PROBABLE LIMIT OF ERROR OF THE POINT DISTANCE-NEIGHBOUR DISTANCE ESTIMATE OF DENSITY

C. L. BATCHELER

Protection Forestry Division, Forest Research Institute, Rangiora

SUMMARY: A formula is proposed for calculating the probable limit of error (PLE) of the density estimate which is obtained from measurements of distance from sample points to the nearest member, its nearest neighbour, and its nearest neighbour.

It takes the form:

$$PLE = t\bar{A}D/\sqrt{N},$$

(where t is Student's t, $\bar{A}$ is a measure of non-randomness, D is the estimate of density, and N is the number of sample points) and is considered to be analogous to the usual bounded plot sampling error formula $tS/\sqrt{N}$ for the specific case where density averages one per plot. Using Student's t for the 95 percent level of probability, 90 percent of the density estimates, drawn from 40 computer-simulated populations were within the ranges given by the proposed formula. Similarly, 92 percent of 36 estimates obtained from 11 paper-dot populations and 76 percent of 93 estimates obtained from 25 natural populations were within the proposed PLE. In some of the remaining cases, the extent by which the proposed PLE's were less than the calculated difference between true density and D were quite trivial. Some anomalies probably arose from sampling difficulties.

INTRODUCTION

In an earlier paper (Batcheler, 1973) formulae are given for calculating population density from distances measured up to a chosen maximum distance R, or without any limit (in which case R is infinity), from each of N sample points (Fig. 1). Defining the nearest member to the point as I_p, its nearest neighbour as I_n, and I_n's nearest neighbour (exclusive of I_p) as I_m, the distances measured are as follows:—

r_p, if I_p is R or less away from the sample point, the distance from the point to $I_p(r_p)$ is measured, r_p^2 is the square of r_p, and p is the number of these measurements;

r_n, if r_p is R or less, the distance from I_p to I_n is measured provided it is also R or less, and n is the number of these measurements;

r_m, if r_n is R or less, the distance from I_n to $I_m(r_m)$ is measured provided it is also R or less, and m is the number of these measurements.

These data are used to calculate the following statistics:—

$$f = p/N \quad \ldots (1)$$

$$d = p/[\pi(\Sigma r_p^2 + (N - p)R^2)] \quad \ldots (2)$$

$$A_1 = \frac{1}{E(CV)} \sqrt{[p\,\Sigma r_p^2 - (\Sigma r_p)^2]n^2N/\,\Sigma r_p\,\Sigma r_n p^3} \quad \ldots (3)$$

$$A_2 = \frac{1}{E(CV)} \sqrt{[p\,\Sigma r_p^2 - (\Sigma r_p)^2]m^2N/\,\Sigma r_p\,\Sigma r_m p^2 n} \quad \ldots (4)$$

in which A_1 and A_2 are two indices of non-randomness, and E(CV) is the expected coefficient of variation of r_p (Batcheler, 1973, Table 2, or from $\log_e E(CV) = -1.0319 + 0.4892f^2 - 0.7182f^4 + 0.6095f^6$),

$$a = (1 + 2.473f) \quad \ldots (5)$$

$$b = (1 + 2.717f) \quad \ldots (6)$$

$$D = \frac{d}{2a}(b^{A1} + b^{A2}) \quad \ldots (7)$$

where D is the required estimate of density.*

It was also shown in the 1973 paper that the geometric mean of A_1 and $A_2(A_g)$ is a linear function of the standard deviation (S) of an estimate derived from bounded plots. Using plots which were large enough to include an average of four members, the relationship was $CV_{plots} = 50A_g$. Subsequently, I have found that A_1 and A_2 and hence A_g and the arithmetic mean ($\bar{A}$) are very similar, except when populations are extremely

*This formula for the estimate of D was incorrectly given as $\overline{d/D} = \frac{a}{2}(b^{-A1} + b^{-A2})$ in Batcheler (1973). An Erratum note appears elsewhere in this issue.

aggregated. Therefore the simpler $\overline{A}$ will be considered throughout this paper as the measure of standard deviation.

