

June 13, 2008

Spatial Statistics

This chapter provides a brief introduction to the statistical analysis and modeling of *spatial* data. Spatial data is distinguished by observations that are obtained at spatial locations $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n$ where the \mathbf{s}_i are coordinates in the plane \mathcal{R}^2 or space \mathcal{R}^3 typically.

Time series models attempt to model the correlations between responses at different time points. Similarly, with spatial data, the spatial correlational structure needs to be incorporated and modeled. The field of spatial statistics is a relatively new area of development and remains an area of active statistical research. The importance of spatial dependencies was realized in the early 1900's and some of the methods of experimental design (randomization, blocking, etc.) were established in agricultural studies to control for spatial correlation.

Spatial statistics consists primarily of three main components:

- **Geostatistics** – a spatial process indexed over a continuous space.
- **Spatial Point Patterns** – pertaining to the location of “events” of interest.
- **Lattice Data** – spatial data indexed over a lattice of points.

Below is a short description of these three methods.

1. Geostatistics. Geostatistics began with mining type applications and the prefix “geo” indicates that geostatistics dealt initially with statistics pertaining to the earth. Geostatistics has applications in climatology, environmental monitoring, and geology. One of the most common geostatistical applications in the mining context was *kriging* which deals with predicting the ore grade in a mining area from spatially observed samples. Geostatistical applications have expanded original mining applications to include modeling soil properties, ground water studies, rainfall precipitation, public health, etc.

2. Point Patterns. Data analysis of point patterns corresponds to studies where the interest lies in where events of interest occur. A fundamental question of interest in this context is whether or not the points of interest are occurring at random, or do the points cluster in some manner, or perhaps the points of interest are occurring with some sort of regularity.

3. Lattice Data. Lattice type data provide the closest analogue to time series data. In time series data sets, observations are typically obtained at equally spaced time points. For lattice data, spatial data is obtained over a regularly spaced set of points (irregular lattice type data is a possibility as well). Lattice data often comes in the form of *pixels*, which are small rectangularly shaped regions, often obtained by

remote sensing from satellites or aircraft. Lattice type data for spatial statistics is very similar to the type lattice type data one obtains from medical imaging studies, such as PET scans which yield images of the brain by way of pixels or voxels.

We shall explore each of these three types of spatial data in more detail next.

Geostatistics.

Geostatistical data is characterized by data involving the measure of a variable of interest over a spatial region where the measurement points can vary continuously. The primary tool in geostatistical analysis is the *variogram* which we will describe shortly.

Let D denote a region of interest (a forest, a field, a mining area etc). Each point $\mathbf{s} = (x, y)$ in D can be described by an x and y coordinate in the plane. In applications we want to measure some variable z (water or pH, mercury levels etc.) at locations within the region D . Let $z(\mathbf{s})$ denote the random variable that can be measured at location \mathbf{s} in the region. In practice, measurements are obtained at a finite number n of points (out of a possible infinite collection of points in D). Geostatistical data then looks like:

$$z(\mathbf{s}_1), z(\mathbf{s}_2), \dots, z(\mathbf{s}_n).$$

A very simple model for geostatistical data is

$$z(\mathbf{s}) = \mu + \epsilon(\mathbf{s}). \quad (1)$$

The error $\epsilon(\mathbf{s})$ is assumed to have a mean of zero in which case

$$E[z(\mathbf{s})] = \mu. \quad (2)$$

Another common assumption is that of *homoskedasticity*:

$$\text{var}[z(\mathbf{s})] = \sigma^2 \quad (3)$$

for all points \mathbf{s} in D . Recall that the autocovariance function for a stationary time series only depends on the lag between the time points. Similarly, for geostatistical data, we can consider covariance functions for the response z at two different points \mathbf{s}_1 and \mathbf{s}_2 that depend only on the difference in locations (distance and direction) between the two points:

$$\text{cov}[z(\mathbf{s}_1), z(\mathbf{s}_2)] = c(\mathbf{s}_1 - \mathbf{s}_2), \quad (4)$$

for some function c . Spatial data that satisfy the conditions of (2), (3), (4) is called *second-order stationary*. Additionally, if the covariance function depends only on the difference between the two points, then it is called *intrinsically stationary*. An even stronger condition states that the covariance function depends on the *distance* between the two points \mathbf{s}_1 and \mathbf{s}_2 – a spatial process whose covariance function satisfies this constraint is called *isotropic*. A spatial process is said to be *anisotropic* if the dependence between the response z at two points depends not only on the distance,

but also on the direction of that difference. Anisotropic process are commonly due to the underlying process evolving differentially in space.

The Variogram.

Instead of using the covariance function for spatial data, a related function called the *variogram* is often used for spatial data. Consider the response variable measured at two different locations $z(\mathbf{s}_1)$ and $z(\mathbf{s}_2)$. Intuitively, one may expect that as the two points get further apart (i.e. as the difference $\mathbf{s}_1 - \mathbf{s}_2$ get bigger, the measured difference in the response $z(\mathbf{s}_1) - z(\mathbf{s}_2)$ will also get bigger. The purpose of the variogram is to understand the squared differences

$$(z(\mathbf{s}_1) - z(\mathbf{s}_2))^2$$

between points in D . If the variance of the difference $(z(\mathbf{s}_1) - z(\mathbf{s}_2))$ depends only on the difference $\mathbf{s}_1 - \mathbf{s}_2$, then we can write

$$\text{var}[z(\mathbf{s}_1) - z(\mathbf{s}_2)] = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2),$$

for some function γ . The function 2γ is called the **variogram** (and γ itself is called the semivariogram).

