Paper 337-2010

Everything in Its Place: Efficient Geostatistical Analysis with SAS/STAT® Spatial Procedures

Alexander Kolovos, SAS Institute Inc., Cary NC

ABSTRACT

Got coordinates in your data? Feeling lost in space? Spatial data arise in a wealth of scientific fields; their stochastic structure poses unique analytical challenges that are different from most modeling and prediction methodologies.

SAS/STAT® 9.22 introduces exciting updates to the SAS suite of spatial procedures to augment, simplify, and streamline spatial analysis in one and two dimensions. All spatial procedures offer a wider variety of correlation models than before. Automated semivariogram modeling in the VARIOGRAM procedure reduces the time and effort required to select a suitable correlation structure. Item stores enable you to pass the semivariance information seamlessly to the KRIGE2D procedure for prediction or to the SIM2D procedure for simulation. Option-rich ODS Graphics plots give you the full picture for diagnosing your spatial data and visualizing your prediction and simulation results.

The presentation demonstrates these new features and discusses key elements and underlying complexities of stochastic spatial analysis. With this knowledge, you can take full advantage of the latest features for your studies with spatial attributes.

INTRODUCTION

Spatial data are encountered in a broad variety of scientific and industrial fields, from environmental sciences and public health to meteorology and oil and gas exploration. In many cases, to analyze those data you need to deal with observations for one or more attributes at a series of locations. These types of spatial data are called point-referenced data. Usually, you want to investigate and model correlations among these observations in order to learn about the underlying process. After you produce a suitable mathematical characterization of the correlation behavior for each of the process attributes, you can predict values for each attribute at unsampled spatial locations, produce spatial maps for visualization, and extract secondary information from the findings to use for further analysis.

SAS/STAT software offers a suite of the following spatial procedures for the study of point-referenced data, also known as geostatistics:

- The VARIOGRAM procedure explores and models underlying correlations in spatial data.
- The KRIGE2D procedure uses a given correlation model to interpolate values at unsampled locations.
- The SIM2D procedure simulates spatial processes with a specified correlation structure, optionally conditional on the observed data.

The KRIGE2D and SIM2D procedures perform univariate analyses for one attribute at a time. SAS/STAT 9.2 introduced ODS Graphics to the spatial procedures. ODS Graphics are instrumental for this suite because spatial analysis and the understanding of spatial processes depend greatly on representing, visualizing, and studying spatial data on maps. This paper describes the use of these features in addition to new SAS/STAT 9.22 features of the spatial procedures in its discussion of a typical spatial analysis.

Specifically, this paper uses a geostatistical example in the field of environmental solar energy assessment. The objective is to interpolate a sample of observed values so that you understand the process that produced the observations and you can predict the process at unsampled locations. For the analysis, take the following steps:

1. Investigate your observations for underlying correlations by using the semivariogram, a common tool for studying and modeling spatial correlations. The observations are used to estimate the empirical semivariance, which examines observations in pairs and indicates how dissimilar to each other the observed values become as the distance between them increases.
2. Fit a theoretical semivariance model to the empirical one, to use the specific characteristics of this model for prediction and simulation. The example examines important fitting process complexities and demonstrates how to deal with them.
3. Use the semivariance analysis results to perform prediction and simulation tasks and to answer scientific questions in your analysis.

The detailed example builds on the close connection among the SAS/STAT 9.22 spatial procedures. It also showcases exciting new features, such as the automated semivariance fitting and the use of item stores to pass information between these procedures seamlessly. These features add significant functionality to the sequential stages of spatial analysis. In addition, this example demonstrates the use of ODS Graphics to produce customized graphs.

EMPIRICAL SEMIVARIOGRAM ANALYSIS WITH PROC VARIOGRAM

You are a businessperson looking for new opportunities. Following the call of the new green economy, you consider distributing solar panels to residents and small businesses in a certain area. You need to consider working with spatial data to address the following questions:

- What are the specific suitable locales in which to market your product and install solar panels?
- How large is the area you can consider for your solar development plans?
- What is the probability that any given location in the area is a suitable candidate for development?

A solar panel manufacturer claims that, with adequate incoming sunlight, the panels can pay off their cost within a 10-year window. Specifically, the manufacturer defines “adequate” as an average daily *global solar irradiance* (GSI) value above 18 $MJm^{-2}d^{-1}$ (mega-joules per square meter per day). Therefore you target locations that consistently receive at least 18 $MJm^{-2}d^{-1}$ of daily average GSI (DAGSI) to make this a reasonable, cost-effective solution for your clients.

Since the amount of sunlight received varies during a year, you focus on the spring months for your first analysis; you can subsequently repeat it for the other seasons. A local government agency collects DAGSI measurements from selected locations in the region, and these data are the attributes in your analysis.

To answer the previous questions, you want to know the DAGSI everywhere in the region. You want to analyze maps of solar radiation and find locales with sufficient DAGSI. You can produce such maps by interpolating the observed values or by simulating the DAGSI behavior. But wait! There are many possible ways to fill the gaps. Can you tell which approach is more appropriate? For example, should you use a smooth or a coarse surface to connect the observations? To find an answer, you need to postpone the actual interpolation and simulation tasks. Observations of a spatial attribute are usually correlated on the basis of the underlying natural process that produces the observations. As a result, when you identify the type of correlation among the observed values, you also provide a physical justification for your choice to fill the gaps. The rest of this section describes how to explore your data for such connections and prepare the findings for the subsequent tasks.

Picture the Problem

You are investigating a relatively large, square region, 300 kilometers on each side. The data set daGSIdata, defined in the section “APPENDIX” on page 17, contains the official solar radiation measurements for your target region. The data are expressed in the Easting and Northing coordinates. The variable daGSI contains the DAGSI values.

How do your data look on a map? Do you possess sufficient information to proceed with semivariogram analysis? You can address these questions by inspecting a plot of your observations and a histogram of the distances between all point pairs. To produce these plots, run the VARIOGRAM procedure with the NOVARIOGRAM and NHCLASS= options in the COMPUTE statement. To analyze the spatial correlation between points, PROC VARIOGRAM bundles point pairs in spatial lag classes for its computations. The following statements specify NHCLASS=30 spatial lag classes in the COMPUTE statement to explore the number of pairs per class:

```
ods graphics on;

proc variogram data=daGSIdata plots=equate;
   compute novariogram nhclass=30;
   coord xc=Easting yc=Northing;
   var daGSI;
run;
```

The PLOTS=EQUATE option in the PROC VARIORGAM statement specifies that the default observation plot should have the same scale for both axes. Figure 1 shows the distribution of the observations and their values in the study area. The observations exhibit some mild local variation, which reflects the effect of additional factors in the process; for example, one such factor is the cloud cover variation at different locations in the region.

