5. Comparative Analyses of Point Patterns

Up to this point, our analysis of point patterns has focused on single point patterns, such as the locations of redwood seedlings or lung cancer cases. But often the relevant questions of interest involve relationships between more than one pattern. For example if one considers a forest in which redwoods are found, there will invariably be other species competing with redwoods for nourishment and sunlight. Hence this competition between species may be of primary interest. In the case of lung cancers, recall from Section 1.2 that the lung cancer data for Lancashire was primarily of interest as a reference population for studying the smaller pattern of larynx cancers. We shall return to this example in Section 5.8 below. But for the moment we start with a simple forest example involving two species.

5.1 Forest Example

The 600 foot square section of forest shown in Figure 5.1 below contains only two types of trees. The large dots represent the locations of *oak* trees, and the small dots represent locations of *maple* trees. Although this is a fairly small section of forest, it seems clear that the pattern of oaks is much more clustered than that of maples. This is not surprising, given the very different seed-dispersal patterns of these two types of trees.

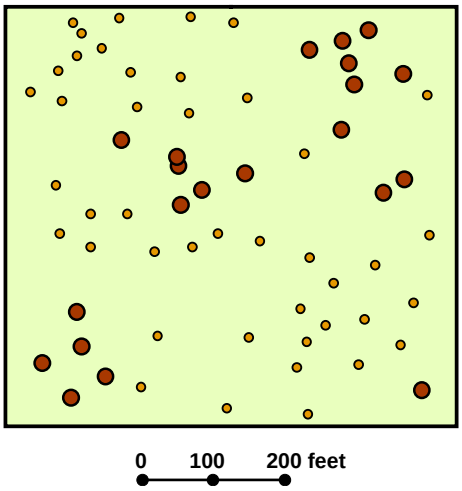

Figure 5.1. Section of Forest

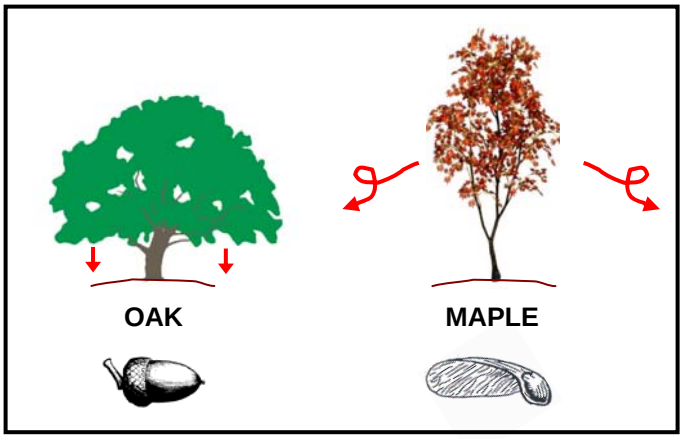

Figure 5.2. Patterns of Seed Dispersal

As shown in Figure 5.2, oaks produce largest acorns that fall directly from the tree, and are only partially dispersed by squirrels. Maples on the other hand produce seeds with individual "wings" that can transport each seed a considerable distance with even the slightest breeze. Hence there are clear biological reasons why the distribution of oaks might be more clustered than that of maples. So how might we test this hypothesis statistically?

5.2 Cross K-Functions

As one approach to this question, observe that if oaks tend to occur in clusters, then one should expect to find that the neighbors of oak trees tend to be other oaks, rather than maples. Hence one way to measure this effect would be to look at the nearest neighbors of oaks and see whether they tend to be predominantly oaks. But as we have already seen in the Bodmin tors example, this does not allow any analysis of the *scale* of oak clusters. Hence a more flexible approach is to extend the above K-function analysis for single populations to a comparable approach for comparing two populations.[1]

The idea is simple. Rather than looking at the expected number of oak trees within distance h of a given oak, we look at the expected number of *maple* trees within distance h of the oak. More generally, if we now consider two point populations, 1 and 2, with respective intensities, λ_1 and λ_2, and denote the members of these two populations by i and j, respectively, then the *cross K-function*, $K_{12}(h)$, for population 1 with respect to population 2 is given for each distance h by the following extension of expression (4.2.1) above:

(5.2.1) $$K_{12}(h) = \tfrac{1}{\lambda_2} E(\text{number of } j\text{-}\textit{events} \text{ within distance } h \text{ of an arbitrary } i\text{-}\textit{event})$$

Notice that there is an asymmetry in this definition, and that in general, $K_{12}(h) \neq K_{21}(h)$. Notice also that the word "additional" in (4.2.1) is no longer meaningful, since populations 1 and 2 are assumed to be distinct. This definition can be formalized in a manner paralleling the single population case as follows. First for any realized point patterns, $S_1 = (s_i : i = 1,..,n_1)$ and $S_2 = (s_i : i = 1,..,n_2)$, from populations 1 and 2 in region R, let $d_{ij} = d(s_i, s_j)$ denote the distance between member i of population 1 and j of population 2 in R. Then for each distance h the indicator function

(5.2.2) $$I_h(d_{ij}) = I_h[d(s_i, s_j)] = \begin{cases} 1 & , d_{ij} \le h \\ 0 & , d_{ij} > h \end{cases}$$

now indicates whether or member j of population 2 is within distance h of a given member i of population 1. In terms of this indicator, the *cross K-function* in (5.2.1) can be formalized [in a manner paralleling (4.3.3)] as

(5.2.3) $$K_{12}(h) = \tfrac{1}{\lambda_2} E\left[\sum_{j=1}^{n_2} I_h(d_{ij})\right]$$

[1] Note that while our present focus is on two populations, analyses of more than two populations are usually formulated either as (i) pairwise comparisons between these populations (as with correlation analyses), or (ii) comparisons between each population and the aggregate of all other populations. Hence the two-population case is the natural paradigm for both these approaches.

where both the size, n_2, of population 2 and the distances $(d_{ij} : j = 1,..,n_2)$ are here regarded as random variables.[2] This function plays a fundamental role in our subsequent comparative analyses of populations.

5.3 Estimation of Cross K-Functions

Given the definition in (5.2.3) it is immediately apparent that cross K-functions can be estimated in precisely the same way as K-functions. First, since the expectation in (5.2.3) does not depend on which random reference point i is selected from population 1, the same argument as in (4.3.4) now shows that for any given size, n_1, of population 1,

$$(5.3.1) \qquad E\left[\sum_{j=1}^{n_2} I_h(d_{ij})\right] = \lambda_2 K_{12}(h) \ , \ i = 1,..,n_1$$

$$\Rightarrow \ \sum_{i=1}^{n_1} E\left[\sum_{j=1}^{n_2} I_h(d_{ij})\right] = n_1 \lambda_2 K_{12}(h)$$

so that for each n_1, $K_{12}(h)$ can be written as[3]

$$(5.3.2) \qquad K_{12}(h) = \tfrac{1}{\lambda_2 n_1} \sum_{i=1}^{n} E\left[\sum_{j=1}^{n_2} I_h(d_{ij})\right]$$

In this form, it is again apparent that for any given realized patterns, $S_1 = (s_{1i} : i = 1,..,n_1)$ and $S_2 = (s_{2j} : j = 1,..,n_2)$, the expected counts in (5.3.2) are naturally estimated by their corresponding *observed counts*, and that the intensities, λ_1 and λ_2, are again estimated by the *observed intensities*,

$$(5.3.3) \qquad \hat{\lambda}_k = \frac{n_k}{a(R)} \ , \ k = 1,2$$

Thus the natural (maximum likelihood) estimate of $K_{12}(h)$ is given by the *sample cross K-function*:

$$(5.3.4) \qquad \boxed{\hat{K}_{12}(h) = \tfrac{1}{\hat{\lambda}_2 n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_h(d_{ij})}$$

---

[2] To be more precise, n_2 is a random integer (count), and for any given value of n_2, the conditional distribution of $[d_{ij} = d(s_i, s_j) : j = 1,..,n_2]$ is then determined by the conditional distribution of the locations, $[s_i, (s_j : j = 1,..,n_2)]$ in R, where s_i is implicitly taken to be the location of a randomly sampled member of population 1.

[3] Technically this should be written as a conditional expectation *given* n_1 [and (4.3.4) should be a conditional expectation *given* n]. But for simplicity, we ignore this additional layer of notation.

5.4 Spatial Independence Hypothesis

We next use these sample cross K-functions as test statistics for comparing populations 1 and 2. Recall that in the single population case, the fundamental question of interest was whether or not the given population was more clustered (or more dispersed) than would be expected if the population locations were completely random. This led to the *CSR hypothesis* as a natural null hypothesis for testing purposes. However, when one compares two populations of random events, the key question is usually whether or not these events influence one another in some way. So here the natural null hypothesis takes the form of *statistical independence* rather than randomness. In terms of cross K-functions, if there are significantly more j-events close to i-events than would be expected under independence, then one may infer that there is some "attraction" between populations 1 and 2. Conversely, if there are significantly fewer j-events close to i-event than expected, then one may infer that there is some "repulsion" between these populations. These basic distinctions between the one-population and two-population cases can be summarized as in Table 5.1 below:

CASE	HYPOTHESIS FRAMEWORK		
One Pop	Clustering ←	Spatial Randomness	→ Dispersion
Two Pops	Attraction ←	Spatial Independence	→ Repulsion

Figure 5.3. Comparison of Hypothesis Frameworks

Next we observe that from a testing viewpoint, the particular appeal of the CSR hypothesis is that one can easily simulate location patterns under this hypothesis. Hence Monte Carlo testing is completely straightforward. But the two-population hypothesis of spatial independence is far more complex. In principle this would not be a problem if one were able to observe many replications of these sets of events, i.e., many replications of joint patterns from populations 1 and 2. But this is almost never the case. Typically we are given a *single* joint pattern (such as the patterns of oaks and maples in Figure 5.1 above) and must somehow detect "departures from independence" using only this single realization. Hence it is necessary to make further assumptions, and in particular, to define "spatial independence" in a manner that allows the distribution of sample cross K-functions to be simulated under this hypothesis. Here we consider two approaches, designated respectively as the *random-shift approach* and the *random-permutation approach*.