Suppose the mean response is constant at each point in the region D . Then

$$\text{var}[z(\mathbf{s}_1) - z(\mathbf{s}_2)] = E[(z(\mathbf{s}_1) - z(\mathbf{s}_2))^2].$$

From the definition of variogram it follows that

$$\gamma(\mathbf{s}_1 - \mathbf{s}_2) = 0.5E[(z(\mathbf{s}_1) - z(\mathbf{s}_2))^2].$$

We can write this last expression as

$$\begin{aligned} 0.5E[(z(\mathbf{s}_1) - z(\mathbf{s}_2))^2] &= 0.5\{E[(z(\mathbf{s}_1) - \mu)^2] + E[(z(\mathbf{s}_2) - \mu)^2] - 2E[(z(\mathbf{s}_1) - \mu)(z(\mathbf{s}_2) - \mu)]\} \\ &= \sigma^2 - \text{cov}[z(\mathbf{s}_1), z(\mathbf{s}_2)] \\ &= \sigma^2(1 - \rho(z(\mathbf{s}_1), z(\mathbf{s}_2))), \end{aligned}$$

assuming the variance is also constant through out the region. Here, ρ is the correlation function between two spatial points. If we let h denote the distance between the two points $z(\mathbf{s}_1)$ and $z(\mathbf{s}_2)$, then we have shown that the variogram for an isotropic process can be written as

$$2\gamma(h) = 2\sigma^2(1 - \rho(h)).$$

One can see immediately from this expression that as h gets large and the correlation becomes small, the variogram takes the value $2\sigma^2$. Figure 1 shows a theoretical variogram. As the distance h gets larger, the variogram values increase indicating that as points get farther apart, the expected difference between the measured response at those two points increases as well.

The Nugget Effect. One would expect that $2\gamma(0)$ should equal zero. That is, $\text{var}[z(\mathbf{s}_1) - z(\mathbf{s}_2)]$ should equal zero if $\mathbf{s}_1 = \mathbf{s}_2$. However, this is usually not the case.

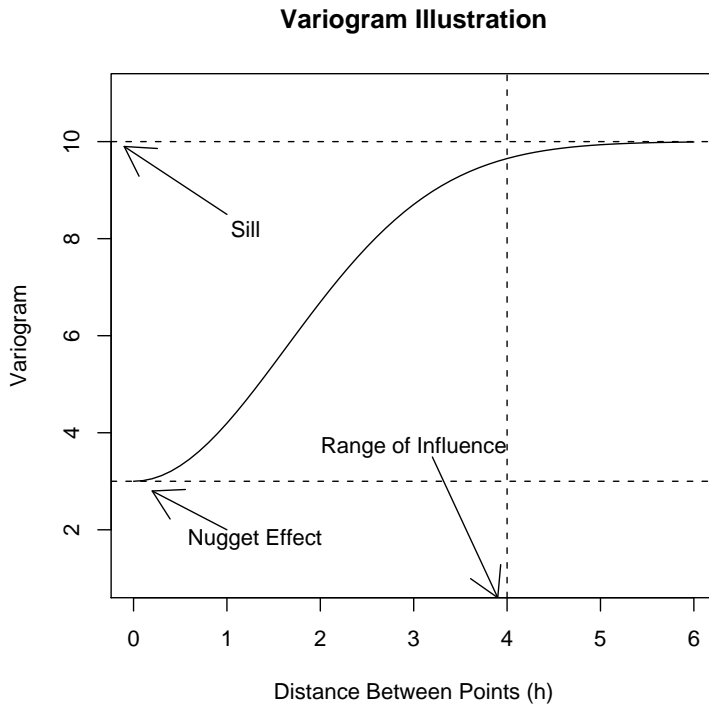


Figure 1: A plot of a theoretical variogram based on a Gaussian model with nugget = 3, sill = 10, and range of influence = 4

As the distance h goes to zero, there tends to be a *nugget effect* due to measurement error and microscale variation. The nugget effect is evident in Figure 1 when $h = 0$.

The Sill. Another aspect of the variogram is the *sill* which corresponds to the maximum height of the variogram curve. The sill is also indicated in Figure 1. As h gets large, the correlation (and hence covariance) between the response at two points separated by a distance h becomes negligible. In this case, the variogram value $2\gamma(h) = \text{var}[z(\mathbf{s}_1) - z(\mathbf{s}_2)] \approx 2\sigma^2$. Thus, the sill in the variogram plot corresponds to two times the variance.

The Range. The range is the distance h such that pairs of sites further than this distance apart are negligibly correlated. The *range of influence* is sometimes defined as the point at which the curve is 95% of the difference between the nugget and the sill.

Models for Variograms

The variogram plot in Figure 1 is for a model variogram. There exist various models for variograms used in practice. One of the more common models is the *Gaussian model* given by the following expression:

$$\text{Gaussian Model} \quad \gamma(h) = c + (S - c)\{1 - e^{-3h^2/a^2}\}, \quad (5)$$

where c is the nugget effect, S is the sill, and a is the range of influence. Note that $\gamma(0) = c$, the nugget effect. Also, as $h \rightarrow \infty$, $\gamma(h)$ approaches S , the sill. Finally, when $h = a$, then $\gamma(a) = c + (S - c)(1 - e^{-3}) \approx c + 0.95(S - c)$.

Some other common models used for variograms are the *spherical model*:

$$\textbf{Spherical Model} \quad \gamma(h) = \begin{cases} c + (S - c)\{1.5(h/a) - 0.5(h/a)^3\}, & \text{for } h \leq a \\ c, & \text{otherwise} \end{cases},$$

the *exponential model*:

$$\textbf{Exponential Model} \quad \gamma(h) = c + (S - c)\{1 - e^{-3h/a}\},$$

and the *power model*:

$$\textbf{Power Model} \quad \gamma(h) = c + Ah^w.$$

In each of these models, the constant c is the nugget effect. Note that the power model increases without bound.

Estimation of the Variogram.

The *classical* method of estimating the variogram (which corresponds to the method of moments estimator) is given by the following formula:

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} [z(\mathbf{s}_i) - z(\mathbf{s}_j)]^2, \quad (6)$$

where $N(h)$ is the set of all distinct pairs of points $z(\mathbf{s}_i), z(\mathbf{s}_j)$ such that $\|\mathbf{s}_i - \mathbf{s}_j\| = h$. In practice, the data is smoothed to generate an estimate of the variogram. For instance, the data can be partitioned into groups where observations in particular groups are within a certain range of distance apart and then using the average squared difference of the points in each group to replace the sum in (6).

