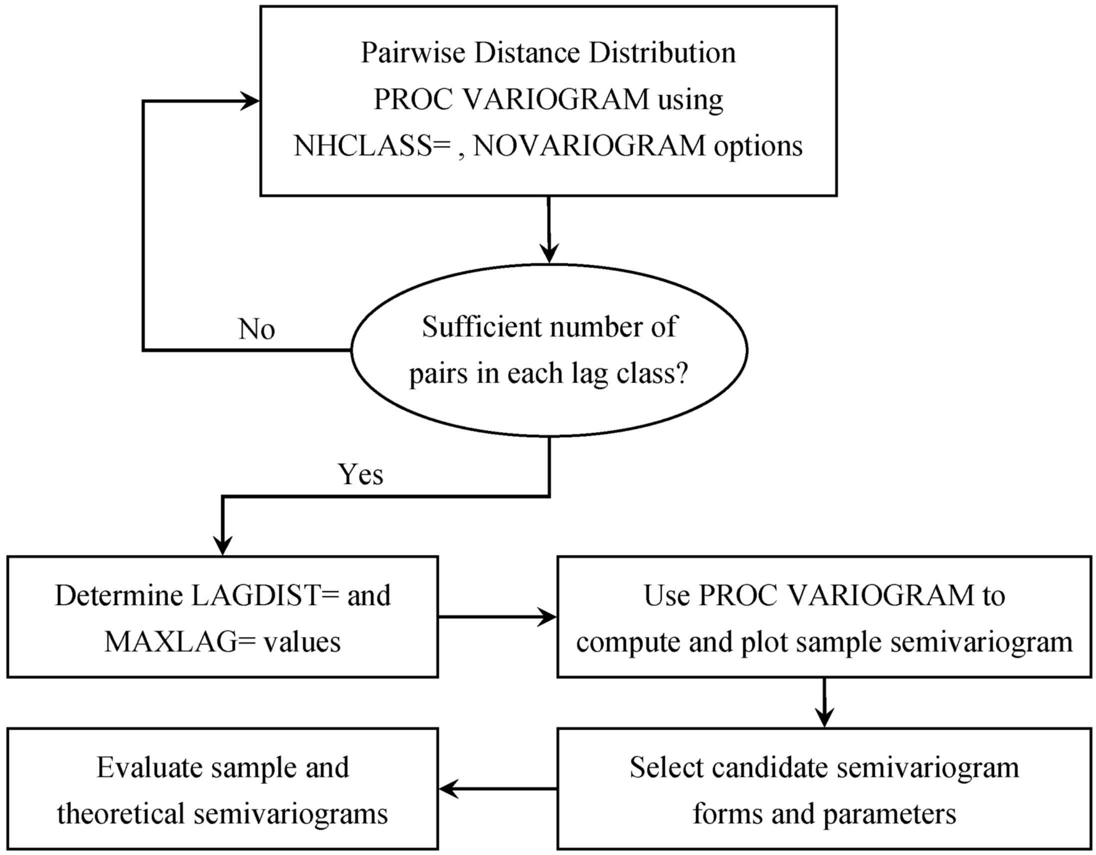
Typically, you choose a theoretical semivariogram model to fit the empirical semivariance in an automated manner. For this task you can use methods such as least squares, maximum likelihood, and robust methods (Cressie 1993, Section 2.6). "Example 95.1: Theoretical Semivariogram Model Fitting" on page 7562 illustrates the fitting process by using ordinary and weighted least squares methods. A different approach is manual fitting, where a theoretical semivariogram model is chosen based on visual inspection of the empirical semivariogram; see, for example, Hohn (1988, p. 25).

In some cases, you might see that using a combination of theoretical models results in a more accurate fit onto the empirical semivariance than using a single model. This is known as model nesting. Nested models, anisotropic models, and the nugget effect increase the scope of theoretical models available. All these concepts are discussed in the section "Theoretical Semivariogram Models" on page 2945 in the KRIGE2D procedure documentation.

Overall, Goovaerts (1997, Section 4.2.4) suggests that fitting a theoretical model should aim to capture the major spatial features. An accurate fit is desirable, but overfitting does not offer advantages, because you might find yourself trying to model possibly spurious details of the empirical semivariogram.

Note the general flow of investigation. The empirical semivariogram is computed after a suitable choice is made for the LAGDISTANCE= and MAXLAGS= options. For computations in more than one directions you can further use the NDIR= option or the DIRECTIONS statement. Potential theoretical models (which can also incorporate nesting, anisotropy, and the nugget effect) are then plotted against the empirical semivariogram and evaluated. The flow of this analytical process is illustrated in Figure 95.10. After a suitable theoretical model is determined, it is used in PROC KRIGE2D for the prediction stage. The prediction analysis is presented in detail in the section "Details of Ordinary Kriging" on page 2960 in the KRIGE2D procedure documentation.

Figure 95.10 Flowchart for Variogram Selection



Theoretical and Computational Details of the Semivariogram

Let $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D \subset \mathcal{R}^2\}$ be a spatial random field (SRF) with n measured values $z_i = Z(\boldsymbol{s}_i)$ at respective locations $\boldsymbol{s}_i$, $i = 1, \ldots, n$. You use the VARIOGRAM procedure because you want to gain insight into the spatial continuity and structure of $Z(\boldsymbol{s})$. A good measure of the spatial continuity of $Z(\boldsymbol{s})$ is defined by means of the variance of the difference $Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)$, where $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are locations in D. Specifically, if you consider $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ to be spatial increments such that $\boldsymbol{h} = \boldsymbol{s}_j - \boldsymbol{s}_i$, then the variance function based on the increments $\boldsymbol{h}$ is independent of the actual locations $\boldsymbol{s}_i, \boldsymbol{s}_j$. Most commonly, the continuity measure used in practice is one half of this variance, better known as the *semivariance* function:

$$\gamma_z(\boldsymbol{h}) = \frac{1}{2}\mathrm{Var}[Z(\boldsymbol{s} + \boldsymbol{h}) - Z(\boldsymbol{s})]$$

or, equivalently,

$$\gamma_z(\boldsymbol{h}) = \frac{1}{2}\left(\mathrm{E}\{[Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s})]^2\} - \{\mathrm{E}[Z(\boldsymbol{s}+\boldsymbol{h})] - \mathrm{E}[Z(\boldsymbol{s})]\}^2\right)$$

The plot of semivariance as a function of $\boldsymbol{h}$ is the *semivariogram*. In extension to its meaning, you might commonly see the term *semivariogram* used instead of the term *semivariance*, as well.

Assume that the SRF $Z(\boldsymbol{s})$ is free of nonrandom (or systematic) surface trends. Then, the expected value $\mathrm{E}[Z(\boldsymbol{s})]$ of $Z(\boldsymbol{s})$ will be a constant for all $\boldsymbol{s} \in \mathcal{R}^2$, and the semivariance expression is simplified to the following:

$$\gamma_z(\boldsymbol{h}) = \frac{1}{2}\mathrm{E}\{[Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s})]^2\}$$

Given the preceding assumption, you can compute an estimate $\hat{\gamma}_z(\boldsymbol{h})$ of the semivariance $\gamma_z(\boldsymbol{h})$ from a finite set of points in a practical way by using the formula

$$\hat{\gamma}_z(\boldsymbol{h}) = \frac{1}{2 \mid N(\boldsymbol{h}) \mid} \sum_{N(\boldsymbol{h})} [Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)]^2$$

where the sets $N(\boldsymbol{h})$ contain all the neighboring pairs at distance $\boldsymbol{h}$:

$$N(\boldsymbol{h}) = \{i, j : \boldsymbol{s}_i - \boldsymbol{s}_j = \boldsymbol{h}\}$$

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

The expression for $\hat{\gamma}_z(\boldsymbol{h})$ is called the *empirical semivariance* (Matheron 1963). This is the quantity that PROC VARIOGRAM computes, and its corresponding plot is the *empirical semivariogram*. According to Cressie (1993, p. 96), the estimate $\hat{\gamma}_z(\boldsymbol{h})$ has approximate variance

$$\mathrm{Var}[\hat{\gamma}_z(\boldsymbol{h})] \simeq \frac{2[\gamma_z(\boldsymbol{h})]^2}{N(\boldsymbol{h})}$$

The empirical semivariance $\hat{\gamma}_z(\boldsymbol{h})$ is also referred to as *classical*. This name is used so that it can be distinguished from the *robust semivariance* estimate $\bar{\gamma}_z(\boldsymbol{h})$ and the corresponding *robust semivariogram*. The robust semivariance was introduced by Cressie and Hawkins (1980) and is described by Cressie (1993, p. 75) as:

$$\bar{\gamma}_z(\boldsymbol{h}) = \frac{\Psi^4(\boldsymbol{h})}{2[0.457 + 0.494/N(\boldsymbol{h})]}$$

In the preceding expression the parameter $\Psi(\boldsymbol{h})$ is defined as:

$$\Psi(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})} \sum_{P_i P_j \in N(\boldsymbol{h})} [Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)]^{\frac{1}{2}}$$

Note that if your data include a surface trend, then the empirical semivariance $\hat{\gamma}_z(\boldsymbol{h})$ is not an estimate of the theoretical semivariance function $\gamma_z(\boldsymbol{h})$. Instead, rather than the spatial increments variance, it represents a different quantity known as *pseudo-semivariance*, and its corresponding plot is a *pseudo-semivariogram*. In principle, pseudo-semivariograms do not provide measures of the spatial continuity. They can thus lead to misinterpretations of the $Z(\boldsymbol{s})$ spatial structure, and are consequently unsuitable for the purpose of spatial prediction. For further information, see the detailed discussion in the section “Empirical Semivariograms and Surface Trends” on page 7551. Under certain conditions you might be able to gain some insight about the spatial continuity with a pseudo-semivariogram. This case is presented in “Example 95.3: Analysis without Surface Trend Removal” on page 7579.

Stationarity

In the combined presence of the previous two assumptions—that is, when $\mathrm{E}[Z(\boldsymbol{s})]$ is constant and spatial increments are used to define $\gamma_z(\boldsymbol{h})$—the SRF $Z(\boldsymbol{s})$ is characterized as *intrinsically stationary* (Cressie 1993, p. 40).

The expected value $\mathrm{E}[Z(\boldsymbol{s})]$ is the first statistical moment of the SRF $Z(\boldsymbol{s})$. The second statistical moment of the SRF $Z(\boldsymbol{s})$ is the *covariance* function between two points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ in $Z(\boldsymbol{s})$, and it is defined as

$$C_z(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathrm{E}\left([Z(\boldsymbol{s}_i) - \mathrm{E}[Z(\boldsymbol{s}_i)]]\left[Z(\boldsymbol{s}_j) - \mathrm{E}[Z(\boldsymbol{s}_j)]\right]\right)$$

Note that when $\boldsymbol{s}_i = \boldsymbol{s}_j = \boldsymbol{s}$, the covariance expression provides the variance at $\boldsymbol{s}$.

The assumption of a constant $\mathrm{E}[Z(\boldsymbol{s})] = m$ means that the expected value is invariant with respect to translations of the spatial location $\boldsymbol{s}$. The covariance is considered invariant to such translations when it depends only on the distance $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$ between any two points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. If both of these conditions are true, then the preceding expression becomes

$$C_z(\boldsymbol{s}_i, \boldsymbol{s}_j) = C_z(\boldsymbol{s}_i - \boldsymbol{s}_j) = C_z(\boldsymbol{h}) = \mathrm{E}\left([Z(\boldsymbol{s}) - m][Z(\boldsymbol{s} + \boldsymbol{h}) - m]\right)$$

When both $\mathrm{E}[Z(\boldsymbol{s})]$ and $C(\boldsymbol{s}_i, \boldsymbol{s}_j)$ are invariant to spatial translations, the SRF $Z(\boldsymbol{s})$ is characterized as *second-order stationary* (Cressie 1993, p. 53).

In a second-order stationary SRF the quantity $C(\boldsymbol{h})$ is the same for any two points that are separated by distance $\boldsymbol{h}$. Based on the preceding formula, for $\boldsymbol{h} = 0$ you can see that the variance is constant throughout a second-order stationary SRF. Hence, second-order stationarity is a stricter condition than intrinsic stationarity.

Under the assumption of second-order stationarity, the semivariance definition at the beginning of this section leads to the conclusion that

$$\gamma_z(\boldsymbol{h}) = C(\mathbf{0}) - C(\boldsymbol{h})$$

which relates the theoretical semivariance and covariance. Note that the empirical estimates of these quantities are not related in exactly the same way, as indicated in Schabenberger and Gotway (2005, Section 4.2.1).

Ergodicity

In addition to the constant $\mathrm{E}[Z(\boldsymbol{s})]$ and the assumption of intrinsic stationarity, *ergodicity* is a necessary third hypothesis to estimate the empirical semivariance. Assume that for the SRF $Z(\boldsymbol{s})$ you have measurements z_i whose sample mean is estimated by $\bar{Z}$. The hypothesis of ergodicity dictates that $\bar{Z} = \mathrm{E}[Z(\boldsymbol{s})]$.

In general, an SRF $Z(\boldsymbol{s})$ is characterized as ergodic if the statistical moments of its realizations coincide with the corresponding ones of the SRF. In spatial analysis we are often interested in the first two statistical moments, and consequently a more relaxed ergodicity assumption is made only for them. See Christakos (1992, Section 2.12) for the use of the ergodicity hypothesis in SRF, and Cressie (1993, p. 57) for a more detailed discussion of ergodicity.

The semivariogram analysis makes implicit use of the ergodicity hypothesis. The VARIOGRAM procedure works with the residual centered values $V(\boldsymbol{s}_i) = v_i = z_i - \bar{Z},\ i = 1, \ldots, n$, where it is assumed that the sample mean $\bar{Z}$ is the constant expected value $\mathrm{E}[Z(\boldsymbol{s})]$ of $Z(\boldsymbol{s})$. This is equivalent to using the original values, since $V(\boldsymbol{s}_i) - V(\boldsymbol{s}_j) = Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)$, which shows the property of the semivariance to filter out the mean. See the section "Semivariance Computation" on page 7550 for the exact expressions PROC VARIOGRAM uses to compute the empirical classical $\hat{\gamma}_z(\boldsymbol{h})$ and robust $\bar{\gamma}_z(\boldsymbol{h})$ semivariances.

Anisotropy

Semivariance is defined on the basis of the spatial increment vector $\boldsymbol{h}$. If the variance characteristics of $Z(\boldsymbol{s})$ are independent of the spatial direction, then $Z(\boldsymbol{s})$ is called *isotropic*; if not, then $Z(\boldsymbol{s})$ is called *anisotropic*. In the case of isotropy, the semivariogram depends only on the length h of $\boldsymbol{h}$ and $\gamma_z(\boldsymbol{h}) = \gamma_z(h)$. Anisotropy is characterized as *geometric*, when the range a_0 of the semivariogram varies in different directions, and *zonal*, when the semivariogram sill c_0 depends on the spatial direction. Either type or both types of anisotropy can be present.

In the more general case, an SRF can be anisotropic. For an accurate characterization of the spatial structure it is necessary to perform individual analyses in multiple directions. Goovaerts (1997, p. 98) suggests an initial investigation in at least one direction more than the working spatial dimensions—for example, at least three different directions in $\mathcal{R}^2$. Olea (2006) supports exploring as many directions as possible when the data set allows.

You might not know in advance whether you have anisotropy or not. If the semivariogram characteristics do not change in different directions, then you assume the SRF is isotropic. If your directional analysis reveals anisotropic behavior in particular directions, then you proceed to focus your analysis on these directions. For example, in an anisotropic SRF in $\mathcal{R}^2$ you should expect to find two distinct directions where you observe the *major axis* and the *minor axis* of anisotropy. The major axis direction is the one in which the semivariogram has maximum range, and hence has the strongest spatial continuity. Conversely, in the minor axis direction the SRF has minimum range and the weakest spatial continuity. See "Example 95.2: An Anisotropic Case Study with Surface Trend in the Data" on page 7566 for a detailed demonstration of a case with anisotropy when you are using PROC VARIOGRAM.

You can find some additional information about anisotropy analysis in the section "Anisotropic Models" on page 2953 in the KRIGE2D procedure documentation.

Pair Formation

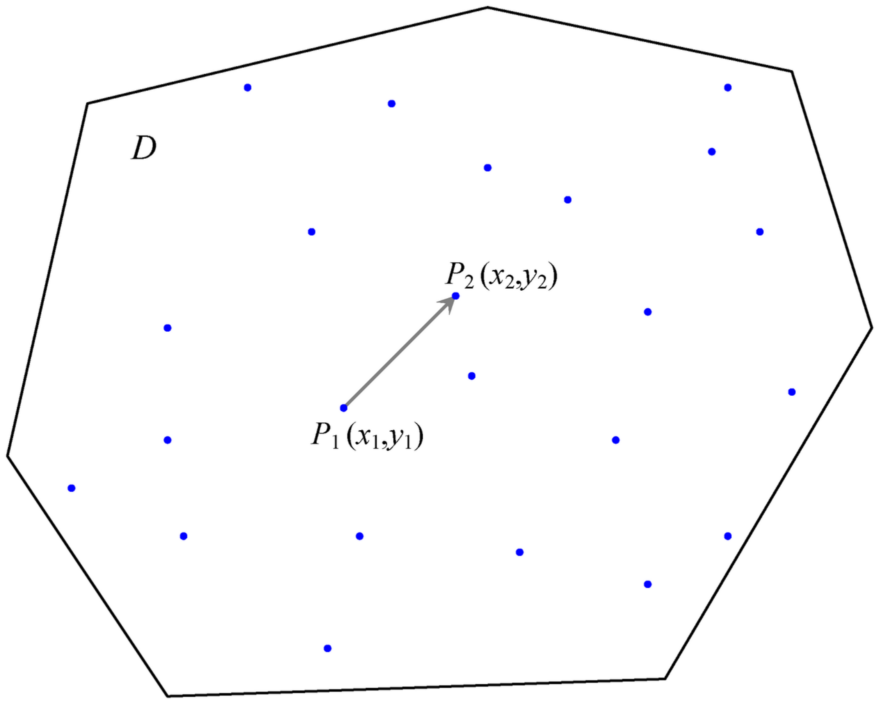
The basic starting point in computing the empirical semivariance is the enumeration of pairs of points for the spatial data. Figure 95.11 shows the spatial domain D and the set of n measurements $z_i, i = 1, \ldots, n$, that have been sampled at the indicated locations in D. Two data points P_1 and P_2, with coordinates $\boldsymbol{s}_1 = (x_1, y_1)$ and $\boldsymbol{s}_2 = (x_2, y_2)$, respectively, are selected for illustration.

A vector, or directed line segment, is drawn between these points. If the length $\mid P_i P_j \mid = \mid \boldsymbol{s}_2 - \boldsymbol{s}_1 \mid = (x_2 - x_1)^2 + (y_2 - y_1)^2$ of this vector is smaller than the specified DEPSILON= value, then the pair is excluded from the continuity measure calculations as the two points P_1 and P_2 are considered to be at zero distance apart (or *collocated*). Spatial collocation might appear due to different scales in sampling, observations made at the same spatial location at different time instances, and errors in the data sets. PROC VARIOGRAM excludes such pairs from the pairwise distance and semivariance computations because they can cause numeric problems in spatial analysis.

If this pair is not discarded on the basis of collocation, it is then classified—first by orientation of the directed line segment $\boldsymbol{s}_2 - \boldsymbol{s}_1$, and then by its length $\mid P_i P_j \mid$. For example, it is unlikely for actual data that the distance $\mid P_i P_j \mid$ between any pair of data points P_i and P_j located at $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, respectively, would exactly satisfy $\mid P_i P_j \mid = \mid \boldsymbol{h} \mid = h$ in the preceding computation of $\hat{\gamma}_z(\boldsymbol{h})$. A similar argument can be made for the orientation of the segment $\boldsymbol{s}_2 - \boldsymbol{s}_1$. Consequently, the pair $P_1 P_2$ is placed into an angle and distance class.

The following subsections give more details about the nature of these classifications. You will also find extensive discussions about the size and the number of classes to consider for the computation of the empirical semivariogram.

Figure 95.11 Selection of Points P_1 and P_2 in Spatial Domain D



Angle Classification

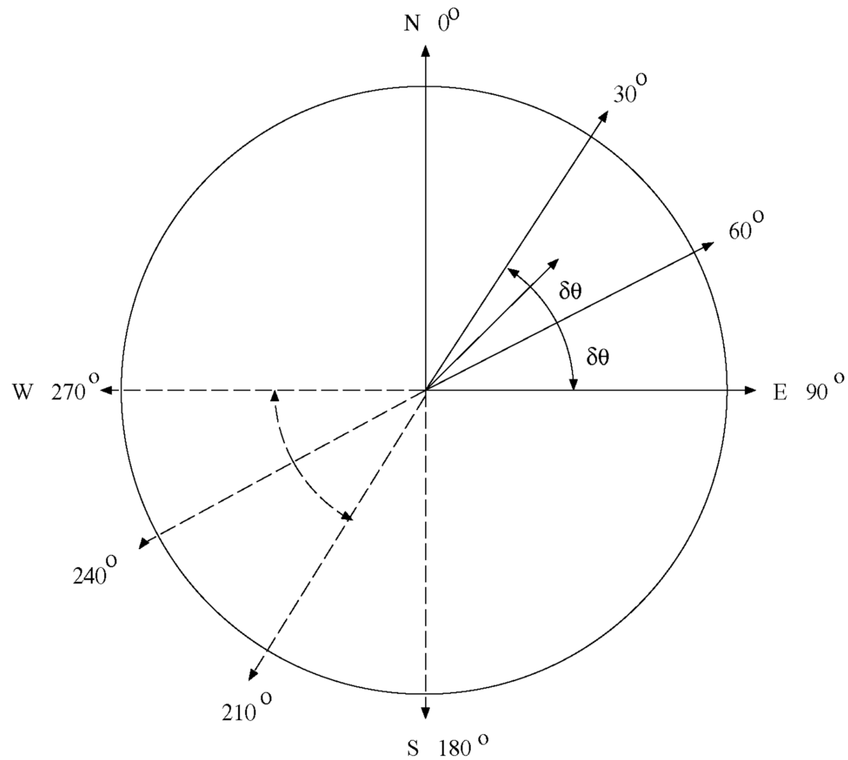
Suppose you specify NDIR=3 in the COMPUTE statement in PROC VARIOGRAM. This results in three angle classes defined by midpoint angles between 0° and 180°: $0° \pm \delta\theta$, $60° \pm \delta\theta$, and $120° \pm \delta\theta$, where $\delta\theta$ is the angle tolerance. If you do not specify an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

$$\delta\theta = \frac{180°}{2 \times \text{NDIR}}$$

For example, if NDIR=3, the default angle tolerance is $\delta\theta = 30°$. When the directed line segment $P_1 P_2$ in Figure 95.11 is superimposed on the coordinate system showing the angle classes, its angle is approximately 45°, measured clockwise from north. In particular, it falls within $[60° - \delta\theta, 60° + \delta\theta) = [30°, 90°)$, the second angle class (Figure 95.12).

