

# Analyzing Spatial Point Patterns Using the New SPP Procedure

Pradeep Mohan and Randy Tobias, SAS Institute Inc.

## ABSTRACT

In many spatial analysis applications (including crime analysis, epidemiology, ecology, and forestry), spatial point process modeling can help you study the interaction between different events and help you model the process intensity (the rate of event occurrence per unit area). For example, crime analysts might want to estimate where crimes are likely to occur in a city and whether they are associated with locations of public features such as bars and bus stops. Forestry researchers might want to estimate where trees grow best and test for association with covariates such as elevation and gradient. This paper describes the SPP procedure, new in SAS/STAT® 13.2, for exploring and modeling spatial point pattern data. It describes methods that PROC SPP implements for exploratory analysis of spatial point patterns and for log-linear intensity modeling that uses covariates. It also shows you how to use specialized functions for studying interactions between points and how to use specialized analytical graphics to diagnose log-linear models of spatial intensity. Crime analysis, forestry, and ecology examples demonstrate key features of PROC SPP.

## INTRODUCTION

This paper introduces the SPP procedure, new in SAS/STAT 13.2, for the analysis of spatial point pattern data. Following are some typical examples of users who might need spatial data analysis:

- crime analysts who have data about the locations of crimes and want to know whether incidents are clustered in some sense
- forestry researchers who have data on both the locations of trees and measured soil nutrient concentrations in a particular area and want to characterize how soil nutrients affect tree growth and shrub density
- geologists who have data about shallow earthquake locations and locations where fracking is performed and want to know whether fracking activity is associated with earthquake occurrence
- insurance analysts who have data from different zip codes and want to determine how variables such as crime rate and average precipitation influence house prices and home insurance risks
- ecologists who study the habitat of a certain species of plant or animal

All of these analysts have spatial data and are seeking answers to “spatial questions,” which can be answered by *spatial data analysis*.

### Data Encountered in Spatial Analysis

The goal of spatial data analysis is to derive insights from the location and context of real world phenomena such as crimes, accidents, vegetation growth, availability of minerals, and so on. Spatial analysts are often concerned with why things happen where they do. They are also concerned with how the occurrence of a phenomenon in a particular region affects nearby regions.

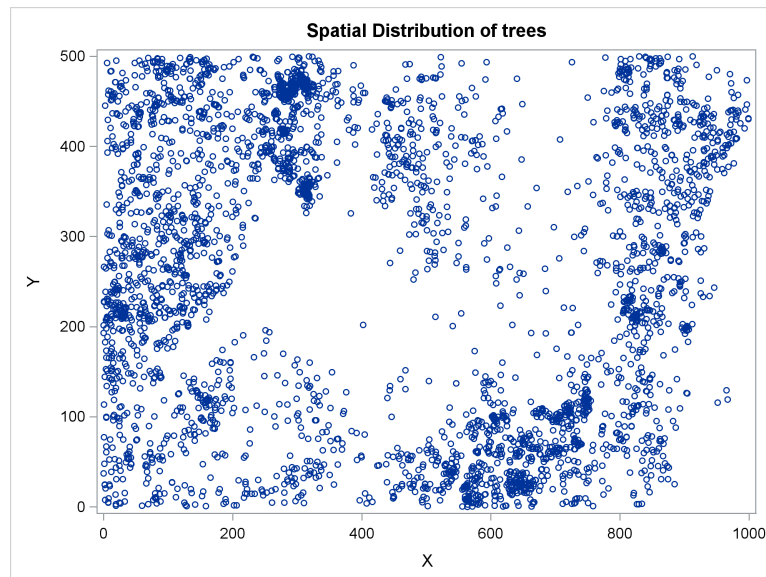
More precisely, the “where” in spatial data analysis usually refers to one of three possibilities:

- a set of randomly occurring discrete locations, such as places where a crime has been reported
- fixed places and their associated data, such as the measurements from sensors that detect pollutants or greenhouse gases
- a geographical region that has associated data, such as the average prices of homes in different counties

The first case is the subject of this paper, and it is called a *spatial point pattern*: a set of discrete spatial locations that can be viewed as one of possibly many realizations of an otherwise continuous and stochastic spatial phenomenon, such as crime likelihood or plant fertility. The second case is point-referenced data and the third is areal data. For more information about the latter two types of spatial data, see Banerjee, Carlin, and Gelfand (2004).

Consider Figure 1, which shows a collection of tree locations from a forest tree census in Panama. The data are part of a much larger data set, which contains the positions of hundreds of thousands of trees that belong to hundreds of species. The spatial point pattern data shown in Figure 1 corresponds to only one species.

**Figure 1** Spatial Point Pattern of Tropical Rain Forest Trees



If, in addition to the location of the trees, there is information about the species of the tree or other properties (such as tree height or diameter), the point pattern is called a *marked point pattern*. If the attribute is a categorical attribute (such as tree species type), the marked point pattern is also called a *multivariate point pattern*.

### Spatial Point Pattern Analysis

So, what is spatial point pattern analysis all about? Considering this forestry example, an analyst might want to answer the following questions:

- How does the pattern affect itself? Is there competition among trees for nutrients, causing tree growth to be either clustered or spread out?
- How does another spatial phenomenon affect the pattern? What effect do soil properties such as nutrient availability have on the spatial distribution of trees?

You can use PROC SPP to answer both of these types of questions.

### Overview and Scope

This paper reviews the basics of spatial point pattern analysis and modeling, provides a description of the features of PROC SPP, and provides examples to illustrate PROC SPP's capabilities. It discusses the types of questions PROC SPP can help you answer, and describes how it does so. A detailed theoretical treatment of spatial point pattern analysis or point process modeling in general is beyond the scope of this paper. For further reading on the methodological and theoretical aspects of analyzing and modeling spatial point patterns, see popular textbooks such as Illian et al. (2008) and Diggle (2013).

The next section of this paper reviews some fundamental concepts in spatial point processes and introduces different aspects of spatial point pattern analysis. The aspects include exploratory analysis, covariate testing, inhomogeneous Poisson process modeling, and residual diagnostics for model criticism and validation. The section "[CASE STUDIES](#)"

on page 16 describes two examples: one (from ecology) that uses a real data set and one (from criminology) that uses a simulated data set.

Although this paper relies heavily on a few examples, drawn in particular from crime and ecology research, these are by no means the only areas where spatial point patterns arise. Other areas that have spatial point patterns include the following, which are shown along with their relevant events of interest:

- insurance, studying where insurable incidents happen (for example, home break-ins)
- transportation safety, studying where vehicle accidents happen
- public health, studying where disease outbreaks happen
- energy, studying where fracking might be associated with shallow earthquakes

These areas are just a small selection. Any time you wonder why something happened where it did, you're asking a question about a spatial point pattern.

### Other SAS/STAT Procedures for Spatial Analysis

PROC SPP joins a suite of SAS/STAT procedures for analyzing point-referenced spatial data. The positions at which point-referenced spatial data are gathered are not random; in fact, they might be very carefully placed to measure some phenomenon accurately. For such data, you can use PROC VARIOGRAM to characterize the measured phenomenon, PROC KRIGE2D to estimate the phenomenon's behavior throughout the region of interest, and PROC SIM2D to simulate new realizations of the phenomenon.

## BASIC CONCEPTS

This section reviews the basic concepts of spatial point processes and provides an overview of what might be called the life cycle of spatial point pattern analysis and modeling. Along the way, you will encounter the following aspects of spatial point pattern analysis: summary statistics for exploratory understanding of a point pattern, testing whether a covariate influences deviation from an assumed model for the point pattern, fitting inhomogeneous Poisson process models, and finally diagnostics for model validation and criticism.

### Mathematical Definitions

Spatial analysis is its own branch of analytics, with its own vocabulary and not a little mathematics. Understanding the concepts in this section will help you see where PROC SPP fits in.

**What is the difference between a spatial point process and a spatial point pattern?** A *spatial point process* is a stochastic process,  $Z(s) : s \in D \subset \mathbb{R}^2$ , where the set  $D$  is random. Some examples of spatial point processes include the Poisson process, Cox process, and Gibbs process. A *spatial point pattern* is a realization of a spatial point process in  $\mathbb{R}^2$ , consisting of an arrangement of points in the random set  $D \subset \mathbb{R}^2$ . There can be multiple realizations (multiple spatial point patterns) of the same spatial point process.

