

# **SAS/STAT<sup>®</sup> 12.3 User's Guide**

## **The VARIOGRAM**

### **Procedure**

### **(Chapter)**



This document is an individual chapter from *SAS/STAT® 12.3 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2013. *SAS/STAT® 12.3 User's Guide*. Cary, NC: SAS Institute Inc.

Copyright © 2013, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

**For a Web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

**U.S. Government Restricted Rights Notice:** Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987).

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

July 2013

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

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

# Chapter 102

## The VARIOGRAM Procedure

### Contents

---

|                                                                      |             |
|----------------------------------------------------------------------|-------------|
| Overview: VARIOGRAM Procedure . . . . .                              | <b>8560</b> |
| Introduction to Spatial Prediction . . . . .                         | 8561        |
| Getting Started: VARIOGRAM Procedure . . . . .                       | <b>8561</b> |
| Preliminary Spatial Data Analysis . . . . .                          | 8562        |
| Empirical Semivariogram Computation . . . . .                        | 8567        |
| Autocorrelation Analysis . . . . .                                   | 8570        |
| Theoretical Semivariogram Model Fitting . . . . .                    | 8571        |
| Syntax: VARIOGRAM Procedure . . . . .                                | <b>8576</b> |
| PROC VARIOGRAM Statement . . . . .                                   | 8576        |
| BY Statement . . . . .                                               | 8583        |
| COMPUTE Statement . . . . .                                          | 8584        |
| COORDINATES Statement . . . . .                                      | 8589        |
| DIRECTIONS Statement . . . . .                                       | 8590        |
| ID Statement . . . . .                                               | 8590        |
| MODEL Statement . . . . .                                            | 8590        |
| PARMS Statement . . . . .                                            | 8603        |
| NLOPTIONS Statement . . . . .                                        | 8607        |
| STORE Statement . . . . .                                            | 8607        |
| VAR Statement . . . . .                                              | 8608        |
| Details: VARIOGRAM Procedure . . . . .                               | <b>8608</b> |
| Theoretical Semivariogram Models . . . . .                           | 8608        |
| Characteristics of Semivariogram Models . . . . .                    | 8609        |
| Nested Models . . . . .                                              | 8613        |
| Theoretical and Computational Details of the Semivariogram . . . . . | 8613        |
| Stationarity . . . . .                                               | 8615        |
| Ergodicity . . . . .                                                 | 8615        |
| Anisotropy . . . . .                                                 | 8616        |
| Pair Formation . . . . .                                             | 8616        |
| Angle Classification . . . . .                                       | 8617        |
| Distance Classification . . . . .                                    | 8619        |
| Bandwidth Restriction . . . . .                                      | 8620        |
| Computation of the Distribution Distance Classes . . . . .           | 8621        |
| Semivariance Computation . . . . .                                   | 8625        |
| Empirical Semivariograms and Surface Trends . . . . .                | 8626        |
| Theoretical Semivariogram Model Fitting . . . . .                    | 8627        |
| Parameter Initialization . . . . .                                   | 8630        |

|                                                                                   |             |
|-----------------------------------------------------------------------------------|-------------|
| Parameter Estimates . . . . .                                                     | 8631        |
| Quality of Fit . . . . .                                                          | 8632        |
| Fitting with Matérn Forms . . . . .                                               | 8634        |
| Autocorrelation Statistics . . . . .                                              | 8635        |
| Autocorrelation Weights . . . . .                                                 | 8635        |
| Autocorrelation Statistics Types . . . . .                                        | 8637        |
| Interpretation . . . . .                                                          | 8638        |
| The Moran Scatter Plot . . . . .                                                  | 8639        |
| Computational Resources . . . . .                                                 | 8640        |
| Output Data Sets . . . . .                                                        | 8640        |
| Displayed Output . . . . .                                                        | 8644        |
| ODS Table Names . . . . .                                                         | 8646        |
| ODS Graphics . . . . .                                                            | 8647        |
| Examples: VARIOGRAM Procedure . . . . .                                           | <b>8648</b> |
| Example 102.1: Aspects of Semivariogram Model Fitting . . . . .                   | 8648        |
| Example 102.2: An Anisotropic Case Study with Surface Trend in the Data . . . . . | 8658        |
| Analysis with Surface Trend Removal . . . . .                                     | 8662        |
| Example 102.3: Analysis without Surface Trend Removal . . . . .                   | 8672        |
| Example 102.4: Covariogram and Semivariogram . . . . .                            | 8680        |
| Example 102.5: A Box Plot of the Square Root Difference Cloud . . . . .           | 8684        |
| References . . . . .                                                              | <b>8688</b> |

---

## 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. The procedure also fits permissible theoretical models to the empirical semivariograms, so that you can use them in subsequent analysis to perform spatial prediction. You can produce plots of the empirical semivariograms in addition to plots of the fitted models. Both isotropic and anisotropic continuity measures are available.

PROC VARIOGRAM also provides the Moran's  $I$  and Geary's  $c$  spatial autocorrelation statistics, in addition to the Moran scatter plot to visualize spatial associations within a specified neighborhood around observations. The procedure produces the OUTVAR=, OUTPAIR=, and OUTDISTANCE= data sets that contain information about the semivariogram analysis. Also, the OUTACWEIGHTS= and the OUTMORAN= output data sets contain information about the autocorrelation analysis.

The VARIOGRAM procedure 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 8647.

---

## 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, in addition to modeling in a broad spectrum of fields (for example, environmental health, environmental pollution, natural resources and energy, hydrology, and 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 often 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 that characterizes 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, in addition to 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 that correspond 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 the automated methods provided by the VARIOGRAM procedure. The model characterizes the type of theoretical semivariance function you use to describe spatial dependence in your data set.

Graphical displays are requested by enabling ODS Graphics. 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 8647.

---

## Preliminary Spatial Data Analysis

The following thick data set is available from the Sashelp library. The data set simulates measurements of coal seam thickness (in feet) taken over an approximately square area. The Thick variable has the thickness values in the thick data set. 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.

```

title 'Spatial Correlation Analysis with PROC VARIOGRAM';

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 be produced.

```
ods graphics on;

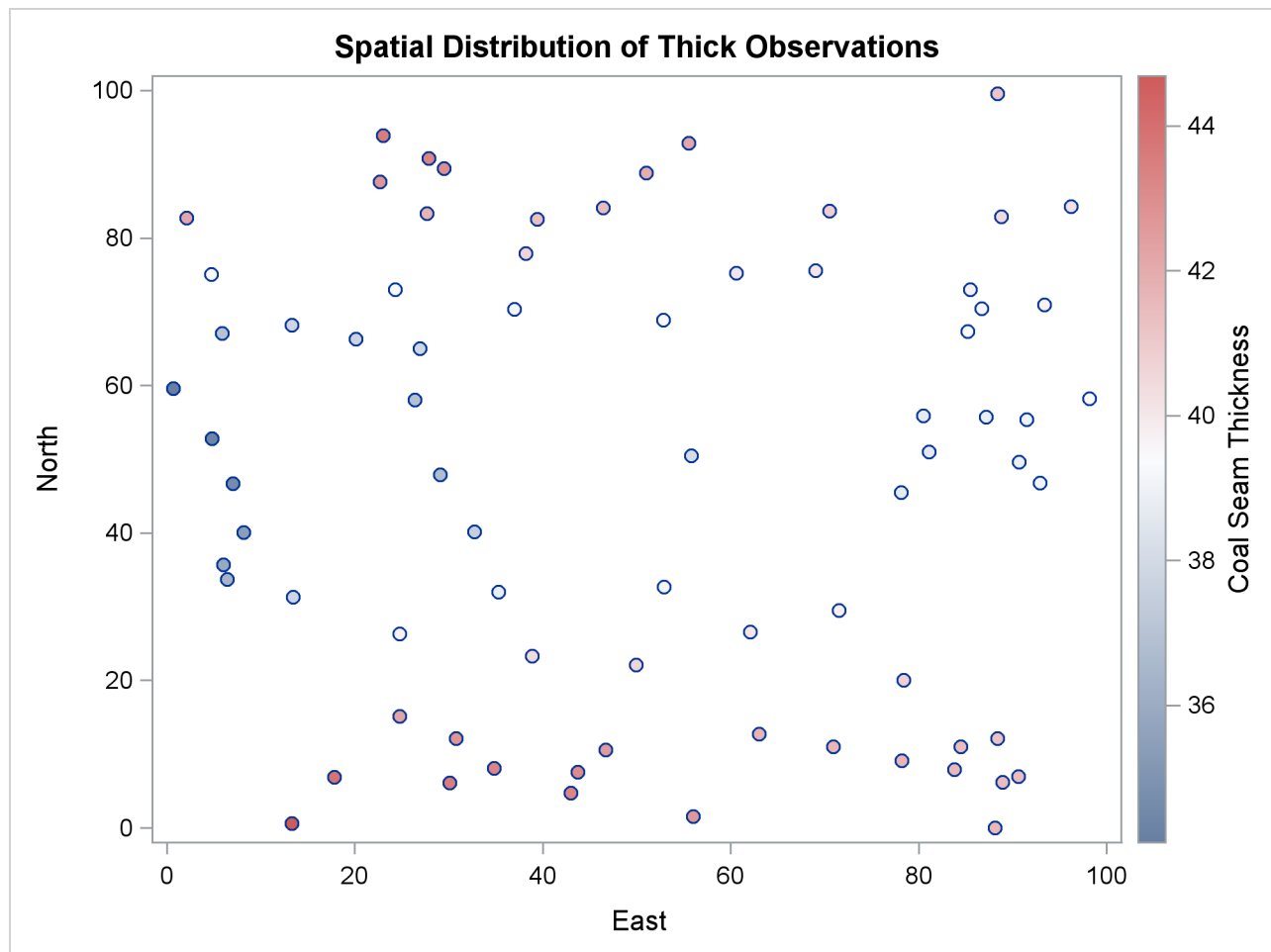
proc variogram data=thick plots=pairs(thr=30);
  compute novariogram nhc=20;
  coordinates xc=East yc=North;
  var Thick;
run;
```

PROC VARIOGRAM produces the table in [Figure 102.1](#) 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 102.1** Number of Observations for the thick Data Set

| Spatial Correlation Analysis with PROC VARIOGRAM |    |
|--------------------------------------------------|----|
| 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 102.2](#). According to the figure, although 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 102.2** 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 102.2](#) 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 that connects any two equidistant pairs of data; this is a property known as isotropy. See “[Example 102.2: An Anisotropic Case Study with Surface Trend in the Data](#)” on page 8658 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 want a sufficient number of distance classes that capture the extent to which your data are correlated and you want each class to 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 8623 for further details.

In the preliminary analysis, you use the option **NHCLASSES=** in the **COMPUTE** statement to help you experiment with these numbers and choose values for the **LAGDISTANCE=** and **MAXLAGS=** options. Here, in particular, you request **NHCLASSES=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 **NHCLASSES=** option, the **NOVARIOGRAM** option produces a pairwise distances table from your observations shown in Figure 102.3, and the corresponding histogram in Figure 102.4. For illustration purposes, you also specify 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.

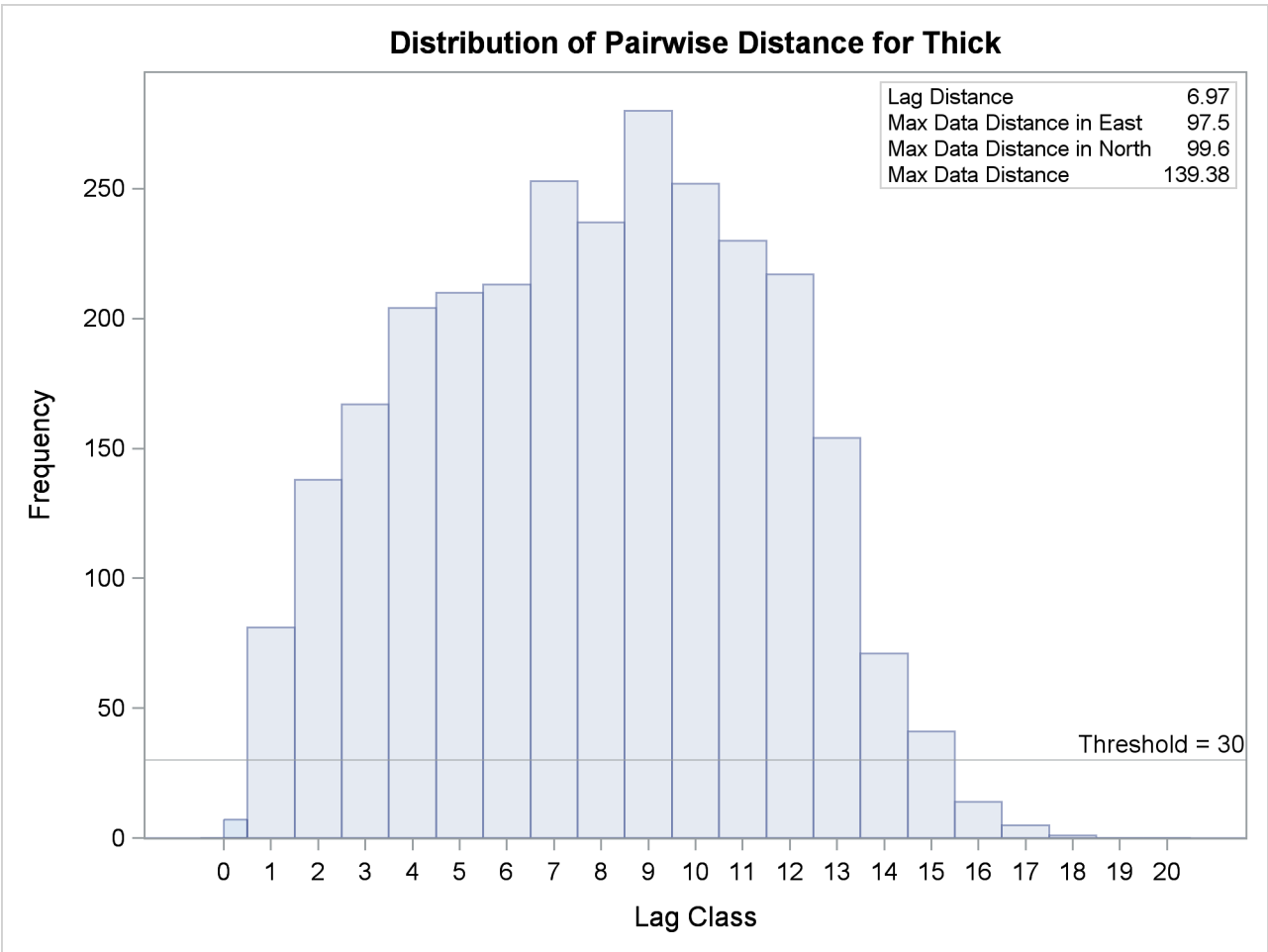
**Figure 102.3** Pairwise Distance Intervals Table

| 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 102.5](#). In particular, the lag distance value is calculated based on your selection of the **NHCLASSES=** option. The last three table entries report the overall maximum distance among your data pairs, in addition to the maximum distances in the main axes directions—that is, the vertical (N–S) axis and the horizontal (E–W) axis. This information is also provided in the inset of [Figure 102.4](#). 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 102.5](#).

**Figure 102.4** Distribution of Pairwise Distances



**Figure 102.5** 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 102.5](#) 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, consider lag classes up to this distance for the empirical semivariogram computations. Given your lag size selection, [Figure 102.3](#) indicates that this distance corresponds to about 10 lags; hence you can set **MAXLAGS=10**.

Overall, for a specific **NHCLASSES=** 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 8624 for more details).

After 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 8584 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 that shows both the classical and the robust empirical semivariograms. See the details of the **PLOTS** option to specify different instances of plots of the empirical semivariogram. The following statements implement the preceding requests:

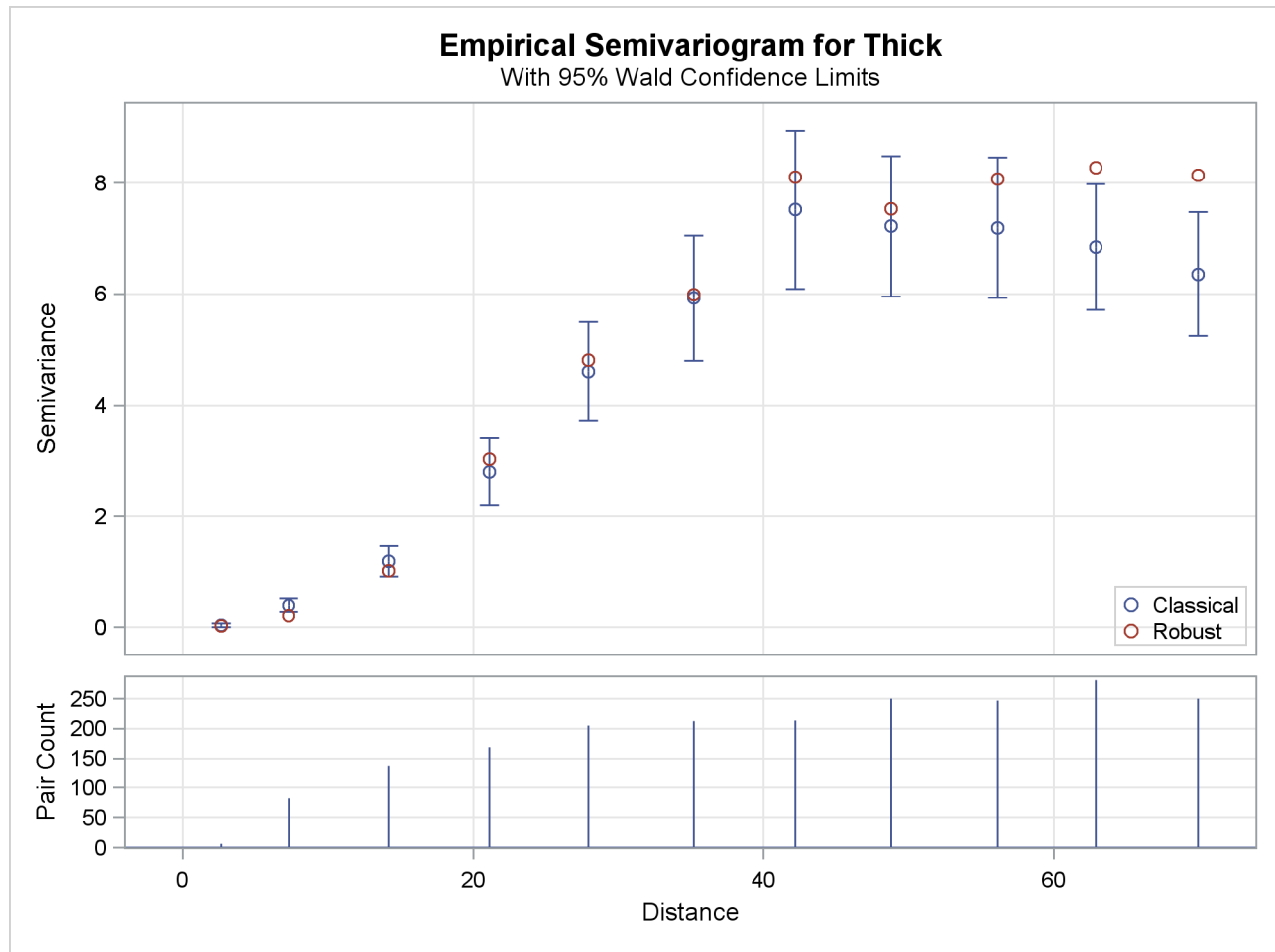
```
proc variogram data=thick outv=outv;  
  compute lagd=7 maxlag=10 cl robust;  
  coordinates xc=East yc=North;  
  var Thick;  
run;
```

Figure 102.6 displays the PROC VARIOGRAM output empirical semivariogram table for the preceding statements. The table displays a total of eleven lag classes, even though you specified `MAXLAGS=10`. The VARIOGRAM procedure always includes a zero lag class in the computations in addition to the `MAXLAGS` classes you request with the `MAXLAGS=` option. Hence, semivariance is actually computed at `MAXLAGS+1` lag classes; see the section “Distance Classification” on page 8619 for more details.

**Figure 102.6** Output Table for the Empirical Semivariogram Analysis

| Spatial Correlation Analysis with PROC VARIOGRAM |               |                     |                        |           |                   |                          |       |
|--------------------------------------------------|---------------|---------------------|------------------------|-----------|-------------------|--------------------------|-------|
| The VARIOGRAM Procedure                          |               |                     |                        |           |                   |                          |       |
| Dependent Variable: Thick                        |               |                     |                        |           |                   |                          |       |
| Empirical Semivariogram                          |               |                     |                        |           |                   |                          |       |
| Lag<br>Class                                     | Pair<br>Count | Average<br>Distance | -----Semivariance----- |           |                   |                          |       |
|                                                  |               |                     | Robust                 | Classical | Standard<br>Error | 95% Confidence<br>Limits |       |
| 0                                                | 7             | 2.64                | 0.028                  | 0.034     | 0.018             | 0                        | 0.069 |
| 1                                                | 82            | 7.29                | 0.210                  | 0.394     | 0.061             | 0.273                    | 0.514 |
| 2                                                | 138           | 14.16               | 1.008                  | 1.179     | 0.142             | 0.901                    | 1.458 |
| 3                                                | 169           | 21.08               | 3.018                  | 2.799     | 0.304             | 2.202                    | 3.396 |
| 4                                                | 205           | 27.93               | 4.811                  | 4.602     | 0.455             | 3.711                    | 5.493 |
| 5                                                | 213           | 35.17               | 5.990                  | 5.928     | 0.574             | 4.802                    | 7.054 |
| 6                                                | 214           | 42.20               | 8.104                  | 7.518     | 0.727             | 6.094                    | 8.943 |
| 7                                                | 250           | 48.78               | 7.533                  | 7.221     | 0.646             | 5.955                    | 8.487 |
| 8                                                | 247           | 56.16               | 8.066                  | 7.195     | 0.647             | 5.926                    | 8.464 |
| 9                                                | 281           | 62.89               | 8.279                  | 6.845     | 0.577             | 5.713                    | 7.976 |
| 10                                               | 250           | 69.93               | 8.144                  | 6.358     | 0.569             | 5.243                    | 7.472 |

Figure 102.7 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.

**Figure 102.7** Classical and Robust Empirical Semivariograms for Coal Seam Thickness Data

The needle plot in the lower part of the [Figure 102.7](#) provides the number of pairs that were used in the computation of the empirical semivariance for each lag class shown. In general, this is a pairwise distribution that is different from the distribution depicted in [Figure 102.4](#). 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.

## Autocorrelation Analysis

You can use the autocorrelation analysis features of PROC VARIOGRAM to compute the autocorrelation Moran's  $I$  and Geary's  $c$  statistics and to obtain the Moran scatter plot. In the following statements, you ask for the Moran's  $I$  and Geary's  $c$  statistics under the assumption of randomization using binary weights, in addition to the Moran scatter plot:

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

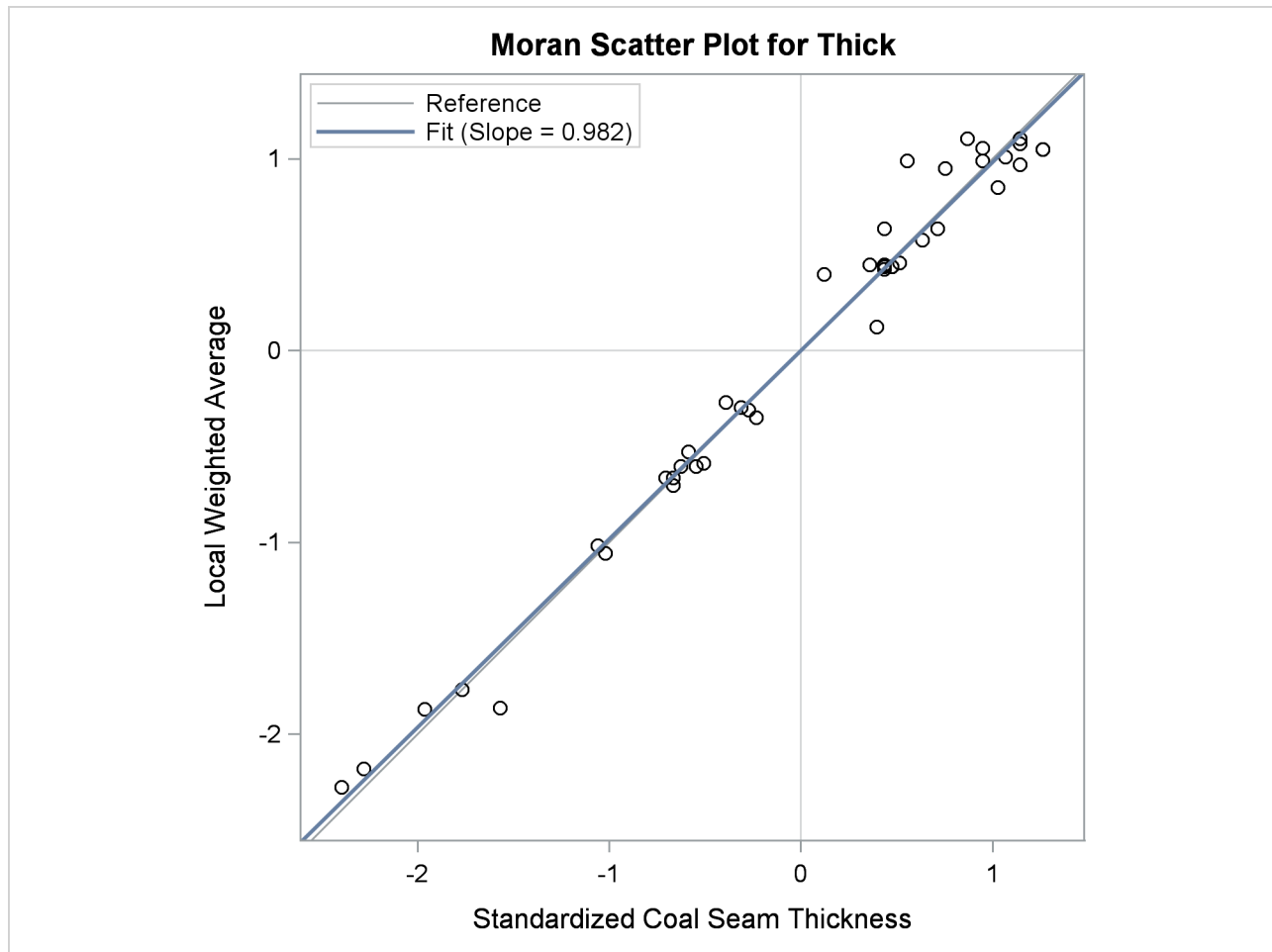
For the autocorrelation analysis with binary weights and the Moran scatter plot, the `LAGDISTANCE=` option indicates that you consider as neighbors of an observation all other observations within the specified distance from it.

Figure 102.8 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. 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 8638 for more details.

**Figure 102.8** Output Table for the Autocorrelation Statistics

| Spatial Correlation Analysis with PROC VARIOGRAM |             |          |          |         |       |         |
|--------------------------------------------------|-------------|----------|----------|---------|-------|---------|
| The VARIOGRAM Procedure                          |             |          |          |         |       |         |
| Dependent Variable: Thick                        |             |          |          |         |       |         |
| Autocorrelation Statistics                       |             |          |          |         |       |         |
| Assumption                                       | Coefficient | Observed | Expected | Std Dev | Z     | Pr >  Z |
| Randomization                                    | Moran's I   | 0.9240   | -0.0244  | 0.145   | 6.53  | <.0001  |
| Randomization                                    | Geary's c   | 0.0162   | 1.0000   | 0.175   | -5.62 | <.0001  |

The requested Moran scatter plot is shown in Figure 102.9. The plot includes all nonmissing observations that have neighbors within the specified `LAGDISTANCE=` distance. The horizontal axis displays the standardized Thick values, and the vertical axis displays the corresponding weighted average of their neighbors. The plot data points are concentrated in the upper right and lower left quadrants defined by the lines  $x = 0$  and  $y = 0$ , and clearly around the axes' diagonal reference line  $y = x$  of slope 1. This fact indicates strong positive spatial association in the thick data set observations. Therefore, for each observation its neighbors within the specified `LAGDISTANCE=` distance have overall similar Thick values to that observation. The plot also displays the linear regression slope, whose value is the Moran's  $I$  coefficient when the binary weights are row-averaged. See the section “The Moran Scatter Plot” on page 8639 for more details about the Moran scatter plot.

**Figure 102.9** Moran Scatter Plot for Coal Seam Thickness Data

## Theoretical Semivariogram Model Fitting

PROC VARIOGRAM features automated semivariogram fitting. In particular, the procedure selects a theoretical semivariogram model to fit the empirical semivariance and produces estimates of the model parameters in addition to a fit plot. You have the option to save these estimates in an item store, which is a binary file format that is defined by the SAS System and that you cannot modify. Then, you can retrieve this information at a later point from the item store for future analysis with PROC KRIGE2D or PROC SIM2D.

The coal seam thickness empirical semivariogram in Figure 102.7 shows first a slow, then rapid, rise from the origin. This behavior suggests that you can approximate the empirical semivariance with 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 8608. Based on this remark, you choose to fit a Gaussian model to your classical semivariogram. Run PROC VARIOGRAM again and

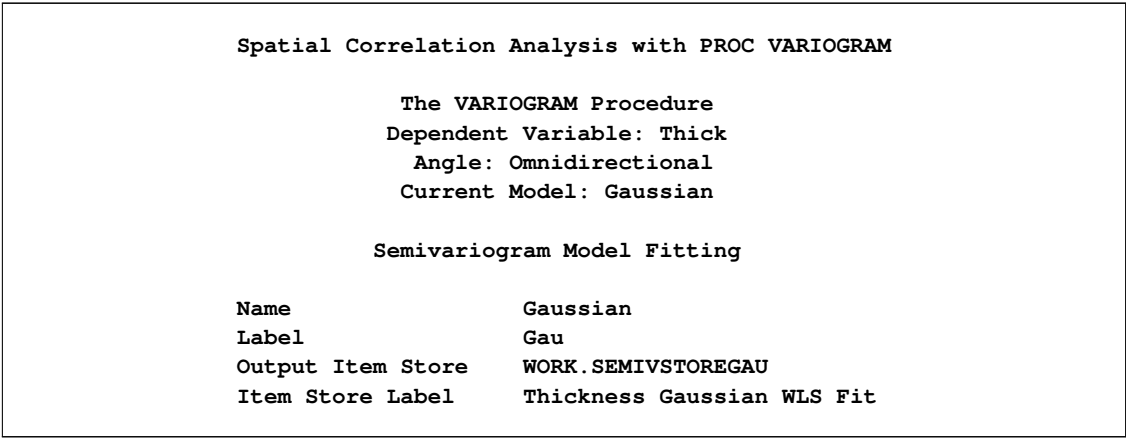
specify the **MODEL** statement with the **FORM=GAU** option. By default, PROC VARIOGRAM uses the weighted least squares (WLS) method to fit the specified model, although you can explicitly specify the **METHOD=** option to request the fitting method. You want additional information about the estimated parameters, so you specify the **CL** option in the **MODEL** statement to compute their 95% confidence limits and the **COVB** option of the **MODEL** statement to produce a table with their approximate covariances. You also specify the **STORE** statement to save the fitting outcome into an item store file with the name SemivStoreGau and a desired label. You run the following statements:

```
proc variogram data=thick outv=outv;
  store out=SemivStoreGau / label='Thickness Gaussian WLS Fit';
  compute lagd=7 maxlag=10;
  coordinates xc=East yc=North;
  model form=gau cl / covb;
  var Thick;
run;

ods graphics off;
```

After you run the procedure you get a series of output objects from the fitting analysis. In particular, Figure 102.10 shows first a model fitting table with the name and a short label of the model that you requested to use for the fit. The table also displays the name and label of the specified item store.

**Figure 102.10** Semivariogram Model Fitting General Information



If you specify no parameters, as in the current example, then PROC VARIOGRAM initializes the model parameters for you with default values based on the empirical semivariance; for more details, see the section “Theoretical Semivariogram Model Fitting” on page 8627. The initial values provided by the VARIOGRAM procedure for the Gaussian model are displayed in the table in Figure 102.11.



**Figure 102.11** Semivariogram Fitting Model Information

| Model Information |               |
|-------------------|---------------|
| Parameter         | Initial Value |
| Nugget            | 0             |
| Scale             | 6.7992        |
| Range             | 34.9635       |

Otherwise, in PROC VARIOGRAM you can specify initial values for parameters with the **PARMS** statement. Alternatively, you can specify fixed values for the model scale and range with the **SCALE=** and **RANGE=** options, respectively, in the **MODEL** statement. A nugget effect is always used in model fitting. Unless you explicitly specify a fixed nugget effect with the **NUGGET=** option in the **MODEL** statement or initialize the nugget parameter in the **PARMS** statement, the nugget effect is automatically initialized to zero. See the section “Syntax: VARIOGRAM Procedure” on page 8576 for more details about how the **MODEL** statement and the **PARMS** statement handle model parameters.

The output in Figure 102.12 comes from the optimization process that takes place during the model parameter estimation. The optimizer produces an optimization information table, information about the optimization technique that is used, optimization-related results, and notification about the optimization convergence.

**Figure 102.12** Fitting Optimization Information

| Optimization Information                                         |                   |                          |              |
|------------------------------------------------------------------|-------------------|--------------------------|--------------|
| Optimization Technique                                           | Dual Quasi-Newton |                          |              |
| Parameters in Optimization                                       | 3                 |                          |              |
| Lower Boundaries                                                 | 3                 |                          |              |
| Upper Boundaries                                                 | 0                 |                          |              |
| Starting Values From                                             | PROC              |                          |              |
| Spatial Correlation Analysis with PROC VARIOGRAM                 |                   |                          |              |
| The VARIOGRAM Procedure                                          |                   |                          |              |
| Dependent Variable: Thick                                        |                   |                          |              |
| Angle: Omnidirectional                                           |                   |                          |              |
| Current Model: Gaussian                                          |                   |                          |              |
| Dual Quasi-Newton Optimization                                   |                   |                          |              |
| Dual Broyden - Fletcher - Goldfarb - Shanno Update (DBFGS)       |                   |                          |              |
| Hessian Computed by Finite Differences (Using Analytic Gradient) |                   |                          |              |
| Optimization Results                                             |                   |                          |              |
| Iterations                                                       | 12                | Function Calls           | 45           |
| Gradient Calls                                                   | 0                 | Active Constraints       | 1            |
| Objective Function                                               | 11.433894152      | Max Abs Gradient Element | 3.0128744E-8 |
| Slope of Search Direction                                        | -3.986332E-8      |                          |              |
| Convergence criterion (GCONV=1E-8) satisfied.                    |                   |                          |              |

The fitting process is successful, and the parameters converge to the estimated values shown in [Figure 102.13](#). For each parameter, the same table also displays the approximate standard error, the degrees of freedom, the  $t$  value, the approximate  $p$ -value, and the requested 95% confidence limits.

**Figure 102.13** Semivariogram Fitting Parameter Estimates

| Parameter Estimates |          |                     |                                      |         |    |         |                   |
|---------------------|----------|---------------------|--------------------------------------|---------|----|---------|-------------------|
| Parameter           | Estimate | Approx<br>Std Error | Approximate 95%<br>Confidence Limits |         | DF | t Value | Approx<br>Pr >  t |
|                     |          |                     | Lower                                | Upper   |    |         |                   |
| Nugget              | 0        | 0                   | 0                                    | 0       | 8  | .       | .                 |
| Scale               | 7.4599   | 0.2621              | 6.8555                               | 8.0643  | 8  | 28.46   | <.0001            |
| Range               | 30.1111  | 1.1443              | 27.4724                              | 32.7498 | 8  | 26.31   | <.0001            |

The approximate covariance matrix of the estimated parameters is displayed in [Figure 102.14](#).

**Figure 102.14** Approximate Covariance Matrix of Parameter Estimates

| Approximate Covariance Matrix |        |        |        |
|-------------------------------|--------|--------|--------|
| Parameter                     | Nugget | Scale  | Range  |
| Nugget                        | 0.0000 | 0.0000 | 0.0000 |
| Scale                         | 0.0000 | 0.0687 | 0.2326 |
| Range                         | 0.0000 | 0.2326 | 1.3094 |

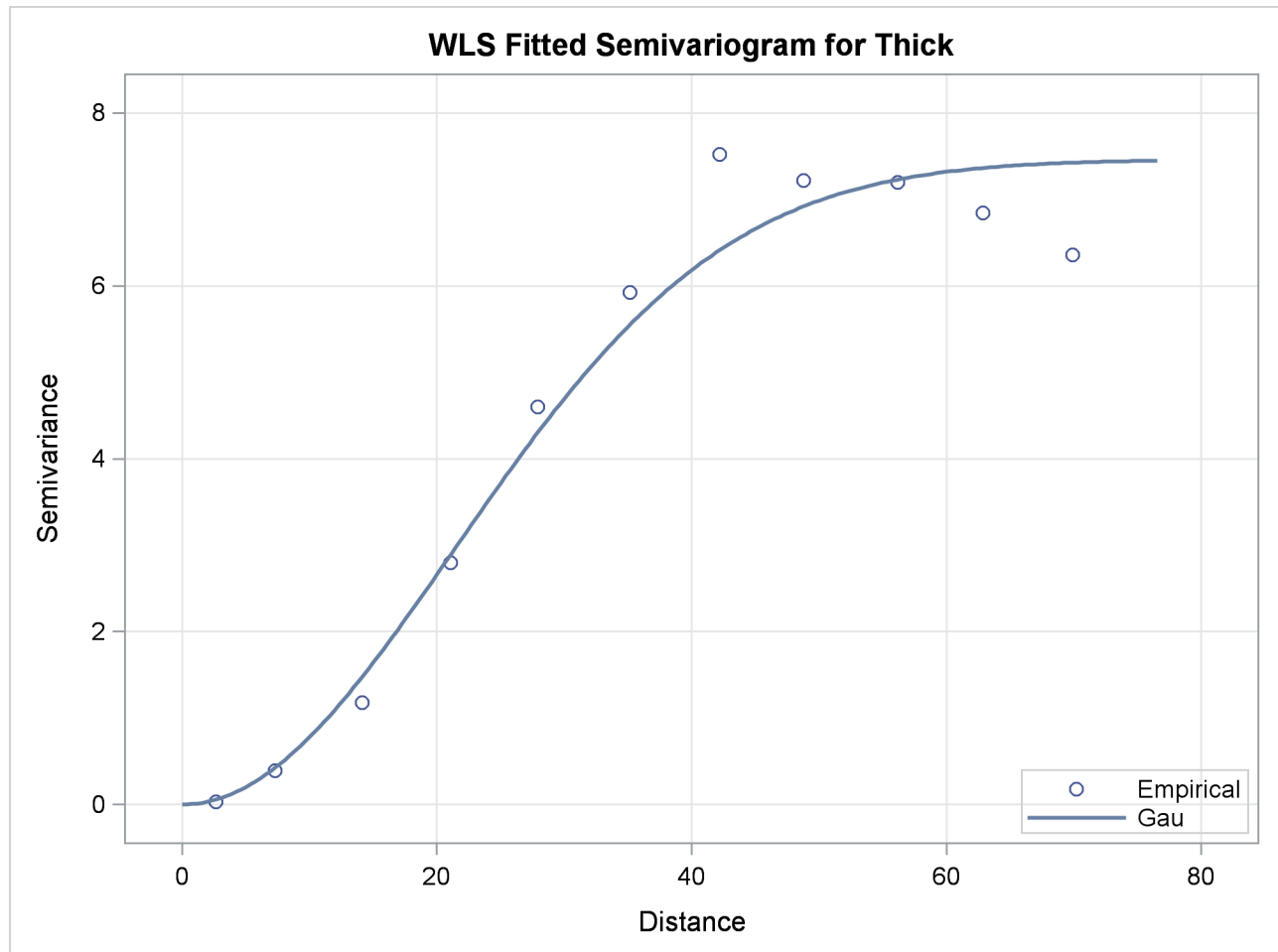
The fitting summary table in [Figure 102.15](#) displays statistics about the quality of the fitting process. In particular, the table shows the weighted error sum of squares in the Weighted SSE column and the Akaike information criterion in the AIC column. See more information about the fitting criteria in section “Quality of Fit” on page 8632.

**Figure 102.15** Semivariogram Model Fitting Summary

| Fit Summary |                 |         |
|-------------|-----------------|---------|
| Model       | Weighted<br>SSE | AIC     |
| Gau         | 11.43389        | 6.42556 |

Figure 102.16 demonstrates the fitted theoretical semivariogram against the empirical semivariance estimates with the weighted least squares method. The fit seems to be more accurate closer to the origin  $h = 0$ , and this is explained as follows: A smaller  $h$  corresponds to smaller semivariance; in turn, this corresponds to smaller semivariance variance, as shown in the section “Theoretical and Computational Details of the Semivariogram” on page 8613. By definition, the WLS optimization weights increase with decreasing variance, which leads to a more accurate fit for smaller distances  $h$  in the WLS fitting results.

**Figure 102.16** Fitted Theoretical and Empirical Semivariogram for Coal Seam Thickness



## Syntax: VARIOGRAM Procedure

The following statements are available in the VARIOGRAM procedure:

```

PROC VARIOGRAM options ;
  BY variables ;
  COMPUTE computation-options ;
  COORDINATES coordinate-variables ;
  DIRECTIONS directions-list ;
  ID variable ;
  MODEL model-options ;
  PARMS parameters-list < / parameters-options > ;
  NLOPTIONS < options > ;
  STORE store-options ;
  VAR analysis-variables-list ;

```

The **COMPUTE** and **COORDINATES** statements are required. The **MODEL** and **PARMS** statements are hierarchical. If you specify a **PARMS** statement, it must follow a **MODEL** statement.

## PROC VARIOGRAM Statement

```

PROC VARIOGRAM options ;

```

The PROC VARIOGRAM statement invokes the VARIOGRAM procedure. Table 102.1 summarizes the options available in the PROC VARIOGRAM statement.

**Table 102.1** PROC VARIOGRAM Statement Options

| Option               | Description                                                              |
|----------------------|--------------------------------------------------------------------------|
| <b>DATA=</b>         | Specifies the input data set                                             |
| <b>IDGLOBAL</b>      | Labels observations across BY groups using ascending observation numbers |
| <b>IDNUM</b>         | Labels observations using the observation number                         |
| <b>NOPRINT</b>       | Suppresses normal display of results                                     |
| <b>OUTACWEIGHTS=</b> | Specifies a data set to store autocorrelation weights information        |
| <b>OUTDISTANCE=</b>  | Specifies a data set to store summary distance information               |
| <b>OUTMORAN=</b>     | Specifies a data set to store Moran scatter plot information             |
| <b>OUTPAIR=</b>      | Specifies a data set to store pairwise point information                 |
| <b>OUTVAR=</b>       | Specifies a data set to store spatial continuity measures                |
| <b>PLOTS</b>         | Specifies the plot display and options                                   |

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

**DATA=SAS-data-set**

specifies a SAS data set that contains the *x* and *y* coordinate variables and the VAR statement variables.

**IDGLOBAL**

specifies that ascending observation numbers be used across BY groups for the observation labels in the appropriate output data sets and the **OBSERVATIONS** plot, instead of resetting the observation number in the beginning of each BY group. The IDGLOBAL option is ignored if no BY variables are specified. Also, if you specify the **ID** statement, then the IDGLOBAL option is ignored unless you also specify the IDNUM option in the **PROC VARIOGRAM** statement.

**IDNUM**

specifies that the observation number be used for the observation labels in the appropriate output data sets and the **OBSERVATIONS** plot. The IDNUM option takes effect when you specify the **ID** statement; otherwise, it is ignored.

**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:** This option temporarily disables the Output Delivery System (ODS); see the section “**ODS Graphics**” on page 8647 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. Use this option with caution when the DATA= data set is large. If  $n$  denotes the number of observations in the DATA= data set, then the OUTACWEIGHTS= data set contains  $[n(n - 1)]/2$  observations.

See the section “**OUTACWEIGHTS=SAS-data-set**” on page 8640 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 8641 for details.

**OUTMORAN=SAS-data-set****OUTM=SAS-data-set**

specifies a SAS data set in which to store information that is illustrated in the Moran plot, namely the standardized value of each observation in the DATA= data set and the weighted average of its local neighbors. You must also specify the **LAGDISTANCE=** and **AUTOCORRELATION** options in the **COMPUTE** statement; otherwise, the OUTMORAN= data set request is ignored.

The OUTMORAN= data set is useful when you want to save the information that is illustrated in the Moran scatter plot. The data set can also contain entries of missing observations with neighbors, although these observations are not displayed in the Moran plot. However, if the only observations with neighbors in your input data set are observations with missing values, then the OUTMORAN= output data set is empty.

See the section “**OUTMORAN=SAS-data-set**” on page 8641 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.

Use this option with caution when your DATA= data set is large. Assume that your DATA= data set has  $n$  observations. When you specify the **NOVARIOGRAM** option in the **COMPUTE** statement, the OUTPAIR= data set is populated with all  $[n(n-1)]/2$  pairs that can be formed with the  $n$  observations.

If the **NOVARIOGRAM** option is not specified, then the OUTPAIR= data set contains only pairs of data that are located within a certain distance away from each other. Specifically, it contains pairs whose distance between observations belongs to a lag class up to the specified **MAXLAGS=** option in the **COMPUTE** statement. Then, depending on your specification of the **LAGDISTANCE=** and **MAXLAGS=** options, the OUTPAIR= data set might contain  $[n(n-1)]/2$  or fewer pairs.

Finally, you can restrict the number of pairs in the OUTPAIR= data set 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 8642 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 8643 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(c1a unpack) semivar semivar(rob))
```

ODS Graphics must be enabled before plots can be requested. For example:

```
ods graphics on;

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

ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section “**Enabling and Disabling ODS Graphics**” on page 600 in Chapter 21, “**Statistical Graphics Using ODS.**”

If ODS Graphics is enabled but you omit 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, in addition to the histogram of the pairwise distances of your data. For an example of the observations plot, see [Figure 102.2](#). For an example of the pairwise distances plot, see [Figure 102.4](#).
- If you omit **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 also 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 semivariogram plot, see [Output 102.7](#). Moreover, if you specify the **MODEL** statement and perform model fitting, then PROC VARIOGRAM also produces a fit plot of the fitted semivariogram. An example of the fit plot is shown in [Figure 102.16](#).

The following *global-plot-options* are available:

#### **ONLY**

suppresses the default plots. Only plots that are 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 coordinates of the axes have equal size units.

#### **FITPLOT** <(fitplot-options)>

##### **FIT** <(fitplot-options)>

requests a plot that shows the model fitting results against the empirical semivariogram. By default, **FITPLOT** displays one plot of the fitted model (or a panel of plots for different angles in the anisotropic case).

If you specify the **FORM=AUTO** option in the **MODEL** statement, then each class of equivalent fitted models is displayed with a different curve on the plot. The best fitting model class is chosen based on the criteria that you specify in the **CHOOSE** option of the **MODEL** statement, and a

thicker line on top of any other curve is shown for it. The plot legend shows the ranked classes by displaying the label of the representative model of each class in the plot. If appropriate, the number of additional models in the same equivalence class also shows within parentheses.

You can specify the following *fitplot-options*:

**NCLASSES=***number*

**NCLASSES=ALL**

specifies the maximum number of classes to display on the fit plot, where *number* is a positive integer. The default is NCLASSES=5 for nonpaneled plots and NCLASSES=3 for paneled plots. The option takes effect when you specify the **FORM=AUTO** option in the **MODEL** statement, and it is ignored when you fit one single model. If you specify NCLASSES=ALL or a larger number than the available classes, then all available classes are shown on the fit plot. If you specify multiple instances of the NCLASSES= option, then only the last specified instance is honored.

**UNPACK**

suppresses paneling in paneled fit plots. By default, fit plots appear in a panel, when appropriate.

**MORAN** < (*moran-options*) >

**MOR** < (*moran-options*) >

produces a Moran scatter plot of the observations with nonmissing values. For more details about this plot, see the section “[The Moran Scatter Plot](#)” on page 8639. In addition to the Moran scatter plot points, the plot also displays the fit line for the linear regression of the weighted average on the standardized observation values, the regression fit line slope, and a reference line with slope equal to 1. The MORAN plot has the following *moran-options*:

**LABEL** < ( *label-options* ) >

labels the observations. The label is the ID variable if the **ID** statement is specified; otherwise, it is the observation number. The *label-options* can be one or more of the following:

**HH**

specifies that labels show for observations in the upper right (high-high) plot quadrant of positive spatial association.

**HL**

specifies that labels show for observations in the lower right (high-low) plot quadrant of negative spatial association.

**LH**

specifies that labels show for observations in the upper left (low-high) plot quadrant of negative spatial association.

**LL**

specifies that labels show for observations in the lower left (low-low) plot quadrant of positive spatial association.

If you specify multiple instances of the MORAN option and you specify the LABEL suboption in any of those, then the resulting Moran scatter plot displays the observations labels. By default, when you specify none of the *label-options*, the PLOTS=MORAN(LABEL) request puts labels in all observations.



**ROWAVG=***rowavg-option*

specifies the flag value for row-averaging of weights in the computation of the weighted average. The *rowavg-option* can be either of the following:

**OFF**

specifies that autocorrelation weights not be row-averaged.

**ON**

specifies that row-averaged autocorrelation weights be used.

The default behavior is ROWAVG=ON. If you specify the ROWAVG= option more than once in the same MORAN plot request, then the behavior is set to ROWAVG=ON unless any of the instances is ROWAVG=OFF.

When you specify the PLOTS=MORAN option, you must specify both the [AUTOCORRELATION](#) and the [LAGDISTANCE=](#) options in the [COMPUTE](#) statement to produce the Moran scatter plot. For more information about the plot, see the section “[The Moran Scatter Plot](#)” on page 8639.

**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 is 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.

**LABEL** < ( *label-option* ) >

labels the observations. The label is the ID variable if the [ID](#) statement is specified; otherwise, it is the observation number. The *label-option* can be one of the following:

**EQ=***number*

specifies that labels show for any observation whose value is equal to the specified *number*.

**MAX=***number*

specifies that labels show for observations with values smaller than or equal to the specified *number*.

**MIN=***number*

specifies that labels show for observations with values equal to or greater than the specified *number*.

If you specify multiple instances of the OBSERVATIONS option and you specify the LABEL suboption in any of those, then the resulting observations plot displays the observations labels. If more than one *label-option* is specified in multiple LABEL suboptions, then the prevailing *label-option* in the resulting OBSERVATIONS plot emerges by adhering to the choosing order: MIN, MAX, EQ.

**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 displays the observations with missing values.

If you omit any of the `GRADIENT`, `OUTLINE`, and `OUTLINEGRADIENT` suboptions, the `OUTLINEGRADIENT` is the default suboption. If you specify multiple instances of the `OBSERVATIONS` option or multiple suboptions for `OBSERVATIONS`, then the resulting observations plot honors 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 that is 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 102.22](#).

**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 that is 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 is created if you specify the `PAIRS` option within a `PLOTS` option. If you specify multiple instances of the `PAIRS` option, the resulting plot has the following features:

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

**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 of observations in groups that are 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. If you specify more than one BY statement, only the last one specified is used.

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 NOTSORTED or DESCENDING option 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 BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in 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.

Table 102.2 summarizes the options available in the COMPUTE statement.

**Table 102.2** COMPUTE Statement Options

| Option          | Description                                                                                 |
|-----------------|---------------------------------------------------------------------------------------------|
| ALPHA=          | Specifies the confidence level                                                              |
| ANGLETOLERANCE= | Specifies the tolerance                                                                     |
| AUTOCORRELATION | Calculates autocorrelation statistics                                                       |
| BANDWIDTH=      | Specifies the bandwidth                                                                     |
| CL              | Requests confidence limits                                                                  |
| DEPSILON=       | Specifies the distance value for declaring that two distinct points are zero distance apart |
| LAGDISTANCE=    | Specifies the basic distance unit that defines the lags                                     |
| LAGTOLERANCE=   | Specifies the tolerance around the LAGDISTANCE= value                                       |
| MAXLAGS=        | Specifies the maximum number of lag classes to be used                                      |
| NDIRECTIONS=    | Specifies the number of angle classes to be used                                            |
| NHCLASSES=      | Specifies the number of distance classes to be used                                         |
| NOVARIOGRAM     | Prevents the computation of the continuity measures                                         |
| OUTPDISTANCE=   | Specifies the cutoff distance for writing observations to the OUTPAIR= data set             |
| ROBUST          | Calculates a robust version of the semivariance                                             |

### **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 in (0, 1), and the confidence level is  $1 - \text{number}$ . 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  $180^\circ / (2n_d)$ , where  $n_d$  is the NDIRECTIONS= specification. If you do not specify the NDIRECTIONS= option or the DIRECTIONS statement, ANGLETOLERANCE= is ignored.

See the section “Theoretical and Computational Details of the Semivariogram” on page 8613 for further information.

**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 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's *c* statistics.

**MORAN | MOR**

specifies use of the Moran's *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 8621. By default, nonnormalized values are used in the computations.

**POWER**=number**POW**=number

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**=number**SCA**=number

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](#)” on page 8635 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’s  $I$  and Geary’s  $c$  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 are 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 8635 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 102.23](#) for a visual representation of the bandwidth.

If you omit the `BANDWIDTH=` option, no restriction occurs. If you omit the `NDIRECTIONS=` option or the `DIRECTIONS` statement, `BANDWIDTH=` is ignored.

## CL

requests confidence limits for the classical semivariance estimate. The lower bound of the confidence limits 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= $\Delta\epsilon$` , then pairs of points  $P_1$  and  $P_2$  for which the distance between them  $|P_1 P_2| < \Delta\epsilon$  are excluded from the continuity measure calculations. The default value of the `DEPSILON=` option is 100 times the machine precision; this product is approximately  $1\text{E}-10$  on most computers.

**LAGDISTANCE=***distance-unit*

**LAGDIST=***distance-unit*

**LAGD=***distance-unit*

specifies the basic distance unit that defines 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. In this case, a distance tolerance of  $\delta x$  accommodates 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 `NHCLASSES=` option.

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

See the section “[Theoretical and Computational Details of the Semivariogram](#)” on page 8613 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 8613 for more details.

If you omit 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 to be used in constructing the continuity measures in addition to a zero lag class; see also the section “[Distance Classification](#)” on page 8619. This option excludes any pair of points  $P_1$  and  $P_2$  for which the distance between them,  $|P_1 P_2|$ , exceeds the MAXLAGS= value times the [LAGDISTANCE=](#) value.

You can determine candidate values for the MAXLAGS= option by plotting or displaying the [OUT-DISTANCE=](#) 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 in which 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  $0^\circ$  (or its reciprocal,  $180^\circ$ ),  $60^\circ$  (or its reciprocal,  $240^\circ$ ), and  $120^\circ$  (or its reciprocal,  $300^\circ$ ). 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  $\theta = 180^\circ/n_d$ , and the angle classes are

$$(k\theta - \delta\theta, k\theta + \delta\theta) \quad k = 0, \dots, n_d - 1$$

If you omit 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 8590, provides greater control over the angle classes.

See the section “[Theoretical and Computational Details of the Semivariogram](#)” on page 8613 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 NHCLASSES= option is used to produce the distance intervals table, the histogram of pairwise distances, and the [OUT-DISTANCE=](#) data set. See the [OUTDISTANCE=](#) option, the section “[OUTDIST=SAS-data-set](#)”



on page 8641, and the section “Theoretical and Computational Details of the Semivariogram” on page 8613 for more information.

The default value is NHCLASSES=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 omit the **OUTPDISTANCE=** option, all distinct pairs are written. This option is ignored if you omit the **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 that contains 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 that contains 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 that contains 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  $45^\circ$  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  $\theta_1 = 0^\circ$ ,  $\theta_2 = 60^\circ$ , and  $\theta_3 = 120^\circ$ , with corresponding angle tolerances  $\delta\theta_1 = 22.5^\circ$ ,  $\delta\theta_2 = 12.5^\circ$ , and  $\delta\theta_3 = 22.5^\circ$ , 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);
```

---

## ID Statement

**ID** *variable* ;

The ID statement specifies which variable to include for identification of the observations in the **OUTPAIR=** and the **OUTACWEIGHTS=** output data sets. The ID statement variable is also used for the labels and tool tips in the **OBSERVATIONS** plot.

In the VARIOGRAM procedure you can specify only one ID variable in the ID statement. If no ID statement is given, then PROC VARIOGRAM uses the observation number in the data sets and the **OBSERVATIONS** plot.

---

## MODEL Statement

**MODEL** *fitting-options* < / *model-options* > ;

You specify the MODEL statement if you want to fit a theoretical semivariogram model to the empirical semivariogram data that are produced in the **COMPUTE** statement. You must have nonmissing empirical semivariogram estimates at a minimum of three lags to perform model fitting.

Table 102.3 summarizes the options available in the MODEL statement.

**Table 102.3** MODEL Statement Options

| Option                 | Description                                                          |
|------------------------|----------------------------------------------------------------------|
| <b>Fitting Options</b> |                                                                      |
| ALPHA=                 | Specifies confidence level                                           |
| CHOOSE=                | Ranks the fitted models and chooses the optimally fit model          |
| CL                     | Constructs a <i>t</i> -type confidence interval                      |
| EQUIVTOL=              | Specifies a positive upper value tolerance                           |
| FIT=                   | Specifies which type of empirical semivariogram to fit               |
| FORM=                  | Specifies the functional form (type) of the semivariogram model      |
| MDATA=                 | Specifies the input data set                                         |
| METHOD=                | Fits a theoretical model to the empirical semivariance               |
| NEPSILON=              | Adds a minimal nugget effect                                         |
| NUGGET=                | Specifies the nugget effect                                          |
| RANGE=                 | Specifies the range parameter                                        |
| RANGELAG=              | Uses consecutive nonmissing empirical semivariance lags to fit model |
| RANKEPS=               | Specifies the minimum threshold to compare fit quality               |
| SCALE=                 | Specifies the scale parameter in semivariogram models                |
| SMOOTH=                | Specifies the positive smoothness parameter $\nu$                    |
| <b>Model Options</b>   |                                                                      |
| COVB                   | Requests the covariance matrix                                       |
| CORRB                  | Requests the approximate correlation matrix                          |
| DETAILS                | Produces different levels of output                                  |
| GRADIENT               | Displays the gradient of the objective function                      |
| MTOGTOL=               | Specifies a threshold value for the smoothness parameter             |
| NOFIT                  | Suppresses the model fitting process                                 |
| NOITPRINT              | Suppresses the display of the iteration history table                |

You can choose to perform a fully automated fitting or to fit one model with specific forms. In the first case you simply specify a list of forms or no forms at all. All suitable combinations are tested, and the result is the model that produces the best fit according to specified criteria. In the second case you specify one theoretical semivariogram model, and you have more control over its parameters for the fitting process.

Furthermore, you can specify a theoretical semivariogram model in two ways:

- You explicitly specify the **FORM** option and any of the options **SCALE**, **RANGE**, and **NUGGET** in the MODEL statement.
- You can specify an **MDATA=** data set. This data set contains variables that correspond to the **FORM** option and to any of the options **SCALE**, **RANGE**, **NUGGET**, and **SMOOTH**. You can also use an **MDATA=** data set to request a fully automated fitting.

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

The MODEL statement has the following *fitting-options*:

**ALPHA=***number*

requests that a *t*-type confidence interval be constructed for each of the fitting parameters with confidence level  $1 - \textit{number}$ . The value of *number* must be in (0, 1); the default is 0.05 which corresponds to the default confidence level of 95%. If the CL option of the MODEL statement is not specified, then ALPHA= is ignored.

**CHOOSE=***criterion*

**CHOOSE=**(*criterion1* . . . *criterionk*)

specifies that if the fitting task has more than one model to fit, then PROC VARIOGRAM ranks the fitted models and chooses the optimally fit model according to one or more available criteria.

If you want to use multiple fitting criteria, then the order in which you specify them in the CHOOSE= option defines how they are applied. This feature is useful when fitting suggests that two or more models perform equally well according to a certain criterion. For example, if two models are equivalent according to *criterioni*, then they are further ranked in the list based on the next criterion, *criterionj*, where  $j = i + 1$ .

Each *criterion* can be one of the following:

**AIC**

specifies Akaike's information criterion.

**SSE**

specifies the weighted sum of squares error for each fitted model when METHOD=WLS, and the residual sum of squares error for each fitted model when METHOD=OLS.

**STATUS**

classifies models based on their fitting process convergence status. CHOOSE=STATUS places on top models for which the fitting process is successful.

By default, the models are ranked in the fit summary table with the best fitted model at the top of the list, based on the criteria that you specify in the CHOOSE= option. This model is the fit choice of PROC VARIOGRAM for the particular fitting task. If you omit the CHOOSE= option, then the default behavior is CHOOSE=(SSE AIC).

Regardless of the specified fitting criteria, models for which the fitting process is unsuccessful always appear at the bottom of the fit summary table. For more details about the fitting criteria, see the section "Fitting Criteria" on page 8632. After multiple models are ranked, they are further categorized in classes of equivalence depending on whether any two models calculate the same semivariance value at the same distance for a series of different distances. For more details, see the section "Classes of Equivalence" on page 8634.

If you specify the same criterion multiple times in the CHOOSE= option, then only the first instance is used for the ranking process and any additional ones are ignored. If you specify only one model to fit in the MODEL statement and you specify the CHOOSE= option, then the option is ignored.

**CL**

requests that *t*-type confidence limits be constructed for each of the fitting parameters estimates. The confidence level is 0.95 by default; this can be changed with the **ALPHA=** option of the **MODEL** statement.

**EQUIVTOL=etol-value****ETOL=etol-value**

specifies a positive upper value tolerance to use when categorizing multiple models in classes of equivalence. For this categorization, the VARIOGRAM procedure computes the sum of absolute differences of semivariances for pairs of consecutively ranked models. If the sum is lower than the EQUIVTOL= value for any such model pair, then these two models are deemed to be equivalent. As a result, the EQUIVTOL= option can affect the number and size of classes of equivalence in the fit summary table. Smaller values of the EQUIVTOL= parameter result in a more strict model comparison and can lead to a higher number of classes of equivalence. For more details, see the section “Classes of Equivalence” on page 8634.

The default value for the EQUIVTOL= parameter is  $10^{-3}$ . The EQUIVTOL= option applies when you fit multiple models with the **FORM=AUTO** option of the **MODEL** statement; otherwise, it is ignored.

The EQUIVTOL= option is independent of the ranking results from the **RANKEPS=** option of the **MODEL** statement. This means that you could possibly have models listed but not ranked in the fit summary table, and still have equivalence classes assigned according to the order in which the models appear in the table.

**FIT=fit-type-options**

specifies which type of empirical semivariogram to fit. You can choose between the following *fit-type-options*:

**CLASSICAL****CLA**

fits a model for the classical empirical semivariance.

**ROBUST****ROB**

fits a model for the robust empirical semivariance. This option can be used only when the **ROBUST** option is specified in the **COMPUTE** statement.

The default value is **FIT=CLASSICAL**.

**FORM=form****FORM=(form1, ..., formk)****FORM=AUTO** (*auto-options*)

specifies the functional form (type) of the semivariogram model. The supported structures are two-parameter models that use the sill and range as parameters. The Matérn model is an exception that makes use of a third smoothing parameter  $\nu$ .

The **FORM=** option is required when you specify the **MODEL** statement. You can perform fitting of a theoretical semivariogram model either explicitly or in an automated manner. For the explicit specification you specify suitable model forms in the **FORM=** option. For an automated fit you specify the **FORM=****AUTO** option which has the **AUTO(MLIST=)** and **AUTO(NEST=)** suboptions. You can read more details in the following two subsections.

### Explicit Model Specification

You can explicitly specify a theoretical semivariogram model to fit by using any combination of one, two, or three forms. Use the syntax with the single *form* to specify a non-nested model. Use the syntax with  $k$  structures *form<sub>i</sub>*,  $i = 1, \dots, k$ , to specify up to three nested structures ( $k \leq 3$ ) in a semivariogram model. Each of the forms can be any of the following:

**CUBIC | EXPONENTIAL | GAUSSIAN | MATERN |  
PENTASPHERICAL | POWER | SINEHOLEEFFECT | SPHERICAL  
CUB | EXP | GAU | MAT | PEN | POW | SHE | SPH**  
specify a form.

All of these forms are presented in more detail in the section “Theoretical Semivariogram Models” on page 8608. In addition, you can optionally specify a nugget effect for your model with the **NUGGET** option in the **MODEL** statement.

For example, the syntax

**FORM=GAU**

specifies a model with a single Gaussian structure. Also, the syntax

**FORM= (EXP , SHE , MAT)**

specifies a nested model with an exponential, a sine hole effect, and a Matérn structure. Finally

**FORM= (EXP , EXP)**

specifies a nested model with two structures both of which are exponential.

**NOTE:** In the documentation, models are named either by using their full names or by using the first three letters of their structures. Also, the names of different structures in a nested model are separated by a hyphen (-). According to this convention, the previous examples illustrate how to specify a GAU, an EXP-SHE-MAT, and an EXP-EXP model, respectively, with the **FORM=** option.

When you explicitly specify the types of structures, you can fix parameter values or ask PROC VARIOGRAM to select default initial values for the forms parameters by using the **SCALE**, **RANGE**, **NUGGET**, and **SMOOTH** options. You can set your own, non-default initial parameter values by using the **PARMS** statement in combination with an explicitly specified semivariogram model in the **MODEL** statement.

### Automated Model Selection

Use the **FORM=****AUTO** option to request the highest level of automation in the best fit selection of the parameters. If you specify **FORM=****AUTO**, any of the **SCALE**, **RANGE**, or **SMOOTH** options that are also specified are ignored. When you specify the **FORM=****AUTO** option, you cannot specify the **PARMS**

statement for the corresponding [MODEL](#) statement. As a result, when you use the FORM=AUTO option, you cannot fix any of the model parameters and PROC VARIOGRAM sets initial values for them.

The AUTO option has the following *auto-options*:

**MLIST=***mform*

**MLIST=**(*mform1*, ..., *mformp*)

specifies one or more different model forms to use in combinations during the model fitting process. If you omit the MLIST= suboption, then combinations are made among all available model types. The *mform* can be any of the following eight forms:

**CUBIC | EXPONENTIAL | GAUSSIAN | MATERN |  
PENTASPHERICAL | POWER | SINEHOLEEFFECT | SPHERICAL  
CUB | EXP | GAU | MAT | PEN | POW | SHE | SPH**  
specify a form.

If you use more than one *mform*, then each *mform<sub>i</sub>*,  $i = 1, \dots, p$  must be different from the others in the group of  $p \leq 8$  forms that you specify.

**NEST=***nest-list*

specifies the number of nested structures to use for the fitting. You can choose between the following to specify the *nest-list*:

*n*                      a single value

*m* TO *n*              a sequence in which *m* equals the starting value and *n* equals the ending value

For example,

**NEST=1**

produces the best fit with one single model among all model types specified in the [MLIST=](#) suboption. Also,

**NEST=2 TO 3**

produces the best fit among all combinations of the model types specified in the [MLIST=](#) suboption that result in nested models with two or three structures. The combinations that are tested include repetitions. Hence, if you specify, for example,

**MODEL FORM=AUTO (MLIST=(EXP, SPH) NEST=1 TO 2)**

then the different models that are tested are equivalent to the specifications FORM=EXP, FORM=SPH, FORM=(EXP,EXP), FORM=(EXP,SPH), FORM=(SPH,SPH) and FORM=(SPH,EXP). **NOTE:** The models EXP-SPH and SPH-EXP are taken as two separate models. Although they are mathematically equivalent (see the section “[Nested Models](#)” on page 8613), PROC VARIOGRAM assigns different initial values to the model structures in each case, which can lead to different fitting results. (See the section “[Example 102.1: Aspects of Semivariogram Model Fitting](#)” on page 8648.)

If you omit the NEST suboption, then by default PROC VARIOGRAM searches for the best fit with up to three nested structures in a model. The default behavior is equivalent to

```
NEST=1 TO 3
```

In the VARIOGRAM procedure you can use a maximum of three nested structures to fit an empirical semivariogram; that is,  $n \leq 3$ .

You can use the AUTO value for the form in the **MDATA=** data set, and also in the **FORM=** option. However, in the former case the automation functionality is limited compared to the latter case and the *auto-options* of the **FORM=AUTO** option. In particular, when you specify the form to be AUTO in the **MDATA=** data set, then PROC VARIOGRAM follows only the default behavior and searches among all available forms for the best fit with up to three nested structures in a model.

#### **MDATA=SAS-data-set**

specifies the input data set that contains parameter values for the covariance or semivariogram model. The **MDATA=** data set must contain a variable named **FORM**, and it can optionally include any of the variables **SCALE**, **RANGE**, **NUGGET**, and **SMOOTH**.

The **FORM** variable must be a character variable. It accepts only the **AUTO** value or the *form* values that can be specified in the **FORM=** option in the **MODEL** statement. The **RANGE**, **SCALE**, **NUGGET**, and **SMOOTH** variables must be numeric or missing.

The number of observations present in the **MDATA=** data set corresponds to the level of nesting of the semivariogram model. Each observation line describes a structure of the model you submit for fitting.

If you specify the **AUTO** value for the **FORM** variable in an observation, then you cannot specify additional nested structures in the same data set, and any parameters you specify in the same structure are ignored. In that case, PROC VARIOGRAM performs a crude automated search among all available forms to obtain the best fit with up to three nested structures in a model. You can refine this type of search with additional suboptions when you perform it with the **FORM=AUTO** option instead of the **MDATA=** option in the **MODEL** statement.

When you have a nested model, you might want to specify parameter values for only some of the nested structures. In this case, you must specify the corresponding parameter values for the remaining model structures as missing values.

For example, you can use the following **DATA** step to specify a non-nested model that uses a spherical covariance within an **MDATA=** data set:

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



Then, you can use the `md1` data in the **MODEL** statement of PROC VARIOGRAM as shown in the following statements:

```
proc variogram data=...;
  compute ...;
  model mdata=md1;
run;
```

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

```
proc variogram data=...;
  compute ...;
  model form=sph scale=25 range=10;
run;
```

The following data set `md2` is an example of a nested model:

```
data md2;
  input form $ scale range nugget smooth;
  datalines;
SPH 20 8 5 .
MAT 12 3 5 0.7
GAU . 1 5 .
;
```

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

```
proc variogram data=...;
  compute ....;
  model form=(sph,mat,gau)
        scale=(20,12,.) range=(8,3,1) smooth=0.7 nugget=5;
run;
```

Use the **SMOOTH** variable column in the `MDATA=` data set to specify the smoothing parameter  $\nu$  in the Matérn semivariogram models. The **SMOOTH** variable values must be positive and no greater than 1,000,000. PROC VARIOGRAM sets this upper limit for numerical and performance reasons. In any case, if the fitting process leads the smoothness value to exceed the default threshold value 10,000, then the VARIOGRAM procedure converts the Matérn form into a Gaussian form and repeats the model fitting. To adjust the switching threshold value, you can use the **MTOGTOL=** option in the **MODEL** statement.

If you specify a **SMOOTH** column in the `MDATA=` data set, then its elements are ignored except for the rows in which the corresponding **FORM** is Matérn.

The **NUGGET** variable value is the same for all nested structures. This is the way to specify a nugget effect in the `MDATA=` data set. If you specify more than one nugget value for different structures, then the last nugget value specified is used.

**METHOD=***method-options*

must be specified in the MODEL statement to fit a theoretical model to the empirical semivariance. The METHOD option has the following suboptions:

**OLS**

specifies that ordinary least squares be used for the fitting.

**WLS**

specifies that weighted least squares be used for the fitting.

The default is METHOD=WLS.

**NEPSILON=***min-nugget-factor***NEPS=***min-nugget-factor*

specifies that a minimal nugget effect be added to the theoretical semivariance in the unlikely occasion that the theoretical semivariance becomes zero during fitting with weighted least squares. As explained in the section “[Theoretical and Computational Details of the Semivariogram](#)” on page 8613, the theoretical semivariance is always positive for any distance larger than zero. If a conflicting situation emerges as a result of numerical fitting issues, then the NEPSILON= option can help you alleviate the problem by adding a minimal variance at the distance lag where the issue is encountered. For more details, see the section “[Parameter Initialization](#)” on page 8630.

If you omit the NEPSILON= option, then PROC VARIOGRAM sets a default value of  $10^{-6}$ . If a minimal nugget effect is used, its value is case-specific and is based on the *min-nugget-factor*. Specifically, its value is defined as *min-nugget-factor* times the sample variance of the input data set, or as *min-nugget-factor* when the sample variance is equal to zero.

**NUGGET=***number*

specifies the nugget effect for the model. The nugget effect is due to a discontinuity in the semivariogram as determined by plotting the sample semivariogram; see “[Theoretical Semivariogram Models](#)” on page 8608 for more details. The NUGGET= parameter is a nonnegative number. If you specify a nonmissing value, then it is used as a fixed parameter in the fitting process.

PROC VARIOGRAM assigns a default initial value for the nugget effect in the following cases:

- if you specify a missing value.
- if you omit the NUGGET= option and you do not specify an associated [PARMS](#) statement with initial values for the nugget.

The NUGGET= option is incompatible with the specification of the [PARMS](#) statement for the corresponding [MODEL](#) statement.

**RANGE=***range***RANGE=***(range1, ..., rangek)*

specifies the range parameter in semivariogram models. The RANGE= option is optional. However, if you specify the RANGE= option, then you must provide range values for all structures that you have specified explicitly in the [FORM=](#) option. All nonmissing range values are considered as fixed parameters. PROC VARIOGRAM assigns a default initial value to any of the model structures for which you specify a missing range value. PROC VARIOGRAM assigns default initial values to all model structures if you omit the RANGE= option, unless you specify an associated [PARMS](#) statement and initial values for the range in it.

The range parameter is a positive number, has the units of distance, and is related to the correlation scale of the underlying spatial process.

**NOTE:** If you specify this parameter for a power model, then it does not correspond to a range. For power models, the parameter you specify in the RANGE option is a dimensionless power exponent whose value must range within [0,2) so that the power model is a valid semivariance function.

The RANGE= option is ignored when you specify the FORM=AUTO option. The RANGE= option is incompatible with the specification of the PARMS statement for the corresponding MODEL statement.

