Offprint from

Statistics in Theory and Practice
Essays in Honour of Bertil Matérn

Edited by Bo Ranneby

Diggle, P.J., and C.J.F. ter Braak (1982). Point sampling of binary mosaics in ecology. Pages 107-122 in Statistics in Theory and Practice. Essays in Honour of Bertil Matérn. Ed. B. Ranneby. Swedisch University of Agricultural Sciences. Section of Forest Biometry. S-901 83 Umeå, Sweden, 326 pp.

Erratum: pages 119 and 120 have been interchanged.

Published by

Swedish University of
Agricultural Sciences
Section of Forest Biometry
S - 901 83 Umeå, Sweden

POINT SAMPLING OF BINARY MOSAICS IN ECOLOGY

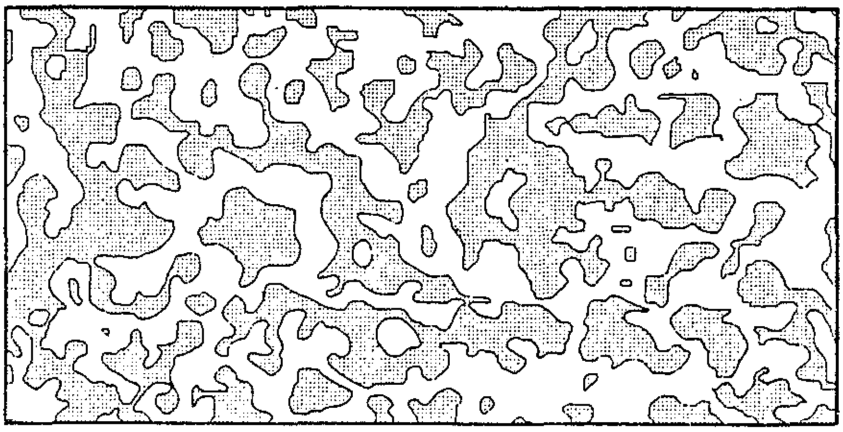
by Peter J. Diggle
Department of Statistics
University of Newcastle upon Tyne
Newcastle upon Tyne. NE1 7RU
England

and Cajo J.F. ter Braak
Institute TNO for Mathematics,
Information Processing and Statistics
P.O. Box 100
6700 AC Wageningen
The Netherlands

1. Introduction

A binary mosaic is a partitioning of a planar region, A say, into two qualitatively different sub-regions (Pielou, 1977; Matérn, 1979). Figure (1) shows an example in which the two phases of the mosaic indicate presence or absence of heather, *Calluna vulgaris*, in a 10m X 20m rectangular plot at Jädraås, Sweden. A different type of example would be one in which a continuous variable determines the phase according as it does or does not exceed some threshold value.

Figure (1). Incidence of *Calluna vulgaris* over a 10m X 20m area (from Diggle, 1981).



When a mosaic is mapped as in Figure (1) it might be possible to fit a stochastic model to the observed map. The estimated values of the parameters in the model then provide summary statistics for the data which can be used to compare ostensibly similar data-sets or, more ambitiously, to relate the observed pattern to possible causal factors. Diggle (1981) gives such an analysis of the data in Figure (1).

In many ecological applications, the region A is extensive and a complete mapping would be prohibitively expensive. In such circumstances, point sampling can be used to obtain data from the mosaic. This consists of recording the phase of the mosaic at each of a finite number of points located within A according to some suitable sampling design. The objectives of a statistical analysis of point sampling data are necessarily more limited than for a complete map, and might in the first instance consist simply of estimating the proportion of the total area of A occupied by each phase of the mosaic.

In §2 we describe two simple models for binary mosaics. One, the L-mosaic, is particularly tractable because it involves a simple exponential correlation function, but seems unrealistic. The second, the C-mosaic, is slightly less tractable, but provides a more plausible model of naturally occuring patterns, especially in vegitation. In §3 we investigate the efficiency of random sampling, frame sampling and systematic sampling designs for the estimation of areal proportions, using C-mosaic models and a simple but plausible cost function. We also use the data in Figure 1 to illustrate how the results on efficiency can be used to construct an appropriate sampling scheme, making crude guesses for the values of the C-mosaic parameters. A more detailed treatment of these topics is given by ter Braak (1980). Finally, §4 discusses briefly how the results relate to previous studies of point sampling methods by Kemp & Kemp (1956) and Rothery (1974).