**What are sampled and mapped point patterns?** Typically, spatial point patterns are viewed as bounded realizations of a spatial point process in  $\mathbb{R}^2$  within a spatial window  $W$ , which might be of any arbitrary shape. Theoretically, the point pattern can be observed beyond  $W$ . A spatial point pattern is called a *sampled point pattern* when all the event observations are recorded only within the window  $W$  and the spatial point pattern can be considered as partially observed. In such cases, the actual boundary of the spatial point process that generates the observed point pattern is unknown. Hence, the sampling window  $W$  might have to be estimated using techniques such as the Ripley-Rasson estimator (Ripley and Rasson 1977). On the other hand, a point pattern is called a *mapped point pattern* when the spatial point process that generates such a pattern operates entirely within a region bounded by  $W$  and the recorded events contain the entire observed realization. For example, the point pattern of trees shown in Figure 1 is a mapped point pattern that corresponds to a forest of trees from Panama. The data were collected as a part of a tropical forest experiment (Condit 1998; Hubbell and Foster 1983; Condit, Hubbell, and Foster 1996). Although most of the examples presented in this paper can be largely viewed as mapped point patterns, PROC SPP also supports sampled point patterns.

**What is the intensity of a spatial point pattern?** The intensity of a spatial point pattern,  $\lambda(s)$ , is defined as the number of events that are observed per unit of area.  $\lambda(s)$  is also called the *first-order intensity* and captures the mean of the spatial point pattern. A point pattern is said to be a realization of a *homogeneous spatial point process* if the first-order intensity in the random set  $D$  is observed to be same throughout every countable subset of  $D$ . However, if  $\lambda(s)$  varies spatially across different countable subsets of  $D$ , then the point pattern is said to be a realization of an *inhomogeneous spatial point process*.

**What is a homogeneous Poisson process?** A *homogeneous Poisson process* (HPP) is a spatial point process that produces a point pattern with a uniform first-order intensity within the random set  $D$ . That is, any two non-overlapping countable subsets of  $D$  have the same first-order intensity, and the numbers of events in these non-overlapping countable subsets are independent and Poisson distributed. Hence, events in an HPP are distributed uniformly and independently throughout the spatial domain that is represented by the random set  $D$ . Illian et al. (2008) refer to the latter property as the *independent scattering property* that makes an HPP completely random. For these reasons, these properties of an HPP are often collectively called *complete spatial randomness* (CSR).

**What is an inhomogeneous Poisson process?** HPPs are the exception rather than the rule among practical phenomena. Most spatial point patterns that are found in practical applications are not realizations of an HPP. There are several reasons why spatial point patterns might deviate from being a realization of an HPP. A common reason is that the first-order intensity is not homogeneous across different countable subsets of the random set  $D$ . In this situation, if the number of events in non-overlapping subregions of the region  $D$  are independent of each other and are Poisson distributed, such a spatial point pattern can be viewed as a realization of an *inhomogeneous Poisson process* (IPP).

**What are the second-order properties of spatial point patterns?** The Poisson process provides a mathematical foundation to construct models for the first-order intensity based on user-specified covariates. However, to capture the dependency between points, you need to characterize a spatial point pattern based on *second-order properties*.

Second-order properties characterize a spatial point pattern by accounting for the dependency between points. The first-order intensity corresponds to the mean function. Likewise, a function that captures the dependency between points in a point pattern is the second-order intensity function (Diggle 2013), also called the second-order product density (Illian et al. 2008). The second-order intensity function can be written as

$$\lambda_2(s_i, s_j) = \lim_{ds_i \rightarrow 0, ds_j \rightarrow 0} \left\{ \frac{\mathbb{E}[N(ds_i)N(ds_j)]}{|ds_i||ds_j|} \right\}$$

A quantity that is related to the second-order intensity is the conditional intensity, which measures the intensity at a particular location  $s_i$  conditional on the intensity at another location  $s_j$ . The conditional intensity can be written as

$$\lambda_c(s_i|s_j) = \lambda_2(s_i, s_j)/\lambda(s_j)$$

Depending on the nature of the second-order intensity, a spatial point process can be labeled stationary or non-stationary. A spatial point process is said to be stationary if the second-order intensity depends only on event location differences; that is,

$$\lambda_2(s_i, s_j) = \lambda_2^*(s_i - s_j)$$

Furthermore, a stationary point process is said to be isotropic if the second-order intensity is invariant to the direction of the point difference and depends only on the absolute distance:

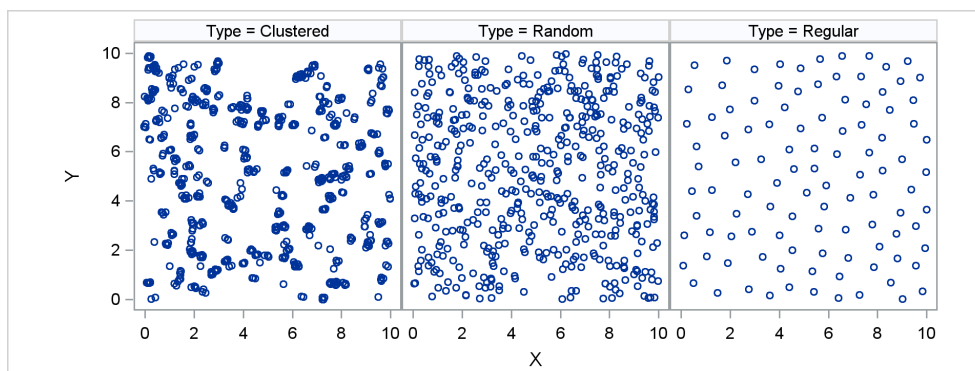
$$\lambda_2(s_i, s_j) = \lambda_2^*(||s_i - s_j||) = \lambda_2^*(h)$$

First and second-order properties allow you to qualitatively characterize spatial point patterns that are generated by spatial point processes.

## Qualitative Characterization of Spatial Point Patterns

Spatial point patterns can be qualitatively characterized as clustered, random, or regular (also called dispersed). The three simulated spatial point patterns in Figure 2 exhibit clustering, randomness, and regularity, respectively.

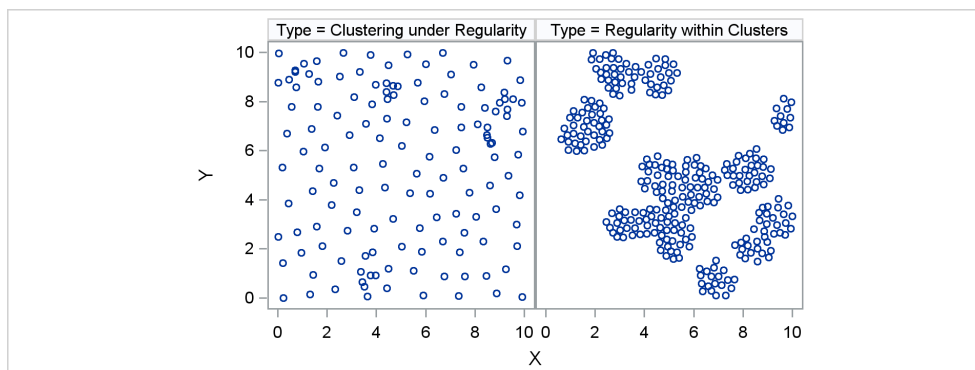
**Figure 2** Clustered, Random, and Regular Spatial Point Patterns



In the center plot of Figure 2, points are scattered haphazardly around the region with a constant intensity, and the position of any one point is independent of the position of all the other points. This is a *completely random pattern*, a realization of a homogeneous point process (HPP) that exhibits complete spatial randomness (CSR). In the left plot of Figure 2, the clustered pattern, you are more likely to find points near other points. In other words, the average distance from any arbitrary point to its nearest-neighbor is less than you'd expect in a completely random pattern. On average, points are closer to their nearest-neighbor than they are in the middle plot. On the other hand, in the right plot, the regular pattern, points are farther away from their nearest neighbors. The section “[Stage 1: Exploration Stage](#)” on page 6 shows how to make these qualitative statements about nearest-neighbor distances precise by using nearest-neighbor and second-order summary statistics.

It is important to understand that clustering and regularity are not just characteristics of a pattern as a whole, but rather characteristics of a pattern at a specific distance scale. The same pattern can exhibit both clustering and regularity at the same time, at different distance scales. In fact, many real-life spatial point patterns behave in just this way. For example, seedlings can be clustered at large-length scales depending on how their parent trees were spaced, but regular on the small scale because they compete for resources. Figure 3 shows two simulated spatial point patterns, each of which exhibits both clustering and regularity in different ways. The left plot shows long-range regularity with short-range clustering, whereas the right plot shows the reverse.

