INSTITUUT TNO VOOR WISKUNDE, INFORMATIEVERWERKING EN STATISTIEK

INSTITUTE TNO FOR MATHEMATICS, INFORMATION PROCESSING AND STATISTICS

BINARY MOSAICS AND POINT QUADRAT SAMPLING IN ECOLOGY

by

Cajo J.F. ter Braak

A thesis submitted for the degree of Master of Science at the University of Newcastle upon Tyne

IWIS-TNO, Wageningen

September 1980

A 80 ST 96 37

CONTENTS

		Page No.
ACKNOWLEDGEMENTS		i
ABSTRACT		ii
GENERAL INTRODUCTION		1
PART I:	BINARY MOSAICS AND POINT QUADRAT SAMPLING IN ECOLOGY.	
CHAPTER 1	INTRODUCTION	6
CHAPTER 2	SAMPLING DESIGN AND EFFICIENCY	
1	INTRODUCTION	8
2	THEORY AND METHODS	
2.1	General Theory	
2.1.1	Estimating the average over a region or the mean of a process.	13
2.1.2	Estimating the average over a finite region.	15
2.1.3	Estimating the mean of a planar stochastic process.	17
2.2	Binary Mosaic Processes	
2.2.1	Definition.	18
2.2.2	L-mosaics.	18
2.2.3	S-mosaics.	22
2.2.4	C-mosaics.	22
2.3	Interpretation of Parameters used in the Comparison of Designs.	26
2.4	Numerical Procedures.	30
3	RESULTS	
3.1	Efficiency of Systematic Sampling.	31
3.2	Efficiency of Frame Sampling.	
3.2.1	Limiting variance for rectangular frames.	32
3.2.2	Frame size and variance per sample point.	32
3.2.3	Optimum frame size for a given cost function.	33
3.2.4	Choice of frame size: an example.	37
4	DISCUSSION	
4.1	Systematic Sampling.	42
4.2	Frame Sampling.	44
5	TABLES AND FIGURES.	

Page No.

CHAPTER 3	STATISTICAL ANALYSIS OF POINT QUADRAT DATA.	
1	INTRODUCTION	48
2	MODEL-FREE METHODS	
2.1	General Considerations.	49
2.2	Estimate of the Covariance Function.	51
2.3	Systematic Sampling.	52
2.4	Frame Sampling.	53
3	MODEL-BASED METHODS	
3.1	Multivariate Binary Distributions on a Lattice.	54
3.2	L-mosaics.	57
3.3	C-mosaics.	59
4	DISCUSSION	65
PART II:	ASSOCIATION BETWEEN PINE AND HEATHER IN A DEVELOPING STAND	
CHAPTER 1	INTRODUCTION	69
CHAPTER 2	FITTING A POINT PROCESS MODEL TO THE PINES	
1	INTRODUCTION	71
2	THEORY AND METHODS	
2.1	Summary Descriptions.	73
2.2	Spatial Point Process Models.	75
2.3	Discrepancy Measure, Estimation and Testing.	80
2.4	Confidence Regions.	82
2.5	Numerical Procedures.	84
3	RESULTS	
3.1	Summary Statistics.	85
3.2	Goodness-of-Fit of Poisson Process and Heather-Based Cox Process.	85
3.3	Gaussian Cluster Process.	
3.3.1	Estimation.	86
3.3.2	Goodness-of-fit.	87
3.3.3	Confidence regions.	88
4	DISCUSSION	88
5	TABLES AND FIGURES	

		Page No.
CHAPTER 3	TESTS FOR ASSOCIATION	
1	INTRODUCTION	89
2	THEORY AND METHODS	90
3	RESULTS	93
4	DISCUSSION	94
5	TABLES AND FIGURES	
CHAPTER 4	DISCUSSION	97
REFERENCES		98

ACKNOWLEDGEMENTS

I would like to thank Dr P.J. Diggle under whose supervision this project became a flourishing piece of work.

This research was carried out while on a year's leave from the Institute TNO for Mathematics, Information Processing and Statistics (IWIS-TNO) in Wageningen, The Netherlands. I am grateful in particular to Mr J.C.A. Zaat, who guided me through the planning stage of this year of study.

I am also pleased to thank Drs J.B. van Biezen and J. Oude Voshaar, who filled my place temporarily at the Research Institute for Nature Management (RIN).

Without the careful typing of Mrs J. Watson this manuscript would not have been ready in time.

ABSTRACT

Point quadrat sampling is a method for estimating the proportional cover p of a plant species over a planar region, whereby the presence or absence of the species is recorded at each of N sample points. Then, if n is the number of points at which the species is present, $\hat{p} = n/N$ is unbiased for p. Expressions for the variance of $\hat{p}$ are given for (1) a random sample of sample points independently and uniformly distributed over a region, (2) a systematic sample of points in a rectangular lattice covering the region, (3) a frame sample of small randomly located frames, each with m points regularly spaced along a line or in a square lattice.

The efficiency of systematic sampling and frame sampling with respect to random sampling is evaluated for two classes of underlying mosaic processes, termed L-mosaics and C-mosaics. For frame sampling the optimum number of points per frame is tabulated, using a realistic cost function.

L-mosaics and C-mosaics induce a Markov process and a renewal process on the line, respectively. These properties allow likelihood based inference for data from linear frames.

A map showing the incidence of heather (Calluna vulgaris) and the locations of pines (Pinus silvestris) on a 10mX20m area in Central Sweden is analysed to test for association between pine and heather. The pines regenerated from seed trees which remained after clear-cutting 21 years previously. Both parametric and nonparametric methods are used to demonstrate significant association. Parametric models show that the pines are clustered even after this association is taken into account. An explanation is proposed.

GENERAL INTRODUCTION

Describing vegetation is not an easy task. Brown (1954), Greig-Smith (1964) and Tothill (1978), among others, discuss various quantitative measures that have been proposed. The concepts involved suggest mathematical idealizations that can be used in models for the spatial distribution of the vegetation over the ground.

Plants sometimes occur as clearly distinct individuals, for example pine seedlings. A natural measure of abundance, then, is number per unit area (density or intensity). We may think of the species as a set of points in the plane, disregarding the third dimension of aerial parts or roots. The points may have a typical spatial distribution over the plane, for example, a clustered as against a random or regular distribution. A biological explanation for a spatial distribution of a plant may be that the plant reacts to a micro-pattern in the environment over the region: or that there is competition or, on the other end of the scale, mutual support between the plants; or that it is the dispersal of seeds or the vegetative propagation that causes the spatial distribution (Greig-Smith, 1979). Quantification of the spatial distribution ('pattern') may help to distinguish between different patterns and, more ambitiously, to relate the type of spatial distribution to possible causal factors. Quantification is most illuminated within a class of models in which each model typifies a pattern. Most patterns of vegetation are not at all regular in a deterministic sense, hence stochastic models appear to be more appropriate than deterministic ones.

A stochastic model consists of a number of rules from which the pattern can be generated. The dynamic character of the model is reflected in the alternative term stochastic process. For example, if plants occur as points, then a stochastic model may be that within any given region

the points are uniformly distributed over the region, i.e. the plants do not react to patterns of other plants or factors, nor do they interact among each other, and thus occur at completely random positions within the area. If, in addition, the number of points in the region is distributed according to a Poisson distribution with a fixed mean, then this model is termed a Poisson process. The points are said to constitute a Poisson process. This is one famous example of a spatial point process, a process that models the spatial distribution of points. The points that are generated are frequently called events to distinguish them from arbitrary points (locations) in a region (Cox and Lewis, 1966; Diggle, 1979). In part II a spatial point process is fitted to young pines occurring in a field partly covered with heather. Notice that though the model may have some dynamic aspect only one 'snapshot' of the pattern is analysed.

Cover is another measure of abundance in vegetation. It is the areal fraction of the ground occupied by the plant. With this definition the concept of a plant as a point no longer makes sense. Instead of being dimensionless, a plant is now a two-dimensional phenomenon, not too much an idealization for, for example, crustose lichens. For higher plants, either the vertical projection of the aerial parts are considered as cover, or the basal area, i.e. the intersection of the plant at ground level (Brown, 1954). A vegetation can then be thought of as a mosaic, a patchwork of occupied areas (patches) and unoccupied areas (gaps),(Pielou, 1964), as for example in a heathland only partially covered by heather (Calluna vulgaris). The shapes and sizes of the patches and gaps, together with their intermingling, are visual phenomena that ask for further explanation of their properties in terms of ecological factors. Again quantification may be a first step. Stochastic

models for mosaics are called mosaic processes. I shall consider only one species at a time, hence binary mosaics, in which each point of a region is classified into one out of two categories, for example, heather and not-heather. The categories are as well called phases of a mosaic (Pielou, 1977).

Two other important measures to describe vegetation are not mentioned further in this study. Frequency refers to the probability of finding a species in an area and is normally estimated by the fraction of quadrats of some fixed size and shape (sic!) in which the species occurs. Frequency is thus a non-absolute measure, i.e. it depends on the mode of sampling, on the size and shape of the quadrat used (Greig-Smith, 1964). Its main advantage is the ease with which presence/absence in a quadrat is recorded. On the other hand, the measurement of biomass (dry matter) is in most ecological applications time-consuming. The measurement is precise, but destructive for the vegetation.

The above measures are not equivalent. Each one highlights a different aspect of the vegetation that may be of ecological importance. For example, cover can be an important measure of the effect of grazing management, fertilizing, burning and growth; the 'cover' of bare ground (i.e. the absence of cover) may indicate danger of soil erosion or the amount of interception of rainfall by the vegetation (Tothill, 1978).

In part I the point quadrat method (Goodall, 1952 and references therein) to estimate cover is introduced. The method consists of recording at, say, N points whether the species is present or absent. The fraction of presences in the N recordings estimates the cover and is unbiased under conditions stated formally by Miles and Davy (1976) in the context of stereological formulae. The point quadrat method

consists of dimensionless samples, point samples, to infer about the cover in two dimensions, to which I restrict myself, or volume in three dimensions (see e.g. Hilliard and Cahn, 1961). The point samples or point quadrats can be randomly distributed over an area, or arranged into networks to sample the region systematically. Another method is to arrange the point samples into clusters, called frames, which can be moved around to sample sections of a far larger area. In Chapter 2 I evaluate the efficiency of the latter two designs relative to random sampling. Of course, the relative efficiency depends on the pattern of the vegetation and in this study the pattern of the vegetation is modelled by binary mosaic processes. The processes do model explicitly the patches and gaps in a vegetation that cause the 'contagion' or correlation between point samples that are not far apart. In earlier work on sampling design (Matérn, 1960) the starting point was the correlation between points at various distances apart, but without guarantee that the correlation function considered corresponded to a feasible and also reasonable model for the vegetation. The evaluation of relative efficiency for these ideal models results in a number of guidelines (Ch 2 § 4) for the choice of sampling design in practice, for real vegetation. The statistical analysis of point quadrat data is dealt with in chapter 3.

In part II the sampling design of the point samples was chosen by Nature: pine seedling established themselves after a clear-cutting on a heathland in Central Sweden. The young pines were recorded in a 10m × 20m area to be either on heather, i.e. in a patch of Calluna vulgaris or off heather, i.e. in a gap. The question at stake is whether there is statistical evidence of association between pine and heather. If there is no association between pine and heather, the pines

are simply a point quadrat sample of the region, although, as will be established, not a random sample. In fact, it is shown that there is statistical evidence of association and, moreover, that the pines are clustered, even if the patchiness of the heather and the association is taken into account.

PART II: ASSOCIATION BETWEEN PINE AND HEATHER IN A DEVELOPING STAND.

CHAPTER 1 INTRODUCTION

"Here are some data. I mapped the occurrence of pine and heather and I want to know whether there is association between pine and heather. I counted the number of pines on heather: the proportion on heather is far bigger than the cover of heather. Is this significant?". The map (Fig. 1) is inspected and our, of course, imaginary conversation may continue with: "Let's assume that the pines are random, then the number of pines on heather is binomial with ...", or "I can't help you, I'm afraid; you have counted the number of pines in only one plot. Statistics can help you, if you do similar counts in a number of independent plots. Let's see, where do these data come from ...". The former approach assumes independence of the pine-locations and this assumption must be realistic or be consistent with the available data, if valid conclusions are to be drawn. The latter approach states that the data must be inconclusive, that independence can be guaranteed by the design of the experiment or the observations and probably assumes that similar data can be collected easily. The aim of this study is to show that spatial analysis can result in insight in the data as presented, even beyond the question of significance of the association. The reader can judge whether the assumptions made are realistic.