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

Figure 95.12 Selected Pair $P_1 P_2$ Falls within the Second Angle Class



If you specify an angle tolerance less than the default, such as ATOL=15°, some point pairs might be excluded. For example, the selected point pair $P_1 P_2$ in Figure 95.12, while closest to the 60° axis, might lie outside $[60 - \delta\theta, 60 + \delta\theta) = [45^\circ, 75^\circ)$. In this case, the point pair $P_1 P_2$ would be excluded from the semivariance computation. This setting can be desirable if you want to reduce interference between neighboring angles. An angle tolerance that is too small might result in too few point pairs in some distance classes for the empirical semivariance estimation (see also the discussion in the section "Choosing the Size of Classes" on page 7547).

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 classes. This has a smoothing effect on the variogram and is useful when there is a small amount of data or the available data are sparsely located. However, in cases of anisotropy the smoothing effect might have the side effect of amplifying weaker anisotropy in some direction and weakening stronger anisotropy in another (Deutsch and Journel 1992, p. 59).

Changes in the values of the BANDW= option have a similar effect. See the section "Bandwidth Restriction" on page 7544 for an explanation of how BANDW= functions.

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 use the DIRECTIONS statement, you should also specify the angle tolerance, too. The default value of the angle tolerance is 45° when a DIRECTIONS statement is used instead of the NDIRECTIONS= option in the COMPUTE statement. This might not be appropriate for a particular set of angle classes. See the section "DIRECTIONS Statement" on page 7531 for more details.

Distance Classification

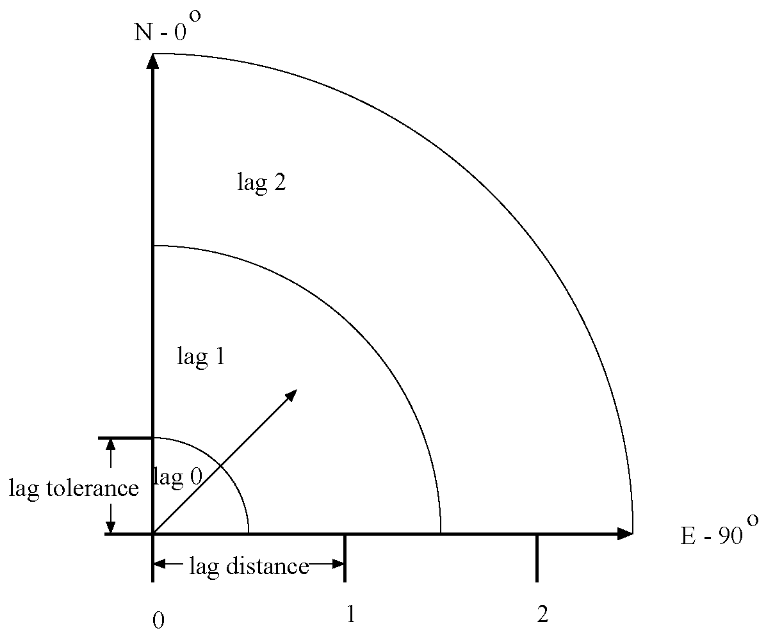
The distance class for a point pair $P_1 P_2$ is determined as follows. The directed line segment $P_1 P_2$ is superimposed on the coordinate system showing the distance or lag classes. These classes are determined by the LAGDISTANCE= option in the COMPUTE statement. Denoting the length of the line segment by $\mid P_1 P_2 \mid$ and the LAGDISTANCE= value by Δ, the lag class L is determined by

$$L(P_1 P_2) = \left\lfloor \frac{\mid P_1 P_2 \mid + 0.5}{\Delta} \right\rfloor$$

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

When the directed line segment $P_1 P_2$ is superimposed on the coordinate system showing the distance classes, it is seen to fall in the first lag class; see Figure 95.13 for an illustration for $\Delta = 1$.

Figure 95.13 Selected Pair $P_1 P_2$ Falls within the First Lag Class

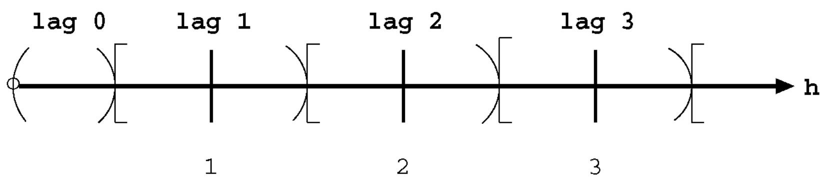


Because pairwise distances are positive, lag class zero is smaller than lag classes $1, \cdots,$ MAXLAG-1. For example, if you specify LAGDISTANCE=1 and MAXLAG=10, and you do not specify a LAGTOL= value in the COMPUTE statement in PROC VARIOGRAM, the 10 lag classes generated by the preceding equation are

$$[0, 0.5), [0.5, 1.5), [1.5, 2.5), \cdots, [8.5, 9.5)$$

This is because the default lag tolerance is half the LAGDISTANCE= value, resulting in no gaps between the distance class intervals. This is shown in Figure 95.14.

Figure 95.14 Lag Distance Axis Showing Lag Classes



On the other hand, if you do specify a distance tolerance with the LAGTOL= 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 LAGDISTANCE=1 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 $P_1 P_2$ might fall in an interval such as

$$| P_1 P_2 | \in [1.25, 1.75)$$

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

In the section "Computation of the Distribution Distance Classes" on page 7545 there is a more extensive discussion of practical aspects in the specification of the LAGDISTANCE= and MAXLAGS= options.

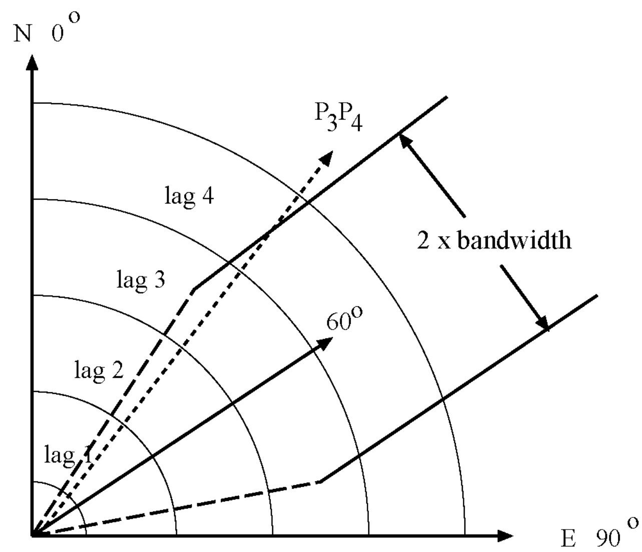
Bandwidth Restriction

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 (Deutsch 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 95.11 and are superimposed on the grid defining distance and angle classes, as shown in Figure 95.15.

The endpoint of vector $P_3 P_4$ falls within the angle class around 60° and the 5th lag class; however, it falls outside the restricted area defined by the bandwidth. Hence, it is excluded from the semivariance calculation.

Figure 95.15 Selected Pair $P_3 P_4$ Falls Outside Bandwidth Limit



Finally, a pair $P_i P_j$ that falls in a lag class larger than the value of the MAXLAGS= option is excluded from the semivariance calculation.

The BANDW= option complements the angle and lag tolerances in determining how point pairs are included in distance classes. Clearly, the number of pairs within each angle/distance class is strongly affected by the angle and lag tolerances and whether BANDW= has been specified. See also the section "Angle Classification" on page 7541 for more details about the effects these rules can have, since BANDW= operates in a manner similar to the ATOL= option.

Computation of the Distribution Distance Classes

This section deals with theoretical considerations and practical aspects when you specify the LAGDISTANCE= and MAXLAGS= options. In principle, these values depend on the amount and spatial distribution of your experimental data.

The value of the LAGDISTANCE= option regulates how many pairs of data are contained within each distance class. In effect, this information defines the pairwise distance distribution (see the following subsection). Your choice of MAXLAGS= specifies how many of these lags you want to include in the empirical semivariogram computation. Adjusting the values of these parameters is a crucial part of your analysis. Based on your observations sample, they determine whether you have sufficient points for a descriptive empirical semivariogram, and they can affect the accuracy of the estimated semivariance, too.

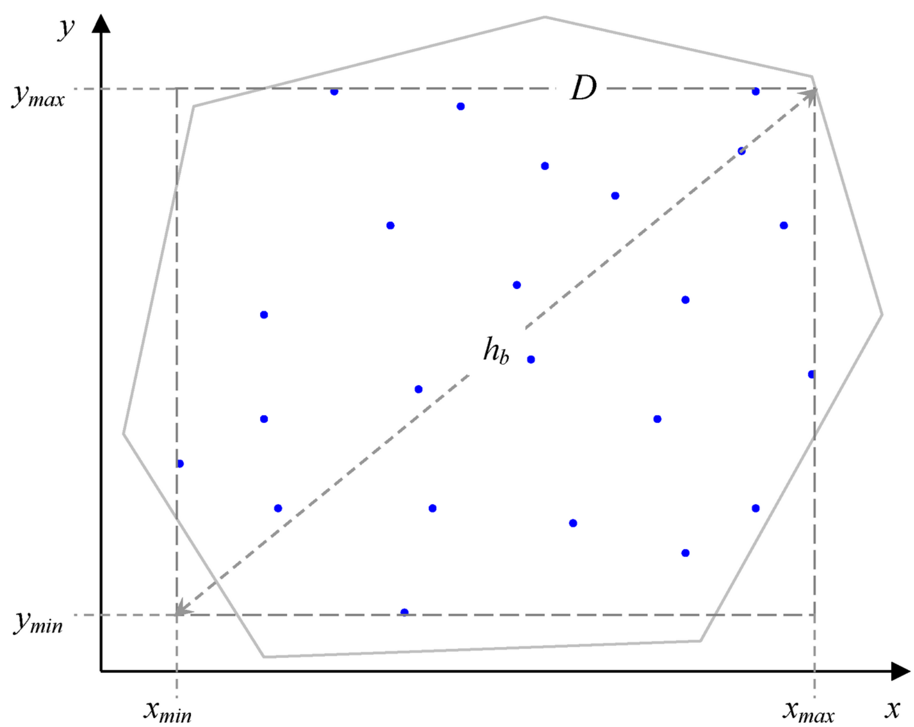
The simplest way of determining the distribution of pairwise distances is to determine the maximum distance h_{max} between any pair of points in your data, and then to divide this distance by some number N of intervals to produce distance classes of length $\delta = h_{max}/N$. The distance $\mid P_1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

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

See Figure 95.16 for an illustration of the bounding rectangle applied to the data of the domain D in Figure 95.11. PROC VARIOGRAM provides you with the sizes of $x_{max} - x_{min}$, $y_{max} - y_{min}$, and h_b. For example, in Figure 95.4 in the preliminary analysis, the specified parameters named "Max Data Distance in East," "Max Data Distance in North," and "Max Data Distance" correspond to the lengths $x_{max} - x_{min}$, $y_{max} - y_{min}$, and h_b, respectively.

Figure 95.16 Bounding Rectangle to Determine Maximum Pairwise Distance in Domain D



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

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

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 (default) 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 $(0, h_0/2)$ and corresponds to the 0th lag. It is half the length of the remaining subintervals, and it often contains the smallest number of pairs. Figure 95.14 shows an example where the lag classes correspond to h_0 =1. This method of partitioning the interval $(0, h_b]$ is used in the empirical semivariogram computation.

Choosing the Size of Classes

When you start with a data sample, the VARIOGRAM procedure computes all the distinct point pairs in the sample. The OUTPAIR= output data set, described in the section "OUTPAIR=*SAS-data-set*" on page 7558, contains information about these pairs. The point pairs are then categorized in classes. The size of each class depends on the common distance that separates consecutive classes. In PROC VARIOGRAM you need to provide this distance value with the LAGDISTANCE= option. Practically, you can define the distance between classes to be about the size of the average sampling distance (Olea 2006).

Under a more scrutinized approach, before you specify a value for the LAGDISTANCE= option, it is helpful to be aware of two issues. First, you should have an estimate of how many classes of data pairs you will need. Each class contributes one point to the empirical semivariogram. Therefore, you need enough classes for an adequate number of points, so that your empirical semivariogram can suggest a suitable theoretical model shape for the description of the spatial continuity. Second, you should keep in mind that a larger number of data pairs in a class can contribute to a more accurate estimate of the corresponding semivariogram point.

The first consideration is a more general issue, and both this and the following subsection address it in detail. Based on the second consideration, the class size problem translates into having a sufficient number of data pairs in each class to produce an accurate semivariance estimate. However, only empirical rules of thumb exist to guide you with this choice. Examples of minimum pairs empirical rules include the suggestion by Journel and Huijbregts (1978, p. 194) to use at least 30 point pairs for each lag class. Also, in a different approach, Chilès and Delfiner (1999, p. 38) increase this number to 50 point pairs.

Obviously, smaller data samples will provide fewer data pairs in the sample. According to Olea (2006), it is difficult to properly estimate a semivariogram with fewer than 50 measurements. The preceding minimum pairs practical rules are useful in cases where small samples are involved. When you work with a relatively small sample, the key is to specify the value of LAGDISTANCE= such that you can strike a balance between the number of the classes you can form and their pairs count. In the coal seam thickness example of the section "Preliminary Spatial Data Analysis" on page 7512, it is not possible to create a desirable large number of classes and maintain an adequate size for each one. On the other hand, there is no practical need to invoke these rules in the case of the much larger sample of ozone concentrations in "Example 95.2: An Anisotropic Case Study with Surface Trend in the Data" on page 7566.

The spatial distribution of the sample might also affect the grouping of pairs into classes. For example, data that are sampled in clusters might prove difficult to classify according to the preceding practical rules. One strategy to address this problem is to accept fewer than 30 pairs for the underpopulated distance classes. Then, at the stage when you determine what theoretical semivariogram model to use, either disregard the corresponding empirical semivariogram points or use them and accept the increased uncertainty.

The VARIOGRAM procedure can help you decide on a suitable class size before you proceed with the empirical semivariogram computation. First, provide a number for the class count by specifying the NHCLASSES= value. Run the procedure with the option NOVARIOGRAM in the COMPUTE statement and examine the distribution data pairs. Use different values of NHCLASSES= to investigate how this will affect the data pairs distribution in each distance class. The pairwise distance intervals table (for example, Figure 95.3) shows the number of pairs in each distance class in the "Number of Pairs" column, and you can use the preceding rule of thumb to adjust the NHCLASSES= value accordingly.

PROC VARIOGRAM displays a rounded value of the distance between the lag bounds as the "Lag Distance" parameter in the pairs information table (see Figure 95.5) or the pairwise distances histogram (see Figure 95.4), which you can use for the LAGDISTANCE= specification. However, this is only one tool. For the semivariogram computation you can specify your own LAGDISTANCE= value based on your experience. Smaller LAGDISTANCE= values result in fewer data pairs in the classes. In that sense, you might find smaller values useful when you work with large samples so that you obtain more semivariogram points. Also, if the LAGDISTANCE= value is too large, you might end up "wasting" too many point pairs in fewer classes at the expense of computing fewer semivariogram points and no significant accuracy gains in the estimation.

As explained earlier, depending on the sample size and its spatial distribution you might have classes with fewer points than what the practical rules advise. Most commonly, the deficient distance classes are the limiting ones close to the origin $\boldsymbol{h} = 0$, as well as the most remote ones at large $\boldsymbol{h}$. The classes near the origin correspond to lags 0 and 1. They are crucial, because the empirical semivariogram in small distances $\boldsymbol{h}$ characterizes the process smoothness and can be further used to detect the presence of a nugget effect. However, as discussed in the section "Distance Classification" on page 7543, lag 0 is half the size of the rest of the classes by definition, so it can be expected to violate the rule of thumb for the number of pairs in a class.

The classes located at higher and extreme distances within a spatial domain are often not accounted for in the empirical semivariogram. The fewer pairs that can be formed in these distances do not allow for an accurate assessment of the spatial correlation, as is explained in the following section.

Spatial Extent of the Empirical Semivariogram

Given your choice for the LAGDISTANCE= value in your spatial domain, the following paragraphs provide guidelines on how many classes to consider when you compute the empirical semivariogram.

Obviously, you want to include no more classes beyond the limit where the pairs count falls below the minimum pairs empirical rule threshold, as discussed in the preceding subsection. PROC VARIOGRAM provides you with a visual way to inspect this upper limit, if you decide to make use of the minimum pairs empirical rule. In particular, specify your threshold choice for the minimum pairs per class by using the THRESHOLD= parameter for the PLOT=PAIRS option.

Then, the procedure produces in the pairwise distances histogram a reference line at the specified THRESHOLD=, which leaves below the line all lags whose pairs count is lower than the threshold value; see, for example, Figure 95.4. The last lag class whose pair population is above the THRESHOLD= is reported in the pairs information table as “Highest Lag With Pairs > Threshold.” Note that this value is not a recommendation for the MAXLAGS= option, but rather is an upper limit for your choice. Detailed information about the pairs count in each class is displayed in the corresponding pairwise distance intervals table, as Figure 95.3 demonstrates.

The preceding suggests that you have an upper limit indication, but you still need some criterion to decide how many lags to include in the semivariogram estimation. The criterion is the extent of spatial dependence in your domain.

Spatial dependence can exist beyond your domain limits. However, you do not have data past your domain scale to define a range for larger-scale spatial dependencies. As you look for pairs of data that are gradually farther apart, the number of pairs naturally decreases with distance. The pairs at the more distant classes might be so few that they are likely to be independent with respect to the spatial dependence scale that you can detect. These will only contribute added noise if you include the largest distances in your empirical semivariogram plot. Note that in the same sense, you cannot explore in detail spatial dependencies in scales smaller than an average minimum distance between your data. The nugget effect is then used to represent microscale correlations whose effect is evident in your working scale.

You specify the spatial dependence extent with commonly used measures such as the *correlation range* (or *correlation length*) ϵ and the *correlation radius* h_c. Both are defined in a similar manner. The correlation range ϵ is the distance at which the covariance is 5% of its value at $\boldsymbol{h} = 0$, and shows that beyond ϵ the covariance is considered to be negligible. The correlation radius h_c is the distance at which the covariance is about half the variance at $\boldsymbol{h} = 0$, and indicates the distance over which significant correlations prevail (Christakos 1992, p. 76). The physical meanings of these measures are similar to that of the semivariogram range. Also, the effective range r_ϵ used in asymptotically increasing semivariance models has essentially the same definition as the correlation range ϵ (see the section “Theoretical Semivariogram Models” on page 7532).

A rough estimate of the correlation extent measures might be available from previous studies of a similar site, or from prior information about related measurements. In such an event, you typically want to consider a maximum pairwise distance that does not exceed the length of two or three correlation radii, or one and a half correlation ranges. You can then specify the MAXLAGS= value on the basis of the lags that fit in that distance.

When there are no estimates of correlation extent measures, you can use first use a crude measure to get started with your analysis: You can typically expect MAXLAGS= to be about half of the lag classes shown in the pairwise distances histogram.

Then, if necessary, you can refine your MAXLAGS= choice by using the following maximum lags rule of thumb: Journel and Huijbregts (1978, p. 194) advise considering lags up to about half of the extreme distance between data in the direction of interest. The VARIOGRAM procedure assists you in this task by providing the overall extreme data distance h_b, as well as the extreme data distances in the vertical and horizontal axes directions. For example, h_b is reported in the pairs information table as “Maximum Data Distance” (see Figure 95.5), and in the pairwise distances histogram as “Max Data Distance” (see Figure 95.4).

Overall, you do not want to deviate significantly from the maximum lags rule of thumb. As was stated earlier, a MAXLAGS= value that takes you well beyond the half-extreme distance between data in a given direction might give you limited accuracy in the empirical semivariance estimates at higher distances. At the other end, a value of MAXLAGS= that is too small might lead you to omit important information about the spatial structure that potentially lies within the range of distances you skipped.

Semivariance Computation

With the classification of a point pair $P_i P_j$ into an angle/distance class, as shown earlier in this section, the semivariance computation proceeds as follows.

Denote all pairs $P_i P_j$ belonging to angle class $[\theta_k - \delta\theta_k, \theta_k + \delta\theta_k)$ and distance class $L = L(P_i P_j)$ as $N(\theta_k, L)$. For example, based on Figure 95.12 and Figure 95.13, $P_1 P_2$ belongs to $N(60°, 1)$.

Let $\mid N(\theta_k, L) \mid$ denote the *number* of such pairs. The component of the standard (or method of moments) semivariance corresponding to angle/distance class $N(\theta_k, L)$ is given by

$$\hat{\gamma}(h_k) = \frac{1}{2 \mid N(\theta_k, L) \mid} \sum_{P_i P_j \in N(\theta_k, L)} [V(\boldsymbol{s}_i) - V(\boldsymbol{s}_j)]^2$$

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 semivariance is given by

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

where

$$\Psi(h_k) = \frac{1}{N(\theta_k, L)} \sum_{P_i P_j \in N(\theta_k, L)} [V(\boldsymbol{s}_i) - V(\boldsymbol{s}_j)]^{\frac{1}{2}}$$

This robust version of the semivariance 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 $h_k, \theta_k, L, N(\theta_k, L), \hat{\gamma}(h)$, and $\bar{\gamma}(h)$.

Empirical Semivariograms and Surface Trends

It was stressed in the beginning of the section "Theoretical and Computational Details of the Semivariogram" on page 7536 that if your data are not free of nonrandom surface trends, then the empirical semivariance $\hat{\gamma}_z(\boldsymbol{h})$ you obtain from PROC VARIOGRAM represents a pseudo-semivariance rather than an estimate of the theoretical semivariance $\gamma_z(\boldsymbol{h})$.

In practice, two major difficulties appear. First, you might have no knowledge of underlying surface trends in your SRF $Z(\boldsymbol{s})$. It can be possible to have this information when you deal with a repetitive phenomenon (Chilès and Delfiner 1999, p. 123), or if you work within a subdomain of a broader region with known characteristics; often, though, this is not the case. Second, even if you suspect the existence of an underlying nonrandom trend, its precise nature might be unknown (Cressie 1993, p. 114, 162).