**Figure 3** Combination of Clustering and Regularity in Spatial Point Patterns



## SPATIAL POINT PATTERN ANALYSIS LIFE CYCLE

It is best if you think of spatial point pattern analysis as having a “life cycle,” which consists of the following stages:

1. Before you can start to fit models for a spatial point pattern's intensity, you might want to perform some preliminary investigation of the spatial point pattern and characterize it qualitatively. You usually do this by

exploring the spatial point pattern and understanding whether the pattern is *clustered*, *regular*, or *random*. To identify the type of pattern, you compute empirical summary statistics (which are functions of inter-point distance) and compare them to theoretical summary statistics. In addition to determining the nature of the point pattern, you can visualize the spatial intensity of the point pattern, usually via nonparametric intensity estimation.

2. If covariates are available in the data that you have gathered, you might want to explore possible relationship between the covariate and the first-order intensity of the spatial point pattern to help explain the nature of the intensity. Covariate dependency testing enables you to accomplish this.
3. After you have understood the available covariates and tested them for a possible influence on the first-order intensity of the spatial point pattern, you can construct log-linear Poisson models for the point pattern's first-order intensity and obtain a fitted intensity estimate. Poisson modeling helps pinpoint how the covariates act and suggests the extent and significance of influence each covariate would have on the first-order intensity.
4. Finally, residual diagnostics reveal whether there are other covariates you might have missed. These residual diagnostics determine whether the chosen covariates and other model components adequately explain the first-order intensity of the point pattern. Residual plots can reveal variables that might be influencing the point pattern but haven't been measured or included in the model.

The remainder of this section discusses these stages of the SPP life cycle in more detail.

## Stage 1: Exploration Stage

Analysis and modeling of spatial point patterns usually begins with exploring the nature of the point pattern and answering a simple question: "Is the point pattern random, clustered, or regular?"

**How do you determine whether a spatial point pattern is random, clustered, or regular?** You can determine whether a spatial point pattern is random, clustered, or regular by examining plots of summary statistics that are computed as a function of the distance between points. There are two classes of summary statistics for diagnosing clustering, regularity, and randomness: statistics that are based on nearest-neighbor distance (Diggle 1979, 2013; Van Lieshout and Baddeley 1996; Baddeley et al. 2000) and statistics that are based on second-order properties (Ripley 1988, 1976).

PROC SPP produces a variety of such plots for the summary statistics. In each plot, the idea is to compare the shape of the empirical distribution function (EDF) of the statistic to the shape that you would expect under the CSR (complete spatial randomness) assumption. Patterns that are clustered or regular relative to CSR will have EDFs that differ systematically from what is expected under CSR. PROC SPP also adds pointwise simulation envelopes to this plot, to help you judge statistically significant departures from CSR. You request the analysis that produces these plots by specifying the appropriate options in the PROCESS statement that defines the point pattern you're studying, as in the following statements:

```
proc spp data=sashelp.bei;
    process p = (x,y / event=Trees) / F G J K L PCF;
run;
```

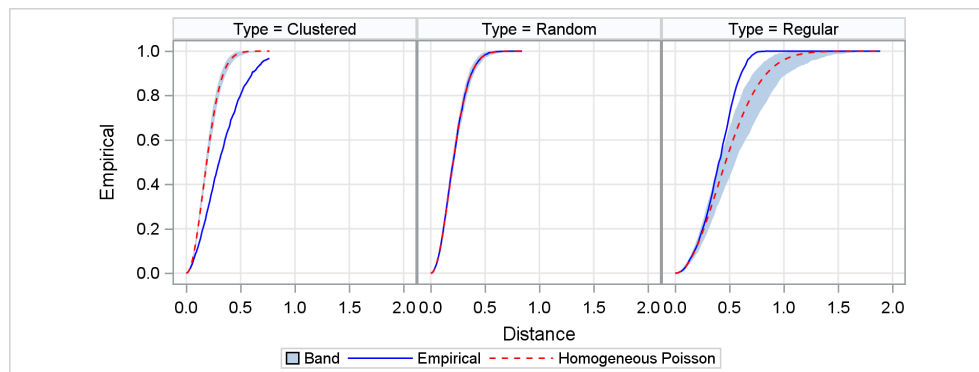
In the preceding call to PROC SPP, the F, G, and J options in the PROCESS statement produce summary statistics that are based on nearest-neighbor distance, and the K, L, and PCF options produce summary statistics that are based on second-order properties. The statistics produced by the F, G, J, K, L, and PCF options are discussed in the following questions.

**What statistics are based on nearest-neighbor distances?** Statistics that are based on nearest-neighbor distances represent the empirical distribution of the number of pairs of nearest-neighbors within particular distance ranges. Depending on the statistic, the distances can be evaluated with respect to other events or with respect to a predefined grid. An example of the former is the *nearest-neighbor G-function* (Diggle 1979, 2013); an example of the latter is the *empty-space F-function* (Diggle 1979, 2013). The F-function and the G-function represent probabilities of finding an event of the point pattern within distance  $r$  of a fixed arbitrary point and a random point, respectively. A third statistic, the J-function, combines the F-function and the G-function; it is the ratio of the probabilities of not finding any points within distance  $r$  from any random point and a fixed arbitrary point (Van Lieshout and Baddeley 1996).

You use the three nearest-neighbor functions, F, G, and J, to assess clustering or regularity by comparing their EDFs (which are based on the observed points) to what you would expect from a pattern that is a realization of CSR. Adding pointwise Monte Carlo envelopes (Baddeley et al. 2014) for HPP realizations help make any diagnosis of departure from CSR statistically valid.

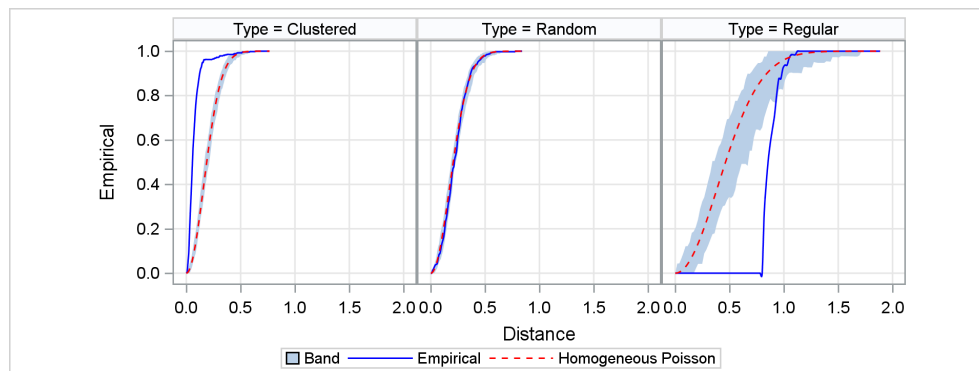
For the empty-space F-function, values of  $\hat{F}(r) > F_{CSR}(r)$  indicate regularity, and values of  $\hat{F}(r) < F_{CSR}(r)$  indicate clustering, with statistical significance if the function is outside the simulated envelopes. For example, Figure 4 shows the F-function that is computed on the clustered, random, and regular spatial point patterns that are shown in Figure 2. You can clearly see how the empirical F-function together with its simulation envelopes helps distinguish the three types of point patterns.

**Figure 4** Empty Space F-Function for Clustered, Random, and Regular Point Patterns



You can use the nearest-neighbor G-function in a similar way. In this case, clustering is indicated by  $\hat{G}(r) > G_{CSR}(r)$  and regularity by  $\hat{G}(r) < G_{CSR}(r)$ , as shown in Figure 5.

