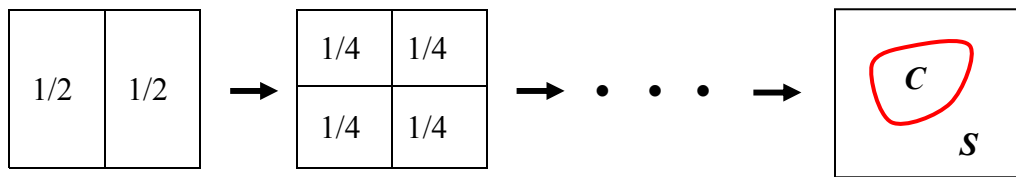


## 2. Models of Spatial Randomness

As with most all statistical analyses, cluster analysis of point patterns begins by asking: What would point patterns look like if points were *randomly distributed*? This requires a statistical model of randomly located points.

### 2.1 Spatial Laplace Principle

To develop such a model, we begin by considering a square region,  $S$ , on the plane and divide it in half, as shown on the left in Figure 1 below:



**Fig. 1. Spatial Laplace Principle**

The *Laplace Principle* of probability theory asserts that if there is no information to indicate that either of two events is more likely, then they should be treated as equally likely, i.e., as having the *same probability* of occurring.<sup>1</sup> Hence by applying this principle to the case of a randomly located point in square,  $S$ , there is no reason to believe that this point is more likely to appear in either left half or the (identical) right half. So these two (mutually exclusive and collectively exhaustive) events should have the same probability,  $1/2$ , as shown in the figure. But if these halves are in turn divided into equal quarters, then the same argument shows that each of these four “occupancy” events should have probability  $1/4$ . If we continue in this way, then the square can be divided into a large number of  $n$  grid cells, each with the same probability,  $1/n$ , of containing the point. Now for any subregion (or *cell*),  $C \subset S$ , the probability that  $C$  will contain this point is at least as large as the sum of probabilities of all grid cells inside  $C$ , and similarly is no greater than the sum of probabilities of all cells that intersect  $C$ . Hence by allowing  $n$  to become arbitrarily large, it is evident that these two sums will converge to the same limit – namely the fractional area of  $S$  inside  $C$ . Hence the probability,  $\Pr(C|S)$  that a random point in  $S$  lies in any cell  $C \subset S$  is *proportional to the area of  $C$* .<sup>2</sup>

$$(1) \quad \Pr(C|S) = \frac{a(C)}{a(S)}$$

Finally, since this must hold for any pair of nested regions  $C \subset R \subset S$  it follows that

<sup>1</sup> This is also known as Laplace’s “Principle of Insufficient Reason”.

<sup>2</sup> This argument in fact simply repeats the construction of area itself in terms of Riemann sums (as for example in Bartle (1975, section 24).

$$(2) \quad \Pr(C | S) = \Pr(C | R) \cdot \Pr(R | S) \Rightarrow \Pr(C | R) = \frac{\Pr(C | S)}{\Pr(R | S)} = \frac{a(C)/a(S)}{a(R)/a(S)}$$

$$\Rightarrow \boxed{\Pr(C | R) = \frac{a(C)}{a(R)}}$$

and hence that the square in (2) can be replaced by any *bounded region*,  $R$ , in the plane. This fundamental proportionality result, which we designate as the *Spatial Laplace Principle*, forms the basis for all models of spatial randomness.

In probability terms, this principle induces a *uniform probability distribution* on  $R$ , describing the location of a single random point. With respect to any given cell,  $C \subset R$ , it is convenient to characterize this event as a *Bernoulli (binary) random variable*,  $X(C)$ , where  $X(C) = 1$  if the point is located in  $C$  and  $X(C) = 0$  otherwise. In these terms, it follows from (2) that the conditional probability of this event (given that the point is located in  $R$ ) must be

$$(3) \quad \Pr[X(C) = 1 | R] = a(C)/a(R),$$

so that  $\Pr[X(C) = 0 | R] = 1 - \Pr[X(C) = 1 | R] = 1 - [a(C)/a(R)]$ .

