Notes for Statistical Analysis of Spatial Point Patterns

October 23, 2015

1. Introduction

1.1 Spatial Point Patterns

Definition: Data in the form of a set of points, irregularly distributed within a region of space are called *spatial point patterns*. Point-locations are called *events* to distinguish them from arbitrary points of the region in question.

Definition: A *stochastic model* assumes that events are generated by some ulderlying random mechanism.

1.2 Sampling

The selection of the study region, A say, mertits some discussion. In some applications, A is objectively determined by the problem in hand, and inferences are required in terms of a process deefined on A itself. More commonly, A is selected from some much larger region. The selection of A ma then be made according to a probability sampling scheme, or it may simply reflect the experimenter's view that A is in some sense representative of the larger region. Either way, but particularly the latter, inferences drawn from an analysis will carry much greater conviction if consistency over replicate data-sets can be demonstrated.

As an alternative to intensive mapping within a single region A, the experimenter may choose to record limited information from a large number of smaller regions, for example the number of event in each region. In this context, the regions are called *quadrats* and the data are referred to as *quadrat counts*.

Quadrat sampling remains poular in plant ecology, but in some contexts it is rather impractical and this has led to the development, in ittialy in the American forestry literature, of a number of *distance methods* for sampling spatial pooints patterns. In these, the basic sampling unit is a poiint, and iformation is recorded in the form of distances to neighbouring events, for example the distances to the first few nearest events. This seems really relevant to what I'm studying and I have downloaded the relevant paper.

We refer to quadrat count and sistance methods as *sparse sampling* methods, to distinguis them from intensive mapping exercises. Apparently, the appropriate techniques for the analysis of data obtained by spare sampling and by intensive mappipng are quite different. Also, analyses of sparsely sampled patterns typically have more limited objectives than do analyses of mapped patterns.

This next paragraph deals with *Markov random fields*, so it doesn't seem super relevant.

Replicated sampling of mapped patterns has beens urprisingly rare until quite recently. Ecological investigations have compared patterns in study regions deliberately selected to represent different environmental conditions, but their are not corresponding studies whic have been designed with a view to establish the consistency of patterns in ostensibly similar regions. Some other, less relevant information follows.

1.3 Edge Effects

Edge effects arise in spatial point pattern analysis when, as is often the case in practice, the region A on which the pattern is observed is part of a larger region on which the larger process operates. The essential difficulty is that unobserved events outside A may interact with observed events within A but, precisely because the events in question are not observed, it is difficult to take proper account of this.

For some kinds of exploratory analysis, edge effects can safely be ignored. We shall discuss when any why this is so at appropriate points. More generally, we can distinguish between three broad approaches to handling edge effects: the use of buffer zones; explicit adjustments to take account of unobserved events; and, when A is rectangular, wrapping A onto a torus by identifying opposite edges.

The books goes on to describe these three methods of edge correction. I have written about two of them in my notebook.

1.4 Complete Spatial Randomness

The hypothesis of *complete spatial randomness*(CSR) for a spatial point pattern asserts that (i) the number of events in any planer region A with area $|A|$ follows a Poisson distribution with mean $\lambda|A|$; (ii) given n events x_i in a region A, the x_i are an independent random sample from the uniform distribution A.The constant λ is referred to as the *intensity*, or the mean number of events per unit area. According to (i), CSR implies that the instensity of events does not vary overy the plane. According to (ii), CSR also implies that there are no interactions amongst the events.

The interest in CSR is that it represents an idealized standard which, if strictly unattainable in practice, may nevertheless be tenable as a convenient first approximation. Most analyses begin with a test of CSR, and there are several good reasons for this. Firstly, a pattern for which CSR is not rejected scarcely merits any further formal statistical analysis. Secondly, tests are used as a means of exploring a set of data, rather than because rejection of CSR is of intrinsic interest. Thirdly, CSR acts as a dividing hypothesis to distinguish between patterns which are broadly classified as 'regular' or 'aggregated.

1.5 Objectives of Statistical Analysis

In any particular application, the objectives of a statistical analysis should be determined by the objectivesin collecting the data i question. What we do after CSR tests will vary depending on context.