2. Binary mosaic processes

(2.1) Definitions

A binary mosaic process is a binary-valued stochastic process $\{Z(\underset{\sim}{x}):\underset{\sim}{x}\in\mathbb{R}^2\}$. Any realisation $\{z(\underset{\sim}{x})\}$ partitions $\mathbb{R}^2$ into a set $S=\{\underset{\sim}{x}:z(\underset{\sim}{x})=1\}$ and the complement $\overline{S}=\{\underset{\sim}{x}:z(\underset{\sim}{x})=0\}$; in order to exclude point and line processes, S must be the closure of an open set in $\mathbb{R}^2$. Note that S is a random set in $\mathbb{R}^2$. Matheron (1975) has developed a very general theory of random sets: Stoyan (1979) gives an excellent review from an applied viewpoint. We shall present explicit results for binary mosaic processes without reference to the more general theory.

We consider only processes which are stationary and isotropic. The expectation of the areal proportion of S in A can then be written as $\mu=E[Z(\underset{\sim}{x})]$, which is independent of $\underset{\sim}{x}$. Further, the covariance function

$$\gamma(t) = E[(Z(\underset{\sim}{x})-\mu)(Z(\underset{\sim}{y})-\mu)]$$

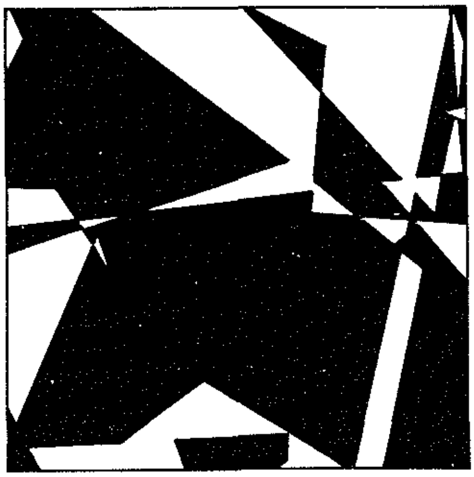
depends only on the distance t between $\underset{\sim}{x}$ and $\underset{\sim}{y}$. The correlation function is $\rho(t)=\gamma(t)/\gamma(0)=\gamma(t)/\{\mu(1-\mu)\}$.

(2.2) The L-mosaic

Following criticism by Bartlett (1964) of an analysis by Pielou (1964), Switzer (1965) showed that an exponential correlation function is admissible for a mosaic process. Consider a homogeneous Poisson process on the infinite cylinder, of intensity $\lambda/4$, and interpret this as a random line process in $\mathbb{R}^2$ with the points of the Poisson process determining the intercepts and orientations of the lines with respect to any fixed axis. These random lines partition $\mathbb{R}^2$ into convex polygonal cells. An L-mosaic (Pielou, 1977) is obtained by independently colouring the cells black $(Z(\underset{\sim}{x})=1)$ or white $(Z(\underset{\sim}{x})=0)$ with probabilities μ and $1-\mu$ respectively. The resulting correlation function is $\rho(t)=\exp(-\lambda t)$. Note that $\rho(t)$ is independent of μ, and that the two phases are treated symmetrically.

Figure (2) shows a partial realisation on the unit square with $\mu = 0.5$ and $\lambda = 100$.

Figure (2). Partial realisation on the unit square of an L-mosaic with $\mu = 0.5$ and $\lambda = 100$



(2.3) The C-mosaic

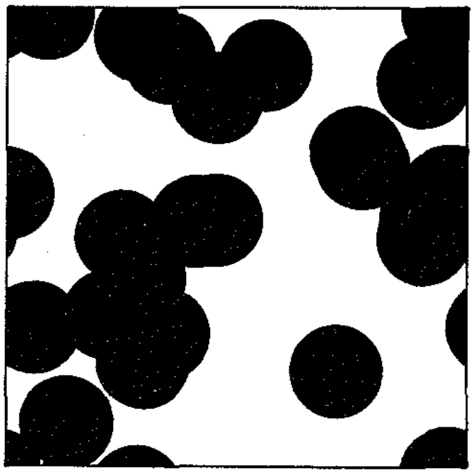
A second model is one in which the set S is the union of countably many closed discs with centres determined by a Poisson process on $\mathbb{R}^2$, of intensity λ, and radii mutually independent and identically distributed according to the distribution function $F(\cdot)$. Expressions for μ and $\rho(t)$ are available, and are particularly simple for the special case of constant radius δ. This gives $\mu = 1 - \exp(-\pi\lambda\delta^2)$ and

$$\rho(t) = \left[\exp\{\lambda V_\delta(t)\} - 1\right]/\{\exp(\pi\lambda\delta^2) - 1\},$$

where $V_\delta(t)$ is the area of intersection of two discs with common radius δ and centres a distance t apart. Note that $\rho(t) = 0$ for $t \geq 2\delta$.