## 2.2 Complete Spatial Randomness

In this context, suppose now that  $n$  points are each located randomly in region  $R$ . Then the second key assumption of spatial randomness is that the locations of these points have *no influence on one another*. Hence if for each  $i = 1, \dots, n$ , the Bernoulli variable,  $X_i(C)$ , now denotes the event that point  $i$  is located in region  $C$ , then under spatial randomness the random variables  $\{X_i(C) : i = 1, \dots, n\}$  are assumed to be *statistically independent* for each region  $C$ . This together with the Spatial Laplace Principle above defines the fundamental hypothesis of *complete spatial randomness (CSR)*, which we shall usually refer to as the *CSR Hypothesis*.

Observe next that in terms of the individual variables,  $X_i(C)$ , the total number of points appearing in  $C$ , designated as the *cell count*,  $N(C)$ , for  $C$ , must be given by the random sum

$$(4) \quad N(C) = \sum_{i=1}^n X_i(C)$$

[It is this *additive* representation of cell counts that in fact motivates the Bernoulli (0-1) characterization of location events above.] Note in particular that since the expected

value of a Bernoulli random variable,  $X$ , is simply  $P(X=1)$ ,<sup>3</sup> it follows (from the linearity of expectations) that the *expected number of points* in  $C$  must be

$$(5) \quad E[N(C) | n, R] = \sum_{i=1}^n E[X_i(C) | R] = \sum_{i=1}^n \Pr[X_i(C)=1 | R]$$

$$= \sum_{i=1}^n \frac{a(C)}{a(R)} = n \cdot \frac{a(C)}{a(R)} = \left( \frac{n}{a(R)} \right) a(C)$$

Finally, it follows from (3) that the under the CSR Hypothesis, the sum of independent Bernoulli variables in (4) is by definition a *Binomial random variable* with distribution given by

$$(6) \quad \Pr[N(C) = k | n, R] = \frac{n!}{k!(n-k)!} \left( \frac{a(C)}{a(R)} \right)^k \left( 1 - \frac{a(C)}{a(R)} \right)^{n-k}, \quad k = 0, 1, \dots, n$$

For most practical purposes, this conditional *cell-count distribution* for the number of points in cell,  $C \subset R$  (given that  $n$  points are randomly located in  $R$ ) constitutes the basic probability model for the CSR Hypothesis.

### 2.3 Poisson Approximation

However, when the reference region  $R$  is large, the exact specification of this region and the total number of points  $n$  it contains will often be of little interest. In such cases it is convenient to remove these conditioning effects by applying the well-known Poisson approximation to the Binomial distribution. To motivate this fundamental approximation in the present setting, imagine that you are standing in a large tiled plaza when it starts to rain, and consider the number of rain drops landing on the tile in front of you during the first ten seconds of rainfall. Here it is evident that this number should not depend on either the size of the plaza itself or the total number of raindrops hitting the plaza. Rather, it should depend on the *intensity* of the rainfall – which should be the same everywhere. This can be modeled in a natural way by allowing both the reference region (plaza),  $R$ , and the total number of points (raindrops landing in the plaza),  $n$ , to become large in such a way that the expected density of points (intensity of rainfall) in each unit area remains the same. In our present case, this *expected density* is given by (5) as

$$(7) \quad \lambda(n, R) = \frac{n}{a(R)}$$

Hence to formalize the above idea, now imagine an increasing sequence of regions  $R_1 \subset R_2 \subset \dots \subset R_m \subset \dots$  and corresponding point totals  $n_1 < n_2 < \dots < n_m < \dots$  that expand such a way that the limiting density

<sup>3</sup> By definition  $E(X) = \sum_x x \cdot p(x) = 1 \cdot p(1) + 0 \cdot p(0) = p(1)$ .