Figure 1 Scatter Plot of the Spatial Distribution of Observations

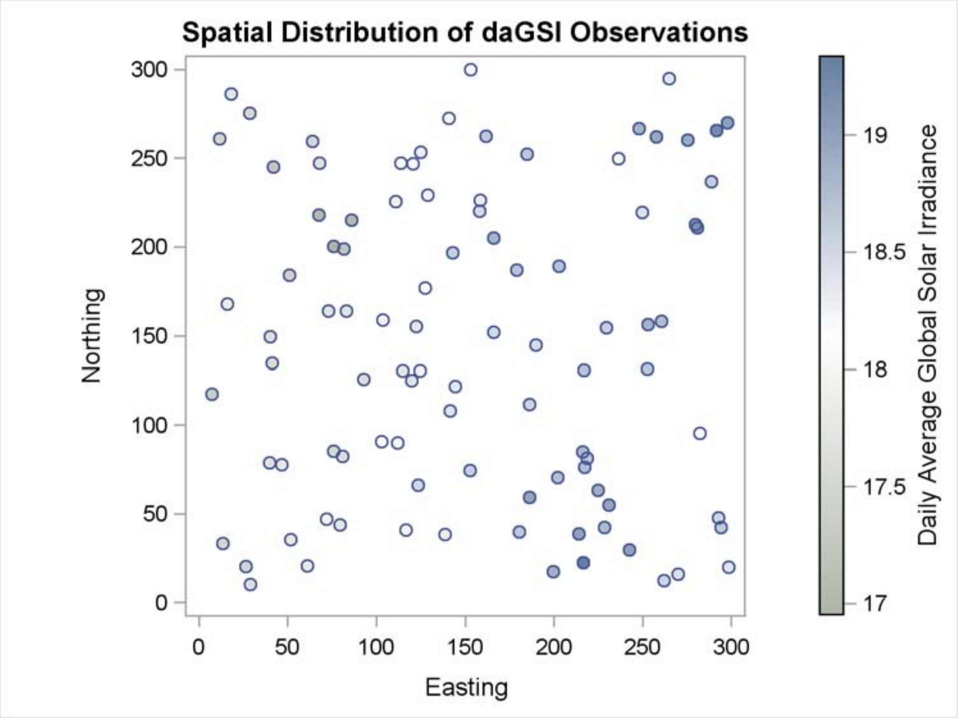
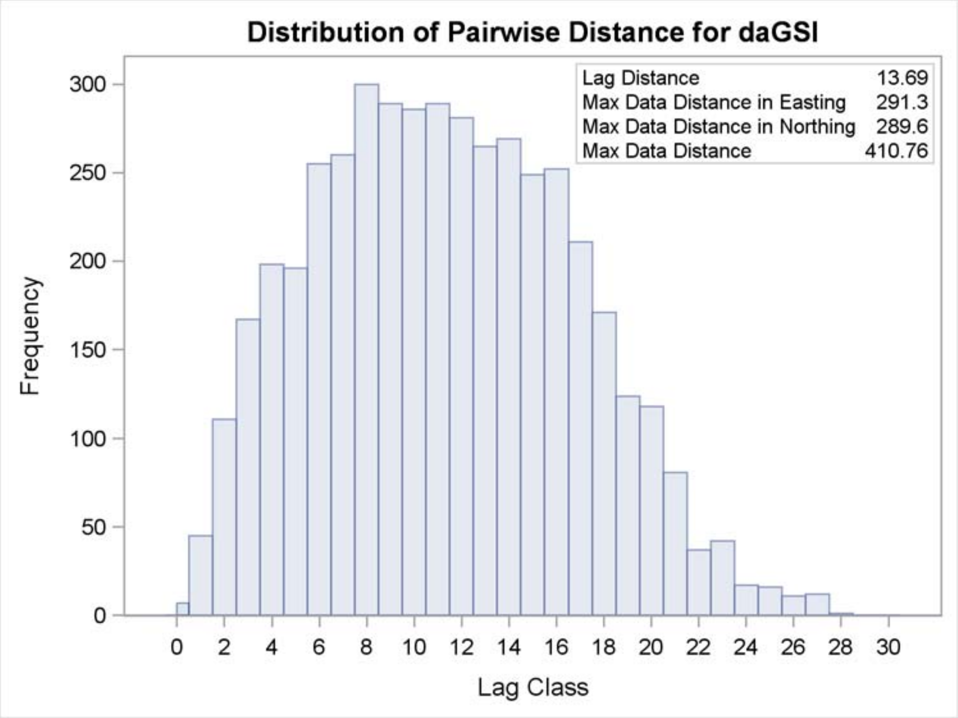


Figure 1 indicates overall higher daGSI values in the eastern part of the domain than in the western part. The vast majority of the observations range between 17 and 19 $MJm^{-2}d^{-1}$. Still, Figure 1 cannot by itself answer your questions conclusively. There are too many holes in the map. A semivariogram analysis can help you understand the correlation of DAGSI across the area. The semivariance analysis relies on the number and the distance distribution of point pairs that can be formed from your observations. Figure 2 shows the histogram of the point pair distances. The next step is to explore the empirical semivariance based on these preliminary impressions from your data.

Figure 2 Histogram of the Point Pair Distances



Exploratory Analysis

First, you estimate the empirical semivariogram from observations. Then, you fit a suitable theoretical model to the empirical semivariogram. After you settle on an adequate semivariogram model, you use it to predict and simulate the behavior of the DAGSI.

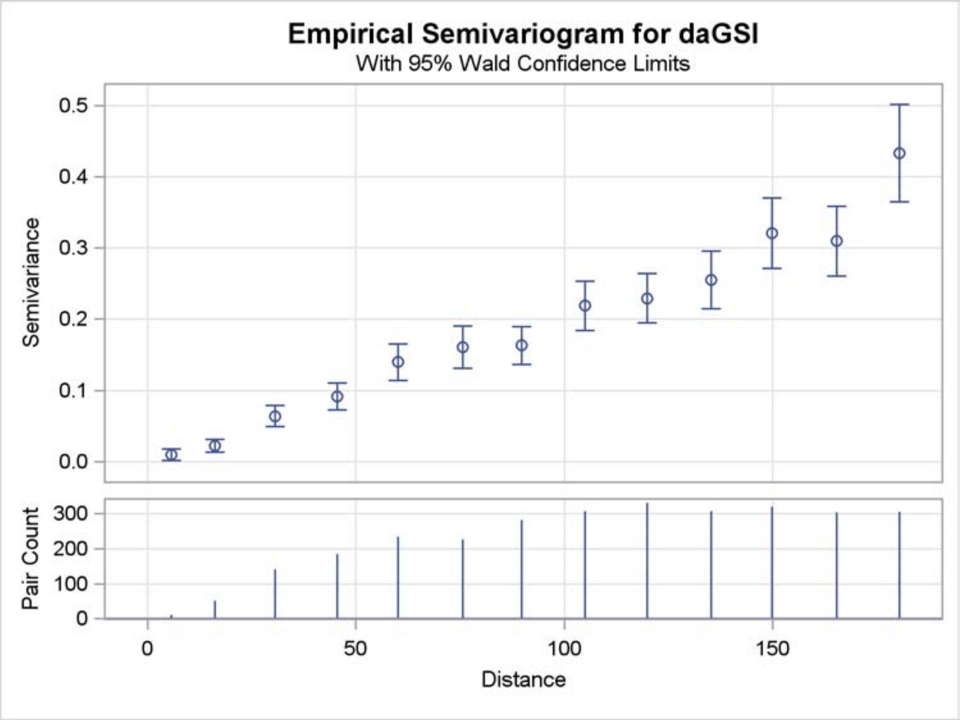
To produce the empirical semivariogram, a rule of thumb is to consider at least several lag classes with no fewer than about 30 point pairs per class for an adequately accurate empirical semivariogram estimate (Chilès and Delfiner 1999; Journel and Huijbregts 1978). Figure 2 indicates that the daGSIdata data set has plenty of observations to satisfy these guidelines.

In order to compute the empirical semivariogram, request an estimate that extends to more than half of your domain size to account for semivariances at a variety of distances (Journel and Huijbregts 1978). Run PROC VARIOGRAM again with the options LAGDISTANCE=15 and MAXLAGS=12 in the COMPUTE statement. In addition, request confidence limits for the estimated semivariance with the CL option in the COMPUTE statement, as shown in the following statements:

```
proc variogram data=daGSIdata plots(only)=semivar;
   compute lagdistance=15 maxlags=12 cl;
   coord xc=Easting yc=Northing;
   var daGSI;
run;
```

Figure 3 depicts the resulting empirical semivariogram for the observations in the daGSIdata data set.

Figure 3 Classical Empirical Semivariogram for daGSI



The points in Figure 3 estimate the function

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

This function expresses how dissimilar two observations are as a funcion of the distance $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$ between them. A basic assumption of geostatistical analysis is that the spatial correlation depends *only* on this distance and not on the locations themselves. Another assumption is that the process has a constant expected value across the region of interest. Together, these two characteristics satisfy the criteria for a process to be *(second-order) stationary* (Cressie 1993). Is that a reasonable assumption for the DAGSI in this case?

Perhaps not. Figure 1 suggests an underlying surface trend in the data in the east-west direction, with higher values in the east, contradicting the assumption of constant expectations across the region. The empirical semivariogram in Figure 3 also indicates a lack of stationarity. A semivariogram from a stationary process usually increases with distance up to a certain point called the *range*, then levels off at a value called the *scale* or *sill*. When there is an underlying trend in your observations, the empirical semivariogram never reaches a sill; instead, it increases monotonically across the entire area of interest (Christakos 1992, section 7.3). The semivariogram in Figure 3 behaves in exactly this way, thus corroborating the notion that there is a surface trend. You need to remove this trend before you continue your spatial analysis.

Removal of Surface Trends

SAS/STAT software provides a number of statistical techniques for fitting trends to spacial surfaces. For this example, your observations suggest the existence of a rather smooth surface, so you use the GLM procedure to specify a simple quadratic trend in its MODEL statement. PROC GLM fits this trend to the daGSIdata data set and saves the detrended data in the resdaGSIdata output data set. The new STORE statement saves the fit so that you can restore the trend later in the predictions of the residual resdaGSI variable. See Tobias and Cai (2010) for more information about the new STORE statement in PROC GLM.