5.5 Random-Shift Approach to Spatial Independence

This approach starts by postulating that each individual population $k = 1, 2$ is generated by a *stationary process* on the plane. If region R is viewed as a window on this process (as in Section 2) and we again represent each process by the collection of cell counts in

R, say $\mathcal{N}_k = \{N_k(C) : C \subset R\}$, $k = 1,2$, then it follows in particular from (2.5.1) that the *marginal* cell-count distribution, $\Pr[N_k(C_h)]$ for population k in any circular cell, C_h, of radius h must be the same for all locations.[4] Hence if we now focus on population 2 and imagine a two-stage process in which (i) a point pattern for population 2 is generated, and (ii) this pattern is then shifted by adding some constant vector, a, to each point, $s_j \to s_j + a$, then the *expected* number of points in C_h would be the same for both stage (i) and stage (ii). Indeed this shift simply changes the location of C_h relative to the pattern (as in Figure 5.5 below) so that by stationarity the expected point count must stay the same.

5.5.1 Spatial Independence Hypothesis for Random Shifts

In this context, the appropriate *spatial independence hypothesis* simply asserts that cell counts for population 2 are not influenced by the locations of population 1, i.e., that for all cells, $C \subset R$,

$$\Pr[N_2(C) = n \mid \mathcal{N}_1] = \Pr[N_2(C) = n] \ , \quad n \geq 0 \tag{5.5.1}$$

where $\Pr[N_2(C) = n \mid \mathcal{N}_1]$ is the conditional probability that $N_2(C) = n$ given all cell counts, $\mathcal{N}_1$, for population 1.[5] Under this hypothesis it then follows that the conditional distribution on the left must also exhibit stationarity, so that if the circular cell, C_h, is centered at the location of a point s_i in population 1, this will make no difference. To illustrate the substantive meaning of this hypothesis in the presence of stationarity, suppose that populations 1 and 2 are plant species in which the root system of species 1 is toxic to species 2, so that no plant of species 2 can survive within two feet of any species 1 plant. Then consider a two stage process in which the plant locations of species 1 and 2 are first generated at random, and then all species 2 plants within two feet of any species 1 plant are removed.[6] Then it is not hard to see that the *marginal* process for population 2 will still exhibit stationarity (since locations of population 1 are equally likely to be anywhere). But the *conditional* process for population 2 *given* the locations of population 1 is highly non-stationary, and indeed must have zero cell counts for all two-foot cells around population 1 sites.

Now returning to the two-stage "shift" process described above, this process suggests a natural way of testing the independence hypothesis in (5.5.1) using sample cross K-functions. In particular, if the given realization of population 2 is randomly shifted in any way, then this should not affect the *expected* counts,

[4] For the present, we implicitly assume that region *R* is "sufficiently large" that edge effects can be ignored.
[5] Note that while there is an apparent asymmetry in this definition between populations 1 and 2, the definition of conditional probability implies that (5.5.1) must also hold with labels 1 and 2 reversed.
[6] This is an instance of what is called a "hard-core" process in the literature (as for example in Ripley, 1977, section 3.2 and Cressie, 1995, section 8.5.4).

(5.5.2) $$E\{N_1[C_h(s_i)]\} = E\left[\sum_{j=1}^{n_2} I_h(d_{ij})\right]$$

of population 2 events within distance h of any population 1 event, s_i. This in turn implies from (5.3.2) that the cross K-function should remain the same for all such shifts (remember that cross K-functions are *expected* values). Hence if one were to randomly sample shifted versions of the given pattern and construct the corresponding statistical population of sample cross K-functions, then this population could be used to test for spatial independence in exactly the same way that the CSR hypothesis was tested using K-functions. This testing scheme is in principle very appealing since it provides a direct test of the spatial independence hypothesis that *preserves the marginal distribution of both populations*.

5.5.2 Problem of Edge Effects

But in its present form, such a test it is not practically possible since we are only able to observe these processes in a *bounded* region, *R*. Thus any attempt to "shift" the pattern for population 2 will require knowledge of the pattern *outside* this window, as shown in Figures 5.4 and 5.5 below. Here the black dots represent unknown sites of population 2 events. Hence any shift of the pattern relative to region *R* will allow the possible entry of unknown population 2 events into the window defined by region *R*.

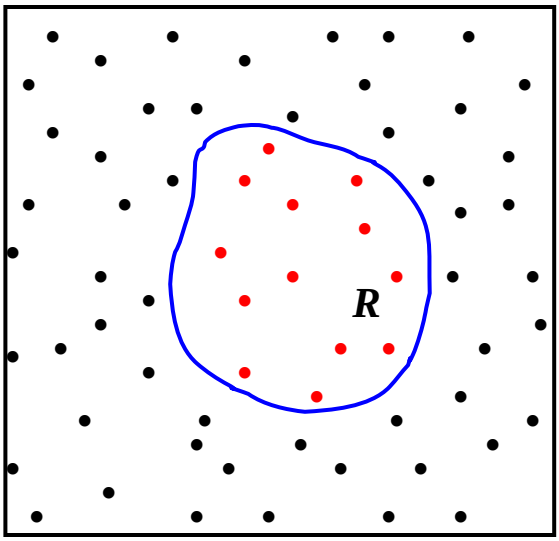

Figure 5.4. Pattern for Population 2

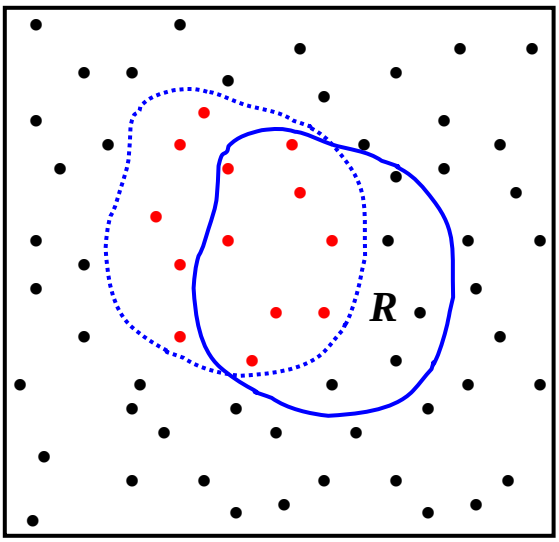

Figure 5.5. Randomly Shifted Pattern

However, it turns out that under certain conditions one can construct a reasonable approximation to this ideal testing scheme. In particular, if the given region *R* is *rectangular*, then there is indeed a way of approximating *stationary* point processes outside the observable rectangular window. To see this, suppose we start with the two point patterns in a rectangular boundary, *R*, as shown in Figure 5.6 below (with pattern 1

= white dots and pattern 2 = black dots).[7] If these patterns are in fact generated by stationary point processes on the plane, then in particular, the realized pattern, $S_2^0 = (s_{2j}^0 : j = 1,..,n_2)$, for population 2 (shown separately in Figure 5.7 below) could equally well have occurred in any shifted version of region *R*.

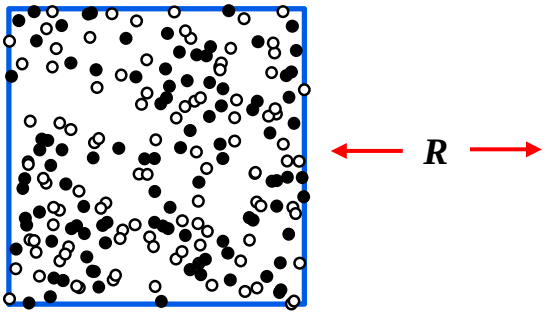

Figure 5.6. Rectangular Region

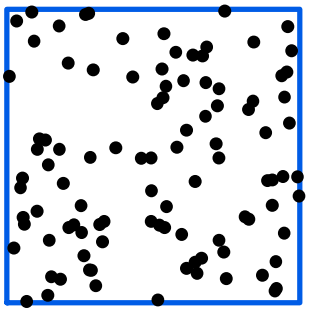

Figure 5.7. Population 2

But since the rectangularity of *R* implies that the entire plane can be filled by a "tiling" of *disjoint* copies of region *R* (also called a "lattice" of translations of *R*) and since this same point pattern can be treated as a typical realization in each copy of *R*, we can in principle extend the given pattern in region *R* to the entire plane by simply reproducing this pattern in each copy of *R* [as shown partially in Figure 5.8 below].[8] We designate this infinite version of pattern S_2^0 by $\tilde{S}_2^0$.

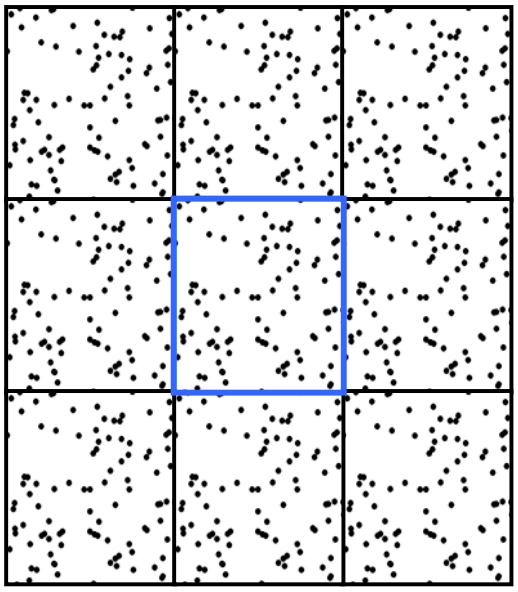

Figure 5.8. Partial Tiling

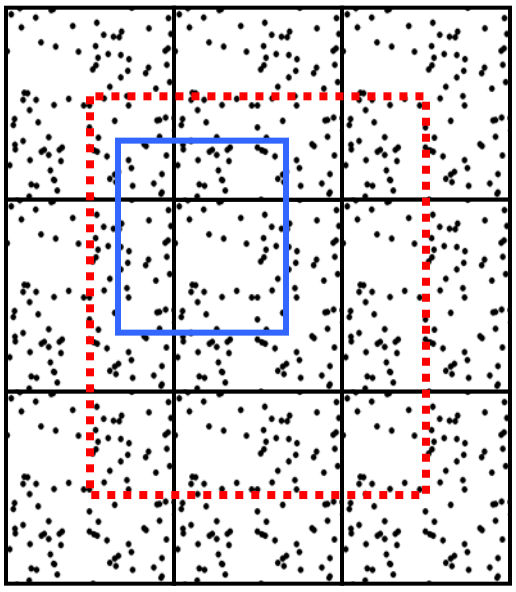

Figure 5.9. Random Shifts

[7] This example is taken from Smith (2004).
[8] Such replications are also called "rectangular patterns with periodic boundary conditions" (see for example Ripley, 1977 and Diggle, 1983, section 1.3).