**RANGELAG=***rlag-list*

**RLAG=***rlag-list*

specifies that you prefer to use the range of consecutive nonmissing empirical semivariance lags in the *rlag-list* for the semivariogram fitting process, instead of using all MAXLAGS+1 lag classes by default. You can specify *rlag-list* in either of the following forms:

- |                      |                                                                                                                                                                                                                                                                                                      |
|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>k</i>             | a single value that designates the width of the selected lag range by starting at lag zero. You must use at least three lags to perform model fitting, so you can specify <i>k</i> within [3, . . . , MAXLAGS+1].                                                                                    |
| <i>m</i> TO <i>n</i> | a sequence in which <i>m</i> equals the starting lag and <i>n</i> equals the ending lag. The parameters <i>m</i> and <i>n</i> must be nonnegative integer numbers to designate lag classes between zero and MAXLAGS. Use at least three lags for model fitting; hence it holds that $n - m \geq 2$ . |

The following two brief examples exhibit the use of the RANGELAG option. These examples assume that you have set the MAXLAGS= option to 9 or higher to indicate nonmissing empirical semivariance estimates at 10 lags or more.

In the first example,

**RANGELAG=8**

uses the empirical semivariance in the first eight lags to fit a theoretical model. Hence, RANGELAG=8 uses only the lag classes zero to seven. This approach enables you to account only for the correlation behavior described by the first *k* empirical semivariogram lag classes.

In the second example,

**RANGELAG=2 TO 9**

specifies that the empirical semivariance values at lag classes zero, one, and after lag class nine are excluded from the model fitting process.

**RANKEPS=***reps-value*

**REPS=***reps-value*

specifies the minimum threshold to compare fit quality of two models for a specific criterion. Beyond this threshold the criterion values become insensitive to comparison. In particular, when you fit multiple models, PROC VARIOGRAM computes for each one the value of the fitting criterion specified in the **CHOOSE=** option of the **MODEL** statement. These values are examined in pairs at the sorting stage. If the difference of a given pair exceeds the *reps-value*, then the sort order of the corresponding models is reversed; otherwise, the two models retain their relative order in the rankings. Hence, the **RANKEPS=** option can affect model ranking in the fit summary table.

The default value for the **RANKEPS=** parameter is  $10^{-6}$  and accounts for the default optimization convergence tolerance at the fitting stage prior to model ranking. The convergence tolerance itself limits the accuracy that you can use to compare two models under a given criterion. As a result, smaller values of the **RANKEPS=** parameter might not lead to a sensible and more strict model comparison because for a smaller *reps-value*, ranking could depend on digits beyond the accuracy limit.

In the opposite end, if the specified *reps-value* turns out to be large compared to the criterion value differences, then it can make the sorting process insensitive to the specified sorting criterion. When this happens, the fit summary table ranking reflects only the order in which different models are examined in the procedure flow. You can tell whether the criterion is bypassed; if it is, then one or more values of the specified criterion might not appear to be sorted in the fit summary table.

The **RANKEPS=** parameter must be a positive number. The **RANKEPS=** option applies when you fit multiple models with the **FORM=AUTO** option of the **MODEL** statement; otherwise, it is ignored.

**SCALE=***scale*

**SCALE=**(*scale1*, ..., *scalek*)

specifies the scale parameter in semivariogram models. The **SCALE=** option is optional. However, if you specify the **SCALE=** option, then you must provide sill values for all structures that you have specified explicitly in the **FORM=** option. All nonmissing scale values are considered as fixed parameters. PROC VARIOGRAM assigns a default initial value to any of the model structures for which you specify a missing scale value. PROC VARIOGRAM assigns default initial values to all model structures if you omit the **SCALE=** option, unless you specify an associated **PARMS** statement with initial values for scale.

The scale parameter is a positive number. It has the same units as the variance of the variable in the **VAR** statement. The scale of each structure in a semivariogram model represents the variance contribution of the structure to the total model variance.

In power models the **SCALE=** parameter does not correspond to a sill because the power model has no sill. Instead, PROC VARIOGRAM uses the **SCALE=** option to designate the slope (or scaling factor) in power model forms. The power model slope has the same variance units as the variable in the **VAR** statement.

The **SCALE=** option is ignored when you specify the **FORM=AUTO** option. The **SCALE=** option is incompatible with the specification of the **PARMS** statement for the corresponding **MODEL** statement.

**SMOOTH=***smooth*

**SMOOTH=**(*smooth1*, ..., *smoothm*)

specifies the positive smoothness parameter  $\nu$  in the Matérn type of semivariance structures. The special case  $\nu = 0.5$  is equivalent to the exponential model, whereas the theoretical limit  $\nu \rightarrow \infty$  gives the Gaussian model.

The SMOOTH= option is optional. When you specify an explicit model in the **FORM=** option with  $m$  Matérn structures, you can provide up to  $m$  smoothness values. You can specify a value for *smoothi*,  $i = 1, \dots, m$  that is positive and no greater than 1,000,000. PROC VARIOGRAM sets this upper limit for the SMOOTH= option values for numerical and performance reasons. In any case, if the fitting process leads the smoothness value to exceed the default threshold value 10,000, then the VARIOGRAM procedure converts the Matérn form into a Gaussian form and repeats the model fitting. To adjust the switching threshold value, you can use the **MTOGTOL=** option in the **MODEL** statement.

If you specify fewer than  $m$  values, then the remaining Matérn structures have their smoothness parameters initialized to missing values. If you specify more than  $m$  values, then values in excess are ignored.

All nonmissing smoothness values are considered as fixed parameters of the corresponding Matérn structures. PROC VARIOGRAM assigns a default initial value to any of the model Matérn structures, if any, for which you specify a missing smoothness value. PROC VARIOGRAM assigns default initial values to all model Matérn structures if you omit the SMOOTH= option, unless you specify an associated **PARMS** statement and initial values for smoothness in it.

The SMOOTH= option is ignored when you specify the **FORM=AUTO** option. The SMOOTH= option is incompatible with the specification of the **PARMS** statement for the corresponding **MODEL** statement.

In addition to the *fitting-options*, you can specify the following *model-options* after a slash (/) in the **MODEL** statement.

### COVB

requests the approximate covariance matrix for the parameter estimates of the model fitting. The COVB option is ignored when you also specify the **DETAILS=ALL** option.

When you specify an explicit model with the **FORM=** option in the **MODEL** statement, the COVB option produces the requested approximate covariance matrix. When you specify the **FORM=AUTO** option in the **MODEL** statement, by default the COVB option produces output only for the selected model, where the choice is based on the criteria that you specify in the **CHOOSE=** option of the **MODEL** statement. If you specify the **DETAILS** option in addition to **FORM=AUTO** in the **MODEL** statement, then the COVB option produces output for each one of the fitted models.

### CORRB

requests the approximate correlation matrix for the parameter estimates of the model fitting. The CORRB option is ignored when you also specify the **DETAILS=ALL** option.

When you specify an explicit model with the **FORM=** option in the **MODEL** statement, the CORRB option produces the requested approximate correlation matrix. When you specify the **FORM=AUTO** option in the **MODEL** statement, by default the CORRB option produces output only for the selected model, where the choice is based on the criteria that you specify in the **CHOOSE=** option of the **MODEL** statement. If you specify the **DETAILS** option in addition to **FORM=AUTO** in the **MODEL** statement, then the CORRB option produces output for each one of the fitted models.

**DETAILS** <= *detail-level*>

requests different levels of output to be produced during the fitting process. You can specify any of the following *detail-level* arguments:

**MOD**

specifies that the default output for all candidate models be produced when the **FORM=**[AUTO](#) option is specified in the **MODEL** statement. If you fit only one explicit model, then the **DETAILS=MOD** option has no effect and is ignored.

**ITR**

requests that a complete iteration history be produced in addition to the default output. The output for **DETAILS=ITR** includes the current values of the parameter estimates, their gradients, and additional optimization statistics.

**ALL**

requests the most detailed level of output when fitting a model. Specifically, except for the default output, the **DETAILS=ALL** option produces optimization statistics in addition to the combined output of the **DETAILS=ITR**, **COVB**, and **CORRB** options.

When you fit multiple models with the **FORM=**[AUTO](#) option in the **MODEL** statement, only the selected model default output is produced. The model selection is based on the criteria that you specify in the **CHOOSE=** option of the **MODEL** statement. With the **DETAILS** option you can produce ODS tables with information about the fitting process of all the models that you fit. Moreover, you can produce output at different levels of detail that you can specify with the *detail-level* argument.

Omitting the **DETAILS** option or specifying the **DETAILS** option without any argument is equivalent to specifying **DETAILS=MOD**.

**GRADIENT**

displays the gradient of the objective function with respect to the parameter estimates in the “Parameter Estimates” table.

**MTOGTOL**=*number***MTOL**=*number*

specifies a threshold value for the smoothness parameter of the Matérn form. Above this threshold, a Matérn form in a model switches to the Gaussian form. The *number* value must be positive and no greater than 1,000,000, which is the smoothness upper bound set by the VARIOGRAM procedure.

By default, if the fitting process progressively increases the Matérn smoothness parameter  $\nu$  without converging to a smoothness estimate, then PROC VARIOGRAM converts the Matérn form into a Gaussian form when smoothness exceeds the default value 10,000. If you specify the *number* value to be greater than the 1,000,000 boundary value, then it is ignored and reset to the default threshold value. For more details about the Matérn-to-Gaussian form conversion, see the section “[Fitting with Matérn Forms](#)” on page 8634.

**NOFIT**

suppresses the model fitting process.

**NOITPRINT**

suppresses the display of the iteration history table when you have also specified the **DETAILS=ITR** or **DETAILS=ALL** option in the **MODEL** statement. Otherwise, the **NOITPRINT** option is ignored.

## PARMS Statement

**PARMS** (*value-list*)...< / *options* > ;

The PARMS statement specifies initial values for the semivariance parameters of a single specified model in the **MODEL** statement. Alternatively, the PARMS statement can request a grid search over several values of these parameters. You must specify the values by starting with the nugget effect parameter. You continue in the order in which semivariogram forms are specified in the **FORM=** option of the **MODEL** statement by specifying for each structure the values for its scale, range, and any other parameters as applicable.

The PARMS statement is optional and must follow the associated **MODEL** statement.

The *value-list* specification can take any of several forms:

|                                  |                                                                                                                             |
|----------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| <i>m</i>                         | a single value                                                                                                              |
| $m_1, m_2, \dots, m_n$           | several values                                                                                                              |
| <i>m</i> to <i>n</i>             | a sequence in which <i>m</i> equals the starting value, <i>n</i> equals the ending value, and the increment equals 1        |
| <i>m</i> to <i>n</i> by <i>i</i> | a sequence in which <i>m</i> equals the starting value, <i>n</i> equals the ending value, and the increment equals <i>i</i> |
| $m_1, m_2$ to $m_3$              | mixed values and sequences                                                                                                  |

You can use the PARMS statement to input fixed values for parameters and also initial values that you want to optimize.

Suppose that you want to fit a semivariogram model with a Matérn component of scale 3, range 20, smoothing parameter 4.5, and an exponential component of unspecified scale and range 15. Assume that you also want to fix all the specified parameter values for the optimization. Including the nugget effect, you have a model with six parameters.

In terms of the PARMS statement, your specifications mean that you have initial values for the second, third, fourth, and sixth parameter in the parameter list. Also, the same specifications imply that you provide no initial values for the first parameter (which corresponds to the nugget effect) and the fifth parameter (which corresponds to the exponential model scale). For these parameters you prefer that PROC VARIOGRAM selects initial values, instead. Since you must specify values for all model parameters in the PARMS statement, you simply specify missing values for the first and fifth parameter. This is the way to request that PROC VARIOGRAM assigns default initial values to parameters. The SAS statements to implement these specifications are as follows:

```
proc variogram data=FirstData;
  < other VARIOGRAM statements >
  model form=(mat,exp);
  parms (.) (3) (20) (4.5) (.) (15) / hold=(2 to 4,6);
run;
```



**NOTE:** The preceding statements are equivalent to the following ones in which the PARMS statement is omitted:

```
proc variogram data=FirstData;
  < other VARIOGRAM statements >
  model form=(mat,exp) scale=(3,.) range(20,15) smooth=4.5;
run;
```

This example might suggest that you can always use either the PARMS or the [MODEL](#) statement to specify the same fitting parameters in the VARIOGRAM procedure. However, the PARMS statement gives you more flexibility in two ways:

- You can set non-default initial parameter values by using the PARMS statement, whereas in the [MODEL](#) statement you can request default initial values only by setting parameters to missing values. For this reason the PARMS statement cannot be specified when the [FORM=AUTO](#) option is specified in the associated [MODEL](#) statement. As an example, the following statements do not have an equivalent without using the PARMS statement, because the first parameter in the PARMS statement list (which corresponds to the [NUGGET](#) parameter) is set to the specific initial value of 2.1 and the fifth parameter (which corresponds to the exponential structure scale) is set to the specific initial value of 0.3.

```
proc variogram data=FirstData;
  < other VARIOGRAM statements >
  model form=(mat,exp);
  parms (2.1) (3) (20) (4.5) (0.3) (15) / hold=(2 to 4,6);
run;
```

- In the [MODEL](#) statement all the nonmissing parameter values that you specify remain fixed. Instead, the PARMS statement considers all values in the specified parameter sets to be subjected to optimization unless you force values to be fixed with the [HOLD=](#) option. In the previous example, you can specify that you want to optimize all of your parameters by skipping the [HOLD=](#) option as shown in the following modified statements:

```
proc variogram data=FirstData;
  < other VARIOGRAM statements >
  model form=(mat,exp);
  parms (2.1) (3) (20) (4.5) (1) (15);
run;
```

When you omit the PARMS statement list and the [PDATA=](#) data set in a PARMS statement, the specification is equivalent to a PARMS statement list where all the parameters have missing initial values. However, if you specify no other option in the PARMS statement, then the PARMS statement is ignored.

In order to avoid ambiguity, you cannot specify the PARMS statement if any of the scale, range, nugget, or smoothness parameters has been specified in the associated [MODEL](#) statement either explicitly or in the [MDATA=](#) data set. This condition is in effect even when you specify an empty PARMS statement.

If you specify more than one set of initial values, a grid of initial values sets is created. PROC VARIOGRAM seeks among the specified sets for the one that gives the lowest objective function value. Then, the procedure uses the initial values in the selected set for the fitting optimization.



The results from the PARMS statement are the values of the parameters on the specified grid. The ODS name of the “Parameter Search” table is ParmSearch.

You can specify the following options after a slash (/) in the PARMS statement:

**HOLD=***value-list*

**EQCONS=***value-list*

specifies which parameter values be constrained to equal the specified values. For example, the following statement constrains the first and third semivariance parameters to equal 0.5 and 12, respectively. The fourth parameter is fixed to the default initial value that is assigned to it by PROC VARIOGRAM.

```
parms (0.5) (3) (12) (.) / hold=1,3,4;
```

The HOLD= option accepts only nonmissing values in its list. If you specify more than the available parameters in the HOLD= option list, then the ones in excess are ignored. If the HOLD= option list has integer values that do not correspond to variables in the [PARMS](#) list, then they are also ignored. Noninteger values are rounded to the closest integer and evaluated accordingly.

When you specify more than one set of parameter initial values, the HOLD= option list applies to the set that gives the lowest objective function value before this set is sent to the optimizer for the fitting.

**LOWERB=***value-list*

specifies lower boundary constraints on the semivariance parameters. The *value-list* specification is a list of numbers or missing values (.) separated by commas. You must list the numbers in the order that PROC VARIOGRAM uses for the semivariance parameters, and each number corresponds to the lower boundary constraint. A missing value instructs PROC VARIOGRAM to use its default constraint.

If you do not specify lower bounds for all of the semivariance parameters, then PROC VARIOGRAM assumes that the remaining parameters are not bounded. If you specify more lower bounds in the *value-list* than the available parameters, then the numbers in excess are ignored. If you specify lower bounds for parameters with missing initial values, then the VARIOGRAM procedure enforces the specified bounds in the fitting process. By default, the lower bound for all parameters is zero.

When you specify the [HOLD=](#) option together with the LOWERB= option, the lower bounds in the LOWERB= option *value-list* that correspond to fixed parameters are ignored. When you specify the [NOBOUND](#) option together with the LOWERB= option, the LOWERB= option is ignored.

**MAXSCALE=***maxscale*

specifies a positive upper threshold for the fitted semivariogram sill. This option imposes a linear constraint on the optimization of the nonfixed semivariogram scale and nugget parameters so that the sum of all scale and nugget parameters does not exceed the specified MAXSCALE= value. The MAXSCALE= constraint is ignored if all the semivariogram scale and nugget parameters are fixed.

**NOBOUND**

requests the removal of boundary constraints on semivariance parameters. For example, semivariance parameters have a default zero lower boundary constraint since they have a physical meaning only for positive values. The NOBOUND option enables the fitting process to derive negative estimates; hence, you need to be cautious with the outcome when you specify this option.

The NOBOUND option has no effect on the power model exponent parameter. The exponent must range within  $[0,2)$  so that the model is a valid semivariance function. Also, the NOBOUND option has no effect on the Matérn smoothness parameter. The options LOWERB= and UPPERB= are ignored if either of them is specified together with the NOBOUND option in the PARMS statement.

**PARMSDATA=SAS-data-set**

**PDATA=SAS-data-set**

specifies that semivariance parameters values be read from a SAS data set. The data set should contain the values in the sequence required by the PARMS statement in either of the following two ways:

- Specify one single column under the variable Estimate (or Est) that contains all the parameter values.
- Use one column for each parameter, and place the  $n$  columns under the Parm1–Parm $n$  variables.

For example, the following two data sets are valid and equivalent ways to specify initial values for the nugget effect and the parameters of the Matérn and exponential structures that have been used in the previous examples in the PARMS statement section:

```
data parData1;
    input Estimate @@;
    datalines;
. 3 20 4.5 . 15
;

data parData2;
    input Parm1 Parm2 Parm3 Parm4 Parm5 Parm6;
    datalines;
. 3 20 4.5 . 15
;
```

If you have the parData1 data set, then you can import this information into the PARMS statement as follows:

```
proc variogram data=FirstData;
    < other VARIOGRAM statements >
    model form=(mat,exp);
    parms / pdata=parData1 hold=(2 to 4,6);
run;
```

You can specify more than one set of initial values in the PDATA= data set by following the preceding guidelines. PROC VARIOGRAM seeks among the specified sets for the one that gives the lowest objective function value. Then, the procedure uses the initial values in the selected set for the fitting optimization.

You can explicitly specify initial parameter values in the PARMS statement or use the PDATA= option, but you cannot use both at the same time.

**UPPERB=***value-list*

specifies upper boundary constraints on the semivariance parameters. The *value-list* specification is a list of numbers or missing values (.) separated by commas. You must list the numbers in the order that PROC VARIOGRAM uses for the semivariance parameters, and each number corresponds to the upper boundary constraint. A missing value instructs PROC VARIOGRAM to use its default constraint.

If you do not specify upper bounds for all of the semivariance parameters, then PROC VARIOGRAM assumes that the remaining parameters are not bounded. If you specify more upper bounds in the *value-list* than the available parameters, then the numbers in excess are ignored. If you specify upper bounds for parameters with missing initial values, then the VARIOGRAM procedure enforces the specified bounds in the fitting process. By default, the scale, range, nugget, and Matérn smoothness parameters have no upper bounds, whereas the power model exponent parameter is lower than two.

When you specify the **HOLD=** option together with the **UPPERB=** option, the upper bounds in the **UPPERB=** option *value-list* that correspond to fixed parameters are ignored. When you specify the **NOBOUND** option together with the **UPPERB=** option, the **UPPERB=** option is ignored.

---

## NLOPTIONS Statement

**NLOPTIONS** < *options* > ;

By default, PROC VARIOGRAM uses the technique **TECH=NRRIDG**, which corresponds to Newton-Raphson optimization with ridging. For more information about the **NLOPTIONS**, see the section “**NLOPTIONS Statement**” on page 482 in Chapter 19, “**Shared Concepts and Topics**.”

---

## STORE Statement

**STORE OUT=***store-name* < / *option* > ;

The **STORE** statement requests that the procedure save the context and results of the semivariogram model fitting analysis in an item store. An item store is a binary file defined by the SAS System. You cannot modify the contents of an item store. The contents of item stores produced by PROC VARIOGRAM can be processed only with the **KRIGE2D** or the **SIM2D** procedure. After you save results in an item store, you can use them at a later time without having to fit the model again.

The *store-name* is a usual one- or two-level SAS name, as for SAS data sets. If you specify a one-level name, then the item store resides in the Work library and is deleted at the end of the SAS session. Since item stores are often used for postprocessing tasks, typical usage specifies a two-level name of the form *libname.membername*. If an item store by the same name as specified in the **STORE** statement already exists, the existing store is replaced.

You can specify the following option in the STORE statement after a slash (/):

**LABEL=***store-label*

specifies a custom label for the item store that is produced by PROC VARIOGRAM. When another procedure processes an item store, the label appears in the procedure's output along with other identifying information.

---

## VAR Statement

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

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

---

## Details: VARIOGRAM Procedure

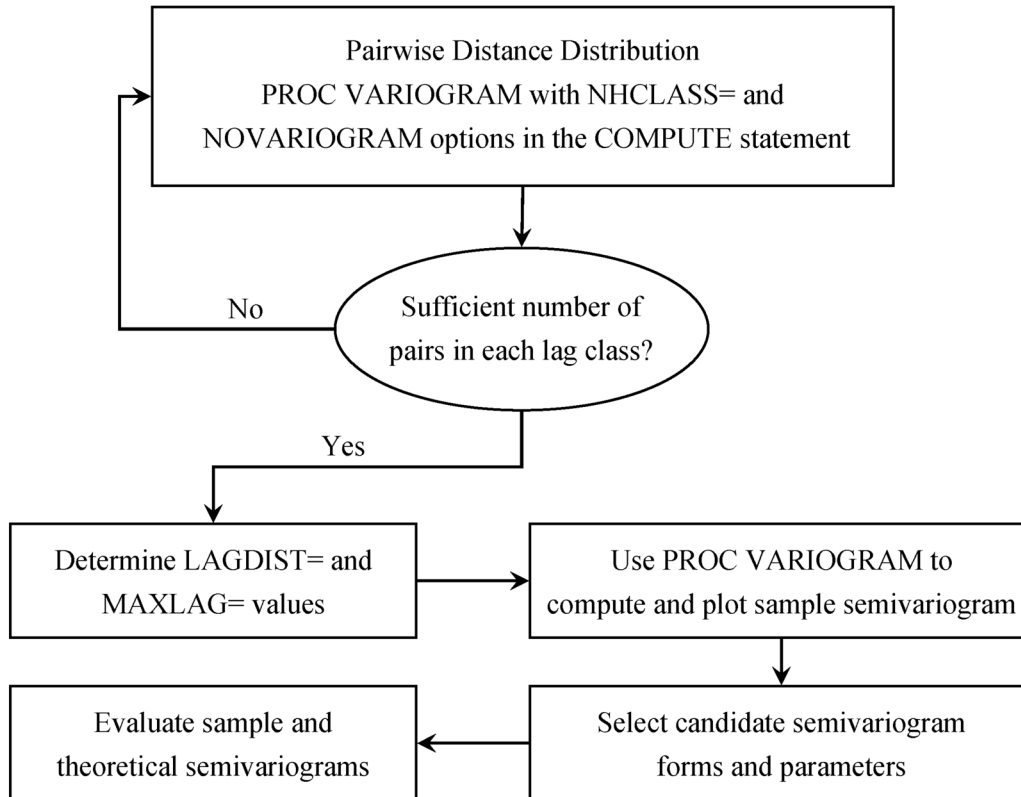
---

### 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 then characterizes the spatial structure of the process.

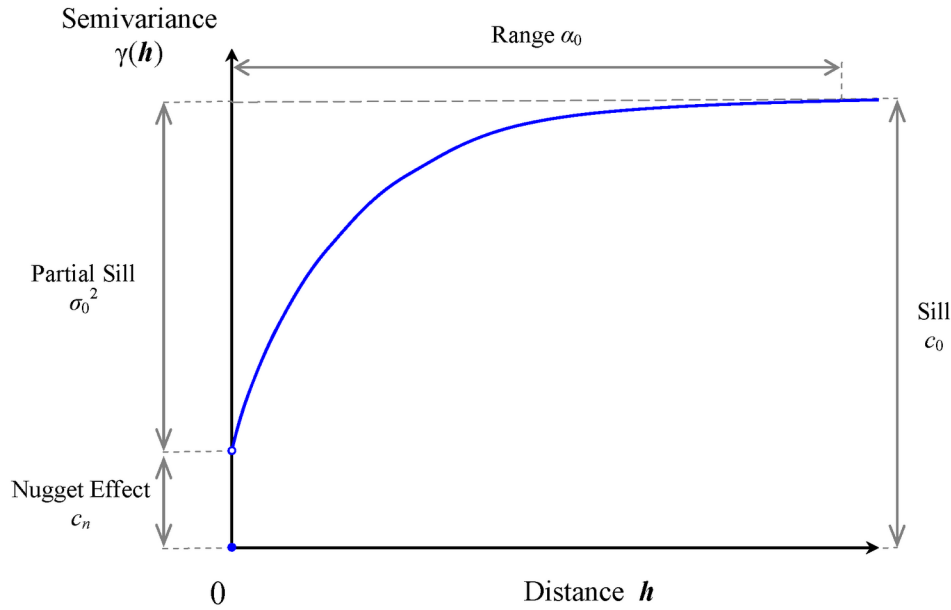
The VARIOGRAM procedure follows a general flow of investigation that leads you from a set of spatial observations to an expression of theoretical semivariance to characterize the SRF continuity. Specifically, the empirical semivariogram is computed after a suitable choice is made for the **LAGDISTANCE=** and **MAXLAGS=** options. For computations in more than one direction you can further use the **NDIRECTIONS=** option or the **DIRECTIONS** statement. Potential theoretical models (which can also incorporate nesting, anisotropy, and the nugget effect) can be fitted to the empirical semivariance by using the **MODEL** statement, and then plotted against the empirical semivariogram. The flow of this analytical process is illustrated in Figure 102.17. 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 3839 in the KRIGE2D procedure documentation.

**Figure 102.17** Flowchart for Semivariogram Selection

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 “[Getting Started: VARIOGRAM Procedure](#)” on page 8561. Equivalently, your data could be residuals of an initial sample that has had a surface trend removed, as portrayed in “[Example 102.2: An Anisotropic Case Study with Surface Trend in the Data](#)” on page 8658. For a closer look at stationarity, see the section “[Stationarity](#)” on page 8615. For details about different stationarity types and conditions see, for example, Chilès and Delfiner (1999, section 1.1.4).

### Characteristics of Semivariogram Models

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 102.18](#) displays a theoretical semivariogram of a spherical semivariance model and points out the semivariogram characteristics.

**Figure 102.18** 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  $r_\epsilon$  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  $h = 0$ ; hence, the nugget effect demonstrates itself as a jump in the semivariance as soon as  $h > 0$  (note in Figure 102.18 the discontinuity of the function at  $h = 0$  in the presence of a nugget effect).

The sill  $c_0$  consists of the nugget effect, if present, and the *partial sill*  $\sigma_0^2$ ; that is,  $c_0 = c_n + \sigma_0^2$ . If the SRF  $Z(s)$  is second-order stationary (see the section “Stationarity” on page 8615), the estimate of the sill is an estimate of the constant variance  $\text{Var}[Z(s)]$  of the field. Nonstationary processes have variances that depend on the location  $s$ . Their semivariance increases with distance; hence their semivariograms have no sill.

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

$$\sum_{i=1}^m \sum_{j=i}^m q_i q_j \gamma_z(s_i - s_j) \leq 0$$

for any number  $m$  of locations  $s_i, s_j$  in  $\mathcal{R}^2$  with  $h = s_i - s_j$ , and any real numbers  $q_i$  such that  $\sum_{i=1}^m q_i = 0$ . PROC VARIOGRAM can use a variety of permissible theoretical semivariogram models.

Specifically, Table 102.4 shows a list of such models that you can use for fitting in the **MODEL** statement of the VARIOGRAM procedure.

**Table 102.4** Permissible Theoretical Semivariogram Models ( $a_0 > 0$ , unless noted otherwise)

| Model Type       | Semivariance                                                                                                                                                                                                                                                                                                                                                                                                             |
|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Exponential      | $\gamma_z(\mathbf{h}) = \begin{cases} 0 & \text{if }  \mathbf{h}  = 0 \\ c_n + \sigma_0^2 \left[ 1 - \exp\left(-\frac{ \mathbf{h} }{a_0}\right) \right] & \text{if } 0 <  \mathbf{h}  \end{cases}$                                                                                                                                                                                                                       |
| Gaussian         | $\gamma_z(\mathbf{h}) = \begin{cases} 0 & \text{if }  \mathbf{h}  = 0 \\ c_n + \sigma_0^2 \left[ 1 - \exp\left(-\frac{ \mathbf{h} ^2}{a_0^2}\right) \right] & \text{if } 0 <  \mathbf{h}  \end{cases}$                                                                                                                                                                                                                   |
| Power            | $\gamma_z(\mathbf{h}) = \begin{cases} 0 & \text{if }  \mathbf{h}  = 0 \\ c_n + \sigma_0^2 \mathbf{h}^{a_0} & \text{if } 0 <  \mathbf{h} , 0 \leq a_0 < 2 \end{cases}$                                                                                                                                                                                                                                                    |
| Spherical        | $\gamma_z(\mathbf{h}) = \begin{cases} 0 & \text{if }  \mathbf{h}  = 0 \\ c_n + \sigma_0^2 \left[ \frac{3}{2} \frac{ \mathbf{h} }{a_0} - \frac{1}{2} \left( \frac{ \mathbf{h} }{a_0} \right)^3 \right] & \text{if } 0 <  \mathbf{h}  \leq a_0 \\ c_0 & \text{if } a_0 <  \mathbf{h}  \end{cases}$                                                                                                                         |
| Cubic            | $\gamma_z(\mathbf{h}) = \begin{cases} 0 & \text{if }  \mathbf{h}  = 0 \\ c_n + \sigma_0^2 \left[ 7 \left( \frac{ \mathbf{h} }{a_0} \right)^2 - \frac{35}{4} \left( \frac{ \mathbf{h} }{a_0} \right)^3 + \frac{7}{2} \left( \frac{ \mathbf{h} }{a_0} \right)^5 - \frac{3}{4} \left( \frac{ \mathbf{h} }{a_0} \right)^7 \right] & \text{if } 0 <  \mathbf{h}  \leq a_0 \\ c_0 & \text{if } a_0 <  \mathbf{h}  \end{cases}$ |
| Pentaspherical   | $\gamma_z(\mathbf{h}) = \begin{cases} 0 & \text{if }  \mathbf{h}  = 0 \\ c_n + \sigma_0^2 \left[ \frac{15}{8} \frac{ \mathbf{h} }{a_0} - \frac{5}{4} \left( \frac{ \mathbf{h} }{a_0} \right)^3 + \frac{3}{8} \left( \frac{ \mathbf{h} }{a_0} \right)^5 \right] & \text{if } 0 <  \mathbf{h}  \leq a_0 \\ c_0 & \text{if } a_0 <  \mathbf{h}  \end{cases}$                                                                |
| Sine hole effect | $\gamma_z(\mathbf{h}) = \begin{cases} 0 & \text{if }  \mathbf{h}  = 0 \\ c_n + \sigma_0^2 \left[ 1 - \frac{\sin(\pi \mathbf{h} /a_0)}{\pi \mathbf{h} /a_0} \right] & \text{if } 0 <  \mathbf{h}  \end{cases}$                                                                                                                                                                                                            |
| Matérn class     | $\gamma_z(\mathbf{h}) = \begin{cases} 0 & \text{if }  \mathbf{h}  = 0 \\ c_n + \sigma_0^2 \left[ 1 - \frac{2}{\Gamma(\nu)} \left( \frac{ \mathbf{h} \sqrt{\nu}}{a_0} \right)^\nu K_\nu \left( 2 \frac{ \mathbf{h} \sqrt{\nu}}{a_0} \right) \right] & \text{if } 0 <  \mathbf{h} , \nu > 0 \end{cases}$                                                                                                                   |

All of these models, except for the power model, are transitive. A transitive model characterizes a random process whose variation reaches the sill value  $c_0$  within a specific range from any location in the field.

The power model is nontransitive and applies to processes whose variance increases with distance. It has no scale and range; instead, it quantifies the process variation by using a positive slope parameter and a dimensionless power exponent  $\alpha$  that indicate how fast the variance increases. The expression for the power model is a valid semivariogram only when the exponent parameter ranges within  $0 \leq \alpha < 2$ . For convenience, PROC VARIOGRAM registers the power model slope parameter under the **SCALE=** option parameters in the **MODEL** statement. For the same reason, the scale and power slope parameters are represented with the common symbol  $\sigma_0^2$  in Table 102.4. Also for convenience, PROC VARIOGRAM registers the power model exponent parameter under the **RANGE=** option parameters. The range and the power exponent parameters are represented with the common symbol  $a_0$  in Table 102.4.



The power model is a generalized case of the linear model, which is not included explicitly in the model set of PROC VARIOGRAM. The linear model is derived from the power model when you specify the exponent  $\alpha = 1$ .

Among the models displayed in Table 102.4, the Matérn (or  $K$ -Bessel) class is a class of semivariance models that distinguish from each other by means of the positive smoothing parameter  $\nu$ . Different values of  $\nu$  correspond to different correlation models. Most notably, for  $\nu = 0.5$  the Matérn semivariance is equivalent to the exponential model, whereas  $\nu \rightarrow \infty$  gives the Gaussian model. Also, Table 102.4 shows that the Matérn semivariance computations use the gamma function  $\Gamma(\nu)$  and the second kind Bessel function  $K_\nu$ .

In PROC VARIOGRAM, you can input the model parameter values either explicitly as arguments of options, or as lists of values. In the latter case, you are expected to provide the values in the order the models are specified in the SAS statements, and furthermore in the sequential order of the scale, range, and smoothing parameter for each model as appropriate, and always starting with the nugget effect. If the parameter values are specified through an input file, then the total of  $n$  parameters should be provided either as one variable named Estimate or as many variables with the respective names Parm1–Parm $n$ .

You can review in further detail the models shown in Table 102.4 in the section “Theoretical Semivariogram Models” on page 3823 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 that extends 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; however, 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.

## Nested Models

When you try to represent an empirical semivariogram by fitting a theoretical model, you might find 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. The semivariance models that result as the sum of two or more semivariance structures are called *nested* models.

In general, a linear combination of permissible semivariance models produces a new permissible semivariance model. Nested models are based on this premise. You can include in a sum any combination of the models presented in Table 102.4. For example, a nested semivariance  $\gamma_z(\mathbf{h})$  that contains two structures, one exponential  $\gamma_{z,\text{EXP}}(\mathbf{h})$  and one spherical  $\gamma_{z,\text{SPH}}(\mathbf{h})$ , can be expressed as

$$\gamma_z(\mathbf{h}) = \gamma_{z,\text{EXP}}(\mathbf{h}) + \gamma_{z,\text{SPH}}(\mathbf{h})$$

If you have a nested model and a nugget effect, then the nugget effect  $c_n$  is a single parameter that is considered jointly for all the nested structures.

Nested models, anisotropic models, and the nugget effect increase the scope of theoretical models available. You can find additional discussion about these concepts in the section “Theoretical Semivariogram Models” on page 3823 in the KRIGE2D procedure documentation.

---

## Theoretical and Computational Details of the Semivariogram

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

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

or, equivalently,

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

The plot of semivariance as a function of  $\mathbf{h}$  is the *semivariogram*. You might also commonly see the term *semivariogram* used instead of the term *semivariance*.

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

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

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

$$\hat{\gamma}_z(\mathbf{h}) = \frac{1}{2 |N(\mathbf{h})|} \sum_{N(\mathbf{h})} [Z(s_i) - Z(s_j)]^2$$

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

$$N(\mathbf{h}) = \{i, j : s_i - s_j = \mathbf{h}\}$$

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

The expression for  $\hat{\gamma}_z(\mathbf{h})$  is called the *empirical semivariance* (Matheron 1963). This is the quantity that PROC VARIOGRAM computes, and its corresponding plot is the *empirical semivariogram*.

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

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

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

$$\Psi(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{P_i P_j \in N(\mathbf{h})} [Z(s_i) - Z(s_j)]^{\frac{1}{2}}$$

According to Cressie (1985), the estimate  $\hat{\gamma}_z(\mathbf{h})$  has approximate variance

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

This approximation is possible by assuming  $Z(s)$  to be a Gaussian SRF, and by further assuming the squared differences in empirical semivariances to be uncorrelated for different distances  $\mathbf{h}$ . Typically, semivariance estimates are correlated because of the underlying spatial correlation among the observations, and also because the same observation pairs might be used for the estimation of more than one semivariogram point, as described in the following subsections. Despite these restrictive assumptions, the approximate variance provides an idea about the semivariance estimate variance and enables fitting of a theoretical model to the empirical semivariance; see the section “[Theoretical Semivariogram Model Fitting](#)” on page 8627 for more details about the fitting process.

**NOTE:** If your data include a surface trend, then the empirical semivariance  $\hat{\gamma}_z(\mathbf{h})$  is not an estimate of the theoretical semivariance function  $\gamma_z(\mathbf{h})$ . 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  $Z(s)$  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 8626. 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 102.3: Analysis without Surface Trend Removal](#)” on page 8672.

## Stationarity

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

The expected value  $E[Z(s)]$  is the first statistical moment of the SRF  $Z(s)$ . The second statistical moment of the SRF  $Z(s)$  is the *covariance* function between two points  $s_i$  and  $s_j$  in  $Z(s)$ , and it is defined as

$$C_z(s_i, s_j) = E([Z(s_i) - E[Z(s_i)]] [Z(s_j) - E[Z(s_j)]])$$

When  $s_i = s_j = s$ , the covariance expression provides the variance at  $s$ .

