THE ANALYSIS OF MAPPED SPATIAL POINT PATTERNS

Edgar Barry Moser, Dept. Experimental Statistics, Louisiana State Univ.

ABSTRACT

A SAS® macro for the analysis of mapped spatial point patterns is described. The nearest neighbor and nearest event distance tests are implemented along with second-order tests described by Ripley (1977). These tests permit the classification of patterns into aggregated, regular, or random point patterns. Monte Carlo tests permit further testing for randomness as well as testing against other alternatives. The SAS/GRAPH® procedures are used to display these functions to aid with their interpretation. A Poisson cluster process is fit to redwood seedling data using PROC NLIN.

INTRODUCTION

Spatial point patterns arise in a variety of situations with examples from ecology, the positions of trees in a forest (Strauss 1975); astronomy, the distribution of galaxies in the universe (Peckles 1973); entomology, the positions of ant nests in a field (Harkness and Isham 1983); and archaeology, the locations of artifacts found at a dig (Graham 1980). Patterns may give insight into the mechanisms structuring the pattern, the interaction of events in the process, prediction of new events, and hypothesis tests of models. This paper will be concerned with 2-dimensional patterns located within a rectangular region, A, and all events will be assumed to be of the same type. In many instances, the analyses can be easily extended to handle shapes other than rectangular and to include multiple event types. The SAS software will be used to perform hypothesis tests for specific spatial pattern models as well as to fit a more complicated spatial model.

Spatial patterns are commonly classified into 1 of 3 mutually exclusive groups: regular, random, or aggregated. The aggregated category is sometimes referred to as contagious or clumped. Often, regular patterns are thought to imply inhibition of events while aggregated patterns suggest attraction of events. Since the random pattern is intermediate of the other two patterns, hypothesis testing of spatial patterns usually begins with the null hypothesis of a random pattern or complete spatial randomness (CSR). The hypothesis of CSR can be stated as 1) the number of events in the region A follows a Poisson distribution with mean \( \lambda A \), and 2) given N(A)=n, the events in A form an independent random sample from the uniform distribution on A, where |A| denotes the area of A, and \( \lambda \) is the intensity or mean number of events per unit area (Diggle 1983). Rejection of CSR implies a regular or aggregated pattern depending upon the direction of rejection.

TESTS FOR COMPLETE SPATIAL RANDOMNESS

Methods for the testing of CSR have been based upon quadrat counts, distance measures, second-order methods, and "test-set" approaches (Ripley 1977). The distance measure and second-order methods further permit spatial pattern model fitting. A SAS macro was written to perform tests of CSR using distance and second-order methods with additional output available for use in model fitting. Additionally, graphical displays are used to further enhance interpretation of results.

The first test incorporated within SPATIAL is based upon the first nearest neighbor (NN) distribution or distance y from an event to the nearest other event in the region A. These distances are found by computing the Delaunay triangulation (Upton and Fingleton 1985) of the events with SAS/GRAPH PROC G3GRID and subsequently, selecting the shortest segment of the triangles associated with each event. The spatial pattern of NN distances (1975) redwood seedling data, along with their Delaunay triangulation, is shown in figures 1 and 2. The vertices of the triangles are the seedlings, or events.

The empirical distribution function of nearest neighbor distances, \( G(y) \), is estimated by

\[
G(y) = \frac{\#(y_i \leq y)/n}{y \geq 0, (1)}
\]

where \( \# \) is the counting operator and n is the number of events in the region A. For large n, the distribution function for NN distances is

\[
G(y) = 1 - \exp(-\lambda y^2), y \geq 0. (2)
\]

Unbiased estimation of \( G(y) \) to incorporate boundary effects is made by

\[
G(y) = \frac{\#(y_i \leq y, d_i > y)/\#(d_i > y), y \geq 0 (3)}
\]

where \( d_i \) is the shortest distance from the event i to the boundary. Donnelly (1978) found that under CSR, the distribution of y, the sample mean NN distance, is approximately normally distributed with mean

\[
E(Y) = \frac{4}{3\pi^2} (n^{-1} |A|)^{1/3} + (0.065+0.037n^{-1})n^{1/3} A(A) (4)
\]

and variance

\[
\text{Var}(Y) = 0.070n^{-2} |A| + 0.037(n^{-1} |A|)^{5/3} A(A). (5)
\]

where A(A) is the length of the boundary of A. Small values of y indicate a regular pattern while large values indicate aggregation. Results for the redwood seedling data using SPATIAL are given in Table 1. CSR is emphatically rejected based upon the mean NN distance.