In this way, we can effectively remove the "edge effects" illustrated in Figure 5.5 above. Moreover, while the "replication process" that generates $\tilde{S}_2^0$ must of course exhibit stronger symmetry properties than the original process for population 2, it can be shown that this process shares the *same mean and covariance structure* as the original process. Moreover, it can also be shown that under the *spatial independence hypothesis*, the cross K-function yielded by this process must be the *same* as for the original process.[9] Hence for the case of rectangular regions, *R*, it is possible to carry this replicated version of the "ideal" testing procedure described above.

5.5.3 Random Shift Test

To make this test explicit, we start by observing that it suffices to consider only *local* random shifts. To see this, note first that if point pattern 1 in Figure 5.6 is designated by $S_1^0 = (s_{1i}^0 : i = 1,..,n_i)$, then shifting $\tilde{S}_2^0$ relative to S_1^0 on the plane is completely equivalent to shifting S_1^0 relative to $\tilde{S}_2^0$. Hence we need only consider shifts of S_1^0. Next observe by symmetry that every distinct rectangular portion $\tilde{S}_2^0$ that can occur in shifted versions of *R* (such as the pattern inside the blue box of Figure 5.8) can be obtained at some position of *R* inside the red dotted boundary shown Figure 5.8. Hence we need only consider random shifts of S_1^0 within this boundary. Again, the blue box in Figure 5.8 represents one such shift (where the white dots for population 1 have been omitted for sake of visual clarity). Hence to construct the desired *random-shift test*, we can use the following procedure:

(i) Simulate N random shifts that will keep rectangle *R* inside the feasible region in Figure 5.9. Then shift all coordinates in S_1^0 by this same amount.

(ii) If $S_2^m = (s_{2j}^m : j = 1,..,n_2^m)$ denotes the pattern for population 2 occurring in random shift $m = 1,..,N$ of rectangle *R* (which will usually be of a slightly different size than S_2^0), then a *sample cross K-function*, $\hat{K}_{12}^m(h)$, can be constructed from S_1^0 and S_2^m. In particular if the relevant set of distance radii is chosen to be $D = \{h_w : w = 1,..,W\}$, then the actual values constructed are $\{\hat{K}_{12}^m(h_w) : w = 1,..,W\}$.

(iii) Finally, if the *observed sample cross K-function*, $\hat{K}_{12}^0(h)$, is constructed in the same way from S_1^0 and S_2^0 (where the latter pattern is equivalent to the "zero shift" denoted by the central box in Figure 5.8), then under the *spatial independence hypothesis*, (5.5.1), each observed value, $\hat{K}_{12}^0(h_w)$, should be a "typical" sample from the list of values $[\hat{K}_{12}^m(h_w) : m = 0,1,..,N]$. Hence (in a manner completely analogous to the single-population tests of CSR), if we now let M_+^0 denote the number of

[9] See the original paper by Lotwick and Silverman (1982) for proofs of these facts.

simulated random shifts, $m = 1,..,N$, with $\hat{K}_{12}^{m}(h_w) \geq \hat{K}_{12}^{0}(h_w)$, then the estimated probability of obtaining a value *as large as* $\hat{K}_{12}^{0}(h_w)$ under this *spatial independence hypothesis* is given by the *attraction p-value*,

(5.5.3) $$\hat{P}_{attraction}(h_w) = \frac{M_{+}^{0} + 1}{N + 1}$$

where small values of $\hat{P}_{attraction}(h_w)$ can be interpreted as implying significant *attraction* between populations 1 and 2 at *scale* h_w.

(iv) Similarly, if M_{-}^{0} denotes the number of simulated random shifts, $m = 1,..,N$, with $\hat{K}_{12}^{m}(h_w) \leq \hat{K}_{12}^{0}(h_w)$, then the estimated probability of obtaining a value *as small as* $\hat{K}_{12}^{0}(h_w)$ under this *spatial independence hypothesis* is given by the *repulsion p-value*,

(5.5.4) $$\hat{P}_{repulsion}(h_w) = \frac{M_{-}^{0} + 1}{N + 1}$$

where small values of $\hat{P}_{repulsion}(h_w)$ can be interpreted as implying significant *repulsion* between populations 1 and 2 at *scale* h_w.

5.5.4 Application to the Forest Example

This testing procedure is implemented in the MATLAB program, **k12_shift_plot.m**, and can be applied to the Forest example above as follows. The forest data appears in the ARCMAP file, **Forest.mxd**, and was exported to the MATLAB workspace, **forest.mat**. The coordinate locations of the $n_1 = 21$ oaks and $n_2 = 43$ maples are given in matrices, **L1** and **L2**, respectively. An examination of these locations in ARCMAP (or in Figure 5.1 above) suggested that a reasonable range of radial distances to consider is from 10 to 330 feet, and the set of (14) distance values, **D = [10:20:270]**,[10] was chosen for analysis. The rectangular region, *R*, in Figure 5.1 is seen in ARCMAP to be defined by the bounding values, (**xmin = -10, xmax = 589, ymin = 20, ymax = 577**). Using these parameters, the command;

>> PVal = k12_shift_plot(L1,L2,xmin,xmax,ymin,ymax,999,D);

yields a vector of *attraction p-values* (5.5.3) at each radial distance in **D** based on 999 simulated random shifts of the maples relative to the oaks. Recall that in this example, an inspection of Figure 5.1 suggested that there are "island clusters" of oaks in a "sea" of

[10] In MATLAB this yields a list **D** of values from **10** to **330** in increments of **20**. (See also p.5-23 below.)

maples. Hence, in terms of attraction versus repulsion, this suggests that there is some degree of *repulsion* between oaks and maples. Thus one must be careful when interpreting the p-value output, **PVal**, of this program.

Recall that as with clustering versus dispersion, unless there are many simulated cross K-function values exactly equal to $\hat{K}_{12}^{0}(h_k)$, we will have $\hat{P}_{replusion}(h_k) \approx 1 - \hat{P}_{attraction}(h_k)$.

Hence one can identify significant repulsion by plotting $\hat{P}_{attraction}(h_k)$ for $k = 1,..,K$ and looking for *large* p-values. This plot is given as screen output for **k12_shift_plot.m**, and is illustrated in Figure 5.10 below for a simulation with $N = 999$:

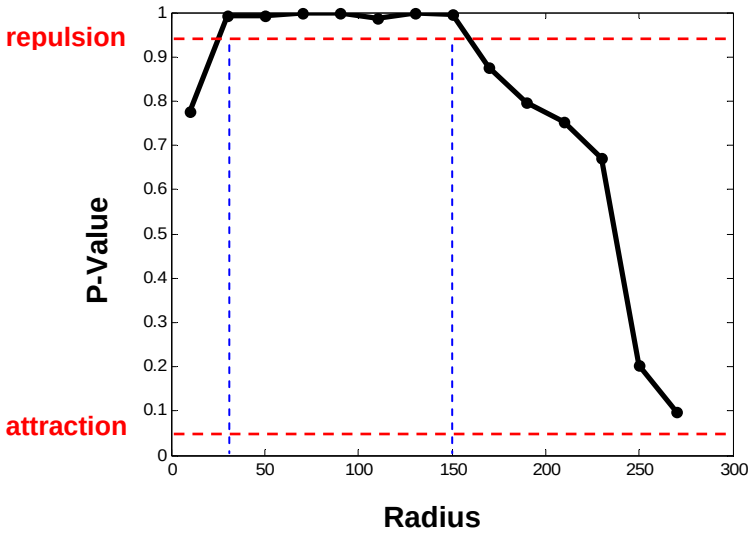

Figure 5.10 Random Shift P-Values

Here the red dashed line on the bottom corresponds to a attraction p-value of .05, so that values below this level denote significant *attraction* at the .05 level. Similarly the red dashed line at the top corresponds to an attraction p-value of .95, so that values above this line denote significant *repulsion* at the .05 level. Hence there appears to be *significant repulsion* between oaks and maples at scales $30 \le h \le 150$. This is seen to be in reasonable agreement with a visual inspection of Figure 5.1 above.

But while this test is reasonable in the present case, this is in large part due to the presence of a *rectangular* region, *R*. More generally, in the cases such as large forests where analyses of "typical" rectangular regions usually suffice, this is not much of a restriction. But for point patterns in regions, *R*, such as the elongated island shown in Figure 5.10, it is clear from the figure that any attempt to reduce *R* to a rectangle might remove most of the relevant pattern data.

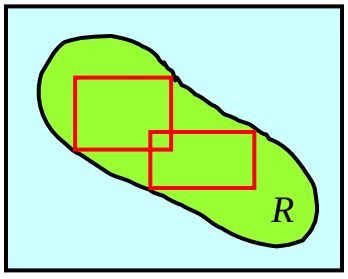

Figure 5.10 Island Example

This island example also raises another important limitation of the random-shift approach when comparing point patterns. Recall that this approach treats the given region, *R*, as a sample "window" from a much larger realization of point patterns, so that the hypothesis of stationarity is at least meaningful in principle. But the shoreline of an island is *physical barrier* between very different ecological systems. So if the point patterns were trees (as in the oak-maple example) then the shoreline is not simply an "edge effect". Indeed the very concept of stationarity is at best artificial in such applications.