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

$$C_z(s_i, s_j) = C_z(s_i - s_j) = C_z(\mathbf{h}) = E([Z(s) - m][Z(s + \mathbf{h}) - m])$$

When both  $E[Z(s)]$  and  $C(s_i, s_j)$  are invariant to spatial translations, the SRF  $Z(s)$  is characterized as *second-order stationary* (Cressie 1993, p. 53).

In a second-order stationary SRF the quantity  $C(\mathbf{h})$  is the same for any two points that are separated by distance  $\mathbf{h}$ . Based on the preceding formula, for  $\mathbf{h} = 0$  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

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

which relates the theoretical semivariance and covariance. Keep in mind 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  $E[Z(s)]$  and the assumption of intrinsic stationarity, *ergodicity* is a necessary third hypothesis to estimate the empirical semivariance. Assume that for the SRF  $Z(s)$  you have measurements  $z_i$  whose sample mean is estimated by  $\bar{Z}$ . The hypothesis of ergodicity dictates that  $\bar{Z} = E[Z(s)]$ .

In general, an SRF  $Z(s)$  is characterized as ergodic if the statistical moments of its realizations coincide with the corresponding ones of the SRF. In spatial analysis you 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  $V(s_i) = v_i = z_i - \bar{Z}$ ,  $i = 1, \dots, n$ , where it is assumed that the sample mean  $\bar{Z}$  is the constant expected value  $E[Z(s)]$  of  $Z(s)$ . This is equivalent to using the original values, since  $V(s_i) - V(s_j) = Z(s_i) - Z(s_j)$ , which shows the property of the semivariance to filter out the mean. See the section “[Semivariance Computation](#)” on page 8625 for the exact expressions PROC VARIOGRAM uses to compute the empirical classical  $\hat{\gamma}_z(\mathbf{h})$  and robust  $\bar{\gamma}_z(\mathbf{h})$  semivariances.

## Anisotropy

Semivariance is defined on the basis of the spatial increment vector  $\mathbf{h}$ . If the variance characteristics of  $Z(\mathbf{s})$  are independent of the spatial direction, then  $Z(\mathbf{s})$  is called *isotropic*; if not, then  $Z(\mathbf{s})$  is called *anisotropic*. In the case of isotropy, the semivariogram depends only on the length  $h$  of  $\mathbf{h}$  and  $\gamma_z(\mathbf{h}) = \gamma_z(h)$ . 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 remain unchanged 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. Typically, these two directions are perpendicular, although they might be at other than right angles when zonal anisotropy is present.

If you can distinguish a maximum and a minimum sill in different directions, then you have a case of zonal anisotropy. The SRF exhibits strongest continuity in the direction of the lowest sill, which is the direction of the major anisotropy axis. If the sill does not change across directions, then the major axis direction of strongest continuity is the one in which the semivariogram has maximum range. See “[Example 102.2: An Anisotropic Case Study with Surface Trend in the Data](#)” on page 8658 for a detailed demonstration of a case with anisotropy when you use PROC VARIOGRAM.

You can find additional information about anisotropy analysis in the section “[Anisotropic Models](#)” on page 3832 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 102.19](#) shows the spatial domain  $D$  and the set of  $n$  measurements  $z_i, i = 1, \dots, n$ , that have been sampled at the indicated locations in  $D$ . Two data points  $P_1$  and  $P_2$ , with coordinates  $\mathbf{s}_1 = (x_1, y_1)$  and  $\mathbf{s}_2 = (x_2, y_2)$ , respectively, are selected for illustration.

A vector, or directed line segment, is drawn between these points. If the length

$$|P_i P_j| = |\mathbf{s}_2 - \mathbf{s}_1| = (x_2 - x_1)^2 + (y_2 - y_1)^2$$

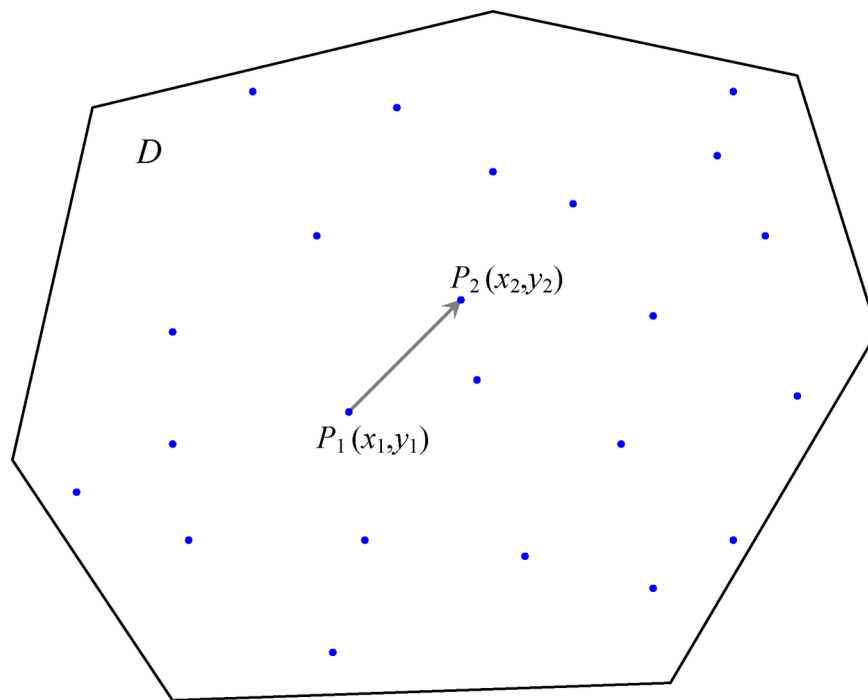
of this vector is smaller than the specified `DEPSILON=` value, then the pair is excluded from the continuity measure calculations because the two points  $P_1$  and  $P_2$  are considered to be at zero distance apart (or *collocated*). Spatial collocation might appear due to different scales in sampling, 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  $\mathbf{s}_2 - \mathbf{s}_1$ , and then by its length  $|P_i P_j|$ . For example, it is unlikely for actual data that the

distance  $|P_i P_j|$  between any pair of data points  $P_i$  and  $P_j$  located at  $s_i$  and  $s_j$ , respectively, would exactly satisfy  $|P_i P_j| = |h| = h$  in the preceding computation of  $\hat{\gamma}_z(h)$ . A similar argument can be made for the orientation of the segment  $s_2 - s_1$ . 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 can also find extensive discussions about the size and the number of classes to consider for the computation of the empirical semivariogram.

**Figure 102.19** Selection of Points  $P_1$  and  $P_2$  in Spatial Domain  $D$



### Angle Classification

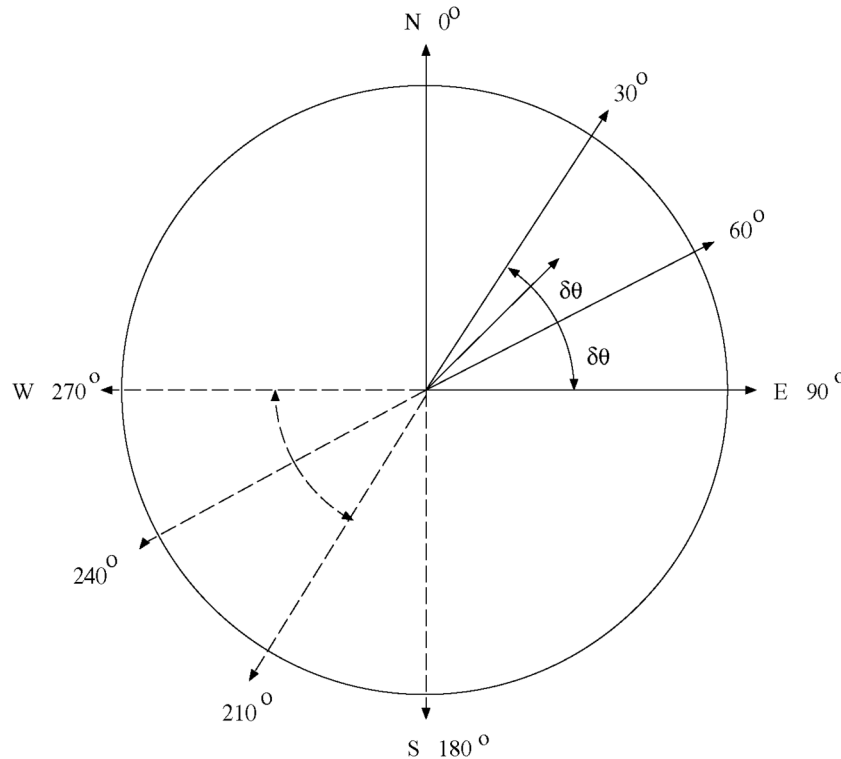
Suppose you specify `NDIRECTIONS=3` in the `COMPUTE` statement in `PROC VARIOGRAM`. This results in three angle classes defined by midpoint angles between  $0^\circ$  and  $180^\circ$ :  $0^\circ \pm \delta\theta$ ,  $60^\circ \pm \delta\theta$ , and  $120^\circ \pm \delta\theta$ , where  $\delta\theta$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

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

For example, if `NDIRECTIONS=3`, the default angle tolerance is  $\delta\theta = 30^\circ$ . When the directed line segment  $P_1 P_2$  in Figure 102.19 is superimposed on the coordinate system that shows the angle classes, its angle is approximately  $45^\circ$ , measured clockwise from north. In particular, it falls within  $[60^\circ - \delta\theta, 60^\circ + \delta\theta) = [30^\circ, 90^\circ)$ , the second angle class (Figure 102.20).

**NOTE:** 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  $225^\circ$  instead of approximately  $45^\circ$ . This does not affect angle class selection; the angle classes  $[60^\circ - \delta\theta, 60^\circ + \delta\theta)$  and  $[240^\circ - \delta\theta, 240^\circ + \delta\theta)$  are the same.

**Figure 102.20** Selected Pair  $P_1 P_2$  Falls within the Second Angle Class



If you specify an angle tolerance less than the default, such as  $ATOL=15^\circ$ , some point pairs might be excluded. For example, the selected point pair  $P_1 P_2$  in Figure 102.20, while closest to the  $60^\circ$  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 8623).

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 only a small amount of data is present 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 **BANDWIDTH=** option have a similar effect. See the section “Bandwidth Restriction” on page 8620 for an explanation of how **BANDWIDTH=** 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, consider specifying the angle tolerance too. The default value of the angle tolerance is  $45^\circ$  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 8590 for more details.

### Distance Classification

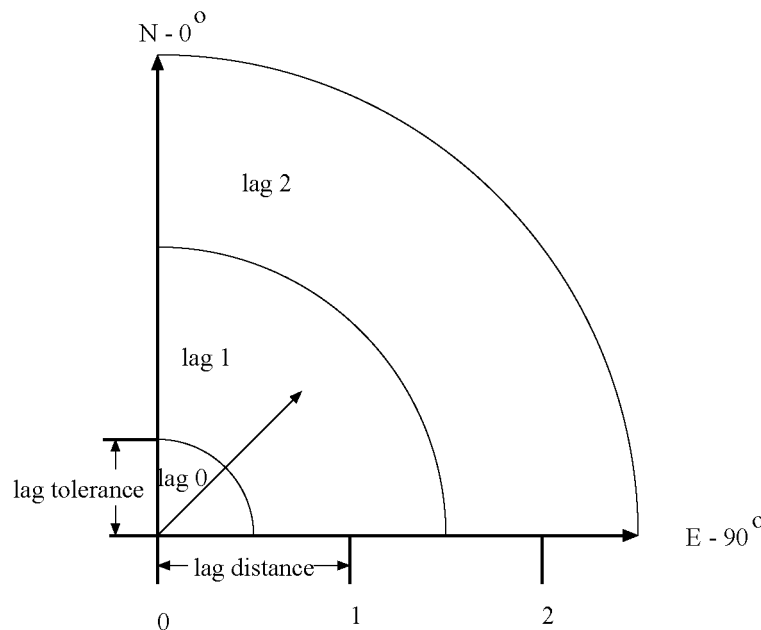
The distance class for a point pair  $P_1P_2$  is determined as follows. The directed line segment  $P_1P_2$  is superimposed on the coordinate system that shows the distance or lag classes. These classes are determined by the **LAGDISTANCE=** option in the **COMPUTE** statement. Denoting the length of the line segment by  $|P_1P_2|$  and the **LAGDISTANCE=** value by  $\Delta$ , the lag class  $L$  is determined by

$$L(P_1P_2) = \left\lfloor \frac{|P_1P_2|}{\Delta} + 0.5 \right\rfloor$$

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

When the directed line segment  $P_1P_2$  is superimposed on the coordinate system that shows the distance classes, it is seen to fall in the first lag class; see **Figure 102.21** for an illustration for  $\Delta = 1$ .

**Figure 102.21** Selected Pair  $P_1P_2$  Falls within the First Lag Class



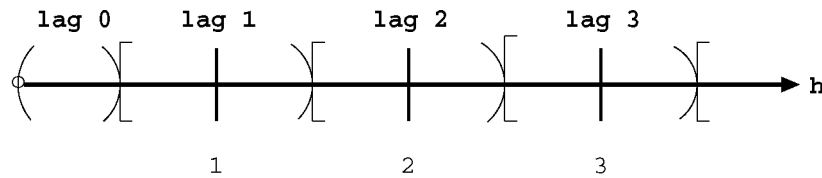
Pairwise distances are positive. Therefore, the line segment  $|P_1P_2|$  might belong to one of the **MAXLAG** lag classes or it could be shorter than half the length of the **LAGDISTANCE=** value. In the last case the segment is said to belong to the lag class zero. Hence, lag class zero is smaller than lag classes  $1, \dots, \text{MAXLAGS}$ . The definition of lag classes in this manner means that when you specify the **MAXLAGS=** parameter, **PROC VARIOGRAM** produces a semivariogram with a total of **MAXLAGS+1** lag classes including the zero lag class. For example, if you specify **LAGDISTANCE=1** and **MAXLAGS=10** and you do not specify a **LAGTOLERANCE=** value in the **COMPUTE** statement in **PROC VARIOGRAM**, the 11 lag classes generated by the preceding equation are

$$[0, 0.5), [0.5, 1.5), [1.5, 2.5), \dots, [9.5, 10.5)$$



The preceding lag classes description is correct under the assumption of the default lag tolerance, which is half the **LAGDISTANCE=** value. Using the default lag tolerance results in no gaps between the distance class intervals, as shown in Figure 102.22.

**Figure 102.22** Lag Distance Axis Showing Lag Classes



On the other hand, if you do specify a distance tolerance with the **LAGTOLERANCE=** option in the **COMPUTE** statement, a further check is performed to see whether the point pair falls within this tolerance of the nearest lag. In the preceding example, if you specify **LAGDISTANCE=1** and **MAXLAGS=10** (as before) and also specify **LAGTOLERANCE=0.25**, the intervals become

$$[0, 0.25), [0.75, 1.25), [1.75, 2.25), \dots, [9.75, 10.25)$$

You might want to avoid this specification because it results in gaps in the lag classes. For example, if a point pair  $P_1 P_2$  falls in an interval such as

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

then it is excluded from the semivariance calculation. The maximum **LAGTOLERANCE=** value allowed is half the **LAGDISTANCE=** value; no overlap of the distance classes is allowed.

See the section “Computation of the Distribution Distance Classes” on page 8621 for a more extensive discussion of practical aspects in the specification of the **LAGDISTANCE=** and **MAXLAGS=** options.

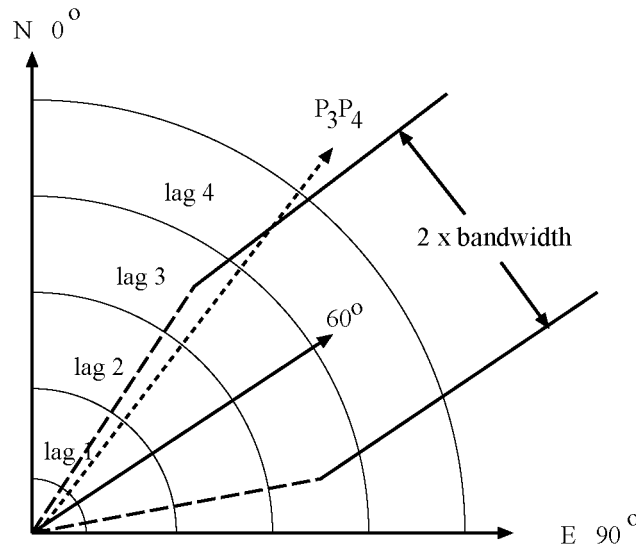
## Bandwidth Restriction

Because the areal segments that are 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 **BANDWIDTH=** option in the **COMPUTE** statement, the lateral, or perpendicular, distance from the axis that defines the angle classes is fixed.

For example, suppose two points  $P_3$ ,  $P_4$  are picked from the domain in Figure 102.19 and are superimposed on the grid that defines distance and angle classes, as shown in Figure 102.23.

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



**Figure 102.23** Selected Pair  $P_3P_4$  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 **BANDWIDTH=** 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 **BANDWIDTH=** has been specified. See also the section “[Angle Classification](#)” on page 8617 for more details about the effects these rules can have, since **BANDWIDTH=** operates in a manner similar to the **ANGLETOLERANCE=** 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.

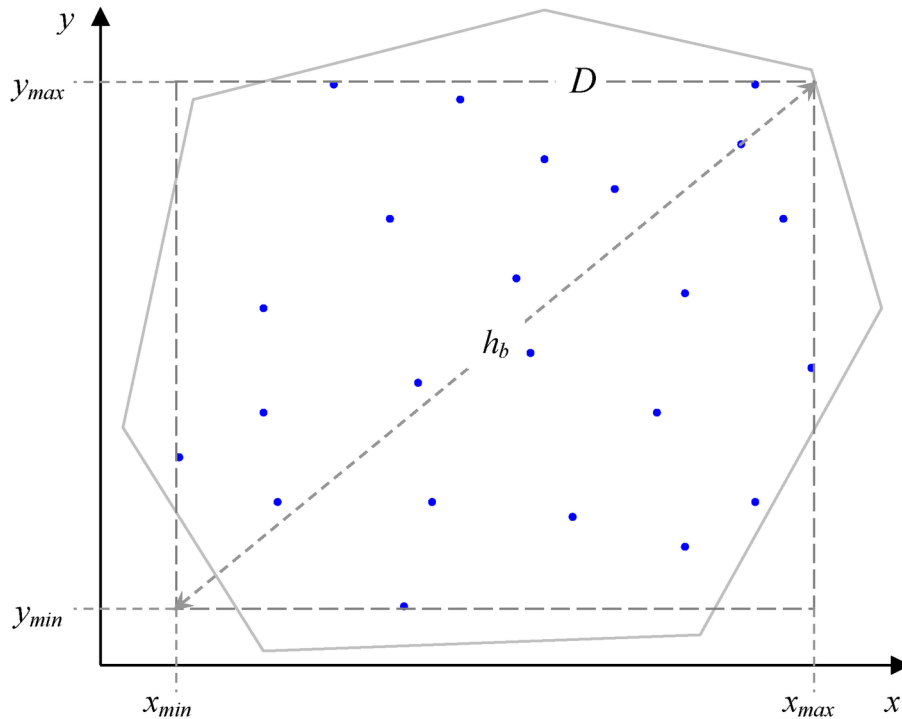
The simplest way of determining the distribution of pairwise distances is to determine the maximum distance  $h_{max}$  between any pair of points in your data, and then to divide this distance by some number  $N$  of intervals to produce distance classes of length  $\delta = h_{max}/N$ . The distance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

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 102.24 for an illustration of the bounding rectangle applied to the data of the domain  $D$  in Figure 102.19. PROC VARIOGRAM provides you with the sizes of  $x_{max} - x_{min}$ ,  $y_{max} - y_{min}$ , and  $h_b$ . For example, in Figure 102.4 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 102.24** 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  $h_0, 2h_0, \dots, Nh_0$ , with a distance tolerance of  $\pm \frac{h_0}{2}$ . The extra subinterval is  $(0, h_0/2)$  and corresponds to lag class zero. It is half

the length of the remaining subintervals, and it often contains the smallest number of pairs. Figure 102.22 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 8642, 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, estimate how many classes of data pairs you might 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, 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 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 8562, 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 102.2: An Anisotropic Case Study with Surface Trend in the Data” on page 8658.

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 parameter affects the data pairs distribution in each distance class. The pairwise distance intervals table (for example, Figure 102.3) 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 102.5) or the pairwise distances histogram (see Figure 102.4), 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  $h = 0$  and the most remote ones at large  $h$ . The classes near the origin correspond to lags 0 and 1. These lags are crucial because the empirical semivariogram in small distances  $h$  characterizes the process smoothness and can help you detect the presence of a nugget effect. However, as discussed in the section “Distance Classification” on page 8619, lag zero 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 `PLOTS=PAIRS` option.

Then, the procedure produces in the pairwise distances histogram a reference line at the specified `THRESHOLD=` value, which leaves below the line all lags whose pairs count is lower than the threshold value; see, for example, Figure 102.4. The last lag class whose pair population is above the `THRESHOLD=` value is reported in the pairs information table as “Highest Lag With Pairs > Threshold.” 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 102.3 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 have no 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. If you include the largest distances in your empirical semivariogram plot, then these pairs only contribute added noise. 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 represents then 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*)  $\epsilon$  and the *correlation radius*  $h_c$ . Both are defined in a similar manner. The correlation range  $\epsilon$  is the distance at which the covariance is 5% of its value at  $\mathbf{h} = 0$ , and shows that beyond  $\epsilon$  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  $\mathbf{h} = 0$ , 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  $r_\epsilon$  used in asymptotically increasing semivariance models has essentially the same definition as the correlation range  $\epsilon$  (see the section “[Theoretical Semivariogram Models](#)” on page 8608).

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 you have 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  $h_b$ , in addition to 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 102.5](#)), and in the pairwise distances histogram as “Max Data Distance” (see [Figure 102.4](#)).

Overall, avoid significant deviations 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 that  $P_i P_j$  belong 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 102.20](#) and [Figure 102.21](#),  $P_1 P_2$  belongs to  $N(60^\circ, 1)$ .

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

$$\hat{\gamma}(h_k) = \frac{1}{2 |N(\theta_k, L)|} \sum_{P_i P_j \in N(\theta_k, L)} [V(s_i) - V(s_j)]^2$$

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

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

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

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

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 8613 that if your data are not free of nonrandom surface trends, then the empirical semivariance  $\hat{\gamma}_z(\mathbf{h})$  you obtain from PROC VARIOGRAM represents a pseudo-semivariance rather than an estimate of the theoretical semivariance  $\gamma_z(\mathbf{h})$ .

In practice, two major difficulties appear. First, you might have no knowledge of underlying surface trends in your SRF  $Z(s)$ . 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, because trend removals with different models are essentially different operations acting upon the values of your original observations. Following the comment by Chilès and Delfiner (1999, section 2.7.3), 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  $\mathbf{h} = 0$ , and it 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 8662 demonstrates how to use PROC GLM (see Chapter 42, “[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 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 8562.

If you observe similar values locally across a particular direction, this is 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  $Z(s)$ . If you suspect that at least one such direction exists for your data, then run PROC VARIOGRAM for a series of directions in the angular vicinity. The trend-free direction, if it exists, coincides with the one whose pseudo-semivariogram exhibits minimal increase with distance; see “[Example 102.3: Analysis without Surface Trend Removal](#)” on page 8672 for a demonstration of this approach. However, you cannot test  $Z(s)$  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 metastatistical decisions and 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 8613, your semivariogram analysis relies fundamentally on the use of trend-free data.

---

## Theoretical Semivariogram Model Fitting

You can choose between two approaches to select a theoretical semivariogram model and fit the empirical semivariance. The first one is manual fitting, in which a theoretical semivariogram model is selected based on visual inspection of the empirical semivariogram. For example, see Hohn (1988, p. 25) and comments from defendants of this approach in Olea (1999, p. 82). The second approach is to perform model fitting 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).

The VARIOGRAM procedure features automated semivariogram model fitting that uses the weighted least squares (WLS) or the ordinary least squares (OLS) method. Use the **MODEL** statement to request that specific model forms or an array of candidate models be tested for optimal fitting to the empirical semivariance.

Assume that you compute first the empirical semivariance  $\gamma_z^*(\mathbf{h})$  at **MAXLAGS**= $k$  distance classes, where  $\gamma_z^*(\mathbf{h})$  can be either the classical estimate  $\hat{\gamma}_z(\mathbf{h})$  or the robust estimate  $\tilde{\gamma}_z(\mathbf{h})$ , as shown in the section “[Theoretical and Computational Details of the Semivariogram](#)” on page 8613. In fitting based on least

squares, you want to estimate the parameters vector  $\theta$  of the theoretical semivariance  $\gamma_z(\mathbf{h})$  that minimizes the sum of square differences  $R(\theta)$  given by the expression

$$R(\theta) = \sum_{i=1}^k w_i^2 [\gamma_z^*(\mathbf{h}_i) - \gamma_z(\mathbf{h}_i; \theta)]^2$$

For  $i = 1, \dots, k$ , the weights are  $w_i^2 = 1/\text{Var}[\gamma_z^*(\mathbf{h}_i)]$  in the case of WLS and  $w_i^2 = 1$  in the case of OLS. Therefore, the parameters  $\theta$  are estimated in OLS by minimizing

$$R(\theta)_{OLS} = \sum_{i=1}^k [\gamma_z^*(\mathbf{h}_i) - \gamma_z(\mathbf{h}_i; \theta)]^2$$

For WLS, Cressie (1985) investigated approximations for the variance of both the classical and robust empirical semivariances. Then, under the assumptions of normally distributed observations and uncorrelated squared differences in the empirical semivariance, the approximate weighted least squares estimate of the parameters  $\theta$  can be obtained by minimizing

$$R(\theta)_{WLS} = \frac{1}{2} \sum_{i=1}^k N(\mathbf{h}_i) \left[ \frac{\gamma_z^*(\mathbf{h}_i)}{\gamma_z(\mathbf{h}_i; \theta)} - 1 \right]^2$$

where  $N(\mathbf{h}_i)$  is the number of pairs of points in the  $i$ th distance lag.

PROC VARIOGRAM relies on nonlinear optimization to minimize the least squares objective function  $R(\theta)$ . The outcome is the model that best fits the empirical semivariogram according to your criteria. The fitting process flow is displayed in [Figure 102.25](#). 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. At the same time, it is important to describe the correlation behavior accurately near the semivariogram origin. As pointed out by Chilès and Delfiner (1999, pp. 104–105), a poor description of spatial continuity at small lags can lead to loss of optimality in kriging predictions and erroneous reproduction of the variability in conditional simulations.

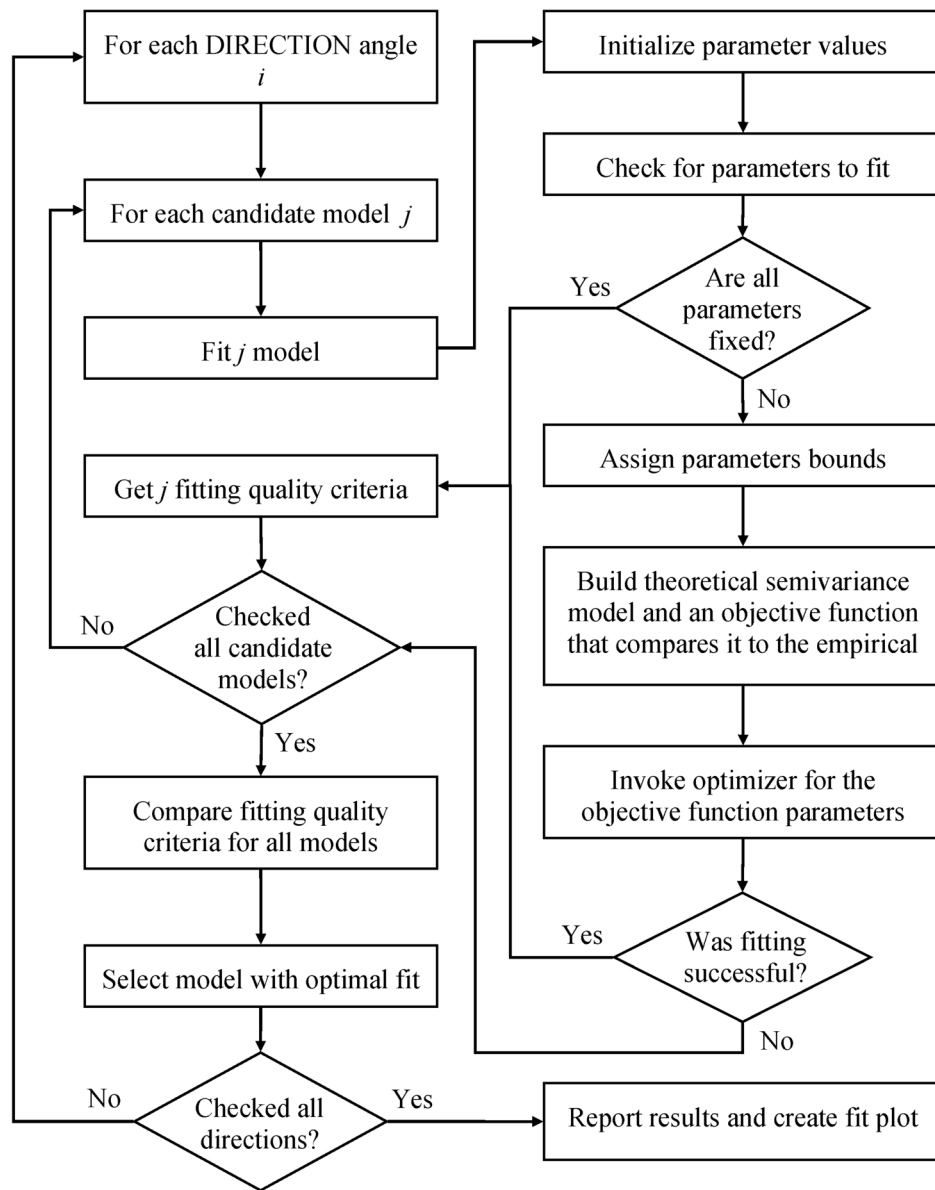
The significance of achieving better accuracy near the semivariogram origin is an advantage of the WLS method compared to OLS. In particular, the semivariance variance decreases when you get closer the origin  $\mathbf{h} = 0$ , as suggested in the section “[Theoretical and Computational Details of the Semivariogram](#)” on page 8613. The WLS weights are expressed as the inverse of this variance; as a result, WLS fitting is more accurate for distances  $\mathbf{h}$  near the semivariogram origin. In contrast, the OLS approach performs a least squares overall best fit because it assumes constant variance at all distances  $\mathbf{h}$ . 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.

Depending on your application, you can use WLS or OLS with PROC VARIOGRAM to fit classical semivariance. Other fitting methods include maximum likelihood approaches that rely crucially on the normality assumption for the data distribution, and the generalized least squares method that offers better accuracy but is computationally more demanding. You can find extensive discussions about these issues in Cressie (1993, section 2.3), Jian, Olea, and Yu (1996), Stein (1988), and Schabenberger and Gotway (2005).



The sections “[Parameter Initialization](#)” on page 8630 and “[Quality of Fit](#)” on page 8632 provide details and insight about semivariogram fitting, in addition to ways to cope with poor fits or no fit at all. These strategies can help you reach a meaningful description of spatial correlation in your problem.

**Figure 102.25** Semivariogram Fitting Process Flowchart



## Parameter Initialization

An important stage when you prepare for the model fitting process is initialization of the model parameters. As stated earlier, nonlinear optimization techniques are used in the fitting process. These techniques assist in the estimation of the model parameters, and being nonlinear means they can be very sensitive to selection of the initial values.

You can specify initial values close to the expected estimates when you have a relatively simple problem, such as in the example of the section “[Getting Started: VARIOGRAM Procedure](#)” on page 8561. In the case of nested models the selection of initial values can be more challenging because you have to assess the level of contribution for each one of the nested components.

The VARIOGRAM procedure features automatic selection of initial values based on the recommendations in Jian, Olea, and Yu (1996). Specifically, if you compute the estimated empirical semivariogram  $\hat{\gamma}_z(\mathbf{h})$  at  $k$  lags, then:

- The default initial nugget effect  $c_{n,0}$  is

$$c_{n,0} = \text{Max} \left[ 0, \hat{\gamma}_z(\mathbf{h}_1) - \frac{\mathbf{h}_1}{\mathbf{h}_2 - \mathbf{h}_1} [\hat{\gamma}_z(\mathbf{h}_2) - \hat{\gamma}_z(\mathbf{h}_1)] \right]$$

- The default initial slope  $\sigma_{0,0}^2$  and initial exponent  $a_{0,0}$  for the power model are

$$\sigma_{0,0}^2 = \frac{[\hat{\gamma}_z(\mathbf{h}_{k-2}) + \hat{\gamma}_z(\mathbf{h}_{k-1}) + \hat{\gamma}_z(\mathbf{h}_k)] / 3 - c_{n,0}}{\mathbf{h}_k - \mathbf{h}_1}$$

$$a_{0,0} = 1$$

- The default initial scale  $\sigma_{0,0}^2$  and initial range  $a_{0,0}$  for all other models are

$$\sigma_{0,0}^2 = \frac{[\hat{\gamma}_z(\mathbf{h}_{k-2}) + \hat{\gamma}_z(\mathbf{h}_{k-1}) + \hat{\gamma}_z(\mathbf{h}_k)]}{3} - c_{n,0}$$

$$a_{0,0} = 0.5\mathbf{h}_k$$

When you use the Matérn form, PROC VARIOGRAM sets the default initial value for the Matérn smoothness to  $\nu_0 = 1$ .

These rules are observed in the case of single, non-nested model fitting, and they are slightly modified to apply for nested model fitting as follows: Assume that you want to fit a nested model composed of  $m$  structures. As stated in the section “[Nested Models](#)” on page 8613, the nugget effect is a single parameter and is independent of the number of nested structures in a model. Also, the sum of the nested structure scales and the nugget effect, if any, must be equal to the total variance. For this reason, PROC VARIOGRAM simply divides the initial scale value it would assign to a non-nested model into  $m$  components  $\sigma_{0,0,1}^2, \dots, \sigma_{0,0,m}^2$ . For the range parameter, the VARIOGRAM procedure sets the initial range  $a_{0,0,1}$  of the first nested structure equal to the value it would assign to a non-nested model initial range. Then, the initial range  $a_{0,0,m}$  of the  $m$ -component is set recursively to half the value of the initial range  $a_{0,0,m-1}$  of the  $(m-1)$ -component.

Your empirical semivariogram must have nonmissing estimates at least at three lags so that you can use the automated fitting feature in PROC VARIOGRAM. Overall, if you specify a model form with  $q$  parameters to fit to an empirical semivariogram with nonmissing estimates at  $k$  lags, then the fitting problem is well-defined only when the degrees of freedom are  $DF = k - q \geq 0$ .

A potential numerical issue is that fitting could momentarily lead the fitting parameters to near-zero semivariance values at lags away from zero distance. The theoretical semivariance is always positive for any distance larger than zero, and this is also a requirement for the numerical computation of  $R(\theta)_{WLS}$  in weighted least squares fitting. Such numerical issues are unlikely but possible, depending on the data set you use and the parameter initial values. If an event of nonpositive semivariance at a given lag occurs during an iteration, then PROC VARIOGRAM transparently adds a minimal amount of variance at that lag for the specific iteration. You can control this amount of variance with the `NEPSILON=` option of the `MODEL` statement. It is recommended that you leave this parameter at its default value.

The section concludes with a reminder of the fitting process sensitivity to the initial parameter values selection. The VARIOGRAM procedure facilitates this selection for you by using the simple rules shown earlier. However, the suggested initial values might not always be the best choice. In simple cases, such as the introductory example in the section “Getting Started: VARIOGRAM Procedure” on page 8561, this approach is very convenient and effective.

In principle, you are strongly encouraged to experiment with initial values. You want to make sure that the fitting process leads the model parameters to converge to estimates that make sense for your problem. When a parameter estimate seems unreasonable on the basis of your problem specification (for example, a model scale might be estimated to be 10 times the size of your sample variance, or the estimate of a range might be zero), PROC VARIOGRAM produces a note to let you know about a potentially ambiguous fit. These issues are examined in more detail in the section “Quality of Fit” on page 8632.

## Parameter Estimates

When the fit process is complete, the VARIOGRAM procedure produces the “Parameter Estimates” table with information about the fitted model parameters. The table includes estimates of the parameters, their approximate standard error, the statistical degrees of freedom DF, the corresponding  $t$  statistic, and its approximate  $p$ -value. For a model with  $q$  parameters that fits an empirical semivariogram of  $k$  nonmissing lags,  $DF = k - q$ .

**NOTE:** Parameter estimates might have nonzero standard errors even in the rather extreme case where  $DF = 0$ . This can typically occur when there are active optimization constraints in the fitting process.

You can request the confidence intervals for the parameter estimates of a fitted model by specifying the `CL` option of the `MODEL` statement. These confidence intervals are computed using the Wald-based formula

$$\hat{\beta}_i \pm \text{stderr}_i \times t(k - q, 1 - \alpha/2)$$

where  $\hat{\beta}_i$  is the  $i$ th parameter estimate,  $\text{stderr}_i$  is its estimated approximate standard error,  $t(k - q, 1 - \alpha/2)$  is the  $t$  statistic with  $DF = k - q$  degrees of freedom. The confidence intervals are only asymptotically valid. The significance level  $\alpha$  used in the construction of these confidence limits can be set with the `ALPHA=` option of the `MODEL` statement; the default value is  $\alpha = 0.05$ .