**Figure 5** Nearest Neighbor G-Function for Clustered, Random, and Regular Point Patterns

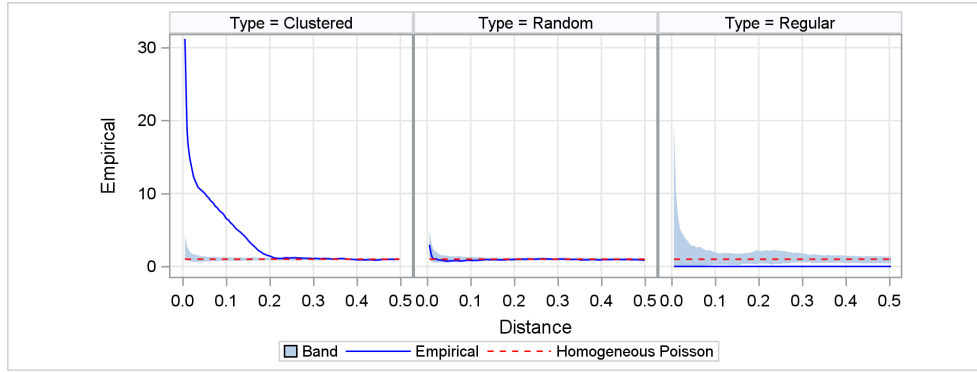


The J-function has a theoretical value of 1 for a realization of the HPP. Values of the empirical J-function,  $\hat{J}(r)$ , indicate regularity when they are greater than 1 and indicate clustering when they are less than 1 (Van Lieshout and Baddeley 1996).

**What statistics are based on second-order properties?** The second-order conditional intensity provides insight into the second-order interactions between pairs of points, based on their interpoint separation. As with the nearest-neighbor distance functions, the second-order summary statistics can be useful in diagnosing clustering or regularity. Second-order distance functions include the K-function (Ripley 1976), the L-function (Besag 1977), and the pair correlation function (PCF) (Stoyan and Stoyan 1994). For example, Figure 6 shows the behavior of the pair correlation for patterns that are shown in Figure 2.



**Figure 6** Pair Correlation Function for Clustered, Random, and Regular Point Patterns



The second-order summary statistics reveal clustering or attraction if the EDF is greater than the theoretical EDF for a CSR process, and significantly so if they are outside the pointwise Monte Carlo envelopes. Similarly, they reveal regularity if the EDF is less than the theoretical function and is outside the pointwise Monte Carlo envelopes. The second-order K, L, and PCF (denoted  $g(r)$ ) functions differ in the choice of the theoretical function as follows:

$$\begin{aligned} K_{CSR} &= \pi r^2 \\ L_{CSR} &= r \\ g_{CSR}(r) &= 1 \end{aligned}$$

The mechanism of interpreting clustering and regularity for the pair correlation function  $g(r)$  is the same as that of the K-function and L-function. Figure 6 show the plots of the pair correlation function. You can clearly see in Figure 6 that the pair correlation function can distinguish the three types of data sets (clustered, random, and regular).

**Cautionary points about using summary statistics:** You need to exercise some caution while exploring your point pattern and interpreting different distance-based summary statistics. Here are some points to consider:

- Even if you take the Monte Carlo simulation envelopes into account, the inferences you make on the basis of these exploratory plots don't constitute formal statistical tests. Such tests have been discussed (Baddeley et al. 2014), but they are not yet available in PROC SPP. Actually, the innocent-until-proven-guilty nature of traditional statistical testing might not often be appropriate for spatial analysis, at least with CSR as a null model. The goodness-of-fit testing in PROC SPP is an area where alternative null models do make significance testing more useful.
- Distance functions merely reveal scales (interpoint or nearest-neighbor distances) at which clustering or regularity are likely to occur. Such a scale is obtained as the distance at which the function sees the maximum deviation from the theoretical curve as shown in the plots of different distance functions. You might find this scale (distance) useful for further analysis of the point pattern or when you study interactions with covariates. However, you should be cautious in attaching any formal scientific meaning to this distance because these distance function plots are not formal inferential tools.
- All distance functions are influenced by edge effects. Edge effects arise because actual spatial point patterns are observed on a study region  $W$  that has fixed dimensions. There might be unobserved points outside the sampling window  $W$ , and these unobserved events potentially interact with events within the study window  $W$ , especially on the edge. Neglecting to somehow take these conditions into account leads to bias in the estimators of the summary statistics. Several types of edge-correction schemes (Stoyan and Stoyan 1994, pp. 280–283) have been explored. The simplest edge correction is border correction (also called minus-sampling), as mentioned by Stoyan and Stoyan (1994, p. 280). PROC SPP implements the border method for all the distance functions.



## Stage 2: Covariate Testing

Often, you collect covariates in the area where you have observed the spatial point pattern and you want to relate these covariates to the intensity. For example, in the forestry application, which has a spatial point pattern of trees, the data set includes covariates that pertain to soil properties (nutrient content, pH, and so on) and topology (elevation, slope, and so on). Testing for overall covariate intensity effects can suggest more detailed intensity models.

Covariate testing usually comes in the second stage of the spatial point pattern analysis life cycle. In the first stage, you got the overall lay of the land, as it were, with your data, determining its general structure—clustered, regular, or random. Now you want to know why it has that structure.

You request covariate testing in PROC SPP by first specifying the PROCESS statement to define the point pattern being studied, then one or more TREND statements to define the covariate trends that are potentially related to the point pattern, and finally the COVTEST statement to perform the covariate testing. The COVTEST statement has a syntax similar to the MODEL statements in other SAS/STAT procedures and in PROC SPP itself. The following statements show an example:

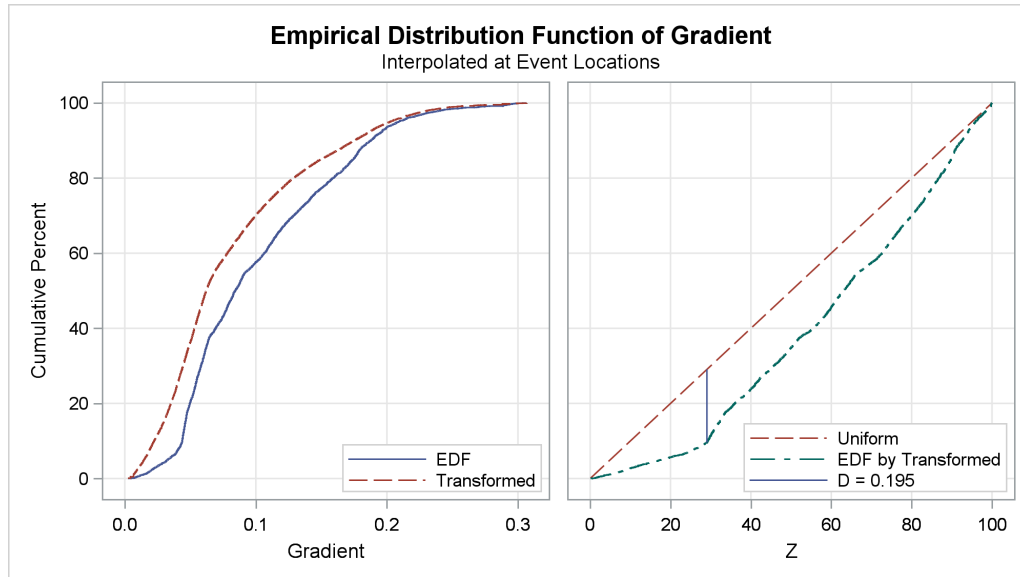
```
proc spp data=sashelp.bei plots=all;
  process trees = (x,y / event=Trees area=(0,0,1000,500));
  trend gradient = field(x,y,gradient);
  trend elevation = field(x,y,elevation);
  trend ph = field(x,y,pH);
  covtest trees = gradient elevation ph;
run;
```

Covariate testing in PROC SPP uses the empirical distribution function (EDF). In particular, PROC SPP provides weighted EDF tests that compute the Kolmogorov-Smirnov  $D$  statistic and the Cramér–von Mises  $W^2$  statistics. As with diagnostics that are based on distance functions, weighted EDF tests are performed with the assumption that the point process that generates the point pattern is an HPP.