$$(8) \quad \lambda = \lim_{m \rightarrow \infty} \lambda(n_m, R_m) = \lim_{m \rightarrow \infty} \frac{n_m}{a(R_m)}$$

exists and is positive. Under this assumption, it is shown in the Appendix (Section 1) that the Binomial probabilities in (6) converge to simple *Poisson probabilities*,

$$(9) \quad \Pr[N(C) = k | \lambda] = \frac{[\lambda a(C)]^k}{k!} e^{-\lambda a(C)}, \quad k = 0, 1, 2, \dots$$

Moreover, by (5) and (8), the expected number of points in any given cell (plaza tile),  $C$ , is now given by

$$(10) \quad E[N(C)] = \lambda a(C)$$

where density  $\lambda$  becomes the relevant constant of proportionality. Finally, if the set of random variables  $\{N(C)\}$  describing cell-counts for every cell of finite area in the plane is designated as a *spatial point process* on the plane, then any process governed by the Poisson probabilities in (9) is designated as a *spatial Poisson process* on the plane. Hence, when extended to the entire plane, the basic model of *complete spatial randomness* (CSR) above corresponds precisely to a spatial Poisson process.

## 2.4 Generalized Spatial Randomness

The basic notion of spatial randomness above was derived from the principle that regions of equal area should have the same chance of containing any given randomly located point. More formally, this *Spatial Laplace Principle* asserts that for any two subregions (cells),  $C_1$  and  $C_2$ , in  $R$ ,

$$(11) \quad a(C_1) = a(C_2) \Rightarrow \Pr[X(C_1) = 1 | R] = \Pr[X(C_2) = 1 | R]$$

However, as was noted in the Housing Abandonment example above, simple area may not always be the most relevant reference measure (backcloth). In particular, while one can imagine a randomly located abandoned house, such houses are very unlikely to appear in the middle of a public park, let alone the middle of a street. So here it makes much more sense to look at the existing *housing distribution*, at to treat a “randomly located abandoned house” as a random sample from this distribution. Here the Laplace principle is still at work, but now with respect to houses. For if housing abandonments are spatially random, then each *house* should have that same chance of being abandoned. Similarly, in the Larynx cancer example, if such cancers are spatially random, then each *individual* should have the same chance of contracting this disease. So here, the existing *population distribution* becomes the relevant reference measure.

To generalize the above notion of spatial randomness, we need only replace “area” with the relevant reference measure, say  $\rho(C)$ , which may be the “number of houses” in  $C$  or the “total population” of  $C$ . As an extension of (11) above, we then have the following *Generalized Spatial Laplace Principle*: For any two subregions (cells),  $C_1$  and  $C_2$ , in  $R$ :

$$(12) \quad \rho(C_1) = \rho(C_2) \Rightarrow \Pr[X(C_1) = 1 | R] = \Pr[X(C_2) = 1 | R]$$

If (11) is now replaced by (12), then one can essentially reproduce all of the results above. There is only one technicality that needs to be mentioned. The basic Laplace argument in Figure 1 above required that we be able to divide the square,  $S$ , into any number of equal-area cells. Hence the simplest way to extend this argument is to assume that the relevant reference measure,  $\rho$ , is *absolutely continuous* in the area measure,  $a$ . Here it suffices to assume that the relevant reference measure can be modeled in terms of a *density function* with respect to area.<sup>4</sup> So if housing (or population) is the relevant reference measure, then we model this in terms of a *housing density* (*population density*) with respect to area. Given this assumption, exactly the same arguments leading to (6) now show that

$$(13) \quad \Pr[N(C) = k | n, R] = \frac{n!}{k!(n-k)!} \left( \frac{\rho(C)}{\rho(R)} \right)^k \left( 1 - \frac{\rho(C)}{\rho(R)} \right)^{n-k}, \quad k = 0, 1, \dots, n$$