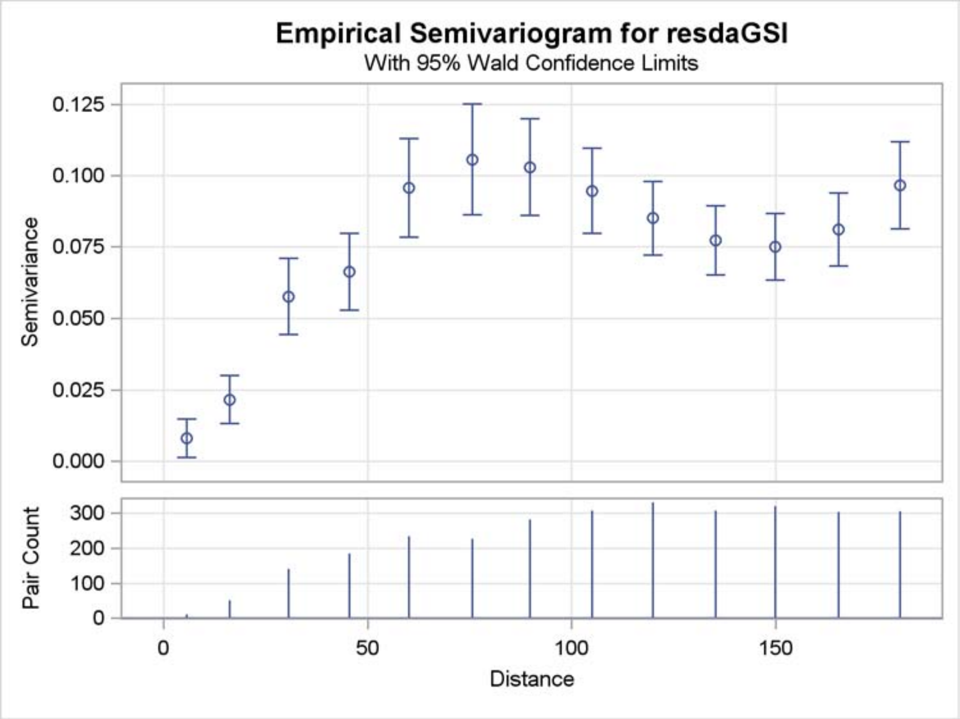
```
proc glm data=daGSIdata plots=none;
   store out=trendStore / label='Trend Analysis Information';
   model daGSI = Easting Easting*Easting Northing Northing*Northing;
   output out=resdaGSIdata predicted=pred residual=resdaGSI;
run;
```

Use the VARIOGRAM procedure to reanalyse the detrended resdaGSI variable of the residuals with the following statements:

```
proc variogram data=resdaGSIdata;
   compute lagdistance=15 maxlags=12 cl;
   coord xc=Easting yc=Northing;
   var resdaGSI;
run;
```

Figure 4 displays the semivariogram of the residuals. It has the range-and-sill structure of a stationary process and indicates a correlation behavior that fluctuates around a sill of about 0.09 variance units.

Figure 4 Classical Empirical Semivariogram for the Detrended DAGSI



Anisotropy

The last step in your exploratory analysis is to check whether correlation changes as you look in different directions in space.

- Does the sill remain the same across directions?
- Does the range change between a smaller and a larger value in two directions at right angles?

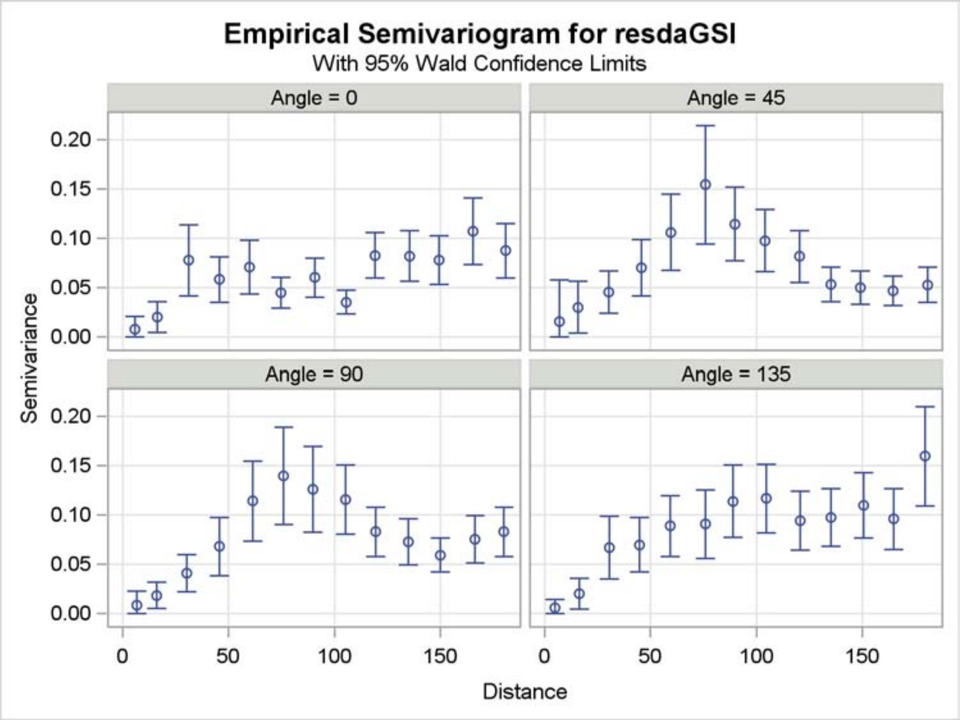
A “yes” answer to either of these questions is an indication of *anisotropy*, and you need to model each direction separately.

Use the VARIOGRAM procedure to inspect the semivariograms for anisotropy in various directions. Specify the option NDIRECTIONS=4 in the COMPUTE statement to inspect semivariograms for anisotropy in four directions ($\theta = 0°$, $\theta = 45°$, $\theta = 90°$, and $\theta = 135°$).

```
proc variogram data=resdaGSIdata plots(only)=semivar;
   compute lagdistance=15 maxlags=12 cl ndirections=4;
   coord xc=Easting yc=Northing;
   var resdaGSI;
run;
```

The panel in Figure 5 is rather inconclusive about anisotropy in the residual daGSI values. The empirical semivariances in each direction indicate a somewhat wiggly behavior at mid and larger distances, but you can tell that sills tend to stabilize in all directions in a relatively narrow neighborhood below 0.10 variance units. The ranges also seem to be of similar size in all directions.

Figure 5 Classical Empirical Semivariogram in Four Directions



In practice, a more thorough anisotropy analysis is recommended. For illustration purposes, you can assume isotropy for the resdaGSI variable. Figure 4 displays the omnidirectional empirical semivariance that you need to model.

FITTING A THEORETICAL MODEL TO THE SEMIVARIOGRAM

Now that you know you have a stationary (residual) process, your next step is to determine a mathematical model for the semivariance of that process. Only certain permissible functions can serve as semivariances (Olea 1999); yet these functions can be combined into many possible semivariogram models. At this stage, you consider the following questions:

- How do you fit a model to your empirical semivariogram?
- How do you choose an appropriate semivariogram model for this data?
- How do you explore and compare alternative models?
- How do you carry the model of your choice on to the next stage of your spatial analysis?

Prior to SAS/STAT 9.22 you could either perform a trial-and-error visual fit or use a different SAS/STAT procedure (such as PROC NLIN) for the fitting part. SAS/STAT 9.22 introduces automated semivariogram fitting in the VARIOGRAM procedure with nonlinear weighted least squares; see discussions about these and additional different methodologies in Schabenberger and Gotway (2005). The new MODEL statement in the VARIOGRAM procedure enables you to select from several model forms to automatically fit the empirical semivariogram in Figure 4. You can request a specific model or choose the best fit among multiple models. You can specify models with one form or models with nested structures.

In addition to the exponential (EXP), Gaussian (GAU), spherical (SPH), and power (POW) semivariance models, SAS/STAT 9.22 introduces the cubic (CUB), pentaspherical (PEN), sine-hole effect (SHE), and Matérn (MAT) forms for both modeling and prediction in the spatial procedures.