Nonlinear least squares can be used to obtain a model estimate of the variogram, such as the Gaussian model in (5), for the variogram.

Nonconstant-Mean Assumption.

One of the implicit assumptions when using a variogram is that the mean of the spatial process is constant. Of course, this assumption may not hold in practice. In particular, a more general model than (1) allows for a nonconstant mean:

$$z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s}), \quad (7)$$

where $\epsilon(\mathbf{s})$ is the random error at point \mathbf{s} and $\mu(\mathbf{s})$ is a non-random mean function. As with time series analysis, a common goal is to estimate the mean function $\mu(\mathbf{s})$ and then model the correlation between the residuals after subtracting off the estimated mean.

Median Polish. One method of estimating the mean function is to use a *median polish*. The use of the median is more “robust” than the use of the mean. Recall that the mean can be strongly influenced by one or more outlying points. The median on the other hand is not strongly influenced by a few extreme points. For more in depth details on the median polish, see Cressie (1993, page 186). The median polish, described briefly here, is for gridded data although it can be modified for an irregular grid. The basic idea of the median polish algorithm is as follows:

1. Compute the median for each row in the grid. Remove the row median from each observation in a given row (i.e. subtract the median value from each observation).
2. Compute the median in each column of the grid. Remove the column median from each observation in a given column.
3. Repeat steps (2) and (3) until convergence.

After the algorithm converges, each of the original observations can be broken down into four parts: an overall median value plus a row effect plus a column effect plus the residual. There exist other types of polishing as well (such as the use of *trimmed means*).

Kriging

Kriging is a method of interpolation named after the South African mining engineer D. G. Krige. We mentioned the problem of prediction briefly in our notes on regression analysis. Kriging is also a prediction problem. Suppose we have a set of geostatistical data $z(\mathbf{s}_1), z(\mathbf{s}_2), \dots, z(\mathbf{s}_n)$. The goal of kriging is to estimate a response $z(\mathbf{s}_0)$ at a point \mathbf{s}_0 . In particular, kriging estimates $z(\mathbf{s}_0)$ are determined by a linear combination of the data values:

$$\hat{z}(\mathbf{s}_0) = \sum a_i z(\mathbf{s}_i), \quad (8)$$

where the a_i 's are weights chosen so that the prediction is unbiased and has the smallest prediction error variance. The weights a_i can be obtained via the variogram. The goal is to minimize the mean squared error $E[(z(\mathbf{s}_0) - \hat{z}(\mathbf{s}_0))^2]$. Let c_{ij} denote the covariance between $z(\mathbf{s}_i)$ and $z(\mathbf{s}_j)$. Note that $c_{ii} = \text{var}(z(\mathbf{s}_i))$. The solution for finding the weights a_i in (8) is found using the method of Lagrange multipliers and is given by

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \\ m \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} c_{10} \\ c_{20} \\ \vdots \\ c_{n0} \\ 1 \end{pmatrix},$$

where m is the Lagrange multiplier. The variogram is needed in the prediction equations used by kriging. For additional details, see Thompson (1992, p 243).

There exists different types of kriging. *Simple kriging* assumes the mean is constant through out the region while ordinary kriging allows for the mean to vary in different parts of the region by only using sample points nearest to the point to be predicted.

Spatial Designs

If the object of the spatial analysis is to predict a value at a new spatial location, then the issue of sampling points arises. What sampling points should be obtained in order to obtain the most accurate predictions? An obvious answer is to sample points near the point you want to make a later prediction. However, if you do not know ahead of time where you want to make predictions, or if you want to make predictions at several locations, how should points be sampled?

The answer is to obtain points that provide the most information possible. Responses for points near each other spatially tend to be highly correlated. If responses are highly correlated, then knowing one response tells us a lot about the other response. That is, we obtain some redundant information. One can show mathematically that predictions become more accurate (in terms of mean squared error) as the correlations between the sampled points becomes smaller. This suggests that the sampled points should be spread out as much as possible, perhaps in a systematic fashion. However, systematic sampling patterns can become very inefficient if there is some sort of periodic variation in the region.

For systematic sampling, Matérn (1986) found that for predicting the mean over a region that a triangular grid was most efficient, more so than a square grid.

In order to estimate the mean concentration of a point-source pollutant over a region, McArthur (1987) found that the most efficient strategy was to use a stratified systematic sampling approach.

Installing Geostatistical Package in R

Paulo Ribeiro Jr. and Peter Diggle have developed a geostatistical package for the software R. To load this package into your R library, first start R on your computer and make sure the computer is connected to the internet. Then, at the R prompt, type the following:

```
install.packages("geoR", contriburl = "http://www.est.ufpr.br/geoR/windows")
```

Spatial Point Patterns.

Consider a region D and in this region we are interested in locations of certain “events”. We may ask ourselves if the events of interest are occurring randomly throughout the area, or if the events tend to cluster together. Some examples may help to illustrate the ideas:

1. A study was done which located nests of the ant *Messor wasmani* in a 240 by 250 square foot area (Harkness and Isham 1983). Additionally, nests of the ant *Cataglyphis bicolor* were also located. Another question of interest besides whether or not the nests are randomly located throughout the region is whether or not there is any evidence of a relationship between the positions for the two species of ants.

2. A study was conducted on the longleaf pine in an old growth forest in Thomas County Georgia (see Rathbun and Cressie 1990). Data on 584 tree locations were obtained in this study. Again, one of the main questions of interest is whether or not the spatial locations of the trees in the forest is random or are they clustered in some way.

The problem with trying to determine the presence of clustering in a spatial data set can be very difficult. Points formed from completely random processes can appear clustered.