The Dirichlet Tesselation

Some information on the dirichlet/voronoi tessalation. I already know what it does, so I'm not taking any notes here.

Monte Carlo Tests

Even simple stochastic models for spatial point patterns lead to intractable distribution theory, and in order to test models against data we shall make extensive use of Monte Carlo tests.

Generally, let u_1 be the observed value of a statistic U and let $u_i \ : \ i = 1, \ldots, s$, be corresponding values generated by independent random sampling from the distribution of U under a simple hypothesis $\mathcal{H}$. Let $u_{(J)}$ denote the jth largest amongst u_i. Then, under $\mathcal{H}$,

$$P\{u_1 = u_{(j)}\} = s^{-1} \ : \ j = 1, \ldots, s,$$

and rejection of $\mathcal{H}$ on the basis that u_1 ranks kth largest or higher gives an exact, on sided test of size k/s. This assumes that the values of the u_i are all different, so that the ranking of u_1 is unambiguous. If U is a discrete random variable, for example a count, tied values are possible and we then adopt the conservative rule of choosing the least extreme rank for u_1. The extension to two-sided tests is clear.

Hope shows that the loss of power resulting from a Monte Carlo implementation is slight, so that s need not be very large. For a one-sided test at the conventional 5% level, $s = 100$ is adequate.

Power loss occurs and can be seen on page 9. The important thing to remember is that the rank is important, not the actual statistic. Other important notes. We need to pick important statistics. We can only perform it on sufficiently simple hypotheses.

The principal advantage to be set against the above is that the investogator need not be contrained by known distribution theory, but rather can and should use informative statistics of their own choosing.

When asympotitic distribution theory is available, Monte Carlo testing provides an exact alternative for smalls amples and a useful check on the applicability of the asymptotic theoy. If the results of classical and Monte Carlo tests are in substantial agreement, little or nothing has been lost; if not, the explanation is usually that the classical test uses inappropriate distributional assumptions.

2. Preliminary Testing

2.1 Tests of complete spatial randomness

Although CSR is of limited interest in itself, there are several good reasons why we might ben an analysis with a test of CSR: rejection of CSR is a minimal prerequisite to any serious attempot to model an observed pattern; tests are used to explore a set of daata and to assist in the formulation of plausible alternatives to CSR; CSR operates as a dividing hypothesis between regular and aggregated patterns.

In view of the avove, the present discussion emphasized two aspects: the value of graphical methods, which will almost always be informative and will sometimes make formal testing unnecessary; and informal combination of serveral complementary tests, to indicate the nature of any departure from CSR. With regard to the second of these, if a single asessment of significance is required the following result is useful. Suppose that the attained significance levels of k not necessarily independent tests of CSR are $p_j : j = 1, \ldots, k$ and let $p_{\min}$ be the smallest such p_j, corresponding to the most significant departure from CSR. Then, under CSR,

$$p \leq P\{p_{\min} \leq p\} \leq kp.$$

For k independent tests, the exact result is

$$P\{p_{\min} \leq p\} = 1 - (1 - p)^k$$

. I worked out why this was the case in my notebook. It's not to hard to figure out. Using multiple tests as part of a diagnostic procedure makes pactical sense only if the various tests examine different aspects of apttern, so that a significant result for one test does not prevent a sensible interpretation of the others.

There are a number of reasons testing for CSR is importance, despite it being unambituous. Most importantly, the need to take account of the inherent dependence amongst multiple measurements derived from a single point pattern.

2.2 Inter-event distances

One possible summary description of a pattern of n events in a region A is the empirical distribution of the $\frac{1}{2}n(n-1)$ inter-event distances, t_{ij}. The corresponding theoretical distribution of the distance T between two events independently and uniformly distributed in A depends on the size and shape of A, bu it expressible in closed form for the most common cases of square or circular A.These distributions are denoted $H(t)$. There are distribution function listed for the two types of page 13, they are too complicated to incorporate into this TeX document.

We know develop a test for CSR based specifically on inter-event distances; the general approach is applicable to other summary desciptions and will reappear in later sections.