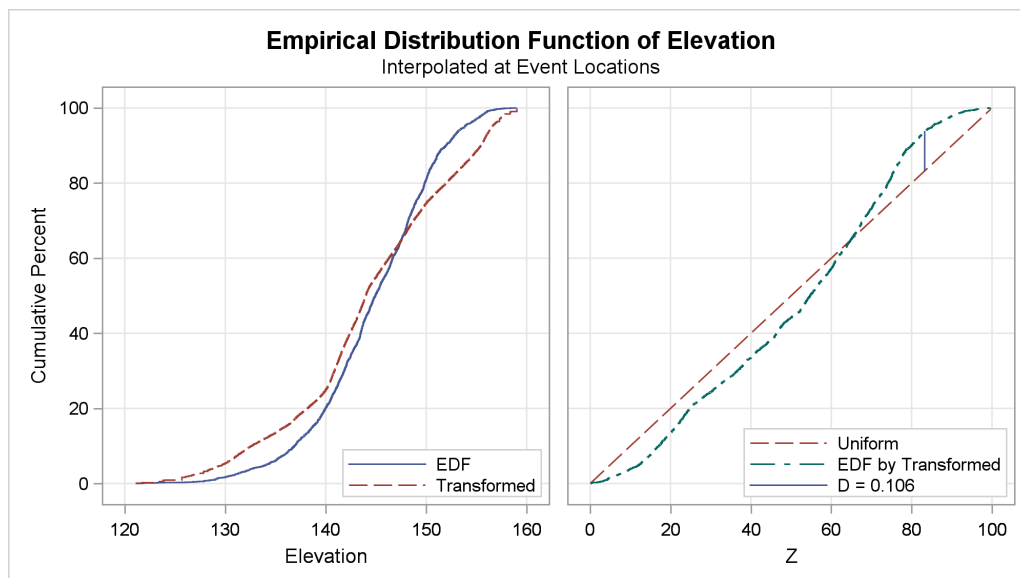
At the heart of the covariate EDF test is the computation of an intensity-weighted EDF of the covariate, as defined in D'Agostino and Stephens (1986). For this approach, the covariate is first interpolated at the event locations in the spatial point pattern. From these interpolated values, a weighted EDF for the covariate is computed by assuming that the underlying spatial point pattern is a realization of the HPP. The SPP procedure uses this weighted EDF to assess goodness of fit by computing a transformed EDF under a probability integral transform (PIT). If the transformed EDF is not uniformly distributed, then you can infer that the underlying model of intensity is not correct and that the point pattern might be influenced by the covariate. You can graphically judge uniformity from the EDF test plot that PROC SPP produces, and you can also make a formal statistical test of uniformity that is based on the Kolmogorov-Smirnov  $D$  statistic or the Cramér–von Mises  $W^2$  statistic.

For example, the preceding PROC SPP statements produce EDF test plots that are shown in [Figure 7](#), [Figure 8](#), and [Figure 9](#) for the three covariates, **Gradient**, **Elevation**, and **pH**, respectively. The plot for **Gradient** shows a definite systematic deviation from uniformity, the plot for **Elevation** less so, and the plot for **pH** even less.

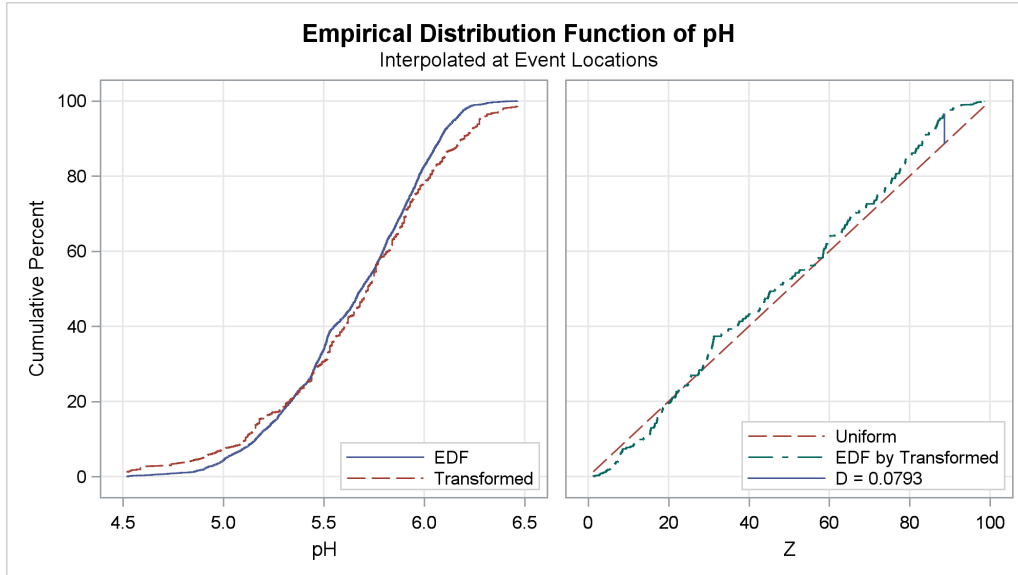
**Figure 7** EDF Test Plot for **Gradient**



**Figure 8** EDF Test Plot for **Elevation**



**Figure 9** EDF Test Plot for pH



Nevertheless, the formal test, which is based on the Kolmogorov-Smirnov  $D$  statistic and shown in Figure 10, indicates that all three covariates have statistically significant effects on the intensity of tree growth.

**Figure 10** Weighted EDF Test Statistic

Kolmogorov-Smirnov Weighted EDF Test for Covariate Values		
D		
Source	Statistic	p Value
Gradient	0.194805	<0.0010
Elevation	0.106492	<0.0010
pH	0.079341	<0.0010

The Kolmogorov-Smirnov  $D$  statistic is the maximum distance between the EDF plot and the transformed EDF plot, as depicted in the EDF test plots Figure 7, Figure 8 and, Figure 9.

### Stage 3: Intensity Modeling

After exploratory analysis and covariate testing, the third stage in the spatial point pattern analysis life cycle is to model the first-order intensity, the rate at which points are expected to occur in a particular region of your study area. You can view this stage as an extension of covariate testing: having determined that certain covariates indicate a departure from an HPP or CSR intensity model, you want to study precisely whether and how they affect the first-order intensity of the point pattern. This is the point at which parametric assumptions are introduced for the nature of the point process that generates the point pattern.

A common assumption is that the underlying point pattern comes from an inhomogeneous Poisson process (IPP) that has a first-order intensity and is a function of the covariates. Under this assumption, you can use the log-linear modeling facilities in PROC SPP to estimate this intensity function. The fitted parametric model for intensity helps you identify the level of influence that each covariate has on the first-order intensity of the spatial point pattern, and it also identifies the statistical significance of that level of influence.

PROC SPP fits parametric models for the first-order intensity, where the first-order intensity function,  $\lambda_\theta(s)$ , is assumed to be log linear in the parameters  $\theta$ . The log-linear function is of the form

$$\lambda_\theta(s) = \theta \cdot Z(s)$$

where  $Z(s)$  is a vector-valued function of location  $s$ , which can include polynomials in the coordinate variables. For the inhomogeneous Poisson process model, the exact log likelihood is

$$\log L(\theta; x) = \sum_{i=1}^n \log \lambda_{\theta}(x_i) - \int_W \lambda_{\theta}(s) ds$$

PROC SPP maximizes an approximation of this function, using quadrature to evaluate the integral. This approximation leads to a log likelihood of the form

$$\log L(\theta; x) \approx \sum_{j=1}^m (y_j \log \lambda_j - \lambda_j) w_j$$

where  $\lambda_j = \lambda_{\theta}(s_j)$  and the list of points  $\{s_j, j = 1, \dots, m\}$  includes both data points and grid points. Log likelihoods of this form show up in several SAS/STAT procedures for analyzing generalized linear models, such as the GENMOD, GLIMMIX, and FMM procedures.

The following statements show you how to specify a parametric intensity model in PROC SPP:

```
proc spp data=sashelp.bei plots=all;
  process trees = (x,y / event=Trees area=(0,0,1000,500));
  trend gradient = field(x,y,gradient);
  trend elevation = field(x,y,elevation);
  trend pH = field(x,y,pH);
  model trees = elevation gradient pH;
run;
```

A parametric intensity model is specified by using the MODEL statement, which includes the covariates on the right side of the equal sign and the point pattern on the left side. The modeling process produces both a table of parameter estimates and a fit summary, as shown in [Figure 11](#).