Spatial point patterns consisting of n events will be compared to completely random point processes. A *complete spatial randomness* (also known as a *homogeneous Poisson process*) is one where, conditional on n events (i.e. points in the region), the events are independently and uniformly distributed over the region.

Statistical methods have been developed to test if a spatial point pattern is random or not. Some of these methods are computer intensive *Distance based methods*.

Figure 2 shows a plot of spatial data that was generated as in accordance to complete spatial randomness. Any pattern that shows up in this plot is due completely to chance and not because of some underlying structure. Figure 3 shows a plot of spatially located points that appears to be spread roughly uniformly throughout the region in a grid-like pattern. Some people may mistakenly consider this random data due to the uniform spacing. However, the data in Figure 3 has very little randomness associated with it. Finally, Figure 4 shows data generated by two clusters of points. Is it evident that there exists two distinct clusters in Figure 4?

Before going into details on some testing procedures, we introduce a couple of definitions.

Intensity Function. For geostatistical data, the mean function $\mu(\mathbf{s})$ was of interest. The analogue of the mean function for spatial point patterns is the intensity function $\lambda(\cdot)$. Let $N_{\mathbf{s}}$ denote the number of events in a square centered at location \mathbf{s} . Consider the ratio

$$\frac{P(N_{\mathbf{s}} > 0)}{\text{area of square}}$$

The intensity function $\lambda(\mathbf{s})$ is the limit of this expression as the area of the square goes to zero. There exist various methods for estimating the intensity function $\lambda(\mathbf{s})$ including nonparametric *kernel* estimators (commonly used to estimate density functions) which are based on smoothing the data.

The K Function. The analogue of the variogram from geostatistics is the K function for spatial point patterns. The K function captures the spatial dependence between different parts of the region where the sampling takes place. The K function is defined to be

$$K(h) = \frac{(\text{Ave. \# events within a distance } h \text{ of each other})}{\lambda}$$

This is estimated empirically by replacing the numerator by the average number of pairs of points within a distance h of each other and replacing the denominator by an estimator of the intensity.

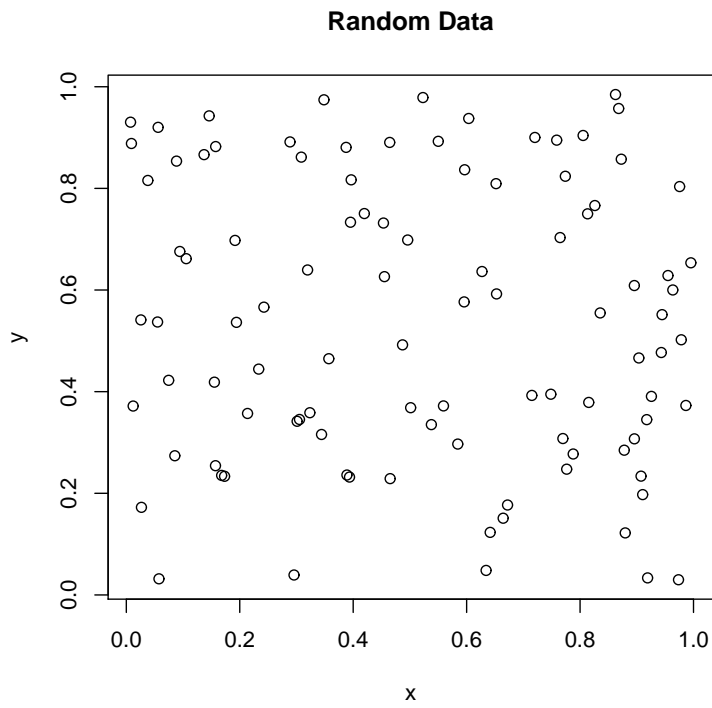


Figure 2: Spatial data generated from random data points.

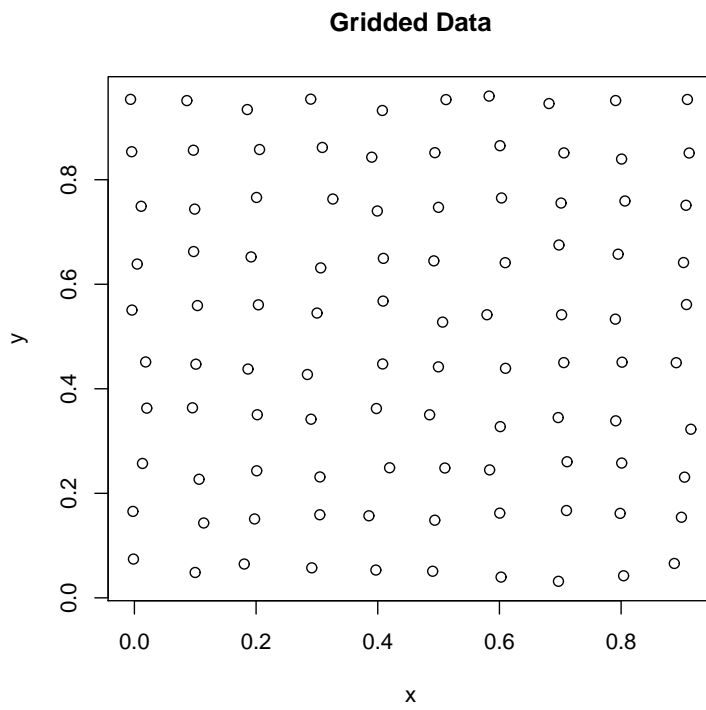


Figure 3: Data in a spatial region forming a grid-like pattern.