Assume that for the particular region A in question, $H(t)$ is known. Calculate the empirical distribution function (EDF) of inter-event distances. This function, $\hat{H}_1(t)$, say, represents the observed proportion of inter-event distances t_{ij} which are at most t; thus,

$$\hat{H}_1(t) = \{\frac{1}{2}n(n-1)\}^{-1}\#(t_{ij} \leq t),$$

where # means 'the number of'. Now prepare a plot of $\hat{H}_1(t)$ as ordinate against $H(t)$ as abscissa. If the data are compatible with CSR, the plot should be roughly linear. To assess the significance or otherwise of departures form linearity, the conventional approach would be to find the sampling distribution of $\hat{H}_1(t)$ under CR, but this is complicated by the dependence between inter-event distances with a common endpoint, and we therefore proceed as follows. Calculate EDFs $\hat{H}_i(t) : i = 2, 3 \ldots, s$, from each of $s-1$ independent simulations of n vents indepednently and uniformly distributed on A, and define *upper* and *lower simulation envelopes*,

$$U(t) = \max\{\hat{H}_i(t)\},$$

$$L(t) = \min\{\hat{H}_i(t)\},$$

, where in each case, i runs from 2 to s. These simulation envelopes can also be plotted against $H(t)$, and havethe property that under CSR, and for each T,

$$P\{\hat{H}_1\}$$

3. Statistical Methods for sparsely sampled patterns

3.1 General Remarks

3.2 Quadrat Counts

The probability distribution of $N(B)$, the number of events in any region with area B, follows a Poisson distribution with mean λB, where λ is the intensity. More explicitly, $N(B)$ is distributed as follows:

$$p_n(B) = \exp(-\lambda B)\{(\lambda B)^n/n!\} \ : \ n = 0, 1, 2, \ldots \quad (3.1)$$

.

3.3 Distance Methods

I think the k-tree sampling is just a distance method. Most early work was concerned with the definition of various types od distance measurement and associated statistics to test CSR or to estimate intensity. Holgate (1965a) marked something of a departure in that he evaluated the power functions of several tests of CSR against theoretical alternatives, thus providing an objective basis for the choice of a method. Developments since 1965 teneded to continue in this vein, investingating the power of tests of CSR(Holgate, 1965b; Besag and Gleaves, 1973; Brown and Holgate, 1974; Diggle et al., 1976; Cox and Lewis, 1976; Diggle, 1977b; Hines and O'Hara Hines, 1979; Byth and Ripley, 1980) or the robustness estimators of intensity (Persson, 1971; Pollard, 1971; Holgate, 1972; Diggle, 1975, 1977a; Cox, 1976; Warren and Batcheler, 1979; Patil et al., 1979; Byth, 1982).

3.3.1 Distribution theory under CSR

When CSR holds, the distribution theory for the varioud distance methods can be derived from the Poissson distribtuion of quadrat counts together with the independence of counts in disjoint regions. From (3.1) , taking B to be the area πx^2 of a disc of radius x, we immediately deduce that the distribution function of the distance X from an arbitrary point (or event) to the nearest (other) event is

$$F(x) = 1 - \exp(-\pi\lambda x * 2) : \ x \geq 0,$$

a result previously given. Notice that πX^2 follows an exponential distribution with parameter λ and $2\pi\lambda X^2$ is therefore distributed as χ^2_2. This can be easily computed.

Various other distance distributions associated with the Poisson process can be derived from (3.1) and the independence of numbers of events in disjoint regions. Let $X_{k,\theta}$ denote the distance from an arbitrary point or event to the kth nearest event within a sector of included angle $\theta \leq 2\pi$ and arbitrary orientation. Let $U_k = \frac{1}{2}\theta X^2_{k,\theta}$ and note that U_k is the area of a sector of included angle θ and radius $X_{k,theta}$. In what follow u_k is an area, not a radius or anything like that. We are testing areas! Then