Various applications of this model are described by Matérn (1960, Ch.3), Marchant & Dillon (1961), Roach (1968), Dupac (1980) and Diggle (1981). We shall call the model a C-mosaic (C for circle) although this is not standard terminology. Figure (3) shows a partial realisation on the unit square with $\mu = 0.5$ and $\delta = 0.1$.

Figure (3). Partial realisation on the unit square of a fixed radius C-mosaic with $\mu = 0.5$ and $\delta = 0.1$.



3. Point sampling designs to estimate cover

(3.1) General considerations

We consider three classes of design for sample points $\underset{\sim}{x}_i : i=1,\ldots,n$ in a region A. In random sampling, the $\underset{\sim}{x}_i$ are independent realisations from the uniform distribution on A. In frame sampling, the $\underset{\sim}{x}_i$ are grouped into sets of m points regularly spaced along a line (linear frames) or of $m \times m$ points in a regular square lattice arrangement (square frames). In either case, frame locations are determined by random sampling. A single linear frame, with m large, is sometimes called a line transect. Finally, in systematic sampling the $\underset{\sim}{x}_i$ form a regular lattice over the whole of A, with the starting point $\underset{\sim}{x}_1$ determined by random sampling.

We define the cover p to the areal proportion $p = |S \cap A|/|A|$. For any of the above designs, the sample mean $\bar{z} = n^{-1} \sum_{i=1}^{n} z(\underset{\sim}{x}_i)$ is unbiased for p with respect to repeated sampling of a single realisation, and unbiased for $\mu = E[Z(\underset{\sim}{x})]$ with respect to repeated realisations of the mosaic process.

The distinction between p and μ is important in the case of systematic sampling when, as the lattice spacing $d \to 0$, $\mathrm{Var}(\bar{z}|p) \to 0$

but

$$\mathrm{Var}(\bar{z}) \to \mathrm{Var}(p) = \mu(1-\mu) \ |A|^{-2} \int\int_{A\times A} \rho(\|\underset{\sim}{x}-\underset{\sim}{y}\|)\, d\underset{\sim}{x}\, d\underset{\sim}{y},$$

which is substantially greater than zero unless $|A|$ is large. In frame sampling or random sampling, we shall assume that observations from different frames or points, respectively, are uncorrelated; this will be approximately so if the distances between frames or points are large, i.e. the sampling is sparse. These designs are often used in vegetation surveys to estimate the cover over a relatively large region A. In such cases, the distinction between p and μ is unimportant, and the assumption of sparse sampling is reasonable.

In the remainder of this section we investigate the efficiency of the three types of sampling design for the estimation of p, in terms of the <u>variance per sample point</u>, $v = n\mathrm{Var}(\bar{z}|p)$. In random sampling, we are simply conducting a sequence of Bernoulli trials with parameter $p \simeq \mu$ and

$$v_r = \mu(1-\mu).$$

For linear frames of m points with spacing d, we need to take account of the correlation within frames, and obtain

$$v_f = \mu(1-\mu)\{1 + 2 \sum_{i=1}^{m-1} (1 - i/m)\rho(di)\}. \qquad (3.1)$$

A similar expression can be written down for square frames. Finally, for systematic sampling we consider a lattice generated by a rectangle with side length $d\alpha^{1/2}$ and $d\alpha^{-\frac{1}{2}}$. Then, the limiting variance per sample point as $A \to \mathbb{R}^2$ is

$$v_s = \mu(1-\mu)\ [\sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \rho\{d\sqrt{(i^2\alpha + j^2\alpha^{-1})}\} - 2\pi\int_0^{\infty} \rho(rd)dr] \quad (3.2)$$

(c.f. Matérn, 1960, Ch.5).