The sinusoidal behavior around the sill, exhibited in the semivariogram in Figure 4, makes the sine-hole effect form a prime candidate for fitting in this case. However, you decide to request an automated model fit. The FORM=AUTO option in the following MODEL statement requests the best semivariogram model for this data by using any one of the listed forms or any sum of two of them—a total of 30 different possible models:

```
model form=auto(mlist=(gau,exp,sph,she,mat) nest=1 to 2);
```

The following statements fit a model to the empirical semivariogram and request a fit plot of the fitted models. The CL option in the MODEL statement requests confidence limits of the estimated parameters. Finally, the new STORE statement saves the semivariance fit in the daGSIsemiv item store. You can subsequently use this item store to transfer the fit model directly to PROC KRIGE2D or PROC SIM2D.

```
proc variogram data=resdaGSIdata plots(only)=(fit);
   compute lagdistance=15 maxlags=12 cl;
   coord xc=Easting yc=Northing;
   var resdaGSI;
   model form=auto(mlist=(gau,exp,sph,she,mat) nest=1 to 2) cl;
   store out=daGSIsemiv / label='Global Solar Irradiance Semivariances';
run;
```

The results from the automatic model fitting in PROC VARIOGRAM are summarized in Figure 6. Models are ordered by the values of the corresponding fit criteria. The first model listed is the one with the lowest weighted sum of squared errors (weighted SSE) and is the selected model (best fit) according to this criterion. Optionally, you can request that models be ranked using Akaike's information criterion (AIC). The ranked models are further categorized into empirical equivalence classes. Each class contains a set of consecutively ranked models with similar semivariance values across a spectrum of distances.

Figure 6 Semivariogram Model Fitting Summary

```
                              The VARIOGRAM Procedure
                           Dependent Variable: resdaGSI
                              Angle: Omnidirectional

                                    Fit Summary

                                         Weighted
                 Class    Model               SSE            AIC

                     1    SHE-Sph         5.28152       -1.70957
                     2    SHE             8.85602        1.00993
                          Mat-SHE         8.85602        7.00993
                          Sph-SHE         8.85607        5.01000
                          Exp-SHE         8.85614        5.01010
                          Gau-SHE         8.85628        5.01031
                          SHE-SHE         8.85629        5.01033
                     3    Gau            20.78000       12.09754
                          Mat-Gau        20.78000       18.09754
                          Exp-Gau        20.78021       16.09767
                          SHE-Gau        20.78032       16.09774
                          Gau-Gau        20.78038       16.09778
                          Gau-Mat        20.78053       18.09787
                          Sph            21.86445       12.75886
                          Sph-Gau        21.86445       16.75886
                          Sph-Sph        21.86448       16.75888
                          Sph-Mat        21.86451       18.75890
                          Sph-Exp        21.86460       16.75896
                          Mat-Sph        21.86464       18.75898
                          Exp-Sph        21.86467       16.75900
                          Gau-Sph        21.86483       16.75909
                     4    Exp            34.04998       18.51744
                          SHE-Exp        34.05005       22.51747
                          Mat-Exp        34.05010       24.51749
                          Exp-Exp        34.05013       22.51750
                          Gau-Exp        34.05026       22.51755
```

PROC VARIOGRAM also displays the parameter estimates for the selected model, as shown in Figure 7.

Figure 7 Semivariogram Fitting Parameter Estimates

```
                                Parameter Estimates

                                  Approximate 95%
                        Approx Confidence Limits                Approx
 Parameter Estimate Std Error     Lower     Upper DF t Value Pr > |t| Gradient

 Nugget     0.000393  0.003447         0  0.008342  8    0.11   0.9120 -0.00597
 SHEScale1   0.06370  0.006048   0.04975   0.07765  8   10.53   <.0001 -0.00373
 SHERange1   57.0261    1.7052   53.0940   60.9583  8   33.44   <.0001 -4.72E-7
 SphScale2   0.02427  0.007727  0.006453   0.04209  8    3.14   0.0138 -0.00482
 SphRange2   33.5239    6.6860   18.1060   48.9418  8    5.01   0.0010 -7.39E-7
```

Finally, the fit plot in Figure 8 shows the shapes of the fitted models listed in Figure 6. The thick line indicates the selected model, whereas each of the other lines designates an equivalence class of one of the other models. The number in each set of parentheses in the inset tells you how many more equivalent models are in that class.

Figure 8 Fitted Theoretical and Empirical Semivariogram for Detrended Global Solar Irradiance

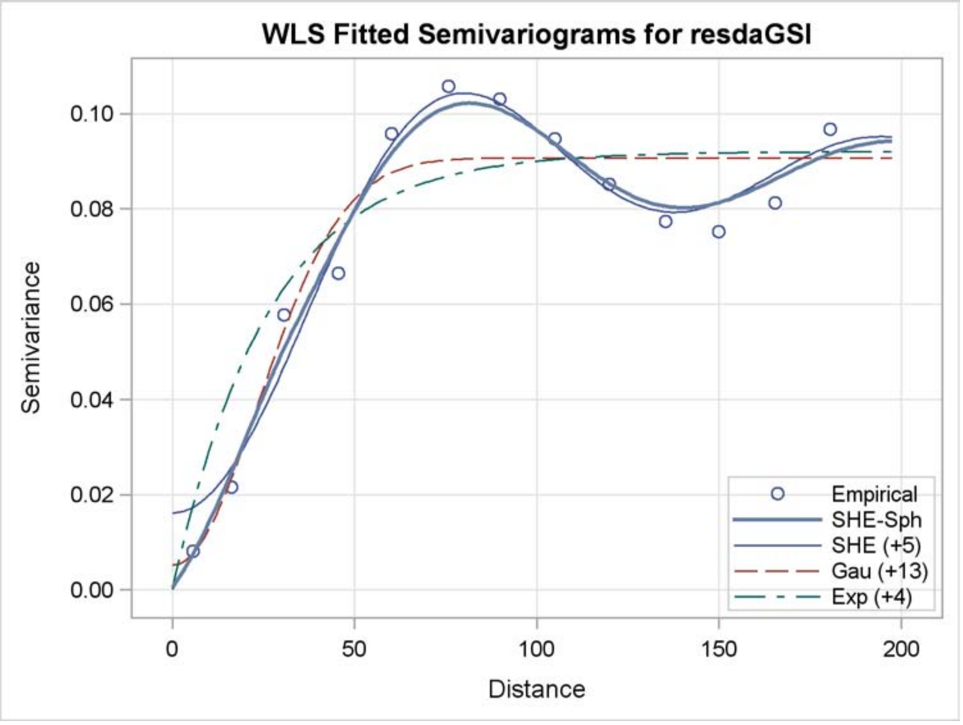


Figure 8 indicates that a simple sine-hole effect model (SHE) can fit most of the empirical semivariogram quite well. Still, the extra spherical (SPH) form enables the nested SHE-SPH model to better fit the empirical semivariogram near the origin. In particular, the nested model is flexible enough to dip all the way to zero at Distance $= 0$, corresponding to the tiny estimate for the "Nugget" parameter in Figure 7. This indicates that the nugget effect is in fact zero. A nugget effect corresponds to measurement error; a nonzero value means that two measurements at the same location can be different. The terminology arises from the mining origins of much geostatistical analysis, in which the incidence of nuggets can add noise to the underlying ore measurements.

You can explore this zero-nugget-effect hypothesis by reanalyzing the semivariance with just the nested SHE-SPH model. Use the HOLD= option in the PARMS statement to constrain the nugget effect to be zero.

```
proc variogram data=resdaGSIdata plots(only)=(fit);
   compute lagdistance=15 maxlags=12 cl;
   coord xc=Easting yc=Northing;
   var resdaGSI;
   model form=(she,sph) cl / covb;
   parms (0) (.) (56 to 58) (.) (33 to 34) / hold=1;
run;
```

The PARMS statement specifies initial values for all of the model parameters. Refering to the parameter estimates in Figure 7, you specify 0 for the nugget effect, a sine-hole effect range between 56 and 58 kilometers, and a spherical range between 33 and 34 kilometers. For the structure scale parameters you specify missing values, so PROC VARIOGRAM uses default initial values for these, according to Jian, Olea, and Yu (1996). PROC VARIOGRAM tries out all possible combinations of the initial values specified for the parameters and selects the best performing set as initial values for the model fit. Finally, the HOLD=1 option specifies that the first parameter—that is, the nugget effect—should be fixed at its initial value of 0.

The results from this reanalysis (not shown) are very similar to the best model fit in Figure 6 and Figure 7, verifying that a nugget effect is probably not required. In general, the PARMS statement gives you the flexibility to explore specific models. In this case, it helped you verify the good fit you obtained earlier with the FORM=AUTO option in the MODEL statement. As you can see, a detailed semivariance fitting analysis might be more than a one-step process; the VARIOGRAM procedure comes with tools that simplify this task and help you make appropriate decisions.