Based on the last remark, the criteria to define the exact form of a surface trend can be subjective. However, statistical methods can identify the presence and remove an estimate of such a trend. Different trend forms can be estimated in your SRF depending on the trend estimation model that you choose. This choice can lead to different degrees of smoothing in the residual random fluctuations. It might also have an effect on the residuals spatial structure characterization, as trend removals with different models are essentially different operations acting upon the values of your original observations. Note the comment by Chilès and Delfiner (1999, Section 2.7.3) that there are as many semivariograms of residuals as there are ways of estimating the trend. The same source also examines the introduction of bias in the semivariance of the residuals as a side effect of trend removal processes. This bias is small when you examine distances close to the origin $\boldsymbol{h} = 0$, and can increase with distance.

Keeping in mind the preceding remarks, an approach you can take is to use one of the many predictive modeling tools in SAS/STAT software to estimate the unknown trend. Then you use PROC VARIOGRAM to analyze the residuals after you remove the trend. If the resulting model does not require too many degrees of freedom (such as if you use a low-order polynomial), then this approach might be sufficient. The section "Analysis with Surface Trend Removal" on page 7570 demonstrates how to use PROC GLM (see Chapter 39, "The GLM Procedure") for that purpose.

Apart from the standard semivariogram analysis, you can attempt to fit a theoretical semivariogram model to your empirical semivariogram if (a) either the analysis itself or your knowledge of the SRF does not clearly suggest the presence of any surface trend, or (b) the analysis can indicate a potentially trend-free direction, along which your data will have a constant mean.

For example, you might observe overall similar values in your data. This can be an indication that your data are free of nonrandom trends, or that a very mild trend is present. The case falls under the preceding option (a). A very mild trend still allows a good determination of the semivariance at short distances according to Chilès and Delfiner (1999, p. 125), and this can be sufficient for your spatial prediction goal. An analysis of this type is assumed in the section "Preliminary Spatial Data Analysis" on page 7512.

If you observe similar values locally across a particular direction, this an instance of option (b). Olea (2006) suggests recognizing a trend-free direction as being perpendicular to the axis of the maximum dip in the values of $Z(\boldsymbol{s})$. If you suspect that there is at least one such direction in your data, then run PROC VARIOGRAM for a series of directions in the angular vicinity. The trend-free direction, if it exists, will coincide with the one whose pseudo-semivariogram exhibits minimal increase with distance; see “Example 95.3: Analysis without Surface Trend Removal” on page 7579 for a demonstration of this approach. However, you cannot test $Z(\mathbf{s})$ for anisotropy in this case, because you can investigate the semivariogram only in the single trend-free direction (Olea 1999, p. 76). Chilès and Delfiner (1999, Section 2.7.4) suggest fitting a theoretical model in a trend-free direction only if the hypothesis of an isotropic semivariogram appears reasonable in your analysis.

As a result, you need to be very cautious when you choose to perform semivariogram analysis on data you have not previously examined for surface trends. In this event, both of the options (a) and (b) that were reviewed in the preceding paragraphs rely mostly on empirical and subjective criteria. As noted in this section, a degree of subjectivity exists in the selection of the surface trend itself. This fact suggests that a significant part of the semivariogram analysis is based on meta-statistical decisions, as well as on your understanding of your data and the physical considerations that govern your study. In any case, as shown in the section “Theoretical and Computational Details of the Semivariogram” on page 7536, your semivariogram analysis relies fundamentally on the use of trend-free data.

Autocorrelation Statistics (Experimental)

Spatial autocorrelation measures offer you additional insight into the interdependence of spatial data. These measures quantify the correlation of an SRF $Z(\mathbf{s})$ with itself at different locations, and they can be very useful whether you have information at exact locations (point-referenced data) or measurements characterizing an area type such as counties, census tracts, zip codes, etc. (areal data).

As in the semivariogram computation, a key issue for the autocorrelation statistics is that you work with a set z_i of measurements, $i = 1, \ldots, n$, that are free of nonrandom surface trends and have a constant mean.

Autocorrelation Weights

In general, the choice of a weighting scheme is subjective. You can obtain different results by using different schemes, options and parameters. PROC VARIOGRAM offers you considerable flexibility in choosing weights that are appropriate for prior considerations such as different hypotheses about neighboring areas, definition of the neighborhood structure, and accounting for natural barriers or other spatial characteristics; see the discussion in Cliff and Ord (1981, p. 17). As stressed for all types of spatial analysis, it is important to have good knowledge of your data. In the autocorrelation statistics, this knowledge will help you avoid spurious correlations when you choose the weights.

The starting point is to assign individual weights to each one of the n data values z_i, $i = 1, \ldots, n$, with respect to the rest. An $n \times n$ matrix of weights is thus defined, such that for any two locations $\mathbf{s}_i$ and $\mathbf{s}_j$, the weight w_{ij} denotes the effect of the value z_i at location $\mathbf{s}_i$ on the value z_j at location $\mathbf{s}_j$. Depending on the nature of your study, the weights w_{ij} need not be symmetric; that is, it can be that $w_{ij} \neq w_{ji}$.

Binary and Nonbinary Weights

The weights w_{ij} can be either binary or nonbinary values. Binary values of 1 or 0 are assigned if the SRF $Z(\boldsymbol{s}_i)$ at one location $\boldsymbol{s}_i$ is deemed to be connected or not, respectively, to its value $Z(\boldsymbol{s}_j)$ at another location $\boldsymbol{s}_j$. Nonbinary values can be used in the presence of more refined measures of connectivity between any two data points P_i and P_j. PROC VARIOGRAM offers a choice between a binary and a distance-based nonbinary weighting scheme.

In the binary weighting scheme the weight $w_{ij} = 1$, if the data pair at $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ is closer than the user-defined distance LAGDISTANCE=, and $w_{ij} = 0$, if $i = j$ or in any other case. For that reason, in the COMPUTE statement, if you specify the WEIGHTS=BINARY suboption of the AUTOCORRELATION option when the NOVARIOGRAM option is also specified, then you must also specify the LAGDISTANCE= option.

The nonbinary weighting scheme is based on the pair distances and is invoked with the WEIGHTS=DISTANCE suboption of the AUTOCORRELATION option. PROC VARIOGRAM uses a variation of the Pareto form functional to set the weights. Namely, the autocorrelation weight for every point pair P_i and P_j located at $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, respectively, is defined as

$$w_{ij} = s\frac{1}{1+ \mid \boldsymbol{h} \mid^p}$$

where $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$, and $p \geq 0$, $s \geq 0$ are user-defined parameters for the adjustment of the weights.

In particular, the power parameter p is specified in the POWER= option of the DISTANCE suboption within the AUTOCORRELATION option. The default value for this parameter is $p = 1$. Also, the scaling parameter s is specified by the SCALE= option in the DISTANCE suboption of the AUTOCORRELATION option. The default value for the scaling parameter is $s = 1$. You can use the p and s parameters to adjust the actual values of the weights according to your needs. Note that variations in the scaling parameter s do not affect the computed values of the Moran's I and Geary's c autocorrelation coefficients that are introduced in the section "Autocorrelation Statistics Types" on page 7554.

Nonbinary Weights with Normalized Distances

PROC VARIOGRAM offers additional flexibility in the DISTANCE weighting scheme through an option to use normalized pair distances. You can invoke this feature by specifying the NORMALIZE option in the DISTANCE suboption of the AUTOCORRELATION option. In this case, the distances used in the definition of the weights are normalized by the maximum pairwise distance h_b (see the section "Computation of the Distribution Distance Classes" on page 7545 and Figure 95.16); the weights are then defined as $w_{ij} = s/[1 + (\mid \boldsymbol{h} \mid / h_b)^p]$.

Note that h_b most likely has a different value for different data sets. Hence, it is suggested that you avoid using the weights you obtain from the preceding equation and a data set for comparisons with the weights you derive from different data sets.

Symmetric and Asymmetric Weights

The weighting schemes presented in the preceding paragraphs are symmetric; that is, $w_{ij} = w_{ji}$ for every data pair at locations $\mathbf{s}_i$ and $\mathbf{s}_j$. However, you can also define asymmetric weights w'_{ij} such that

$$\sum_{j \in J} w'_{ij} = 1$$

for $i = 1, 2, \cdots, n$, where $w'_{ij} = w_{ij} / \sum_{j \in J} w_{ij}$, $i = 1, 2, \cdots, n$. In the distance-based scheme, J is the set of all locations that form point pairs with the point at $\mathbf{s}_i$. In the binary scheme, J is the set of the locations that are connected to $\mathbf{s}_i$ based on your selection of the LAGDISTANCE= option; see Cliff and Ord (1981, p. 18). The weights w'_{ij} are *row-averaged* (or *standardized* by the count of their connected neighbors). You can apply row averaging in weights when you specify the ROWAVG option within either the BINARY or DISTANCE suboptions in the AUTOCORRELATION option.

Autocorrelation Statistics Types

One measure of spatial autocorrelation provided by PROC VARIOGRAM is Moran's I statistic, which was introduced by Moran (1950) and is defined as

$$I = \frac{n}{(n-1)S^2 W} \sum_i \sum_j w_{ij} v_i v_j$$

where $S^2 = (n-1)^{-1} \sum_i v_i^2$, and $W = \sum_i \sum_{j \neq i} w_{ij}$.

Another measure of spatial autocorrelation in PROC VARIOGRAM is Geary's c statistic (Geary 1954), defined as

$$c = \frac{1}{2S^2 W} \sum_i \sum_j w_{ij} (z_i - z_j)^2$$

Note that Moran's I coefficient makes use of the centered variable, whereas the Geary's c expression uses the noncentered values in the summation.

Inference on these two statistic types comes from approximate tests based on the asymptotic distribution of I and c, which both tend to a normal distribution as n increases. To this end, PROC VARIOGRAM calculates the means and variances of I and c. The outcome depends on the assumption made regarding the distribution $Z(\mathbf{s})$. In particular, you can choose to investigate any of the statistics under the *normality* (also known as *Gaussianity*) or the *randomization* assumption. Cliff and Ord (1981) provided the equations for the means and variances of the I and c distributions, as described in the following.

The normality assumption asserts that the random field $Z(\boldsymbol{s})$ follows a normal distribution of constant mean ($\bar{Z}$) and variance, from which the z_i values are drawn. In this case, the I statistics yield

$$\mathrm{E}_g[I] = -\frac{1}{n-1}$$

and

$$\mathrm{E}_g[I^2] = \frac{1}{(n+1)(n-1)W^2}(n^2 S_1 - nS_2 + 3W^2)$$

where $S_1 = 0.5\sum_i \sum_{j \neq i}(w_{ij} + w_{ji})^2$ and $S_2 = \sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2$. The corresponding moments for the c statistics are

$$\mathrm{E}_g[c] = 1$$

and

$$\mathrm{Var}_g[c] = \frac{(2S_1 + S_2)(n-1) - 4W^2}{2(n+1)W^2}$$

According to the randomization assumption, the I and c observations are considered in relation to all the different values that I and c could take, respectively, if the n z_i values were repeatedly randomly permuted around the domain D. The moments for the I statistics are now

$$\mathrm{E}_r[I] = -\frac{1}{n-1}$$

and

$$\mathrm{E}_r[I^2] = \frac{A_1 + A_2}{(n-1)(n-2)(n-3)W^2}$$

where $A_1 = n[(n^2 - 3n + 3)S_1 - nS_2 + 3W^2]$, $A_2 = -b_2[n(n-1)S_1 - 2nS_2 + 6W^2]$. The factor $b_2 = m_4/(m_2{}^2)$ is the coefficient of kurtosis that uses the sample moments $m_k = \frac{1}{n}\sum_i v_i^k$ for $k = 2, 4$. Finally, the c statistics under the randomization assumption are given by

$$\mathrm{E}_r[c] = 1$$

and

$$\mathrm{Var}_r[c] = \frac{B_1 + B_2 + B_3}{n(n-2)(n-3)W^2}$$

with $B_1 = (n-1)S_1[n^2 - 3n + 3 - (n-1)b_2]$, $B_2 = -\frac{1}{4}(n-1)S_2[n^2 + 3n - 6 - (n^2 - n + 2)b_2]$, and $B_3 = W^2[n^2 - 3 - b_2(n-1)^2]$.

Note that when you specify LAGDISTANCE= to be larger than the maximum data distance in your domain, the binary weighting scheme used by the VARIOGRAM procedure leads to all weights $w_{ij} = 1, i \neq j$. In this extreme case the preceding definitions can show that the variances of the I and c statistics become zero under either the normality or the randomization assumption.

A similar effect might occur when you have collocated observations (see the section "Pair Formation" on page 7540). The Moran's I and Geary's c statistics allow for the inclusion of such pairs in the computations. Hence, contrary to the semivariance analysis, PROC VARIOGRAM does not exclude pairs of collocated data from the autocorrelation statistics.

Interpretation

For Moran's I coefficient, $I > \mathrm{E}[I]$ indicates positive autocorrelation. Positive autocorrelation suggests that neighboring values $\mathbf{s}_i$ and $\mathbf{s}_j$ tend to have similar feature values z_i and z_j, respectively. When $I < \mathrm{E}[I]$, this is a sign of negative autocorrelation, or dissimilar values at neighboring locations. A measure of strength of the autocorrelation is the size of the absolute difference $| I - \mathrm{E}[I] |$.

Geary's c coefficient interpretation is analogous to that of Moran's I. The only difference is that $c > \mathrm{E}[c]$ indicates negative autocorrelation and dissimilarity, whereas $c < \mathrm{E}[c]$ signifies positive autocorrelation and similarity of values.

The VARIOGRAM procedure uses the mathematical definitions in the preceding section to provide the observed and expected values, and the standard deviation of the autocorrelation coefficients in the autocorrelation statistics table. The Z scores for each type of statistics are computed as follows:

$$Z_I = \frac{I - \mathrm{E}[I]}{\sqrt{\mathrm{Var}[I]}}$$

for Moran's I coefficient, and

$$Z_c = \frac{c - \mathrm{E}[c]}{\sqrt{\mathrm{Var}[c]}}$$

for Geary's c coefficient. PROC VARIOGRAM also reports the two-sided p-value for each coefficient under the null hypothesis that the sample values are not autocorrelated. Smaller p-values correspond to stronger autocorrelation for both the I and c statistics. However, the p-value does not tell you whether the autocorrelation is positive or negative. Based on the preceding remarks, you have positive autocorrelation when $Z_I > 0$ or $Z_c < 0$, and you have negative autocorrelation when $Z_I < 0$ or $Z_c > 0$.

Computational Resources

The fundamental computation of the VARIOGRAM procedure is binning: for each pair of observations in the input data set, a distance class and an angle class are 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.

Output Data Sets

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

OUTACWEIGHTS=*SAS-data-set*

The OUTACWEIGHTS= data set contains one observation for each pair of points P_1, P_2 in the original data set, where P_1 is different from P_2, with information about the data distance and autocorrelation weight of each point pair.

Note that the OUTACWEIGHTS= data set can be very large, even for a moderately sized DATA= data set. For example, if the DATA= data set has NOBS=500, then the OUTACWEIGHTS= data set has NOBS$(\text{NOBS} - 1)/2$=124,750 observations.

When you perform autocorrelation computations, the OUTACWEIGHTS= data set is a practical way to save the autocorrelation weights for further use.

The OUTACWEIGHTS= data set contains the following variables:

- ACWGHT12, the autocorrelation weight for the pair P_1, P_2
- ACWGHT21, the autocorrelation weight for the pair P_2, P_1
- DISTANCE, the distance between the data in the pair

- V1, the variable value for the first point in the pair
- V2, the variable value for the second point in the pair
- VARNAME, the variable name for the current VAR= variable
- X1, the x coordinate of the first point in the pair
- X2, the x coordinate of the second point in the pair
- Y1, the y coordinate of the first point in the pair
- Y2, the y coordinate of the second point in the pair

When the autocorrelation weights are symmetric, the pair P_1, P_2 has the same weight as the pair P_2, P_1. For this reason, in the case of symmetric weights the OUTACWEIGHTS= data set contains only the autocorrelation weights ACWGHT12.

OUTDIST=*SAS-data-set*

The OUTDIST= data set contains counts for a modified histogram showing the distribution of pairwise distances. This data set provides you with information related to the choice of values for the LAGDISTANCE= option in the COMPUTE statement.

To request an OUTDIST= data set, specify the OUTDIST= data set in the PROC VARIOGRAM statement and the NOVARIOGRAM option in the COMPUTE statement. The NOVARIOGRAM option prevents any semivariogram or covariance computation from being performed.

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

- COUNT, the number of pairs falling into this lag class
- LAG, the lag class value
- LB, the lower bound of the lag class interval
- UB, the upper bound of the lag class interval
- PER, the percent of all pairs falling in this lag class
- VARNAME, 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 $\mid P_1 P_2 \mid \leq D_{max}$ 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, then the OUTPAIR= data set has NOBS(NOBS − 1)/2=124,750 if no OUTPDISTANCE= restriction is given in the COMPUTE statement.

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

The OUTPAIR= data set contains the following variables:

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

OUTVAR=*SAS-data-set*

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

The OUTVAR= data set contains the following variables:

- ANGLE, the angle class value (clockwise from N to S)
- ATOL, the angle tolerance for the lag/angle class
- AVERAGE, the average variable value for the lag/angle class
- BANDW, the bandwidth for the lag/angle class
- COUNT, the number of pairs in the lag/angle class
- COVAR, the covariance value for the lag/angle class
- DISTANCE, the average lag distance for the lag/angle class

- LAG, the lag class value (in LAGDISTANCE= units)
- RVARIO, the sample robust semivariance value for the lag/angle class
- VARIOG, the sample semivariance value for the lag/angle class
- VARNAME, the name of the current VAR= variable

The robust semivariance estimate, RVARIO, is not included in the data set if the user has not specified the option ROBUST in the DIRECTIONS statement.

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.

The OUTVAR= data set contains a line where the LAG variable is -1. The AVERAGE variable in this line displays the sample mean value $\bar{Z}$ of the SRF $Z(\boldsymbol{s})$, and the COVAR variable shows the sample variance $\mathrm{Var}[Z(\boldsymbol{s})]$.

Displayed Output

In addition to the output data sets, the VARIOGRAM procedure produces several output objects. Most of those are produced depending on whether you specify either NOVARIOGRAM or LAGDISTANCE= and MAXLAGS= in the COMPUTE statement. The VARIOGRAM procedure output objects are the following:

- a default "Number of Observations" table that displays the number of observations read from the input data set and the number of observations used in the analysis
- a default map showing the spatial distribution of the observations of the current variable in the VAR statement. The observations are displayed by default with circled markers whose color indicates the VAR value at the corresponding location.
- a table with basic information about the lags and the extreme distance between data pairs, when NOVARIOGRAM is specified
- a table that describes the distribution of data pairs in distance intervals, when NOVARIOGRAM is specified
- a histogram plot of the pairwise distance distribution, when NOVARIOGRAM is specified). The plot also displays a reference line at a user-specified pairs frequency threshold when you specify the THRESHOLD= parameter in the PLOT=PAIRS option. The option PLOT=PAIRS(NOINSET) forces the informational inset that appears in the plot to hide.
- empirical semivariogram details, when NOVARIOGRAM is not specified and LAGDISTANCE= and MAXLAGS= are specified. This table also includes the semivariance estimate variance and confidence limits when CL is specified, and estimates of the robust semivariance when ROBUST is specified.

- plots of the appropriate empirical semivariograms, when NOVARIOGRAM is not specified and LAGDISTANCE= and MAXLAGS= are specified. If you perform the analysis in more than one direction simultaneously, the output is a panel comprising the empirical semivariogram plots for the specified angles. If the semivariograms are nonpaneled, then each plot includes in the lower part a needle plot of the contributing pairs distribution.
- a table that provides autocorrelation statistics, when the options AUTOCORRELATION and LAGDISTANCE= are specified

ODS Table Names

Each table created by PROC VARIOGRAM has a name associated with it, and you must use this name to reference the table when using ODS Graphics. These names are listed in Table 95.4.

Table 95.4 ODS Tables Produced by PROC VARIOGRAM

ODS Table Name	Description	Statement	Option
AutoCorrStats	Autocorrelation statistics information	COMPUTE	AUTOCORR
DistanceIntervals	Pairwise distances matrix	COMPUTE	NOVARIOGRAM
NObs	Number of observations read and used	PROC	default output
PairsInformation	General information about the pairs distribution in classes and data maximum distances in selected directions	COMPUTE	NOVARIOGRAM
SemivariogramTable	Empirical semivariance classes, parameters, and estimates	COMPUTE	LAGD=, MAXLAGS=

ODS Graphics

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

To request these graphs, you must specify the ODS GRAPHICS statement. For additional control of the graphics that are displayed, see the PLOTS= option in the section "PROC VARIOGRAM Statement" on page 7521. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

ODS Graph Names

PROC VARIOGRAM assigns a name to each graph it creates by using ODS Graphics. You can use these names to reference the graphs when using ODS Graphics. The names are listed in Table 95.5.

Table 95.5 ODS Graphics Produced by PROC VARIOGRAM

ODS Graph Name	Plot Description	Statement	Option
ObservationsPlot	Scatter plot of observed data and colored markers indicating observed values	PROC	PLOTS=OBSERV
PairDistPlot	Histogram of the pairwise distance distribution	PROC	PLOTS=PAIRS
Semivariogram	Plots of empirical classical and robust (optional) semivariograms	PROC	PLOTS=SEMIVAR
SemivariogramPanel	Panel of empirical classical and robust (optional) semivariogram plots	PROC	PLOTS=SEMIVAR

To request these graphs, you must specify the ODS GRAPHICS statement in addition to the statements indicated in Table 95.5. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

Examples: VARIOGRAM Procedure

Example 95.1: Theoretical Semivariogram Model Fitting

This example continues the introduction study presented in the section "Getting Started: VARIOGRAM Procedure" on page 7511 by fitting a theoretical semivariogram model to the estimated classical sample semivariogram in Figure 95.8. You will use PROC NLIN to perform the model fit and to compare two different approaches: the ordinary least squares (OLS) and the weighted least squares (WLS) fitting methods.

A review of the coal seam thickness empirical semivariogram in Figure 95.8 shows first a slow, then rapid rise from the origin, suggesting a Gaussian-type form:

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

as shown in the section "Theoretical Semivariogram Models" on page 7532.