Fig. 1 concerns the same 10mX20m area as Fig. 8 in §3.2.4 of Pt I Ch 2 and has been mapped in late July 1978 at Ivantjärnheden in Central Sweden. The heather was coded into a 100X200 array with each cell of size 10cm X 10cm. A cell is scored 'heather' (Calluna vulgaris) if the cover in the cell is more than 50%. The location of the pines (Pinus silvestris) was recorded to the nearest centimetre and classified to be on heather or off heather. There are 150 pines of which 36 are dead.

The median height of the living pines is 5.6m with interquartile range of 6.5m; their median age is 10 yr with interquartile range of 5 yr. The pines on and off heather are very similar, although there is a tendency that the age of pines off heather is less. In the sequel I confine attention to the locations of the pines.

Persson (1978) describes the site in some detail. The stand was regenerated from seed trees, which remained after clear-cutting in 1957. The soil surface had been treated with a tractor-scarifier, resulting in fairly regularly distributed patches of exposed soil (coarse sand with fractions of median sand of glaci-fluvial origin). Calluna vulgaris occupied about 50% of the area and was concentrated, in the main, in the scarified patches. Persson (1978) estimates the diameter of the scarified patches as 1m, in agreement with the diameter of the discs (0.8m) in the model used in §3.2.4 of Pt I Ch 2. The 'gaps' were covered by reindeer lichens, cup lichens, etc. Vaccinium vitis-idaea (Cowberry) was also abundant. In 1972/1973 parts of the stand were cleaned. It is unknown whether the area of Fig. 1 was cleaned, but I decided hereupon to include the dead pines, including stumps, in the analysis.

It stands out from the site description that the scarified patches are important for the occurrence of heather, and thus possibly for the occurrence of pines, and might eventually explain the observed association that I shall study. It is important to realize right from the start this limitation that is common to all observational studies. On the other hand, once the observed features are more clearly understood, experimental work may establish more effectively the causal relations. In Chapter 2 it is established that the pine-locations are neither completely random nor independent when the different numbers of pines on and off heather are taken into account. An alternative model allows for

the clustering of the pines and gives an acceptable fit. In Chapter 3 I show with both parametric and nonparametric methods that the data are not consistent with the hypothesis of no association. A combined model for pine and heather is proposed that accounts for their respective patterns and the association. Finally a methodological point. The map shown as Fig. 1 is split into two square plots, subsequently taken to be of unit side. The two plots, henceforth plot 1 (the left one) and plot 2 (the right one) are used for simple cross-validation. If in a test procedure parameters are to be specified a priori, these parameters are estimated from the other plot. This occurs in Monte-Carlo tests of goodness-of-fit. Statistics for the two plots are as follows. In plot 1 there are 82 pines of which 63 pines are on heather. The cover of heather is 0.4994. In plot 2 there are 68 pines of which 48 pines are on heather. The cover of heather is 0.4953.

CHAPTER 2 FITTING A POINT PROCESS MODEL TO THE PINES.

1. INTRODUCTION

Features of a spatial pattern may be modelled by stochastic processes, but their dynamic definition can contain only some major aspects of the assumed genesis of the pattern, or - even worse - may be irrelevant. In either case goodness-of-fit must necessarily be tested on the observed pattern.

A newly established pine has at least survived the following stages (van der Pijl,1972): maturation of the seed; dispersal, fixation to the soil, germination and establishment. The dispersal of pine seed is mainly by wind but, at least in some Pinus species, birds and squirrels are reported to act as planters by storing seeds. Wind dispersal is over distances up to 2 km (van der Pijl,1972). This suggests a 'Poisson-rain'

of seed in the small area we are concerned with. However, the seeds are likely to land in the lee and fixation is more likely in the scarified patches. Further, neither the germination conditions may be homogeneous nor the protection against severe climatic conditions or animals. The occurrence of heather may be important in nearly all stages. Moreover, pine and heather both have lateral root systems so that nutrient competition may occur in later stages of development. Surprisingly, the model that attempts to explain the clustered pattern of the pines from the patchiness of the heather mosaic alone does not fit. In the absence of precise knowledge of other factors that influence the final location, we must restrict ourselves to simple models that at least acknowledge the random components and describe the observed clustered pattern. A two parameter model is found that gives a reasonable fit, without use of the heather data.

Parameter estimation and goodness-of-fit testing for stochastic processes is hampered by the lack of manageable expressions for the likelihood function: hence must proceed along different lines. Diggle (1979) discusses methods that involve choice of a (functional) summary description of pattern and a measure of discrepancy between summary descriptions. Estimation of parameters then proceeds by minimizing the discrepancy between the summary description expected under the model and the observed one. Testing is by Monte Carlo methods, i.e. by simulating the model and calculating the discrepancy of the expected summary description with the summary description of the simulated pattern. Under the hypothesis that the data are consistent with the model the observed and simulated discrepancy values are exchangeable, hence the rank of the observed discrepancy provides the exact significance level of the test.

2. THEORY AND METHODS

2.1 Summary Descriptions

As there are no clear large scale inhomogeneities in the data nor any obvious directionality, attention is restricted to stationary and isotropic point processes, i.e. translations and rotations do not change the statistical properties of the process. Under these assumptions informative summary descriptions of the processes are (Ripley, 1977; Diggle, 1979)

λ = intensity, i.e. expected number of events per unit area;

$\lambda K(t)$ = expected number of further events within distance t of an arbitrary event;

$F(t)$ = probability that the distance from an arbitrary point to the nearest event is at most t;

$G(t)$ = probability that the distance from an arbitrary event to the nearest other event is at most t. (1)

The intensity describes the first moment of the process and is estimated by the observed number of events per unit area.

The K-function K(t) can be linked with the second moment structure of the process. Let N(A) denote the number of events in $A \subset \mathbb{R}^2$, and $d\underset{\sim}{x}$ an infinitesimal region centred at $\underset{\sim}{x}$. We assume that the process is orderly, i.e. $P\{N(d\underset{\sim}{x}) > 1\} = o(|d\underset{\sim}{x}|)$ where $|.|$ denotes area, so that $\lambda = \lim_{|d\underset{\sim}{x}| \to 0} P\{N(d\underset{\sim}{x}) > 0\}/|d\underset{\sim}{x}|$, and assume further that $E[N(d\underset{\sim}{x})N(d\underset{\sim}{y})] \sim P\{N(d\underset{\sim}{x}) > 0, N(d\underset{\sim}{y}) > 0\}$ as $|d\underset{\sim}{x}|, |d\underset{\sim}{y}| \to 0$. These assumptions are valid for the processes that will be considered and facilitate the interpretation of the second-moment function

$$g(t) = \lim_{|d\underset{\sim}{x}|, |d\underset{\sim}{y}| \to 0} \frac{E[N(d\underset{\sim}{x})N(d\underset{\sim}{y})]}{|d\underset{\sim}{x}|\,|d\underset{\sim}{y}|} \qquad (\|\underset{\sim}{x}-\underset{\sim}{y}\| = t) \tag{2}$$

as the joint probability density for the occurrence of a pair of events distance t apart (Ripley, 1977). Because now $g(t)/\lambda$ is the conditional

intensity of an event at $\underset{\sim}{x}$ given an event at $\underset{\sim}{o}$ ($\|\underset{\sim}{x}\| = t$) we have the following relation between K(t) and g(t)

$$\lambda K(t) = \int_0^t \int_0^{2\pi} \frac{g(r)}{\lambda} r \, d\vartheta dr = \frac{2\pi}{\lambda} \int_0^t g(r) r \, dr \tag{3}$$

or conversely (Ripley, 1977)

$$g(t) = \frac{\lambda^2}{2\pi t} \frac{dK(t)}{dt} \tag{4}$$

The advantage of K(t) over g(t) or the covariance density $c(t) = g(t) - \lambda^2$ (Cox and Lewis, 1966) is that a cumulative function like K(t) does not need smoothing when estimated from data. Given the definition of $\lambda K(t)$ it follows immediately that

$\lambda^2 K(t)$ = expected number of ordered pairs of events a distance at most t apart with the first event of each pair in a given region of unit area,

because

$$\sum_{n=0}^{\infty} n \, P\{N(A) = n\}[\lambda K(t)] = \lambda^2 |A| K(t) \, . \tag{5}$$

This result suggests how K(t) can be estimated from the empirical distribution of pairwise distances between events in the data. Given the sampling region, each ordered pair of events $(\underset{\sim}{x},\underset{\sim}{y})$ is given a weight $k(\underset{\sim}{x},\underset{\sim}{y})$ which is the reciprocal of the proportion of the perimeter of the circle centred on $\underset{\sim}{x}$ and passing through $\underset{\sim}{y}$, that is within the sampling region. Then the sum of these weights over all ordered pairs, less than t apart, divided by the area of the sampling region is an unbiased estimator for $\lambda^2 K(t)$, provided t is such that the above mentioned proportion of the perimeter cannot be zero for any point in the sampling region (Ripley, 1977); for the unit square: $t < \frac{1}{2}\sqrt{2}$. If λ is unknown its estimate is substituted, whence an estimator for K(t) that is slightly biased.

The F-function F(t) and G-function G(t) are nearest-neighbour distribution functions. Unbiased estimators for the F- and G-function are derived from the empirical sampling point-event distances and event-event distances with a correction for edge-effects. The correction ignores sample points or events which are closer than t to the boundary of the sampling region. If x_i and d_i are the respective distances from the i^{th} of m sample points to the nearest event and to the nearest point on the boundary of the sampling region then the unbiased estimator is (Ripley, 1977)

$$\hat{F}(t) = \#(x_i \leq t,\ d_i > t)/\#(d_i > t), \tag{6}$$

with the analogous formula for $\hat{G}(t)$ with x_i and d_i the distances of the i^{th} event to the nearest neighbour and to the boundary, respectively.

The F- and G-function restrict attention to one scale of spatial interaction, the scale determined by the nearest-neighbour distributions. The K-function allows inspection of different scales as in the nested block analysis of Greig-Smith (1964).

In practice the summary-functions are estimated for a finite number of equidistant values of t. The interval-width used for the pine data is 0.005 in terms of the unit square, i.e. 5 cm intervals in the field. F(t) is estimated using a square grid of 9×9 points.

2.2 Spatial Point Process Models

The homogeneous planar Poisson process, shortly Poisson process, is the simplest spatial point process. It has a single parameter, the intensity $\lambda(>0)$, and lacks any form of spatial dependence. In a Poisson process the number of events in a region with area A has a Poisson distribution with mean λA, while conditional on the number of events in a region, the events are distributed independently and identically, uniformly in the region. These characteristics enable simulation of a Poisson process.

It follows from these definitions that the expected number of further events in a circle centred at an arbitrary event is $\lambda\pi t^2$ and that the probability that no further events occur is $\exp(-\lambda\pi t^2)$. Hence the summary descriptions are

$$\begin{aligned} K(t) &= \pi t^2 \\ F(t) &= G(t) = 1 - \exp(-\pi\lambda t^2)\ . \end{aligned} \tag{7}$$

Instead of a constant intensity the point process may have a variable intensity over the plane. This leads to inhomogeneous planar Poisson processes, determined by the intensity function $\lambda(\underset{\sim}{x})$, with $\underset{\sim}{x}$ a location in the plane. In an inhomogeneous Poisson process the number of events in disjoint regions are independent, while the number of events in a region is Poisson distributed with a mean equal to the integral of the intensity function $\lambda(\underset{\sim}{x})$ over that region. Obviously the inhomogeneous Poisson process is not stationary. However, the process can be made stationary by assuming a stochastic model $\Lambda(\underset{\sim}{x})$ for the intensity function. The resulting point process is called a Cox process or doubly stochastic Poisson process (Matérn, 1971; Grandell, 1976). Given a realisation of $\Lambda(\underset{\sim}{x})$ a Cox process is an inhomogeneous Poisson process.