5.6. Random-Labeling Approach to Spatial Independence

An approach which overcomes many of these problems is based on an alternative characterization of multiple-population processes. Rather than focusing on the individual processes generating patterns $S_1 = (s_{1i} : i = 1,..,n_1)$ and $S_2 = (s_{2j} : j = 1,..,n_2)$ above, one can characterize this realized joint pattern in and entirely different way. Suppose we let $n = n_1 + n_2$ denote the total number of events generated, and associate with each event, $i = 1,..,n$, a pair (s_i, m_i) where $s_i \in R$ is the *location* of event i in *R*, and $m_i \in \{1,2\}$ is a *marker* (or *label*) denoting whether event i is of type 1 or 2. Stochastic processes generating such pairings of joint locations and labels for each event are called *marked point processes*.[11] The Forest example above can be regarded as the realization of a marked point process where the number of events is $n = 21 + 43 = 64$, and the possible labels for each event are "oak" and "maple". Clearly each realized set of values, $[(s_i, m_i) : i = 1,..,n]$, yields a complete description of a joint pattern pair (S_1, S_2) above. The key advantage of this particular characterization is that it allows the location process to be *separated* from the distribution of event types.

This is particularly relevant in situations where the location process is complex, or where the set of feasible locations may involve a host of unobserved restrictions. As a simple illustration, suppose that in the Forest example there were in fact a number of subsurface rock formations, denoted by the gray regions in Figure 5.11, that prevented the growth of any large trees in these areas. Then even if these rock formations are not observed (and thus impossible to model), the *observed* locations of trees must surely avoid these areas. Hence if one were to *condition* on these observed locations, then it would still possible to analyze certain relations between oaks and maples without the need to model all feasible locations.

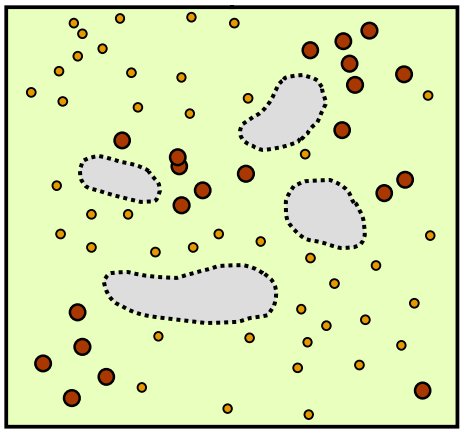

Figure 5.10 Location Restrictions

[11] The following development is based on the treatment in Cox and Isham (1980). For a nice overview discussion, see Diggle (2003.pp.82-83), and for a deeper analysis of marked spatial point processes, see Cressie (1993, section 8.7).

More generally, by conditioning on the observed set of locations, one can compare a wide variety of point populations without the need to identify alternative locations at all. Not only does this circumvent all problems related to the shape of region, *R*, but it also avoids the need to identify specific land-use constraints (such street networks or zoning restrictions) that may influence the locations of relevant point events (like housing sales or traffic accidents).

5.6.1 Spatial Indistinguishability Hypothesis

To formalize an appropriate notion of spatial independence for population comparisons in the context of marked point processes, we start by considering the joint distribution of a set of n marked events,

$$(5.6.1) \qquad \Pr[(s_i, m_i) : i = 1,..,n] = \Pr[(s_1,..,s_n),(m_1,..,m_n)]$$

$$= \Pr[(m_1,..,m_n) \mid (s_1,..,s_n)] \cdot \Pr(s_1,..,s_n)$$

where $\Pr(s_1,..,s_n)$ denotes the *marginal* distribution of event locations, and where $\Pr[(m_1,..,m_n) \mid (s_1,..,s_n)]$ denotes the *conditional* distribution of event labels given their locations.[12] If $\Pr(m_1,..,m_n)$ denotes the corresponding marginal distribution of event labels, then the relevant hypothesis of *spatial independence* for our present purposes asserts simply that *event labels are not influenced by their locations*. i.e., that

$$(5.6.2) \qquad \boxed{\Pr[(m_1,..,m_n) \mid (s_1,..,s_n)] \equiv \Pr(m_1,..,m_n)}$$

for all locations $s_1,..,s_n \in R$ and labels $m_1,..,m_n \in \{1,2\}$. In the Forest example above, for instance, the hypothesis that there is no spatial relationship between oaks and maples is here taken to mean that the given set of tree locations, $(s_1,..,s_n)$, tell us nothing about whether these locations are occupied by oaks or maples. Hence the only locational assumption implicit in this hypothesis is that any observed tree location *could* be occupied by either an oak or a maple. Note also that this doesn't mean that oaks and maples are equally likely events. Indeed if there are many more maples than oaks, then all of this information is captured in the distribution of labels, $\Pr(m_1,..,m_n)$.

As with the random shift approach (where the marginal distributions of each population were required to be stationary), we do require one additional assumption about the marginal distribution of labels, $\Pr(m_1,..,m_n)$. Note in particular that the indexing of events, $1,2,..,n$, only serves to distinguish them, and that their particular ordering has no

[12] For simplicity we take the number of events, *n*, to be fixed. Alternatively, the distributions in (5.6.1) can all be viewed as being conditioned on *n*.

relevance whatsoever.[13] Hence the likelihood of labeling events, $(m_1,..,m_n)$, should not depend on which event is called "1", and so on. This *exchangeability* condition can be formalized by requiring that for all permutations $(\pi_1,..,\pi_n)$ of the subscripts $(1,..,n)$,[14]

(5.6.3) $$\Pr(m_{\pi_1},..,m_{\pi_n}) = \Pr(m_1,..,m_n)$$

These two conditions together imply that the point processes generating populations 1 and 2 are essentially indistinguishable. Hence we now designate the combination of conditions, (5.6.2) and (5.6.3) as the *spatial indistinguishability hypothesis* for populations 1 and 2. This hypothesis will form the basis for many of the tests to follow.

5.6.2 Random Labeling Test

To test the spatial indistinguishability hypothesis, [(5.6.2),(5.6.3)], our objective is to show that for any observed set of locations $(s_1,..,s_n)$ and population sizes n_1 and n_2 with $n_1 + n_2 = n$, all possible labelings of events must be *equally likely* under this hypothesis. This in turn will give us an exact sampling distribution that will allow us to construct Monte Carlo tests of (5.6.2).

To do so, we begin by observing that in the same way that stationarity of marginal distributions was inherited by conditional distributions in (5.5.1) above, it now follows that exchangeability of labeling events in (5.6.3) is inherited by the corresponding conditional events in (5.6.2). To see this, observe simply that for any given set of locations $(s_1,..,s_n)$ and subscript permutation $(\pi_1,..,\pi_n)$ it follows at once from (5.6.2) and (5.6.3) that

(5.6.4) $$\Pr[(m_{\pi_1},..,m_{\pi_n})\,|\,(s_1,..,s_n)] = \Pr(m_{\pi_1},..,m_{\pi_n})$$

$$= \Pr(m_1,..,m_n) = \Pr[(m_1,..,m_n)\,|\,(s_1,..,s_n)]$$

To complete the desired task, it is enough to observe that for any two labelings, $(m_1,..,m_n)$ and $(m'_1,..,m'_n)$ consistent with n_1 and n_2 we must have

(5.6.5) $$(m'_1,..,m'_n) = (m_{\pi_1},..,m_{\pi_n})$$

for some permutation, $(\pi_1,..,\pi_n)$. Hence if the conditional distribution of such labels given both $(s_1,..,s_n)$ and (n_1,n_2) is denoted by $\Pr[\,\cdot\,|\,(s_1,..,s_n),n_1,n_2]$, then it follows that:

[13] However, if one were to model the immergence of new events (such as new disease victims or new housing sales), then this ordering would indeed play a significant role.

[14] For example, possible permutations of $(1,2,3)$ include $(\pi_1,\pi_2,\pi_3) = (2,1,3)$ and $(\pi_1,\pi_2,\pi_3) = (3,2,1)$.

(5.6.6) $$\Pr[(m'_1,..,m'_n)\,|\,(s_1,..,s_n),n_1,n_2] = \Pr[(m_{\pi_1},..,m_{\pi_n})\,|\,(s_1,..,s_n),n_1,n_2]$$
$$= \Pr[(m_1,..,m_n)\,|\,(s_1,..,s_n),n_1,n_2]$$

Moreover, since these conditional labeling events are mutually exclusive and collectively exhaustive, it also follows that this set of permutations must yield a well-defined conditional probability distribution, i.e. that:

(5.6.7) $$\sum_{(\pi_1,..,\pi_n)} \Pr[(m_{\pi_1},..,m_{\pi_n})\,|\,(s_1,..,s_n),n_1,n_2] = 1$$

Finally, recalling that the number of permutations of $(1,..,n)$ is given by $n!$, we may conclude from (5.6.6) and (5.6.7) that for any observed event locations, $(s_1,..,s_n)$, and event labels, $(m_1,..,m_n)$, with corresponding population sizes, n_1 and n_2, we have the following exact conditional distribution for all permutations $(\pi_1,..,\pi_n)$ of these labels under the *spatial indistinguishability hypothesis*:[15]

(5.6.8) $$\boxed{\Pr[(m_{\pi_1},..,m_{\pi_n})\,|\,(s_1,..,s_n),n_1,n_2] \equiv \frac{1}{n!}}$$

This provides us with the desired sampling distribution for testing this hypothesis. In particular, the following procedure yields a *random-labeling test* of (5.6.2) that closely parallels the random-shift test above:

(i) Given *observed* locations, $(s_1,..,s_n)$, and labels $(m_1,..,m_n)$ with corresponding population sizes, n_1 and n_2, simulate N random permutations $[\pi_1(\tau),..,\pi_n(\tau)]$, $\tau = 1,..,N$ of $(1,..,n)$,[16] and form the permuted labels $(m_{\pi_1(\tau)},..,m_{\pi_n(\tau)})$, $\tau = 1,..,N$ [which is equivalent to taking a sample of size *N* from the distribution in (5.6.8)].