By experimentation, you find that a sill of $c_0 = 7.5$ and a range of $a_0 = 30,000$ feet (effective range $r_\epsilon = \sqrt{3}a_0 \approx 52,000$ feet) provide a reasonable fit of the preceding semivariogram model. You can use these values as initial guesses in the least squares fitting process. For the fitting you need the output data set that is produced by the parameters used in the section "Theoretical Semivariogram Models" on page 7532. The following statements skip the autocorrelation statistics and the creation of plots, and produce the required output data for the semivariogram table. Note that for the fitting process you need information about the semivariance variance, which you obtain by specifying the CL option in the COMPUTE statement.

```
title 'Theoretical Semivariogram Model Fitting Example';
ods graphics on;

proc variogram data=thick outv=outv plots=none;
   compute lagd=7 maxlag=10 robust cl;
   coordinates xc=East yc=North;
   ods output SemivariogramTable=sv;
   var Thick;
run;
```

Since MAXLAG=10, you computed the empirical semivariogram at 11 points (see also Figure 95.8). You would like to obtain a smooth theoretical semivariogram plot, so you need to estimate the theoretical model at more points on the horizontal (distance) axis. The following statements create a sequence of such distance points from 0 to 70,000 feet and space them 500 feet apart:

```
data pv;
   do Distance = 0 to 70 by 0.5;
      Semivariance = .;
      output;
   end;
run;

data sv; set sv pv; by Distance;
run;
```

PROC NLIN performs its own analysis based on the Gaussian model you provided as input. By invoking the NLIN procedure twice, as shown in the following statements, you obtain the estimates for the theoretical model parameters for the OLS and WLS fitting methods. Notice in the WLS case that the weights are defined as the inverse of the computed semivariance variance.

```
proc nlin data=sv;
   parms Range=30
         Sill=7.5;
   model Semivariance =
         Sill*(1-exp(-Distance*Distance/(Range*Range)));
   output out=OLS p=OLS;
run;

proc nlin data=sv;
   parms Range=30
         Sill=7.5;
   model Semivariance =
         Sill*(1-exp(-Distance*Distance/(Range*Range)));
   _weight_ = 1/SemivarianceStdErr/SemivarianceStdErr;
   output out=WLS p=WLS;
run;
```

Output 95.1.1 shows part of the NLIN procedure output that displays the model parameters statistics for the OLS methods.

Output 95.1.1 Ordinary Least Squares Fitting Parameter Estimates

```
              Theoretical Semivariogram Model Fitting Example

                            The NLIN Procedure

                                  Approx      Approximate 95% Confidence
         Parameter      Estimate   Std Error                Limits

         Range           27.1706      1.9211     22.8247      31.5165
         Sill             7.1628      0.2781      6.5337       7.7919
```

The corresponding PROC NLIN output for the WLS method is displayed in the following Output 95.1.2.

Output 95.1.2 Weighted Least Squares Fitting Parameter Estimates

```
              Theoretical Semivariogram Model Fitting Example

                            The NLIN Procedure

                                  Approx      Approximate 95% Confidence
         Parameter      Estimate   Std Error                Limits

         Range           30.6239      1.7382     26.6917      34.5560
         Sill             7.2881      0.4082      6.3646       8.2116
```

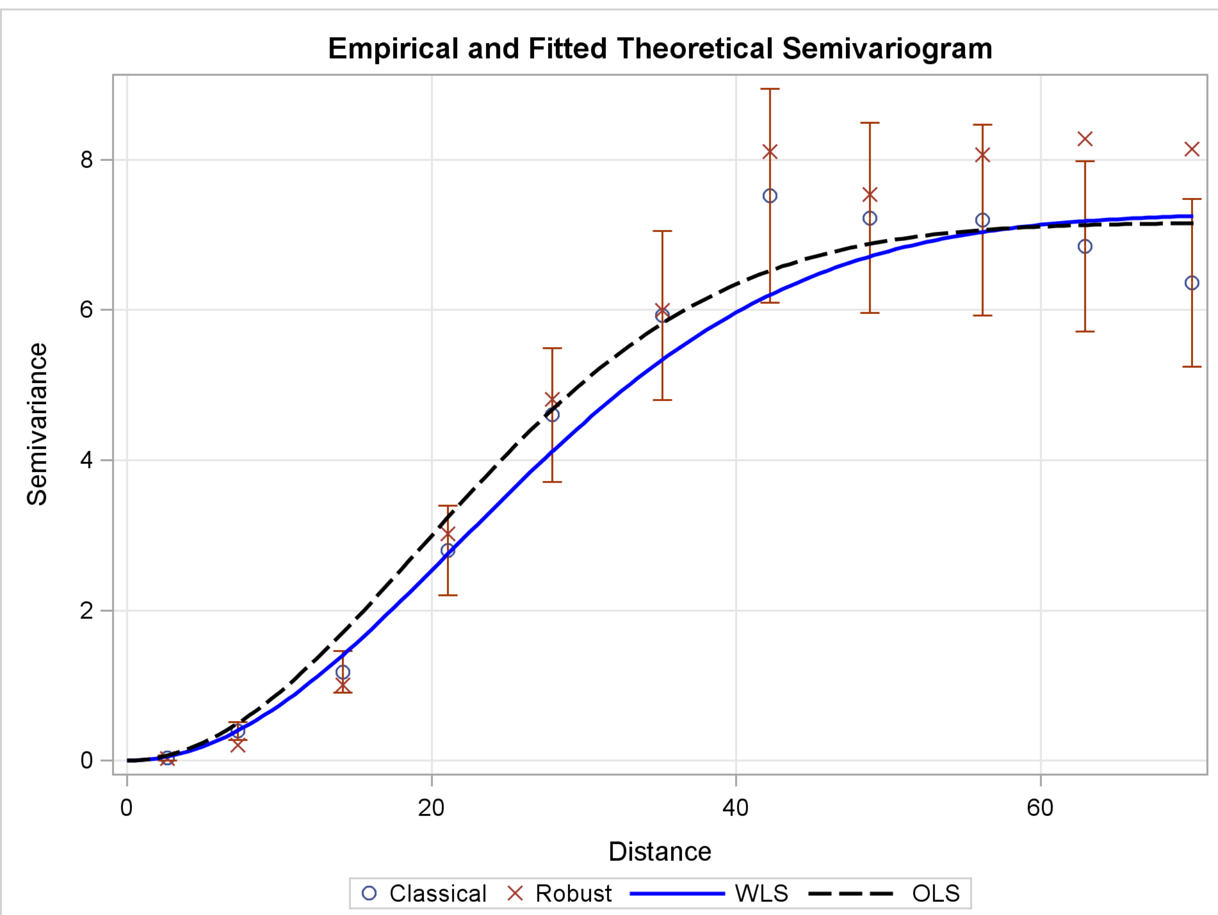
Finally, you visualize the outcome of the fitting analysis. You start with a DATA step to arrange the WLS and OLS data in the same data set. Then you use the SGPLOT procedure to produce a plot showing the empirical and fitted theoretical semivariograms. This sequence is exhibited in the following statements:

```
data pv;
   merge WLS OLS;
run;

proc sgplot data=pv;
   title "Empirical and Fitted Theoretical Semivariogram";
   xaxis label = "Distance" grid;
   yaxis label = "Semivariance" grid;
   scatter y=Semivariance x=Distance /
           markerattrs = GraphData1(symbol=circle)
           name = 'SemiVarClassical'
           yerrorupper = UpperCLSemivariance
           yerrorlower = LowerCLSemivariance;
   scatter y=RobustSemivariance x=Distance /
           markerattrs = GraphData2(symbol=X)
           name = 'SemiVarRobust';
```

```
      series x=Distance y=WLS /
             lineattrs = (thickness=2px color=blue)
             name = 'SemivarPredWgh';
      series x=Distance y=OLS /
             lineattrs = (thickness=2px color=black
                          pattern=MediumDash)
             name = 'SemivarPredUnw';
      discretelegend 'SemiVarClassical' 'SemiVarRobust'
                     'SemivarPredWgh' 'SemivarPredUnw';
   run;

   ods graphics off;
```

Output 95.1.3 demonstrates the difference between the ordinary and weighted least squares fitting results: WLS achieves a more accurate fit closer to the empirical points with the smaller variances, because the weights are expressed as the inverse of these variances. In the presence of a sizable population per distance class, you expect the points with lower variance to be situated close to the origin, as the semivariance variance expression suggests in the section "Theoretical and Computational Details of the Semivariogram" on page 7536. Hence, you expect in general the WLS method to perform with increased accuracy at short distances because it acknowledges the smaller variance at small h. In contrast, the OLS approach performs a least squares overall best fit as it assumes constant variance.

Output 95.1.3 Fitted Theoretical and Empirical Semivariogram for Coal Seam Thickness Data



The WLS method is preferred over the OLS method because it is important to obtain accurate estimates of the spatial continuity closer to the origin $\boldsymbol{h} = 0$. Another advantage of WLS over OLS is that OLS falsely assumes that the differences in the optimization process are normally distributed and independent. However, WLS has the disadvantage that the weights depend on the fitting parameters.

Other fitting methods include maximum likelihood approaches that rely crucially on the normality assumption for the data distribution, and the generalized least squares method, which offers better accuracy but is computationally more demanding. You can find extensive discussions of these issues in Cressie (1993, Section 2.3), Jian, Olea, and Yu (1996), Stein (1988), and Schabenberger and Gotway (2005).

Example 95.2: An Anisotropic Case Study with Surface Trend in the Data

This example shows how to examine data for nonrandom surface trends and anisotropy. You will use simulated data where the variable is atmospheric ozone (O_3) concentrations measured in Dobson units (DU). The coordinates are offsets from a point in the southwest corner of the measurement area, with the east and north distances in units of kilometers (km). You will be working with 300 measurements in a square area of 100 km × 100 km.

The following statements read the data set.

```
title 'Semivariogram in Anisotropic Case With Trend Removal Example';
data ozoneSet;
   input East North Ozone @@;
   datalines;
   34.9 68.2 286  39.2 12.5 270  44.4 37.7 275  90.5 27.0 282
   91.1 40.8 285  98.6 61.6 294  61.8 26.7 281  64.0 11.5 274
   22.4 26.5 274  89.3 18.3 279  32.3 28.3 274  31.1 53.1 279
   43.0 17.5 272  79.3 42.3 283  99.9 57.9 291   1.8 24.1 273
   81.7 73.5 294  22.9 32.0 273  64.9 67.5 292  76.5 56.3 285
   78.7 11.7 276  61.8 99.3 307  49.1 86.6 299  40.0 35.8 273
   69.3  3.8 278  23.4  9.3 270  66.3 94.3 304  71.3  6.5 275
    9.7 54.4 280  85.2 81.7 300  30.3 60.9 284  94.6 94.3 309
   10.6 10.3 271  73.0 43.0 280   4.9 50.7 280  19.0 79.4 289
    2.4 73.1 287  77.7 25.2 278   8.4 27.1 276  93.5 19.7 279
    0.2 34.5 275  50.4 91.3 302  55.7 26.2 279  50.3  2.3 274
   16.3 84.4 293  19.0  6.9 272  57.1 92.3 303  61.0  0.4 275
   10.7 18.7 271  15.2 43.5 277  67.0 87.4 301  79.0 54.0 285
   36.0 53.3 279  58.3 52.1 282  56.6 79.7 294  40.4 32.4 275
   48.9 64.1 286  54.0 54.9 281  27.5 48.5 279  36.4 30.3 275
   10.5 31.0 273  87.0 39.4 283  47.9 37.5 274  64.7 63.4 288
    0.5 90.8 294  22.8 22.4 275  31.1 78.8 291  93.6 49.8 290
    2.5 39.3 273  83.6 25.6 282  49.8 24.1 278  73.1 91.8 305
   30.5 90.6 297  26.0 61.2 284  58.4 66.2 289  30.5  4.3 273
   38.3 85.6 298  89.2 96.6 309  53.4  6.3 275  27.3 12.8 271
   43.4 56.5 281  99.5 86.9 305  85.8 22.8 281  83.0 10.9 278
   24.8 16.7 271  51.1 18.8 275  59.0 54.3 283  35.5 91.4 298
   18.1 56.0 279  78.0 36.4 277  56.8  6.9 275  21.1 44.5 277
```

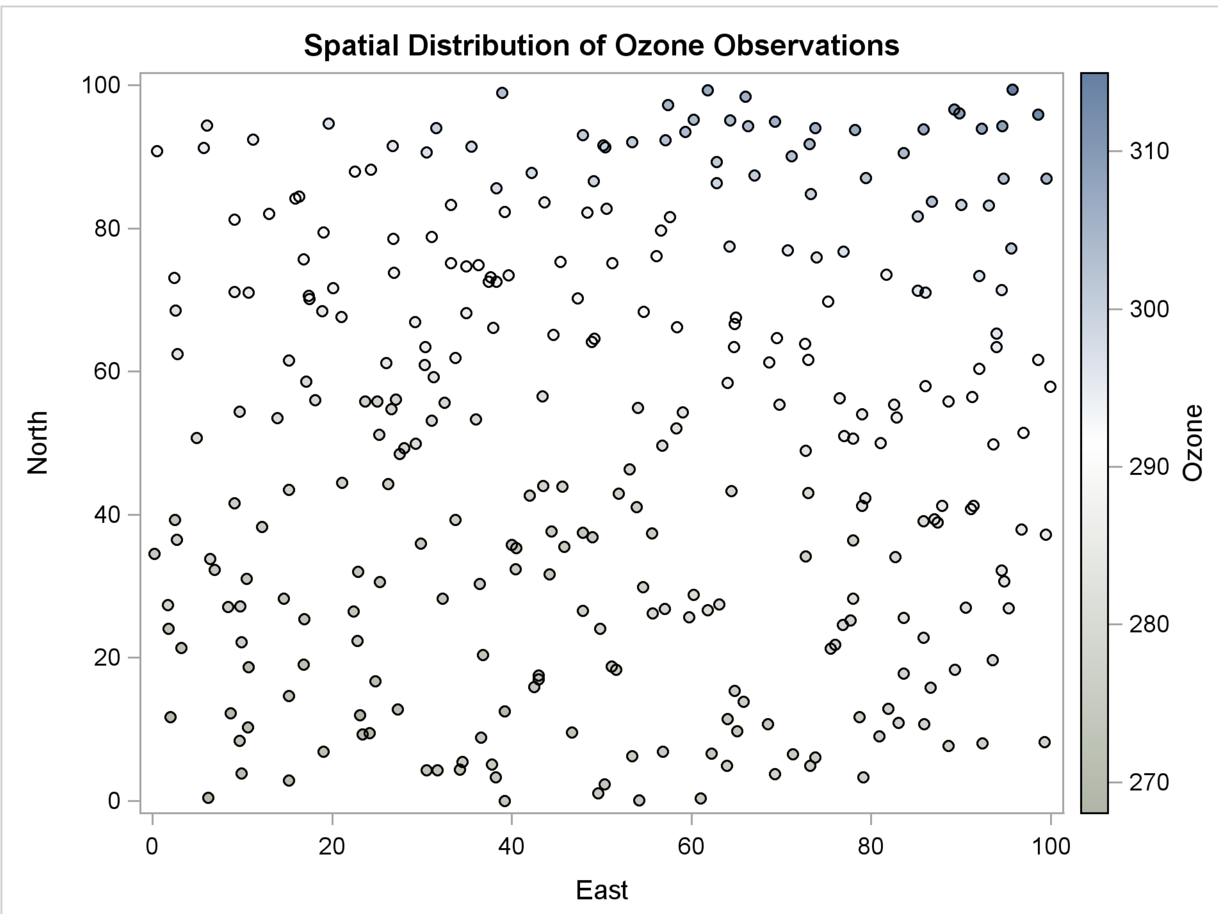
```
   73.9 75.9 296  54.2  0.1 274  33.2 75.1 290  38.2  3.3 274
   15.2 14.7 272  15.9 84.2 292  60.2 95.2 304   9.8 27.2 276
   91.2 56.4 289  94.7 86.9 303  56.7 49.6 281  24.2  9.5 270
   43.0 17.0 272  85.9 10.7 278  53.9 41.1 276  30.4 63.4 286
   62.8 86.3 299  76.8 24.6 279  31.6 94.0 300  26.9 73.8 287
   18.9 68.4 284  99.4 37.2 285  79.1  3.3 277  34.9 74.7 289
    6.4 33.8 277  48.4 82.2 294  86.0 58.0 289  92.0 60.4 293
   50.2 91.6 300  12.2 38.3 275  72.7 48.9 283  82.7 34.1 279
   77.0 51.0 286  86.6 15.8 278  42.0 42.7 277  99.3  8.2 278
   17.4 70.6 286  11.2 92.4 295  60.2 28.8 280  92.0 73.3 297
   25.3 30.6 273  36.6  8.9 274  34.2  4.4 273  26.6 54.7 278
    1.7 27.4 278  49.6  1.1 275  62.8 89.3 301  28.0 49.3 279
   51.2 75.1 293  59.3 93.5 304  83.6 90.5 304  79.4 87.0 302
   78.0 28.3 281  16.8 19.1 272   9.1 81.2 292  23.7 55.8 277
   75.5 21.3 279  64.4 43.3 279  38.9 98.9 303  22.5 87.9 293
   96.7 37.9 285  92.3 93.9 308  16.9 25.4 273  15.2 61.5 283
   73.8 94.0 306  57.4 97.2 305  73.2  4.9 276  39.2 82.3 294
   95.7 99.4 315  66.0 98.4 306  95.3 26.9 283  45.4 75.3 291
   64.8 15.4 276  69.8 55.4 284  36.3 74.9 290   9.9 22.2 276
   65.8 13.9 276  13.0 82.0 293  95.6 77.2 301  32.5 55.6 279
   45.8 35.5 275  62.2  6.6 274  25.2 51.2 279  92.4  8.1 277
   40.5 35.3 273   9.9  3.9 271  43.5 44.0 278  68.6 61.3 287
   64.2 77.5 296  57.6 81.6 294  69.5 64.7 291  64.3 95.1 304
    2.8 62.4 283  33.2 83.3 294  10.7 71.0 285  24.3 88.2 294
   94.5 32.2 283  21.0 67.6 286  20.1 71.6 286  85.2 71.3 296
   94.8 30.7 283  53.4 92.0 301  81.0 50.0 287  54.6 29.9 277
   71.1 90.1 303  15.2  2.9 271  83.6 17.8 278  76.0 21.8 279
   55.6 37.4 275  86.7 83.7 303  43.6 83.6 295  44.2 31.7 274
   90.0 83.3 300   6.2  0.5 270  42.2 87.7 298  31.7  4.3 273
   91.4 41.2 285  78.0 50.6 286  27.1 56.1 278  72.6 63.9 291
   29.3 49.9 281  49.0 36.9 275  13.9 53.5 280  93.1 83.2 300
   73.0 61.6 289  63.1 27.5 280  38.3 72.5 287  72.7 34.2 277
    6.9 32.3 274  17.1 58.6 280  19.6 94.6 297   2.7 36.5 276
   34.5  5.5 275  98.6 95.9 313   9.1 71.1 285  88.6 55.8 287
   26.8 78.5 289  64.8 66.6 292  59.7 25.7 280  47.3 70.2 288
    6.1 94.4 296  50.5 82.7 296   9.1 41.6 276  86.0 71.0 296
   75.2 69.8 293  73.3 84.8 300  42.5 15.9 274  56.1 76.1 292
   87.9 41.2 285  65.1  9.8 274  79.0 41.2 282  44.6 65.1 287
   54.7 68.3 289  57.0 26.8 279   8.7 12.3 270  33.7 61.9 286
   25.0 55.8 278  69.3 94.9 306  49.2 64.6 287  78.2 93.7 307
   47.9 26.6 277  96.9 51.4 292  39.6 73.4 287  37.9 66.1 285
   94.5 71.4 296  51.6 18.3 276  37.6 73.2 287  68.5 10.7 274
   46.7  9.6 273  87.4 38.9 282  45.6 43.9 277  70.7 76.9 296
   82.8 53.6 287  82.5 55.4 286  37.8  5.1 275  89.8 96.1 309
   63.9  4.9 276   2.0 11.7 270  31.3 59.2 282  93.9 65.3 296
   47.9 93.0 301  29.9 36.0 274  14.6 28.3 274  17.5 70.1 286
    2.6 68.5 282  23.1 12.0 268  36.8 20.4 273  80.9  9.0 276
   39.2  0.0 274  26.2 44.3 276  81.9 12.9 277   3.2 21.4 272
   76.9 76.7 297  88.6  7.7 277   9.7  8.4 273  26.7 91.5 296
   73.8  6.1 276  33.7 39.3 276  64.0 58.4 286   5.7 91.2 295
   85.8 93.8 307  85.8 39.1 281  93.9 63.4 295  53.1 46.3 278
   51.9 42.9 277  16.8 75.7 288  29.2 66.9 285  37.4 72.5 287
   ;
run;
```

The initial step is to explore the data set by inspecting the data spatial distribution. Run PROC VARIOGRAM, specifying the NOVARIOGRAM option in the COMPUTE statement as follows:

```
ods graphics on;

proc variogram data=ozoneSet;
   compute novariogram nhc=35;
   coord xc=East yc=North;
   var Ozone;
run;
```

The result is a scatter plot of the observed data shown in Output 95.2.1. The scatter plot suggests an almost uniform spread of the measurements throughout the prediction area. No direct inference can be made about the existence of a surface trend in the data. However, the apparent stratification of ozone values in the northeast–southwest direction might indicate a nonrandom trend.

Output 95.2.1 Ozone Observation Data Scatter Plot



You will need to define the size and count of the data classes by specifying suitable values for the LAGDISTANCE= and MAXLAGS= options, respectively. Compared to the smaller sample of thickness data used in “Example 95.1: Theoretical Semivariogram Model Fitting” on page 7562, the larger size of the ozoneSet data results in more densely populated distance classes for the same value of the NHC= option. Once you experiment with a variety of values for the NHC= option, you can adjust LAGDISTANCE= to have a relatively small number. Then you can account for a large value of MAXLAGS= so that you obtain many sample semivariogram points within your data correlation range. Specifying these values requires some exploration, for which you might need to return to this point from a later stage in your semivariogram analysis. For illustration purposes you now specify NHC=35.

Your choice of NHC=35 yields the pairwise distance intervals table in Output 95.2.2 and the corresponding histogram in Output 95.2.3.