If $\Lambda(\underset{\sim}{x})$ is stationary and isotropic with mean λ, variance σ^2 and correlation function r(u), then the associated Cox process is stationary and isotropic with joint density

$$g(t) = E[\Lambda(\underset{\sim}{x})\Lambda(\underset{\sim}{y})] = \lambda^2 + \sigma^2\, r(u) \qquad (\|x-y\| = u) \tag{8}$$

whence with (3)

$$K(t) = \pi t^2 + \frac{2\pi\sigma^2}{\lambda^2}\int_0^t ur(u)du \tag{9}$$

Explicit expressions for the nearest neighbour distributions are difficult to obtain (Matérn, 1971).

The inhomogeneous Poisson process and the Cox process are suitable models to express the dependence of the intensity of pines on the occurrence of heather. Let the intensity in the patches and gaps be λ_1 and λ_0, respectively. Conditionally on the number of events N, only the relative intensity is important. Assuming $\lambda_0 < \lambda_1$, we can simulate the process by locating successive pines completely at random in the area while discarding pines that fall in a gap with probability $1 - \lambda_0/\lambda_1$. We proceed in this way until eventually N pines are produced. The ratio λ_0/λ_1 is estimated by the observed intensity ratio. If the mosaic of the heather is kept fixed we have an inhomogeneous Poisson process, but if in each simulation the mosaic of the heather is taken to be an independent realization of a mosaic process, we have a Cox process. Diggle (in prep.) fitted successfully a C-mosaic to the heather in which the patches of heather are the union of countably many closed discs with mutually independent and identically distributed radii and centres determined by a Poisson process (cf. Pt I Ch 2 §2.2.4). In the model fitted by Diggle (in prep.) the distribution of the radii is a three-parameter Weibull distribution,

$$H(r) = 1 - \exp\{-\rho(r-\delta)^{\kappa}\} \qquad r > \delta \ . \tag{10}$$

The resulting Cox process mirrors the inhomogeneity of the environment for the pines: patches of heather are more favourable for the pines than gaps. Because of the patchiness of the heather the pattern of pines will be clustered relative to the Poisson process ($r(u) \geq 0$ in (9)).

Contagion is another cause of clustered patterns. Flexible and tractable models are the Poisson cluster models (Neyman and Scott, 1958) that produce aggregated patterns by the following mechanism:

(1) Parent events constitute a Poisson process with intensity $1/\rho$.

(2) Each parent event produces a random number M of daughter events, independently and identically distributed for each parent.

(3) The position of each daughter relative to her parent is independently and identically distributed according to a bivariate distribution G(.).

I take the final process to consist of daughters only and assume that G(.) is radially symmetric and gives distribution H(t) of the distance between two arbitrary sisters. Further, let $\mu = EM$ and $\lambda = \mu/\rho$ then, with $\underset{\sim}{x}$ an arbitrary event,

$$\begin{aligned} \lambda K(t) &= E(\#\text{ of further events within t from } \underset{\sim}{x}) \\ &= E(\#\text{ of sisters within t from } \underset{\sim}{x}) \\ &\quad + E(\#\text{ of further events from other parents within t from } \underset{\sim}{x}). \end{aligned} \tag{11}$$

The first term on the right-hand side is

$$\sum_{n=0}^{\infty} P(\underset{\sim}{x} \text{ has } (n-1) \text{ sisters}) E(\#\text{ of sisters within t from } \underset{\sim}{x} \mid n)$$

$$= \sum_{n=0}^{\infty} \frac{nP(n)}{\mu} \left[(n-1)H(t)\right] = \frac{E\,M(M-1)}{\mu} H(t) \tag{12}$$

where P(n) is the probability that a parent has n daughters. The second term is simply $\lambda\pi t^2$ because the locations of different families are independent (Bartlett, 1975). Hence,

$$K(t) = \pi t^2 + \rho \frac{E\,M(M-1)}{\mu^2} H(t) \tag{13}$$

For the pines I specify M to be Poisson distributed and G(.) as the radially symmetric Gaussian distribution with density

$$g(x_1, x_2) = (2\pi\sigma^2)^{-1} \exp\{-(x_1^2 + x_2^2)/2\sigma^2\} \tag{14}$$

Hence, $E\,M(M-1) = \mu^2$, $H(t) = 1 - \exp(-t^2/4\sigma^2)$ and thus

$$K(t) = \pi t^2 + \rho\{1 - \exp(-t^2/4\sigma^2)\} \ . \tag{15}$$

This will be referred to as the Gaussian cluster process with parameters ρ and σ. Note that μ is a redundant parameter because $K(t)$ is scale-invariant, that $1/\rho$ and $1/\sigma$ are the number of clusters and their tightness, respectively and that, conditionally on the number of events N and the number of parents, the N daughters are allocated independently and randomly amongst the parents. Expressions for the nearest-neighbour distributions are available (Bartlett, 1975) but are not very enlightening.

The Gaussian cluster process is formally equivalent to the Cox process with random intensity function

$$\Lambda(\underset{\sim}{x}) = \mu \sum_{i=1}^{\infty} g(\underset{\sim}{x}-\underset{\sim}{X}_i) \qquad (16)$$

where the $\underset{\sim}{X}_i$ are the points of a Poisson process. Any attempt to distinguish from the observed pattern between contagion and heterogeneity is thus futile.

From super-position of independent point processes we can derive new processes. Let the subscript i(i = 1, 2) refer to the defining processes. The intensity of the resulting process, without a subscript, is $\lambda = \lambda_1 + \lambda_2$ and the K-function is

$$\lambda K(t) = \frac{\lambda_1}{\lambda}\{\lambda_1 K_1(t) + \lambda_2 \pi t^2\} + \frac{\lambda_2}{\lambda}\{\lambda_2 K_2(t) + \lambda_1 \pi t^2\} \qquad (17)$$

with $K_i^*(t) = K_i(t) - \pi t^2$ we get

$$K(t) = \pi t^2 + \{\lambda_1^2 K_1^*(t) + \lambda_2^2 K_2^*(t)\}/\lambda^2 \qquad (18)$$

so that for a Poisson cluster process (i=1) superimposed with a Poisson process (i=2)

$$K(t) = \pi t^2 + \rho \frac{\lambda_1^2}{\lambda^2} \frac{E\,M(M-1)}{\mu^2} H(t) \qquad (19)$$

i.e. the K-function of the superposed process is indistinguishable from the K-function of a Poisson cluster process with an identical distribution of M but a different number of parents $\lambda^2/(\lambda_1^2 \rho)$. For the pine-data

this property may well be advantageous in that a Poisson rain of seed from wind dispersal together with an independent dispersal agent that causes clustered pattern still gives a K-function within scope of the K-function of a Gaussian cluster process. The F- and G-functions do not have this property. For example, we have

$$1 - F(t) = P(\text{no event from either process within t from } \underset{\sim}{o})$$
$$= \{1 - F_1(t)\}\{1 - F_2(t)\} \qquad (20)$$

With a Gaussian cluster process and a Poisson process the parameter λ_2 of the Poisson process cannot be absorbed in the parameters ρ, μ or σ.

2.3 Discrepancy Measure, Estimation and Testing

Parameter estimation and goodness-of-fit testing will be based on the K-function. When in Chapter 3 the fitted model is used to test the association-hypothesis, the inter-event distances of the pattern determine the variance of the number of pines on heather and the K-function has been designed to summarize these distances. Fortunately, the K-function is more tractable than the F- and G-function.

The discrepancy between model and data is taken to be

$$d(\theta) = \int_0^{0.1} \{\hat{K}_\theta^{\frac{1}{2}}(t) - \hat{K}^{\frac{1}{2}}(t)\}^2 dt \qquad (21)$$

where θ is the parameter of the model. The square root transformation is chosen to stabilize the variance of $\hat{K}(t)$, at least under the Poisson model (Besag in discussion to Ripley, 1977; Silverman, 1978). The effect of a different choice of transformation and upper bound of integration will be discussed for the Gaussian cluster process. For the pine data the integration in (21) is replaced by a summation with interval width 0.005.

Parameter estimation proceeds by numerical minimization of (21) with the Simplex algorithm (Nelder and Mead, 1965). This method fails if (21) is insensitive to changes in θ as for example in the model where a Gaussian cluster process is superposed with an independent Poisson process; the parameters λ_2 and ρ in §2.2 are confounded in the K-function, but not in the F- and G-function.

Assessment of the goodness-of-fit of a model with a prescribed value of θ proceeds by a Monte Carlo test. A simulation of the model results in a pattern of events from which the K-function can be estimated. Then the discrepancy (21) is calculated. The discrepancy calculated for the data and the values of discrepancy for m-1 simulations give m values that are exchangeable under the hypothesis. The rank of the discrepancy for the data provides the exact significance level of the Monte Carlo test.

The limitations of a Monte Carlo test are clear. It tests only a simple hypothesis and parameters need to be given a priori. The last problem is circumvented by estimation of parameters in one plot of the data and testing the goodness-of-fit with these estimates in the other plot. The (overall) intensity parameter, however, is removed by conditioning on the observed number of events. The tests are therefore conditional tests, that may differ from unconditional tests (cf. Ripley, 1977).

For the Cox process the expression for the K-function is rather inconvenient. Therefore, instead of the theoretical K-function the mean of the m estimated K-functions of simulations and data is used in the goodness-of-fit test. This does not affect the exchangeability of the discrepancy values.

The Monte Carlo test depends in its detail on the formal definition of the discrepancy. A more informal assessment may supplement the test as provided by a graph of the theoretical summary description, its estimate from the data and the upper and lower simulation envelopes (Ripley, 1977). The envelopes are pointwise minima and maxima of the functional summary description. The number of simulations in this study is either 19 or 99. Tests for the pine data are accompanied by graphs based on $K(t)^{\frac{1}{2}}/\pi$, which is linear in t for the Poisson process. Disadvantages of the simulation envelopes are that they depend on the number of simulations and have a high pointwise variance. Alternatively, lower and upper quartiles, or the percentage point that is chosen to be relevant for the tests, can be given.

2.4 Confidence Regions

Lack of knowledge of the distribution of the statistics we use for point estimation hampers extension to interval estimation. I propose a pragmatic approach that has attractive properties under ideal conditions.

A 95% confidence region is a stochastic region in the parameter space that contains the true parameter of the process with 95% probability. A confidence region can be constructed by a possibly infinite number of tests, one for each θ in the parameter space, of the simple hypothesis that θ is the true parameter of the process. The confidence region consists of those values of θ for which the hypothesis is not rejected at the 5% level of significance. The construction of the test is arbitrary as long as no optimum properties for the confidence regions are required. As such the Monte Carlo tests based on (21) could be used, with a minimum of 19 simulations for each θ to guarantee the 95%-coverage property of the confidence region.

In constructing a confidence region we partition the parameter space into two exclusive sets, the set S of plausible values of θ and the set $\overline{S}$ of implausible values of θ. If S is closed, possibly after redefinition

of boundary points, then we want in fact to map a binary mosaic with a minimal number of sample points, i.e. tests for values of θ. As our test would be a Monte Carlo test, 'observation errors' occur; the problem of mapping a mosaic is discussed by Switzer (1971). Mapping S would be enormously facilitated if S is convex, so that a crude search over the parameter space suffices in practice. This approach is not followed up any further.

The problem to which confidence regions are an answer, is in general how precisely parameters can be estimated given an estimation procedure and how the precision can be estimated from the data. Given a model with prescribed parameters the distribution of the estimator of the parameters can be determined in principle by simulation of the model. The 'spread' of this distribution determines how precisely the parameters can be estimated for this model, for a given estimation procedure. The point estimate derived for the data provides under suitable conditions the approximate value of the true parameter. Simulation of the distribution for the process with this point estimate as true parameter will be particularly revealing. Under a number of assumptions this distribution can be used to construct a confidence region.