**Figure 11** Parameter Estimates Table

### The SPP Procedure

Poisson Parameter Estimates				
Parameter	Estimate	Standard Error	z Value	Approx Pr >  z
Intercept	-6.7420	0.3859	-17.47	<.0001
Elevation	0.02856	0.002473	11.55	<.0001
Gradient	6.6207	0.2672	24.77	<.0001
pH	-0.5166	0.05048	-10.23	<.0001

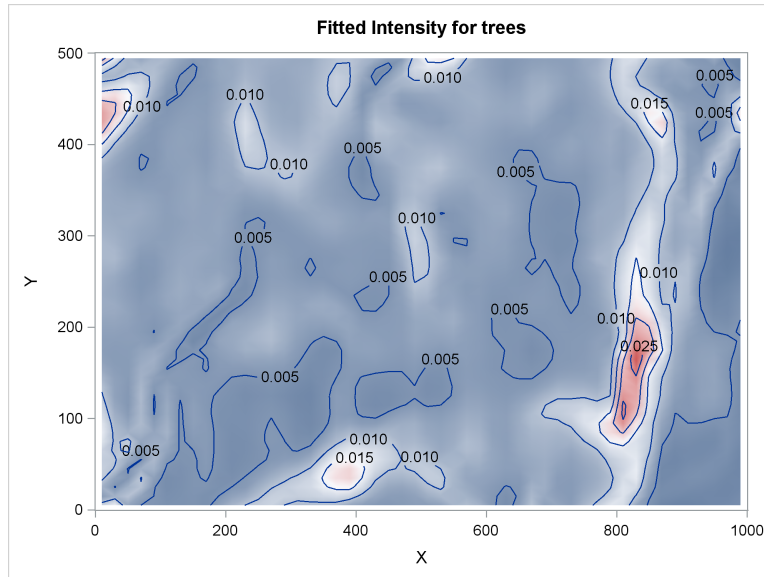
  

Fit Statistics	
Criterion	Value
-2 Log Likelihood	42203.4
AIC (smaller is better)	42211.4
BIC (smaller is better)	42238.3

The parameter estimates show (on a logarithmic scale) how each model effect, including the intercept, influences the point process intensity. In this case, all three covariates are significant, but the  $z$  value for **Gradient** is about twice as large as it is for the other two covariates, indicating that Gradient is more important in explaining where trees grow.

[Figure 12](#) shows a heat map of the fitted intensity with overlaid contours to indicate the response surface of the fitted model.

**Figure 12** Intensity Estimates of Tree Presence in Study Area



#### Stage 4: Model Assessment and Validation

In the third stage of the spatial point pattern analysis life cycle, you fit a particular intensity model that uses covariate information to explain the point pattern. The final stage in the life cycle is to examine closely how well this intensity model accounts for the point pattern, either to decide that the model fits well and that the underlying phenomenon is fully accounted for by the covariates, or to decide that you need to gather other covariates or revisit your assumptions. The SPP procedure provides two ways to explore how well your intensity model fits the data: a formal goodness-of-fit test and more general residual diagnostics.

**Goodness-of-Fit Testing:** The formal goodness-of-fit test addresses the following question: “How likely is it that this spatial point pattern was generated by this fitted intensity model?” To answer this question, PROC SPP uses simulation to generate realizations of the inhomogeneous Poisson process under the fitted model and to assess the size of their Pearson residuals relative to the residuals that are observed from the actual data. The Pearson residuals are defined as

$$r_P = \frac{O_c - E_c}{\sqrt{E_c}}$$

where  $O_c$  is the observed count in each quadrat (which is based on the data) and  $E_c$  is the expected count in each quadrat (which is computed as the average of counts over the simulation). Based on these observed and expected counts, a  $\chi^2$  statistic is computed and a Pearson  $\chi^2$  test is performed. A small  $p$ -value indicates that the data are not likely to be generated by the model.

The following statements show how to use PROC SPP to fit a model and request a simulation-based goodness-of-fit test for the tree data set BEI:

```
proc spp data=sashelp.bei plots=all;
  process trees = (x,y / event=Trees area=(0,0,1000,500))
    / quadrat(4,4 /details);
  trend gradient = field(x,y,gradient);
  trend elevation = field(x,y,elevation);
  trend pH = field(x,y,pH);
  model trees = elevation gradient pH / gof;
run;
```

To perform a goodness-of-fit analysis, you use the GOF option in the MODEL statement. [Figure 13](#) shows the Pearson  $\chi^2$  test for goodness of fit that is based on the Pearson residuals. The test has a very small  $p$ -value, which indicates that you might need more than just the covariates **Elevation**, **Gradient**, and **pH** to explain the observed data.

**Figure 13** Goodness of Fit for Model with **Elevation**, **Gradient**, and **pH****The SPP Procedure**

Pearson Chi-Square Test for Goodness-of-Fit			
Dispersion			
DF	Index	Chi-Square	Pr > ChiSq
15	105.33	1579.92	<.0001

If the goodness of fit reveals that the model does not do a good job of explaining the observed spatial point pattern, then your model might be missing some important effects. Using PROC SPP, you can examine further residual diagnostics to help you pinpoint these effects, as discussed in the next section.

**Residual Diagnostics:** Residual diagnostics help you understand whether the model is adequate to explain the observed spatial point pattern. Close inspection of the residual diagnostics is particularly important if goodness of fit reveals that your intensity model does not do a good job of accounting for the actual data. You might be able to identify additional covariates and incorporate them to improve your intensity model.

There are several different types of residual diagnostics in spatial point pattern analysis. However, PROC SPP supports only the raw residuals. The raw residuals over a region  $W$  are defined as

$$r_{\hat{\theta}}(W) = n(x \cap W) - \int_W \hat{\lambda}(s, x) ds$$

where  $\hat{\lambda}$  is the estimated intensity, which depends log-linearly on the estimated parameters  $\hat{\theta}$ . As suggested by Baddeley et al. (2005), the discretized version of the raw residuals is

$$r_j = z_j - w_j \lambda_j$$

at the quadrature points,  $u_j$ , where  $z_j$  is an indicator equal to 1 if  $u_j$  is a data point or equal to 0 if  $u_j$  is a dummy point,  $w_j$  is the quadrature weight that is attached to  $u_j$ , and  $\lambda_j = \hat{\lambda}(u_j, x)$  is the conditional intensity of the fitted model at  $u_j$ .

To diagnose whether there is a hidden trend that the model has not accounted for, you look at the smoothed residual plot. Residual plots are analogous to the plots computed by PROC GENMOD to assess covariate misspecification. The smoothed residual plot is a spatial residual intensity plot that shows a smooth spatial surface computed as

$$s(u) = \hat{\lambda}(u) - \tilde{\lambda}(u)$$

where  $\hat{\lambda}(u)$  is a nonparametric kernel estimate of the intensity and  $\tilde{\lambda}(u)$  is a smoothed version of the parametric estimate of the intensity according to the fitted model.  $\hat{\lambda}(u)$  can be written as

$$\hat{\lambda}(u) = e(u) \sum_{i=1}^{n(x)} k(u - x_i)$$

where,  $e(u)$  is an edge correction. Similarly,  $\tilde{\lambda}(u)$  can be written as

$$\tilde{\lambda}(u) = e(u) \int_W k(u - s) \lambda_{\hat{\theta}}(s) ds$$

where,  $\lambda_{\hat{\theta}}(s)$  is the fitted estimate of the first-order intensity based on a model.

If the fitted model is correct, the kernel estimate and the kernel smoothed estimate of the fitted intensity should be approximately equal. Positive values of  $s(u)$  suggest that the model underestimates the intensity at those locations, and negative values suggest that the model overestimates the intensity at those locations (Baddeley and Turner 2005).

The following statements request residuals for the model of tree intensity that involves the covariates **Gradient**, **Elevation**, and **pH**. Residual analysis is requested by the RESIDUAL option in the MODEL statement, which includes a bandwidth value (B=70) for computing the nonparametric kernel estimate of the intensity.

```

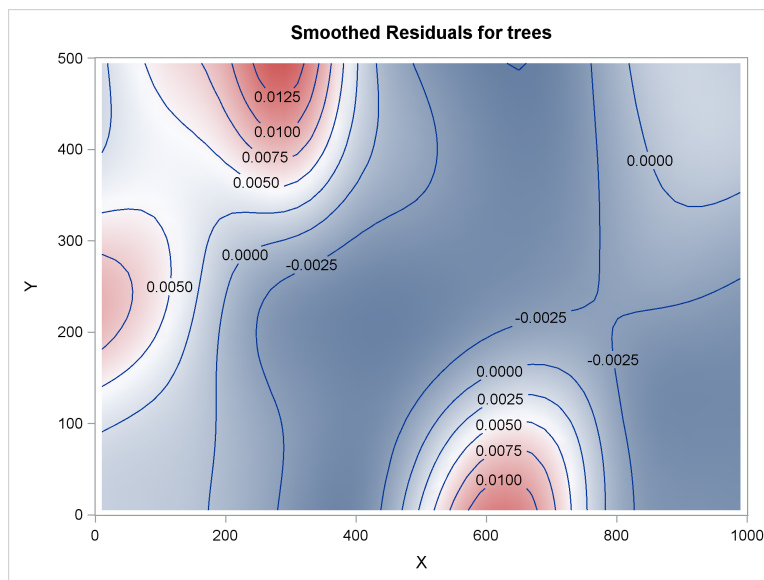
proc spp data=sashelp.bei plots(unpack)=all;
  process trees = (x,y / event=Trees area=(0,0,1000,500));
  trend gradient = field(x,y,gradient);
  trend elevation = field(x,y,elevation);
  trend pH = field(x,y,pH);
  model trees = elevation gradient pH /residual(b=70);
run;

```

Figure 14 shows the smoothed residual plot of the raw residuals. This plot clearly reveals a trend in the residuals that shows areas where the model might be misspecified. For example, in the top left, where  $X = 200$  and  $Y$  is between 400 and 500, the specified model (which involves the covariates) underestimates the tree distribution. The same effect is indicated in the center left, where  $Y$  is between 200 and 300, and also toward the bottom right, where  $X = 600$ .