When you model semivariograms, keep in mind the remark of the 18th-century Enlightenment French writer and philosopher Voltaire, "*Le mieux est l'ennemi du bien*" [The best is the enemy of the good] ("La Bégueule," 1772). In spatial analysis an *overall good* fit is more important than the *most accurate* fit. In particular, you should choose a fit that follows the empirical behavior more closely towards the low-distance values in the semivariogram. For example, the fit plot in Figure 8 indicates that Equivalence Class 3, represented by the Gaussian model, could also satisfy the latter criterion. In general, you should try different lag arrangements and model specifications to obtain a clear empirical semivariogram shape and an overall good fit with the simplest model possible. You should avoid overfitting a model, remembering that the target empirical semivariance is itself only an estimate (Goovaerts 1997, section 4.2.4).

PREDICTION OF THE GLOBAL SOLAR IRRADIANCE

The observation plot in Figure 1 hints at areas where the detrended GSI has peaks and valleys. Your sales assessment requires a more thorough representation of potential target sites. Can you complete the area picture with values for locations without observations? The KRIGE2D procedure can do this for you with its predictive features, now that you have a spatial correlation model for the resdaGSI variable.

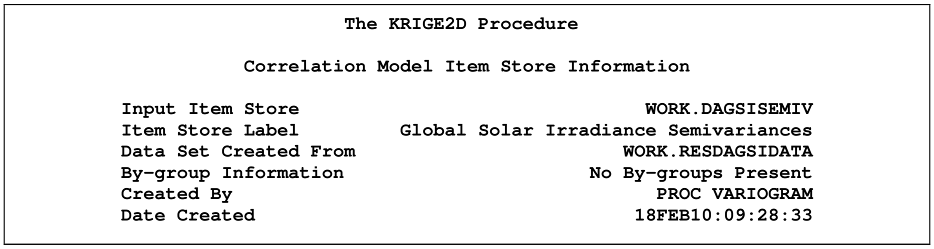
Prediction with Selected Model

The correlation information from PROC VARIOGRAM that you saved in the daGSIsemiv item store makes it simple to provide the correlation model to PROC KRIGE2D. Specify the input item store name in the IN= option in the new RESTORE statement, and specify the INFO option to produce basic output about the store models. The new STORESELECT option in the MODEL statement enables you to use any of the fitted models in the input item store. In the following statements, you specify the STORESELECT option by itself; by default, PROC KRIGE2D uses for prediction the top-ranked fitted model in the item store:

```
proc krige2d data=resdaGSIdata outest=resdaGSIpred plots=equate;
   restore in=daGSIsemiv / info(det);
   coordinates xc=Easting yc=Northing;
   predict var=resdaGSI;
   model storeselect;
   grid x=0 to 300 by 7.5 y=0 to 300 by 7.5;
run;
```

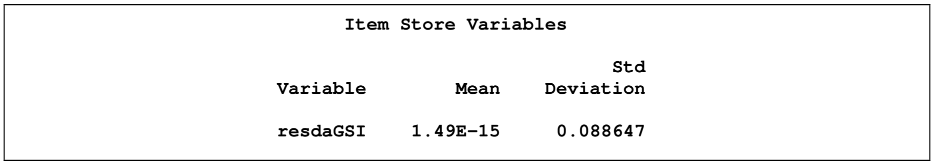
The GRID statement specifies a rectangular prediction grid with a horizontal and vertical distance of 7.5 kilometers between nodes; that is, you define a grid with 41 nodes on each grid side for a total of 1,681 nodes in your study area. The output in Figure 9 shows general information about the daGSIsemiv input item store, and Figure 10 shows the table with information about the item store variables. Observe the zero mean value of the detrended GSI resdaGSI variable that was passed on to PROC KRIGE2D from PROC VARIOGRAM.

Figure 9 PROC KRIGE2D and Input Item Store General Information



```
                        The KRIGE2D Procedure

               Correlation Model Item Store Information

      Input Item Store                              WORK.DAGSISEMIV
      Item Store Label           Global Solar Irradiance Semivariances
      Data Set Created From                       WORK.RESDAGSIDATA
      By-group Information                     No By-groups Present
      Created By                                     PROC VARIOGRAM
      Date Created                                 18FEB10:09:28:33
```

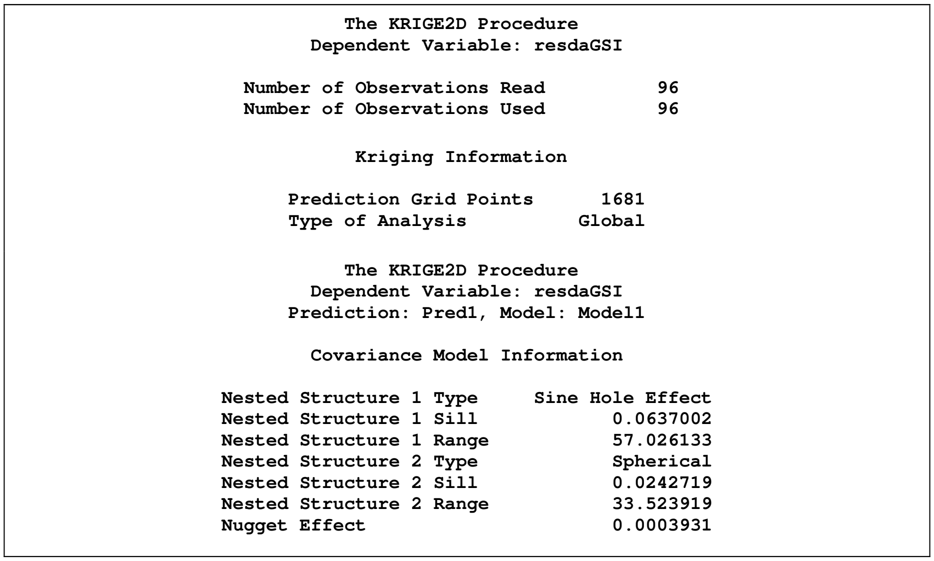
Figure 10 Variables in the Input Item Store



```
                        Item Store Variables

                                                Std
                  Variable        Mean    Deviation

                  resdaGSI    1.49E-15     0.088647
```

The KRIGE2D procedure performs the specified prediction task with the SHE-SPH model because it is the top-ranking model in the daGSIsemiv item store. Figure 11 displays the number of observations table and information about the kriging process input, which includes the parameter values of the selected model.

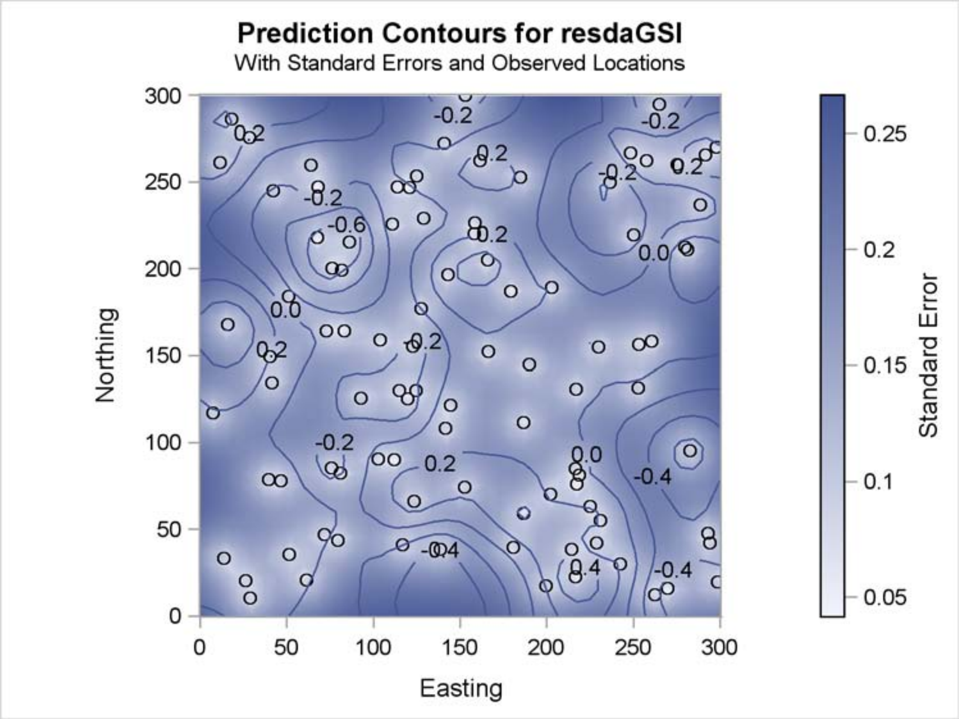
Figure 11 Number of Observations, Kriging, and Model Information Tables



```
                    The KRIGE2D Procedure
                 Dependent Variable: resdaGSI

          Number of Observations Read          96
          Number of Observations Used          96

                     Kriging Information

              Prediction Grid Points      1681
              Type of Analysis          Global

                    The KRIGE2D Procedure
                 Dependent Variable: resdaGSI
              Prediction: Pred1, Model: Model1

                 Covariance Model Information

      Nested Structure 1 Type     Sine Hole Effect
      Nested Structure 1 Sill            0.0637002
      Nested Structure 1 Range           57.026133
      Nested Structure 2 Type            Spherical
      Nested Structure 2 Sill            0.0242719
      Nested Structure 2 Range           33.523919
      Nugget Effect                      0.0003931
```

Finally, Figure 12 shows the plot of the predicted residual resdaGSI values and their standard errors. The residual GSI predictions are much smaller than the observed GSI shown in Figure 1. Not surprisingly, standard errors are larger at grid locations farther away from the observations.