Suppose that for every θ the distribution of $\hat{\theta}$ is a bivariate Normal distribution with mean θ and covariance matrix S that does not depend on the value of θ - a rather restrictive assumption. Then, $\hat{\theta}$ as estimated from the data, is a bivariate Normal quantity, while S can be estimated from, say, m simulations. The ellipsoid

$$(\theta-\hat{\theta})\hat{S}^{-1}(\theta-\hat{\theta}) \leq \frac{2(m-1)}{m-2} F^{0.05}_{2,m-2} \qquad (22)$$

is a 95% confidence region for θ where $F^{0.05}_{2,m-2}$ is the 5% point of the F-distribution with 2 and m-2 degrees of freedom.

To satisfy the normality assumption transformation of the parameter may help. In the pine data a confidence ellipsoid is constructed for the logarithm of the parameters of the Gaussian cluster process. The assumption that S does not depend on θ should at least hold in the region of the parameter space to which point estimates are confined with high probability. I recommend that S is estimated from simulations with the point estimate as true parameter. No further checks on the assumptions have been made for the pine-data. Note that the total number of events is treated as an ancillary statistic on which is conditioned.

2.5 Numerical Procedures

Simulations of the processes are conditioned on the number of events observed in the plots. If necessary, periodic boundary conditions are imposed, i.e. the unit square is wrapped around a torus and the location of an event is (x_1 modulo 1, x_2 modulo 1). These conditions avoid edge-distortion.

Computer programs were written in FORTRAN IV and APL and run on the IBM 370 of the Northumbrian Universities Multiple Access Computer (NUMAC) at Newcastle upon Tyne. Routines for random number generation and the Simplex algorithm were taken from the NAG FORTRAN library (Anon, 1977). New FORTRAN programs (FGHAT and KHAT) were written to calculate the F-, G- and K-function. Ripley (1977) reported numerical instability in his procedure to estimate K(t). This instability arose from the way in which the weights for each pair of events were calculated. In KHAT the problem is avoided and the result is a far more efficient program. As the practical interest in K(t) is limited to the smaller values of t, KHAT is written for $t \leq 0.5$ with arbitrary interval width. Note that the weights thus cannot exceed four.

3. RESULTS

3.1 Summary Statistics.

Fig. 1 shows the positions of the pines in plot 1 and plot 2. Figs. 2 and 3 give the F- and G-functions of nearest neighbour distances for the two plots. For comparison the theoretical curve for the Poisson process (7) is given in Fig. 2 while this curve is used as abscissa in Fig. 3. The number of event-to-nearest-event distances, as shown by G(t), for values of t below 0.1 (1m in the field) exceeds the expected number under complete spatial randomness. The distribution of point-to-nearest-event distances, F( ), does show clustering but not as markedly as G(.).

Figs. 4a and 5a show the estimated K-function of event-event distances with for comparison the parabolic K-function of the Poisson process and the K-function of a Gaussian cluster process. In Figs. 4b, c and 5b, c transformations are shown for which the Poisson process is the zero-function. Notice the similarity between the overall shape of the K-function in plot 1 and plot 2. On small scale (below t = 0.05) there is marked clustering but on a larger scale (t between about 0.1 and 0.2) there appears to be some regularity in the pattern of the pines that is interesting in view of earlier remarks about scarification.

3.2 Goodness-of-Fit of Poisson Process and Heather-Based Cox Process.

The discrepancy (21) between the data and the Poisson process is 0.026 and 0.030 for plot 1 and plot 2, respectively. The Monte Carlo test based on 99 simulations has a level of significance of 0.01 for both plots; hence, the hypothesis, that the pines are completely randomly distributed, is rejected. The simulation envelopes are shown in Fig. 6. The K-function of the data lies outside the simulation envelopes for values of t below about 0.08.

The hypothesis that the pines are completely randomly distributed, but with different intensities on and off heather, is tested by simulation of the Cox process with a C-mosaic for the heather as random two-state intensity function. Diggle (in prep.) estimated the intensity of the centres of the discs as 221 and 211 and the parameters of the Weibull distribution (10) of the radii as $(\delta, \kappa, \rho) = (0.0281, 0.8471, 144.7)$ and $(0.0226, 1.011, 128.4)$ for plot 1 and plot 2, respectively. The intensities of pines off heather relative to the intensities on heather are estimated as 0.3009 and 0.4089 for plot 1 and plot 2, respectively. The Monte Carlo test in which the parameter estimates of the plots are not exchanged has level of significance 0.01 for both plots. Exchanging parameter estimates would give a worse fit. The envelopes and mean of 99 simulations are shown in Fig. 7. The discrepancies are 0.024 and 0.027. Note that the K-function of the Poisson process and this Cox process hardly differ, hence estimation of the relative intensities via the K-function would lead to unsatisfactory estimates.

3.3 Gaussian Cluster Process

3.3.1 Estimation

The two parameters of the Gaussian cluster process are (ρ, σ) with ρ the reciprocal of the intensity of the parent process and σ the standard deviation of the normal distribution which governs the spread of daughters around a parent. Estimates for (ρ, σ) are derived by minimizing the discrepancy (21). To avoid negative estimates for ρ or σ during the search of the Simplex algorithm the parameter-space is transformed to $(\log \rho, \log \sigma)$. A convenient initial estimate for ρ can be derived by noting that $\max_t \{K(t) - \pi t^2\} = \rho$. A number of initial estimates (ρ and σ both ranging between 0.01 and 0.10) were tried to enhance the chance of finding

5. TABLES AND FIGURES

Table 1 Point estimation of (ρ, σ) of Gaussian cluster processes

Upper limit of integration	K(.)		$\{K(.)\}^{\frac{1}{2}}$				log $\{K(.)\}$	
	plot 1		plot 1		plot 2		plot 1	
	ρ	σ	ρ	σ	ρ	σ	ρ	σ
0.025	0.035	0.027	0.010	0.011	0.008	0.010	0.018	0.019
0.05	0.010	0.012	0.010	0.011	0.011	0.014	0.010	0.014
0.075	0.010	0.011	0.010	0.012	0.011	0.013	0.010	0.014
0.1	0.008	0.011	0.009	0.012	0.010	0.012	0.010	0.013
0.2	0.002	0.004	0.006	0.008	0.007	0.009	0.008	0.012
0.3	0.002	0.004	0.005	0.008	0.006	0.008	0.008	0.012
0.4	0.003	0.007	0.005	0.008	0.006	0.009	0.008	0.012
0.5	0.010	0.180	0.005	0.008	0.006	0.008	0.008	0.012

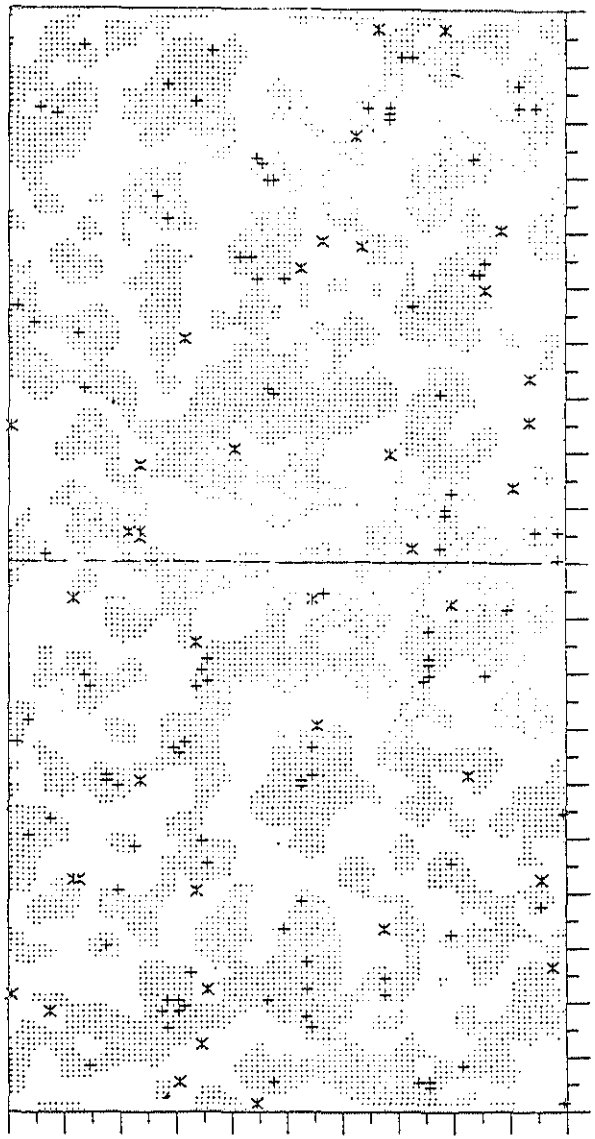

Figure 1 Pinus silvestris (crosses) and Calluna vulgaris (dots) in a 10m×20m area at Ivantjärnheden in Central Sweden (+ : on heather; • : off heather).

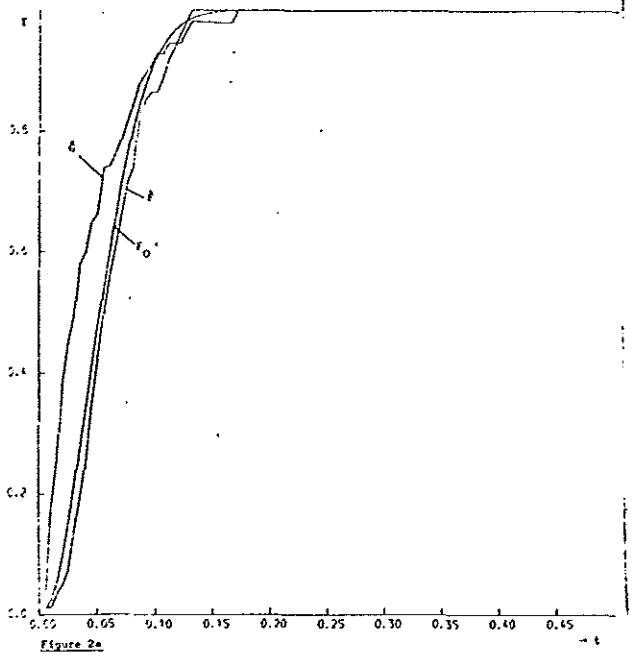

Figure 2a

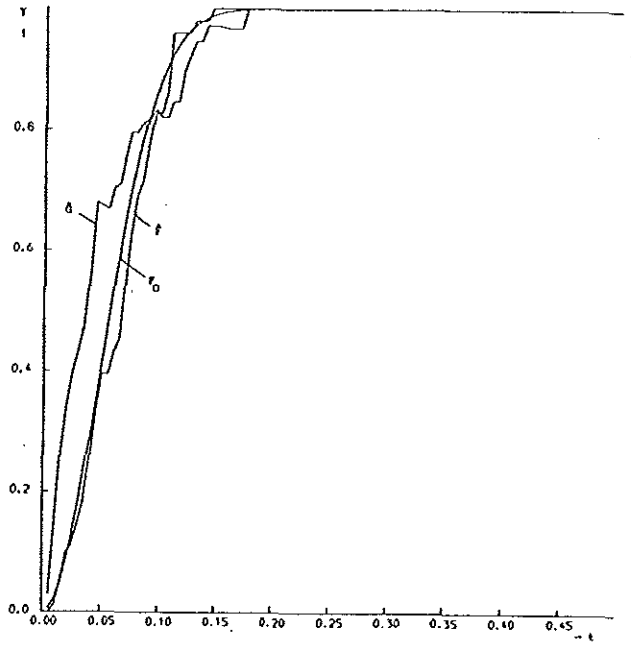

Figure 2b

Figure 2 Distribution of nearest neighbour distances for the pines. $\hat{F}(t)$: point-event distances; $\hat{G}(t)$: event-event distances; $F_0(t)$: marginal distribution function under complete spatial randomness.