(3.2) <u>Systematic sampling</u>

Table (1) shows the relative variance per sample point, $w = v_s/v_r$, for fixed radius C-mosaics with various values of the mosaic parameters μ and δ, and one sample point per unit area, $d=1$. A square lattice ($\alpha=1$) is always more efficient than random sampling, with progressive gains in efficiency as δ increases and the mosaic becomes more "coarse-grained"; equivalently, since the unit of measurement is arbitrary, for a given C-mosaic the relative efficiency of systematic to random sampling increases with the number of sample points per unit area. For a rectangular lattice, systematic sampling may be less efficient than random sampling, a phenomenon noted also by Matérn (1960, Ch.5) who used an exponential correlation function corresponding to an L-mosaic model. Matérn suggests that rectangular lattices may be economically efficient, but does not declare an explicit cost function.

Certain anomalies occur in Table (1) because of the fixed radius in the C-mosaic model. Generally, systematic sampling becomes less efficient as α increases. However, for $\delta = 0.25$ the efficiency is constant for $1 \leq \alpha \leq 4$ because over this range a disc can be touched by at most one sample point. Moreover, the efficiency can increase with α: for $\delta = 0.75$ a disc can be touched by four sample points if $\alpha=1$ but by at most three points if $\alpha=2$.

The parameter α has an interesting alternative interpretation in terms of anisotropic mosaic processes (Matérn, 1960). For example, the C-mosaic can be extended to include a Poisson field of ellipses with axes of length $\alpha^{\frac{1}{2}}\delta$ and $\alpha^{-\frac{1}{2}}\delta$, and fixed orientation. Sampling this process using a square lattice aligned with the orientation of the ellipses is equivalent to sampling an isotropic C-mosaic using a rectangular lattice. For anisotropic mosaics square lattice systematic sampling is therefore not always more efficient than random sampling (Table 1).

Table (1): Relative variance per sample point of systematic and random sampling for fixed radius C-mosaics with one sample point per unit area.

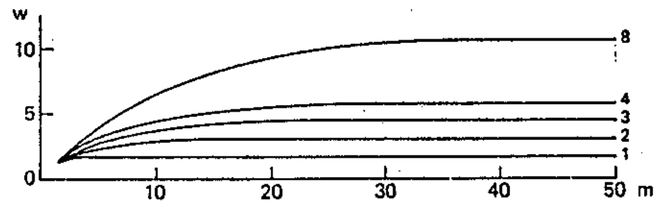
		δ				
α	μ	0.25	0.50	0.75	1.00	2.00
	0.1	0.809	0.237	0.185	0.157	0.085
1	0.5	839	356	252	206	107
	0.9	905	620	454	360	187
	0.1	0.809	0.585	0.165	0.161	0.134
2	0.5	839	624	277	229	154
	0.9	905	735	528	424	241
	0.1	0.809	0.994	0.845	0.327	0.174
4	0.5	839	978	876	472	241
	0.9	905	944	920	746	438
	0.1	1.566	2.617	3.313	3.635	1.128
16	0.5	1.462	2.404	3.067	3.419	1.632
	0.9	1.230	1.886	2.434	2.813	2.486

(3.3) Frame sampling

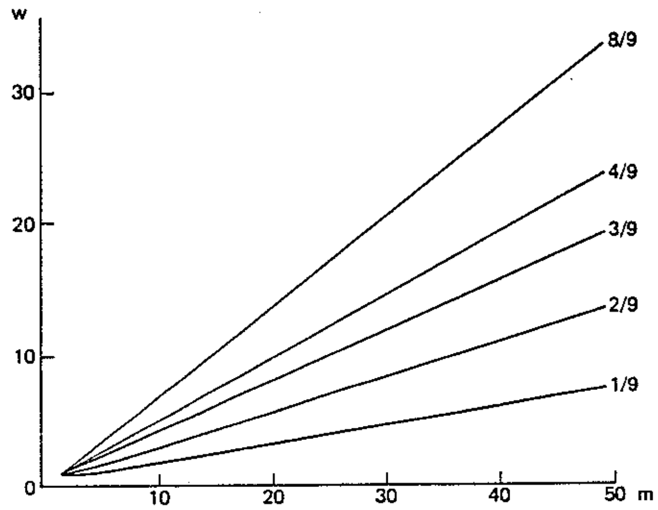
Figure (4) shows $w = v_f/v_r$ as a function of m, the number of points in a linear frame, for fixed radius C-mosaics with $\mu = 0.5$ and various values of δ. In Figure (4a), the spacing between successive points in the frame is $d = 1$, for all m. After an initially rapid increase in w, the curves flatten out as the longer frames dilute the effect of the correlation within frames. If the frame is a physical instrument such as the point frame sometimes used in vegetation analysis (Goodall, 1952), it might be more natural to keep the total frame length $\ell = (m-1)d$ fixed. Figure (4b) shows that in this case w increases more sharply with m, which is to be expected because d decreases. The values of δ illustrated in Figure (4a) and (4b) are chosen so that corresponding pairs represent the same mosaic process if the recording instrument is a ten-point frame.