(ii) If $S_1^\tau = (s_{1i}^\tau : i = 1,..,n_1)$ and $S_2^\tau = (s_{2j}^\tau : j = 1,..,n_2)$ denote the patterns for populations 1 and 2 obtained from the joint realization, $[(s_1,..,s_n),(m_{\pi_1(\tau)},..,m_{\pi_n(\tau)})]$, and if $\hat{K}_{12}^\tau(h)$ denotes the *sample cross K-function* resulting from (S_1^τ, S_2^τ), then choose a relevant set of distance radii, $D = \{h_w : w = 1,..,W\}$, and calculate the sample cross K-function values, $\{\hat{K}_{12}^\tau(h_w) : w = 1,..,W\}$ for each $\tau = 1,..,N$.

[15] It should be noted that since $m_i \in \{1,2\}$ for each $i = 1,..,n$, many permutations $(m_{\pi_1},..,m_{\pi_n})$ will in fact be identical. Hence the probability of each *distinct* realization is $n_1!n_2!/n!$. But since it is easier to sample random permutations (as discussed in the next footnote) we choose to treat each permutation as realization.

[16] This is in fact a standard procedure in most software. In MATLAB, a random permutation of the integers $(1,..,n)$ is obtained with the command **randperm(n)**.

(iii) Finally, if the *observed sample cross K-function*, $\hat{K}_{12}^{0}(h)$, is constructed from the observed patterns, S_1^0 and S_2^0, then under the *spatial indistinguishability hypothesis* each observed value, $\hat{K}_{12}^{0}(h_w)$, should be a "typical" sample from the list of values $[\hat{K}_{12}^{\tau}(h_w): \tau = 0,1,..,N]$. Hence if we now let M_{+}^{0} denote the number of simulated random relabelings, $\tau = 1,..,N$, with $\hat{K}_{12}^{\tau}(h_w) \geq \hat{K}_{12}^{0}(h_w)$, then the estimated probability of obtaining a value *as large as* $\hat{K}_{12}^{0}(h_w)$ under this hypothesis is again given by the *attraction p-value* in (5.5.3) above.

(iv) Similarly, if M_{-}^{0} denotes the number of simulated random relabelings, $\tau = 1,..,N$, with $\hat{K}_{12}^{\tau}(h_w) \leq \hat{K}_{12}^{0}(h_w)$, then the estimated probability of obtaining a value *as small as* $\hat{K}_{12}^{0}(h_w)$ under this hypothesis is again given by the *repulsion p-value* in (5.5.4) above.

Before applying this test it is of interest to ask why simulation is required at all. Since the distribution in (5.6.8) is constant, why not simply calculate the values, $\Pr[\hat{K}_{12}^{\tau}(h_w) \geq \hat{K}_{12}^{0}(h_w)]$ for each $w = 1,..,W$? The difficulty here is that since there is no simple analytical expression for these probabilities, one must essentially enumerate the sample space of relabelings and check these inequalities case by case. But even for patterns as small as $n_1 = 10 = n_2$ the number of *distinct* relabelings to be checked is seen to be $20!/(10! \times 10!) = 184,756$. So even for small patterns, there are sufficiently many distinct relabelings to make Monte Carlo simulation the most efficient procedure for testing purposes.

Finally it is important to stress that while this random-labeling approach is clearly more flexible than the random-shift approach above, this flexibility is not achieved without some costs. In particular, the most appealing feature of the random shift test was its ability to preserve many key properties of the *marginal distributions* for populations 1 and 2. In the present approach, where the joint distribution is recast in terms of a location and labeling process, all properties of these marginal distributions are lost. So (as observed by Diggle, 2003, p.83) the present marked-point-process approach is most applicable in cases where there is a natural separation between location and labeling of population types. In the context of the Forest example above, a simple illustration would be the analysis of a disease affecting say maples. Here the two populations might be "healthy" and "diseased" maples. So in this case there is a single location process involving all maple trees, followed by a labeling process which represents the spread of disease among these trees.[17]

[17] An example of precisely this type involving "Myrtle Wilt", a disease specific to myrtle trees, is part of Assignment 2 in this course.

5.6.3 Application to the Forest Example

In a manner paralleling the random-shift test, this random-relabeling test is implemented in the MATLAB program, **k12_perm_plot.m**. If the observed locations of populations 1 and 2 are again denoted by **L1** and **L2**, and if **D** again denotes the set of selected radial distances, then a screen plot of attraction p-values for **999** simulations is now obtained by the command (where the final argument, "**1**", specifies that a random seed is to be used):

>> k12_perm_plot(L1,L2,999,D,1);

If this test is applied to the Forest example with the somewhat larger set of radial distance values, **D = [10:20:330]**, then a typical result is shown in Figure 5.11 below:

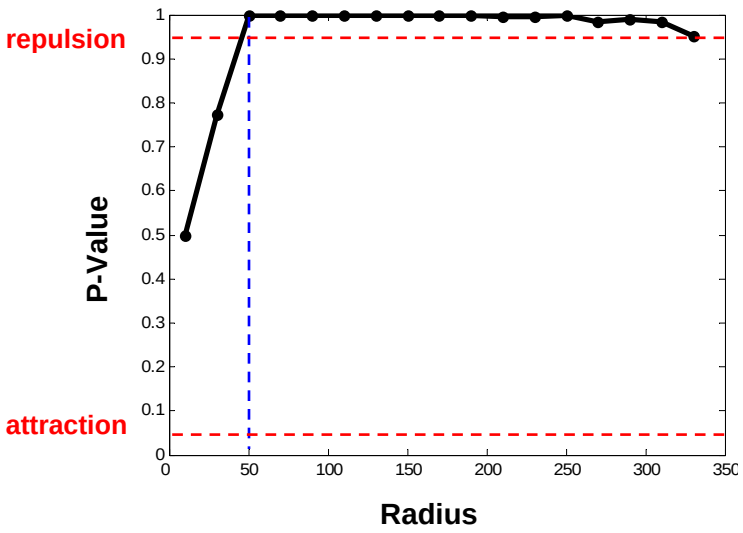

Figure 5.11 Random Relabeling P-Values

Here we see that the results are qualitatively similar to the random-shift test for short distances, but that repulsion is dramatically more extreme for long distances. Indeed significant repulsion now persists up to the largest possible relevant scale of 300 feet (= Dmax/2). Part of the reason for this can be seen in Figure 5.12 below, where a partial tiling of the maple pattern in Figure 5.1 is shown.

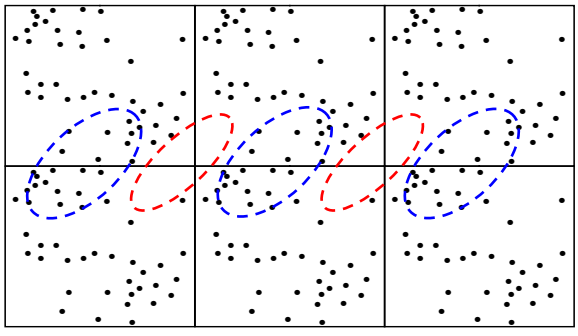

Figure 5.12 New Maple Structure

Even this small portion of the tiling reveals an additional hidden problem with the random-shift approach. For while this replication process statistically preserves the *means* of sample cross K-functions, the *variance* of these functions tends to increase. The reason for this is that tiling by its very nature tends to create *new structure* near the boundaries of the rectangle region, *R*.[18] In the present case, the red ellipses in Figure 5.11 represent larger areas devoid of maples than those in *R* itself (created mainly by the combination of empty areas in the lower left and upper right corners of *R*). Similarly the blue ellipses represent new clusters of maples larger that those in *R*. The result of this new structure in the present case is to make the tiled pattern, $\tilde{S}_2^0$, of maples appear somewhat more clustered at larger scales. This in turn yields higher levels of repulsion between oaks (S_1^0), and maples $(\tilde{S}_2^0)$ at these larger scales for most simulated shifts. The result of this is to make the *observed* level of repulsion between S_1^0 and S_2^0 appear relatively less significant at these larger scales, as reflected in the plot of Figure 5.10.[19]

5.7. Analysis of Spatial Similarity

The two procedures above allowed us to test whether there was significant "attraction" or "repulsion" between two patterns. This focuses on their *joint* distribution. Alternatively, we might simply compare their *marginal* distributions by asking: How *similar* are the spatial point patterns S_1 and S_2? For instance, in the Forest example of Figure 5.1 we started off with the observation that the oaks appear to be much more clustered than the maples. Hence rather than characterizing this relative clustering as repulsion between the two populations, we might simply ask whether the pattern of oaks, S_1, is more clustered than the pattern of maples, S_2.

But while the original (univariate) sample K-functions, $\hat{K}_1(h)$ and $\hat{K}_2(h)$, provide natural measures of individual population clustering, it is not clear how to compare these two values statistically. Note that since the population values, $K_1(h)$ and $K_2(h)$, are simply mean values (for any given h), one might be tempted to conduct a standard difference-between-means test. But this could be very misleading, since such tests assume that the two underlying populations (in this case S_1 and S_2) are *independently* distributed. As we have seen above, this is generally false. Hence the key task here is to characterize "complete similarity" in a way that will allow deviations from this hypothesis to be tested statistically.