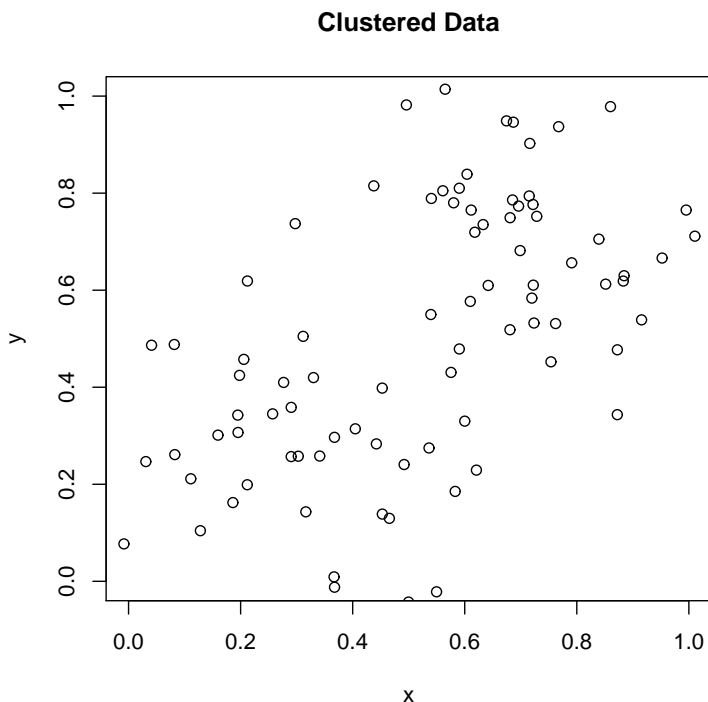


Figure 4: Clustered spatial data generated from two underlying clusters.

Quadrat Methods.

The basic idea of quadrat methods is to divide the region D into subsets often rectangular in shape (but other shapes are used as well), and then counting the number of events in each of the subsets. The use of quadrat counts can be used to access whether there is any spatial pattern in the data. If clustering is present in the data, then one would expect quadrats with higher counts to be located near each other. On the other hand, if the quadrat counts are spread out over the region, then there is evidence of uniformity.

If the events correspond to spatial randomness, then the quadrat counts should conform to this randomness as well. Recall that count data is often modeled by a Poisson distribution. One can show that for complete spatial randomness the quadrat counts will indeed have a Poisson distribution. Recall that the Poisson distribution has the following density function:

$$f(x) = e^{-\mu} \mu^x / x!, \quad x = 0, 1, 2, \dots,$$

where μ is the rate parameter (and also the mean and variance of the distribution).

Because the mean and variance of a Poisson distribution are equal, we can define a test statistic based on this property. For all the quadrats in the data set, compute the mean count \bar{x} and the sample variance s^2 of the quadrat counts. Then the quantity

$$R = s^2 / \bar{x},$$

should be close to one in value if the counts are Poisson distributed. Deviations of R from 1 indicate deviations from spatial randomness. In particular,

- If R is too small (less than 1) then the mean is bigger than the variance. This indicates that the counts appear more uniform than expected from a strictly random process. If the events are more spread out than what one would expect from a random scatter of points, then that is an indication of *evenness*.
- If R is too big (greater than 1), then the variance is bigger than the mean. This is an indication of events to accumulate together which is indicative of clustering.

Of course, the natural question is: what constitutes “too big” or “too small”? To answer this question, the behavior (i.e. sampling distribution) of R needs to be known when the null hypothesis is true. If R is standardized as

$$T = \frac{(R - 1)}{\sqrt{2/(n - 1)}},$$

then T follows a t -distribution on $n - 1$ degrees of freedom approximately under the hypothesis of complete randomness.

Example. Returning to Figure 2, Figure 3, and Figure 4, let us test for randomness using the quadrat method by dividing each region into square quadrats. The quadrat grid is superimposed on the data in Figure 5, Figure 6, and Figure 7. For the random data, the counts in each of the quadrats are (going down the first column and then the second column, etc).

The quadrat counts for the data in Figure 2 are:

2	1	0	1	0	1	1	1	3	0
3	2	1	3	1	2	1	2	1	1
0	0	0	1	2	0	0	1	2	0
1	2	0	1	2	0	1	0	1	3
2	1	1	0	0	1	1	0	1	2
1	1	1	1	1	0	0	0	0	2
1	0	2	3	1	2	0	3	2	3
1	3	0	2	1	1	0	2	1	1
1	0	0	0	0	0	3	0	1	1
1	0	1	0	0	0	1	0	0	2

Quadrat counts from the random spatial pattern.

The mean and variance of the quadrat counts for the random data in Figure 5 are $\bar{x} = 1.000$ and $s^2 = 0.8889$. The ratio of the variance to the mean is $R = 0.8889$ and the t -test statistic is

$$t = \frac{(R - 1)}{\sqrt{2/(n - 1)}} = \frac{(0.8889 - 1)}{\sqrt{2/(100 - 1)}} = -0.7817$$

which gives a two-tailed p -value of $p = 0.4362$ indicating that the hypothesis of random data is consistent for the data shown in Figure 2.

The quadrat counts for the data in Figure 3 are:

1	1	1	0	1	2	1	1	0	1
1	0	1	2	1	1	0	2	0	1
2	1	0	2	1	0	1	1	1	1
2	1	0	1	2	0	2	1	0	1
0	1	2	0	1	2	0	2	0	1
1	2	1	0	2	0	1	1	2	0
2	0	2	1	1	1	1	1	0	1
1	2	0	1	1	2	0	2	0	1
0	2	0	2	1	1	0	2	1	0
1	1	1	1	1	0	2	0	2	0

Quadrat counts for gridded spatial pattern

The mean and variance of the quadrat counts for the gridded data in Figure 6 are $\bar{x} = 0.95$ and $s^2 = 0.53283$. The ratio of the variance to the mean is $R = 0.56087$ and the t -test statistic is

$$t = \frac{(R - 1)}{\sqrt{2/(n - 1)}} = \frac{(0.56087 - 1)}{\sqrt{2/(95 - 1)}} = -3.0105$$

which yields a two-tailed p -value of $p = 0.0034$ indicating very strong evidence that the spatial data is not random in Figure 3. In fact, since $r = 0.56097 < 1$, this indicates that the counts are more uniform than would be expected by chance.