(a) Plot 1 (b) Plot 2

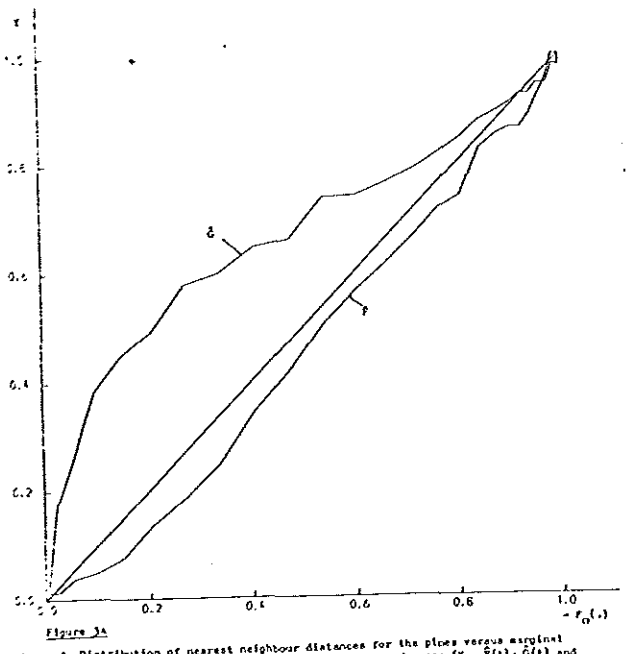

Figure 3a

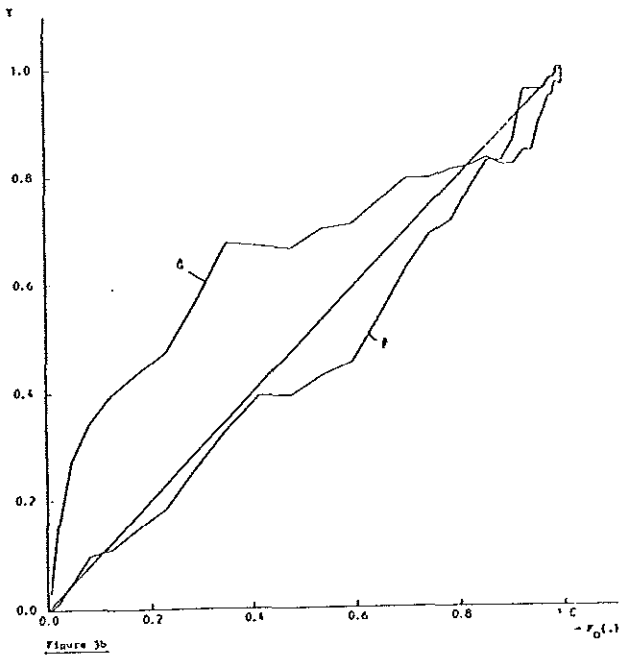

Figure 3b

Figure 3 Distribution of nearest neighbour distances for the pines versus marginal distribution function under complete spatial randomness (Y = $\hat{F}(t)$, $\hat{G}(t)$ and $F_0(t)$; codes as in Fig. 2.)

(a) Plot 1 (b) Plot 2

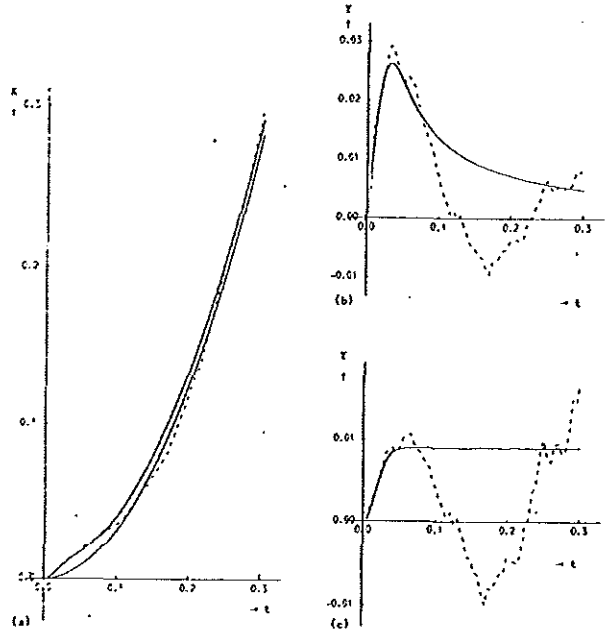

Figure 4 K-function of inter-pine distances in plot 1. (--- $\hat{K}(t)$; —— $K(t)$ for the Gaussian cluster process with $\rho = 0.009$ and $\sigma = 0.012$)
(a) $K(t)$; the parabolic function is the K-function of a Poisson process;
(b) $Y(t) = \{K(t)/\pi\}^{\frac{1}{2}} - t\}$;
(c) $Y(t) = K.t) - \pi t^2$.

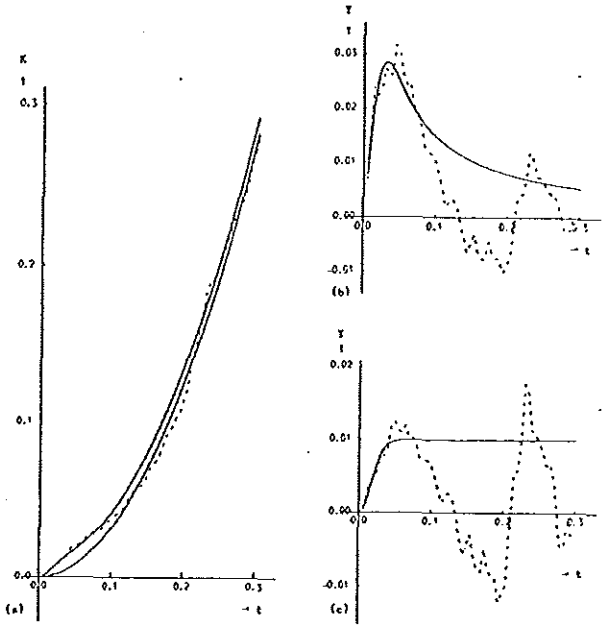

Figure 5 K-function of inter-pine distances in plot 2. Codes as in Fig. 4. ($\rho = 0.010$; $\sigma = 0.012$).

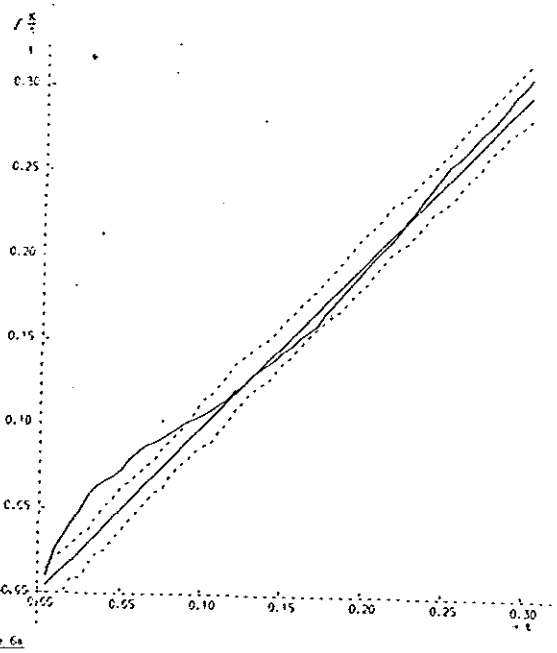

Figure 6a

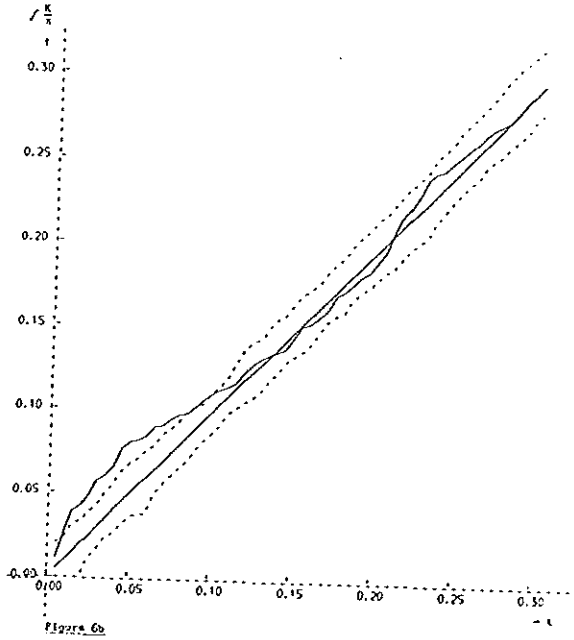

Figure 6b

Figure 6 Poisson process. Observed and theoretical K-function (——); envelopes of 99 simulations (---).
(a) Plot 1 (b) Plot 2

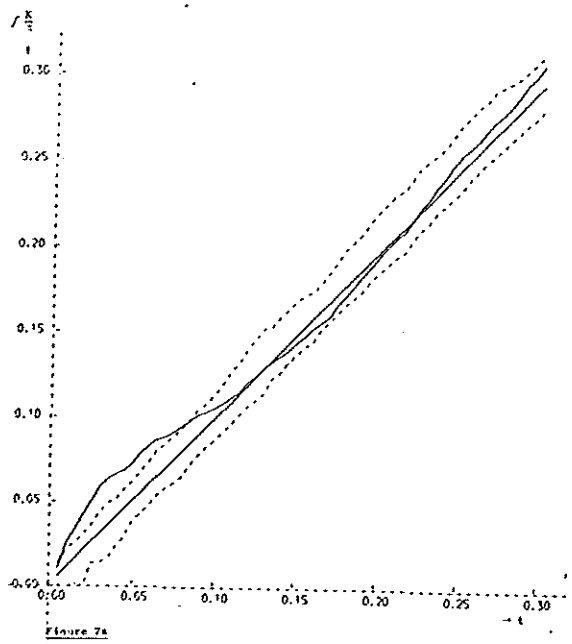

Figure 7a

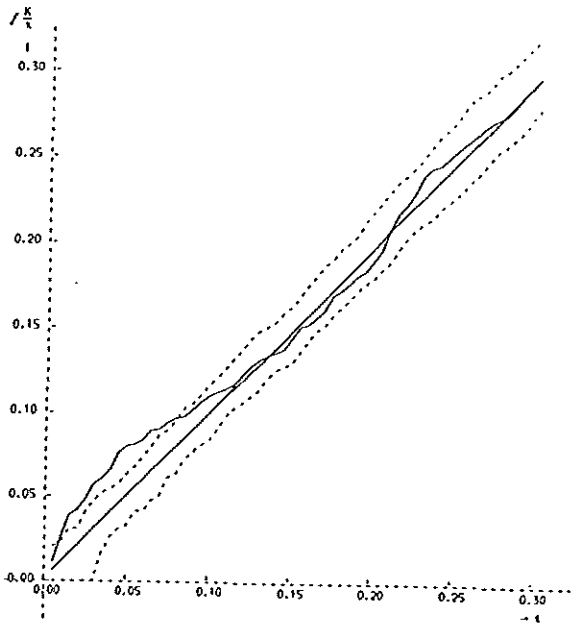

Figure 7b

Figure 7 Heather-based Cox process. Observed and mean K-function (——), envelopes of 99 simulations (---).

(a) Plot 1; relative intensity $\lambda_0/\lambda_1 = 0.3009$;

(b) Plot 2; relative intensity $\lambda_0/\lambda_1 = 0.4089$.