Here the basic strategy is to interpret "complete similarity" to mean that both point patterns are generated by the *same spatial point process*. Hence if the sizes of S_1 and S_2 are given respectively by n_1 and n_2, then our null hypothesis is simply that the

[18] For additional discussion of this point see Diggle (2003, p.6).
[19] Lotwick and Silverman noted this same phenomenon in their original paper (1982, p.410), where they concluded that such added structure will tend to "show less discrepancy from independence" and thus yield a relatively conservative testing procedure.

combination of these two patterns, $S = [(s_{1i} : i = 1,..,n_1),(s_{2j} : j = 1,..,n_2)]$, is in fact a *single population* realization of size $n = n_1 + n_2$, i.e., $S = (s_1,..,s_{n_1},s_{n_1+1},..,s_n)$. If this were true, then it would not matter which subset of n_1 samples was labeled as "population 1". It should be clear from the above discussion that a natural way to formulate this hypothesis is to treat the combined process as a *marked point process*.[20] In this framework, the relevant null hypothesis is simply that given observed locations, $(s_1,..,s_n)$ and labels $(m_1,..,m_n)$ with n_1 occurrences of "1" and n_2 occurrences of "2", each permutation of these labels is *equally likely*. But this is precisely the assertion in expression (5.6.8) above. Hence in the context of marked point processes, the *joint* distribution of labels $(m_1,..,m_n)$ given locations $(s_1,..,s_n)$ and population sizes, n_1 and n_2, is here seen to be precisely the *spatial indistinguishability hypothesis*.

However, the present focus is on the *marginal* distributions of populations 1 and 2 rather than the dependency properties of their joint distribution. Hence the natural test statistics are the sample K-functions, $\hat{K}_1(h)$ and $\hat{K}_2(h)$, for each marginal distribution rather than the sample cross K-function. Note moreover that if both samples are indeed coming from the same population, then $\hat{K}_1(h)$ and $\hat{K}_2(h)$ should be estimating the *same* K-function, say $K(h)$, for this common population. Hence if these sample K-functions were *unbiased* estimates, then by definition the individual K-functions, $K_i(h) = E[\hat{K}_i(h)]$, $i = 1,2$, would be the same. In this context, "complete similarity" would thus reduce to the simple null hypothesis: $H_0 : K_1(h) = K_2(h)$. However, as noted in section 4.3, this simplification is only appropriate for stationary isotropic processes with Ripley corrections. Thus, in view of the fact that hypothesis (5.6.2) is perfectly meaningful for any point process, we choose to adopt a more flexible approach.

To do so, we first note that even in the absence of stationarity, the sample K-functions, $\hat{K}_1(h)$ and $\hat{K}_2(h)$, continue to be reasonable measures of clustering (or dispersion) within populations. Hence to test for relative clustering (or dispersion) it is still natural to focus on the *difference* between these sample measures,[21] which we now define to be

$$\Delta(h) = \hat{K}_1(h) - \hat{K}_2(h) \qquad (5.7.1)$$

Hence the relevant *spatial similarity hypothesis* for our present purposes is that the observed difference obtained from (5.7.1) is not statistically distinguishable from the random differences obtained from realizations of the conditional distribution of labels under the spatial indistinguishability hypothesis [(5.6.2),(5.6.3)].

[20] Indeed this is the reason why the analysis of joint distributions above was developed *before* considering the present comparison of marginal distributions.

[21] Note that one could equally well consider the *ratio* of these measures, or equivalently, the difference.of their logs.

5.7.1 Spatial Similarity Test

If we simulate random relabelings in (5.6.8) to obtain a sampling distribution of $\Delta(h)$ under this spatial similarity hypothesis, then the observed difference can simply be compared with this distribution. In particular, if the observed difference is unusually large (small) relative to this distribution, then it can reasonably be inferred that subpopulation 1 is significantly more clustered (dispersed) than subpopulation 2. This procedure can now be formalized by the following simple variation of the random relabeling test, which we designate as the *spatial similarity test*:

(i) Given *observed* locations, $(s_1,..,s_n)$, and labels $(m_1,..,m_n)$ with corresponding population sizes, n_1 and n_2, simulate N random permutations $[\pi_1(\tau),..,\pi_n(\tau)]$, $\tau = 1,..,N$ of $(1,..,n)$, and construct the corresponding the label permutations $(m_{\pi_1(\tau)},..,m_{\pi_n(\tau)})$, $\tau = 1,..,N$

(ii) If $S_1^\tau = (s_{1i}^\tau : i = 1,..,n_1)$ and $S_2^\tau = (s_{2j}^\tau : j = 1,..,n_2)$ denote the population patterns obtained from the joint realization, $[(s_1,..,s_n),(m_{\pi_1(\tau)},..,m_{\pi_n(\tau)})]$, $\tau = 1,..,N$, and if the corresponding sample difference function is denoted by $\Delta^\tau(h) = \hat{K}_1^\tau(h) - \hat{K}_2^\tau(h)$, then for the given set of relevant radial distances, $D = \{h_w : w = 1,..,W\}$, calculate the sample difference values, $\{\Delta^\tau(h_w) : w = 1,..,W\}$ for each $\tau = 1,..,N$.

(iii) Finally, if the *observed* sample difference function, $\Delta^0(h) = \hat{K}_1^0(h) - \hat{K}_2^0(h)$, is constructed from the observed patterns, S_1^0 and S_2^0, then under the *spatial similarity hypothesis*, each observed value, $\Delta^0(h_w)$, should be a "typical" sample from the list of values $[\Delta^\tau(h_w) : \tau = 0,1,..,N]$. Hence if we now let m_+^0 denote the number of simulated random relabelings, $\tau = 1,..,N$, with $\Delta^\tau(h_w) \geq \Delta^0(h_w)$, then the probability of obtaining a value *as large as* $\Delta^0(h_w)$ under this hypothesis is estimated by the following *relative clustering p-value* for population 1 versus population 2:

(5.7.2) $$\hat{P}_{r\text{-}clustered}^{12}(h) = \frac{m_+^0 + 1}{N + 1}$$

(iv) Similarly, if m_-^0 denotes the number of simulated random relabelings, $\tau = 1,..,N$, with $\Delta^\tau(h_w) \leq \Delta^0(h_w)$, then the probability of obtaining a value *as small as* $\Delta^0(h_w)$ under this hypothesis is estimated by the following *relative dispersion p-value* for population 1 versus population 2:

(5.7.3) $$\hat{P}_{r\text{-}dispersed}^{12}(h) = \frac{m_-^0 + 1}{N + 1}$$

5.7.2 Application to the Forest Example

This spatial similarity test is implemented in the MATLAB program, **k2_diff_plot.m**. Here it is convenient to adopt the marked-point-process format by defining a single list of locations, **loc**, in which the first **n1** locations correspond to population 1 and all remaining locations correspond to population 2. Hence both of these populations are identified by simply specifying **n1**. If **D** again denotes the set of selected radial distances used for the Forest example above, then a screen plot of *relative clustering p-values* for **999** simulations is now obtained by the command:

>> k2_diff_plot(loc,n1,sims,D,1);

The output of a typical run is shown in Figure 5.13 below:

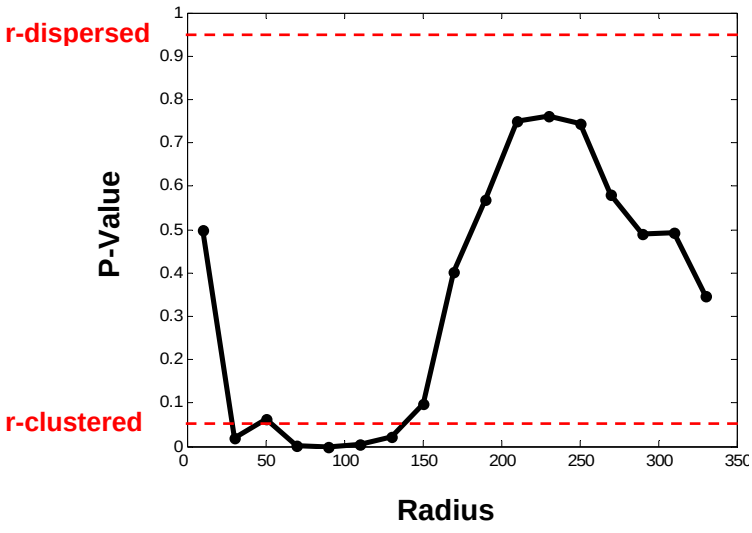

Figure 5.13. Relative Clustering of Oaks

This confirms the informal observation above that oaks are indeed more clustered than maples, for scales consistent with a visual inspection of Figure 5.1.

5.8 Larynx and Lung Cancer Example

While the simple Forest example above was convenient for developing a wide range of techniques for analyzing bivariate point populations, the comparison of Larynx and Lung cancer cases in Lancashire discussed in Section 1 is a much richer example. Hence we now explore this example in some detail. First we analyze the overall relation between these two patterns, using a variation of the spatial similarity analysis above. Next we restrict this analysis to the area most relevant for the Incinerator in Figure 1.9. Finally, we attempt to isolate the cluster near this Incinerator by a new method of local K-function analysis that provides a set of *exact* local clustering p-values.

5.8.1 Overall Comparison of the Larynx and Lung Cancer Populations

Given the Larynx Cancer population of $n_1 = 57$ cases, and Lung Cancer population of $n_2 = 917$ cases, we could in principle use **k2_diff_plot** to compare these populations. But the great difference in size between these populations makes this somewhat impractical. Moreover, it is clear that the Larynx cancer population in Figure 1.7 above is of primary interest in the present example, and that Lung cancers serve mainly as an appropriate reference population for testing purposes. Hence we now develop an alternative testing procedure that is designed precisely for this type of analysis.

Subsample Similarity Hypothesis

To do so, we again start with the hypothesis that Larynx and Lung cancer cases are samples from the same statistical population. But rather than directly compare the small Larynx population with the much larger Lung population, we simply observe that if the Larynx cases could equally well be *any* subsample of size n_1 from the larger joint population, $n = n_1 + n_2$, then the observed sample K-function, $\hat{K}_1(h)$, should be typical of the sample K-functions obtained in this way. Hence, in the context of marked point processes, the present *subsample similarity hypothesis* asserts that for any given realization $[(s_1,..,s_n),(m_1,..,m_n)]$, the value $\hat{K}_1(h)$ obtained from the n_1 locations with $m_i = 1$ is statistically indistinguishable from the same sample K-function obtained by randomly permuting these labels.

Test of the Subsample Similarity Hypothesis

The corresponding test of this subsample similarity hypothesis can be formalized as follows variation of the spatial similarity test procedure above:

(i) Same as for the spatial similarity test.