These positive trends in the smoothed residual plots actually indicate that new covariates might be needed to account for the observed intensity in these areas.

**Figure 14** Smoothed Residual Diagnostic Plot



What if, after all, there are no covariates that you can use to ameliorate trends in the residuals? Trends unaccounted for by the model can indicate several things. Maybe you need to collect more data to find appropriate covariates. Alternatively, the underlying assumption of an inhomogeneous Poisson process might be questionable. In the former case, the domain scientist needs to study the area closely to look for possibly newer explanatory factors. In the latter case, an inhomogeneous Poisson model assumption for the underlying point pattern needs to be revisited.

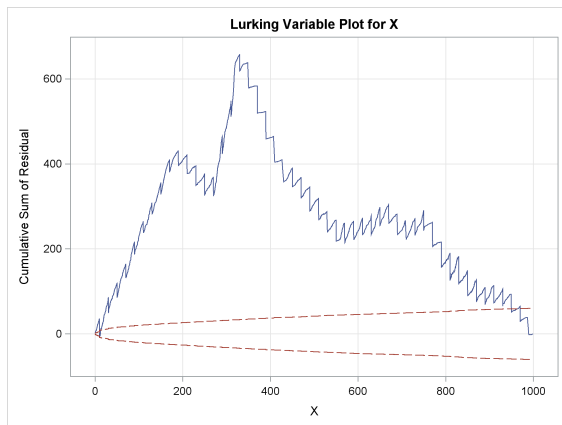
**Lurking Variable Plots:** You use a lurking variable plot to determine whether a potential covariate influences your point pattern and thus should be included in your intensity model. PROC SPP generates lurking variable plots by default for all model covariates and coordinate variables.

Figure 15 shows the lurking variable plot for the coordinate variable  $X$ . The dashed red lines are the standard deviation intervals, and the blue continuous line is the cumulative sum of raw residuals. In this case, the cumulative sum of residuals deviates from the standard deviation intervals, indicating that  $X$  does influence the tree distribution in some way that is not captured by the specified model.

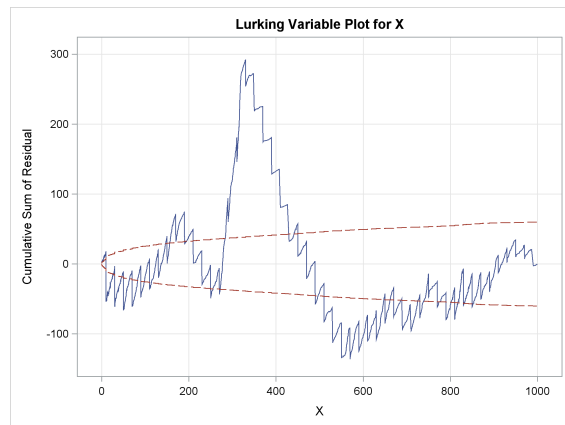
To fit an intensity model that captures this lurking effect, you can include a polynomial trend in the model by using the  $POLY(n)$  option in the MODEL statement, where  $n$  specifies the degree of the polynomial. The lurking variable plot for  $X$  in this new model, shown in Figure 16, certainly deviates much less from its standard error bars, although there still might be significant unexplained features of the pattern around  $X=300$ .



**Figure 15** No X Trend in Model



**Figure 16** Polynomial X Trend in Model



## CASE STUDIES

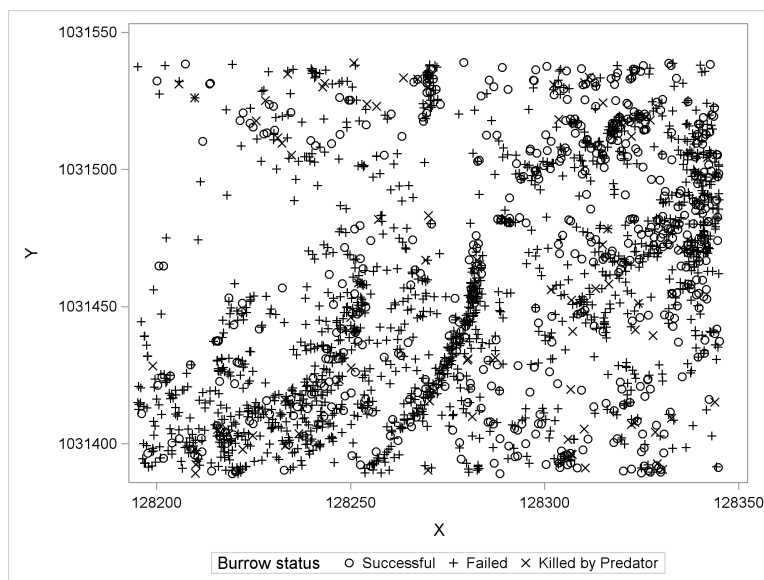
This section demonstrates how you can use PROC SPP to answer analysis questions in practical applications such as ecology and criminology. Although the case studies are compressed, you can easily visualize PROC SPP's analytical life cycle being applicable to these cases as well.

### Case Study 1: Modeling Isopod Burrows

The data for this example are from Banerjee and Gelfand (2002) and are included in the SasHELP library as SasHELP.Burrows.

Suppose you are an entomologist who is studying isopods ("pill bugs") in the Negev desert in Israel. Isopods are a diverse and important component of many different habitats, and they can be a bellwether for overall environmental health. You have data, shown in Figure 17, on the location of various types of isopod burrows in a particular study region.

**Figure 17** Locations of Isopod Burrows



For each burrow, you know whether it still contains live isopods ("Successful"), has been emptied by an isopod predator, or has failed for some other environmental reason. Different types of burrow are displayed with different markers in Figure 17.

You also have data about certain covariates, which are measured on a grid of one-square-meter units over the

study region. These covariates include topographical characteristics (such as elevation, aspect, and slope) and environmental characteristics (such as the proportion of rocks and shrubs as well as the amount of dew in each square-meter unit of the study area). You think that these covariates might affect either the environmental viability of isopod burrows or the range of potential isopod predators.

The following statements analyze the burrow intensity model for each type of burrow separately:

```
proc spp data=sashelp.burrows plots(unpack)=all;
  where (status in (1, .));
  process burrows=(X, Y /Event=Burrow)
    / quadrat(5,5 / details);
  trend elevation= field(X,Y,Elevation  );
  trend aspect   = field(X,Y,Aspect    );
  trend slope    = field(X,Y,Slope      );
  trend dew      = field(X,Y,Dew        );
  trend rock     = field(X,Y,PercentRock);
  trend shrub    = field(X,Y,PercentShrub);
  model burrows = elevation aspect slope dew rock shrub
    / gof residual(b=10);
run;
```

This code analyzes the sample data set Sashelp.Burrows, using a WHERE statement to restrict attention only to successful burrows (**Status=1**), together with the covariates (**Status=.**). You can use similar code with **Status=0** or **Status=-1** instead of **Status=1** to instead analyze failed and preyed-upon burrows, respectively.

One way to summarize the results of the model fit is to examine just the  $p$ -values for the covariates, as shown in Table 1. These results indicate that although several covariates affect the location of successful burrows and especially of failed burrows, they are not significant for the location of burrows that have been preyed upon, and thus presumably not for the ranges of predator activity.