Specify the `COVB` and the `CORRB` options in the `MODEL` statement to request the approximate covariance and approximate correlation matrices of the fitted parameters, respectively. These matrices are based on the optimization process results. In agreement with reporting similar optimization output in SAS/STAT software, parameters with active restraints have zeros in the corresponding rows and columns in the covariance and correlation matrices, and display 1 in the correlation matrix diagonal.

## Quality of Fit

The VARIOGRAM procedure produces a fit summary table to report about the goodness of fit. When you specify multiple models to fit with the **FORM=**[AUTO](#) option in the **MODEL** statement, the VARIOGRAM procedure uses two processes to rank the fitted models: The first one depends on your choice among available fitting criteria. The second one is based on an operational classification of equivalent models in classes. The two processes are described in more detail in the following subsections.

Overall, no absolutely correct way exists to rank and classify multiple models. Your choice of ranking criteria could depend on your study specifications, physical considerations, or even your personal assessment of fitting performance. The VARIOGRAM procedure provides you with fitting and comparison features to facilitate and help you better understand the fitting process.

### Fitting Criteria

The fit summary table ranks multiple models on the basis of one or more fitting criteria that you can specify with the **CHOOSE=** option of the **MODEL** statement, as explained in the section “[Syntax: VARIOGRAM Procedure](#)” on page 8576. Currently, the VARIOGRAM procedure offers two numerical criteria (for which a smaller value indicates a better fit) and a qualitative criterion:

- The residual sum of squares error (SSE) is based on the objective function of the fitting process. When the specified method is weighted least squares, the sum of squares of the weighted differences (WSSE) is computed according to the expression

$$\text{WSSE} = \sum_{i=1}^k w_i^2 [\gamma_z^*(\mathbf{h}_i) - \gamma_z(\mathbf{h}_i; \boldsymbol{\theta})]^2$$

where  $\gamma_z^*(\mathbf{h}_i)$  can be either the classical or robust semivariance estimate of the theoretical semivariance  $\gamma_z(\mathbf{h}_i; \boldsymbol{\theta})$  at the  $i$ th lag and the weights  $w_i^2$  are taken at lags  $i = 1, \dots, k$ . When you specify the **METHOD=**[OLS](#) option in the **MODEL**, the weights  $w_i^2 = 1$  for  $i = 1, \dots, k$ , and the SSE is expressed as

$$\text{SSE} = \sum_{i=1}^k [\gamma_z^*(\mathbf{h}_i) - \gamma_z(\mathbf{h}_i; \boldsymbol{\theta})]^2$$

- Akaike’s information criterion (AIC) is included in the fit summary table when there is at least one nonfixed parameter. In its strict definition, AIC assumes that the model errors are normally and independently distributed. This assumption is not correct in the semivariance fitting analysis. However, the AIC can be also defined in an operational manner on the basis of the weighted squared error sum WSSE as

$$\text{AIC} = k \ln \left( \frac{\text{WSSE}}{k} \right) + 2q$$

for  $k$  lags and  $q$  model parameters; see, for example, Olea (1999, p. 84). The operational definition of the AIC is provided as an additional criterion for the comparison of fitted models in PROC VARIOGRAM.

The AIC expression suggests that when you specify multiple models with the **FORM=**[AUTO](#) option in the **MODEL** statement, all models with the same number of parameters are ranked in the same way by the AIC and the WSSE criteria. Among models with the same WSSE value, AIC ranks higher the ones with fewer parameters.

- The third qualitative criterion enables you to classify multiple models based on their convergence status. A model is sent to the bottom of the ranking table if the parameter estimation optimization fails to converge or fitting is unsuccessful due to any other issue. These two cases are distinguished by the different notes they produce in the fit summary ODS table. If you specify the [STORE](#) statement to save the fitting output in an item store, then models that have failed to fit are not passed to the item store.

With respect to convergence status, PROC VARIOGRAM ranks higher those models that have successfully completed the fitting process. It might occur that the selection of parameter initial values, physical considerations about the forms that are used for the fit, or numerical aspects of the nonlinear optimization could result in ambiguous fits. For example, you might see that model parameters converge at or near their boundary values, or that parameters have unreasonably high estimates when compared to the empirical semivariogram characteristics. Then, the fit summary table designates such fits as questionable.

You might not need to take any action if you are satisfied with the fitting results and the selected model. You can investigate questionable fits in one or more of the following ways:

- If a form in a nested model makes no contribution to the model due to a parameter at or near its boundary value, then you could have a case of a degenerate fit. When you fit multiple models, a model with degenerate fit can collapse to the more simple model that does not include the noncontributing form. The VARIOGRAM procedure includes in its fit summary all models that are successfully fit. In such cases you can ignore degenerate fits. You can also try subsequent fits of individual models and exclude noncontributing forms or use different initial values.
- Unreasonably low or high parameter estimates might be an indication that the current initial values are not a good guess for the nonlinear optimizer. In most cases, fitting an empirical semivariogram gives you the advantage of a fair understanding about the value range of your parameters. Then, you can use the [PARMS](#) statement to specify a different set of initial values and try the fit again.
- Try replacing the problematic form with another one. A clear example is that you can expect a very poor fit if you specify an exponential model to fit an empirical semivariogram that suggests linear behavior.

Eventually, if none of the aforementioned issues exist, then a model is ranked in the highest positions of the fit summary table. You can combine two or more of the fitting criteria to manage classification of multiple fitted models in a more detailed manner.

In some cases you might still experience a poor quality of fit or no fit at all. If none of the earlier suggestions results in a satisfactory fit, then you could decide to re-estimate the empirical semivariogram for your same input data. The following actions can produce different empirical semivariograms to fit a theoretical model to:

- If you compute the semivariogram for different angles and you experience optimization failures, try specifying explicitly the same direction angles with different tolerance or bandwidth value in the [DIRECTIONS](#) statement.
- Modify slightly the [LAGDISTANCE=](#) option in the [COMPUTE](#) statement to obtain a different empirical semivariogram.

Finally, it is possible to have models in the fitting summary table ranked in a way that seemingly contradicts the specification in the **CHOOSE=** option of the **MODEL** statement. Consider an example with the default behavior **CHOOSE=(SSE AIC)**, where you might observe that models have the same SSE values but are not ranked further as expected by the AIC criterion. A closer examination of such cases typically reduces this issue to a matter of the accuracy shown in the table. That is, the displayed accuracy of the SSE values might hide additional decimal digits that justify the given ranking.

In such scenarios, discrimination of models at the limits of numerical accuracy might suggest that you choose a model of questionable fit or a nested structure over a more simple one. You can then review the candidate models and exercise your judgment to select the model that works best for you. If all values of a criterion are equal, then the ranking order is simply the order in which models are examined unless more criteria follow that can affect the ranking.

### **Classes of Equivalence**

The fit summary that is produced after fitting multiple models further categorizes the ranked models in classes of equivalence. Equivalence classification is an additional investigation that is unrelated to the ranking criteria presented in the previous subsection; it is an operational criterion that provides you with a qualitative overview of multiple model fit performance under given fitting conditions.

To examine model equivalence, the VARIOGRAM procedure computes the semivariances for each one of the fitted models at a set of distances. For any pair of consecutively ranked models, if the sum of their semivariance absolute differences at all designated distances is smaller than the tolerance specified by the **EQUIVTOL=** parameter, then the two models are deemed equivalent and placed in the same class; otherwise, they are placed in different classes. Equivalence classification depends on the existing ranking; hence the resulting classes can differ when you specify different ranking criteria in the **CHOOSE=** option of the **MODEL** statement.

The equivalence class numbers start at 1 for the top-ranked model in the fit summary table. You can consider the top model of each equivalence class to be a representative of the class behavior. When you specify that fit plots be produced and there are equivalence classes, the plot displays the equivalence classes and the legend designates each one by its representative model.

Consequently, if an equivalence class contains multiple members after a fit, then all of its members produce in general the exact same semivariogram. A typical reason could be that the fitting process estimates of scale parameters are at or close to their zero boundaries in one or more nested forms in a model. In such cases, the behavior of this model reduces to the behavior of its nested components with nonzero parameters. When one or more models share this situation or have the same contributing nested forms, they could end up as members of the same equivalence class depending on the ranking criteria.

It is not necessary for all models in the same equivalence class to produce the exact same semivariogram. If a fit of two obviously different forms involves semivariance values that are small enough for the equivalence criterion to be satisfied by the default value of the **EQUIVTOL=** option, then you might need to specify an even smaller value in the **EQUIVTOL=** option to rank these two models in separate equivalence classes.

### **Fitting with Matérn Forms**

When you use a Matérn form in the fitting process, it is possible that the fitting optimizer might encounter numerical difficulties if it tries to push the smoothing parameter  $\nu$  towards increasingly high values. The VARIOGRAM procedure addresses this issue by imposing an amply elevated upper bound of 1,000,000 on the smoothness values it processes. The section “[Characteristics of Semivariogram Models](#)” on page 8609

mentions that  $\nu \rightarrow \infty$  gives the Gaussian model. In the scenario of progressively increasing smoothness values, PROC VARIOGRAM acknowledges that the Matérn form behavior tends asymptotically to become Gaussian and replaces automatically the Matérn with a Gaussian form in the model. Subsequently, fitting resumes with the resulting model.

If you explore fitting of multiple models, then any duplicate models that might occur due to Matérn-to-Gaussian form conversions are fitted only once. Also, if a nested model has more than one Matérn form, then the fitting process checks one of them at a time about whether they need to be replaced by a Gaussian form. Consequently, following the switch of one Matérn form, the fitting process starts anew with the resulting model before any decisions for additional form conversions are made.

Replacement of the Matérn form with the Gaussian form occurs by default when  $\nu > 10,000$ . However, you can control this threshold value with the `MTOGTOL=` parameter of the `MODEL` statement. Practically, the Matérn form starts to resemble the Gaussian behavior for  $\nu$  values that are about  $\nu > 10$ . If you encounter such conversions of the Matérn form into Gaussian and you prefer to set a lower  $\nu$  threshold for the conversion than the default, you might experience improved code performance because computation of the Matérn semivariance can be numerically demanding.

---

## Autocorrelation Statistics

Spatial autocorrelation measures offer you additional insight into the interdependence of spatial data. These measures quantify the correlation of an SRF  $Z(s)$  with itself at different locations, and they can be very useful whether you have information at exact locations (point-referenced data) or measurements that characterize an area type such as counties, census tracts, zip codes, and so on (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,  $i = 1, \dots, n$ , 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 can 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  $z_i$ ,  $i = 1, \dots, n$ , with respect to the rest. An  $n \times n$  matrix of weights is thus defined, such that for any two locations  $s_i$  and  $s_j$ , the weight  $w_{ij}$  denotes the effect of the value  $z_i$  at location  $s_i$  on the value  $z_j$  at location  $s_j$ . Depending on the nature of your study, the weights  $w_{ij}$  need not be symmetric; that is, it can be true 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  $Z(s_i)$  at one location  $s_i$  is deemed to be connected or not, respectively, to its value  $Z(s_j)$  at another location  $s_j$ . 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  $s_i$  and  $s_j$  is closer than the user-defined distance that is defined by the **LAGDISTANCE=** option, 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  $s_i$  and  $s_j$ , respectively, is defined as

$$w_{ij} = s \frac{1}{1 + |\mathbf{h}|^p}$$

where  $\mathbf{h} = s_i - s_j$  and  $p \geq 0$  and  $s \geq 0$  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. 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 8637.

### 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 8621 and Figure 102.24); the weights are then defined as  $w_{ij} = s/[1 + (|\mathbf{h}|/h_b)^p]$ .

Most likely,  $h_b$  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 one 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  $s_i$  and  $s_j$ . However, you can also define asymmetric weights  $w'_{ij}$  such that

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

for  $i = 1, 2, \dots, n$ , where  $w'_{ij} = w_{ij} / \sum_{j \in J} w_{ij}$ ,  $i = 1, 2, \dots, n$ . In the distance-based scheme,  $J$  is the set of all locations that form point pairs with the point at  $s_i$ . In the binary scheme,  $J$  is the set of the locations that are connected to  $s_i$  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

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

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

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

These expressions indicate 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  $Z(s)$ . 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  $Z(s)$  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

$$E_g[I] = -\frac{1}{n-1}$$

and

$$E_g[I^2] = \frac{1}{(n+1)(n-1)W^2} (n^2 S_1 - n S_2 + 3W^2)$$

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

$$E_g[c] = 1$$

and

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

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

$$E_r[I] = -\frac{1}{n-1}$$

and

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

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

$$E_r[c] = 1$$

and

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

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  $B_3 = W^2[n^2 - 3 - b_2(n-1)^2]$ .

If 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 8616). 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,  $I > E[I]$  indicates positive autocorrelation. Positive autocorrelation suggests that neighboring values  $s_i$  and  $s_j$  tend to have similar feature values  $z_i$  and  $z_j$ , respectively. When  $I < E[I]$ , 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  $|I - E[I]|$ .

Geary’s  $c$  coefficient interpretation is analogous to that of Moran’s  $I$ . The only difference is that  $c > E[c]$  indicates negative autocorrelation and dissimilarity, whereas  $c < E[c]$  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

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

for Moran’s  $I$  coefficient, and

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

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  $Z_c > 0$ .

## The Moran Scatter Plot

The Moran scatter plot (Anselin 1996) is a useful visual tool for exploratory analysis, because it enables you to assess how similar an observed value is to its neighboring observations. Its horizontal axis is based on the values of the observations and is also known as the response axis. The vertical Y axis is based on the weighted average or spatial lag of the corresponding observation on the horizontal X axis. **NOTE:** The term *spatial lag* in the current context is unrelated to the concept of the semivariogram lag presented in the section “[Distance Classification](#)” on page 8619.

The Moran scatter plot provides a visual representation of spatial associations in the neighborhood around each observation. You specify a neighborhood size with the `LAGDISTANCE=` option in the `COMPUTE` statement. The observations are represented by their standardized values; therefore only nonmissing observations are shown in the plot. For each one of those, the VARIOGRAM procedure computes the weighted average, which is the weighted mean value of its neighbors. Then, the centered weighted average is plotted against the standardized observations. As a result, the scatter plot is centered on the coordinates (0, 0), and distances in the plot are expressed in deviations from the origin (0, 0).

Depending on their position on the plot, the Moran plot data points express the level of spatial association of each observation with its neighboring ones. Conceptually, these characteristics differentiate the Moran plot from the semivariogram. The latter is typically used in geostatistics to depict spatial associations across the whole domain as a continuous function of a distance metric.

You can find the data points on the Moran scatter plot in any of the four quadrants defined by the horizontal line  $y = 0$  and the vertical line  $x = 0$ . Points in the upper right (or high-high) and lower left (or low-low) quadrants indicate positive spatial association of values that are higher and lower than the sample mean, respectively. The lower right (or high-low) and upper left (or low-high) quadrants include observations that exhibit negative spatial association; that is, these observed values carry little similarity to their neighboring ones.

When you use binary, row-averaged weights for the creation of the Moran scatter plot and in autocorrelation statistics, Moran's  $I$  coefficient is equivalent to the regression slope of the Moran scatter plot. That is, when you specify

```
PLOTS=MORAN (ROWAVG=ON)
```

in the `PROC VARIOGRAM` statement and

```
AUTOCORR (WEIGHTS=BINARY (ROWAVERAGING) )
```

in the `COMPUTE` statement, then the regression line slope of the Moran scatter plot is the Moran's  $I$  coefficient shown in the section “[Autocorrelation Statistics Types](#)” on page 8637. In this sense, the Moran's  $I$  coefficient has a global character, whereas the Moran scatter plot provides you with a more detailed exploratory view of the autocorrelation behavior of the individual observations.

This detailed view can reveal outliers with respect to the regression line slope of the Moran scatter plot. Outliers, if present, can function as leverage points that affect the value of Moran's  $I$  coefficient. As noted by Anselin (1996), such extremes indicate local instability in spatial association. This instability can be caused either by problems with the autocorrelation weights matrix or by fine-scale characteristics of the spatial structure, which are beneath the current observation structure.

---

## 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.

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  $\text{NOBS}(\text{NOBS} - 1)/2 = 124,750$  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
- ID1, the ID variable value or observation number for the first point in the pair
- ID2, the ID variable value or observation number for the second point 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.

If no ID statement is specified, then the corresponding observation number is assigned to each one of the variables ID1 and ID2, instead.

### OUTDIST=SAS-data-set

The OUTDIST= data set contains counts for a modified histogram that shows 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 that fall 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 that fall in this lag class
- VARNAME, the name of the current VAR variable

### OUTMORAN=SAS-data-set

The OUTMORAN= data set contains the standardized value (or response) of each observation and the weighted average of its  $N$  neighbors, based on a neighborhood within a LAGDISTANCE= distance from the observation. To request this data set, specify the OUTMORAN= data set in the PROC VARIOGRAM statement, in addition to the AUTOCORRELATION and LAGDISTANCE= options in the COMPUTE statement.

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

- DISTANCE, the value of the neighborhood radius, which is specified with the LAGDISTANCE= option
- ID, the ID variable value or observation number for the current observation
- N, the number of neighbors within the specified DISTANCE from the current observation
- RESPONSE, the standardized value of the current observation
- STDWAVG, the standardized weighted average of the neighbors for the current observation
- V, the variable value of the current observation
- VARNAME, the variable name for the current VAR variable
- X, the  $x$  coordinate of the current observation
- Y, the  $y$  coordinate of the current observation
- WAVG, the weighted average of the neighbors for the current observation

For zero neighbors in the neighborhood of a nonmissing observation, the corresponding value of the variable N=0 and the variables STDWAVG and WAVG are assigned missing values. Observations with missing values are included in the OUTMORAN= data set if they have neighbors and only if nonmissing observations with neighbors also exist in the same data set.

### OUTPAIR=SAS-data-set

When you specify the NOVARIOGRAM option in the COMPUTE statement, the OUTPAIR= data set contains one observation for each distinct pair of points  $P_1, P_2$  in the original data set. Otherwise, the OUTPAIR= data set might have fewer observations, depending on the values you specify in the LAGDISTANCE= and MAXLAGS= options and whether you specify the OUTPDISTANCE= option in the COMPUTE statement.

If the NOVARIOGRAM option is not specified in the COMPUTE statement, then the OUTPAIR= data set contains one observation for each distinct pair of points that are up to a distance within MAXLAGS= away from each other. If you also specify the OUTPDISTANCE= $D_{max}$  option in the COMPUTE statement, then all pairs  $P_1, P_2$  in the original data set that satisfy the relation  $|P_1 P_2| \leq D_{max}$  are written to the OUTPAIR= data set.

Given the aforementioned specifications, 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 could have up to  $\text{NOBS}(\text{NOBS} - 1)/2 = 124,750$  observations 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
- ID1, the ID variable value or observation number for the first point in the pair
- ID2, the ID variable value or observation number for the second point 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

If no **ID** statement is specified, then the corresponding observation number is assigned to each one of the variables ID1 and ID2, instead.

### **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 or angle class
- AVERAGE, the average variable value for the lag or angle class
- BANDW, the bandwidth for the lag or angle class
- COUNT, the number of pairs in the lag or angle class
- COVAR, the covariance value for the lag or angle class
- DISTANCE, the average lag distance for the lag or angle class
- LAG, the lag class value (in **LAGDISTANCE=** units)
- RVARIO, the sample robust semivariance value for the lag or angle class

- STDERR, the approximate standard error of the sample semivariance estimate
- VARIOG, the sample semivariance value for the lag or angle class
- VARNAME, the name of the current VAR variable

The robust semivariance estimate, RVARIO, is not included in the data set if you omit the option ROBUST in the COMPUTE 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  $Z(s)$ , and the COVAR variable shows the sample variance  $\text{Var}[Z(s)]$ .

---

## Displayed Output

In addition to the output data sets, the VARIOGRAM procedure produces a variety of output objects. Most of these 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 that shows 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 PLOTS=PAIRS option. The option PLOTS=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 that contains 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
- the Moran scatter plot of the standardized observation values against the weighted averages of their neighbors, when the options **PLOTS=MORAN**, **AUTOCORRELATION**, and **LAGDISTANCE=** are specified

When you specify the **MODEL** statement and request a fit of a theoretical model to the empirical semivariogram, the **VARIOGRAM** procedure also produces the following default output:

- a table with some general fitting information, in addition to the output item store if you have specified one with the **STORE** statement
- a table with more specific information about the selected model's parameters and their initial values
- a table with general information about the optimization that provides the fitting parameters of the selected model
- a table with the optimization process output and a table with the convergence status of the optimization process, if you have specified a single model to fit
- a "Parameter Estimates" table with information about the fitted parameters estimates
- a "Fit Summary" table that reports the fit quality of all models you requested to fit
- plots of fitted theoretical semivariogram models. If you perform model fitting in more than one direction angle or for more than one variable in your **DATA=** data set, then the output is a panel that contains all fitted models for the respective directions or variables.

Additional output can be produced in model fitting if you specify a higher level of output detail with the **DETAILS** option in the **MODEL** statement. This output can be information tables for each separate model when you specify multiple models to fit, tables with more details about the optimization process, and the covariance and correlation matrices of the model parameter estimates. The complete listing of the PROC **VARIOGRAM** output follows in the section "**ODS Table Names**" on page 8646 and the section "**ODS Graph Names**" on page 8647.

## ODS Table Names

Each table created by PROC VARIOGRAM has a name associated with it, and you must use this name to refer to the table when using ODS Graphics. These names are listed in Table 102.6.

**Table 102.6** ODS Tables Produced by PROC VARIOGRAM

| ODS Table Name            | Description                                                                                                   | Required Statement | Option             |
|---------------------------|---------------------------------------------------------------------------------------------------------------|--------------------|--------------------|
| AutoCorrStats             | Autocorrelation statistics information                                                                        | COMPUTE            | AUTOCORRELATION    |
| ConvergenceStatus         | Status of optimization at conclusion                                                                          | MODEL              | Default output     |
| CorrB                     | Approximate correlation matrix of model parameter estimates                                                   | MODEL              | CORRB              |
| CovB                      | Approximate covariance matrix of model parameter estimates                                                    | MODEL              | COVB               |
| DistanceIntervals         | Pairwise distances matrix                                                                                     | COMPUTE            | NOVARIOGRAM        |
| FitGenInfo                | General fitting information                                                                                   | MODEL              | Default output     |
| FitSummary                | Fitting process summary                                                                                       | MODEL              | Default output     |
| InputOptions              | Optimization input options                                                                                    | MODEL              | DETAILS=ALL        |
| IterHist                  | Iteration history                                                                                             | MODEL              | DETAILS=ITR        |
| IterStop                  | Optimization-related results                                                                                  | MODEL              | Default output     |
| Lagrange                  | Information about Lagrange multipliers                                                                        | MODEL              | DETAILS=ALL        |
| ModelInfo                 | Model information                                                                                             | MODEL              | Default output     |
| NObs                      | Number of observations read and used                                                                          | PROC               | Default output     |
| OptInfo                   | Optimization information                                                                                      | MODEL              | Default output     |
| PairsInformation          | General information about the pairs distribution in classes and data maximum distances in selected directions | COMPUTE            | NOVARIOGRAM        |
| ParameterEstimates        | Model fitting solution and statistics                                                                         | MODEL              | Default output     |
| ParameterEstimatesResults | Parameter estimates and gradient information                                                                  | MODEL              | DETAILS=ALL        |
| ParameterEstimatesStart   | More detailed model information                                                                               | MODEL              | DETAILS=ITR        |
| ParmSearch                | Parameter search values                                                                                       | MODEL              | Default output     |
| ProblemDescription        | Information at the optimization start                                                                         | MODEL              | DETAILS=ITR        |
| ProjGrad                  | Projected gradient information                                                                                | MODEL              | DETAILS=ALL        |
| SemivariogramTable        | Empirical semivariance classes, parameters, and estimates                                                     | COMPUTE            | LAGD=,<br>MAXLAGS= |

## ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, “[Statistical Graphics Using ODS](#).”

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “[Enabling and Disabling ODS Graphics](#)” on page 600 in Chapter 21, “[Statistical Graphics Using ODS](#).”

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section “[A Primer on ODS Statistical Graphics](#)” on page 599 in Chapter 21, “[Statistical Graphics Using ODS](#).”

For additional control of the graphics that are displayed, see the [PLOTS](#) option in the section “[PROC VARIOGRAM Statement](#)” on page 8576.

## ODS Graph Names

PROC VARIOGRAM assigns a name to each graph it creates by using ODS Graphics. You can use this name to refer to the graph when using ODS Graphics. You must also specify the [PLOTS=](#) option indicated in [Table 102.7](#).

**Table 102.7** Graphs Produced by PROC VARIOGRAM

| ODS Graph Name                     | Plot Description                                                                 | Statement | Option                        |
|------------------------------------|----------------------------------------------------------------------------------|-----------|-------------------------------|
| <a href="#">FitPanel</a>           | Panel of one or more classes of fitted semivariograms in different angles        | PROC      | <a href="#">PLOTS=FIT</a>     |
| <a href="#">FitPlot</a>            | Plot of one or more classes of fitted semivariograms                             | PROC      | <a href="#">PLOTS=FIT</a>     |
| <a href="#">MoranPlot</a>          | Scatter plot of standardized observed values against weighted averages           | PROC      | <a href="#">PLOTS=MORAN</a>   |
| <a href="#">ObservationsPlot</a>   | Scatter plot of observed data and colored markers that indicates observed values | PROC      | <a href="#">PLOTS=OBSERV</a>  |
| <a href="#">PairDistPlot</a>       | Histogram of the pairwise distance distribution                                  | PROC      | <a href="#">PLOTS=PAIRS</a>   |
| <a href="#">Semivariogram</a>      | Plots of empirical classical and robust (optional) semivariograms                | PROC      | <a href="#">PLOTS=SEMIVAR</a> |
| <a href="#">SemivariogramPanel</a> | Panel of empirical classical and robust (optional) semivariogram plots           | PROC      | <a href="#">PLOTS=SEMIVAR</a> |

## Examples: VARIOGRAM Procedure

### Example 102.1: Aspects of Semivariogram Model Fitting

This example helps you explore aspects of automated semivariogram fitting with PROC VARIOGRAM. The test case is a spatial study of arsenic (As) concentration in drinking water.

Arsenic is a toxic pollutant that can occur in drinking water because of human activity or, typically, due to natural release from the sediments in water aquifers. The World Health Organization has a standard that allows As concentration up to a maximum of 10  $\mu\text{g/l}$  (micrograms per liter) in drinking water.

In general, natural release of arsenic into groundwater is very slow. Arsenic concentration in water might exhibit no significant temporal fluctuations over a period of a few months. For this reason, it is acceptable to perform a spatial study of arsenic with input from time-aggregated pollutant concentrations. This example makes use of this assumption for its data set `logAsData`. The data set consists of 138 simulated observations from wells across a square area of 500 km  $\times$  500 km. The variable `logAs` in the `logAsData` data set is the natural logarithm of arsenic concentration. Often, the natural logarithm of arsenic concentration (`logAs`) is used as the random variable to facilitate the analysis because its distribution tends to resemble the normal distribution.

The goal is to explore spatial continuity in the `logAs` observations. The following statements read the `logAs` values from the `logAsData` data set:

```
title 'Semivariogram Model Fitting of Log-Arsenic Concentration';

data logAsData;
  input East North logAs @@;
  label logAs='log(As) Concentration';
  datalines;
193.0 296.6 -0.68153   232.6 479.1   0.96279   268.7 312.5 -1.02908
  43.6   4.9  0.65010   152.6  54.9   1.87076   449.1 395.8  0.95932
310.9 493.6 -1.66208   287.8 164.9 -0.01779   330.0   8.0  2.06837
225.7 241.7  0.15899   452.3  83.4 -1.21217   156.5 462.5 -0.89031
  11.5  84.4 -0.24496   144.4 335.7  0.11950   149.0 431.8 -0.57251
234.3 123.2 -1.33642    37.8 197.8 -0.27624   183.1 173.9 -2.14558
149.3 426.7 -1.06506   434.4  67.5 -1.04657   439.6 237.0 -0.09074
  36.4 175.2 -1.21211   370.6 244.0  3.28091   452.0  96.5 -0.77081
247.0  86.8  0.04720   413.6 373.2  1.78235   253.5 291.7  0.56132
129.7 111.9  1.34000   352.7  42.1  0.23621   279.3  82.7  2.12350
382.6 290.7  0.86756   188.2 222.8 -1.23308   382.8 154.5 -0.94094
304.4 309.2 -1.95158   337.5 387.2 -1.31294   490.7 189.8  0.40206
159.0 100.1 -0.22272   245.5 329.2 -0.26082   372.1 379.5 -1.89078
417.8  84.1 -1.25176   173.9 407.6 -0.24240   121.5 107.7  1.54509
453.5 313.6  0.65895   143.5 346.7 -0.87196   157.4 125.5 -1.96165
371.8 353.2 -0.59464   358.9 338.2 -1.07133     8.6 437.8  1.44203
395.9 394.2 -0.24144   149.5  58.9  1.17459   453.5 420.6 -0.63951
182.3  85.0  1.00005    21.0 290.1  0.31016    11.1 352.2 -0.88418
131.2 238.4 -0.57184   104.9   6.3  1.12054   247.3 256.0  0.14019
428.4 383.7  0.92448   327.8 481.1 -2.72543   199.2  92.8 -0.05717
453.9 230.1  0.16571   205.0 250.6  0.07581   459.5 271.6  0.93700
```

```

229.5 262.8 1.83590 370.4 228.6 2.96611 330.2 281.9 1.79723
354.8 388.3 -3.18262 406.2 222.7 2.41594 254.4 393.1 2.03221
96.7 85.2 -0.47156 407.2 256.8 0.66747 498.5 273.8 1.03041
417.2 471.4 -1.42766 368.8 424.3 -0.70506 303.0 59.1 1.43070
403.1 264.1 1.64554 21.2 360.8 0.67094 148.2 78.1 2.15323
305.5 310.7 -1.47985 228.5 180.3 -0.68386 161.1 143.3 1.07901
70.5 155.1 0.54652 363.1 282.6 -0.43051 86.0 472.5 -1.18855
175.9 105.3 -2.08112 96.8 426.3 1.56592 475.1 453.1 -1.53776
125.7 485.4 1.40054 277.9 201.6 -0.54565 406.2 125.0 -1.38657
60.0 275.5 -0.59966 431.3 494.6 -0.36860 399.9 399.0 -0.77265
28.8 311.1 0.91693 166.1 348.2 -0.49056 266.6 83.5 0.67277
54.7 356.3 0.49596 433.5 460.3 -1.61309 201.7 167.6 -1.40678
158.1 203.6 -1.32499 67.6 230.4 1.14672 81.9 250.0 0.63378
372.0 50.7 0.72445 26.4 264.6 1.00862 300.1 91.7 -0.74089
303.0 447.4 1.74589 108.4 386.2 1.12847 55.6 191.7 0.95175
36.3 273.2 1.78880 94.5 298.3 -2.43320 366.1 187.3 -0.80526
130.7 389.2 -0.31513 37.2 324.2 0.24489 295.5 211.8 0.41899
58.6 206.2 0.18495 346.3 142.8 -0.92038 484.2 215.9 0.08012
451.4 415.7 0.02773 58.9 86.5 0.17652 212.6 363.9 0.17215
378.7 407.6 0.51516 265.9 305.0 -0.30718 123.2 314.8 -0.90591
26.9 471.7 1.70285 16.5 7.1 0.51736 255.1 472.6 2.02381
111.5 148.4 -0.09658 440.4 375.0 1.23285 406.4 19.5 1.01181
321.2 65.8 -0.02095 466.4 357.1 -0.49272 2.0 484.6 0.50994
200.9 205.1 0.43543 30.3 337.0 1.60882 297.0 12.7 1.79824
158.2 450.7 0.05295 122.8 105.3 1.53936 417.8 329.7 -2.08124
;

```

First you want to inspect the `logAs` data for surface trends and the pairwise distribution. You run the `VARIOGRAM` procedure with the `NOVARIogram` option in the `COMPUTE` statement. You also request the `PLOTS=PAIRS(MID)` option, which prompts the pair distance plot to display the actual distance between pairs, rather than the lag number itself, in the midpoint of the lags. You use the following statements:

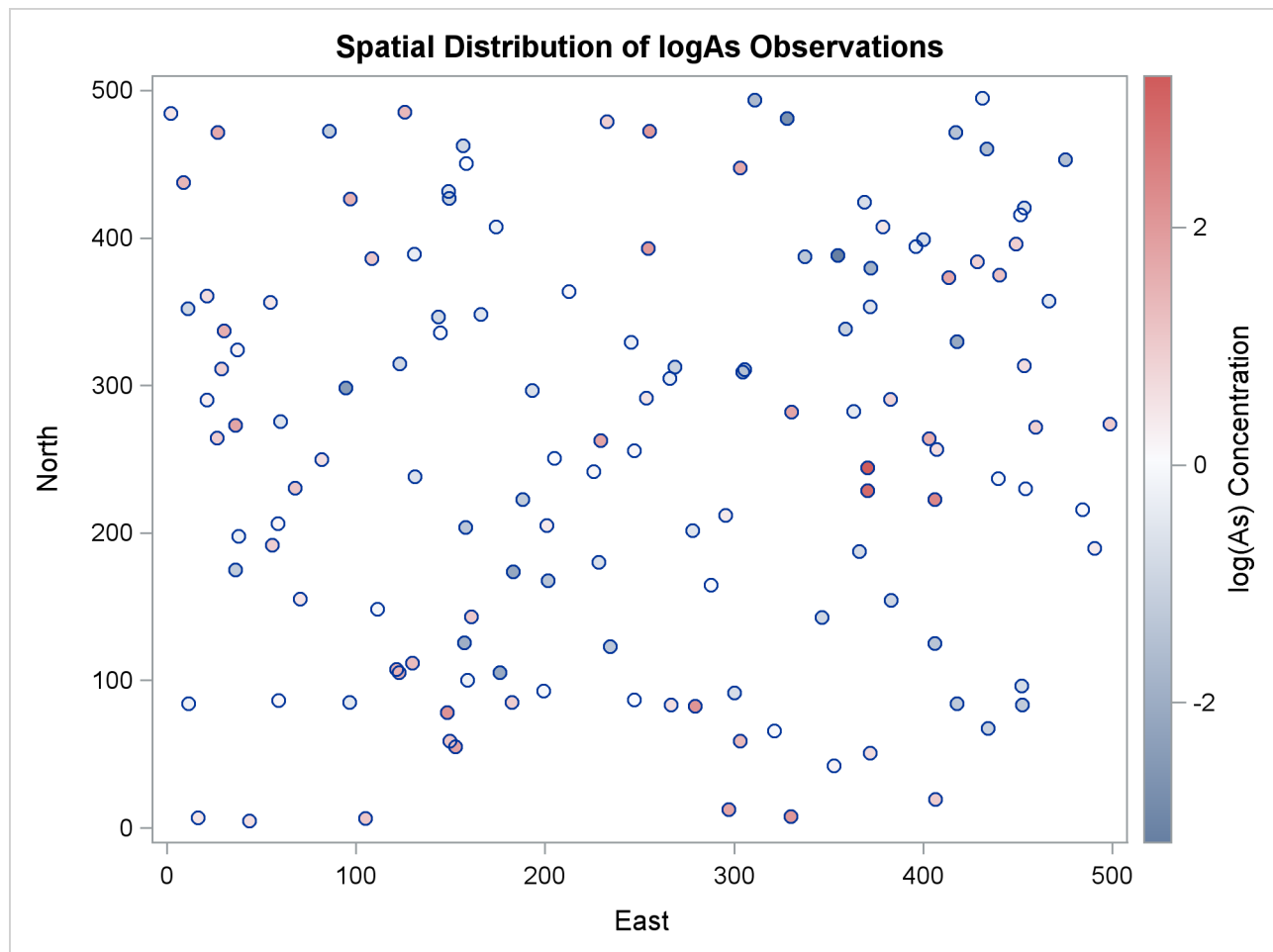
```

ods graphics on;

proc variogram data=logAsData plots=pairs(mid);
  compute novariogram nhc=50;
  coord xc=East yc=North;
  var logAs;
run;

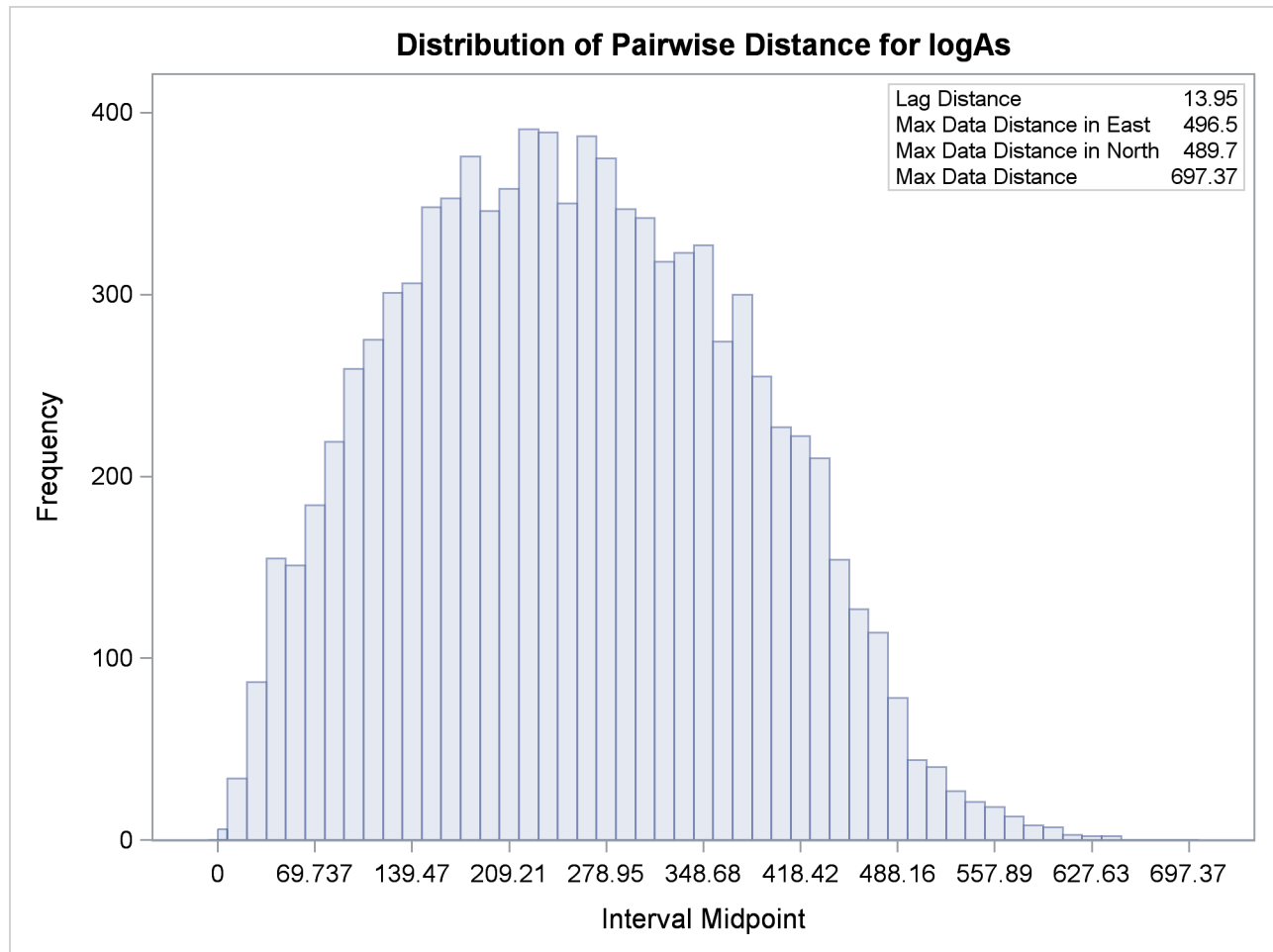
```

The observations scatter plot in [Output 102.1.1](#) shows a rather uniform distribution of the locations in the study domain. Reasonably, neighboring values of `logAs` seem to exhibit some correlation. There seems to be no definite sign of an overall surface trend in the `logAs` values. You can consider that the observations are trend-free, and proceed with estimation of the empirical semivariance.