The quadrat counts for the data in Figure 4 are:

0	0	0	0	1	0	2	2	1	0
0	0	0	0	1	2	1	0	0	0
0	0	1	0	0	2	6	3	1	1
0	0	1	0	0	0	2	2	3	1
0	0	0	1	0	1	2	3	0	1
2	1	2	1	0	2	0	1	1	0
0	3	2	0	2	1	0	0	1	0
2	1	1	3	2	1	1	0	0	0
0	2	1	1	2	1	0	0	0	0
0	0	0	1	0	0	0	0	0	0

Quadrat counts for the clustered spatial data.

The mean and variance of the quadrat counts for the clustered data in Figure 7 are $\bar{x} = 0.79$ and $s^2 = 1.0969$. The ratio of the variance to the mean is $R = 1.388$, and the t -test statistic is

$$t = \frac{(R - 1)}{\sqrt{2/(n - 1)}} = \frac{(1.388 - 1)}{\sqrt{2/(79 - 1)}} = 2.426$$

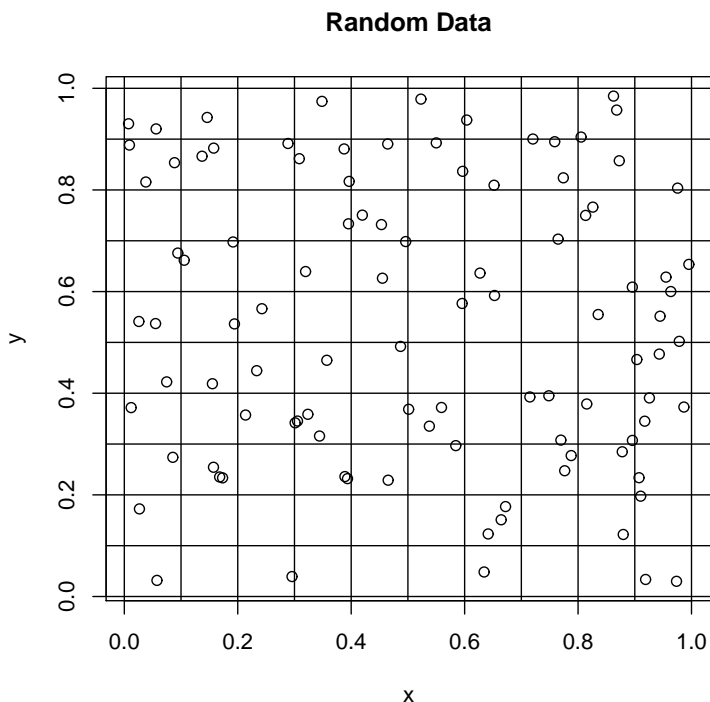


Figure 5: Spatial data generated from random data points with square quadrats.

which yields a two-tailed p -value of $p = 0.0176$. Thus, we reject the hypothesis of random data shown in Figure 4. Because $R = 1.388 > 1$, we conclude that the counts are more variable than one would expect by chance which is consistent with clustered data.

Mantel Matrix Randomization Test. Another way of testing for a random scatter of points in the region is by use of a *randomization* test which is a computer intensive method. The test is carried out as follows:

1. For any two quadrats i and j , compute the distance between them and denote this distance by $d_{i,j}$. These distances can be collected together into a matrix:

$$\mathbf{D} = \begin{pmatrix} 0 & d_{1,2} & d_{1,3} & \dots & d_{1,n} \\ d_{2,1} & 0 & d_{2,3} & \dots & d_{2,n} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ d_{n,1} & d_{n,2} & d_{n,3} & \dots & 0 \end{pmatrix}$$

Note that the matrix \mathbf{D} is a symmetric matrix and that the main diagonal terms are all zero.

2. Similarly, construct a matrix where the i th row and j th column element is the absolute difference in the counts between quadrats i and j :

$$c_{i,j} = |c_i - c_j|.$$

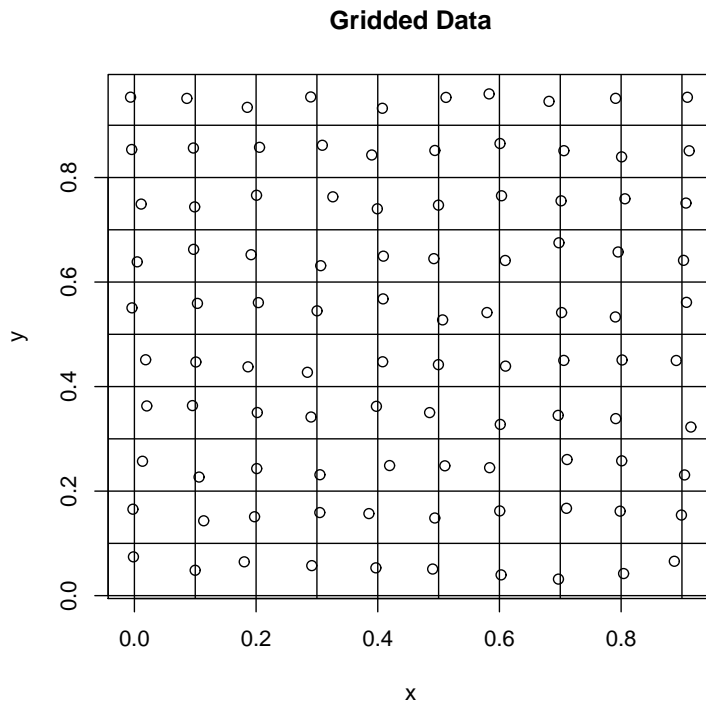


Figure 6: Data in a spatial region forming a grid-like pattern with square quadrats.