**Table 1** Analysis of Intensity Model of Covariates for Burrows

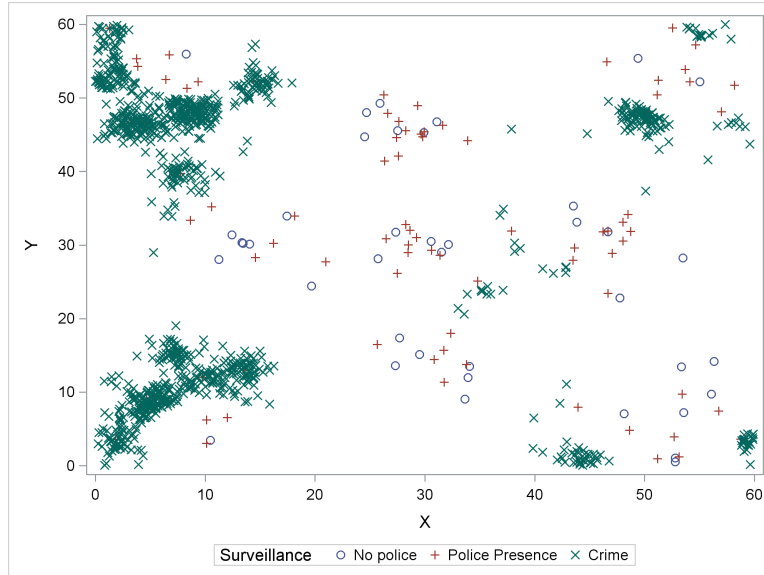
Parameter	Successful	Failed	Preyed upon
Intercept	0.9302	<.0001	0.0320
Elevation	0.7134	<.0001	0.0437
Aspect	<.0001	0.0055	0.8891
Slope	0.0210	<.0001	0.9062
Dew	0.3463	<.0001	0.5548
PercentRock	0.9515	<.0001	0.6487
PercentShrub	<.0001	<.0001	0.5183

## Example 2: Does Police Presence Help Reduce Crime?

The use of spatial analysis tools to study crime has been an active area of recent research (Levine 2014; Leitner 2013). This example uses simulated crime report data to demonstrate the types of analyses that can provide actionable inferences.

Suppose you are a crime analyst who is studying how the placement of police near bars affects crime. Working with law enforcement in your city, you have gathered data on where the bars are, where the police are (that is, which bars have dedicated police surveillance), and where the crimes are. These data are shown in Figure 18, where crimes are depicted by an  $\times$ , bars with police surveillance are depicted by circles, and bars without police surveillance are depicted by plus signs.

**Figure 18** Bar locations and Crime Incidents



Notice that crime events seem to be reported around bars that do not have any police presence, and not around bars that do. To analyze this relationship more precisely, first append to the data the minimum distances to each type of bar, **dUnwatchedBar** and **dWatchedBar**, for each point in a grid over the study area. This is easiest to do by using SAS/IML function like the following:

```
start mindist(a,b);          /* Minimum distances from B to A */
  free d;
  do i = 1 to nrow(a);
    d = d // min(sqrt( (shape(a[i,],nrow(b),ncol(b)) - b[,##] ) ));
  end;
  return(d);
finish;
```

You can use these distances with PROC SPP to analyze how proximity to a bar influences the intensity of crimes, as shown in the following statements:

```
proc spp data=Crimes plots=all;
  process Crimes = (X , Y / area=(0,0,60,60) Event=Event);
  trend dUnwatchedBar = field(X,Y,dUnwatchedBar);
  trend dWatchedBar = field(X,Y,dWatchedBar );
  model Crimes = dUnwatchedBar dWatchedBar / residual(b=10);
run;
```

These statements first define both the **Crimes** point process and two covariate trends that are based on the minimum distance to the two types of bar, and then model the intensity of **Crimes** as a function of these distances. Figure 19 shows the parameter estimates for the intensity model.

**Figure 19** Parameter Estimates Table for the Intensity of Crimes

### The SPP Procedure

Poisson Parameter Estimates				
Parameter	Estimate	Standard Error	z Value	Approx Pr >  z
Intercept	-0.3789	0.1008	-3.76	0.0002
dUnwatchedBar	-1.5615	0.04533	-34.45	<.0001
dWatchedBar	0.4657	0.01061	43.90	<.0001

The estimated parameters tell you what happens to the intensity as the respective distances increase; in other words, they tell you the *opposite* of what happens as you get closer to each type of bars. So the interpretation is that the crime rate increases as you get closer to an unwatched bar at about three times the rate that it decreases as you get closer to a bar with police surveillance (this is indeed precisely the effects that the simulation was constructed to have).

## CONCLUSION

The SPP procedure, introduced in SAS/STAT 13.2, expands the spatial analysis capabilities in SAS/STAT software to enable you to study spatial point patterns, which are commonplace in many practical applications. PROC SPP enables you to take your data through the spatial point pattern analysis life cycle, including exploration, covariate testing, intensity modeling, and model assessment. The final stage in this life cycle leads directly to more understanding about how your point pattern data was generated.

## REFERENCES

- Baddeley, A., and Turner, R. (2005). "Spatstat: An R Package for Analyzing Spatial Point Patterns." *Journal of Statistical Software* 12:1–42.
- Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005). "Residual Analysis for Spatial Point Processes." *Journal of the Royal Statistical Society, Series B* 67:616–666.
- Baddeley, A. J., Diggle, P., Hardegen, A., Lawrence, T., Milne, R., and Nair, G. (2014). "On tests of spatial pattern based on simulation envelopes." *Ecological Monographs* 84(3):477–489.
- Baddeley, A. J., Kerscher, M., Schladitz, K., and Scott, B. T. (2000). "Estimating the J function without edge correction." *Statistica Neerlandica* 54(3):315–328.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman & Hall/CRC.
- Banerjee, S., and Gelfand, A. (2002). "Prediction, interpolation, and regression for spatially misaligned data points." *Sankhya: The Indian Journal of Statistics* 64:227–245.
- Besag, J. (1977). "Discussion of Dr Ripley's paper." *Journal of the Royal Statistical Society, Series B* 39:193–195.
- Condit, R. (1998). *Tropical Forest Census Plots: Methods and Results from Barro Colorado Island, Panama, and a Comparison with Other Plots*. Berlin: Springer-Verlag.
- Condit, R., Hubbell, S. P., and Foster, R. B. (1996). "Changes in Tree Species Abundance in a Neotropical Forest: Impact of Climate Change." *Journal of Tropical Ecology* 12:231–256.
- D'Agostino, R. B., and Stephens, M., eds. (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker.
- Diggle, P. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, Taylor & Francis. ISBN 9781466560239.
- Diggle, P. J. (1979). "On Parameter Estimation and Goodness-of-Fit Testing for Spatial Point Patterns." *Biometrics* 35:87–101.
- Hubbell, S. P., and Foster, R. B. (1983). "Diversity of Canopy Trees in a Neotropical Forest and Implications for the Conservation of Tropical Trees." In *Tropical Rain Forest: Ecology and Management*, edited by S. J. Sutton, T. C. Whitmore, and A. C. Chadwick, 25–41. Oxford: Blackwell.
- Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Hoboken, NJ: John Wiley & Sons.
- Leitner, M. (2013). *Crime Modeling and Mapping Using Geospatial Technologies*. Geotechnologies and the Environment. Springer. ISBN 9789400749979.
- Levine, N. (2014). "CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v 4.01)." Ned Levine & Associates, Houston, Texas, and the National Institute of Justice, Washington, D.C.

- Ripley, B. D. (1976). "The Second-Order Analysis of Stationary Point Processes." *Journal of Applied Probability* 13:255–266.
- Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.
- Ripley, B. D., and Rassin, J.-P. (1977). "Finding the Edge of a Poisson Forest." *Journal of Applied Probability* 14:483–491.
- Stoyan, D., and Stoyan, H. (1994). *Fractals, random shapes, and point fields: methods of geometrical statistics*. Wiley series in probability and mathematical statistics: Applied probability and statistics. John Wiley & Sons. ISBN 9780471937579.
- Van Lieshout, M. N. M., and Baddeley, A. J. (1996). "A Nonparametric Measure of Spatial Interaction in Point Patterns." *Statistica Neerlandica* 50:344–361.

## ACKNOWLEDGMENTS

The authors thank Anne Baxter for editorial assistance.

## CONTACT INFORMATION

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

Pradeep Mohan	Randy Tobias
SAS Institute Inc.	SAS Institute Inc.
SAS Campus Drive	SAS Campus Drive
Cary, NC 27513	Cary, NC 27513
Pradeep.Mohan@sas.com	Randy.Tobias@sas.com

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 trademarks of their respective companies.