Output 95.2.2 Pairwise Distance Intervals Table

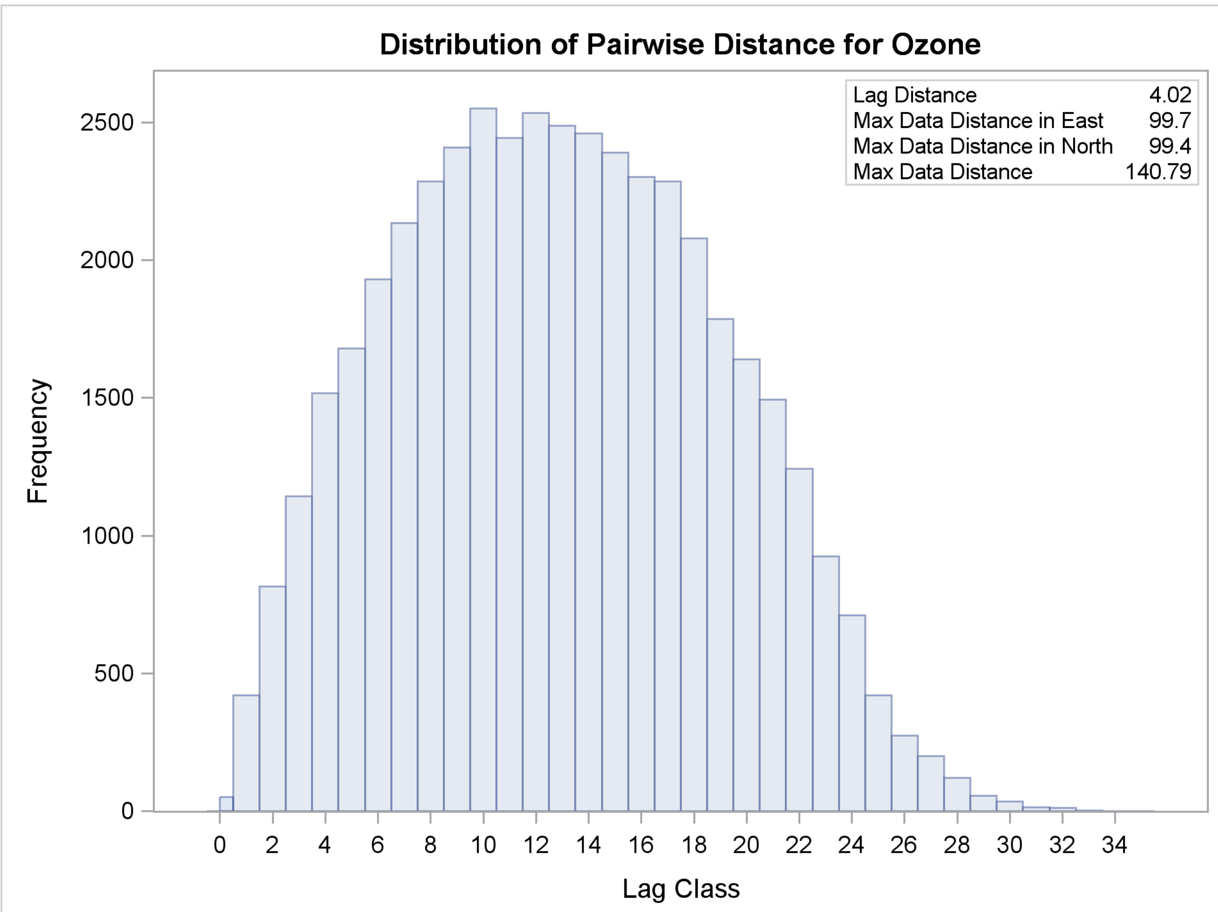
Pairwise Distance Intervals

Lag Class	Bounds		Number of Pairs	Percentage of Pairs
0	0.00	2.01	52	0.12%
1	2.01	6.03	420	0.94%
2	6.03	10.06	815	1.82%
3	10.06	14.08	1143	2.55%
4	14.08	18.10	1518	3.38%
5	18.10	22.12	1680	3.75%
6	22.12	26.15	1931	4.31%
7	26.15	30.17	2135	4.76%
8	30.17	34.19	2285	5.09%
9	34.19	38.21	2408	5.37%
10	38.21	42.24	2551	5.69%
11	42.24	46.26	2444	5.45%
12	46.26	50.28	2535	5.65%
13	50.28	54.30	2487	5.55%
14	54.30	58.33	2460	5.48%
15	58.33	62.35	2391	5.33%
16	62.35	66.37	2302	5.13%
17	66.37	70.39	2285	5.09%
18	70.39	74.41	2079	4.64%
19	74.41	78.44	1786	3.98%
20	78.44	82.46	1640	3.66%
21	82.46	86.48	1493	3.33%
22	86.48	90.50	1243	2.77%
23	90.50	94.53	925	2.06%
24	94.53	98.55	710	1.58%
25	98.55	102.57	421	0.94%
26	102.57	106.59	274	0.61%
27	106.59	110.62	200	0.45%
28	110.62	114.64	120	0.27%
29	114.64	118.66	55	0.12%
30	118.66	122.68	35	0.08%
31	122.68	126.71	14	0.03%
32	126.71	130.73	11	0.02%
33	130.73	134.75	2	0.00%
34	134.75	138.77	0	0.00%
35	138.77	142.80	0	0.00%

Notice the overall high pair count in the majority of classes in Output 95.2.2. You can see that even for higher values of NHC= the classes are still sufficiently populated for your semivariogram analysis according to the rule of thumb stated in the section "Choosing the Size of Classes" on page 7547. Based on the displayed information in Output 95.2.3, you specify LAGDISTANCE=4 km. You can further experiment with smaller lag sizes to obtain more points in your sample semivariogram.

You will return to the MAXLAGS= specification later. The important step now is to investigate the presence of trends in the measurement. The following section makes a suggestion about how to remove surface trends from your data, and then continues the semivariogram analysis with the detrended data.

Output 95.2.3 Distribution of Pairwise Distances for Ozone Observation Data



Analysis with Surface Trend Removal

You can use a SAS/STAT predictive modeling procedure to extract surface trends from your original data. If your goal is spatial prediction, you can continue processing the detrended data for the prediction tasks, and at the end you can reinstate the trend at the prediction locations to report your analysis results.

In general, the exact form of the trend is unknown, as discussed in the section "Empirical Semivariograms and Surface Trends" on page 7551. In this case, the spatial distribution of the measurements shown in Figure 95.2.1 suggests that you can use a quadratic model to describe the surface trend like the one that follows:

$$T(\text{East}, \text{North}) = f_0 + f_1\,[\text{East}] + f_2\,[\text{East}]^2 + f_3\,[\text{North}] + f_4\,[\text{North}]^2$$

The following statements show how to invoke the GLM procedure for your ozone data, and how to extract the preceding trend from them:

```
proc glm data=ozoneSet;
   model ozone = East East*East North North*North;
   output out=gmout predicted=pred residual=ResidualOzone;
run;
```

Among other output, PROC GLM produces estimates for the parameters $f_0, \ldots, f_4$ in the preceding trend model. Output 95.2.4 shows the table with the parameter estimates. In this table, the coefficient f_0 corresponds to the intercept estimate, and the rest of the coefficients correspond to their matching variables; for example, the estimate in the line of "East*East" refers to f_2 in the preceding model. For more information about the syntax and the PROC GLM output, see Chapter 39, "The GLM Procedure."

Output 95.2.4 Parameter Estimates for the Surface Trend Model

Semivariogram in Anisotropic Case With Trend Removal Example

The GLM Procedure

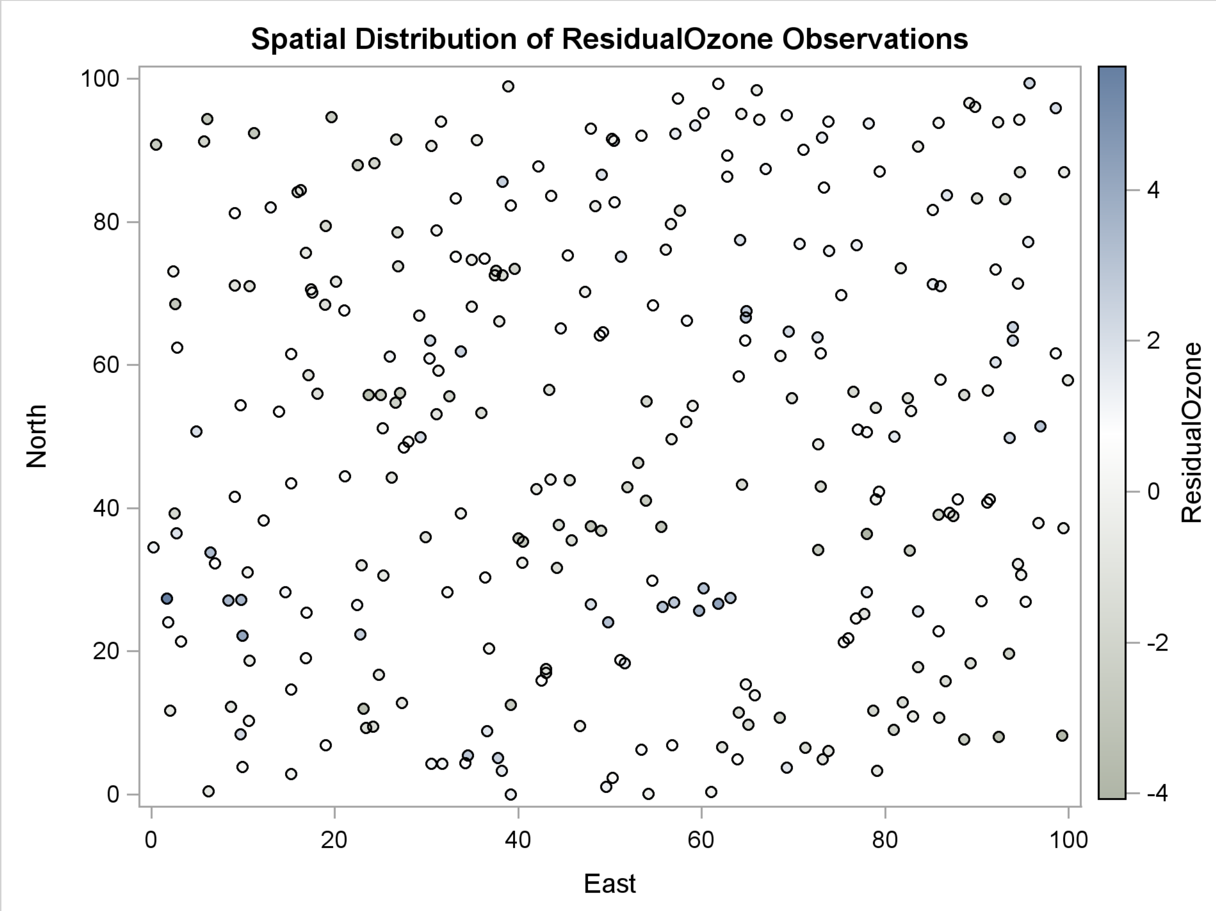
Dependent Variable: Ozone

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	270.6798273	0.40595731	666.77	<.0001
East	0.0065148	0.01360281	0.48	0.6323
East*East	0.0010726	0.00012987	8.26	<.0001
North	-0.0369159	0.01297491	-2.85	0.0047
North*North	0.0035587	0.00012659	28.11	<.0001

The detrending process leaves you with the GMOUT data set, which contains the ResidualOzone data residuals. You can run PROC VARIOGRAM again as follows, with the NOVARIOGRAM option to inspect the detrended residuals. Note that you ask only for the OBSERVATIONS plot, shown in Output 95.2.5.

```
proc variogram data=gmout plots(only)=observ;
  compute novariogram nhc=35;
  coord xc=East yc=North;
  var ResidualOzone;
run;
```

Output 95.2.5 Ozone Residual Observation Data Scatter Plot



Before you proceed with the empirical semivariogram computation and model fitting, examine your data for anisotropy. This investigation is necessary to portray the spatial structure of your SRF accurately. If anisotropy exists, it will manifest itself as different ranges and/or sills for the empirical semivariograms in different directions.

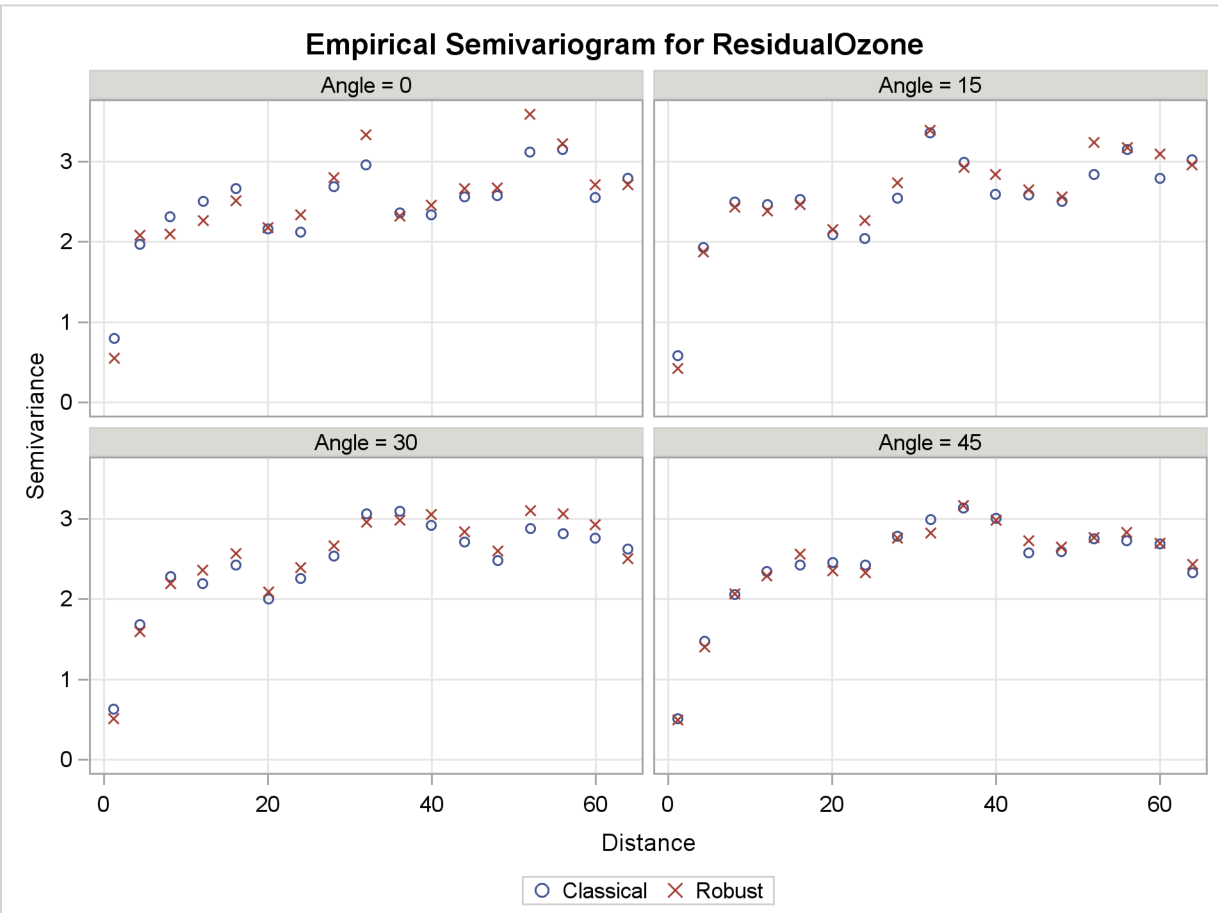
You want detail in your analysis, so you ask for the empirical semivariance in 12 directions by specifying NDIRECTIONS=12. Based on the NDIRECTIONS= option, empirical semivariograms are produced in increments of the base angle $\theta = 180°/12 = 15°$.

You also choose ANGLETOLERANCE=22.5 and BANDWIDTH=20. A different choice of values will produce different empirical semivariograms, because these options can regulate the number of pairs that are included in a class. You do not want these parameters to have values that are too small, so that you can allow for an adequate number of point pairs per class. On the other hand, the higher the values of these parameters, the more data pairs coming from closely neighboring directions will be included in each lag. In this case, there is a risk of losing information along the particular direction. This can be a side effect because you will incorporate data pairs from a broader spectrum of angles, and thus potentially amplify weaker anisotropy or weaken stronger anisotropy, as noted in the section "Angle Classification" on page 7541. You can experiment with different ANGLETOLERANCE= and BANDWIDTH= values to reach this balance with your data, if necessary.

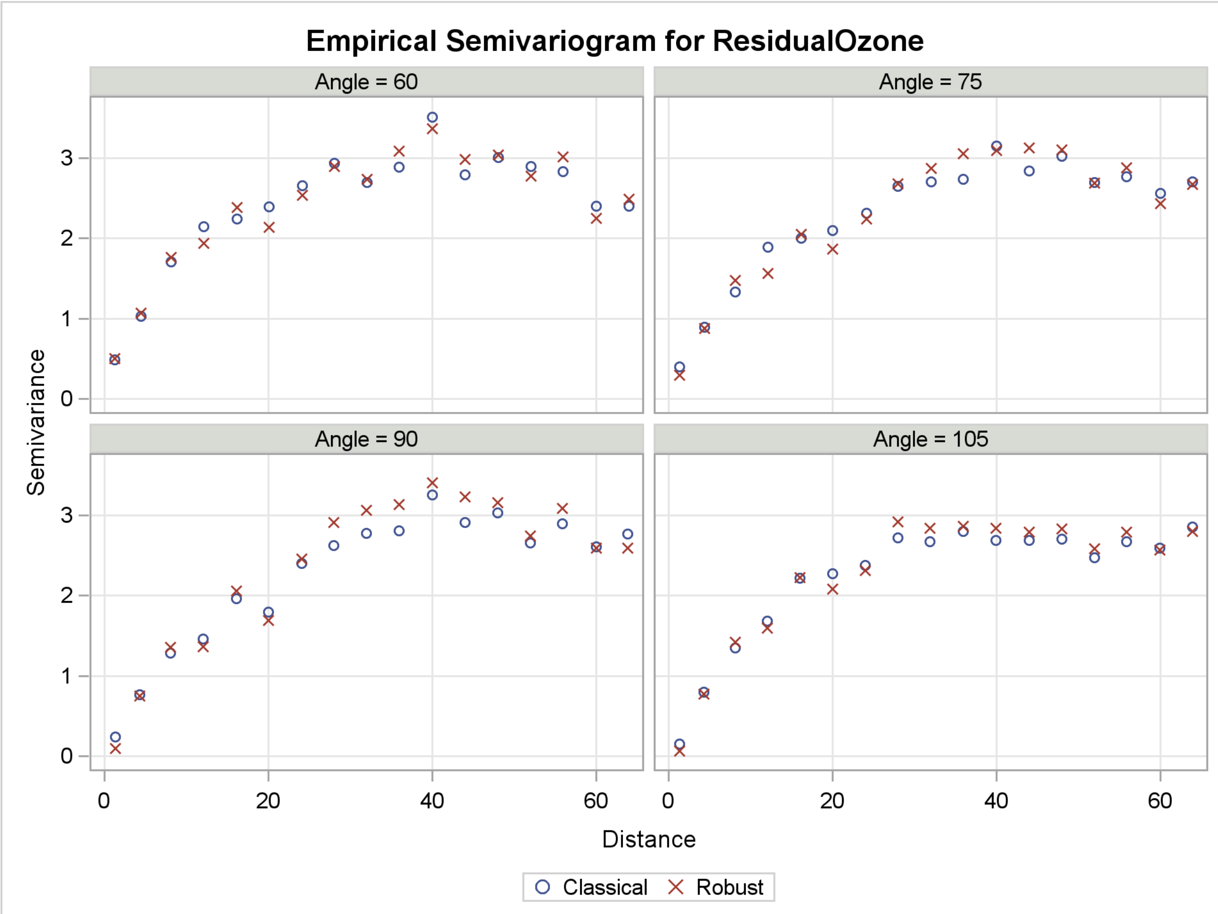
With the following statements you ask to display only the SEMIVAR plots in the specified number of directions. Multiple empirical semivariograms are placed by default in panels, as Output 95.2.6 shows. If you want an individual plot for each angle, then you need to further specify the plot option SEMIVAR(UNPACK).

```
proc variogram data=gmout plot(only)=semivar;
   compute lagd=4 maxlag=16 ndir=12 atol=22.5 bandw=20 robust;
   coord xc=East yc=North;
   var ResidualOzone;
run;
```

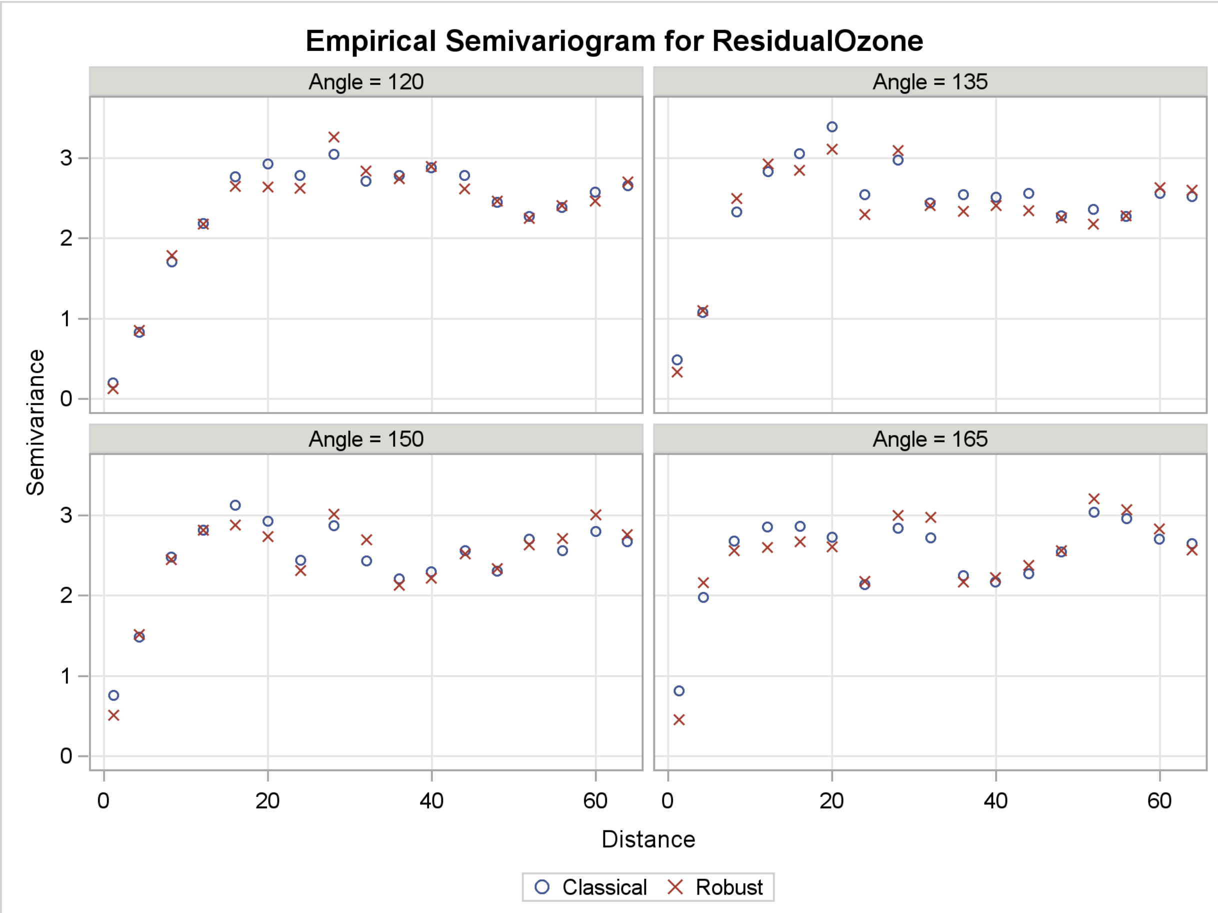
Output 95.2.6 Ozone Empirical Semivariograms with $0° \leq \theta < 180°$ and $\delta\theta = 15°$



Output 95.2.6 *continued*



Output 95.2.6 *continued*



Note that in some of the directions, such as for $\theta = 0°$, the directional plots tend to exhibit a somewhat noisy structure. This behavior can be due to the pairs distribution across the particular direction. Specifically, based on the LAGDISTANCE= choice there might be insufficient pairs present in a class. Also, depending on the ANGLETOLERANCE= and BANDWIDTH= values, too many pairs might be considered from neighboring angles that potentially follow a modified structure. These are factors that can increase the variability in the semivariance estimate. A different explanation might lie in the existence of outliers in the data set; this aspect is further explored in "Example 95.5: A Box Plot of the Square Root Difference Cloud" on page 7592.