**Output 102.1.1** logAs Observation Data Scatter Plot

The observed logAs values go as high as 3.28091, which corresponds to a concentration of  $26.6 \mu\text{g/l}$ . In fact, only three observations exceed the health standard of  $10 \mu\text{g/l}$  (or about 2.3 in the log scale), and they are situated in relatively neighboring locations to the east of the domain center.

Based on the discussion in section “[Preliminary Spatial Data Analysis](#)” on page 8562, the pair distance plot in [Output 102.1.2](#) suggests that you could consider pairs that are anywhere around up to half the maximum pairwise distance of about 700 km.

**Output 102.1.2** Distribution of Pairwise Distances for logAs Data

After some experimentation with values for the `LAGDISTANCE=` and `MAXLAGS=` options, you actually find that a lag distance of 5 km over 40 lags can provide a clear representation of the logAs semivariance. With respect to [Output 102.1.2](#), this finding indicates that in the current example it is sufficient to consider pairs separated by a distance of up to 200 km. You run the following statements to obtain the empirical semivariogram:

```
proc variogram data=logAsData plots(only)=semivar;
  compute lagd=5 maxlag=40;
  coord xc=East yc=North;
  var logAs;
run;
```

The first few lag classes of the logAs empirical semivariance table are shown in [Output 102.1.3](#).

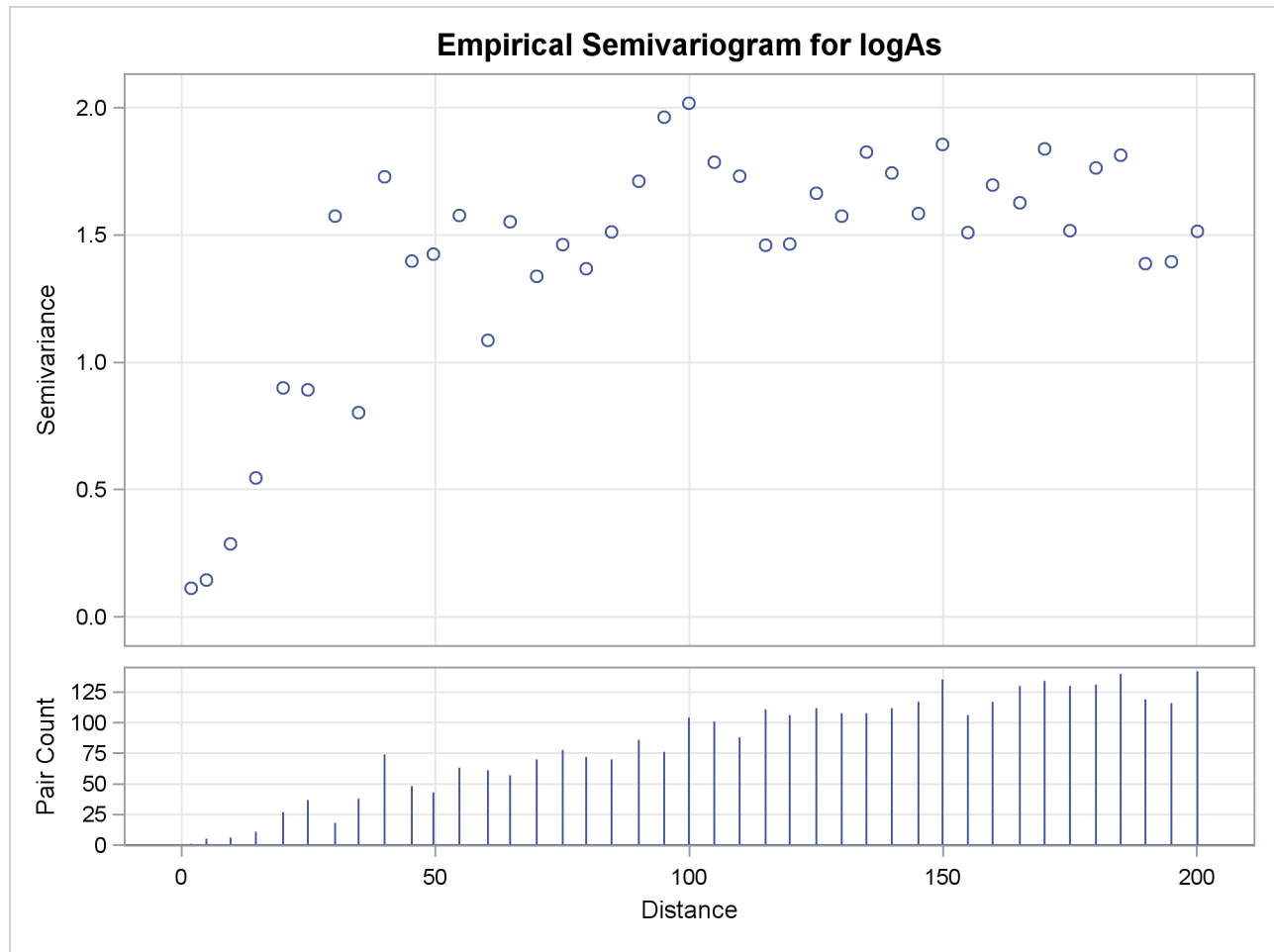
**Output 102.1.3** Partial Output of the Empirical Semivariogram Table for logAs Data

| Semivariogram Model Fitting of Log-Arsenic Concentration |            |                  |              |  |
|----------------------------------------------------------|------------|------------------|--------------|--|
| The VARIOGRAM Procedure                                  |            |                  |              |  |
| Dependent Variable: logAs                                |            |                  |              |  |
| Empirical Semivariogram                                  |            |                  |              |  |
| Lag Class                                                | Pair Count | Average Distance | Semivariance |  |
| 0                                                        | 1          | 1.9              | 0.111        |  |
| 1                                                        | 5          | 4.9              | 0.145        |  |
| 2                                                        | 6          | 9.7              | 0.286        |  |
| 3                                                        | 11         | 14.6             | 0.545        |  |
| 4                                                        | 27         | 20.0             | 0.900        |  |

[Output 102.1.3](#) and [Output 102.1.4](#) indicate that the logarithm of arsenic spatial correlation starts with a small nugget effect around 0.11 and rises to a sill value that is most likely between 1.4 and 1.8. The rise could be of exponential type, although the smooth increase of semivariance close to the origin could also suggest Gaussian behavior. You suspect that a Matérn form might also work, since its smoothness parameter  $\nu$  can regulate the form to exhibit an intermediate behavior between the exponential and Gaussian forms.



**Output 102.1.4** Empirical Semivariogram for logAs Data



You can investigate all of the preceding clues with the model fitting features of PROC VARIOGRAM. The simplest way to fit a model is to specify its form in the **MODEL** statement. In this case, you have the added complexity of having more than one possible candidate. For this reason, you use the **FORM=AUTO** option that picks the best fit out of a list of candidates. Within this option you specify the **MLIST=** suboption to use the exponential, Gaussian, and Matérn forms. You also specify the **NEST=** suboption to request fitting of a model with up to two nested structures. Eventually, you specify the **PLOTS=FIT** option to produce a plot of the fitted models. The **STORE** statement saves the fitting output into an item store you name **SemivAsStore** for future use. You apply these specifications with the following statements:

```
proc variogram data=logAsData plots(only)=fit;
  store out=SemivAsStore / label='LogAs Concentration Models';
  compute lagd=5 maxlag=40;
  coord xc=East yc=North;
  model form=auto(mlist=(exp,gau,mat) nest=1 to 2);
  var logAs;
run;

ods graphics off;
```

The table of general information about fitting is shown in [Output 102.1.5](#). The table lets you know that 12 model combinations are to be tested for weighted least squares fitting, based on the three forms that you specified.

**Output 102.1.5** Semivariogram Model Fitting General Information

| Semivariogram Model Fitting of Log-Arsenic Concentration |                                     |
|----------------------------------------------------------|-------------------------------------|
| The VARIOGRAM Procedure                                  |                                     |
| Dependent Variable: logAs                                |                                     |
| Angle: Omnidirectional                                   |                                     |
| Semivariogram Model Fitting                              |                                     |
| Model                                                    | Selection from 12 form combinations |
| Output Item Store                                        | WORK.SEMIVASSTORE                   |
| Item Store Label                                         | LogAs Concentration Models          |

The combinations include repetitions. For example, you specified the GAU form; hence the GAU-GAU form is tested, too. The model combinations also include permutations. For example, you specified the GAU and the EXP forms; hence the GAU-EXP and EXP-GAU models are fitted separately. According to the section “[Nested Models](#)” on page 8613, it might seem that the same model is fitted twice. However, in each of these two cases, each structure starts the fitting process with different parameter initial values. This can lead GAU-EXP to a different fit than EXP-GAU leads to, as seen in the fitting summary table in [Output 102.1.6](#). The table shows all the model combinations that were tested and fitted. By default, the ordering is based on the weighted sum of squares error criterion, and you can see that the lowest values in the Weighted SSE column are in top slots of the list.

**Output 102.1.6** Semivariogram Model Fitting Summary

| Fit Summary |         |                 |           |
|-------------|---------|-----------------|-----------|
| Class       | Model   | Weighted<br>SSE | AIC       |
| 1           | Gau-Gau | 25.42435        | -9.59246  |
|             | Gau-Mat | 25.42482        | -7.59169  |
| 2           | Exp-Gau | 25.97835        | -8.70865  |
| 3           | Exp-Mat | 26.36846        | -6.09754  |
| 4           | Mat     | 26.37519        | -10.08708 |
| 5           | Gau     | 26.78629        | -11.45296 |
| 6           | Exp     | 28.01200        | -9.61851  |
|             | Exp-Exp | 28.01200        | -5.61850  |
|             | Mat-Exp | 28.01200        | -3.61850  |
|             | Gau-Exp | 28.01200        | -5.61850  |

Note the leftmost Class column in [Output 102.1.6](#). As explained in detail in section “Classes of Equivalence” on page 8634, when you fit more than one model, all fitted models that compute the same semi-variance are placed in the same class of equivalence. For example, in this fitting example the top ranked GAU-GAU and GAU-MAT nested models produce indistinguishable semivariograms; for that reason they are both placed in the same class 1 of equivalence. The same occurs with the EXP, GAU-EXP, EXP-EXP, and MAT-EXP models in the bottom of the table. By default, PROC VARIOGRAM uses the AIC as a secondary classification criterion; hence models in each equivalence class are already ordered based on their AIC values.

Another remark in [Output 102.1.6](#) is that despite submitting 12 model combinations for fitting, the table shows only 10. You can easily see that the combinations MAT-GAU and MAT-MAT are not among the listed models in the fit summary. This results from the behavior of the VARIOGRAM procedure in the following situation: A parameter optimization takes place during the fitting process. In the present case the optimizer keeps increasing the Matérn smoothness parameter  $\nu$  in the MAT-GAU model. At the limit of an infinite  $\nu$  parameter, the Matérn form becomes the Gaussian form. For that reason, when the parameter  $\nu$  is driven towards very high values, PROC VARIOGRAM automatically replaces the Matérn form with the Gaussian. This switch converts the MAT-GAU model into a GAU-GAU model. However, a GAU-GAU model already exists among the specified forms; consequently, the duplicate GAU-GAU model is skipped, and the fitted model list is reduced by one model. A similar explanation justifies the omission of the MAT-MAT model from the fit summary table.

In our example, the nested Gaussian-Gaussian model is the fitting selection of the procedure based on the default ranking criteria. [Output 102.1.7](#) displays additional information about the selected model. In particular, you see the table with general information about the Gaussian-Gaussian model, the initial values used for its parameters, and information about the optimization process for the fitting.

**Output 102.1.7** Fitting and Optimization Information for Gaussian-Gaussian Model

| Semivariogram Model Fitting |                   |
|-----------------------------|-------------------|
| Name                        | Gaussian-Gaussian |
| Label                       | Gau-Gau           |
| Model Information           |                   |
| Parameter                   | Initial Value     |
| Nugget                      | 0.0903            |
| GauScale1                   | 0.6709            |
| GauRange1                   | 100.0             |
| GauScale2                   | 0.6709            |
| GauRange2                   | 50.0230           |
| Optimization Information    |                   |
| Optimization Technique      | Dual Quasi-Newton |
| Parameters in Optimization  | 5                 |
| Lower Boundaries            | 5                 |
| Upper Boundaries            | 0                 |
| Starting Values From        | PROC              |

The estimated parameter values of the selected Gaussian-Gaussian model are shown in [Output 102.1.8](#).

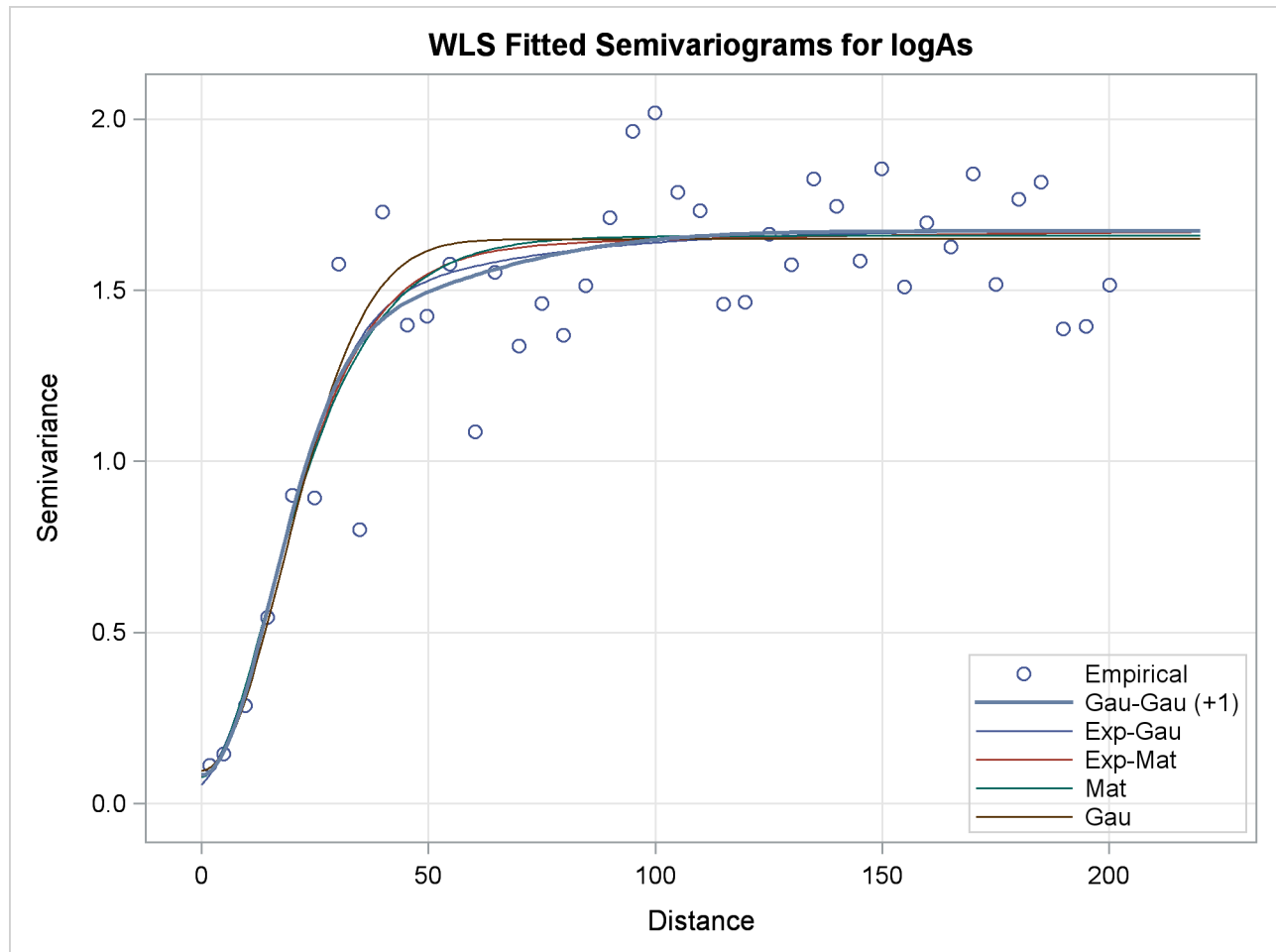
**Output 102.1.8** Parameter Estimates of the Fitting Selected Model

| Parameter Estimates |          |                  |    |         |                |
|---------------------|----------|------------------|----|---------|----------------|
| Parameter           | Estimate | Approx Std Error | DF | t Value | Approx Pr >  t |
| Nugget              | 0.08308  | 0.05097          | 36 | 1.63    | 0.1118         |
| GauScale1           | 0.3277   | 0.2077           | 36 | 1.58    | 0.1234         |
| GauRange1           | 62.3127  | 19.8488          | 36 | 3.14    | 0.0034         |
| GauScale2           | 1.2615   | 0.2070           | 36 | 6.10    | <.0001         |
| GauRange2           | 21.4596  | 3.2722           | 36 | 6.56    | <.0001         |

By default, when you specify more than one model to fit, PROC VARIOGRAM produces a fit plot that compares the first five classes of the successfully fit candidate models. The model that is selected according to the specified fitting criteria is shown with a thicker line in the plot.

You can modify the number of displayed equivalence classes with the **NCLASSES=** suboption of the **PLOTS=FIT** option. When you have such comparison plots, PROC VARIOGRAM displays the representative model from each class of equivalence.

The default fit plot for the current model comparison is shown in [Output 102.1.9](#). The legend informs you there is one more model in the first class of equivalence, as the fitting summary table indicated earlier in [Output 102.1.6](#).

**Output 102.1.9** Fitted Theoretical and Empirical logAs Concentration Semivariograms for the Specified Models

In the present example, all fitted models in the first five classes have very similar semivariograms. The selected Gaussian-Gaussian model seems to have a relatively larger range than the rest of the displayed models, but you can expect any of these models to exhibit a near-identical behavior in terms of spatial correlation. As a result, all models in the displayed classes are likely to lead to very similar output, if you proceed to use any of them for spatial prediction.

In that sense, semivariogram fitting is a partially subjective process, for which there might not exist only one single correct answer to solve your problem. In the context of the example, on one hand you might conclude that the selected Gaussian-Gaussian model is exactly sufficient to describe spatial correlation in the arsenic study. On the other hand, the similar performance of all models might prompt you to choose instead a more simple non-nested model for prediction like the Matérn or the Gaussian model.

Regardless of whether you might opt to sacrifice the statistically best fit (depending on your selected criteria) to simplicity, eventually you are the one to decide which approach serves your study optimally. The model fitting features of PROC VARIOGRAM offer you significant assistance so that you can assess your options efficiently.

## Example 102.2: An Anisotropic Case Study with Surface Trend in the Data

This example shows how to examine data for nonrandom surface trends and anisotropy. You 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 work with the ozoneSet data set that contains 300 measurements in a square area of 100 km  $\times$  100 km.

The following statements read the data set:

```

title 'Semivariogram Analysis in Anisotropic Case With Trend Removal';

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
;

```

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:

```

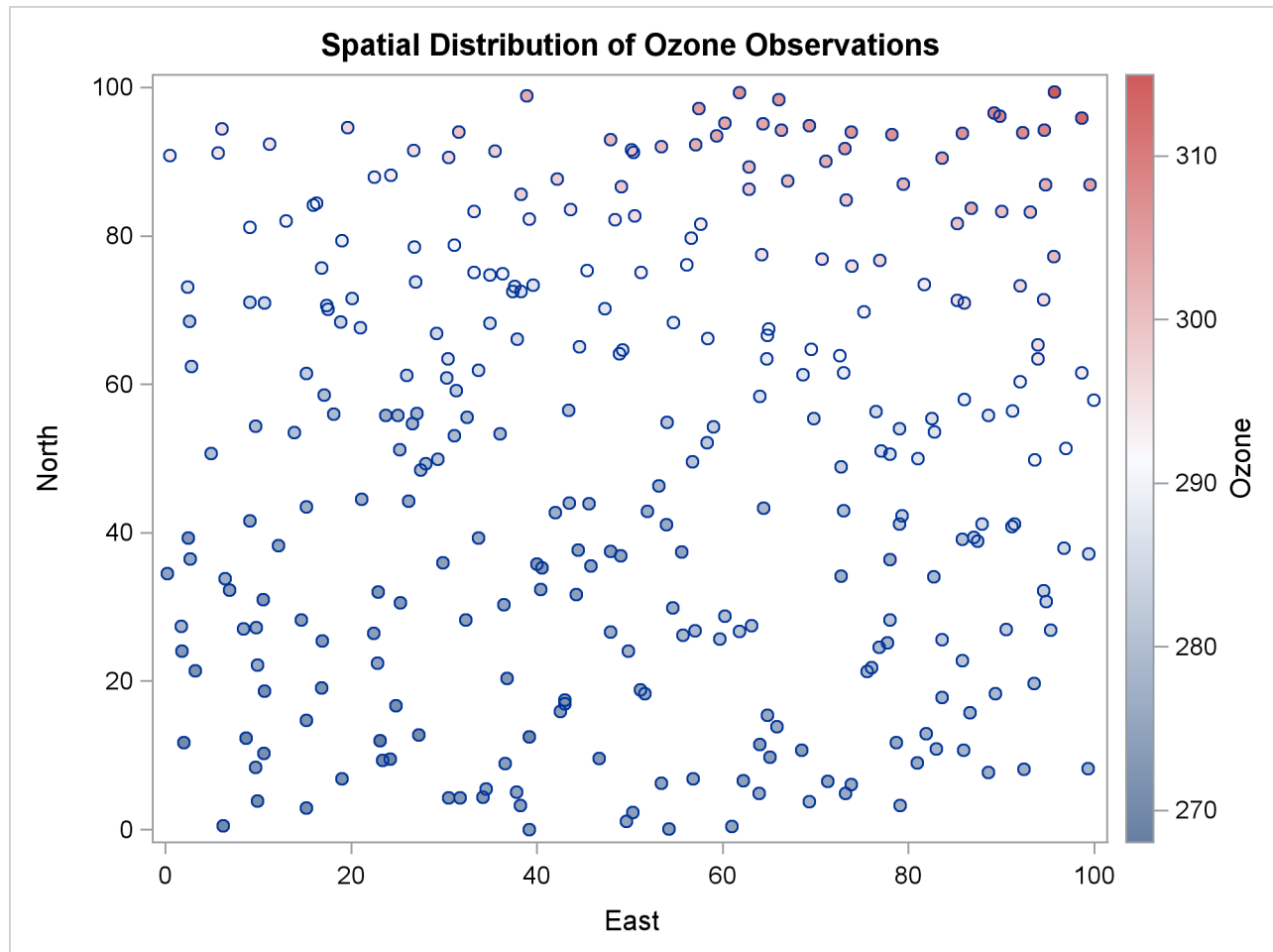
ods graphics on;

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 102.2.1](#). 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 102.2.1** Ozone Observation Data Scatter Plot



You 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 “[Getting Started: VARIOGRAM Procedure](#)” on page 8561, the larger size of the `ozoneSet` data results in more densely populated distance classes for the same value of the `NHCLASSES=` option. After you experiment with a variety of values for the `NHCLASSES=` 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 `NHCLASSES=35`.

Your choice of `NHCLASSES=35` yields the pairwise distance intervals table in [Output 102.2.2](#) and the corresponding histogram in [Output 102.2.3](#).

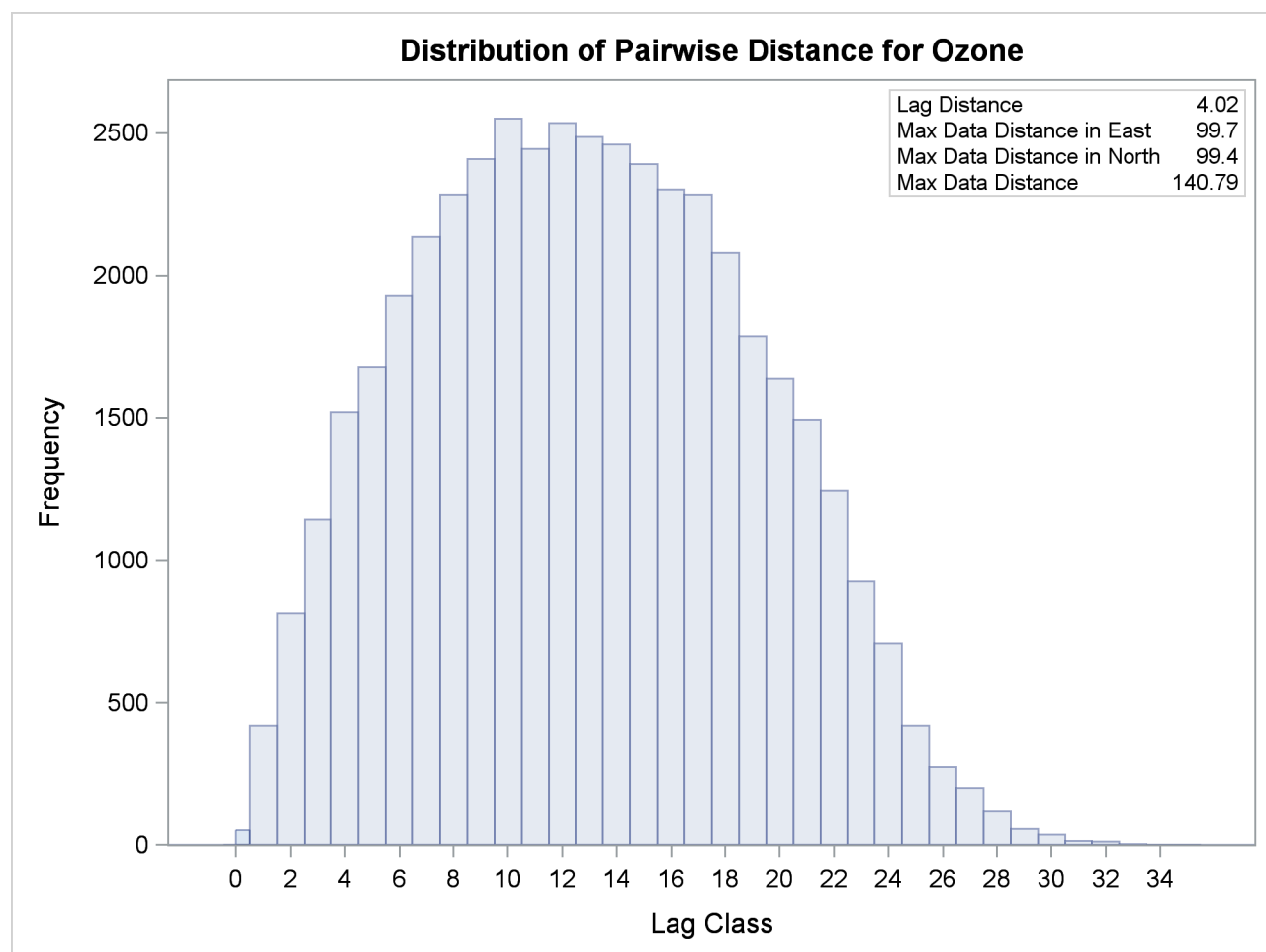


**Output 102.2.2** Pairwise Distance Intervals Table

| 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 102.2.2](#). You can see that even for higher values of `NHCLASSES=` 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 8623. Based on the displayed information in [Output 102.2.3](#), you specify `LAGDISTANCE=4` km. You can further experiment with smaller lag sizes to obtain more points in your sample semivariogram.

You can focus on the `MAXLAGS=` specification at a later point. 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 102.2.3** 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 8626. In this case, the spatial distribution of the measurements shown in [Figure 102.2.1](#) 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 plots=none;
  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  $f_0, \dots, f_4$  in the preceding trend model. [Output 102.2.4](#) 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 42, “[The GLM Procedure](#).”

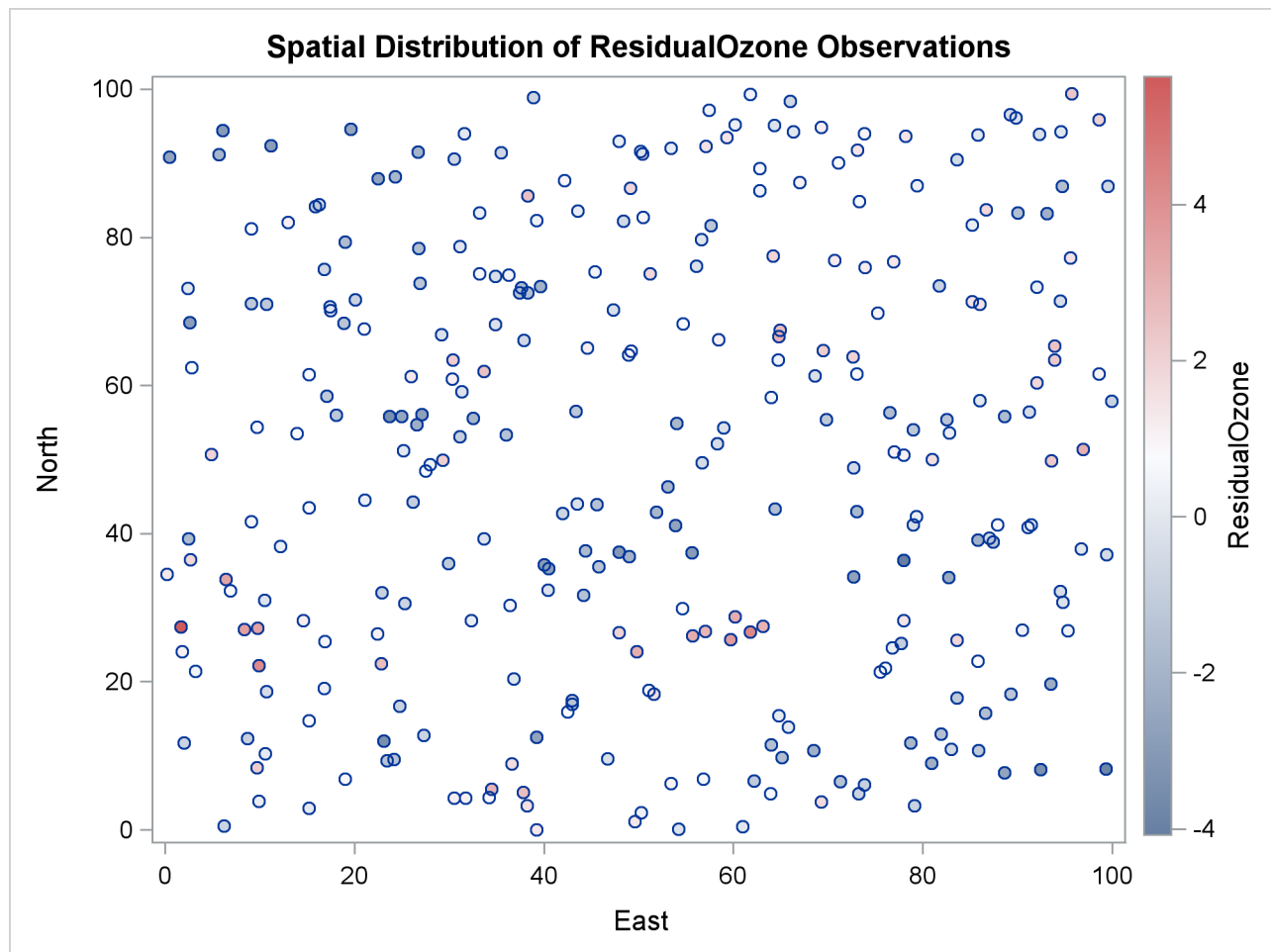
**Output 102.2.4** Parameter Estimates for the Surface Trend Model

| Semivariogram Analysis in Anisotropic Case With Trend Removal |             |                |         |         |
|---------------------------------------------------------------|-------------|----------------|---------|---------|
| 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. This time you run PROC VARIOGRAM again with the [NOVARIOGRAM](#) option to inspect the detrended residuals, and with a request only for the observations plot, as follows:

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

The requested observations plot is shown in [Output 102.2.5](#).

**Output 102.2.5** 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 manifests itself as different ranges or sills or both 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  $\theta = 180^\circ/12 = 15^\circ$ .

You also choose `ANGLETOLERANCE=22.5` and `BANDWIDTH=20`. A different choice of values produces different empirical semivariograms, because these options can regulate the number of pairs that are included in a class. Avoid assigning values that are too small to these parameters so that you can allow for an adequate number of point pairs per class. At the same time, the higher the values of these parameters are, the more data pairs that come from closely neighboring directions are included in each lag. Therefore, values for the `ANGLETOLERANCE=` and `BANDWIDTH=` options that are too high pose a risk of losing information along the particular direction. The side effect occurs because you incorporate data pairs from a broader spectrum of angles; thus, you potentially amplify weaker anisotropy or weaken stronger anisotropy, as noted in the section “Angle Classification” on page 8617. 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 102.2.6](#) 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;
  coord xc=East yc=North;
  var ResidualOzone;
run;
```

**Output 102.2.6** Ozone Empirical Semivariograms with  $0^\circ \leq \theta < 180^\circ$  and  $\delta\theta = 15^\circ$



Output 102.2.6 *continued*

**Output 102.2.6** *continued*



The panels in [Output 102.2.6](#) suggest that in some of the directions, such as for  $\theta = 0^\circ$ , 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 102.5: A Box Plot of the Square Root Difference Cloud](#)” on page 8684.

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 102.2.6](#) suggest that during shifting from  $\theta = 0^\circ$  to  $\theta = 90^\circ$ , the empirical semivariogram range increases. Beyond the angle  $\theta = 90^\circ$ , the range starts decreasing again until the whole circle is traversed at  $180^\circ$  and small range values are encountered around the N–S direction at  $\theta = 0^\circ$ . The sill seems to remain overall the same. This analysis suggests the presence of anisotropy in the ozone concentrations, with the major axis oriented at about  $\theta = 90^\circ$  and the minor axis situated perpendicular to the major axis at  $\theta = 0^\circ$ .

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 8624. Following the information displayed in [Output 102.2.3](#), 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 102.2.2](#) indicates that within this range of distances you can specify `MAXLAGS=` to be between 12 and 17 lags. In this example you specify `MAXLAGS=16`.

At this point you are ready to continue with fitting theoretical semivariogram models to the empirical semivariogram in the selected directions of  $\theta = 0^\circ$  and  $\theta = 90^\circ$ . By trying out different models, you see that an exponential one is suitable for your empirical data:

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

For the purpose of the present example, it is reasonable to assume a constant nugget effect equal to zero, based on the empirical semivariograms shown in [Output 102.2.6](#). The same output suggests that the model scale is likely to be above 2, and that the range might be relatively small in  $\theta = 0^\circ$ . You specify the `PARMS` statement to set initial values for the exponential model parameters and account for these considerations.