$$P(U_1 > u_1) = P\{N(u_1) = 0\} = e^{-\lambda u_1},$$

again using (3.1). Furthermore, for any $u_2 > u_1$,

$$P(U_2 > U_1 | U_1 = u_1) = P\{N(u_2 - u_1) = 0\} = \exp\{-\lambda(u_2 - u_1)\},$$

so that the conditional probability density function of U_2, gien $U_1 = u_1$, is $\lambda \exp(-\lambda(u_2 - u_1))$. and the joint pdf of $(U_1, U)2$ is

$$f_2(u_1, u_2) = \lambda^2 \exp(-\lambda u_2) : 0 < u_1 < u_2.$$

. Essentially the same argument gives the joint pdf of $(U_1, U_2, \ldots, U_k)$ for any k as

$$f_k(u_1, \ldots, u_k) = \lambda^k \exp(-\lambda u_k) : 0 < u_1 < \cdots < u_l, \quad (3.5)$$

.

I think the following section could be informative to, though different from, the current project. It deals with non-nicely covered areas. Cox and Lewis consider the joint distribution of random variables X and Y defined as the respective distances from an arbitrary point O to the nearest event, at P say, and from P to the nearest other event, Q. Then

$$P(Y > y|X = x) = P[N\{A(x, y)\} = 0],$$

where

$$A(x, y) = \pi y^2 - (\varphi y^2 + \theta x^2 - xysin\varphi) \;\; (3.6)$$

is the area of the shaded region in Figure 3.1, $\cos\varphi = y/(2x)$ and $\theta + 2\varphi = \pi$

Notice in particular that P is *not* an arbitrary event; the selection procedure for P is biased in favour of the more isolated events in the population. Some other findings follow, but I'm undersure whether the will help.

3.3.2 Tests of CSR

Remarks of the comprritive power of different tests represent an overview of results in Diffle et al. (1976), Hines and O'Hara Hines (1979) and Byth and Ripley (1980) .The original papers give more details.

The first test won't play particulary nicely with our data. It assumes we can pick random sample point and go to the nearest event and also pick events and go to the nearest event. But we already have the data. They authors pick up on this in the following paragraph.

The second test assumes we have first and second event distances, which we don't

The third test seems promising. Eberhardt(1967) considered only point-event distances x_i, and proposed an index

$$e = m\sum x_i^2/(\sum x_i)^2.$$

They also say the $\sqrt{m(e-1}/(m-1)$ is the samle coefficient of variation of the distances, but I can't get that to come out in my work. Hines and O'Hara Hines (1979) provide critical values to test CSR but the test must be weak against aggregated alternatives, because the distribution of e under CSR applies also to a process of randomly distributed point clusters, in which each single event of a completely random pattern is replaced by a fixed or random number of coincident events. By the same arguemtn, any scale-free statistics based only on measurements of the distance from a sample point to the nearest event must be suspect. *I see where they are getting at here. But it seems that this randomly distributed point cluster model could actually be really helpful in testing aggregation. I should look into this.*

The remainder of the tests are interesting, but sampling scheme specific. It seems like $k = 1$ tree sampling isn't super ideal.

0.0.1 3.3.3 Estimators of Intensity

Suppose now that distances are measured from each sample point to the nearest, second nearest, ... kth nearest event. Then (3.5) shows that under CSR the distances $x_{ki} : i = 1, 2, \ldots, m$ to kth nearest events are sufficient for λ, and the the maximum likelihood estimator of $\gamma = \lambda^{-1}$ is

$$\hat{\gamma}_k = \pi(\sum x_{ki}^2)/(km)$$

which is unbiased with variance $\gamma^2/(km)$. An increase in the value of k gives an estimator which has smaller variance, but whose application in the field is more time consuming. The more subtle question of robustness to departures from CSR will be considered shortly. The change from λ to γ as the parameter of interest makes for ease of presentation, but also seems natural for a distance-based method of estimation, since squared distances effectively measure areas. Holgate showed that if the total quadrat area in the quadrat count estimator. *I stopped taking notes here. the result is highly biased when CSR does not hold, as it often will not. Also, the variance is big when $k = 1$ as in our case.*

1 4. Spatial Point Processes