Figure (4). Relative variance per sampling point of frame and random sampling for fixed radius C-mosaics.

(a) fixed spacing $d = 1$, $\delta = 1, 2, 3, 4, 8$.



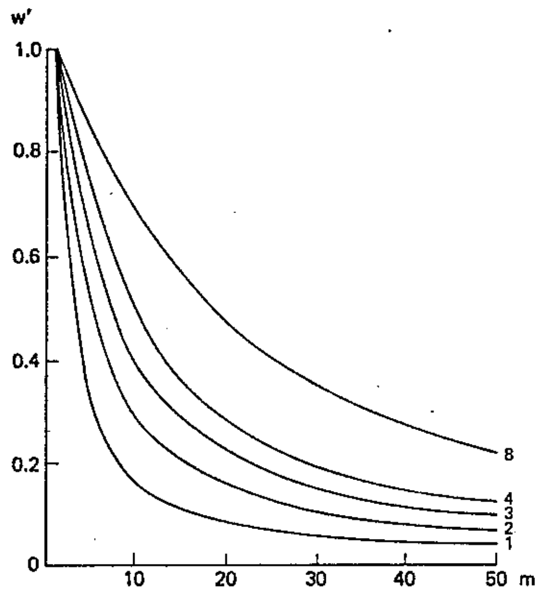
(b) fixed frame length $\ell = 1$, $\delta = 1/9, 2/9, 3/9, 4/9, 8/9$.



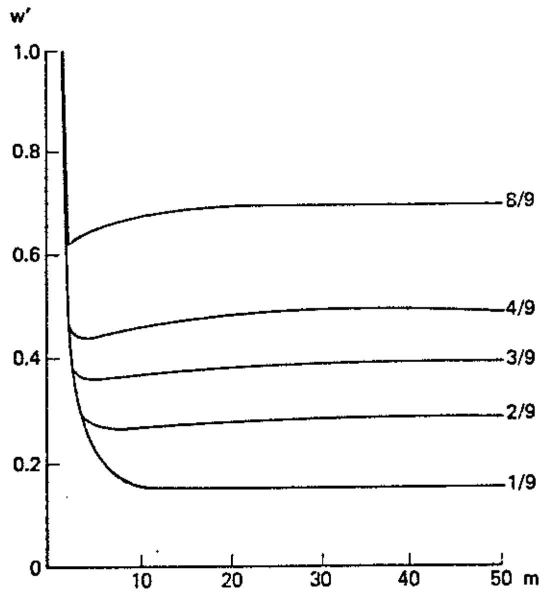
If recording is cheap, but travelling between random locations is expensive, an appealing measure of efficiency is $w' = w/m$, which measures the variance using n frames containing m points each, relative to the variance using n single points. Figure (5a) shows how, for a fixed spacing $d = 1$ of points within a frame, w' decreases steadily with m. In the case of fixed frame length, Figure (5b) shows how w' initially decreases sharply, but is thereafter fairly constant. One surprising feature in the case of fixed frame length is that w/m does not always decrease monotonically,

Figure (5). Relative variance per frame for fixed radius C-mosaics.

(a) fixed spacing, $d = 1$, $\delta = 1, 2, 3, 4, 8$.



(b) fixed length, $\ell = 1$, $\delta = 1/9, 2/9, 3/9, 4/9, 8/9$.



which implies that taking additional observations actually increases the variance of the estimator. The explanation for this is that the sample mean is not a fully efficient estimator because it discards information provided by the spatial arrangement of points within each frame. Rothery (1974) observes a similar phenomenon for the exponential correlation function associated with an L-mosaic. Alternative estimators, for example the maximum likelihood estimator for μ, use the spatial information but are impractical for routine use.