Similarly, if we now let  $\lambda(n, R) = n/\rho(R)$  and again assume the existence of limiting positive density

$$(14) \quad \lambda = \lim_{m \rightarrow \infty} \lambda(n_m, R_m) = \lim_{m \rightarrow \infty} \frac{n_m}{\rho(R_m)}$$

as the reference region becomes larger, then the same argument for (9) [in Section 1 of the Appendix] now shows that

$$(15) \quad \Pr[N(C) = k | \lambda] = \frac{[\lambda \rho(C)]^k}{k!} e^{-\lambda \rho(C)}, \quad k = 0, 1, 2, \dots$$

Spatial point processes governed by Poisson probabilities of this type (i.e., with non-uniform reference measures) are often referred to as *nonhomogeneous spatial Poisson processes*. Hence we shall often refer to this as the *nonhomogeneous CSR Hypothesis*.

<sup>4</sup> More formally, it is assumed that there is some “density” function,  $f$ , on  $R$  such that  $\rho$  is the integral of  $f$ , i.e., such that for any cell,  $C \subset R$ ,  $\rho(C) = \int_C f(x) dx$ .

## 2.5 Spatial Stationarity

Finally we consider a number of weaker versions of the spatial randomness model that will also prove to be useful. First observe that some processes may in fact be “Laplace like” in the sense that they look the same everywhere, but may not be completely random. A simple example is provided by the cell centers in Figure 1 of Section 1 above. Here one can imagine that if the microscope view were shifted to the left or right on the given cell slide, the basic pattern of cell centers would look very similar. Such point processes are said to be *stationary*. To make this notion more precise, it is convenient to think of each subregion  $C \subset R$  as a “window” through which one can see only part of larger point process on all of region  $R$ . In these terms, the most important notion of stationarity for our purposes is one in which the process seen in  $C$  remains the same no matter how we move this window. Consider for example the pattern of trees in a large rain-forest,  $R$ , part of which is shown in Figure 2 below. Here again this pattern is much too uniform to be completely random, but nonetheless appears to be the same everywhere. Suppose that the relevant subregion,  $C$ , under study corresponds to the small square in the lower left. In these terms, the appropriate notion of stationarity for our purposes amounts to the assumption that the cell-count distribution in  $C$  will remain the

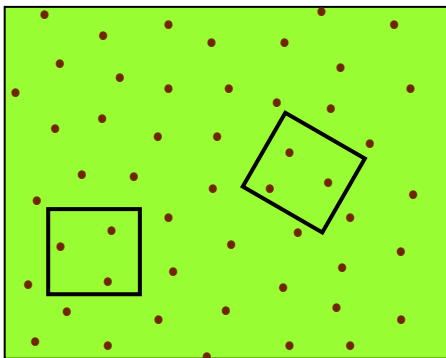


Fig.2. Isotropic Stationarity

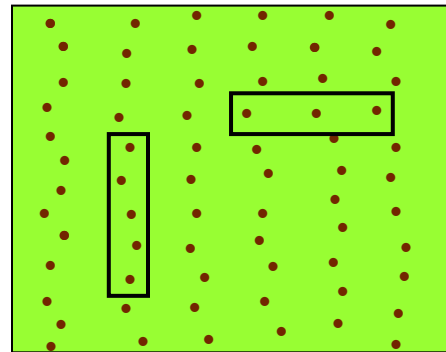


Fig.3. Anisotropic Stationarity

same no matter where this subregion is located. For example the tilted square shown in the figure is one possible relocation (or copy) of  $C$  in  $R$ . More generally if cell,  $C_2$ , is simply a translation and/or rotation of cell,  $C_1$ , then these cells are said to be geometrically *congruent*, written  $C_1 \cong C_2$ . Hence our formal definition of *stationarity* asserts that the cell-count distributions for congruent cells are the same, i.e., that for any  $C_1, C_2 \subset R$

$$(16) \quad C_1 \cong C_2 \Rightarrow \Pr[N(C_1) = k] = \Pr[N(C_2) = k], \quad k = 0, 1, \dots$$

Since the directional orientation of cells make no difference, this is also called *isotropic stationarity*. There is a weaker form of stationarity in which directional variations are

allowed, i.e., in which (16) is only required to hold for cells that are translations of one another. This type of *anisotropic stationarity* is illustrated by the tree pattern in Figure 3, where the underlying point process tends to produce vertical alignments of trees (more like an orchard than a forest). Here the variation in cell counts can be expected to differ depending on cell orientation. For example the vertical cell in Figure 3 is more likely to contain extreme point counts than its horizontal counterpart. (We shall see a similar distinction made for continuous stationary processes in Part II of this NOTEBOOK.)