(ii) If $S_1^\tau = (s_{1i}^\tau : i = 1,..,n_1)$ denotes the population pattern obtained from the joint realization, $[(s_1,..,s_n),(m_{\pi_1(\tau)},..,m_{\pi_n(\tau)})]$, and if the corresponding sample K-function is $\hat{K}_1^\tau(h)$, then for the given set of relevant radial distances, $D = \{h_w : w = 1,..,W\}$, calculate the sample K-function values, $\{\hat{K}_1^\tau(h_w) : w = 1,..,W\}$ for each $\tau = 1,..,N$.

(iii) Finally, if the *observed* sample K-function, $\hat{K}_1^0(h)$, is constructed from the observed patterns, S_1^0 and S_2^0, then under the *subsample similarity hypothesis*, each observed value, $\hat{K}_1^0(h_w)$, should be a "typical" sample from the list of values $[\hat{K}^\tau(h_w) : \tau = 0,1,..,N]$. Hence if we now let m_+^0 denote the number of simulated random relabelings, $\tau = 1,..,N$, with $\hat{K}_1^\tau(h_w) \geq \hat{K}_1^0(h_w)$, then the probability of

obtaining a value *as large as* $\hat{K}_1^0(h_w)$ under this hypothesis is estimated by the following *clustering p-value* for population 1:

(5.8.1) $$\hat{P}^1_{clustered}(h) = \frac{m_+^0 + 1}{N+1}$$

(iv) In a similar manner, if m_-^0 denotes the number of simulated random relabelings, $\tau = 1,..,N$, with $\hat{K}_1^\tau(h_w) \le \hat{K}_1^0(h_w)$, then the probability of obtaining a value *as small as* $\hat{K}_1^0(h_w)$ under this hypothesis is estimated by the following *dispersion p-value* for population 1:

(5.8.2) $$\hat{P}^1_{dispersed}(h) = \frac{m_-^0 + 1}{N+1}$$

Hence under this testing procedure, significant clustering (dispersion) for population 1 means that the observed pattern of size n_1 is more clustered (dispersed) than would be expected if it were a typical subsample from the larger pattern of size n. Note that while this test is in principle possible for subpopulations of any size less than n, it only makes statistical sense when n_1 is sufficiently small relative to n to allow a meaningful sample of alternative subpopulations. Moreover, when n_1 is much smaller than n, the present Monte Carlo test is considerably more efficient in terms of computing time then the full spatial similarity test above

Application to Larynx and Lung Cancers

This testing procedure is implemented in the MATLAB program, **k2_global_plot.m**. (Here "global" refers to the global nature of this pattern analysis. We consider a local version later.) Before carrying out the analysis, it is instructive to construct a sample subpopulation pattern, S_1^τ, for visual comparison with the observed pattern, S_1^0, of Larynx cancers. The MATLAB workspace, **Larynx.mat**, contains the full set of $n = 57 + 917 = 974$ locations in the matrix, **loc**, where the $n_1 = 57$ Layrnx cancer cases are at the top. A random subpopulation of size n_1 can be constructed in MATLAB by the following command sequence:

>> list = randperm(974);

>> sublist = list(1:57);

>> sub_loc = loc(sublist,:);

The first command produces a random permutation, **list**, of the indices **(1,...,974)** and the second command selects the first 57 values of **list** and calls them **sublist**. Finally, the last

command creates a matrix, **sub_loc**, of the corresponding locations in **loc**. While this procedure is a basic component of the program, **k2_global_plot.m**, it is useful to perform these commands manually in order to see an explicit example. This coordinate data can then be imported to ARCMAP and compared visually with the given Larynx pattern as shown in Figures 5.14 and 5.15 below:[22]

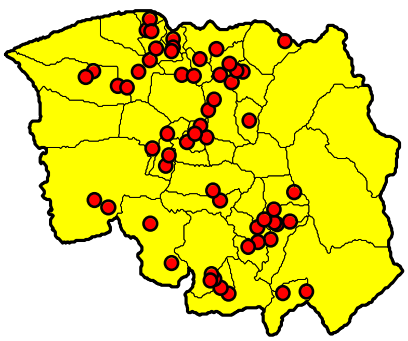

Fig.5.14. Observed Larynx Cases

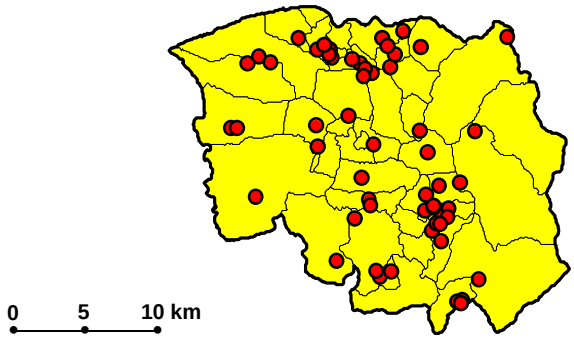

Fig.5.15. Sampled Larynx Cases

This visual comparison suggests that there may not be much difference between the overall pattern of observed Larynx cancers and typical subsamples of the same size from the combined population of Larynx and Lung cancers.

To confirm this by a statistical test, it remains only to construct an appropriate set of radial distances, **D**, for testing purposes. Here it is instructive to carry out this procedure explicitly by using the following command sequence:

>> Dist = dist_vec(loc);

>> Dmax = max(Dist);

>> d = Dmax/2;

>> D = [d/20:d/20:d];

The first command uses the program, **dist_vec**, to calculate the vector of $n(n-1)/2$ distinct pairwise distances among the n locations. The second command identifies the maximum, **Dmax**, of all these distances, and the third command used the "**Dmax/2**" rule of thumb in expression (4.5.1) above to calculate the maximum radial distance for the test. Finally, some experimentation with the test results suggests that the p-value plot should include 20 equally spaced distance values up to **Dmax/2**. This can be obtained by

[22] Note also that these subpopulations can be constructed directly in MATLAB. The relevant boundary file is stored in the matrix, **larynx_bd**, so that subpopulation, **sub_loc**, can be displayed with the command, **poly_plot(larynx_bd,sub_loc).** See Section 9 of the Appendix to Part I for further details.

the last command, which constructs a list of numbers starting at the value, **d/20**, and proceeding in increments of size **d/20** until the number **d** is reached.

Given this set of distances, **D**, a statistical test of the *subsample similarity hypothesis* for this example can be carried out with the command:

>> k2_global_plot(loc,n1,999,D,1);

A typical result is shown in Figure 5.16 below:

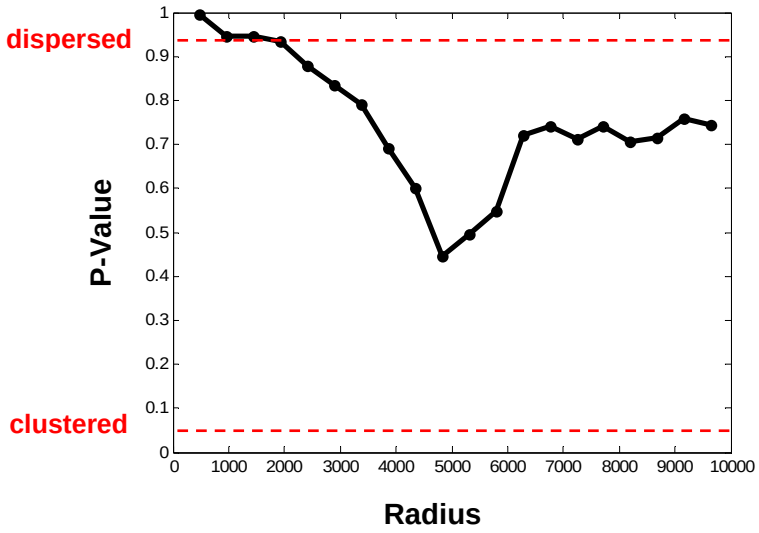

Figure 5.16. P-Values for Larynx Cases

Here we can see that, except at small distances, there is no significant difference between the observed pattern of Larynx cases and random samples of the same size from the combined population. Moreover, since the default p-values calculated in this program are the *clustering p-values* in (5.8.1), the portion of the plot above .95 shows that Larynx cases are actually significantly *more dispersed* at small distances than would be expected from random subsamples. An examination of Figures 1.7 and 1.8 suggests that, unlike Lung cancer cases which (as we have seen in Section 4.7.3) are distributed in a manner roughly proportional to population, there appear to be somewhat more Larynx cases in less populated outlying areas than would be expected for Lung cancers. This is particularly true in the southern area, which contains the Incinerator. Hence we now focus on this area more closely.

5.8.2 Local Comparison in the Vicinity of the Incinerator

To focus in on the area closer to the Incinerator itself, we start with the observation that heavier exhaust particles are more likely to affect the larynx (which is high in the throat). Hence while little is actually known about either the exact composition of exhaust fumes from this Incinerator or the exact coverage of the exhaust plume, it seems reasonable to

suppose that heavier exhaust particles are mostly concentrated within a few kilometers of the source. Hence for purposes of the present analysis, a maximum range of 4000 meters (≈ 2.5 miles) was chosen.[23] This region is shown in Figure 5.17 below as a circle of radius 4000 meters about the Incinerator (which is again denoted by a red cross as in Figure 1.9):

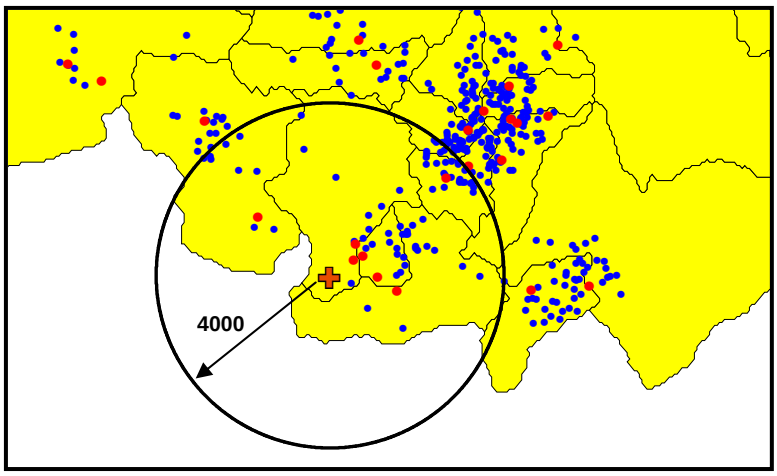

Figure 5.17. Vacinity of the Incinerator

If the coordinate position of the Incinerator is denoted by **Incin**,[24] then one can identify those cases that are within 4000 meters of **Incin** by means of the customized MATLAB program, **Radius_4000.m**. Open the workspace, **layrnx.mat**, and use the command:

>> OUT = Radius_4000(Incin,Lung,Larynx);

Here **Lung** and **Larynx** denote the locations of the Lung and Larynx cases, respectively. The output structure, **OUT**, includes the locations of Lung and Larynx cases within 4000 meters of **Incin**, along with their respective distances from **Incin**. Here it can be seen by inspection that the number of Larynx cases is **n1 = 7**. The total number of cases in this area is **n = 75**. The appropriate inputs for **k2_global_plot** above can be obtained from **OUT** as follows:

>> loc_4000 = OUT.LOC;

>> n1_4000 = length(OUT.L1);

[23] This is in rough agreement with the distance influence function, $f(d)$, estimated by Diggle, Gatrell and Lovett (1978, Figure 7), which is essentially flat for $d \geq 4$ kilometers.