This behavior is relatively mild here and should not obstruct your goal to study anisotropy in your data. You can also perform individual computations in any direction. By doing so, you can fine-tune the computation parameters and attempt to obtain smoother estimates of the sample semivariance.

Further in this study, the directional plots in Output 95.2.6 suggest that during shifting from $\theta = 0°$ to $\theta = 90°$, the empirical semivariogram range increases. Beyond the angle $\theta = 90°$, the range starts decreasing again until the whole circle is traversed at 180° and small range values are encountered around the N–S direction at $\theta = 0°$. Note that the sill seems to remain overall the same. This analysis suggests that there is anisotropy in the ozone concentrations, with the major axis oriented at about $\theta = 90°$ and the minor axis situated perpendicular to the major axis at $\theta = 0°$.

The multidirectional analysis requires that for a given LAGDISTANCE= you also specify a MAXLAGS= value. Since the ozone correlation range might be unknown (as assumed here), you can apply the rule of thumb that suggests use of the half-extreme data distance in the direction of interest, as explained in the section "Spatial Extent of the Empirical Semivariogram" on page 7548. Following the information displayed in Output 95.2.3, for different directions this distance varies between $99.4/2 = 49.7$ and $140.8/2 = 70.4$ km. In turn, the pairwise distances table in Output 95.2.2 indicates that within this range of distances you can specify MAXLAG= to be between 12 and 17 lags. In this example you specify MAXLAG=16.

At this point you are ready to continue with fitting theoretical semivariogram models to the empirical semivariogram in the selected directions of $\theta = 0°$ and $\theta = 90°$. This process follows the exact steps shown in "Example 95.1: Theoretical Semivariogram Model Fitting" on page 7562. You apply weighted least squares fitting. In the following, PROC NLIN is used for the simultaneous theoretical semivariogram fitting in both directions of interest. By trying out different models, you see that an exponential one is suitable for your empirical data:

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

where $\gamma_z(0) = 0$ and $a_0 > 0$. First, you run PROC VARIOGRAM to create the necessary input information for PROC NLIN. The NLIN procedure needs information about the semivariance variance, which you obtain when you specify the CL option in the COMPUTE statement. You use the following statements:

```
proc variogram data=gmout plot=none;
   compute lagd=4 maxlag=16 robust cl;
   directions 0(22.5,10) 90(22.5,10);
   coord xc=East yc=North;
   ods output SemivariogramTable=svg;
   var ResidualOzone;
run;
```

You also request a vector of points throughout the Distance axis where PROC NLIN estimates the theoretical model values for the two selected directions. Note that essentially you need such a vector for each one of these directions. Then, the output of PROC VARIOGRAM is combined with the added Distance points into the PROC NLIN input data set, as shown in the following statements:

```
data pv;
   do Angle = 0 to 90 by 90;
      do Distance = 0 to 64 by 0.5;
         Semivariance = .;
         output;
      end;
   end;
run;

data svg; set svg pv; by Angle Distance;
run;
```

PROC NLIN requires initial values for the fitting parameters. For the $\theta = 0°$ direction, Output 95.2.6 suggests that a range $a_{00,init} = 4$ km (effective range $r_{\epsilon 0} = 12$ km) and a partial sill ${\sigma_0}^2{}_{0,init} = 2.5$ can be used. Output 95.2.6 indicates that in the direction of $\theta = 90°$, a range of $a_{090,init} = 16$ km (effective range $r_{\epsilon 9}0 = 48$ km) and a partial sill ${\sigma_0}^2{}_{90,init} = 3$ can be used. Both cases indicate that there is no significant nugget effect present. The nugget effect is initialized to the value $c_{n init} = 0$. Note that you use one nugget effect value for all directions; that is, it is considered isotropic. The following statements implement these considerations, and PROC SGPLOT is eventually called to create the sample and theoretical semivariogram plots:

```
proc nlin data=svg;
   parms Nugget=0 Range0=4   PartSill0=2.5
                  Range90=16 PartSill90=3;
   if (Angle=0) then
      y = Nugget + PartSill0*(1-exp(-Distance/Range0));
   else if (Angle=90) then
      y = Nugget + PartSill90*(1-exp(-Distance/Range90));
   model Semivariance = y;
   _weight_ = 1/SemivarianceStdErr/SemivarianceStdErr;
   output out=WLS p=WLSprediction;
run;

proc sgplot data=WLS;
   title "Empirical and Fitted Theoretical Semivariogram";
   xaxis label = "Distance" grid;
   yaxis label = "Semivariance" grid;
   scatter y=Semivariance x=Distance /
           name='SemiVarClassical'
           group=Angle;
   series x=Distance y=WLSprediction /
          lineattrs=(thickness=2px color=black)
          name='SemivarPredWgh'
          group=Angle;
   discretelegend 'SemiVarClassical' 'SemivarPredWgh';
run;

ods graphics off;
```

Output 95.2.7 shows the PROC NLIN estimates of the fitting parameters. From this table, you can easily compute the estimates for the sills in $\theta = 0°$ and $\theta = 90°$, which are $c_{00} = 2.5339$ and $c_{090} = 2.8703$, respectively. The fitted and empirical semivariograms are displayed in Output 95.2.8.

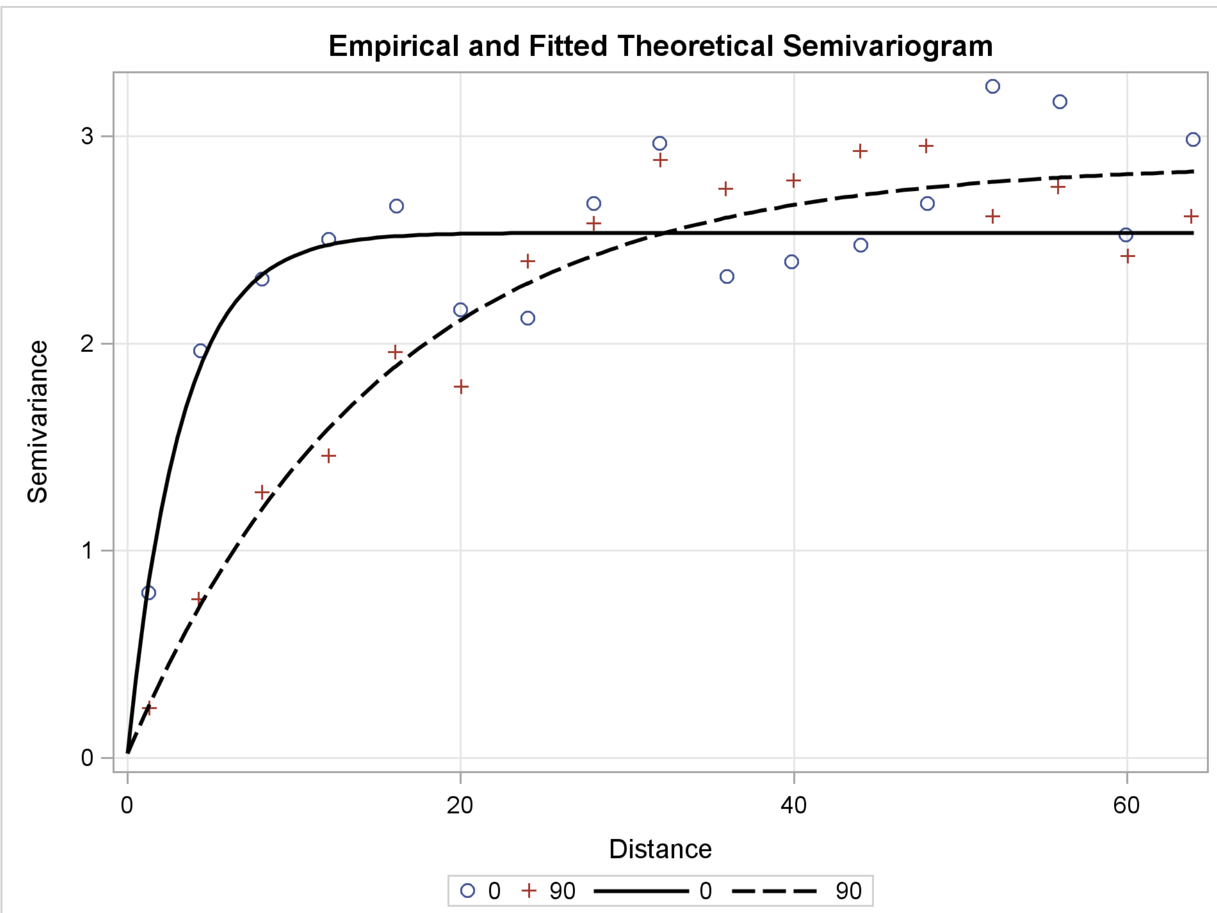
Output 95.2.7 Weighted Least Squares Fitting Parameter Estimates for the Selected Directions $\theta = 0°$ and $\theta = 90°$

Semivariogram in Anisotropic Case With Trend Removal Example

The NLIN Procedure

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
Nugget	0.0214	0.1517	-0.2888	0.3316
Range0	3.2073	1.1413	0.8730	5.5415
PartSill0	2.5125	0.1677	2.1696	2.8554
Range90	15.0962	2.7113	9.5510	20.6414
PartSill90	2.8489	0.1740	2.4930	3.2048

Output 95.2.8 Fitted Theoretical and Empirical Semivariogram for the Ozone Data in the $\theta = 0°$ and $\theta = 90°$ Directions



Conclusively, your semivariogram analysis on the detrended ozone data suggests that the ozone SRF exhibits anisotropy in the perpendicular directions of N–S ($\theta = 0°$) and E–W ($\theta = 90°$). The sills in the two directions of anisotropy are similar in size, whereas the range in the major axis is about 4.5 times larger than the minor axis range.

Example 95.3: Analysis without Surface Trend Removal

This example uses PROC VARIOGRAM without removing potential surface trends in a data set, in order to investigate a distinguished spatial direction in the data. In doing so, this example also serves as a guide to examine under which circumstances you might be able to bypass the effect of a trend on a semivariogram. Typically, though, for theoretical semivariogram estimations you follow the analysis presented in "Example 95.2: An Anisotropic Case Study with Surface Trend in the Data" on page 7566.

As explained in the section "Details: VARIOGRAM Procedure" on page 7532, when you compute the empirical semivariance for data that contain underlying surface trends, the outcome is the pseudo-semivariance. Pseudo-semivariograms are not estimates of the theoretical semivariogram; hence, they do not provide any information about the spatial continuity of your SRF.

However, in the section "Empirical Semivariograms and Surface Trends" on page 7551 it is mentioned that you might still be able to perform a semivariogram analysis with potentially non-trend-free data, if you suspect that your measurements might be trend-free across one or more specific directions. The example demonstrates this approach.

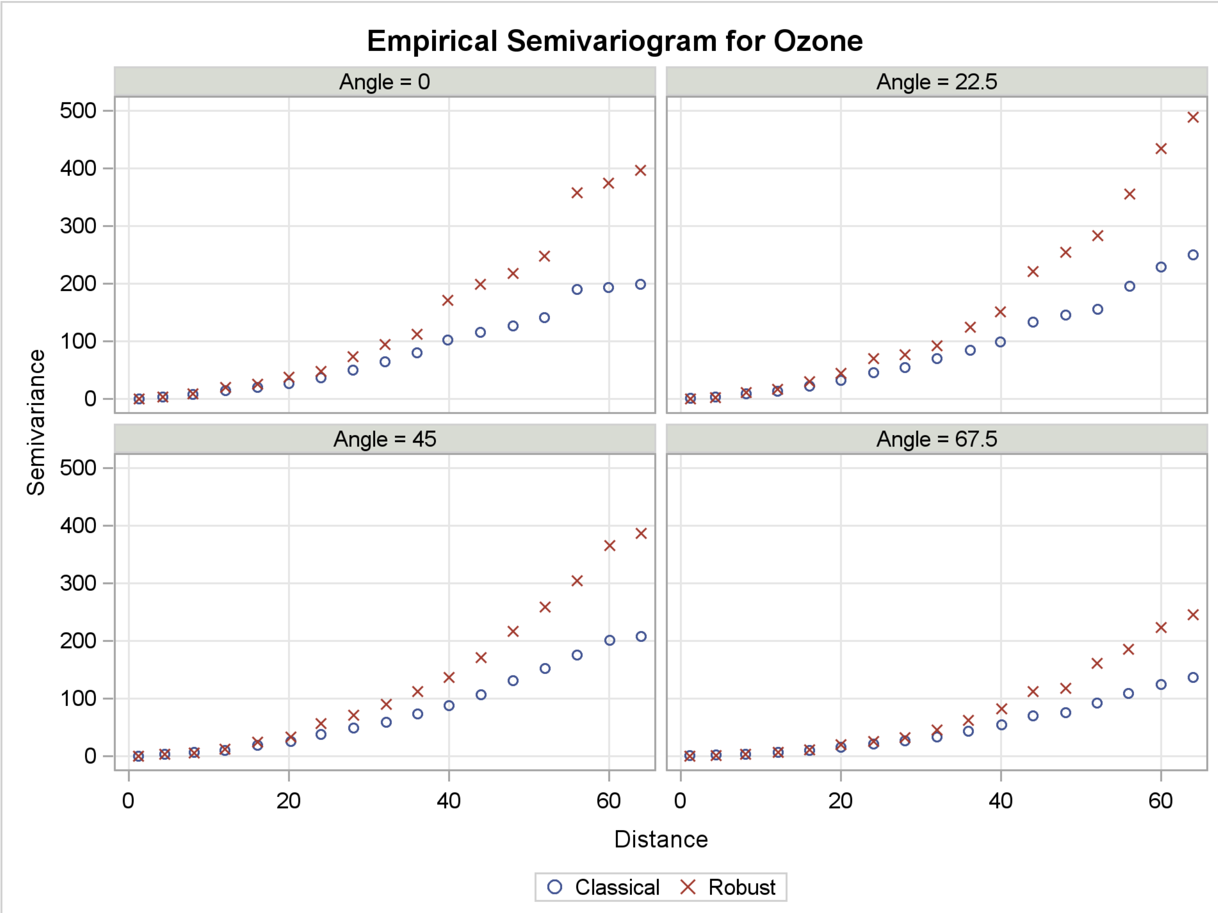
Reconsider the ozone data presented at the beginning of "Example 95.2: An Anisotropic Case Study with Surface Trend in the Data" on page 7566. The spatial distribution of the data is shown in Figure 95.2.1, and the pairwise distance distribution for NHC=35 is illustrated in Figure 95.2.3. This exploratory analysis suggested a LAGDISTANCE=4 km, and Figure 95.2.2 indicated that for this LAGDISTANCE= you can consider a value of MAXLAGS=16.

Recall from the section "Empirical Semivariograms and Surface Trends" on page 7551 that you need to investigate the empirical semivariogram of the data in a few different directions, in order to identify a trend-free direction. If such a direction exists, then you can proceed with this special type of analysis. The following statements employ NDIRECTIONS=8 to examine eight directions:

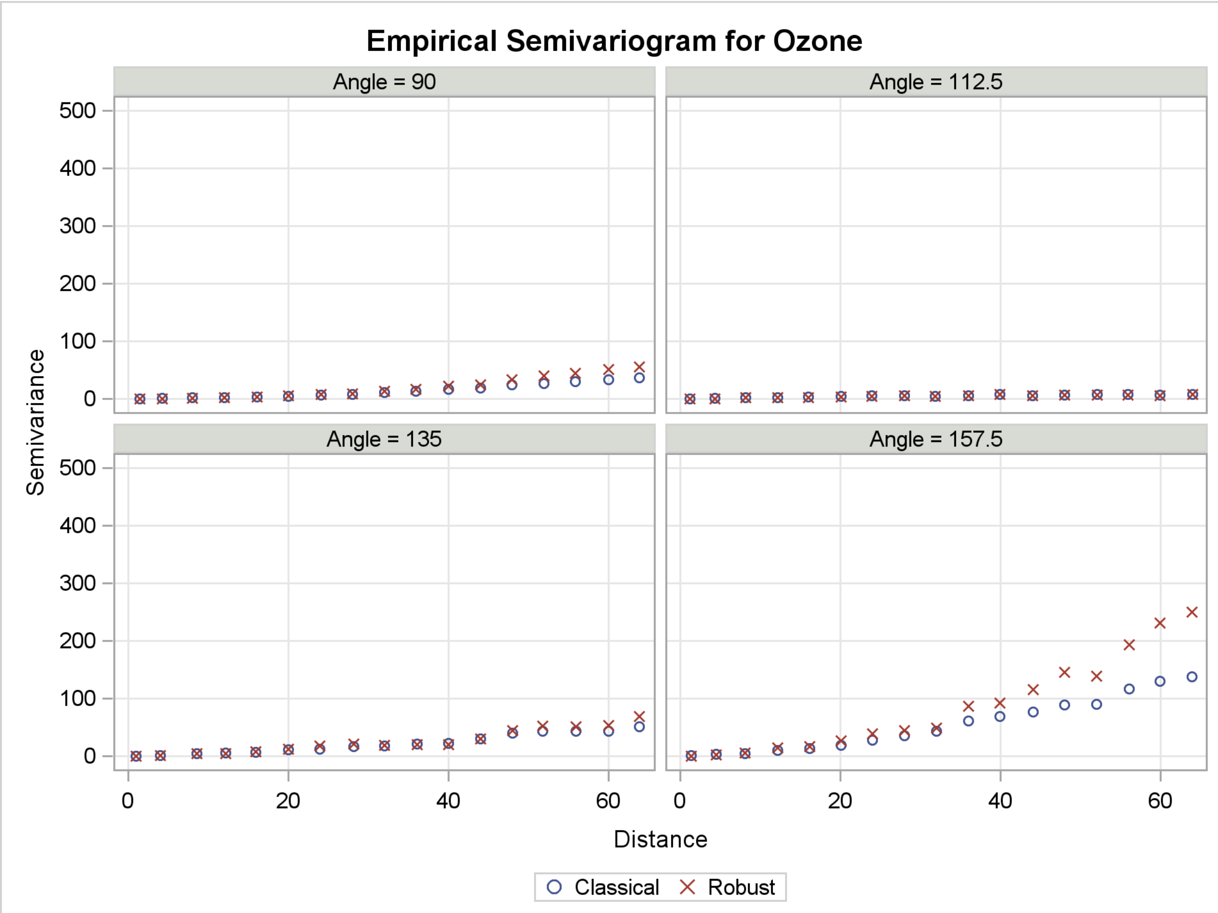
```
title 'Semivariogram Without Trend Removal Example';
ods graphics on;

proc variogram data=ozoneSet plot(only)=semivar;
   compute lagd=4 maxlag=16 ndirections=8 robust;
   coord xc=East yc=North;
   var Ozone;
run;
```

By default, the range of 180° is divided into eight equally distanced angles: $\theta = 0°$, $\theta = 22.5°$, $\theta = 45°$, $\theta = 67.5°$, $\theta = 90°$, $\theta = 112.5°$, $\theta = 135°$, and $\theta = 157.5°$. The resulting empirical semivariograms for these angles are shown in Output 95.3.1.

Output 95.3.1 Ozone Empirical Semivariograms with $0° \leq \theta < 180°$ and $\delta\theta = 22.5°$



Output 95.3.1 *continued*



The figures in Output 95.3.1 suggest that in all directions there is an overall continuing increase with distance of the semivariance. As explained in the section "Theoretical Semivariogram Models" on page 7532, this can be an indication of systematic trends in the data. However, in the direction of $\theta = 112.5°$ there is a clear indication that the increase rate, if any, is smaller than the corresponding rates across the rest of the directions. You then want to search whether there exists a trend-free direction in the neighborhood of this angle.