In particular, you assign an initial value of zero to the nugget effect. Then you request a grid search for the range and scale parameters, so that the optimal initial values set is selected for the parameter estimation in each of the two angles  $\theta = 0^\circ$  and  $\theta = 90^\circ$ . By inspecting the empirical semivariograms in [Output 102.2.6](#), you specify the value list 2, 2.5, and 3 for the scale, and the values from 5 to 25 with a step of 10 for the range. In addition, you specify the parameter 1 in the `HOLD=` option to designate the nugget effect parameter as a constant. According to these specifications, you use the following statements:

```
proc variogram data=gmount plot(only)=fit;
  compute lagd=4 maxlag=16;
  directions 0(22.5,10) 90(22.5,10);
  coord xc=East yc=North;
```



```

model form=exp;
parms (0.) (2 to 3 by 0.5) (5 to 25 by 10) / hold=(1);
var ResidualOzone;
run;

ods graphics off;

```

The VARIOGRAM procedure repeats the fitting process for each one of the selected directions. First, in  $\theta = 0^\circ$  the parameter search table in [Output 102.2.7](#) shows you which value combinations are tested initially to choose the one that gives the lowest objective function value.

**Output 102.2.7** Parameter Search for the Selected Direction  $\theta = 0^\circ$

| Semivariogram Analysis in Anisotropic Case With Trend Removal |        |       |       |                       |
|---------------------------------------------------------------|--------|-------|-------|-----------------------|
| The VARIOGRAM Procedure                                       |        |       |       |                       |
| Dependent Variable: ResidualOzone                             |        |       |       |                       |
| Angle: 0                                                      |        |       |       |                       |
| Current Model: Exponential                                    |        |       |       |                       |
| Parameter Search                                              |        |       |       |                       |
| Set                                                           | Nugget | Scale | Range | Objective<br>Function |
| 1                                                             | 0      | 2     | 5     | 391.06593             |
| 2                                                             | 0      | 2     | 15    | 1740.0                |
| 3                                                             | 0      | 2     | 25    | 5167.5                |
| 4                                                             | 0      | 2.5   | 5     | 64.86565              |
| 5                                                             | 0      | 2.5   | 15    | 664.03665             |
| 6                                                             | 0      | 2.5   | 25    | 2480.5                |
| 7                                                             | 0      | 3     | 5     | 72.86743              |
| 8                                                             | 0      | 3     | 15    | 305.53306             |
| 9                                                             | 0      | 3     | 25    | 1305.0                |

From this search, the combination of scale equal to 2.5 and a range of size 5 is passed as initial values to the model fitting process. This result is reflected in the model information table shown in [Output 102.2.8](#).

**Output 102.2.8** Model Initial Values for the Selected Direction  $\theta = 0^\circ$

| Model Information |                  |        |
|-------------------|------------------|--------|
| Parameter         | Initial<br>Value | Status |
| Nugget            | 0                | Fixed  |
| Scale             | 2.5000           |        |
| Range             | 5.0000           |        |

Fitting is successful, and among the output objects you can see the estimated parameters and the fit summary tables for the direction  $\theta = 0^\circ$  in [Output 102.2.9](#).

**Output 102.2.9** Weighted Least Squares Fitting Parameter Estimates and Summary for the Selected Direction  $\theta = 0^\circ$ 

| Parameter Estimates |                 |                     |          |         |                   |
|---------------------|-----------------|---------------------|----------|---------|-------------------|
| Parameter           | Estimate        | Approx<br>Std Error | DF       | t Value | Approx<br>Pr >  t |
| Scale               | 2.6657          | 0.03830             | 15       | 69.60   | <.0001            |
| Range               | 3.7277          | 0.5609              | 15       | 6.65    | <.0001            |
| Fit Summary         |                 |                     |          |         |                   |
| Model               | Weighted<br>SSE |                     | AIC      |         |                   |
| Exp                 | 43.35103        |                     | 19.91399 |         |                   |

A corresponding parameter search takes place for the direction  $\theta = 90^\circ$ . The respective table and the choice of initial values for fitting in the direction  $\theta = 90^\circ$  are shown in [Output 102.2.10](#).

**Output 102.2.10** Parameter Search and Model Initial Values for the Selected Direction  $\theta = 90^\circ$ 

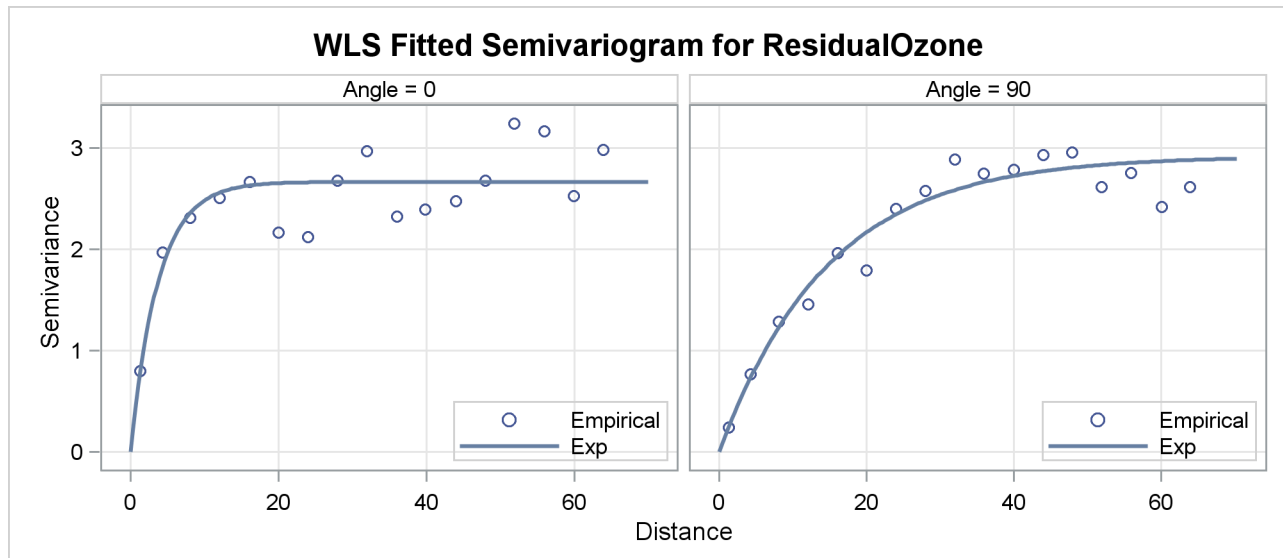
| Parameter Search  |                  |        |       |                       |
|-------------------|------------------|--------|-------|-----------------------|
| Set               | Nugget           | Scale  | Range | Objective<br>Function |
| 1                 | 0                | 2      | 5     | 302.54551             |
| 2                 | 0                | 2      | 15    | 635.93338             |
| 3                 | 0                | 2      | 25    | 1996.0                |
| 4                 | 0                | 2.5    | 5     | 95.09939              |
| 5                 | 0                | 2.5    | 15    | 104.56776             |
| 6                 | 0                | 2.5    | 25    | 662.06813             |
| 7                 | 0                | 3      | 5     | 155.50670             |
| 8                 | 0                | 3      | 15    | 20.48482              |
| 9                 | 0                | 3      | 25    | 190.30599             |
| Model Information |                  |        |       |                       |
| Parameter         | Initial<br>Value | Status |       |                       |
| Nugget            | 0                | Fixed  |       |                       |
| Scale             | 3.0000           |        |       |                       |
| Range             | 15.0000          |        |       |                       |

[Output 102.2.11](#) displays the estimated parameters and the fit summary for the direction  $\theta = 90^\circ$ .

**Output 102.2.11** Weighted Least Squares Fitting Parameter Estimates and Summary for the Selected Direction  $\theta = 90^\circ$ 

| Parameter Estimates |              |                  |         |         |                |
|---------------------|--------------|------------------|---------|---------|----------------|
| Parameter           | Estimate     | Approx Std Error | DF      | t Value | Approx Pr >  t |
| Scale               | 2.9199       | 0.07007          | 15      | 41.67   | <.0001         |
| Range               | 14.7576      | 0.9530           | 15      | 15.49   | <.0001         |
| Fit Summary         |              |                  |         |         |                |
| Model               | Weighted SSE |                  | AIC     |         |                |
| Exp                 | 19.12246     |                  | 6.00005 |         |                |

The fitted and empirical semivariograms for the selected directions are displayed in the panel of [Output 102.2.12](#).

**Output 102.2.12** Fitted Theoretical and Empirical Semivariogram for the Ozone Data in the  $\theta = 0^\circ$  and  $\theta = 90^\circ$  Directions

Conclusively, your semivariogram analysis on the detrended ozone data suggests that the ozone SRF exhibits anisotropy in the perpendicular directions of N–S ( $\theta = 0^\circ$ ) and E–W ( $\theta = 90^\circ$ ).

The sills in the two directions of anisotropy are similar in size. By inspecting again the empirical semivariograms in [Output 102.2.6](#), you could make the reasonable assumption that you have a case of geometric anisotropy, where the range in the major axis is about 4.5 times larger than the minor axis range. If you would like to use these PROC VARIOGRAM results for predictions, then you would need to specify a single scale value for the geometric anisotropy sill. In this case you could choose an arbitrary value for the constant scale from the narrow interval formed by the estimated scales in the previous results. For example, you can

specify the **PARMS** statement modified as shown in the following statement to approximate a common scale for the semivariance in all directions:

```
parms (0.) (2.7) (5 to 25 by 10) / hold=(1,2);
```

As an alternative, you can use PROC VARIOGRAM to fit an exponential model to all different angles examined in this example, and then select the constant scale value to be the mean of the scales across all directions.

---

### Example 102.3: 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 102.2: An Anisotropic Case Study with Surface Trend in the Data](#)” on page 8658.

As explained in the section “[Details: VARIOGRAM Procedure](#)” on page 8608, 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 provide no information about the spatial continuity of your SRF.

However, in the section “[Empirical Semivariograms and Surface Trends](#)” on page 8626 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 102.2: An Anisotropic Case Study with Surface Trend in the Data](#)” on page 8658. The spatial distribution of the data is shown in [Figure 102.2.1](#), and the pairwise distance distribution for NHCLASSES=35 is illustrated in [Figure 102.2.3](#). This exploratory analysis suggested a LAGDISTANCE=4 km, and [Figure 102.2.2](#) indicated that for this LAGDISTANCE= you can consider a value of MAXLAGS=16.

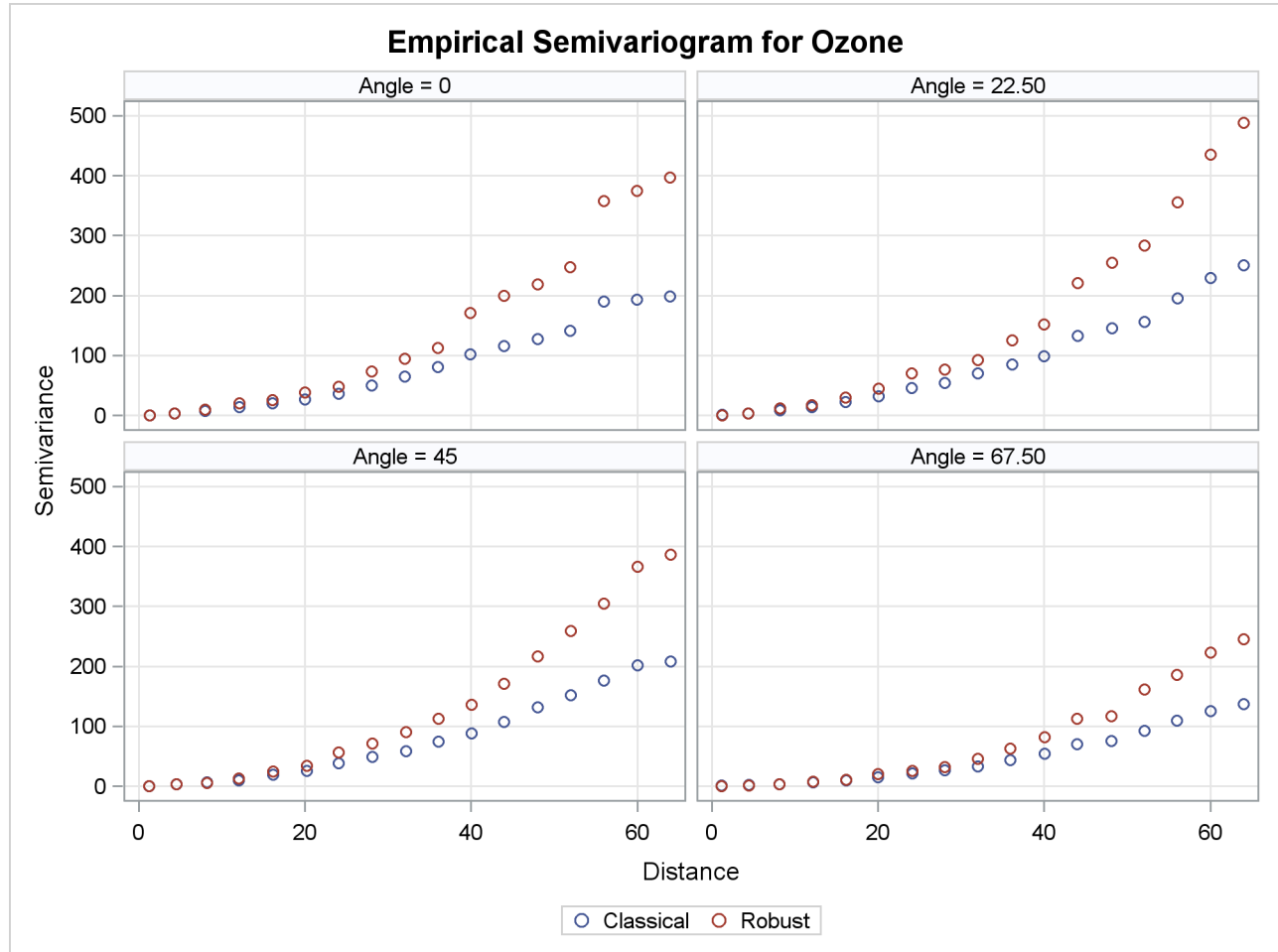
Recall from the section “[Empirical Semivariograms and Surface Trends](#)” on page 8626 that you need to investigate the empirical semivariogram of the data in a few different directions in order to identify a trend-free direction. If such a direction exists, then you can proceed with this special type of analysis. The following statements employ NDIRECTIONS=8 to examine eight directions:

```
ods graphics on;

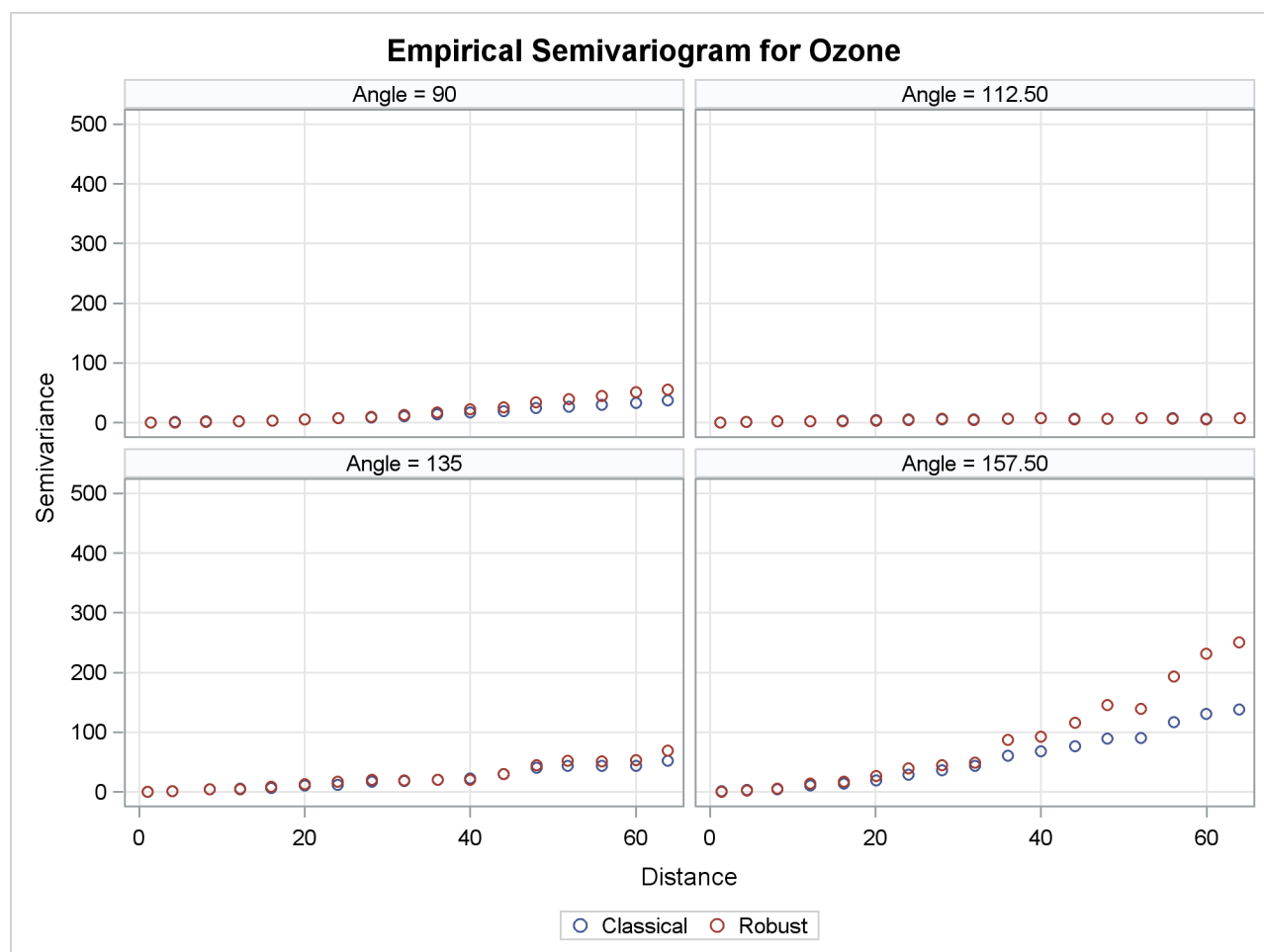
proc variogram data=ozoneSet plot(only)=semivar;
  compute lagd=4 maxlag=16 ndirections=8 robust;
  coord xc=East yc=North;
  var Ozone;
run;
```

By default, the range of  $180^\circ$  is divided into eight equally distanced angles:  $\theta = 0^\circ$ ,  $\theta = 22.5^\circ$ ,  $\theta = 45^\circ$ ,  $\theta = 67.5^\circ$ ,  $\theta = 90^\circ$ ,  $\theta = 112.5^\circ$ ,  $\theta = 135^\circ$ , and  $\theta = 157.5^\circ$ . The resulting empirical semivariograms for these angles are shown in [Output 102.3.1](#).

**Output 102.3.1** Ozone Empirical Semivariograms with  $0^\circ \leq \theta < 180^\circ$  and  $\delta\theta = 22.5^\circ$



Output 102.3.1 continued



The figures in [Output 102.3.1](#) suggest an overall continuing increase with distance of the semivariance in all directions. As explained in the section “[Theoretical Semivariogram Models](#)” on page 8608, this can be an indication of systematic trends in the data. However, the direction of  $\theta = 112.5^\circ$  clearly indicates 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.

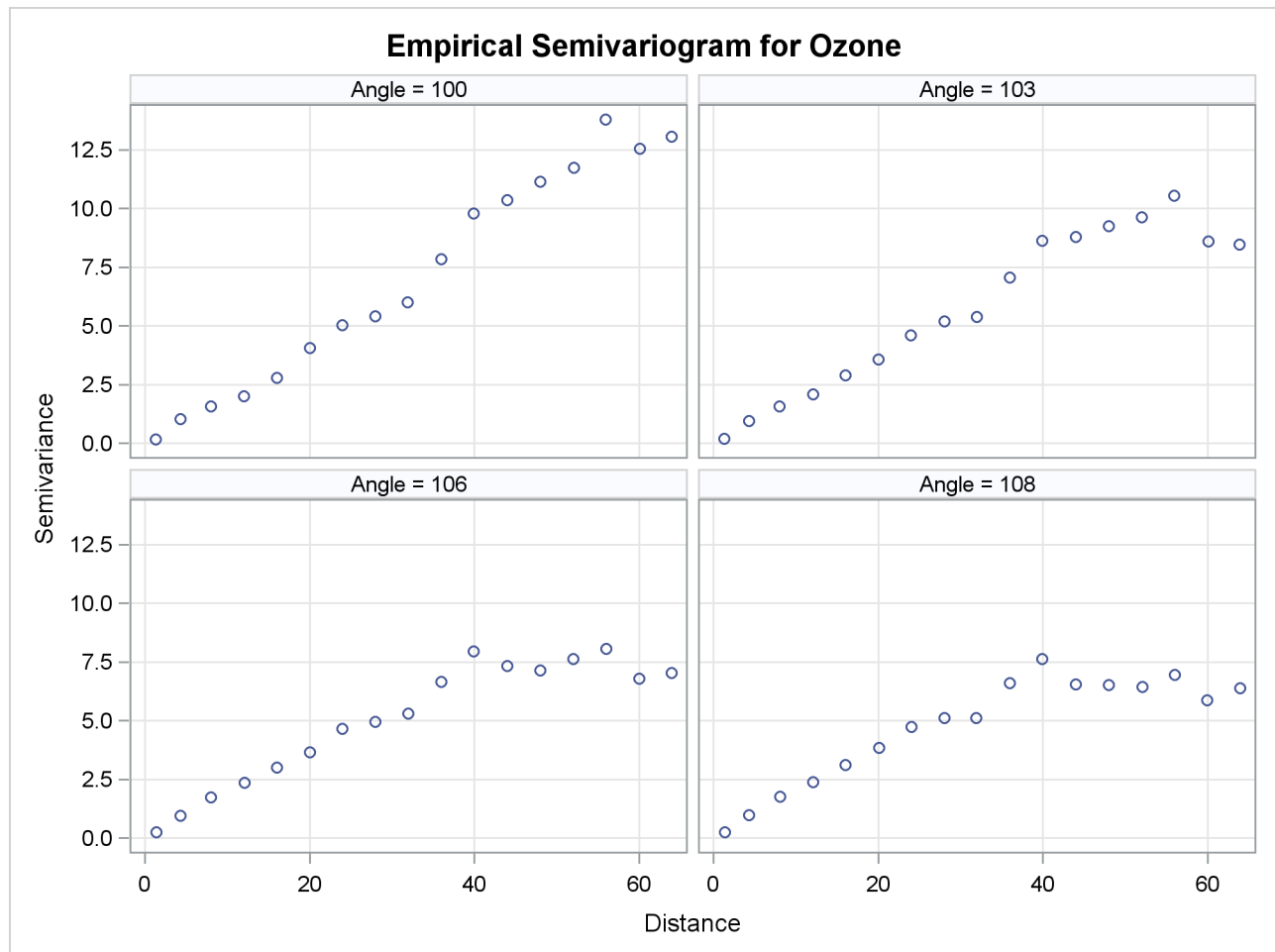
Run PROC VARIOGRAM again, specifying several directions within an interval of angles where you want to close in and you suspect the existence of a trend-free direction. In the following step you specify ANGLETOL=15°, which is smaller than the default value of 22.5°, and you also specify BANDWIDTH=10 km. The smaller values help with minimization of the interference with neighboring directions, as discussed in the section “[Angle Classification](#)” on page 8617.

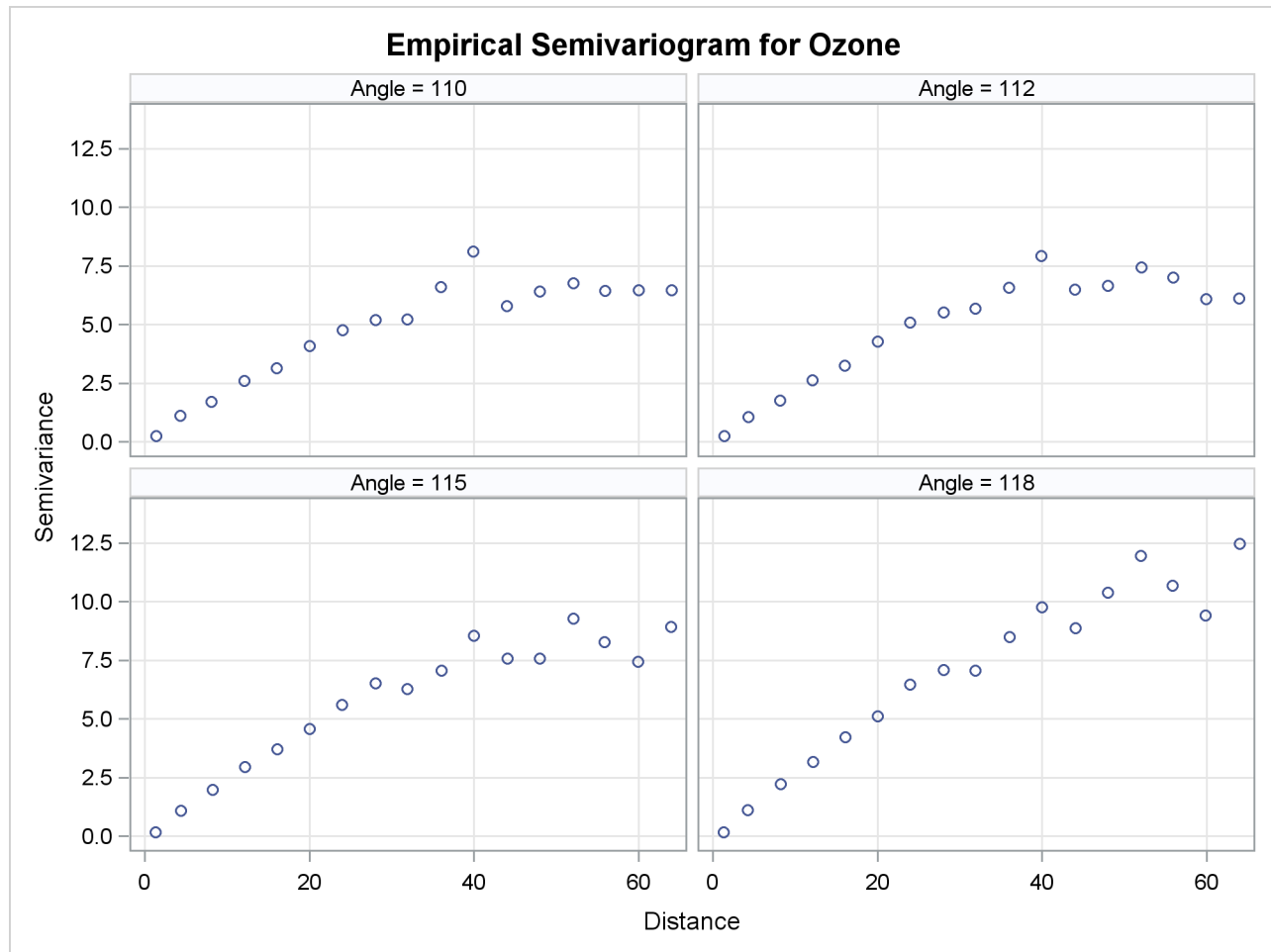
The aforementioned considerations are addressed in the following statements:

```
proc variogram data=ozoneSet plot(only)=semivar(c1a);
  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  $\theta = 100^\circ$  and  $\theta = 118^\circ$ . The pseudo-semivariograms in [Output 102.3.2](#) and [Output 102.3.3](#) 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 102.3.2** Ozone Empirical Semivariograms in  $100^\circ$ ,  $103^\circ$ ,  $106^\circ$ , and  $108^\circ$



**Output 102.3.3** Ozone Empirical Semivariograms in 110°, 112°, 115°, and 118°

You can 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  $\theta = 108^\circ$  to very likely be the trend-free one you are looking for.

From a physical standpoint, the trend-free direction, if it exists, is expected to be perpendicular to the direction of the maximum dip in the values of the ozone field, as mentioned in the section “[Empirical Semivariograms and Surface Trends](#)” on page 8626. If you cross-examine the ozone data distribution in [Output 102.2.1](#), 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  $\theta = 108^\circ$  and to perform a theoretical model fit.

The semivariance in [Output 102.3.2](#) 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 spherical model

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



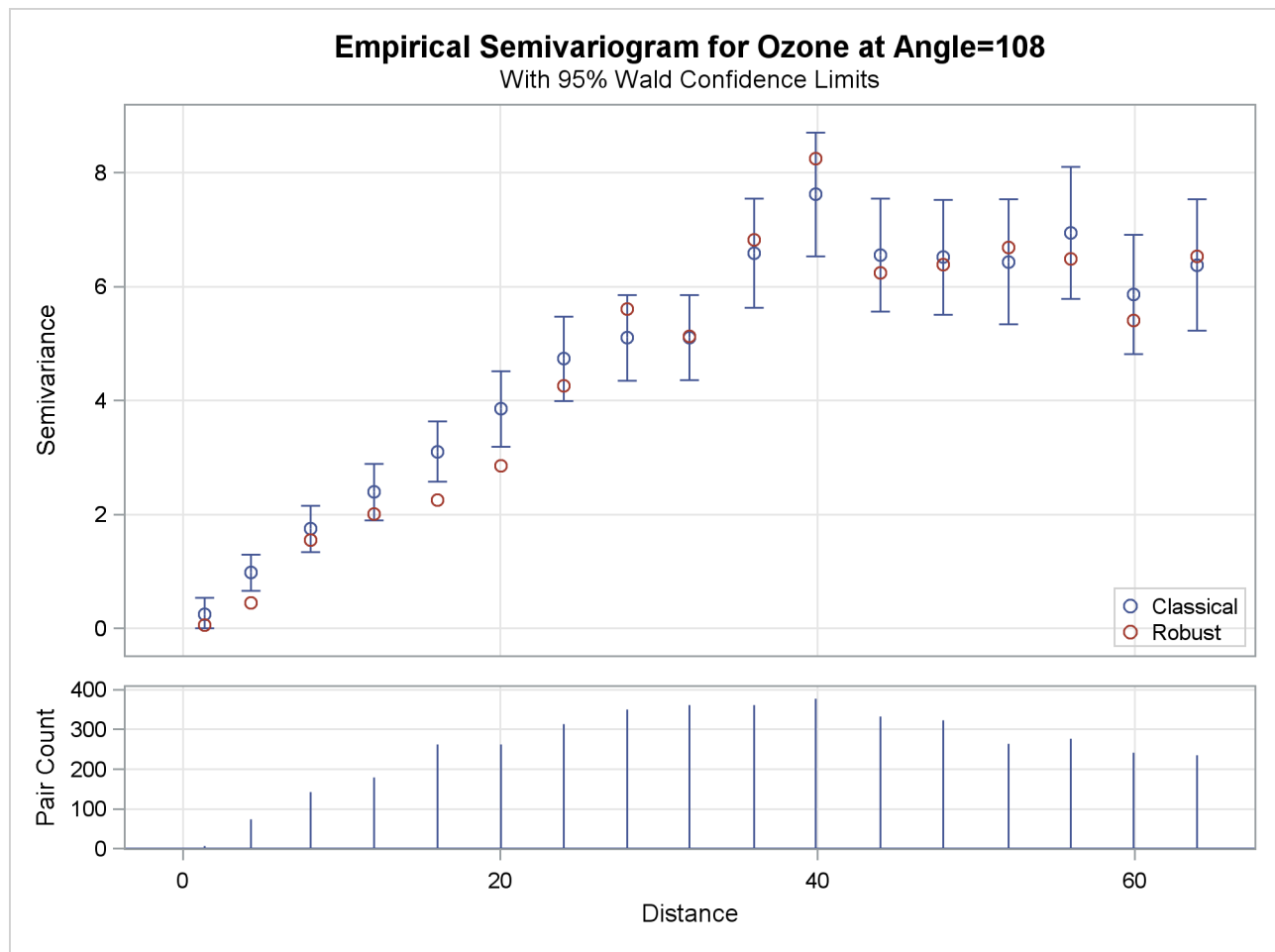
where  $\gamma_z(0) = 0$  and  $a_0 > 0$ . The empirical semivariograms also suggest that there does not seem to be a nugget effect. Assume that in this example you are interested in what the fitting process concludes about the nugget effect, so you skip the `NUGGET=` option in the `MODEL` statement. You also let PROC VARIOGRAM provide initial values for the rest of the model parameters. Eventually, you use the `PLOTS` option to inspect the classical and robust empirical semivariograms in the selected direction and to produce a plot of the fitted model. The following statements implement these considerations:

```
proc variogram data=ozoneSet plot(only)=(semivar fit);
  compute lagd=4 maxlag=16 robust cl;
  directions 108(15,10);
  coord xc=East yc=North;
  model form=sph;
  var ozone;
run;

ods graphics off;
```

The classical and robust empirical semivariograms in the selected direction  $\theta = 108^\circ$  are displayed in Figure 102.3.4.

**Output 102.3.4** Ozone Classical and Robust Empirical Semivariograms in  $\theta = 108^\circ$

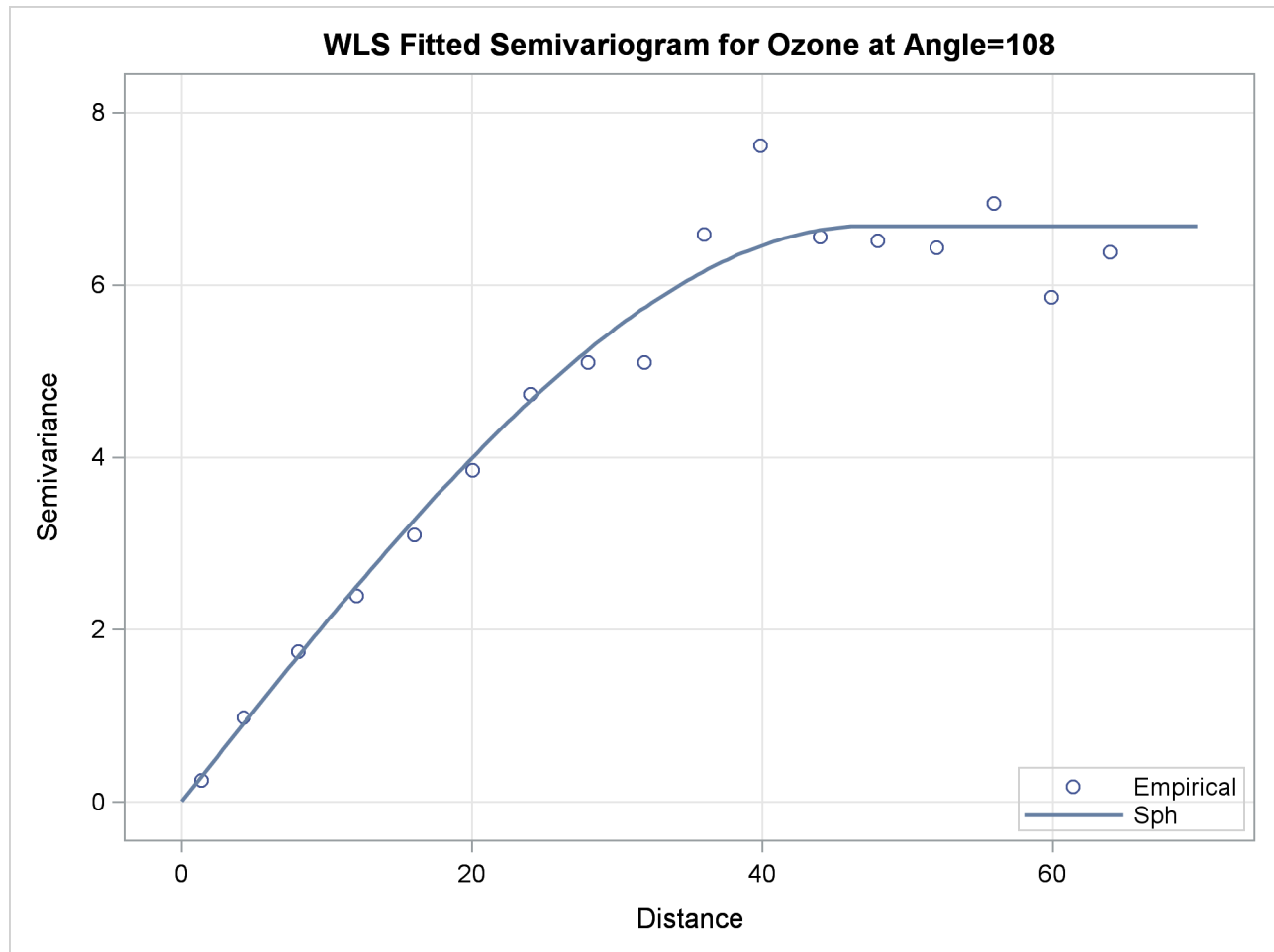


The output continues with information about the fitting process, which terminates successfully and produces the estimated parameters and the fit summary tables shown in [Output 102.3.5](#). The near-zero nugget parameter estimate indicates that you can consider the process to be practically free of nugget effect.

**Output 102.3.5** Weighted Least Squares Fitting Parameter Estimates and Summary in  $\theta = 108^\circ$

| Parameter Estimates |                 |                     |         |         |                   |
|---------------------|-----------------|---------------------|---------|---------|-------------------|
| Parameter           | Estimate        | Approx<br>Std Error | DF      | t Value | Approx<br>Pr >  t |
| Nugget              | 0.006260        | 0.09449             | 14      | 0.07    | 0.9481            |
| Scale               | 6.6791          | 0.1741              | 14      | 38.37   | <.0001            |
| Range               | 47.3012         | 2.0776              | 14      | 22.77   | <.0001            |
| Fit Summary         |                 |                     |         |         |                   |
| Model               | Weighted<br>SSE |                     | AIC     |         |                   |
| Sph                 | 13.13869        |                     | 1.61991 |         |                   |

The fitted and empirical semivariograms for the selected direction  $\theta = 108^\circ$  are displayed in [Output 102.3.6](#).

**Output 102.3.6** Fitted Theoretical and Empirical Semivariogram for the Ozone Data in  $\theta = 108^\circ$ 

A comparative look at the empirical and fitted semivariograms in [Output 102.3.6](#) and [Output 102.2.12](#) 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  $\theta = 108^\circ$  of the current scenario and the one shown in [Output 102.2.6](#) in the nearly identical direction  $\theta = 105^\circ$ . 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, the semivariogram analysis treats the trend-free ozone set and the original ozone measurements as different quantities. The process of detrending the original Ozone values is a transformation of these values into the trend-free values of ResidualOzone. Any existing spatial correlation in the original data is not necessarily retained within the transformed data. Depending on the transformation features, 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 8626, 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. However, 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 8662. 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 102.4: Covariogram and Semivariogram

The covariance that was reviewed in the section “[Stationarity](#)” on page 8615 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 8613, 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 8608. 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  $Z(s)$  in a square domain of  $100 \times 100$  ( $10^6$  km<sup>2</sup>). The following DATA step defines the data locations:

```
title 'Covariogram and Semivariogram';

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 85, “[The SIM2D Procedure](#).” The Gaussian SRF implies full knowledge of the SRF expected value  $E[Z(s)]$  and variance  $\text{Var}[Z(s)]$  at every location  $s$ . The following statements simulate an isotropic, second-order stationary SRF with constant expected value and variance throughout the simulation domain:

```
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  $E[Z(s)] = 30$  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  $\text{Var}[Z(s)] = c_n + \sigma_0^2 = 8$ . 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  $a_e = 3a_0 = 30$  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 sval;
run;
```

For each distance lag you take the average of the empirical measures over the number of simulations. PROC SORT prepares 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 creates 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;
```

The plot of the average empirical semivariance and covariance of the preceding analysis is shown in [Output 102.4.1](#). 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  $\text{Var}[Z(s)] = 8$  of the SRF.

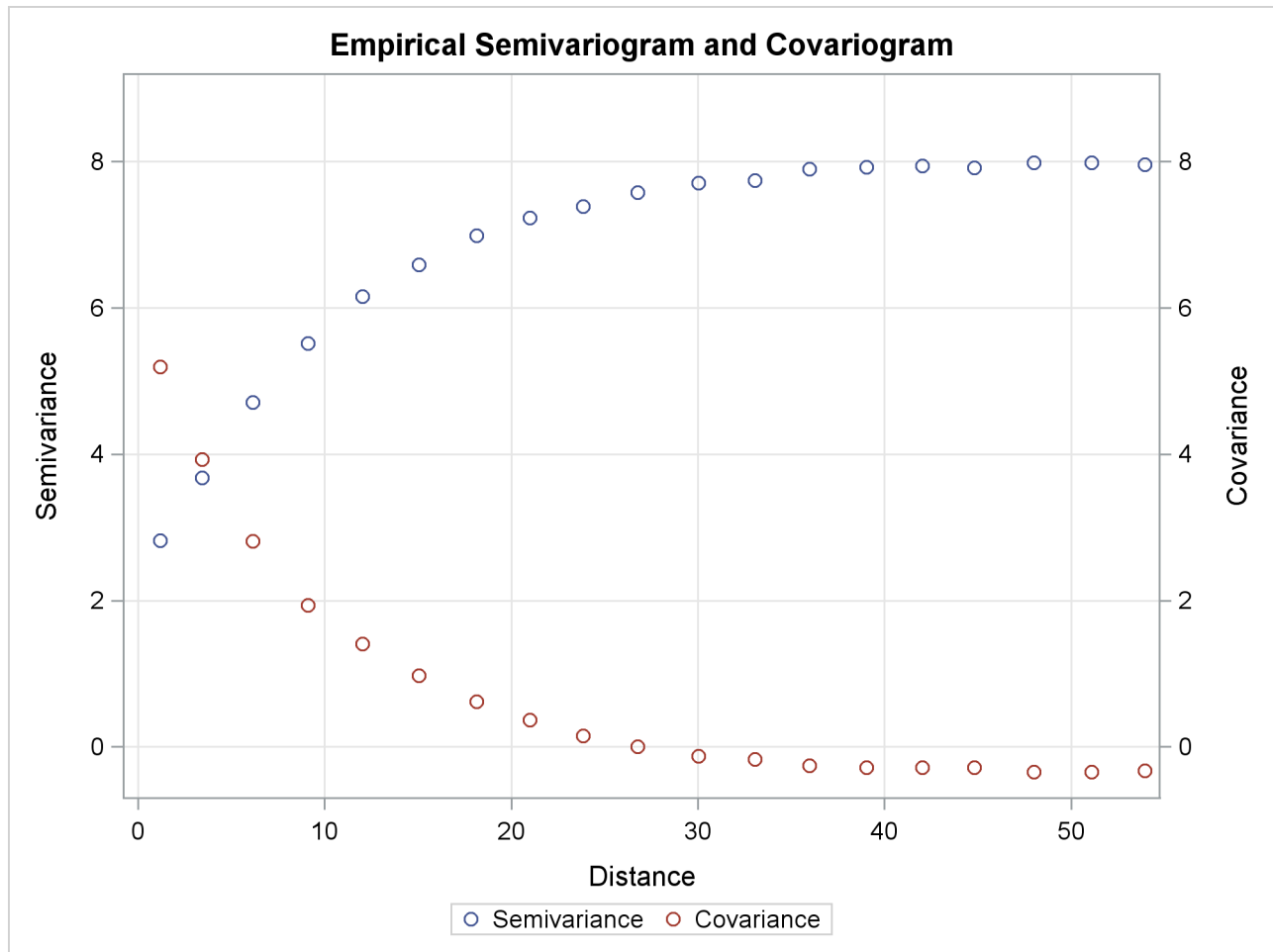
The simulated SRF is second-order stationary, so you expect at each lag the sum of the empirical semivariance and covariance to approximate the field variance  $\text{Var}[Z(s)]$ , as explained in the section “[Stationarity](#)” on page 8615. This behavior is evident in [Output 102.4.1](#).

This example concludes 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 102.4.1](#), where you expect to see that  $C(\mathbf{0}) = \text{Var}[Z(s)]$ , but the empirical covariogram cannot have a point at exactly  $\mathbf{h} = \mathbf{0}$ . A practical way to investigate for a nugget effect when you use empirical covariograms is as follows: 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);
run;
```

**Output 102.4.1** Average Empirical Semivariogram and Covariogram from 500 Simulations

**Output 102.4.2** 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 includes 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  $\mathbf{h} = \mathbf{0}$  to account for the corresponding nugget effect in the semivariogram.

**Output 102.4.2** 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 increases with distance  $h$ , as [Output 102.3.1](#) shows in “[Example 102.3: Analysis without Surface Trend Removal](#)” on page 8672. 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 8613 clearly show that the sample mean  $\bar{Z}$  and the SRF expected value  $E[Z(s)]$  are not important for the computation of the semivariance, but either one is necessary for the covariance. Hence, the semivariance expression filters the mean, and this behavior is especially useful when the mean is unknown. On the other hand, if  $E[Z(s)]$  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 102.5: A Box Plot of the Square Root Difference Cloud

The Gaussian form selected for the semivariogram in the section “[Getting Started: VARIOGRAM Procedure](#)” on page 8561 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  $Z(s)$ ,  $s \in \mathcal{R}^2$ , the square root difference cloud for a particular direction  $e$  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  $e$  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 8561, 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
;

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:

```
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;
```

The preceding statements produce [Output 102.5.1](#).

**Output 102.5.1** Summary of Results

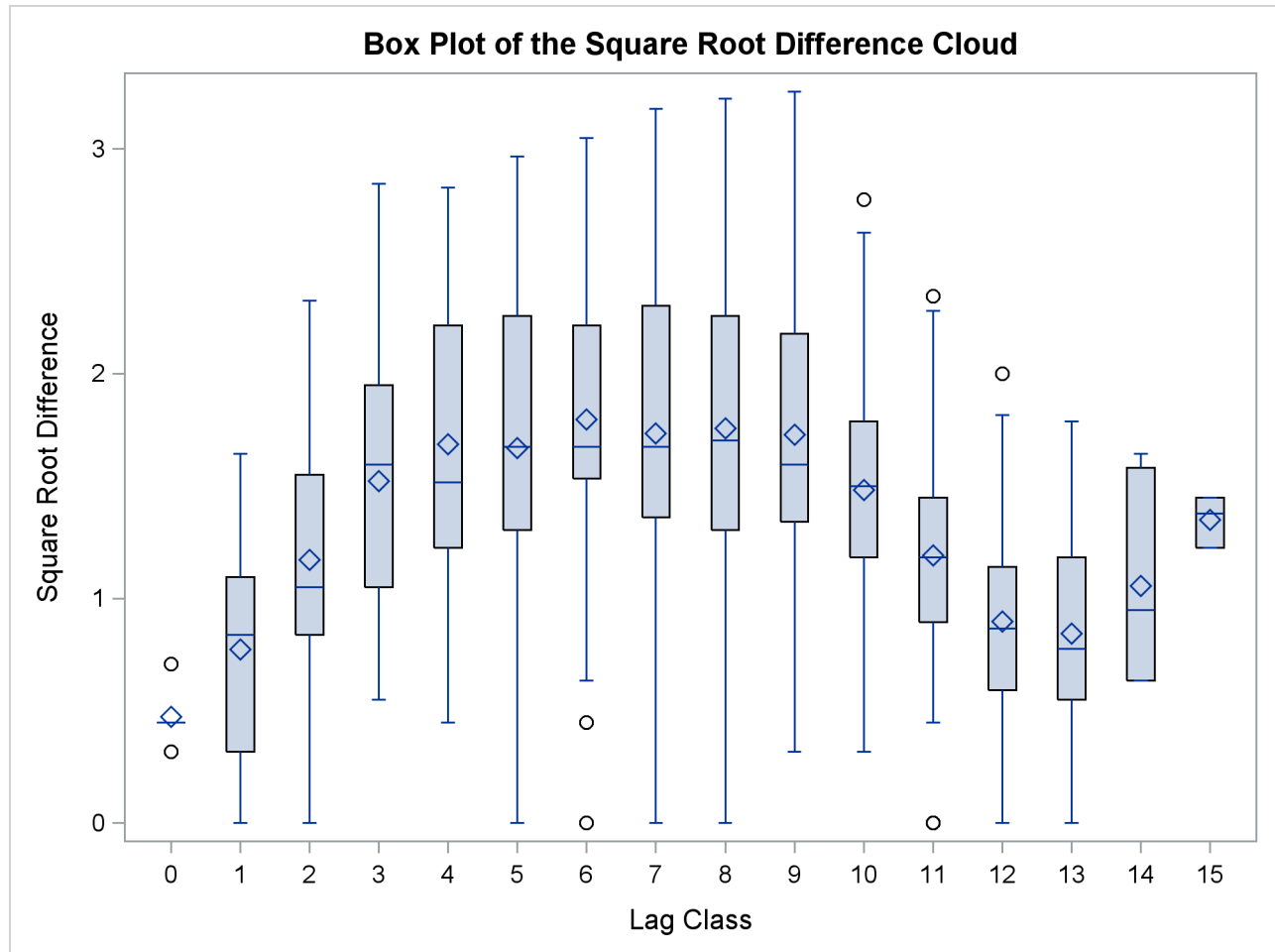
| Square Root Difference Cloud Example<br>Summary of Results |     |         |         |  |
|------------------------------------------------------------|-----|---------|---------|--|
| lag_                                                       | n   | mean    | std     |  |
| class                                                      |     |         |         |  |
| 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;
```

Output 102.5.2 suggests that outliers, if any, do not appear to be adversely affecting the empirical semivariogram in the N–S direction for the coal seam thickness data. The conclusion from Output 102.5.2 is consistent with our previous semivariogram analysis of the same data set in the section “Getting Started: VARIOGRAM Procedure” on page 8561. The effect of the isolated outliers in lag classes 6 and 10–12 in Output 102.5.2 is demonstrated as the divergence between the classical and robust empirical semivariance estimates in the higher distances in Output 102.7. The difference in these estimates comes from the definition of the robust semivariance estimator  $\hat{\gamma}_z(\mathbf{h})$  (see the section “Theoretical and Computational Details of the Semivariogram” on page 8613), which imposes a smoothing effect on the outlier influence.

**Output 102.5.2** Box Plot of the Square Root Difference Cloud



## References

- Anselin, L. (1996), "The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association," in M. Fischer, H. Scholten, and D. Unwin, eds., *Spatial Analytical Perspectives on GIS*, 111–125, London: Taylor and Francis.
- 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. (1985), "Fitting Variogram Models by Weighted Least Squares," *Mathematical Geology*, 17(5), 563–570.
- 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. E. (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 and 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," *Biometrika*, 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

- AIC, *see* fit criteria (VARIOGRAM)
- Akaike information criterion, *see* fit criteria (VARIOGRAM)
- angle
  - classes (VARIOGRAM), 8586, 8588, 8590, 8617, 8618
  - tolerance (VARIOGRAM), 8584, 8588, 8590, 8617, 8618
- anisotropy
  - geometric (VARIOGRAM), 8616
  - major axis (VARIOGRAM), 8616, 8668
  - minor axis (VARIOGRAM), 8616, 8668
  - VARIOGRAM procedure, 8588, 8616, 8627, 8658
  - zonal (VARIOGRAM), 8616
- autocorrelation
  - Geary's *c* coefficient (VARIOGRAM), 8560, 8585, 8637
  - Moran scatter plot (VARIOGRAM), 8570, 8577, 8639
  - Moran's *I* coefficient (VARIOGRAM), 8560, 8585, 8637
  - VARIOGRAM procedure, 8560, 8635
- autocorrelation weights
  - row-averaged (VARIOGRAM), 8570, 8636, 8639
  - standardized (VARIOGRAM), 8636
  - VARIOGRAM procedure, 8635
- bandwidth
  - VARIOGRAM procedure, 8586, 8590, 8620
- boundary constraints
  - VARIOGRAM procedure, 8605, 8607
- class, *see* angle classes (VARIOGRAM), *see* fit equivalence classes (VARIOGRAM), *see* lag classification (VARIOGRAM)
- collocation
  - VARIOGRAM procedure, 8616, 8638
- confidence level
  - VARIOGRAM procedure, 8584, 8592
- confidence limits
  - VARIOGRAM procedure, 8587, 8593
- constraints
  - boundary (VARIOGRAM), 8605, 8607
  - scale (VARIOGRAM), 8605
- correlation
  - length (VARIOGRAM), 8625
  - radius (VARIOGRAM), 8625
  - range (VARIOGRAM), 8625
- covariance
  - VARIOGRAM procedure, 8560, 8615, 8680
- covariogram
  - VARIOGRAM procedure, 8680
- cubic semivariance model
  - VARIOGRAM procedure, 8594, 8611
- distance, *see* lag (VARIOGRAM)
  - classification (VARIOGRAM), 8619
- equivalence class, *see* fit equivalence classes (VARIOGRAM)
- ergodicity
  - VARIOGRAM procedure, 8615
- estimation
  - VARIOGRAM procedure, 8561
- exploratory data analysis
  - VARIOGRAM procedure, 8562
- exponential semivariance model
  - VARIOGRAM procedure, 8594, 8611
- fit, *see also* semivariogram theoretical model fitting (VARIOGRAM)
  - automated (VARIOGRAM), 8627
  - criteria (VARIOGRAM), 8632
  - equivalence classes (VARIOGRAM), 8634
  - quality (VARIOGRAM), 8632
- Gaussian semivariance model
  - VARIOGRAM procedure, 8594, 8611
- Geary's *c* coefficient, *see* autocorrelation
- initial values
  - VARIOGRAM procedure, 8603, 8630
- isotropy, *see* anisotropy (VARIOGRAM)
  - VARIOGRAM procedure, 8564, 8588, 8616
- kriging
  - ordinary kriging (VARIOGRAM), 8561
- lag
  - classification (VARIOGRAM), 8565, 8619
  - count (VARIOGRAM), 8565, 8621
  - distance (VARIOGRAM), 8564, 8587, 8621, 8624
  - number of point pairs in (VARIOGRAM), 8623
  - pairwise distance (VARIOGRAM), 8564
  - tolerance (VARIOGRAM), 8587, 8620
  - VARIOGRAM procedure, 8564

Matérn, *see also* semivariogram theoretical models (VARIOGRAM)  
 model fitting (VARIOGRAM), 8634

Matérn semivariance model  
 VARIOGRAM procedure, 8594, 8611

mean trend, *see* surface trend (VARIOGRAM)

measures of spatial continuity  
 VARIOGRAM procedure, 8560, 8613, 8680

model fitting, *see* semivariogram theoretical model fitting (VARIOGRAM)  
 VARIOGRAM procedure, 8648

modeling, *see* semivariogram theoretical model fitting (VARIOGRAM)

Moran scatter plot, *see* autocorrelation

Moran's *I* coefficient, *see* autocorrelation

nested models  
 VARIOGRAM procedure, 8613

nonrandom trend, *see* surface trend

normality assumption  
 VARIOGRAM procedure, 8585, 8637

nugget effect  
 VARIOGRAM procedure, 8598, 8610, 8613

ODS graph names  
 VARIOGRAM procedure, 8647

ODS Graphics  
 VARIOGRAM procedure, 8578

ODS table names  
 VARIOGRAM procedure, 8646

output data sets  
 VARIOGRAM procedure, 8577, 8578, 8623, 8640–8642

pairwise distance, *see also* lag classification (VARIOGRAM)  
 distribution (VARIOGRAM), 8621  
 VARIOGRAM procedure, 8565

panels (VARIOGRAM procedure), *see* plots (VARIOGRAM procedure)

pentaspherical semivariance model  
 VARIOGRAM procedure, 8594, 8611

plots (VARIOGRAM procedure)  
 Fit, 8647  
 Fit panel, 8647  
 Moran scatter plot, 8647  
 Observations, 8647  
 Pairs, 8647  
 Pairwise distance distribution, 8647  
 panels, 8579, 8665, 8673, 8675  
 Semivariogram, 8647  
 Semivariogram panel, 8647

point pairs  
 VARIOGRAM procedure, 8564, 8614, 8617

power semivariance model

VARIOGRAM procedure, 8594, 8611

prediction  
 VARIOGRAM procedure, 8561

preliminary data analysis, *see* exploratory data analysis

pseudo-semivariance  
 VARIOGRAM procedure, 8614

pseudo-semivariogram  
 VARIOGRAM procedure, 8614, 8672

randomization assumption  
 VARIOGRAM procedure, 8585, 8637

range  
 effective (VARIOGRAM), 8610  
 practical (VARIOGRAM), 8610  
 VARIOGRAM procedure, 8609

scale constraints  
 VARIOGRAM procedure, 8605

semivariance, *see also* semivariogram  
 classical (VARIOGRAM), 8567, 8614  
 computation (VARIOGRAM), 8625  
 empirical (VARIOGRAM), 8614  
 robust (VARIOGRAM), 8567, 8589, 8614  
 theoretical models, 8611  
 variance (VARIOGRAM), 8614  
 VARIOGRAM procedure, 8560, 8613, 8680

semivariogram  
 analysis (VARIOGRAM), 8562  
 and covariogram (VARIOGRAM), 8680  
 computation (VARIOGRAM), 8562  
 empirical (VARIOGRAM), 8562, 8614  
 parameters (VARIOGRAM), 8609  
 robust (VARIOGRAM), 8614  
 theoretical model fitting, 8571, 8590, 8603, 8627, 8668  
 theoretical models (VARIOGRAM), 8562, 8564, 8608  
 VARIOGRAM procedure, 8560, 8613

sill  
 VARIOGRAM procedure, 8609

sine hole effect semivariance model  
 VARIOGRAM procedure, 8594, 8611

spatial continuity  
 VARIOGRAM procedure, 8560, 8561, 8613, 8614, 8672, 8680

spatial dependence, *see* spatial continuity

spatial lag, *see also* autocorrelation Moran scatter plot (VARIOGRAM)  
 VARIOGRAM procedure, 8639

spatial prediction  
 VARIOGRAM procedure, 8561, 8614

spatial random field  
 autocorrelation (VARIOGRAM), 8635

- ergodicity (VARIOGRAM), 8615
- intrinsically stationary (VARIOGRAM), 8615
- isotropic (VARIOGRAM), 8616, 8627
- realization (VARIOGRAM), 8608
- spatial continuity (VARIOGRAM), 8613, 8672
- VARIOGRAM procedure, 8608
- spatial structure, *see* spatial continuity
- spherical semivariance model
  - VARIOGRAM procedure, 8594, 8611
- square root difference cloud
  - VARIOGRAM procedure, 8684
- SRF, *see* spatial random field
- SSE, *see* fit criteria (VARIOGRAM)
- standard errors
  - VARIOGRAM procedure, 8561
- stationarity
  - intrinsic (VARIOGRAM), 8615
  - second-order (VARIOGRAM), 8615, 8682
  - VARIOGRAM procedure, 8609
- stochastic analysis
  - VARIOGRAM procedure, 8561
- stochastic modeling, *see* semivariogram theoretical
  - model fitting (VARIOGRAM), *see* spatial prediction (VARIOGRAM)
- stochastic spatial prediction, *see* spatial prediction (VARIOGRAM)
- surface trend
  - VARIOGRAM procedure, 8560, 8564, 8614, 8626, 8658, 8662
- systematic trend, *see* surface trend
- tables (VARIOGRAM procedure)
  - Approximate Correlation Matrix, 8646
  - Approximate Covariance Matrix, 8646
  - Autocorrelation Statistics, 8570, 8644, 8646
  - Convergence Status, 8646
  - Empirical Semivariogram, 8568, 8644, 8646
  - Fit Summary, 8646
  - Fitting General Information, 8646
  - Iteration History, 8646
  - Lagrange Multipliers, 8646
  - Model Information, 8646
  - Number of Observations, 8644
  - Optimization Information, 8646
  - Optimization Input Options, 8646
  - Optimization Results, 8646
  - Pairs Information, 8566, 8623–8625, 8644, 8646
  - Pairwise Distance Intervals, 8565, 8623–8625, 8644, 8646, 8660, 8668
  - Parameter Estimates, 8646
  - Parameter Estimates Results, 8646
  - Parameter Search, 8646
  - Problem Description, 8646
  - PROC VARIOGRAM statements, 8576
  - Projected Gradient, 8646
  - Starting Parameter Estimates, 8646
- theoretical semivariogram models, *see* semivariogram (VARIOGRAM)
- tolerance, *see* angle tolerance (VARIOGRAM), *see* lag tolerance (VARIOGRAM)
- trend, *see* surface trend (VARIOGRAM)
- uncertainty
  - VARIOGRAM procedure, 8561
- VARIOGRAM procedure
  - angle classes, 8586, 8588, 8590, 8617, 8618
  - angle tolerance, 8584, 8588, 8590, 8617, 8618
  - anisotropy, 8588, 8658
  - autocorrelation, 8560, 8635
  - autocorrelation weights, 8635
  - bandwidth, 8586, 8590, 8620
  - boundary constraints, 8605, 8607
  - collocation, 8616, 8638
  - confidence level, 8584, 8592
  - confidence limits, 8587, 8593
  - correlation measures, 8625
  - covariance, 8560, 8615, 8680
  - covariance matrix, 8601
  - cubic semivariance model, 8594, 8611
  - cutoff distance, 8589
  - DATA= data set, 8576
  - distance classes, 8621, 8623, 8624
  - distance classification, 8619
  - distance interval, 8565
  - ergodicity, 8615
  - estimation, 8561
  - examples, 8562, 8648, 8658, 8672, 8680, 8684
  - exploratory data analysis, 8562
  - exponential semivariance model, 8594, 8611
  - Gaussian semivariance model, 8594, 8611
  - Geary's *c* coefficient, 8560, 8585, 8637
  - grid search, 8603
  - initial values, 8603, 8630
  - input data set, 8576
  - isotropy, 8588
  - lag, 8564, 8621, 8623
  - lag distance, 8587, 8621
  - lag tolerance, 8587
  - Matérn semivariance model, 8594, 8611
  - measures of spatial continuity, 8560, 8613, 8680
  - model fitting, 8571, 8590, 8603, 8627, 8648, 8668
  - Moran scatter plot, 8570, 8577, 8639
  - Moran's *I* coefficient, 8560, 8585, 8637
  - nested models, 8613
  - normality assumption, 8585, 8637
  - nugget effect, 8598, 8610, 8613



- ODS graph names, 8647
- ODS Graphics, 8578
- ODS table names, 8646
- ordinary kriging, 8561
- OUTACWEIGHTS= data set, 8577
- OUTDIST= data set, 8577, 8623
- OUTMORAN= data set, 8577
- OUTPAIR= data set, 8578
- output data sets, 8577, 8578, 8640–8642
- OUTVAR= data set, 8578
- pairwise distance, 8565, 8587
- panel plots, 8578
- pentaspherical semivariance model, 8594, 8611
- point pairs, 8564, 8614, 8617
- power semivariance model, 8594, 8611
- prediction, 8561
- pseudo-semivariance, 8614
- pseudo-semivariogram, 8614, 8672
- randomization assumption, 8585, 8637
- scale constraints, 8605
- semivariance, 8560, 8613, 8680
- semivariance computation, 8625
- semivariance, classical, 8567, 8614
- semivariance, empirical, 8614
- semivariance, robust, 8567, 8589, 8614
- semivariance, variance, 8614
- semivariogram, 8560, 8613
- semivariogram analysis, 8562
- semivariogram and covariogram, 8680
- semivariogram computation, 8562
- semivariogram effective range, 8610
- semivariogram parameters, 8609
- semivariogram range, 8609
- semivariogram sill, 8609
- semivariogram, empirical, 8562, 8614
- semivariogram, robust, 8614
- sine hole effect semivariance model, 8594, 8611
- spatial continuity, 8560, 8561, 8613, 8614, 8672, 8680
- spatial lag, 8639
- spatial prediction, 8561, 8614
- spatial random field, 8608, 8613, 8615, 8627, 8635, 8672
- spherical semivariance model, 8594, 8611
- square root difference cloud, 8684
- standard errors, 8561
- stationarity, 8609, 8615, 8682
- stochastic analysis, 8561
- surface trend, 8560, 8564, 8614, 8626, 8658, 8662
- theoretical semivariogram models, 8562, 8564, 8571, 8608, 8611, 8668
- uncertainty, 8561
- weighted average, 8570, 8639

- VARIOGRAM procedure, plots
  - Fit, 8647
  - Fit panel, 8647
  - Moran scatter plot, 8647
  - Observations, 8647
  - Pairs, 8647
  - Pairwise distance distribution, 8647
  - Semivariogram, 8647
  - Semivariogram panel, 8647
- VARIOGRAM procedure, tables
  - Approximate Correlation Matrix, 8646
  - Approximate Covariance Matrix, 8646
  - Autocorrelation Statistics, 8570, 8644, 8646
  - Convergence Status, 8646
  - Empirical Semivariogram, 8568, 8644, 8646
  - Fit Summary, 8646
  - Fitting General Information, 8646
  - Iteration History, 8646
  - Lagrange Multipliers, 8646
  - Model Information, 8646
  - Number of Observations, 8644
  - Optimization Information, 8646
  - Optimization Input Options, 8646
  - Optimization Results, 8646
  - Pairs Information, 8566, 8623–8625, 8644, 8646
  - Pairwise Distance Intervals, 8565, 8623–8625, 8644, 8646, 8660, 8668
  - Parameter Estimates, 8646
  - Parameter Estimates Results, 8646
  - Parameter Search, 8646
  - Problem Description, 8646
  - PROC VARIOGRAM statements, 8576
  - Projected Gradient, 8646
  - Starting Parameter Estimates, 8646
- weighted average
  - VARIOGRAM procedure, 8570, 8639
- weighted least squares, *see* fit criteria (VARIOGRAM)
- WSSE, *see* fit criteria (VARIOGRAM)



# Syntax Index

- ALPHA= option
  - COMPUTE statement (VARIogram), [8584](#)
  - MODEL statement (VARIogram), [8592](#)
- ANGLETOLERANCE= option
  - COMPUTE statement (VARIogram), [8584](#)
- AUTOCORRELATION option
  - VARIogram procedure, COMPUTE statement, [8585](#)
- AUTOCORRELATION STATISTICS= option
  - VARIogram procedure, COMPUTE statement, [8585](#)
- BANDWIDTH= option
  - COMPUTE statement (VARIogram), [8586](#)
- BY statement
  - VARIogram procedure, [8583](#)
- CHOOSE= option
  - MODEL statement (VARIogram), [8592](#)
- CL option
  - COMPUTE statement (VARIogram), [8587](#)
  - MODEL statement (VARIogram), [8593](#)
- COMPUTE statement
  - VARIogram procedure, [8584](#)
- COORDINATES statement
  - VARIogram procedure, [8589](#)
- CORRB option
  - MODEL statement (VARIogram), [8601](#)
- COVB option
  - MODEL statement (VARIogram), [8601](#)
- DATA= option
  - PROC VARIogram statement, [8576](#)
- DEPSILON= option
  - COMPUTE statement (VARIogram), [8587](#)
- DETAILS option
  - MODEL statement (VARIogram), [8602](#)
- DIRECTIONS statement
  - VARIogram procedure, [8590](#)
- EQCONS= option
  - PARMS statement (VARIogram), [8605](#)
- EQUIVTOL= option
  - MODEL statement (VARIogram), [8593](#)
- FIT option
  - MODEL statement (VARIogram), [8593](#)
- FORM= option
  - MODEL statement (VARIogram), [8593](#)
- GRADIENT option
  - MODEL statement (VARIogram), [8602](#)
- HOLD= option
  - PARMS statement (VARIogram), [8605](#)
- ID statement
  - VARIogram procedure, [8590](#)
- IDGLOBAL option
  - PROC VARIogram statement, [8577](#)
- IDNUM option
  - PROC VARIogram statement, [8577](#)
- LABEL= option
  - STORE statement (VARIogram), [8608](#)
- LAGDISTANCE= option
  - COMPUTE statement (VARIogram), [8587](#)
- LAGTOLERANCE= option
  - COMPUTE statement (VARIogram), [8587](#)
- LOWERB= option
  - PARMS statement (VARIogram), [8605](#)
- MAXLAGS= option
  - COMPUTE statement (VARIogram), [8588](#)
- MAXSCALE= option
  - PARMS statement (VARIogram), [8605](#)
- MDATA= option
  - MODEL statement (VARIogram), [8596](#)
- MODEL statement
  - VARIogram procedure, [8590](#)
- MTOGTOL= option
  - MODEL statement (VARIogram), [8602](#)
- NDIRECTIONS= option
  - COMPUTE statement (VARIogram), [8588](#)
- NEPSILON= option
  - MODEL statement (VARIogram), [8598](#)
- NHCLASSES= option
  - COMPUTE statement (VARIogram), [8588](#)
- NLOPTIONS statement
  - VARIogram procedure, [8607](#)
- NOBOUND option
  - PARMS statement (VARIogram), [8605](#)
- NOFIT option
  - MODEL statement (VARIogram), [8602](#)
- NOITPRINT option
  - MODEL statement (VARIogram), [8602](#)
- NOPRINT option
  - PROC VARIogram statement, [8577](#)

NOVARIOGRAM option  
 COMPUTE statement (VARIOGRAM), [8589](#)

NUGGET= option  
 MODEL statement (VARIOGRAM), [8598](#)

OUTACWEIGHTS= option  
 PROC VARIOGRAM statement, [8577](#)

OUTDISTANCE= option  
 PROC VARIOGRAM statement, [8577](#)

OUTMORAN= option  
 PROC VARIOGRAM statement, [8577](#)

OUTPAIR= option  
 PROC VARIOGRAM statement, [8578](#)

OUTPDISTANCE= option  
 COMPUTE statement (VARIOGRAM), [8589](#)

output data sets  
 VARIOGRAM procedure, [8560](#)

OUTVAR= option  
 PROC VARIOGRAM statement, [8578](#)

PARMS statement  
 VARIOGRAM procedure, [8603](#)

PARMSDATA= option  
 PARMS statement (VARIOGRAM), [8606](#)

PDATA= option  
 PARMS statement (VARIOGRAM), [8606](#)

PLOTS option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8578](#)

PLOTS(ONLY) option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8579](#)

PLOTS(UNPACKPANEL) option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8579](#)

PLOTS=ALL option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8579](#)

PLOTS=EQUATE option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8579](#)

PLOTS=FITPLOT option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8579](#)

PLOTS=MORAN option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8580](#)

PLOTS=NONE option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8581](#)

PLOTS=OBSERVATIONS option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8581](#)

PLOTS=PAIRS option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8582](#)

PLOTS=SEMIVARIOGRAM option  
 VARIOGRAM procedure, PROC VARIOGRAM statement, [8583](#)

PROC VARIOGRAM statement, *see* VARIOGRAM procedure

RANGE= option  
 MODEL statement (VARIOGRAM), [8598](#)

RANGLAG= option  
 MODEL statement (VARIOGRAM), [8599](#)

RANKEPS= option  
 MODEL statement (VARIOGRAM), [8600](#)

ROBUST option  
 COMPUTE statement (VARIOGRAM), [8589](#)

SCALE= option  
 MODEL statement (VARIOGRAM), [8600](#)

SMOOTH= option  
 MODEL statement (VARIOGRAM), [8601](#)

STORE statement  
 VARIOGRAM procedure, [8607](#)

UPPERB= option  
 PARMS statement (VARIOGRAM), [8607](#)

VAR statement  
 VARIOGRAM procedure, [8608](#)

VARIOGRAM procedure, [8560](#)  
 output data sets, [8560](#)  
 syntax, [8576](#)

VARIOGRAM procedure, BY statement, [8583](#)

VARIOGRAM procedure, COMPUTE statement, [8584](#)  
 ALPHA= option, [8584](#)  
 ANGLETOLERANCE= option, [8584](#)  
 AUTOCORRELATION option, [8585](#)  
 AUTOCORRELATION STATISTICS= option, [8585](#)  
 BANDWIDTH= option, [8586](#)  
 CL option, [8587](#)  
 DEPSILON= option, [8587](#)  
 LAGDISTANCE= option, [8587](#)  
 LAGTOLERANCE= option, [8587](#)  
 MAXLAGS= option, [8588](#)  
 NDIRECTIONS= option, [8588](#)  
 NHCLASSES= option, [8588](#)  
 NOVARIOGRAM option, [8589](#)  
 OUTPDISTANCE= option, [8589](#)  
 ROBUST option, [8589](#)

VARIOGRAM procedure, COORDINATES statement, [8589](#)  
 XCOORD= option, [8589](#)  
 YCOORD= option, [8589](#)

VARIOGRAM procedure, DIRECTIONS statement,  
8590

VARIOGRAM procedure, ID statement, 8590

VARIOGRAM procedure, MODEL statement, 8590

ALPHA= option, 8592

CHOOSE= option, 8592

CL option, 8593

CORRB option, 8601

COVB option, 8601

DETAILS option, 8602

EQUIVTOL= option, 8593

FIT option, 8593

FORM= option, 8593

GRADIENT option, 8602

MDATA= option, 8596

MTOGTOL= option, 8602

NEPSILON= option, 8598

NOFIT option, 8602

NOITPRINT option, 8602

NUGGET= option, 8598

RANGE= option, 8598

RANGELAG= option, 8599

RANKEPS= option, 8600

SCALE= option, 8600

SMOOTH= option, 8601

VARIOGRAM procedure, NLOPTIONS statement,  
8607

VARIOGRAM procedure, PARMS statement, 8603

EQCONS= option, 8605

HOLD= option, 8605

LOWERB= option, 8605

MAXSCALE= option, 8605

NOBOUND option, 8605

PARMSDATA= option, 8606

PDATA= option, 8606

UPPERB= option, 8607

VARIOGRAM procedure, PROC VARIOGRAM  
statement, 8576

DATA= option, 8576

IDGLOBAL option, 8577

IDNUM option, 8577

NOPRINT option, 8577

OUTACWEIGHTS= option, 8577

OUTDISTANCE= option, 8577

OUTMORAN= option, 8577

OUTPAIR= option, 8578

OUTVAR= option, 8578

PLOTS option, 8578

PLOTS(ONLY) option, 8579

PLOTS(UNPACKPANEL) option, 8579

PLOTS=ALL option, 8579

PLOTS=EQUATE option, 8579

PLOTS=FITPLOT option, 8579

PLOTS=MORAN options, 8580

PLOTS=NONE option, 8581

PLOTS=OBSERVATIONS option, 8581

PLOTS=PAIRS option, 8582

PLOTS=SEMIVARIOGRAM option, 8583

VARIOGRAM procedure, STORE statement, 8607

LABEL= options, 8608

VARIOGRAM procedure, VAR statement, 8608

XCOORD=option

COORDINATES statement (VARIOGRAM),  
8589

YCOORD=option

COORDINATES statement (VARIOGRAM),  
8589



## Your Turn

---

We welcome your feedback.

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



# SAS® Publishing Delivers!

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

## SAS® Press

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

**[support.sas.com/saspress](http://support.sas.com/saspress)**

## SAS® Documentation

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

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

**[support.sas.com/publishing](http://support.sas.com/publishing)**

## SAS® Publishing News

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

**[support.sas.com/spn](http://support.sas.com/spn)**



**THE  
POWER  
TO KNOW®**