[24] This position is given in the ARCMAP layer, **incin_loc.shp**, as **Incin = (354850,413550)** in meters.

Hence choosing **D_4000 = [400:200:4000]** to be an appropriate set of radial distances, a test of the *subsample similarity hypothesis* for this subpopulation can be run for **999** simulations with the command:

>> k2_global_plot(loc_4000,n1_4000,999,D_4000,1);

Here a typical result is shown in Figure 5.18 below:

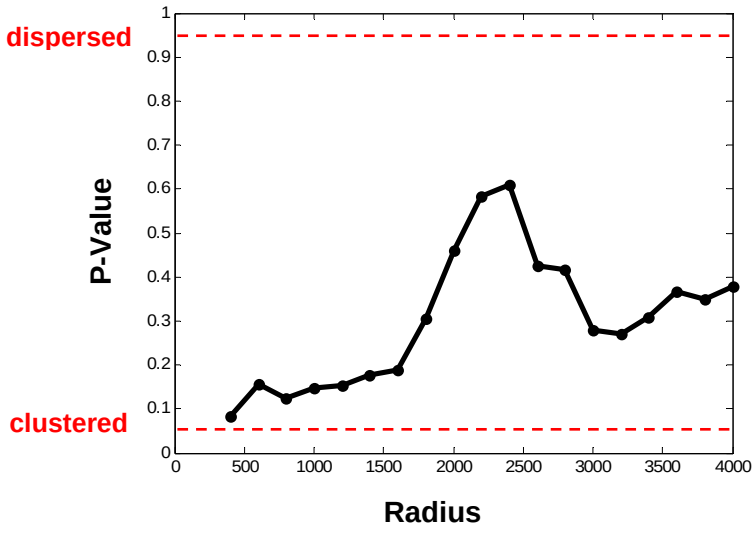

Figure 5.18. P-Values for Incinerator Vicinity

This plot is seen to be quite different from the global plot of Figure 5.16 above. In particular, there is now some weakly significant *clustering* at scales below 500 meters. This suggests that while the global pattern of Larynx cases exhibits no significant clustering relative to the combined population of Larynx and Lung cases, the picture is quite different when cases are restricted to the vicinity of the Incinerator. In particular, the strong cluster of three Larynx cases nearest to the Incinerator in Figure 5.17 would appear to be a contributing factor here.

5.8.3 Local Cluster Analysis of Larynx Cases

This leads to the third and final phase of our analysis of this problem. Here we consider a local analysis of clustering which is a variation of the local K-function analysis in Section 4.8 above. We again adopt the spatial indistinguishability hypothesis that Larynx and Lung cases are coming from the same point process, but now focus on each individual Larynx case by considering the conditional distribution of all other labels given this Larynx case.

To motivate this approach, we start by considering an enlargement of Figure 5.17 in Figure 5.19 below that focuses on the cluster of three Larynx cases closest to the Incinerator. Here we choose upper most case, labeled s_{1i} in the figure, and consider a

circular region of radius $h = 400$ meters about this case. There are seen to be six other cases within distance h of s_{1i}, of which two are also Larynx cases. Hence it is of interest to ask how likely it is to find at least two other Larynx cases within this small set of cases near s_{1i}.

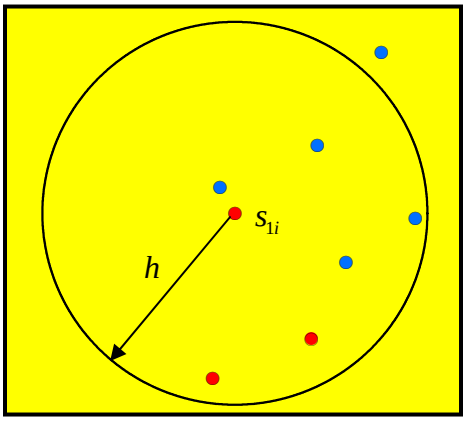

Figure 5.19. Neighborhood of Larynx Case

To determine the probability of this event, we start by removing the 4000-meter restriction and return to the full population of cancer cases, $n = n_1 + n_2 = 974$ with $n_1 = 57$. If we again adopt the null hypothesis of *subsample similarity* (so that Larynx cases could equally well be any subsample of size n_1 from the full population of n cases), then under this hypothesis one can calculate the *exact* probability of this event. To start with, if there are c other cases within distance h of case, s_{1i}, and c_1 of these belong to population 1, then under the subsample similarity hypothesis, this event can be regarded as a random sample of size c from the population of $n-1$ other cases which contains exactly c_1 of the $n_1 - 1$ other population 1 cases. Hence the probability of this event is given by the general *hypergeometric* probability:

$$(5.8.3) \qquad p(k \mid m, K, M) = \frac{\binom{K}{k}\binom{M-K}{m-k}}{\binom{M}{m}} = \frac{\left(\frac{K!}{k!(K-k)!}\right)\left(\frac{(M-K)!}{(m-k)!(M-K-m+k)!}\right)}{\left(\frac{M!}{m!(M-m)!}\right)}$$

where in the present case, $k = c_1, K = n_1 - 1, m = c,$ and $M = n - 1$. Finally, to construct the desired event probability as stated above, observe that if we let the random variable, C_1, denote the number of population 1 cases within distance h of s_{1i}, then the chance of observing *at least* c_1 cases from population 1 is given by the sum:

(5.8.4) $$P(c_1 \mid c, n_1, n) = \text{Prob}(C_1 \geq c_1 \mid c, n_1, n) = \sum_{k=c_1}^{c} p(k \mid c, n_1 - 1, n-1)$$

It is this cumulative probability, $P(c_1 \mid c, n_1, n)$, that yields the desired event probability. In the specific case above where $c_1 = 2, c = 6, n_1 = 57,$ and $n = 974$, we see that this probability is given by

(5.8.5) $$P(2 \mid 6, 57, 974) = .042$$

Hence if the subsample similarity hypothesis were true, then it would be quite surprising to find at least two Larynx cases with this subpopulation of six cases. In other words, for the given pattern of Larynx and Lung cases, there appears to be *significant clustering* of Larynx cases near s_{1i} at the $h = 400$ meter scale.

Thus to construct a general testing procedure for local clustering (or dispersion) of Larynx cases, it suffices to calculate the event probabilities in (5.8.4) for every observed Larynx location, s_{1i}, at every relevant radial distance, h. This procedure is implemented in the MATLAB program, **k2_local_exact.m**.[25] In the present case, if we consider only the single radial distance, **D = 400**, and again use the location matrix, **loc**, then the set of clustering p-values at each of the **n1 = 57** Larynx locations is obtained with the command:

>> [P,C,C1] = k2_local_exact(loc,n1,400);

Here **P** is the vector of p-values at each location, and **C** and **C1** are the corresponding vectors of total counts, c, and population 1 counts, c_1, at each location.

To gain further perspective on the significance of the cluster in Figure 5.19 above, one can compare distances of cases to the Incinerator with the corresponding p-values as follows:

>> L = [Incin;Larynx];

>> dist_L = dist_vec(L);

>> dist = dist_L(1:57);

>> COMP = [P,dist];

>> COMP = sortrows(COMP,1);

>> COMP(1:7,:)

P	dist
0.0094077	**693.80**
0.029091	**910.34**
0.042038	**1002.90**
0.29995	**12512.00**
0.34049	**14858.00**
0.41478	**13744.00**
0.48083	**14982.00**

[25] In the MATLAB directory for the class, there is also a Monte Carlo version of this program, **k2_local.m.** By running these two programs for the same data set (say with 999 simulations) you can see that exact calculations tend to be orders of magnitude faster than simulations – when they are possible.

The first command stacks the Incinerator location on top of the Larynx locations in a matrix, **L**. The second and third commands then identify the relevant distances (i.e., from **Incin** to all locations in **Larynx**) as the first 57 distances, **dist**, produced by **dist_vec(L)**. The fourth and fifth commands combine **P** with **dist** in the matrix, **COMP**, and then sort rows of **COMP** by **P** from low to high. Finally the last command displays the first seven rows of this sorted version of **COMP**, as shown in the box on the right.

The first three rows (in red) are the three *closest* Larynx cases to the Incinerator, as can be verified in ARCMAP (and can also be seen in Figure 5.17 above).[26] Moreover, the ordering of p-values shows that these are the *only three locations that exhibit significant clustering*. Hence this result suggests that there may indeed be some relation between the Incinerator and nearby Larynx cases.

[26] Note that the case just below these three is almost as close to the Incinerator as one of these three. But this case has only a single Lung case within 400 meters, and hence exhibits *no* clustering at this scale.