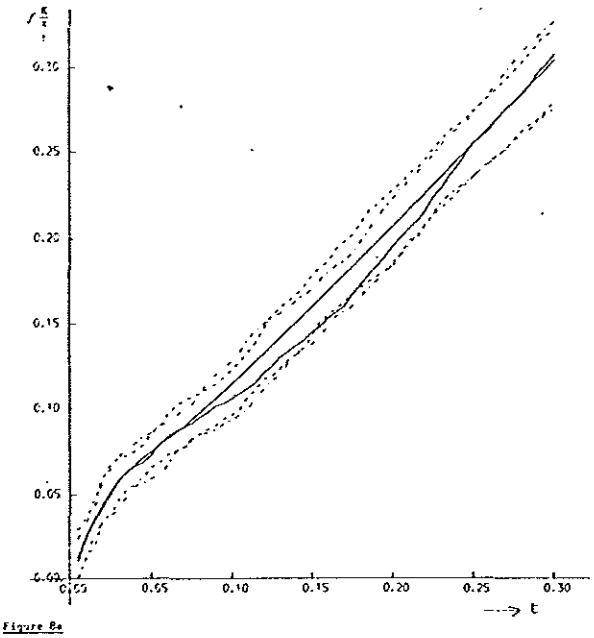

Figure 8a

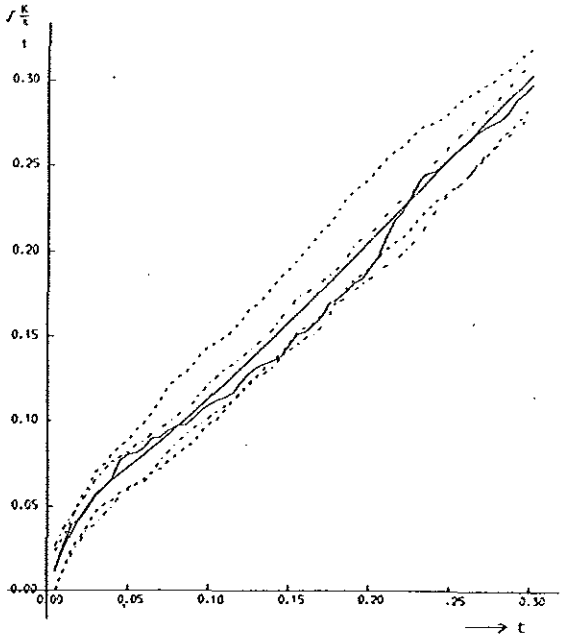

Figure 8b

Figure 8 Gaussian cluster process. Observed and theoretical K-functions (——), envelopes of 19 simulations with Poisson (---) and fixed (-·-·-) number of parents.

(a) Plot 1; $\rho = 0.010$, $\sigma = 0.012$.

(b) Plot 2; $\rho = 0.009$, $\sigma = 0.012$.

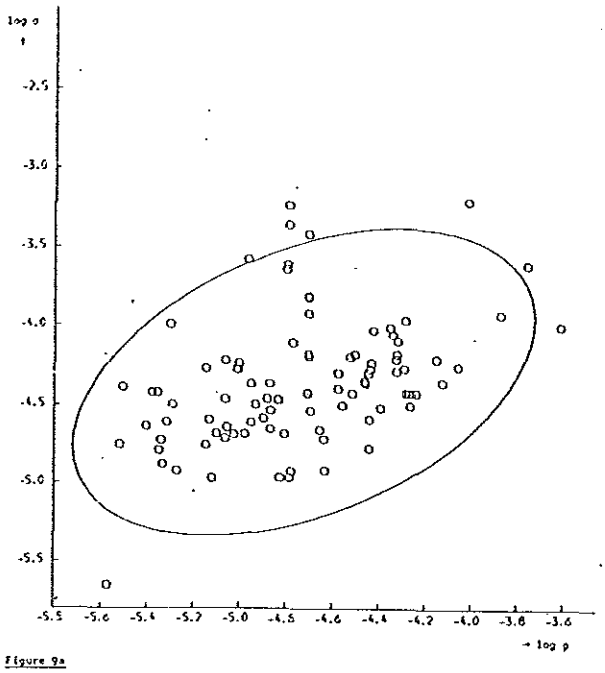

Figure 9a

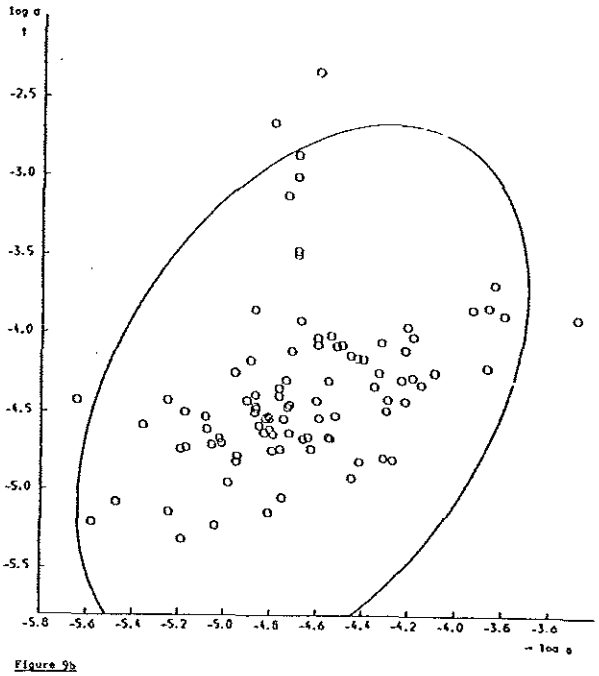

Figure 9b

Figure 9 Scatter diagram with 95% confidence ellipsoid of log σ vs log ρ for Gaussian cluster process with a Poisson number of parents.

(a) Plot 1: $\rho = 0.009$, $\sigma = 0.012$.

(b) Plot 2: $\rho = 0.010$, $\sigma = 0.012$.

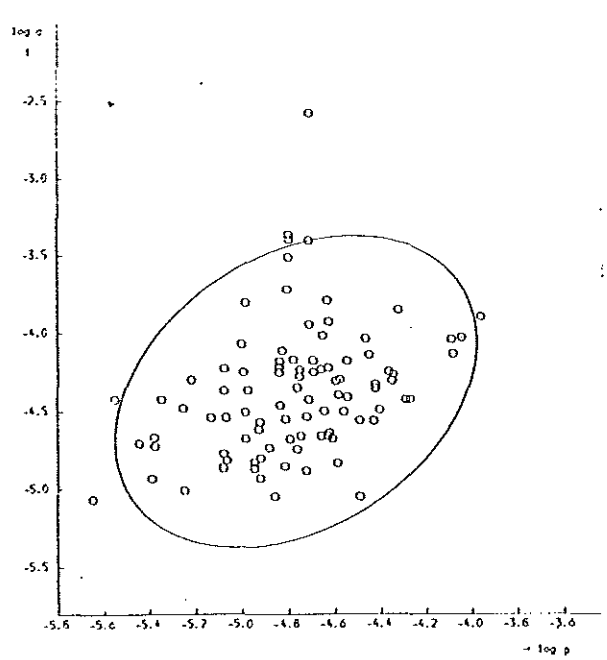

Figure 10a

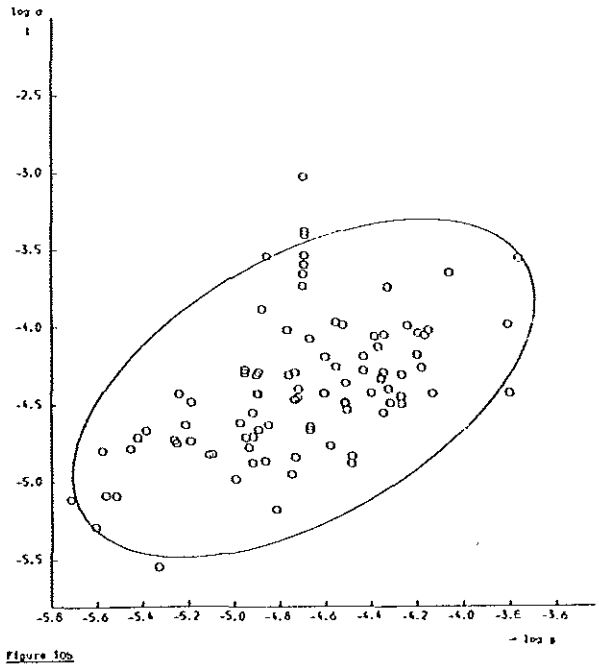

Figure 10b

Figure 10 As Fig. 9 for Gaussian cluster process with a fixed number of parents.

the global minimum. Table 1 shows the effect on the estimates of slightly different definitions of the discrepancy. Use of K(.) instead of $K^{\frac{1}{2}}(.)$ in (21) increases the dependence of the estimates on the range of integration. The estimates with $K^{\frac{1}{2}}(.)$ and a range of integration of 0.1 were chosen as final estimates. Note that above t = 0.1 the estimated K-function lies for some t below the K-function of the Poisson process, while this is impossible for the theoretical K-function of a Poisson cluster process. Extending the range of integration means trying to let the model fit over a range of distances that cannot be fitted properly. The final estimates, used throughout Part II, are (ρ, σ) = 0.009, 0.012) and (0.010, 0.012) for plot 1 and 2 respectively, with discrepancies of 0.0005 and 0.0008.

3.3.2 Goodness-of-fit

The parameter estimates of plot 2 were used in simulations to assess the goodness-of-fit of the model in plot 1 and vice-versa. With either a fixed or Poisson number of parents in 19 simulations the Monte Carlo test based on (21) has levels of significance of 0.95 and 1.0 for plot 1 and 2. However, the simulation envelopes (Fig. 8) touch the K-function of the data for values of t of about 0.15. If the range of integration in (21) is increased to 0.25 the Monte Carlo test gives still levels of significance of 0.75 (2 times) in plot 1 and 0.55 and 0.75 in plot 2 for 19 simulations with a Poisson and fixed number of parents respectively. Statistically, the second order properties of data appear to be consistent with a Gaussian cluster process.

In plot 2 the envelopes of the simulations with a Poisson number of parents are much wider than of the simulations with a fixed number of parents (Fig. 8b). Obviously the introduction of variability in the

number of parents increases the overall variability. However, further simulations showed that Fig. 8b is untypical: the pointwise standard deviations in K(t) based on 100 simulations are almost equal for simulations with fixed and Poisson number of parents.

3.3.3 Confidence regions

Fig. 9 shows estimates of the logarithms of the parameters for 100 simulations of the Gaussian cluster process (Poisson number of parents). The true parameters of the process are the point estimates (0.009, 0.012) and (0.010, 0.012) for plot 1 and 2 respectively. The 95% confidence ellipsoids assume Normality in the logarithm of the parameters. With a fixed number of parents (Fig. 10) the confidence ellipsoids have a smaller area. The confidence ellipsoid in Fig. 9b is distorted by an extreme estimate. If this estimate is excluded, the Figs. suggest plausible values between 44 and 270 for the number of clusters (ρ^{-1}) and between 5cm and 30cm for the spread (σ) of the clusters around their centre.

4. DISCUSSION

The shape of the distribution of inter-pine distances, K(.) is very similar in the two plots and indicates clustering on small scale and regularity on larger scale. The small scale clustering has been modelled by a Gaussian cluster process. The large scale regularity appears to be compatible with sampling fluctuations in simulations of this model. That the regularity occurs in both plots of pines is discomforting but shows the value of the data-splitting exercise. The Redwood data discussed in Ripley (1977) and Diggle (1978) show a similar shape of K(.).

The parameter estimates for the Gaussian cluster model quantify concisely the second-order properties of the observed pattern and,

together with their confidence region, provide a basis for comparison of pattern in other developing stands of pine or in other variables. The clustering is much more pronounced than expected on the basis of the mosaic pattern of the heather (Figs (6)-(8)) and its scale is much smaller than the scale of patches of heather. The model is formulated in terms of contagion but (16) shows that an interpretation in terms of heterogeneity is equally satisfactory. Note that it is possible to subject the model to further tests based on the nearest-neighbour distributions or any other relevant summary description, but the danger of data-dredging should be recognised.

CHAPTER 3 TESTS FOR ASSOCIATION

1. INTRODUCTION

In this chapter the question is examined whether there is statistical evidence that there is association between pine and heather in Fig.1. Among possible measures of association I choose the number of pines on heather.

Assume temporarily that the pines are completely randomly distributed in the gaps with intensity λ_0 and completely randomly distributed in the patches of heather with intensity λ_1. Then the hypothesis of no association is equivalent with the hypothesis $\lambda_0=\lambda_1$, which can be tested with a binomial test, because N_1, the number of pines in heather is binomially distributed with parameters n and p. Here n is the total number of pines and $p = \lambda_1 p_1/(\lambda_0 p_0+\lambda_1 p_1)$ with p_1 the cover of heather and $p_0 = 1-p_1$. If $\lambda_0 = \lambda_1$ then $p = p_1$. For the data, standardized Normal deviates are on this basis 4.86 and 3.40 for plots 1 and 2, thus the hypothesis is convincingly rejected. However, the assumptions are falsified in the previous chapter, where the heather-based Cox process has

been rejected as a plausible model. Because the pines are clustered the variance of N_1 will be greater than under the binomial model and it is not clear beforehand how large the effect of clustering on the variance will be. What is needed is a more reasonable model for pine and heather expressing the hypothesis of no association. I shall consider a permutation model based on random shifts of the observed patterns and a parametric model based on independent processes, the C-mosaic process for the heather and the Gaussian cluster process for the pines as described in the previous chapter. The variance of N_1 under these models will be compared with the variance under simplified models for which the distribution theory is known.