Figure 12 Prediction Contours and Standard Error for the resdaGSI Variable



Prediction with Alternative Model

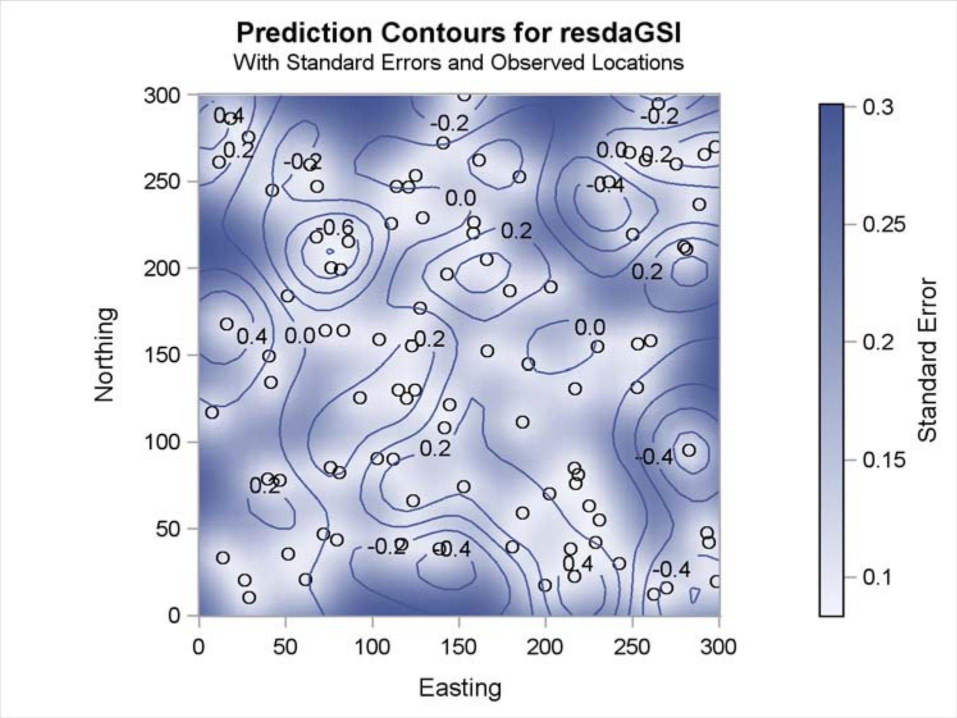
The Equivalence Class 3 models displayed in Figure 6 are also good candidates to use for prediction, as discussed earlier. In the following statements you run PROC KRIGE2D again. You specify MODEL=GAU in the STORESELECT option in the MODEL statement to perform prediction with a representative from Equivalence Class 3 and to compare the prediction results in both cases.

```
proc krige2d data=resdaGSIdata outest=resdaGSIpredAlt plots=equate;
   restore in=daGSIsemiv;
   coordinates xc=Easting yc=Northing;
   predict var=resdaGSI;
   model storeselect(model=gau);
   grid x=0 to 300 by 7.5 y=0 to 300 by 7.5;
run;
```

Figure 13 displays the new prediction plot. Compared to the selected model SHE-SPH prediction plot in Figure 12, prediction with the Gaussian model results in slightly steeper gradients of the resdaGSI variable values. In some locations, Gaussian model predictions tend to be more extreme, but overall both maps exhibit quite similar patterns in the predicted values.

With respect to the standard error, the Gaussian model prediction errors are generally higher than the corresponding ones produced with the SHE-SPH model. The standard error surface of the selected SHE-SPH model in Figure 12 suggests smoother error variation than the Gaussian prediction errors, whereas in Figure 13 errors tend to increase faster by moving farther away from the observation locations.

Figure 13 Prediction Contours and Standard Error for the resdaGSI Variable



The comparison suggests that both predictions produce similar results. This is reasonable, considering that they both use valid, sensible correlation models. The subjective nature of spatial analysis does not declare an absolute winner in this comparison. To choose between them, you usually consult additional resources for particular characteristics of the natural process in your study.

Surface Trend Restoration and Analysis

At this point, you have obtained predictions of the detrended DAGSI. However, the prediction of the DAGSI itself is of more direct interest. To compute the predicted daGSI variable, combine the estimated trend for the DAGSI with the predicted residual process for resdaGSI.

The following statements use the PLM procedure (Tobias and Cai 2010), new in SAS/STAT 9.22, to apply the quadratic model that you fit with the GLM procedure to the grid of node locations in the predicted residual process resdaGSIpred data set:

```
proc plm restore=trendStore;
   score data=resdaGSIpred(rename=(gxc=Easting gyc=Northing))
         out=predGridTrend;
run;
```

The following DATA step adds the trend estimate in the Estimate variable and the residual prediction in the resdaGSI variable to produce the DAGSI prediction:

```
data daGSIpred;
   set predGridTrend;
   daGSI = Predicted + Estimate;
run;
```

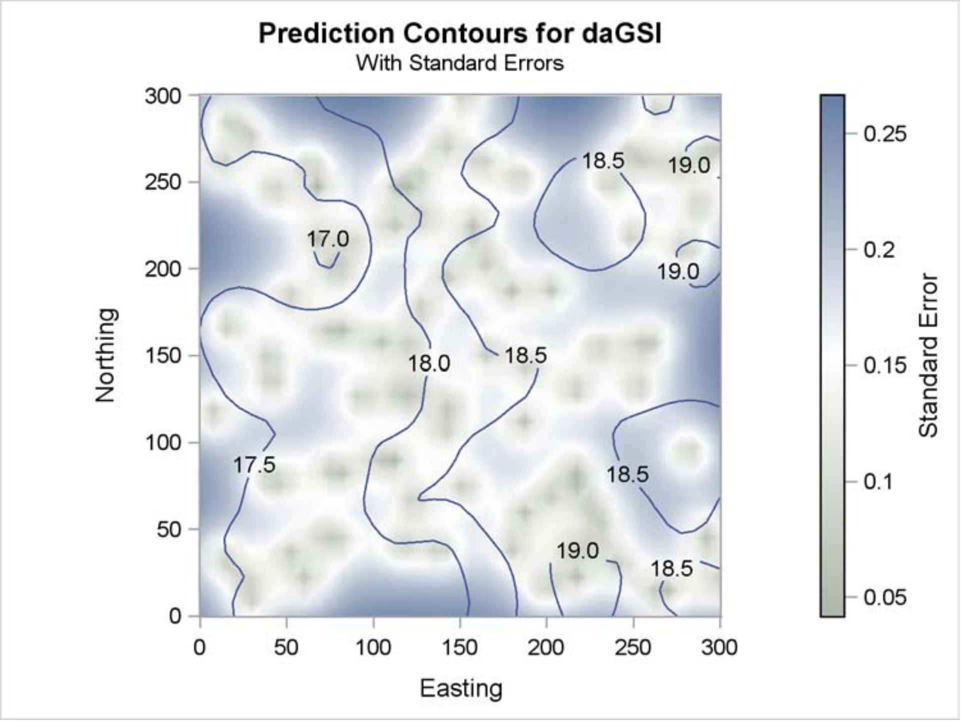
You can create a plot of these predictions by using the ODS Graph Template Language and the SGRENDER procedure to define a specialized plot that is similar to the ones produced by the KRIGE2D procedure in Figure 12 and Figure 13:

```
proc template;
   define statgraph surfacePlot;
      dynamic _VARX _VARY _VAR1 _VAR2 _TITLE _SUBTITLE _LEGENDLABEL;
      BeginGraph;
      entrytitle _TITLE;
      entrytitle _SUBTITLE / textattrs=GraphValueText;
      layout overlayequated /
         xaxisopts  = (offsetmin=0 offsetmax=0)
         yaxisopts  = (offsetmin=0 offsetmax=0)
         equatetype = equate;
      contourplotparm x=_VARX y=_VARY z=_VAR1 /
         contourtype=gradient name='spatploterr';
      contourplotparm x=_VARX y=_VARY z=_VAR2 /
         contourtype=LabeledLine nhint=10 name='spatplot';
      continuouslegend 'spatploterr' / title=_LEGENDLABEL;
      endlayout;
      EndGraph;
   end;
run;

proc sgrender data=daGSIpred template=surfacePlot;
   dynamic _VARX        = 'Easting'
           _VARY        = 'Northing'
           _VAR1        = 'StdErr'
           _VAR2        = 'daGSI'
           _TITLE       = 'Prediction Contours for daGSI'
           _SUBTITLE    = 'With Standard Errors'
           _LEGENDLABEL = 'Standard Error';
   label Easting  = 'Easting'
         Northing = 'Northing';
run;
```