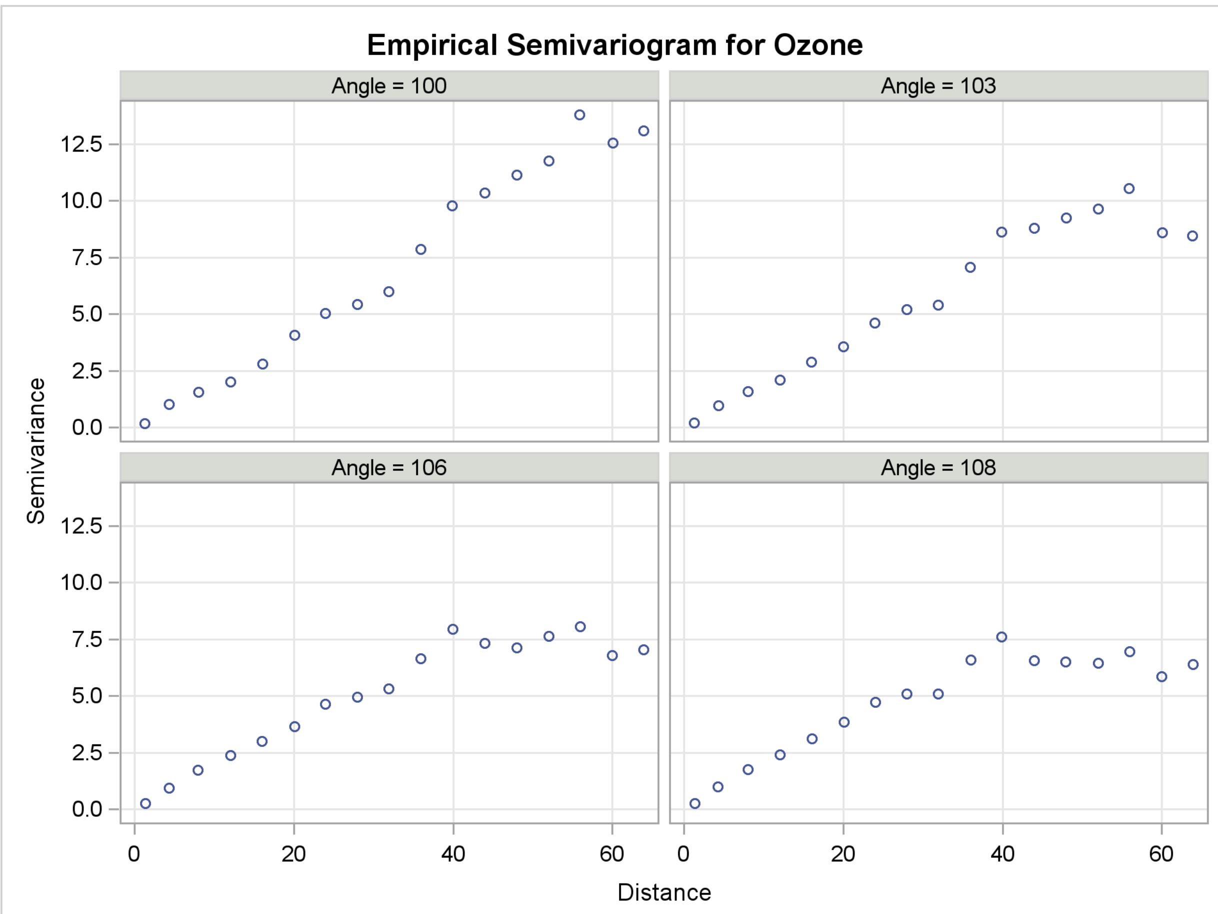
Run PROC VARIOGRAM again, specifying several directions within an interval of angles where you want to close in and suspect the existence of a trend-free direction. In the following step you use an ANGLETOL=15°, which is smaller than the default value of 22.5°, as well as a BANDWIDTH=10 km. The smaller values help with minimization of the interference with neighboring directions, as discussed in the section "Angle Classification" on page 7541.

The aforementioned considerations are addressed in the following statements:

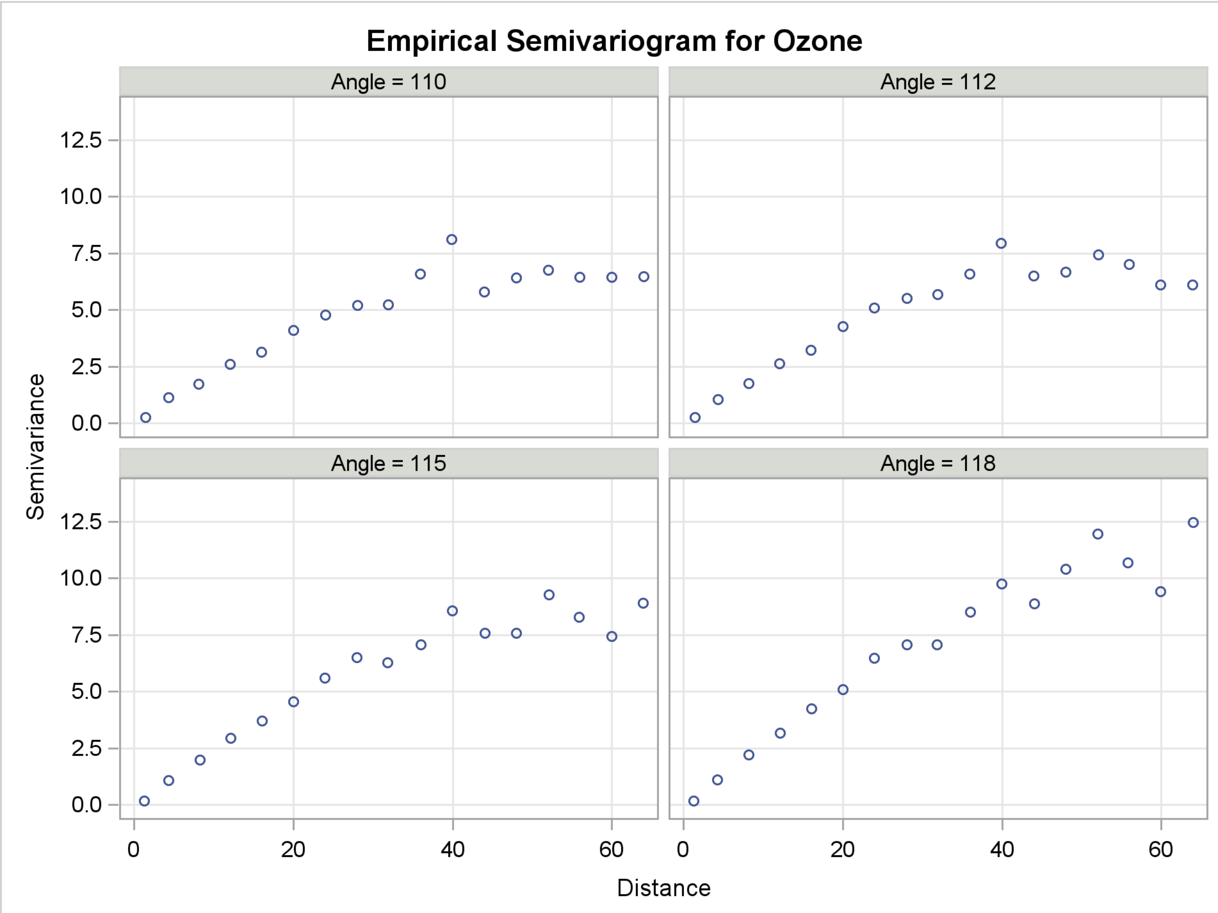
```
proc variogram data=ozoneSet plot(only)=semivar(cla);
   compute lagd=4 maxlag=16 robust;
   directions 100(15,10) 103(15,10)
              106(15,10) 108(15,10)
              110(15,10) 112(15,10)
              115(15,10) 118(15,10);
   coord xc=East yc=North;
   var Ozone;
run;
```

Your analysis has brought you to examine a narrow strip of angles within $\theta = 100°$ and $\theta = 118°$. The pseudo-semivariograms in Output 95.3.2 and Output 95.3.3 indicate that at the boundaries of this strip, the angles display increasing semivariance with distance. On the other hand, within this interval there are directions across which the semivariance is tentatively reaching a sill, and these are potential candidates to be trend-free directions.

Output 95.3.2 Ozone Empirical Semivariograms in $100°$, $103°$, $106°$, and $108°$



Output 95.3.3 Ozone Empirical Semivariograms in 110°, 112°, 115°, and 118°



You could further investigate this angle spectrum in more detail. For example, you can monitor additional angles in between, or use a smaller LAGDISTANCE= and increased MAXLAGS= values to single out the most qualified candidate. For the purpose of this example, you can consider the direction $\theta = 108°$ to very likely be the trend-free one you are looking for.

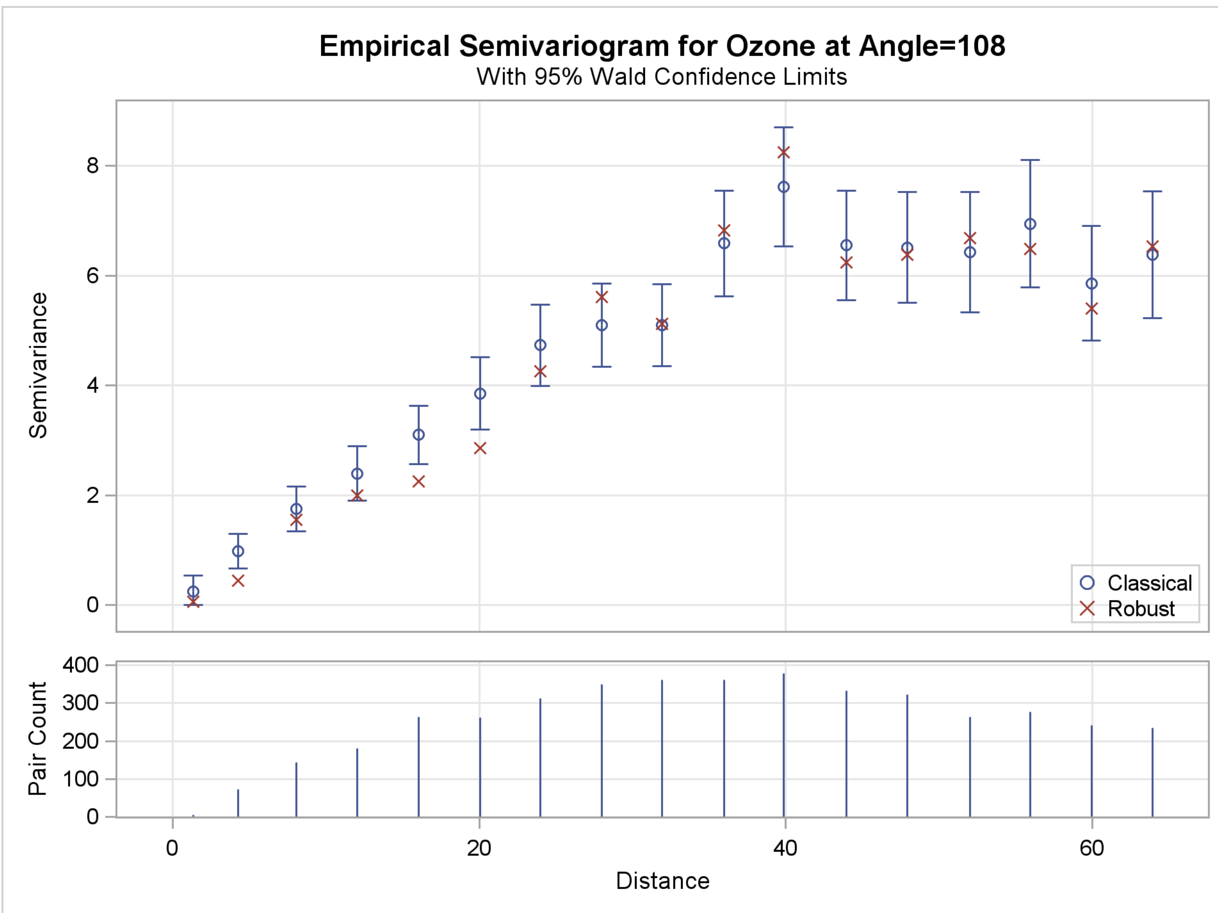
From a physical standpoint, it is also mentioned in the section "Empirical Semivariograms and Surface Trends" on page 7551 that you should expect the trend-free direction, if it exists, to be perpendicular to the direction of the maximum dip in the values of the ozone field. If you cross-examine the ozone data distribution in Output 95.2.1, the figure suggests that this direction exists and is slightly tilted clockwise with respect to the E–W axis. This direction emerges from the mild stratification of the ozone values in your data distribution. The ozone concentrations across it are similar when compared to surrounding directions, and as such, it has been identified as a trend-free direction.

Your next step is to obtain the empirical semivariogram in the suspected trend-free direction of $\theta = 108°$, and to perform a theoretical model fit as shown in "Example 95.1: Theoretical Semivariogram Model Fitting" on page 7562 and "Example 95.2: An Anisotropic Case Study with Surface Trend in the Data" on page 7566.

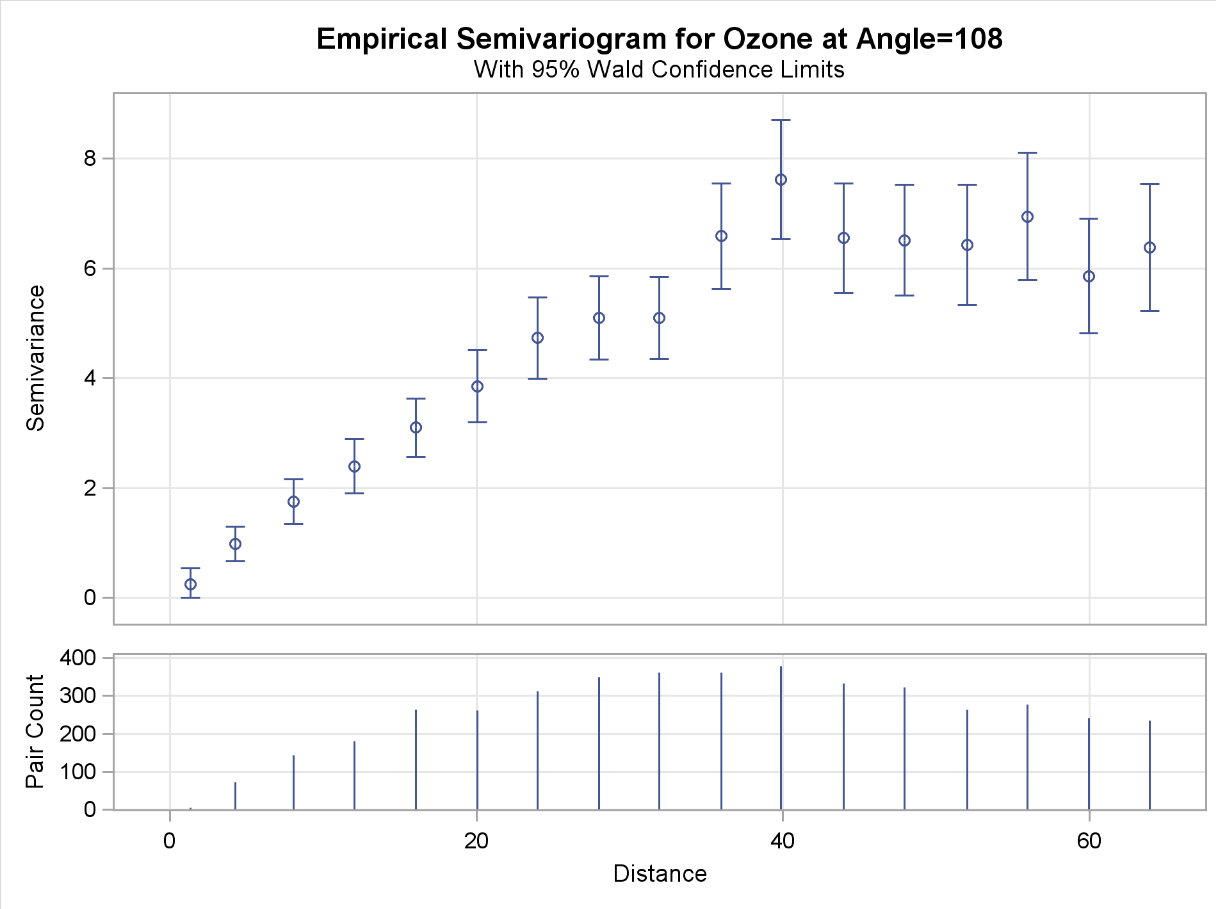
First, assume that you want to inspect the classical and robust empirical semivariograms in the selected direction, as well as a separate plot of the classical one and its confidence limits. Use the PLOT option as shown in the following statements to produce the two empirical semivariograms for your selected angle. The resulting plots are displayed in Output 95.3.4 and Output 95.3.5.

```
proc variogram data=ozoneSet
               plot(only)=(semivar semivar(cla));
   compute lagd=4 maxlag=16 robust cl;
   directions 108(15,10);
   coord xc=East yc=North;
   ods output SemivariogramTable=sv;
   var ozone;
run;
```

Output 95.3.4 Ozone Classical and Robust Empirical Semivariograms in $\theta = 108°$



Output 95.3.5 Ozone Classical Empirical Semivariogram in $\theta = 108°$



Then, you create a vector of points on the Distance axis to provide enough theoretical model estimation locations for a smooth plot. Also, you combine this data set with the output of PROC VARIOGRAM to create the input data set for PROC NLIN. The statements you use are as follows:

```
data pv;
   do Distance = 0 to 64 by 0.5;
      Semivariance = .;
      output;
   end;
run;

data sv; set sv pv; by Distance;
run;
```

Finally, as in the previous examples, you employ PROC NLIN to fit a theoretical model to the empirical semivariogram shown in Output 95.3.5. Note that you have not omitted specifying the CL option earlier in the COMPUTE statement in PROC VARIOGRAM, so that you can provide PROC NLIN with semivariance variance information. The semivariance exhibits a slow, almost linear rise at short distances and seems to be reaching the sill fast, rather than asymptotically.

You can accommodate this behavior by using the following spherical model:

$$\gamma_z(h) = \begin{cases} c_n + {\sigma_0}^2 \left[\frac{3}{2}\frac{h}{a_0} - \frac{1}{2}(\frac{h}{a_0})^3 \right], & \text{for } 0 < h \le a_0 \\ c_0, & \text{for } a_0 < h \end{cases}$$

where $\gamma_z(0) = 0$ and $a_0 > 0$. PROC NLIN requires initial values for the model parameters. A reasonable initial choice for the sill value is about 7, and for the range $a_{0init} = 40$ km. Though there does not seem to be a nugget effect, you use an initial value $c_{ninit} = 0$ and you initialize the partial sill by using ${\sigma_0}^2{}_{init} = 7$. The following statements implement these considerations. Consequently, PROC SGPLOT is invoked to create the empirical and theoretical semivariogram plots.

```
proc nlin data=sv;
   parms Nugget=0 Range=40 PartSill=7;
   if (Distance<Range)
      then y = Nugget + PartSill*(1.5*(Distance/Range) -
                                  0.5*(Distance/Range)**3);
      else y = Nugget + PartSill;
   model Semivariance = y;
   _weight_ = 1/SemivarianceStdErr/SemivarianceStdErr;
   output out=WLS p=WLSprediction;
run;

proc sgplot data=WLS;
   title "Empirical and Fitted Theoretical Semivariogram";
   xaxis label = "Distance" grid;
   yaxis label = "Semivariance" grid;
   scatter y=Semivariance x=Distance /
           markerattrs = GraphData1(symbol=circle)
           name='SemiVarClassical';
   series x=Distance y=WLSprediction /
          lineattrs = (thickness=2px color=black)
          name = 'SemivarPredWgh';
   discretelegend 'SemiVarClassical' 'SemivarPredWgh';
run;

ods graphics off;
```

PROC NLIN fits the requested model, for which the range, partial sill, and nugget effect estimates are shown in Figure 95.3.6. Clearly, there is an almost negligible nugget effect based on the weighted least squares PROC NLIN analysis. From the fitted values the estimate for the theoretical semivariogram sill is $c_0 = 6.55601$. Figure 95.3.7 displays the fitted and empirical semivariograms in the selected direction $\theta = 108°$.

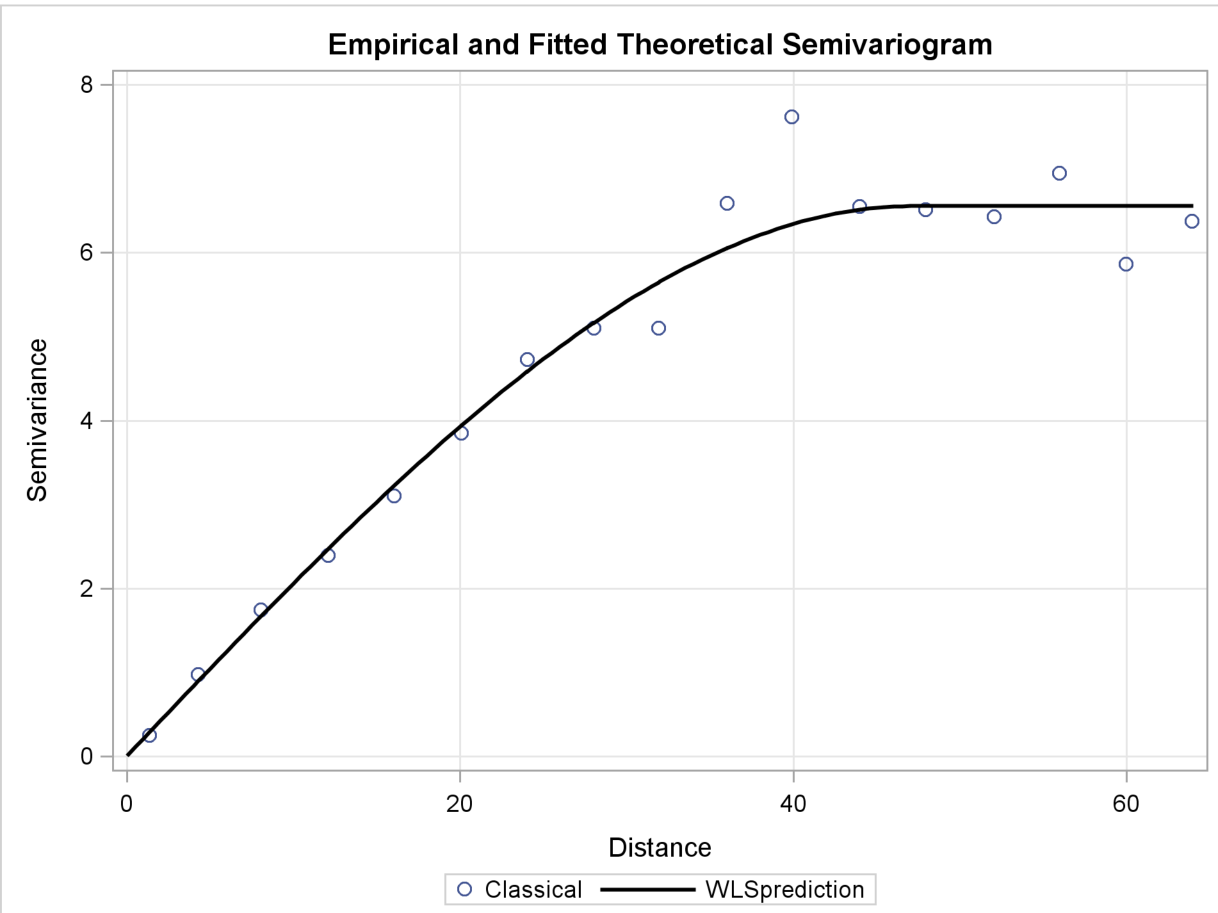
Output 95.3.6 Weighted Least Squares Fitting Parameter Estimates for $\theta = 108°$

Semivariogram Without Trend Removal Example

The NLIN Procedure

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
Nugget	0.00991	0.1111	-0.2284	0.2482
Range	47.0906	2.7872	41.1126	53.0685
PartSill	6.5461	0.2208	6.0725	7.0197

Output 95.3.7 Fitted Theoretical and Empirical Semivariogram for Ozone Data in the Direction $\theta = 108°$



A comparative look at the empirical and fitted semivariograms in Output 95.3.7 and Output 95.2.8 suggests that the analysis of the trend-free ResidualOzone produces a different outcome from that of the original Ozone values. In fact, a more suitable comparison can be made between the semivariograms in the assumed trend-free direction $\theta = 108°$ of the current scenario, and the one shown in Output 95.2.6 in the nearly identical direction $\theta = 105°$. It might seem unreasonable that these two semivariograms are produced both in the same ozone study and in a narrow band of directions free of apparent surface trends, yet they bear no resemblance. However, the lack of similarity in these plots stems from operating on two different data sets where the outcome depends on the actual data values.