A second method of testing for CSR uses the points to nearest event distances (NE) where points are placed on the area, usually in a regular grid pattern, and the distances x from each point to its nearest event are measured.
TEST FOR COMPLETE SPATIAL RANDOMNESS

| TOTAL AREA:          | 1.00000000 |
| TOTAL BOUNDARY LENGTH: | 4.00000000 |
| TOTAL NUMBER OF EVENTS: | 62.00000000 |
| MIN. NEAREST NEIGHBOR DISTANCE: | 0.01600000 |
| MAX. NEAREST NEIGHBOR DISTANCE: | 0.11000000 |
| MEAN NEAREST NEIGHBOR DISTANCE: | 0.03853779 |
| EXPECTED MEAN VALUE UNDER CSR: | 0.06713452 |
| VARIANCE UNDER CSR: | 0.00002310 |
| TEST STATISTIC FOR CSR: | -5.94392237 |
| PROB. > | TEST STAT. | UNDER CSR: | 0.00000000 |

Table 1. A test of CSR for the redwood data using the test statistic of Donnelly (1978).

Again, the empirical distribution function for nearest event distances is,

\[ F(x) = \frac{\#(X_t \leq x)}{m}, \quad x \geq 0, \]  

where \( m \) is the number of points in the grid. Unbiased estimation incorporating boundary effects is performed similarly to the NN estimation. Under CSR, \( F(x) \) is approximately,

\[ F(x) = 1 - \exp(-\lambda x^2), \quad x \geq 0, \]  

where \( \lambda = n|A|^{-1} \).

Monte Carlo tests are used to evaluate \( G(y) \) and \( F(x) \) by computing these functions for \( k \) patterns of \( n \) events simulated under CSR and comparing the functions estimated from the observed data to simulation envelopes constructed from the \( k \) simulations. A simulation envelope consists of the largest and smallest values taken over all simulations computed for each value of \( y \), say, of \( G(y) \), and \( x \) for \( F(x) \). A test with 95% confidence results from 19 simulations, while 99 simulations is used for 99% confidence. For each function, \( G(y) \) and \( F(x) \), the simulation envelopes and empirical distribution functions are plotted versus \( G(y) \) and \( F(x) \), respectively, from the simulations. If the EDF, at any point, is outside the simulation envelope, then CSR is rejected. In practice, a practical range of \( y \)-values (\( x \)-values) is selected rather than testing over large ranges of \( y \) (\( x \)). NN EDF values above the envelope suggest an aggregated pattern and those below suggest a regular pattern.

A test for CSR using the second-order properties of the spatial pattern is based on a function defined by Ripley (1977) as,

\[ K(t) = \lambda^{-1}, \quad \text{Expected number of additional events within a distance } t \text{ of an arbitrary event.} \]

Under CSR,

\[ K(t) = \pi t^2, \quad t \geq 0, \]  

and an unbiased estimator is,

\[ R(t) = n^{-2} |A| \sum_{i,j}^{-1} I_{e(x)}(u_{ij}), \]  

where \( I_{e(x)}(u_{ij}) \) is the indicator function for the event \( u_{ij} \).
where $u_{ij}$ is the distance between events $i$ and $j$, $I_t(u) = 1$ if $u < t$ and 0 otherwise, and $w_{ij}$ is the proportion of the circumference of a circle centered at the event $i$ with radius $u_{ij}$ which lies within $A$ (Diggle 1983). Monte Carlo methods can again be used to compare this estimate with the CSR hypothesis. By plotting $L(t) = t - [R(t)/t]^2$ versus $t$, an initial appraisal can be made since under CSR, $L(t)$ versus $t$ should be a horizontal line at $L(t) = 0$. An example analysis of the redwood seedling data using the spatial macro is given in figure 3. The NN and NE EDs, Ripley's $K(t)$ and $L(t)$ functions, and simulation envelopes for each are shown. The envelopes were constructed using 19 simulations. To produce this analysis, the SPATIAL macro was invoked using the statement,

```
SPATIAL X Y DATA=REDWOOD GMAX=200 GTCRIT=0.25 FMAX=200 FTCRIT=0.35 KMAX=200 KTCRIT=0.25 MONTE NSIMS=19;
```

### SPATIAL PATTERN MODELS

To demonstrate the fitting of a spatial model, a Poisson cluster process (see Diggle 1983 for a description of this process),

$$K(t) = \pi t^2 + e^{-t^2/(4\sigma^2)}$$

was fitted to the redwood data using the methods given by Diggle (1983). In this model, $\pi$ is the mean number of parents per unit area and $\sigma^2$ is the mean squared distance of an offspring from its parent. The discrepancy measure,

$$D(\theta) = \int [(K(t)/\pi - (K(t;\theta))^2) dt$$

was minimized using Simpson's composite rule (Burden et al. 1981, p.149) and was implemented using PROC NLIN. The range of integration is from 0 to $t_0$, where $t_0$ is the largest distance of interest. The power transformation, $c$, was taken at 0.25 as suggested by Diggle (1983) for aggregated patterns. The SPATIAL macro generated the $K(t)$ values for values of $t$ from 0 to 0.25 and placed them into a data set named RIPLEY. The following DATA step and PROC NLIN statements are used to fit model (10) for the parameters $\pi$ and $\sigma^2$,

```
%LET c=0.25;
/* TRANSFORMATION CONSTANT FOR TUNING FIT */

DATA RIPLEY;
/* COMPUTE THE WEIGHTS FOR SIMPSON'S METHOD */
SET RIPLEY NOBS=NOBSS;
IF N=1 THEN W=1;
ELSE IF N>NOBSS THEN W=1;
ELSE IF MOD(N,2)=0 THEN W=4;
ELSE W=2;
IF K=0 THEN KMOD=0;
ELSE KMOD=K**&C;
RUN;
```