Figure 14 shows the resulting contour plot of the predicted daGSI and its standard error. Although the error values are the same as in the prediction plot of the residuals in Figure 12, the actual restored trend values dominate the much smaller residual predictions. For this reason, if you repeat the previous steps to produce the corresponding plot for prediction with the Gaussian model, you end up with a map (not shown here) whose prediction contours are almost identical to Figure 14.

The GLM procedure helped you remove an estimate of the underlying surface trend in your DAGSI observations. You might consider that estimate as one that possibly follows your observations too closely to allow for larger fluctuations in the daily average GSI residuals. Larger fluctuations might lead to visibly more diverse DAGSI predictions for different correlation models. In all of these cases you need to exercise judgement and intuition to obtain sensible results by repeating and adjusting your analysis.

Figure 14 Prediction Contour Map for the DAGSI and Standard Error Surface



For the scope of this analysis, Figure 14 indicates that, at least in the spring season, more than half of the area in the eastern part of the domain receives amounts of DAGSI above the panel manufacturer specification of 18 $MJm^{-2}d^{-1}$. This is good news because these locations seem to have good potential for solar development. The standard error surface indicates error values throughout the study domain that range at most around 1% of the daGSI variable values. Expectedly, standard errors are smaller near the observations. In practice, you would now combine these results with other forms of geographical information, and take into account additional factors such as the local terrain and vegetation cover. In any case, you have a first solid assessment of site identification to help you with your sales plans.

SITE ASSESSMENT WITH SIMULATION

In the preceding sections you saw how to use PROC VARIOGRAM to model spatial covariance and how to use such covariance models in PROC KRIGE2D to predict spatial processes. Spatial simulation can help you understand even better the behavior of the daGSI variable and answer more of your spatial problem questions. You simulate the solar irradiance process with the selected correlation structure to examine the range of different likely realizations for the distribution of the DAGSI values in the area. The SIM2D procedure provides you with the tools for this task.

Daily Average GSI Simulation

In this example, you use PROC SIM2D in its conditional simulation mode, which is the most useful mode for data analysis. The idea is to compute not just a single predicted residual DAGSI surface, as with PROC KRIGE2D, but rather a large number of possible realizations, conditional on the observed data and taking the covariance model into account. The syntax for PROC SIM2D is quite similar to that for PROC KRIGE2D, with the SIMULATE statement in SIM2D essentially taking the role of the PREDICT and MODEL statements in KRIGE2D. You condition the simulation on the resdaGSI variable in the VAR= option in the SIMULATE statement. The NUMREAL= option in the SIMULATE statement requests 1,000 simulated realizations of the resdaGSI process.

```
proc sim2d data=resdaGSIdata outsim=resdaGSIsim;
   restore in=daGSIsemiv;
   coordinates xc=Easting yc=Northing;
   simulate var=resdaGSI numreal=1000 storeselect;
   grid x=0 to 300 by 7.5 y=0 to 300 by 7.5;
run;
```

PROC SIM2D displays an information table about the data and model, similar to Figure 11, and creates a data set resdaGSIsim that contains the simulated realizations. There is also a default simulation plot of the means of the residuals resdaGSI variable, but it, too, looks very much like Figure 12 produced by PROC KRIGE2D. There is a good theoretical reason for this: the expected value of the conditional simulation in PROC SIM2D is precisely the predicted value from PROC KRIGE2D, and the standard deviation of the simulations is asymptotically the same as the standard error for the kriging prediction. As opposed to the smoothing effect of prediction and averaged realizations, each individual realization exhibits one possible behavior of the process with increased detail at the node level and within the guiding limits of the specified correlation model. The point is that simulation via PROC SIM2D shows its utility best when it is used for more complex statistics for the realizations. The following discussion is one example of a more elaborate question you can ask of this simulated data.

Site Analysis and Assessment

The PROC SIM2D simulation resdaGSIsim output data set is a valuable resource. You use it to estimate the percentage of the target region area that is suitable for your sales plans. In addition, it can help you produce a map of the probability that each location in the area satisfies the solar manufacturer feasibility criterion.

Start by sorting the simulation data in order to merge it with the simGridTrend data set with the quadratic trend estimates from PROC PLM. The Predicted variable is the PROC PLM trend estimate and the svalue variable is the simulated resdaGSI value from the SIM2D procedure resdaGSIsim output data set. Create the above18 variable, which flags all simulated values above 18 $MJm^{-2}d^{-1}$. The above18perc variable expresses the same number in a more convenient format for percentage reporting. The following DATA step performs the necessary steps to prepare the input data set for the MEANS procedure:

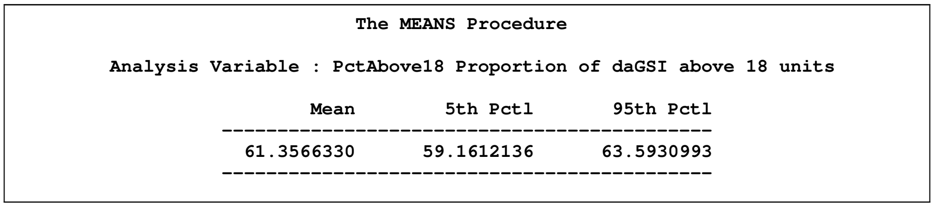
```
proc sort data=resdaGSIsim;
   by gxc gyc;
data daGSIallSim;
   merge predGridTrend
         resdaGSIsim(rename=(gxc=Easting gyc=Northing));
   by Easting Northing;
   daGSI      = Predicted + svalue;
   above18    = (daGSI>18);
   above18pct = (daGSI>18)*100;
run;
```

The first goal is to answer the question, what percentage of the area is suitable for solar development based on the spring season analysis? First use the CLASS statement in PROC MEANS to estimate the percentage in each realization, and then use PROC MEANS again to estimate the true percentage and its 5% and 95% confidence limits. The following statements perform these tasks:

```
proc means data=daGSIallSim noprint;
   class _ITER_;
   ways 1;
   var above18pct;
   output out=daGSIabove18data mean=PctAbove18;
proc means data=daGSIabove18data mean p5 p95;
   var PctAbove18;
   label PctAbove18="Proportion of daGSI above 18 units";
run;
```

According to the results in Figure 15, based on the averaged simulated DAGSI during the spring months alone, about 61.3% of your study area is eligible for development of solar projects. At the 90% confidence level, this estimate is expected to range between about 59.1% and 63.6%.