(3.4) Optimum frame size for a particular cost function

In practical terms, any assessment of the relative efficiency of frame sampling and random sampling must involve cost considerations. For n points in f linear frames of m points each, Rothery (1974) uses the cost function

$$C = an + bf = n(a + bm^{-1}), \tag{3.3}$$

where a represents the cost of analysing a single point and b the cost of locating a point or frame at random within A. Note that this ignores the cost of moving from point to point within a frame. With this cost function and $b = 1$, the relative variance of frame sampling to random sampling for fixed total cost is $w^* = v_f(a+m^{-1})/\{v_r(a+1)\}$, where v_f is given by (3.1) and depends on both m and d, the spacing of successive points within a frame. The optimal value of m depends on a, d and on the C-mosaic parameters μ and δ.

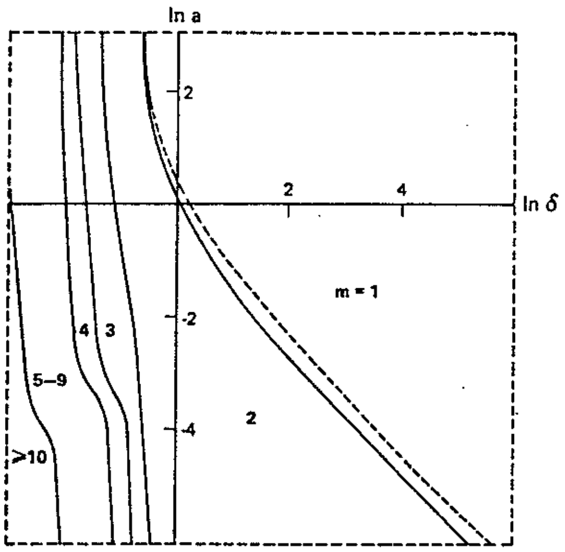
We consider first a frame of fixed length with spacing $d = (m-1)^{-1}$ between successive points. Figure 6 sketches the regions in the (δ,a)-space corresponding to different optimal values of m when $\mu = 0.5$; different values of μ give similar pictures. The exact boundaries between the different values of m show minor irregularities which may again be attributable to the cut-off in the correlation function of the fixed radius C-mosaic.

For a frame with fixed spacing $d = 1$ between successive points, we conjecture that the optimal value of m is either 1 (random sampling) or ∞ (line transect sampling). The boundary between the two cases is included in Figure (6), again with $\mu = 0.5$. The basis of our conjecture is that for given values of a, μ and δ the problem can be solved by finite enumeration, and we have not found any combination of parameter values which gives an optimal result other than $m = 1$ or ∞.

Figure (6). Optimal number m of points per frame for C-mosaics with $\mu = 0.5$.

______ boundaries between different values of m for frames of fixed length $\ell = 1$.

------ boundary between $m = 1$ and $m = \infty$ for frame with fixed spacing $d = 1$.



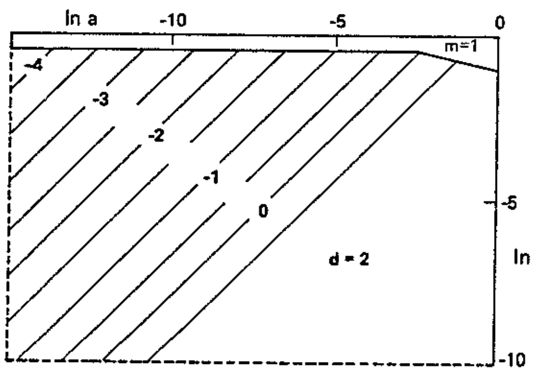
The relative variance of line transect to random sampling can be deduced from Figure (6). For given δ, let a_δ denote the value of a for which $m = 1$ and ∞ are equally efficient for fixed total cost. Then, for any other value of a the relative variance for fixed total cost is $w^* = (1+a_\delta^{-1})/(1+a^{-1})$.

The optimal spacing between successive points can be found when a cost c for moving from point to point per unit distance within a frame is incorporated in the cost function to give

$$C = an + bf + cd(m-1)f = n\{(a+cd) + (b-cd)/m\}. \tag{3.4}$$

Because the cost function (3.4) is of the same form as (3.3), the optimal value of m for fixed d is either 1 or ∞. The optimal spacing d_0 depends only on the ratio a/c. For $d < 2$, $\ln(d_0)$ is approximately linearly related to $\ln(a/c)$ as sketched in Figure (7) for $\delta = b = 1$ and $\mu = 0.5$; as in Figure (6) minor irregularities occur but are not shown in the sketch. Because different points within a frame generate independent observations whenever $d \geq 2$, $d > 2$ can never be optimal. If the cost of moving from point to point within a frame is high, random sampling is more efficient than line transect sampling as indicated in Figure (7).