2. THEORY AND METHODS

Without modelling the patterns of pine and heather the minimal assumption under which a test is available is stationarity. Stationarity implies that the inter-pine difference-vectors are complete minimal sufficient statistics. Conditionally on these differences we may think of the pines as a fixed irregular network of points. Under the null hypothesis of no association every shift of the network with respect to the mosaic of the heather is equally probable, hence these shifts specify a permutation distribution (Cox and Hinkley, 1974, §6.2). I base a Monte Carlo test on the permutation distribution. If the observed number of pines on heather is extreme with respect to the number on heather after each of, say, 19 random shifts the data are not consistent with the null hypothesis at 5% level of significance in a one-sided test. Let X be bivariate uniformly distributed in the unit square, then a random shift is defined by adding X to the locations of the pines and applying the periodic boundary conditions (Ch 2 §2.5). The border effect violates the conditionality argument slightly but apart from this, the patterns

of pine and heather are kept unchanged.

In the parametric approach the patterns are taken to be realisations of stochastic models. Under the null hypothesis the two stochastic processes are independent so that the definition of the processes implies the null distribution of the number of pines on heather. The Monte Carlo test is based on simulations of the Gaussian cluster process for the pines and the C-mosaic process of the heather (Ch 2) and counts for each simulation the number of pines on heather.

The parametric approach has the disadvantage that more assumptions are needed before the test can be carried out, but the advantage that the patterns are summarized in parameters that give insight into their nature and allow comparison with ostensibly similar data-sets.

If the variance of the number of pines on heather (N_1) is thought to be a sufficient basis for a test, then a number of other approaches are possible that do not require simulations. The variance of N_1 has been derived analytically under simplified models. The simplest model is, of course, the binomial model with

$$\mathrm{var}(N_1) = np(1-p)$$

More realistic models can be derived from the Gaussian cluster process. Fitting the Gaussian cluster process for the pines gives a value for σ (0.012) which suggests very tight clusters. Consequently, if the parent is on(off) heather, then the daughters are on(off) heather with high probability.

Under the assumption that this probability equals unity ($\sigma = 0$) more tractable models are obtained. The number of parents will be based on the previously fitted value of ρ. Attention is restricted here to the hypothesis of no association. Three models are considered. In the first two models I condition on the observed number of events, while

the number of parents is either a fixed number or is Poisson distributed. In the third model I do not condition and both the number of parents and total number of events will be Poisson distributed.

In the first model both the number of parents and the number of events is fixed. The number of parents on heather, O, then is binomially distributed, $O \sim Bi(m,p_1)$ with m (for the unit region) the nearest integer to ρ^{-1}, the total number of parents, and p_1 the cover of heather. The n events observed are to be assigned at random to the parents, thus, given O, the number of pines is binomial, $N_1 \sim Bi(n, O/m)$. The marginal distribution of N_1 is a compound distribution which I call the binomial-binomial distribution, symbolically (cf. Johnson and Kotz, 1969)

$$Bi(n, O/m) \underset{O}{\wedge} Bi(m,p_1) \qquad (1)$$

This distribution is not mentioned in Johnson and Kotz (1969), but the variance of N_1 can be derived by standard methods based on the probability generating function, or directly from the conditional mean and variance,

$$\mathrm{var}(N_1) = n\, p_1(1-p_1)\left(1 + \frac{n-1}{m}\right) \qquad (2)$$

In the second model the number of parents is not fixed but is Poisson distributed with mean ρ^{-1}. Therefore the previous distribution is compounded over m, symbolically

$$Bi(n, O/M) \underset{O}{\wedge} Bi(M, p_1) \underset{M}{\wedge} Poi(\rho^{-1}) \qquad (3)$$

Here a slight problem arises for M=O, as there was (unmentioned) in the simulations; I discard such realizations and modify the Poisson distribution accordingly. The variance of N_1 then becomes

$$\mathrm{var}(N_1)^* = n\, p_1(1-p_1)\left\{1 + (n-1)\left[\exp(\rho^{-1}) - 1\right]^{-1} \sum_{m=1}^{\infty} \frac{\rho^{-m}}{m\, m!}\right\} \qquad (4)$$

The summation in (4) equals $Ei(\rho^{-1}) - \gamma + \log \rho$ where Ei(.) is the exponential integral and γ Euler's constant (Abramowitz & Stegun, 1964,

5.1.10). For small ρ (say $\rho < 0.02$)

$$\text{var}(N_1) \approx n\, p_1(1-p_1)(1 + \rho(n-1)) \qquad (5)$$

hardly differs from (2).

In the third model both the number of parents and the number of daughters are Poisson distributed. Then the distribution of N_1 is a generalized distribution, the Poisson-Poisson distribution,

$$\text{Poi}(p_1\rho^{-1}) \vee \text{Poi}(\mu) \qquad (6)$$

where $p_1\rho^{-1}$ is the mean number of parents on heather and μ the mean number of daughters per parent. The variance of N_1 is (Johnson and Kotz, 1969; Pielou, 1977)

$$\text{var}(N_1) = p_1\rho^{-1}\mu(1+\mu) = p_1 n(1 + \rho n) \qquad (7)$$

where n now is the expected number of pines in the area; thus (7) is greater than (2) and (5) and shows the effect of conditioning.

Notice that in the above described formulae for the variance of N_1 the areal proportion of heather is fixed while this proportion is stochastic in the simulations.

3. RESULTS

The distribution of N_1, the number of pines on heather is shown in Fig. 1 based on 250 simulations of the population model with C-mosaic and Gaussian cluster process (Poisson number of parents). Note that no parameters of C-mosaic or cluster process are exchanged between the plots. The associated Monte Carlo test (one-sided) rejects the hypothesis of no association below the 1% level, as does the test based on the randomization model or the model with fixed instead of Poisson number of parents. The variances of N_1 as estimated from these 250 simulations of each model are given in Table 1. In addition the variance of N_1 is estimated as based on the binomial, Poisson-Poisson and binomial-binomial distributions

(Table 1). For comparison the population model with fixed number of parents and σ very small ($\sigma = 0.0001$) is included.

It is surprising that the variance of N_1 with a fixed number of parents is much smaller (F-test, $P < 0.05$, two-sided) than with a Poisson number of parents as the difference is negligible if $\sigma = 0$ (cf. (2) with (5)). However, two additional runs of simulations give variances of 27.7 and 33.4 for the model with Poisson number of parents and 30.3 and 34.0 for the model with fixed number of parents. The difference is thus well within the simulation fluctuation. The difference between the variance for $\sigma = 0$ (binomial-binomial distribution) and $\sigma = 0.012$ must be due to the variation in proportion of heather in the simulations. The Poisson-Poisson distribution gives, of course, the largest variance. Standard normal deviates are for this model 2.62 and 1.90, the latter just not significant at 5% in a two-sided test.

4. DISCUSSION

The model fitting in Chapter 2 could just be seen as giving a concise description of the observed pattern or the second order statistics thereof. However, when the models are used in statistical tests this restrictive view is not enough: the variability of the relevant phenomena should be mirrored in the model. This variability determines the distribution of N_1 with which the observed number of pines on heather is compared.

The range of models considered all reject the hypothesis of no association. The permutation model requires minimal assumptions, it does not even assume isotropy. It is at first sight surprising how close the variance of N_1 for this model is to the variance for the parametric models. The explanation is that, under the assumptions of stationarity,

isotropy and no association, the variance of N_1 depends only on the covariance function c(t) of the heather mosaic and the distribution of inter-pine distances (Pt I Ch 2 (5))

$$\mathrm{var}(N_1) = n\, c(o) + \sum_{k \neq l} \sum E_{\{\underset{\sim}{x}_i\}} \quad c(\|\underset{\sim}{x}_k - \underset{\sim}{x}_l\|) \qquad (8)$$

where the $\{\underset{\sim}{x}_i\}$ are the locations of the pines. As both parametric models are fitted to match the second-order properties, they give about the same variance as the non-parametric model. Moreover, with the pines an irregular but fixed network of pins that sample a stationary mosaic process, the covariance function of the mosaic can be estimated, preferably from a separate point quadrat sample, and hence (8) (cf. Pt I Ch 3 §2.4). Of course, this approach gives a test of significance of the association only, while the parametric approach gave insight into the patterns of pine and heather as well.

Estimation of (8) with sample points distributed according to a Gaussian cluster process - a rather curious sampling design - is not easy: the distribution of inter-event distances within the unit-square differs from its stationary analog K(t) and the analytical expression for it will be complex (cf. Bartlett, 1964).

The statistical evidence for association motivates the construction of a combined model for pine and heather that describes their pattern and accounts for the association. Such a model can be based on the idea of an interrupted point process (Stoyan, 1979). Assume we have two stationary, isotropic spatial processes, Π a point process and $Z(\underset{\sim}{x})$ a stochastic process with realized values between 0 and 1 ($0 \leq z(\underset{\sim}{x}) \leq 1$). The interrupted process Π_I is defined by the rule: retain an event at $\underset{\sim}{x}$ with probability $z(\underset{\sim}{x})$ and discard it with probability $1 - z(\underset{\sim}{x})$. We have already seen one example: the heather-based Cox process (Ch 2) was an interrupted point process with Π a Poisson process with intensity λ_1

and $Z(\underset{\sim}{x})$ defined by the binary mosaic process $Z_0(\underset{\sim}{x})$ of the heather: $Z(\underset{\sim}{x}) = \lambda_0/\lambda_1$ if $Z_0(\underset{\sim}{x}) = 0$ ($\underset{\sim}{x}$ off heather) and $Z(\underset{\sim}{x}) = 1$ if $Z_0(\underset{\sim}{x}) = 1$ ($\underset{\sim}{x}$ on heather). If Π is replaced by the Gaussian cluster process then the second-order statistics fit, and the number of pines on heather as well.

The K-function of an interrupted point process can be derived. Let $C^*(t) = E\{Z(\underset{\sim}{x})Z(\underset{\sim}{y})\}$ with $\|\underset{\sim}{x}-\underset{\sim}{y}\| = t$, then $C^*(t)$ is the probability that two events of Π, distance t apart, both survive. With g(t) and $g_I(t)$ the joint probability densities of the occurrence of a pair of events, distance t apart, of Π and Π_I, respectively ((2) in Ch 2), we have

$$g_I(t) = g(t)C^*(t) \qquad (9)$$

Let $Z(\underset{\sim}{x})$ have expected value p, variance σ_z^2 and correlation function r(t), so that $C^*(t) = p^2 + \sigma_z^2\, r(t)$. Define $K'(t) = \frac{dK(t)}{dt}$, then with (3) and (4) in Ch 2,

$$K_I(t) = K(t) + \frac{\sigma_z^2}{p^2} \int_0^t K'(u)r(u)du \qquad (10)$$

The parameter estimates of Π_I can be based on (10), and will in general not be identical with the parameter estimates of Π. I expect that updating the parameter estimates for ρ and σ for the pines changes the estimates only slightly because the K-function of Π_I in (10) does not change much (cf. K-functions of Poisson process and heather-based Cox process in Ch 2). Unfortunately, we can still not fit the large-scale regularity in the point pattern in this way: the C-mosaic has correlation function $r(u) \geq 0$, hence $K_I(t) \geq K(t) \geq \pi t^2$.

CHAPTER 4 DISCUSSION

Statistical analysis of contagion and heterogeneity where the latter is caused by a known process is difficult but the present analysis shows that it is possible. The analysis has not explained the genesis of Fig. 1 but has shown that there is evidence of association between pine and heather and, moreover, that the pattern of the pine is clustered beyond the effect of association; hence, that the pattern of pine is subject to other factors as well. The conclusions of an observational study cannot reach further and experiments, if well designed, may proceed from here to establish the causal factors involved.