Figure 15 Area Percentage That Satisfies the Cost-Effectiveness Criterion daGSI > 18



```
                        The MEANS Procedure

    Analysis Variable : PctAbove18 Proportion of daGSI above 18 units

                     Mean         5th Pctl        95th Pctl
            ------------------------------------------------
              61.3566330       59.1612136       63.5930993
            ------------------------------------------------
```

Based on your analysis assumptions, can you map the solar development potential over the entire area? You answer this question by computing the probability that the DAGSI is greater than 18 $MJm^{-2}d^{-1}$ for every node in your simulation grid. You only need to call the MEANS procedure to average the above18 variable in the daGSIallSim data set over each grid node. Use the following statements:

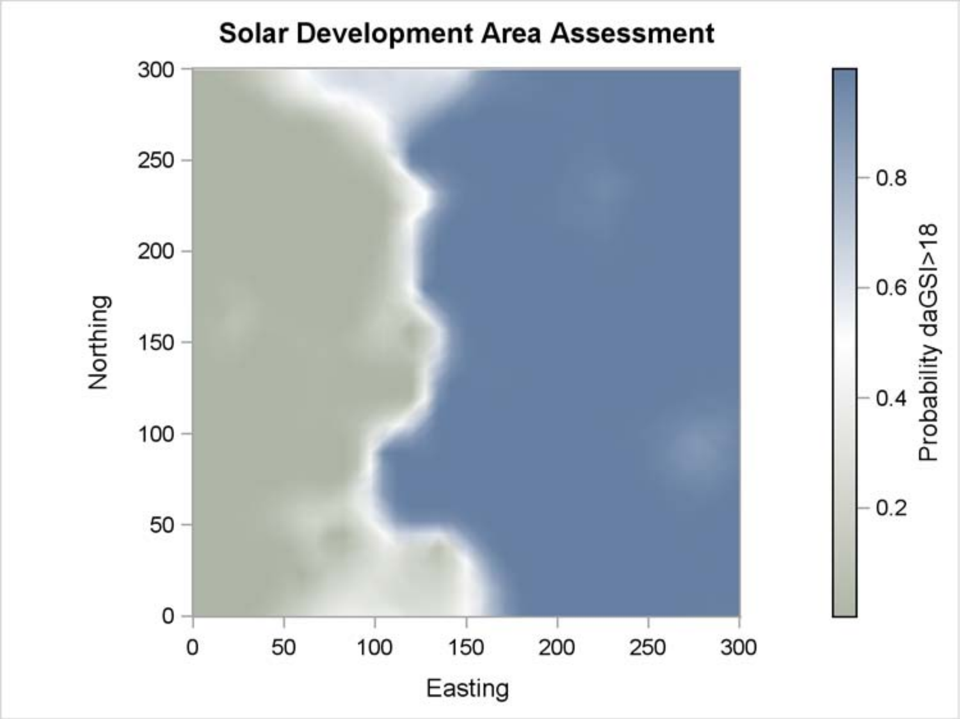
```
proc means data=daGSIallSim noprint;
   by Easting Northing;
   var above18;
   output out=daGSIabove18loci mean=probAbove18;
run;
```

The following statements use the Graph Template Language and PROC SGRENDER to create the probability plot:

```
proc template;
   define statgraph probabilityPlot;
      dynamic _VARX _VARY _VAR _TITLE _LEGENDLABEL;
      BeginGraph;
      entrytitle _TITLE;
      layout overlayequated /
         xaxisopts  = (offsetmin=0 offsetmax=0)
         yaxisopts  = (offsetmin=0 offsetmax=0)
         equatetype = equate;
      contourplotparm x=_VARX y=_VARY z=_VAR /
         contourtype=gradient name='probsurface';
      continuouslegend 'probsurface' / title=_LEGENDLABEL;
      endlayout;
      EndGraph;
   end;
run;

proc sgrender data=daGSIabove18loci template=probabilityPlot;
   dynamic _VARX        = 'Easting'
           _VARY        = 'Northing'
           _VAR         = 'probAbove18'
           _TITLE       = 'Solar Development Area Assessment'
           _LEGENDLABEL = 'Probability daGSI>18';
   label Easting  = 'Easting'
         Northing = 'Northing';
run;

ods graphics off;
```

Figure 16 shows the map of the probability that solar development is cost-effective, based on your current assumptions. You can easily identify in Figure 16 the eastern part of the area that was earlier recommended for solar development by the KRIGE2D prediction plot in Figure 14.

Figure 16 Plot of the Probability That DAGSI Is Greater Than 18 $MJm^{-2}d^{-1}$



CONCLUSION

Geostatistical spatial analysis involves two steps: determining the spatial correlation model for your data, and then using the model to make inferences about the underlying spatial process that are conditional on the observed data. Spatial analysis tasks are greatly enhanced in the suite of spatial procedures in SAS/STAT 9.22.

Automatic nonlinear fitting of theoretical semivariance models enables you to choose a semivariance model from a broad and flexible variety of forms. ODS Graphics plots help you navigate this highly qualitative process and assist your decision making. In addition, the updated suite of spatial procedures streamlines the sequential steps in spatial analysis and ties these procedures tightly together with the introduction of item stores.

This presentation focuses on these new SAS/STAT features which enable you to perform efficient geostatistical analysis. Some additional new features in these procedures include the extension of the KRIGE2D and SIM2D mapping capabilities to one-dimensional problems, the new ID statement for labeling observations, and the Moran plot in PROC VARIOGRAM. For more information about all of the new features in the SAS/STAT spatial procedures, see Chapter 1, “What’s New in SAS/STAT 9.22” (*SAS/STAT User’s Guide*).

APPENDIX

The following DATA step creates simulated daily average global solar irradiance data for the area of interest (in $MJm^{-2}d^{-1}$ units):

```
data daGSIdata;
   input Easting Northing daGSI @@;
   label daGSI='Daily Average Global Solar Irradiance';
   datalines;
    13.6   33.4  17.4906   116.6   40.9  17.9709    86.0  215.3  17.0232
   248.1  266.9  18.8413    40.2  149.3  17.6422   249.9  219.5  18.4011
   298.6   19.8  18.3800   161.6  262.3  18.6537   165.9  204.9  18.8376
   236.3  249.9  18.2338   141.5  108.1  18.3977   253.0  156.4  18.8378
   218.6   81.2  18.6493    46.5   77.8  17.7590   292.8   47.7  18.5304
    83.0  164.0  17.7159   110.8  225.6  17.9139   202.9  189.1  18.7072
   199.5   17.5  18.8718   103.7  158.7  17.9594   184.7  252.5  18.6570
    76.0  200.3  16.9481    51.5   35.6  17.7746   158.1  220.3  18.4775
   230.9   55.0  18.9772   158.5  226.4  18.3128   252.8  131.5  18.6633
    61.2   20.6  17.8255   279.6  212.6  19.0338   294.1   42.3  18.6408
    80.9   82.3  17.6355   124.8  253.5  18.3031    18.0  286.4  17.7586
   229.7  154.7  18.5950    71.8   47.0  17.9481   280.9  210.9  19.1626
    92.9  125.8  17.6257    28.6  275.5  17.5359   186.4  111.7  18.5684
    16.1  167.7  17.8476   114.9  130.2  17.7519   140.7  272.4  18.1968
   122.5  155.3  17.7771    63.8  259.7  17.4883   262.1   12.3  18.5385
   257.6  262.2  18.9482   128.8  229.1  17.9638   269.8   15.9  18.4247
   288.6  236.9  18.6137   124.5  130.2  17.9114     7.3  117.2  17.3567
   153.0  299.9  18.0288    75.8   85.2  17.5313    67.9  247.3  17.6824
    81.7  199.0  17.2191   275.3  260.2  18.9168   180.6   39.6  18.6395
    28.8   10.3  17.6430   217.2   76.1  18.7481   282.4   95.2  18.0833
   143.0  196.7  18.5825    42.0  245.1  17.2660   179.1  187.1  18.6990
    41.3  134.6  17.5906   217.0  130.8  18.6883   166.1  152.1  18.5032
   186.4   59.2  18.9922   264.9  294.8  18.3716   228.6   42.1  18.7740
   214.1   38.5  19.0145   123.5   66.2  18.5217    26.3   20.4  17.4753
    39.7   78.8  17.6835   113.8  247.1  18.0232   225.0   63.3  18.8826
   242.4   29.8  19.0110   260.5  158.1  18.7582   297.8  270.0  19.0336
   111.8   90.1  18.2008    79.6   43.7  17.7467   102.8   90.7  18.0879
    51.1  184.0  17.4158   216.3   85.0  18.7803    11.7  261.0  17.5346
   291.6  265.7  19.2310   119.8  125.1  17.7093   127.4  176.8  18.1153
   152.7   74.3  18.5789   202.1   70.3  18.7343    67.7  218.1  17.0247
   138.5   38.3  17.8701    72.9  163.9  17.6970   189.7  144.7  18.4584
   144.3  121.6  18.3561   120.7  246.9  18.0552   216.6   22.4  19.3410
   ;
```

REFERENCES

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.

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

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

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

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

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

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

Tobias, R. and Cai, W. (2010), “Introducing PROC PLM and Postfitting Analysis for Very General Linear Models in SAS/STAT 9.22,” in *Proceedings of the SAS Global Forum 2010 Conference*, Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Alexander Kolovos
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-2165
alexander.kolovos@sas.com