Figure (7). Contours of $\ln(d_0)$ for line transect sampling of a C-mosaic with $\mu = 0.5$ and $\delta = 1$.



(3.5) <u>An example</u>

As an illustration, we used random sampling and frame sampling to estimate the cover of heather in Figure (1). As a first approximation, a fixed radius C-mosaic is a plausible model (Gimingham, 1972). Rough <u>a priori</u> guesses for the mosaic parameters were $\mu = 0.5$, $\delta = 40$cm. Each dot in Figure (1) represents an area 10cm X 10cm, so a convenient spacing of points within a frame is 10cm. In the cost function (3.3) with $b = 1$, a reasonable value for a was 0.05. Using Figure (6) with $\ln(\delta) = \ln(40/10) = 1.4$ and $\ln(a) = -3$ we deduce that line transect sampling would be much more cost-effective than single points. A sample of 17 line transects with 50 points each gave an estimated cover of 0.5035, with an empirical standard error of 0.0490, compared with the true cover value 0.4974. A sample of 50 single points took approximately the same length of time and gave an estimate 0.4912 with empirical standard error 0.0663. Note that the estimates and standard errors make no assumptions about the underlying mosaic process, which is needed only to deduce the optimal sampling design. One advantage of using a qualitatively plausible process such as the C-mosaic is that guessing appropriate parameter values becomes a reasonable proposition.

4. Discussion

The results in this paper are based on an assumed stationary, isotropic process. Rothery (1974) adopts a similar viewpoint, whereby observations within a frame are correlated, but the cover μ is constant between frames. In contrast, Kemp & Kemp (1956) propose a mixed binomial model for frame sampling, which assumes that observations within a frame are independent but that μ varies randomly between frames.

Properties of stationarity and isotropy can always be induced into the statistical problem by random location and orientation of frames, but this results in an unnecessary loss of information. If the sampling region is thought to be heterogeneous, stratification can be used to give a more precise estimate of the overall cover together with information about both large-scale (between strata) and small-scale (within frames) spatial variation(Goodall, 1952; Matérn, 1960). An alternative to stratification, which achieves the same aims, is to locate frames systematically. Using square frames with a fixed orientation allows also for the analysis of possible directional effects.

The cost analysis in §(3.4) assumes that the only objective is to estimate cover. However, in many cases the sampling design should also cater for estimation of the standard error and, with future sampling in mind, for an assessment of factors which influence efficiency. Such an assessment requires checks on stationarity, on isotropy, on the goodness-of-fit of the assumed mosaic model and on the cost function. Moreover, a second objective may be to model the mosaic pattern. As the emphasis switches from estimation of cover to description of spatial pattern, so the optimum sampling design switches from random points or linear frames to square frames or, with a reduction in the target region, to systematic sampling or complete mapping.

References

Bartlett, M.S. (1964). A note on spatial pattern. Biometrics, **20**, 891-2.

Diggle, P.J. (1981). Binary mosaics and the spatial pattern of heather. Biometrics, **37**, 531-9.

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

Gimingham, C.H. (1972). Ecology of Heathlands. London: Chapman & Hall.

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

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

Marchant, J.C. & Dillon, P.L.P. (1961). Correlation between random-dot samples and the photographic emulsion. J. Opt. Soc. Amer. **51**, 641-4.

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

Matérn, B. (1979). The analysis of ecological maps as mosaics. In Spatial and Temporal Analysis in Ecology eds. R.M. Cormack & J.K. Ord, 271-88. Fairland: International Co-operative Publishing House.

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

Pielou, E.C. (1964). The spatial patterns of two-phase patchworks of vegetation. Biometrics, **20**, 156-67.

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

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

Rothery, P. (1974). The number of pins in a point quadrat frame. J. Appl. Ecol., **11**, 745-754.

ter Braak, C.J.F. (1980). Binary mosaics and point quadrat sampling in ecology. M.Sc. thesis. University of Newcastle, U.K. Report A 80 ST 96 37 of IWIS-TNO, P.O. Box 100, 6700 AC Wageningen, the Netherlands.

Stoyan, D. (1979). Applied stochastic geometry: a survey. Biom. J., **21**, 693-713.

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