Now the formal conclusions have been stated I propose an explanation for the pattern. A tractor scarified the field parallel to the long side and made furrows at 2m distance of each other. The fixation and germination conditions for the pines and possibly the heather were more favourable in the furrows than outside. On the map the pines tend to lie along lines (more evidently in plot 1 than in plot 2) while the K-function of both plots is periodic. (The permutation test did not assume isotropy!) The association with heather is due to the furrows, but the heather has grown out of the furrow to cover the rest of the scarified patches and possibly beyond. Of course, my explanation is no better than any one elses, until more is known either of the history of Fig. 1 or of the ecology of pine and heather by designed experiments.

Table 1. Variance of number of pines on heather under hypothesis of no association.

Model	Plot 1	Plot 2
Parametric Models:		
C-mosaic and Gaussian cluster process		
Poisson no. of parents	37.9	28.3
fixed no. of parents	27.2	24.3
fixed no. of parents, small σ (0.0001)	41.8	34.9
Permutation Model	39.4	29.5
Simplified Models:		
binomial	20.5	17.0
binomial-binomial (-Poisson)	35.4	28.4
Poisson-Poisson	71.1	56.6

Figure 1 Null distribution of number of pines on heather in parametric model. (C-mosaic and Gaussian cluster model with Poisson number of parents)(• : the observed number).

(a) Plot 1

(b) Plot 2

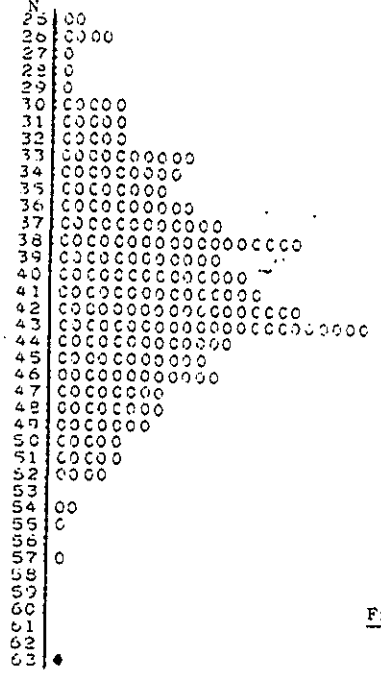

Figure 1a

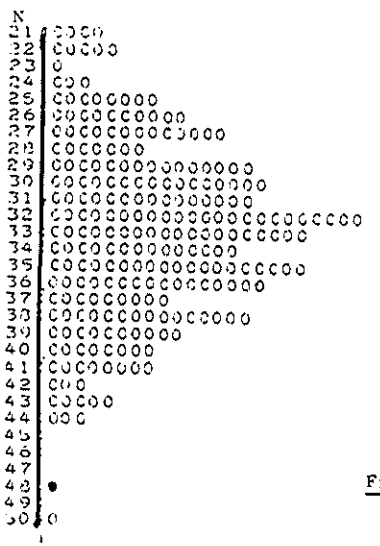

Figure 1b

REFERENCES

Abramowitz, M. and Stegun, I.A. (1964). Handbook of Mathematical Functions. U.S. National Bureau of Standards.

Anon (1977). NAG FORTRAN LIBRARY MANUAL. MARK 6 (FLM6). Numerical Algorithm Group, Oxford.

Armitage, P. (1949). An overlap problem arising in particle counting. Biometrika, 36, 257-266.

Bartlett, M.S. (1964). Spectral analysis of two-dimensional point processes. Biometrika, 44, 299-311.

Bartlett, M.S. (1971). Two-dimensional nearest neighbour systems and their ecological applications. In: Statistical Ecology, Vol. 1 (Patil, G.P. Pielou, E.C., Waters, W.E., eds.) Pennsylvania State University Press, University Park, 179-194.

Bartlett, M.S. (1975). The Statistical Analysis of Spatial Pattern. Chapman and Hall, London.

Bartlett, M.S. (1978). An Introduction to Stochastic Processes. 3rd Edn. Cambridge University Press, Cambridge.

Bartlett, M.S. and Besag, J.E. (1969). Correlation properties of some nearest-neighbour models. Bull. Int. Statist. Inst. 43, Book 2, 191-193.

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems. J.R. Statist. Soc. B, 36, 192-236.

Besag, J.E. (1975). Statistical analysis of non-lattice data. The Statistician, 24, 179-95.

Brown, D. (1954). Methods of Surveying and Measuring Vegetation. Commonwealth Bureau of Pastures and Field Crops, Bulletin 42, Commonwealth Agricultural Bureaux, Hurley.

Cochran, W.G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. Ann. Math. Statist. 17, 164-77.

Cochran, W.G. (1977). Sampling Techniques. 3rd Edn. John Wiley & Sons, New York.

Cox, D.R. (1972). The analysis of multivariate binary data. Appl. Statist. 21, 113-120.

Cox, D.R. and Hinkley, D.V. (1974). Theoretical Statistics. Chapman and Hall, London.

Cox, D.R. and Lewis, P.A.W. (1966). The Statistical Analysis of Series of Events. Chapman and Hall, London.

Diggle, P.J. (1977). The detection of random heterogeneity in plant populations. Biometrics, 33, 390-94.

Diggle, P.J. (1979). On parameter estimation and goodness-of-fit testing for spatial point patterns. Biometrics, 35, 87-101.

Diggle, P.J. (in prep.) Statistical analysis of binary mosaics.

Diggle, P.J. and Matérn, B. (1980). On sampling designs for the study of point-event nearest neighbour distributions in R^2. Scand. J. Statist. 7, 80-84.

Dupač, V. (1980). Parameter estimation in the Poisson field of discs. Biometrika, 67, 187-90.

Feller, W. (1968). An Introduction to Probability Theory and Its Applications. Volume I 3rd Edn. John Wiley & Sons, New York.

Fisser, H.G. and Van Dyne, G.M. (1966). Influence of number and spacing of points on accuracy and precision of basal cover estimates. J. Range Mgmt. 19, 205-11.

Goldsmith, F.B. (1973). The vegetation of exposed sea cliffs at South Stack, Anglesey. I: The multivariate approach. J. Ecol. 61, 787-818.

Goldsmith, F.B. and Harrison, C.M. (1976). In: Methods of Plant Ecology (Chapman, S.N., ed.). Blackwell Scientific Publications, Oxford.

Goodall, D.W. (1952). Some considerations in the use of point quadrats for the analysis of vegetation. Aust. J. Sci. B, 5, 1-41.

Grandell, J. (1976). Doubly Stochastic Poisson Processes. Lecture Notes in Mathematics, 529. Springer-Verlag, Berlin.

Greig-Smith, P. (1964). Quantitative Plant Ecology. 2nd Edn. Butterworths Scientific Publications, London.

Greig-Smith, P. (1979). Pattern in vegetation. J. Ecol. 67, 755-779.

Heslehurst, M.R. (1971). The point quadrat method of vegetation analysis: a review. Univ. Reading, Dept. Agric., Study No. 10.

Hillard, J.E. and Cahn, J.W. (1961). Evaluation of procedures in quantitative metallography for volume fraction analysis. Trans. Met. Soc. AIME, 221, 344-352.

Johnson, N.L. and Kotz, S. (1969). Discrete Distributions. Wiley-Interscience, New York.

Kemp, C.D. and Kemp, A.W. (1956). The analysis of point quadrat data. Aust. J. Bot. 4, 167-74.

Kendall, M.G. and Moran, P.A.P. (1963). Geometrical Probability. Griffin, London.

Klotz, J. (1973). Statistical inference in Bernoulli trials with dependence. Ann. Statist., 1, 373-379.

Kupper, L.L. and Haseman, J.K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments. Biometrics, 34, 69-76.

Lackritz, J.R. and Scheaffer, R.L. (1980). Point sampling in areal-fraction analysis. Commun. Statist.- Theor. Meth., A 9.(4), 355-374.

Mack, C. (1954). The expected number of clumps when convex laminae are placed at random and with random orientation on a plane area. Proc. Cam. Phil. Soc., 50, 581-585.

Matérn, B. (1960). Spatial variation. Medd. Statens Skogsforskningsinst. 49 (3), 1-144.

Matérn, B. (1971). Doubly stochastic Poisson processes in the plane. In: Statistical Ecology, Vol. 1. (Patil, G.P., Pielou, E.C., Waters, W.E., eds.) Pennsylvania State University Press, University Park, 195-213.

Matheron, G. (1975). Random Sets and Integral Geometry. John Wiley & Sons, New York.

McIntyre, G.A. (1953). Estimation of plant denstiy using line transects. Ecology, 41, 319-30.

Miles, R.E. (1976). Estimating aggregate and overall characteristics from thick sections by transmission microscopy. J. Microsc., 107, 227-233.

Miles, R.E. and Davy, P.J. (1976). Precise and general conditions for the validity of a comprehensive set of stereological fundamental formulae. J. Microsc. 107, 211-226.

Miller, J.B. (1964). An integral equation from phytology. J. Aust. Math. Soc. 4, 397-402.

Müller-Dombois, D. and Ellenberg, H. (1974). Aims and Methods of Vegetation Ecology. John Wiley & Sons, New York.

Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization. Computer J. 7, 308-313.

Neyman, J. and Scott, E.L. (1958). Statistical approach to problems of cosmology. J.R. Statist. Soc. B 20, 1-43.

Persson, H. (1978). Root dynamics in a young Scots pine stand in Central Sweden. Oikos, 30, 508-519.

Pielou, E.C. (1964). The spatial pattern of two-phase patchworks of vegetation. Biometrics, 20, 156-167.

Pielou, E.C. (1977). Mathematical Ecology. John Wiley & Sons, New York.

van der Pijl, L. (1972). Principles of Dispersal in Higher Plants. 2nd Edn. Springer-Verlag, Berlin.

Reid, W.P. (1955). Distribution of sizes of spheres in a solid from a study of slices of the solid. J. Math. Phys. 34, 95-102.

Ripley, B.D. (1977). Modelling spatial patterns (with discussion). J.R. Statist. Soc. B, 39, 172-212.

Roach, S.A. (1968). The Theory of Random Clumping. Methuen, London.

Robinson, P. (1955). The estimation of ground cover by the point quadrat method. Ann. Bot., 19, 59-66.

Silverman, B.W. (1978). Distances on circles, toruses and spheres. J. Appl. Prob. 15, 136-43.

Smith, W.L. (1958). Renewal theory and its ramifications. J.R. Statist. Soc. B, 20, 243-301.

Solomon, H. (1953). Distribution of the measure of a random two-dimensional set. Ann. Math. Statist. 24, 650-6.

Stoyan, D. (1979). Interrupted point processes. Biom. J. 21, 607-10.

Switzer, P. (1965). A random set process in the plane with a Markovian property. Ann. Math. Statist. 36, 1859-63.

Switzer, P. (1971). Mapping a geographically correlated environment. In: Statistical Ecology, Vol. 1 (Patil, G.P., Pielou, E.C., Waters, W.E., eds.) Pennsylvania State University Press, University Park, 235-269.

Tothill, J.C. (1978). Measuring botanical composition of grasslands. In: Measurement of Grassland Vegetation and Animal Production ('t Mannetje, L.,ed.) Commonwealth Bureau of Pastures and Field Crops, Bulletin 53, Commonwealth Agricultural Bureaux, Hurley.

Vere-Jones, D. (1968). Some applications of probability generating functionals to the study of input-output streams. J.R. Statist. Soc. B, 30, 321-33.

Warren-Wilson, J. (1963). Estimation of foliage denseness and foliage angle by inclined point quadrats. Aust. J. Bot. 11, 95-105.

Watson, G.S. (1971). Estimating functionals of particle size distributions. Biometrika, 58, 483-490.

Wicksell, S.D. (1925). The corpuscle problem, Part I. Biometrika, 17, 84-99.

Williams, D.A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 31, 949-952.

Williams, R.M. (1956). The variance of the mean of systematic samples. Biometrika, 43, 137-48.

Winkworth, R.E. (1955). The use of point quadrats for the analysis of heathland. Austr. J. Bot. 3, 68-81.

Winkworth, R.E. and Goodall, D.W. (1962). A crosswire sighting tube for point quadrat analysis. Ecology 43, 342-343.