PROC NLIN DATA=RIPLEY METHOD=MARQUARDT;
WEIGHT=W;
/* MINIMIZE DISCREPANCY USING SIMPSON'S METHOD */
PARAMETERS P=25.00 S2=0.002;
BOUNDS P>0, S2>0;
/* CONSTRUCT TEMPORARY VARIABLES */
T2=T*T;
CI=(4C-1);
B=2-T2/4;
TEMP1=EXP(B/S2);
TEMP2=T-TEMP1;
TEMP3=3.41592653*T2+TEMP2/P;
IF TEMP3=0 THEN DO;
RIGHTC=0; RIGHTCI=0;
END;
ELSE DO;
RIGHTC=TEMP3**&C; RIGHTCI=TEMP3**CI;
END;
MODEL KMOD=RIGHTC;
DER.P=-&C/P*P*TEMP2*RIGHTCI;
DER.S2=B*P/S2*TEMP2*RIGHTCI;
TITLE2 'NON-LINEAR FIT OF POISSON CLUSTER PROCESS'
RUN;

The resulting estimates, $(\hat{\pi}, \hat{\sigma^2})$, from the redwood seedling data are $(24.88, 0.001855)$ with asymptotic 95% confidence intervals of $19.95 < \pi < 29.81$ and $0.001175 < \sigma^2 < 0.002535$. This suggests a process with 20 to 30 parent trees. The observed and predicted values of $K(t)$ are shown in figure 4. Most of the model lack-of-fit is confined to the very small inter-event distances.

![Figure 4. Observed and predicted $K(t)$ functions for a Poisson cluster process model fitted with PROC NLIN.](image)
Figure 3. Simulation envelopes of complete spatial randomness for the redwood seedlings data about the nearest neighbor and nearest event EDFs and Ripley's K(t) and L(t) functions.
THE SPATIAL MACRO

The macro SPATIAL has, as parameters, the data set name containing the point pattern, the X and Y location variables, the number of intervals used in estimating G(y), F(x), and K(t), the maximum y, x, and t of interest, and the number of grid points for estimating F(x). For hypothesis testing, Monte Carlo simulations of CSR can be requested. Several options permit listings and plots of the data as well as SAS/GRAPH plots of the estimated functions as demonstrated in the figures. The nearest neighbor and K(data) sets can optionally be saved for model fitting. In addition, the routine for generating the CSR point pattern can easily be replaced for testing other hypothesized point patterns. More specifically, the macro interface is,

\%MACRO SPATIAL(X,Y,DATA= LAST ,MAXX=1,MAXY=1, GMAX=40,GTCRIT=0,FMAX=40,FTCRIT=0, M=10, KMAX=40 , KTCRIT=0, KONLY=0, NEARNAME=_NEGHBR_, KNAME= _KFUNC ,MONTE=0,SEED=0,NSIMS=19, LIST=0,SASGRAPH=0,CHART=0,MAP=0) / STMT;

where MAXX and MAXY give the maximum X and Y map dimensions with the origin assumed at (0,0). The G-, F-, and KMAX values denote the number of intervals to be used in computing their respective functions with maximum distances of G-, F-, and KTCRIT, respectively. M specifies the number of points in the NE measures. The KONLY option requests that only the K(t) functions be computed. The NEARNAME and KNAME specify the optional data set names for keeping the nearest neighbor and K-function estimates. The MONTE option requests the simulation of the CSR point process for NSIMS simulations and the SEED value is used as the initial seed for the random number generator. LIST requests listings of the function estimates, CHART requests histograms of the functions, and MAP requests a map of the spatial pattern and the Delaunay triangulation when the SASGRAPH keyword is also given. SASGRAPH requests plots and charts using the SAS/GRAPH procedures.

These techniques can be extended to handle other than rectangular regions, especially since Monte Carlo tests are not affected by the boundaries. Also, it is possible to extend these techniques to patterns in higher dimensions, as well as to multiple type patterns. The macro is written in modular form to permit modification of the present routines or addition of other tests.

The SAS macro SPATIAL is available from the author upon request at the address:
Edgar Barry Moser
Dept. of Experimental Statistics
Louisiana State University
Baton Rouge, LA 70803

REFERENCES


SAS and SAS/GRAPH are registered trademarks of SAS Institute Inc., Cary, NC, USA.