More specifically, in the eyes of semivariogram analysis the trend-free ozone set and the original ozone measurements are treated as different quantities. The process of detrending the original Ozone values is a transformation of these values into the trend-free ones of ResidualOzone. Any existing spatial correlation in the original data is not directly memorized into the transformed data, but is rather affected by the transformation features. In principle, the emerging data set has its own characteristics, as demonstrated in this example.

A final remark concerns the issue of isotropy. Based on the details presented in the section "Empirical Semivariograms and Surface Trends" on page 7551, your knowledge of the spatial structure of the ozoneSet data set is limited to the selected trend-free direction you indicated in the present example. You can generalize this outcome for all spatial directions only if you consider the hypothesis of isotropy in the ozone field to be reasonable. Note that you cannot infer the assumption of anisotropy in the present example based on the analysis in the section "Analysis with Surface Trend Removal" on page 7570. Again, the reason is that you currently use the observed Ozone values, whereas the ResidualOzone data in the previous example emerged from a transformation of the current data. Hence, you have essentially two data sets that do not necessarily share the same properties.

Example 95.4: Covariogram and Semivariogram

The covariance that was reviewed in the section "Stationarity" on page 7538 is an alternative measure of spatial continuity that can be used instead of the semivariance. In a similar manner to the empirical semivariance that was presented in the section "Theoretical and Computational Details of the Semivariogram" on page 7536, you can also compute the empirical covariance. The covariograms are plots of this quantity and can be used to fit permissible theoretical covariance models, in correspondence to the semivariogram analysis presented in the section "Theoretical Semivariogram Models" on page 7532. This example displays a comparative view of the empirical covariogram and semivariogram, and examines some additional aspects of these two measures.

You consider 500 simulations of an SRF $Z(\boldsymbol{s})$ in a square domain of 100×100 (10^6 km^2). The following DATA step defines the data locations:

```
title 'Covariogram and Semivariogram Example';
data dataCoord;
   retain seed 837591;
   do i=1 to 100;
      East = round(100*ranuni(seed),0.1);
      North = round(100*ranuni(seed),0.1);
      output;
   end;
run;
```

For the simulations you use PROC SIM2D, which produces Gaussian simulations of SRFs with user-specified covariance structure—see Chapter 79, "The SIM2D Procedure." The Gaussian SRF implies full knowledge of the SRF expected value $\mathrm{E}[Z(\boldsymbol{s})]$ and variance $\mathrm{Var}[Z(\boldsymbol{s})]$ at every location $\boldsymbol{s}$. The following statements simulate an isotropic, second-order stationary SRF with constant expected value and variance throughout the simulation domain:

```
ods graphics on;

proc sim2d outsim=dataSims;
   simulate numreal=500 seed=79750
            nugget=2 scale=6 range=10 form=exp;
   mean 30;
   grid gdata=dataCoord xc=East yc=North;
run;
```

Here, the SIMULATE statement accommodates the simulation parameters. The NUMREAL= option specifies that you want to perform 500 simulations, and the SEED= option specifies the seed for the simulation random number generator. You use the MEAN statement to specify the expected value $\mathrm{E}[Z(\boldsymbol{s})] = 30$ units of Z. You also specify two variance components. The first is the nugget effect, and you use the NUGGET= option to set it to $c_n = 2$. The second is the partial sill ${\sigma_0}^2 = 6$ that you specify with the SCALE= option. The two variance components make up the total SRF variance $\mathrm{Var}[Z(\boldsymbol{s})] = c_n + {\sigma_0}^2 = 8$. You assume an exponential covariance structure to describe the field spatial continuity, where ${\sigma_0}^2$ is the sill value, and its range $a_0 = 10$ km (effective range $a_\epsilon = 3a_0 = 30$ km) is specified by the RANGE= option. The option FORM= specifies the covariance structure type.

The empirical semivariance and covariance are computed by the VARIOGRAM procedure, and are available either in the ODS output semivariogram table (as variables Semivariance and Covariance, respectively) or in the OUTVAR= data set. In the following statements you obtain these variables by using the OUTVAR= data set of the VARIOGRAM procedure:

```
proc variogram data=dataSims outv=outv noprint;
   compute lagd=3 maxlag=18;
   coord xc=gxc yc=gyc;
   by _ITER_;
   var svalue;
run;
```

For each distance lag you take the average of the empirical measures over the number of simulations. PROC SORT is used to prepare the input data for PROC MEANS, which produces these averages and stores them in the dataAvgs data set. This sequence is performed with the following statements:

```
proc sort data=outv;
   by lag;
run;

proc means data=outv n mean noprint;
   var Distance variog covar;
   by lag;
   output out=dataAvgs mean(variog)=Semivariance
                       mean(covar)=Covariance
                       mean(Distance)=Distance;
run;
```

The SGPLOT procedure is used to create the plot of the average empirical semivariogram and covariogram, as in the following statements:

```
proc sgplot data=dataAvgs;
   title "Empirical Semivariogram and Covariogram";
   xaxis label = "Distance" grid;
   yaxis  label = "Semivariance" min=-0.5 max=9 grid;
   y2axis label = "Covariance"   min=-0.5 max=9;
   scatter y=Semivariance x=Distance /
           markerattrs = GraphData1
           name='Semivar'
           legendlabel='Semivariance';
   scatter y=Covariance x=Distance /
           y2axis
           markerattrs = GraphData2
           name='Covar'
           legendlabel='Covariance';
   discretelegend 'Semivar' 'Covar';
run;

ods graphics off;
```

The plot of the average empirical semivariance and covariance of the preceding analysis is shown in Output 95.4.1. Note that the high number of simulations led to averages of empirical continuity measures that accurately approximate the simulated SRF characteristics. Specifically, the empirical semivariogram and covariogram both exhibit clearly exponential behavior. The semivariogram sill is approximately at the specified variance $\text{Var}[Z(\mathbf{s})] = 8$ of the SRF.

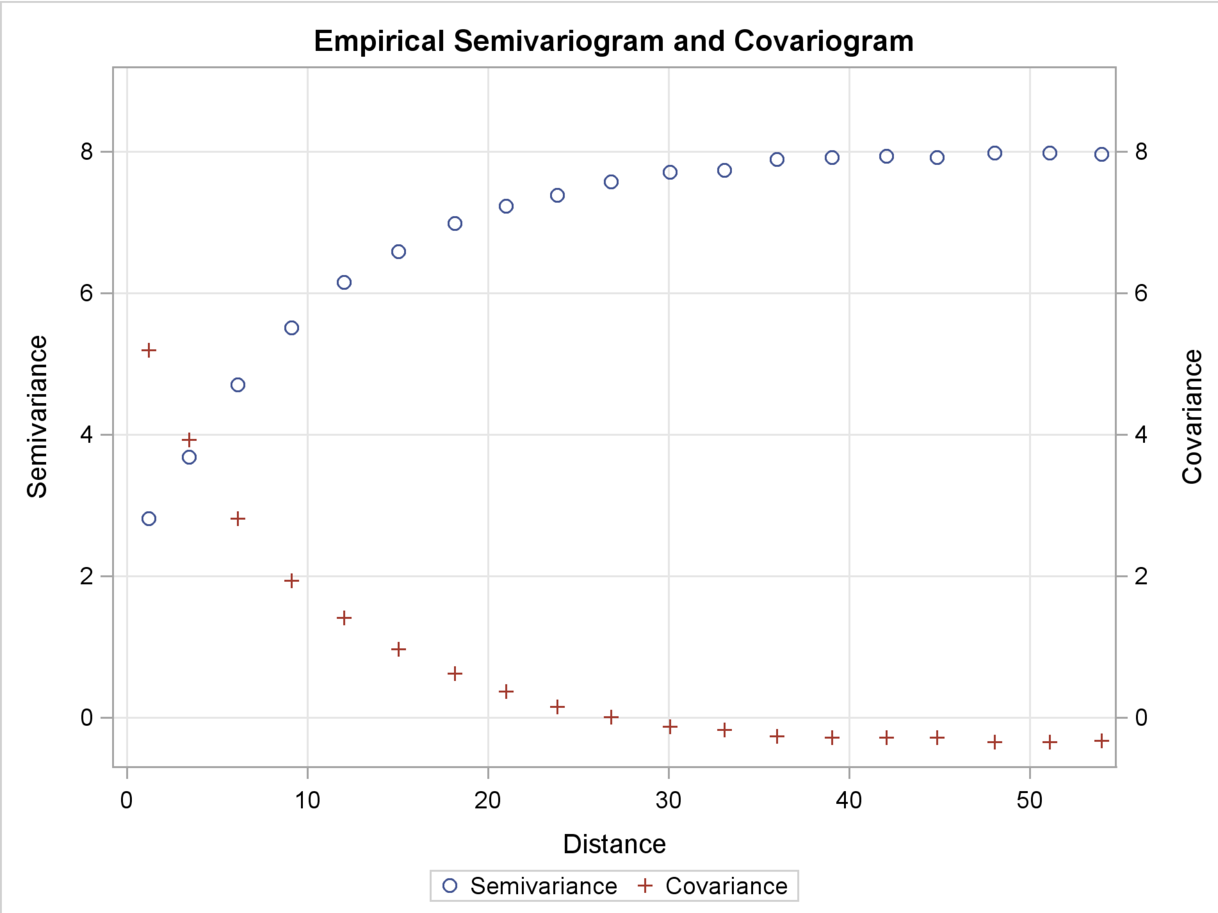
The simulated SRF is second-order stationary, so you expect at each lag the sum of the empirical semivariance and covariance to approximate the field variance $\text{Var}[Z(\mathbf{s})]$, as explained in the section "Stationarity" on page 7538. This behavior is evident in Output 95.4.1.

We conclude with a discussion of basic reasons why the empirical semivariogram analysis is commonly preferred to the empirical covariance analysis. A first reason comes from the assumptions that are necessary to compute each of these two measures. The condition of intrinsic stationarity that is required in order to define the empirical semivariogram is less restrictive than the condition of second-order stationarity that is required in order to consider the covariance function as a parameter of the process.

Also, an empirical semivariogram can indicate whether a nugget effect is present in your data sample, whereas the empirical covariogram itself might not reveal this information. This point is illustrated in Output 95.4.1, where you expect to see that $C(\mathbf{0}) = \text{Var}[Z(\mathbf{s})]$, but the empirical covariogram cannot have a point at exactly $\mathbf{h} = \mathbf{0}$. There is a practical way to investigate for a nugget effect when you use empirical covariograms. Recall that the OUTVAR= data set provides you with the sample variance (shown in the COVAR column for LAG=-1), as the following statement shows:

```
/* Obtain the sample variance from the data set ----------------*/

proc print data=dataAvgs (obs=1);
```

Output 95.4.1 Average Empirical Semivariogram and Covariogram from 500 Simulations



Output 95.4.2 is a partial output of the dataAvgs data set, which contains averages of the OUTVAR= data set, and shows the computed average $C(\mathbf{0})$ in the Covariance column. The combination of the empirical covariogram and the $C(\mathbf{0})$ value can help you fit a theoretical covariance model that will include any nugget effect, if present. See also the discussion in Schabenberger and Gotway (2005, Section 4.2.2) about the Matérn definition of the covariance function that is related to this issue. In particular, this definition provides for an additional variance component in the covariance expression at $\boldsymbol{h} = \mathbf{0}$ to account for the corresponding nugget effect in the semivariogram.

Output 95.4.2 Partial Outcome of the dataAvgs Data Set

Empirical Semivariogram and Covariogram

Obs	LAG	_TYPE_	_FREQ_	Semivariance	Covariance	Distance
1	-1	0	500	.	7.74832	.

In addition to the preceding points, if the SRF is nonstationary, the empirical semivariogram indicates that the SRF variance will increase with distance $\boldsymbol{h}$, as Output 95.3.1 shows in "Example 95.3: Analysis without Surface Trend Removal" on page 7579. In that case it makes no sense to compute the empirical covariogram. Specifically, the covariogram could provide you with an estimate of the sample variance, which is not sufficient to indicate that the SRF might not be stationary (see also Chilès and Delfiner 1999, p. 31).

Finally, the definitions of the empirical semivariance and covariance in the section "Theoretical and Computational Details of the Semivariogram" on page 7536 clearly show that the sample mean $\bar{Z}$ and the SRF expected value $\mathrm{E}[Z(\boldsymbol{s})]$ are not important for the computation of the semivariance, but either one is necessary for the covariance. Hence, the semivariance expression filters the mean, which is especially useful when it is unknown. On the other hand, if $\mathrm{E}[Z(\boldsymbol{s})]$ is unknown and the empirical covariance is computed based on the sample mean $\bar{Z}$, this can induce additional bias in the covariance computation.

Example 95.5: A Box Plot of the Square Root Difference Cloud

The Gaussian form chosen for the semivariogram in the section "Getting Started: VARIOGRAM Procedure" on page 7511 is based on consideration of the plots of the sample semivariogram. For the coal thickness data, the Gaussian form appears to be a reasonable choice.

However, it can often happen 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 SRF $Z(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{R}^2$, the square root difference cloud for a particular direction $\boldsymbol{e}$ is given by

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

for a given lag distance h. In the actual computation, all pairs $P_1 P_2$ of points P_1, P_2 within a distance tolerance around h and an angle tolerance around the direction $\boldsymbol{e}$ 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 section "Getting Started: VARIOGRAM Procedure" on page 7511, this example uses a basic LAGDISTANCE=7, with a distance tolerance of 3.5, and a direction of N–S, with an angle tolerance ATOL=30°.

First, use PROC VARIOGRAM to produce an OUTPAIR= data set. Then use a DATA step to subset this data by choosing pairs within 30° of N–S. In addition, compute lag class and square root difference variables, as the following statements show:

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

ods graphics on;

proc variogram data=thick outp=outp noprint;
   compute novariogram;
   coordinates xc=East yc=North;
   var Thick;
run;

data sqroot;
   set outp;
   /*- Include only points +/- 30 degrees of N-S -------*/
   where abs(cos) < 0.5;
   /*- Unit lag of 7, distance tolerance of 3.5 --------*/
   lag_class=int(distance/7 + 0.5000001);
   sqr_diff=sqrt(abs(v1-v2));
run;

proc sort data=sqroot;
   by lag_class;
run;
```

Next, summarize the results by using the MEANS procedure. The statements follow, and the output is shown in Output 95.5.1.

```
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;
```

Output 95.5.1 Summary of Results

Square Root Difference Cloud Example
Summary of Results

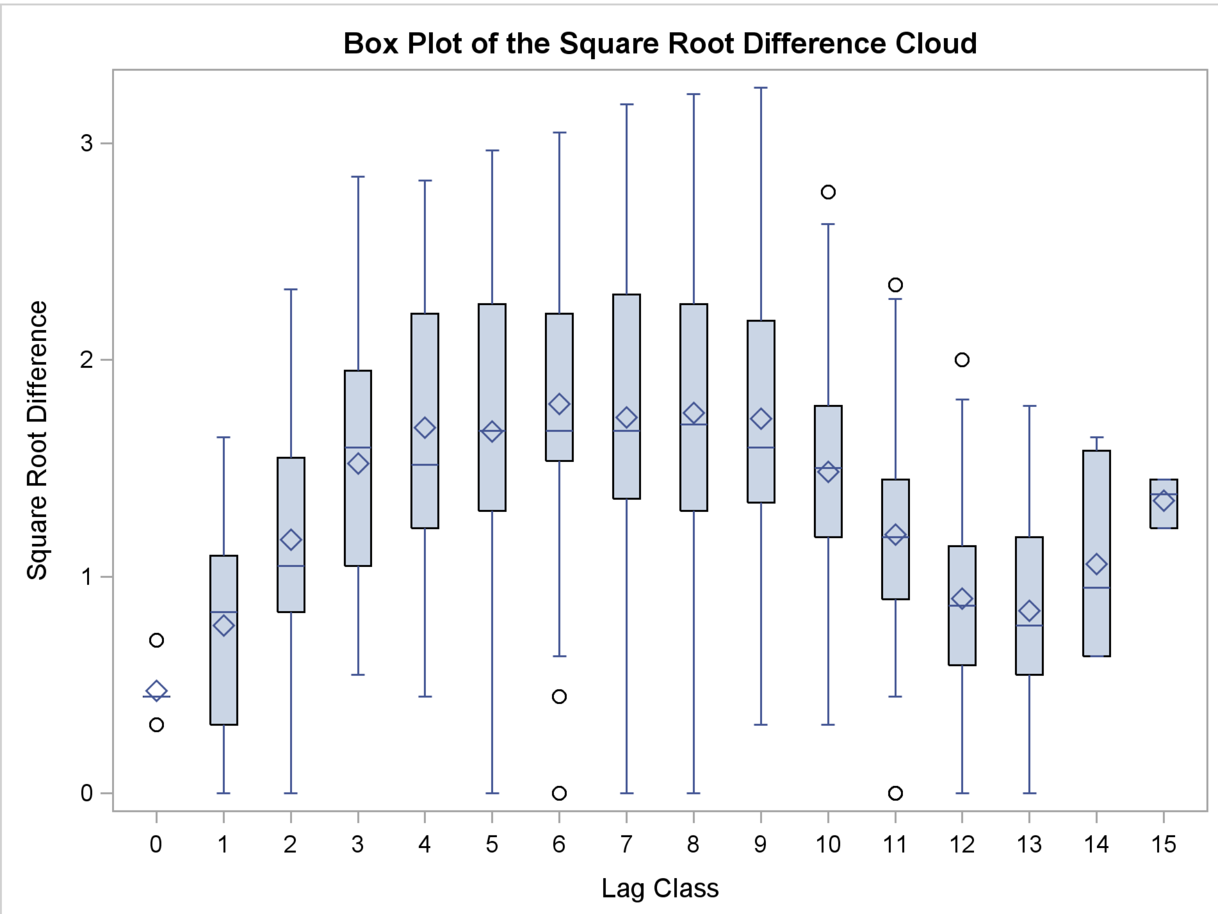
lag_ class	n	mean	std
0	5	0.47300	0.14263
1	31	0.77338	0.41467
2	51	1.17052	0.47800
3	58	1.52287	0.51454
4	65	1.68625	0.58465
5	65	1.66963	0.68582
6	80	1.79693	0.62929
7	88	1.73334	0.73191
8	83	1.75528	0.68767
9	108	1.72901	0.58274
10	80	1.48268	0.48695
11	84	1.19242	0.47037
12	68	0.89765	0.42510
13	38	0.84223	0.44249
14	7	1.05653	0.42548
15	3	1.35076	0.11472

Finally, present the results in a box plot by using the SGPLOT procedure. The box plot facilitates the detection of outliers. The statements are as follows:

```
proc sgplot data=sqroot;
   xaxis label = "Lag Class";
   yaxis label = "Square Root Difference";
   title "Box Plot of the Square Root Difference Cloud";
   vbox sqr_diff /category=lag_class;
run;

ods graphics off;
```

Output 95.5.2 suggests that there do not appear to be any outliers adversely affecting the empirical semivariogram in the N–S direction for the coal seam thickness data. The conclusion from Output 95.5.2 is consistent with our previous semivariogram analysis of the same data set in the section "Getting Started: VARIOGRAM Procedure" on page 7511. The effect of the isolated outliers in lag classes 6 and 10–12 in Output 95.5.2 is demonstrated as the divergence between the classical and robust empirical semivariance estimates in the higher distances in Output 95.8. The difference in these estimates comes from the definition of the robust semivariance estimator $\bar{\gamma}_z(\boldsymbol{h})$ (see the section "Theoretical and Computational Details of the Semivariogram" on page 7536), which imposes a smoothing effect on the outlier influence.

Output 95.5.2 Box Plot of the Square Root Difference Cloud



References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman & Hall/CRC.

Chilès, J. P. and Delfiner, P. (1999), *Geostatistics-Modeling Spatial Uncertainty*, New York: John Wiley & Sons.

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

Cliff, A. D. and Ord, J. K. (1981), *Spatial Processes: Models and Applications*, London: Pion Ltd.

Cressie, N. and Hawkins, D. M. (1980), "Robust Estimation of the Variogram: I," *Mathematical Geology*, 12(2), 115–125.

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

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

Geary, R. C. (1954), "The Contiguity Ratio and Statistical Mapping," *The Incorporated Statistician*, 5, 115–145.

Goovaerts, P. (1997), *Geostatistics for Natural Resources Evaluation*, New York: Oxford University Press.

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

Jian, X., Olea, R. A., and Yu, Y.-S. (1996), "Semivariogram Modeling by Weighted Least Squares," *Computers & Geosciences*, 22(4), 387–397.

Journel, A. G. and Huijbregts, C. J. (1978), *Mining Geostatistics*, New York: Academic Press.

Matheron, G. (1963), "Principles of Geostatistics," *Economic Geology*, 58, 1246–1266.

Moran, P. A. P. (1950), "Notes on Continuous Stochastic Phenomena," *Biometrica*, 37, 17–23.

Olea, R. A. (1999), *Geostatistics for Engineers and Earth Scientists*, Boston: Kluwer Academic.

Olea, R. A. (2006), "A Six-Step Practical Approach to Semivariogram Modeling," *Stochastic Environmental Research and Risk Assessment*, 20(5), 307–318.

Schabenberger, O. and Gotway, C. A. (2005), *Statistical Methods for Spatial Data Analysis*, Boca Raton, FL: Chapman & Hall/CRC.

Stein, M. L. (1988), "Asymptotically Efficient Prediction of a Random Field with a Misspecified Covariance Function," *Annals of Statistics*, 16, 55–63.

Subject Index

Syntax Index

Your Turn

We welcome your feedback.

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