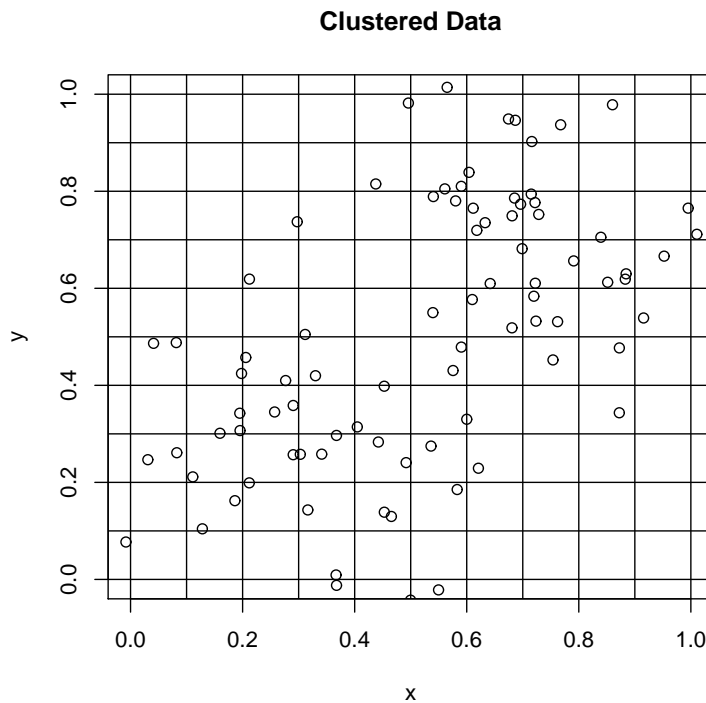


Figure 7: Clustered spatial data generated from two underlying clusters with square quadrats.

Thus, define \mathbf{C} as

$$\mathbf{C} = \begin{pmatrix} 0 & c_{1,2} & c_{1,3} & \dots & c_{1,n} \\ c_{2,1} & 0 & c_{2,3} & \dots & c_{2,n} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ c_{n,1} & c_{n,2} & c_{n,3} & \dots & 0 \end{pmatrix}$$

3. Once these matrices are constructed, simply compute the Pearson correlation r between the pairs $(d_{i,j}, c_{i,j})$. The question of interest now is whether or not r is more extreme than what one would expect from a random scattering of events to quadrats.

To test the hypothesis of randomness, what one does is randomly re-assign the quadrat counts to quadrats. This can be done easily with a computer. To understand the randomization, consider writing each of the quadrat counts on cards and then shuffle the cards. Next, deal the cards out to each of the quadrats – this will randomly re-assign the counts to quadrates. After randomly assigning the counts, repeat the above 3 steps to get a new correlation value, call it r_1 . Now do many more random re-assignments of counts and in each case generate the corresponding correlations r_2, r_3, \dots, r_N where N is some large number (say $N = 10,000$). From this randomized distribution of correlations, observe where the original correlation r from the original data falls. A p -value for this test simply the proportion of randomized correlations that exceed the observed correlation r in magnitude.

Instead of grouping the events artificially into quadrats, we can instead deal directly with where the events actually occur in the region. The following are some tests for randomness in a spatial point pattern.

Nearest Neighbor Methods. The idea here is to compute the distance from each point to its nearest neighbor. Do this for all points and compute the average of these nearest-neighbor distances. Next, simulate a set of n points from a completely random point processes and compute this same nearest-neighbor average distance. We can continue to generate simulated data sets that are completely random. For each of these simulated data sets, one can compute the nearest-neighbor average distance. These simulated nearest-neighbor average distances can then be used to determine the sampling distribution of the nearest-neighbor average distance when the null hypothesis of complete spatial randomness holds. Generally, a large number of simulated data sets need to be generated in order to obtain a good approximation to the null-distribution of the test statistic (e.g. 1000 to 10,000 data sets). The nearest-neighbor average distance for the original data can then be compared to this null distribution to access whether or not the hypothesis of complete spatial randomness can be rejected. In particular, if the nearest-neighbor average distance for the original data falls in the extreme tail of the null distribution, then that is evidence against the null hypothesis. This type of procedure is an example of a *Monte Carlo* test.

Using only the nearest neighbor in the test just described is not very efficient use of the data. One can also consider measuring distances to the nearest second, third, fourth neighbor and so on. Let Q_1 equal the mean distance of each point to its

nearest neighbor, and let Q_2 equal the mean distance of each point to its second nearest neighbor. We can compute the statistics Q_1, Q_2, \dots, Q_k , and so on. Again, Monte Carlo methods can be used to simulate the null distribution of Q_1, Q_2, \dots , and compare then compare the values of Q_1, Q_2, \dots , from the actual data to the null-distribution.

One difficulty that arises with these nearest-neighbor tests deals with the boundary of the area of interest, called *Edge Effects*. Nearest neighbor distances for events near the boundary will tend to be larger than points near the center of the region because points near the boundary can only have nearest neighbors in a direction away from the boundary. There are a few common ways of dealing with this problem such as constructing a “guard” area inside the perimeter of the region and only points inside the guard area are considered. These edge effects can have at times considerable influence on the statistical results (see Figure 8.12 of Cressie (1993, page 613)).

Spatial Models for Lattice Data.

We have considered spatial data in a region D . In this section, the region D is a finite (or countable) collection of spatial sites where observations are made. The collection of points in D is known as a lattice. The spatial sites in a lattice are typically identified using its *longitude* (x) and its *latitude* (y).

Example. Consider an epidemiological study in the state of Ohio and let the lattice consist of the counties of Ohio. Each county can be identified by the longitude and latitude of the county seat.

Another aspect of lattice data is that of a *neighborhood*. In the Ohio counties example, the neighborhood of a given county will be the collection of counties that are nearby, say in terms of distance. For instance, the neighborhood of Greene County can be defined as all counties within 50 miles. The distance metric can be Euclidean distance, but other distance metrics can be used. For example, in an urban area, “city-block” distance may be the appropriate metric: $|x_i - x_j| + |y_i - y_j|$. Another way of defining a neighborhood is to simply define it to be the set of counties that border a given county.