One basic consequence of both forms of stationarity is that *mean* point counts continue to be *proportional to area*, as in the case of complete randomness, i.e. that

$$(17) \quad E[N(C)] = \lambda \cdot a(C)$$

where  $\lambda$  is again the expected point density (i.e., expected number of points per unit area). To see this, note simply that the basic Laplace argument in Figure 1 of Section 1 depends only on similarities among individual cells in uniform grids of cells. But since such cells are all translations of one another, it now follows from (16) that they all have the same cell-count distributions, and hence have the same means. So by the same argument above (with cell occupancy probabilities now replaced by mean point counts) it follows that such mean counts must gain be proportional to area. Thus while there can be many types of statistical dependencies between counts in congruent cells (as in the uniform tree patterns above), the *expected* numbers of points must be the same in each.

One final point should be made about stationarity. This concept implicitly assumes that the reference region,  $R$ , is sufficiently large to ensure that the relevant cells  $C$  never intersect the boundary of  $R$ . Since this rarely happens in practice, the present notion of stationarity is best regarded as a convenient fiction. For example, suppose that in the rain-forest illustrated in Figure 2 above there is actually a lake, as shown in Figure 4 below. In this case, any copies of the given (vertical) cell that lie in the lake will of course contain no trees. More generally, those cells that intersect that lake are likely to have fewer trees, such as the tilted cell in the figure. Here it is clear that condition (16) cannot possibly hold. Such violations of (16) are often referred to as *edge effects*.

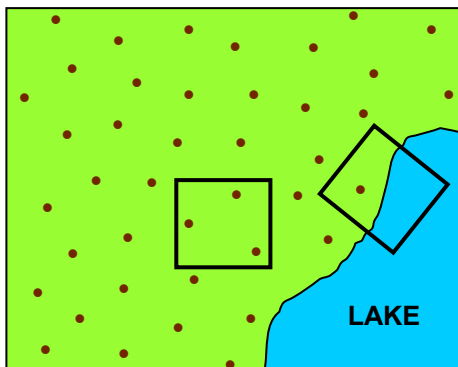


Fig.4. Actual Landscape

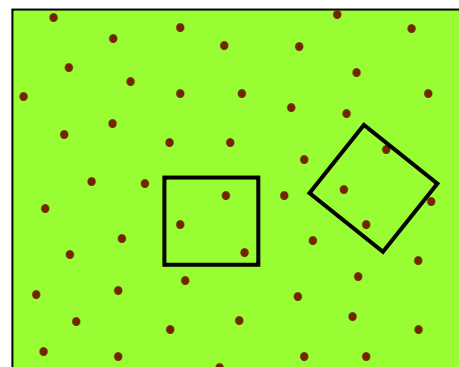


Fig.5. Stationary Version

Here there are two approaches that one can adopt. The first is to disallow any cells that intersect the lake, and thus to create a *buffer zone* around the lake. While this is no doubt effective, it has the disadvantage of excluding some points near the lake. If the forest,  $R$ , is large, this will probably make little difference. But if  $R$  is small (say not much bigger than the section shown) then this amounts to throwing away valuable data. An alternative approach is to ignore the lake altogether and to imagine a “stationary version” of this landscape, such as that shown in Figure 5. Here there are seen to be more points than were actually counted in this cell. So the question is then how to *estimate* these missing points. A method for doing so (known as *Ripley’s correction* ) will be discussed further in Section 4.3 below.