There are three characteristics of lattice data that need to be considered:

1. Is the lattice *regular* or *irregular*? The Ohio Counties are irregular. On the other hand, consider a study of yield from an agricultural experiment. A regular lattice of points can be placed over the field to determine where the sampling will take place.
2. Do the locations in the lattice refer to “points” or “regions”? In the Ohio counties, each county is a region.
3. Is the response measured at the lattices “discrete” such as a count or “continuous”?

Initial Data Analysis for Lattice Data. A simple way to illustrate lattice data graphically is to use a *choropleth map*. A choropleth map is a map which shows regions or areas which have the same characteristics. For instance, in an epidemiological study of a particular disease in Ohio counties, we could record the number of instances of the disease in each county and then look at the rates in each county (# cases/county population). Counties with similar rates will then be shaded the same color. However, choropleth maps of rates as described here can be quite misleading because of varying populations within counties. The variability of the rates for smaller counties will be much higher than for counties with large populations.

If the data are on a regular lattice, it is useful to “smooth” the data. One method for doing this is to apply the *median polish* that was discussed previously.

Depending on the type of data being analyzed, a transformation may be helpful. For instance, if we are analyzing count data, such as the number of cases of a disease in the population, the distribution of the counts is binomial. The variance of a binomial random variable depends on the number of trials n , which in this case is the population size. Often, a square-root transformation is helpful to get rid of the dependence between mean and variance of the measurements.

The next two sub-sections pertain to continuous response variables at the lattice points.

Spatial Markov Process.

In the Time Series chapter, the term *Markov* was introduced to describe a time series where the response at time t depended only on the response immediately preceding it at time $t - 1$ plus a random error. A spatial analogue of this sort of dependence states that the response z at the point (i, j) in a square lattice in the plane depends only on the responses at the four nearest-neighbor sites.

Spatial Autoregressive Models

Recall also from the Time Series chapter the definition of an autoregressive processes of order p stipulates the model

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \epsilon_t, \quad (9)$$

i.e. that the response y_t at time t is dependent on the responses at the p previous time periods. The errors ϵ_t are often assumed to be independent with a Gaussian (i.e. normal) distribution with mean zero and constant variance at all time points. For lattice data, suppose the lattice points are denoted $\mathbf{s}_1, \dots, \mathbf{s}_n$. Then the spatial analogue of the autoregressive model is given by:

$$z(\mathbf{s}_i) = \mu_i + \sum_{j=1}^n b_{ij}(z(\mathbf{s}_j) - \mu_j) + \epsilon_i, \quad i = 1, \dots, n.$$

The coefficients b_{ij} determine the extent of spatial dependence. (Note that $b_{ii} = 0$ for all i .) The errors ϵ_i are assumed to be independent with constant variance and mean zero. Additionally, the errors are often assumed to have a Gaussian distribution.

Inference procedures for spatial models typically use the method of maximum likelihood, assuming Gaussian errors. *Likelihood Ratio Tests* can then be performed to test various hypotheses. However, due to spatial dependencies, it is sometimes difficult to verify that maximum likelihood estimators will have approximately normal distributions for large sample sizes.

Remote Sensing.

Satellites and aircraft can be used to obtain images from which spatial data can be produced. Remote sensing can be used for inventory of natural resource (such as crop yields), or monitoring the effects of forest clearing and erosion. Such images are often contaminated by errors (i.e. noise) such as dust, ozone, water vapor etc. Part of the analysis of the images deals with restoring and interpreting the images. Other related applications include images generated from magnetic resonance (MRI) imaging and positron emission tomography (PET). The objects in an image vary continuously (colors, lines, etc.) but the images are usually captured in a discrete, digitized format. As before, let D represent the region of interest. For lattice data, D will consist of a set of lattice points. Recall that the goal is to determine the “true” image. At a lattice point \mathbf{s} , let $\theta(\mathbf{s})$ denote the true image. There are different types of images determined by the values θ can assume:

1. *Dichotomous Images* – the only possible values for θ are 0 and 1 at all lattice points.
2. *Polychotomous Images* – For these, θ can take any value in the set $\{1, 2, \dots, K\}$ for $K \geq 2$.
3. *Nonnegative Images* – θ can assume any positive real number value: $\theta > 0$.

Here are some terms associated with remote sensing:

Restoration: One of the statistical goals in remote sensing is to estimate the true values of θ in D based on data $y(\mathbf{s}_1), \dots, y(\mathbf{s}_n)$ obtained from the image.

Segmentation: The partitioning of the lattice D into regions corresponding to different types of surfaces etc. The goal is that the groups in the partition are homogeneous in some respect.

Classification: The problem of assigning a pixel (or region) a label (e.g. water surface, sandy soil, agricultural surface, etc.). The multivariate topic of *discriminant analysis* is the statistical tool used for classifying pixels into one of several possible classes. The classical methods of discriminant analysis assume the observations are independent of one another. This is clearly a false assumption for spatial data due to spatial correlations. Smoothing techniques can be used to incorporate spatial information for discriminant analysis. One common approach is to use a moving average (or median) *filter* (see the time series notes) where the smoothed response at a lattice point is equal to a weighted average (or median) of nearby lattice points. The nearest points get the highest weights.

References.

- Cressie, N. (1993), *Statistics for Spatial Data*, Revised Edition, Wiley: New York.
- Harkness, R. D. and Isham, V. (1983), “A bivariate spatial point pattern of ants’ nests,” *Applied Statistics*, **32**, 293—303.
- Matérn, B. (1986), *Spatial Variation*, 2nd Edition. Berlin: Springer-Verlag.
- McArthur, R. D. (1987), “An Evaluation of Sample Designs for Estimating a Locally Concentrated Pollutant,” *Communications in Statistics – Simulation and Computation*, **16**, 735–759.
- Rathbun, S. L. and Cressie, N. (1990), “A space-time survival point process for a longleaf pine forest in southern Georgia,” *Journal of the American Statistical Association*.
- Thompson, S. K. (1992), *Sampling*, Wiley: New York.