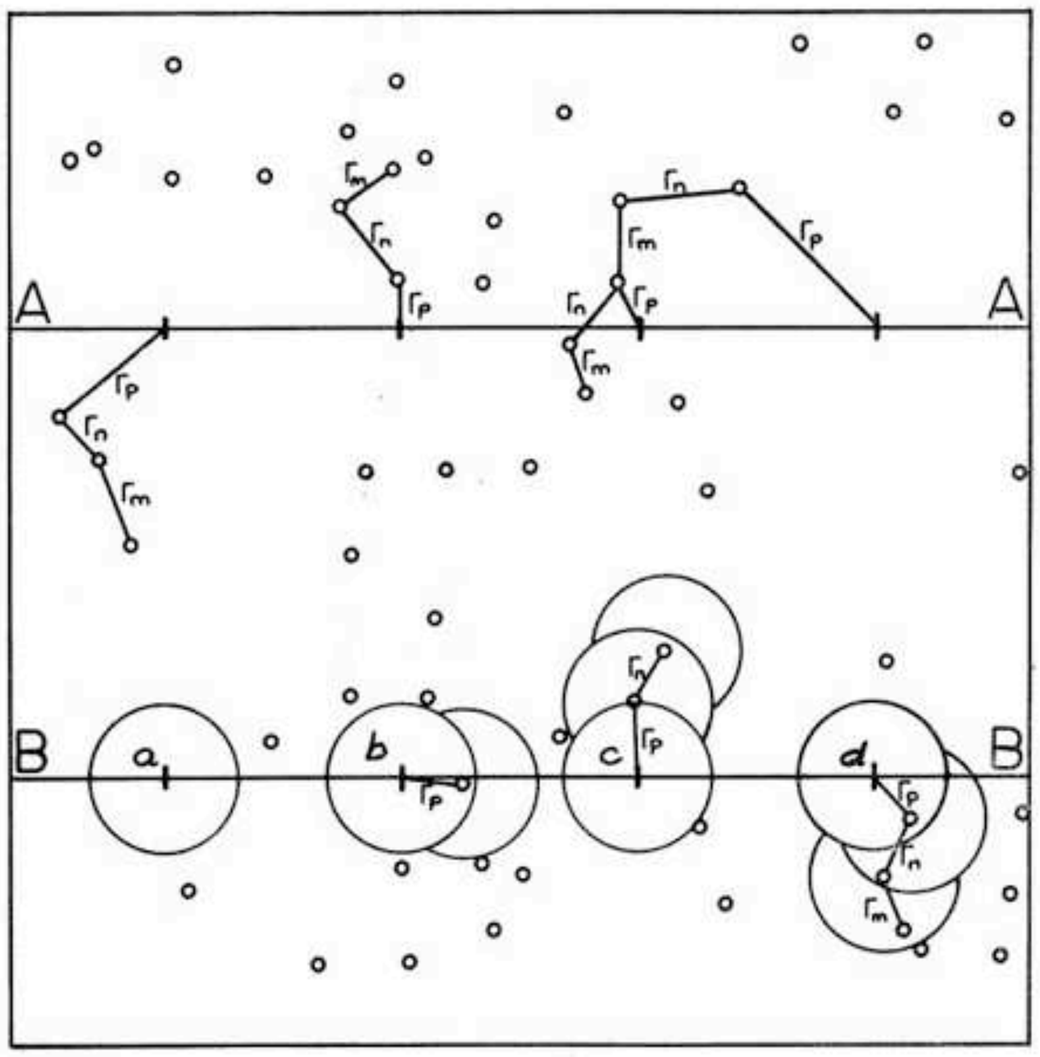

FIGURE 1. *Diagram of sampling schemes when no limit R is imposed (line A), and when a limit is imposed (line B). For line B the four possible events at a point are illustrated: (a) nearest member beyond R—no distances measured; (b) $r_p \leqslant R$, therefore measured; (c) r_p and $r_n \leqslant R$, therefore measured; (d) r_p, r_n and r_m all measured $\leqslant R$. In A, $N = p = n = m = 4$. In B, $N = 4$, $p = 3$, $n = 2$ and $m = 1$.*

Despite the obvious linear relationship between S and $\overline{A}$, I did not realise until recently that this relationship might offer a simple way of calculating a probable limit of error of D. Since the corrected point distance technique (CPD) intrinsically bases the density estimate on the distance to the nearest member (and uses the r_n and r_m data to correct bias), the sample obtained is analogous to bounded plots of unit size. Where bounded plots of unit size are sampled, $S^2/\bar{x} = 1$ under Poisson assumptions, and S will also equal 1. Similarly, in CPD sampling, $\overline{A}$ has been found to nearly equal 1 (actually 0.948) when the population is random so that D is 1 and unbiased. Therefore, since D is unit density, $\overline{A}D/\sqrt{N}$ can be considered as an empirical analogue of the standard error for a sample of plots, and at a specified level, the probable limit of its error (PLE) will be approximately

$$PLE = tAD/\sqrt{N}$$

where t is "Student's" t.

Empirical testing of this hypothesis is the subject of this paper.

TEST DATA

The data include the results drawn from analyses of 40 computer-simulated populations, 11 paper-dot populations, and 25 field experiments described in the earlier paper (Batcheler, 1973) in connection with developing and testing formulae for calculating D. The only feature of them which warrants repetition here is to emphasise the consequences of estimating "true density" as a bench-mark for evaluating the accuracy of D and its PLE. Even in the case of the simulated populations, estimates of the density parameter ($\pm$ probable sampling error) were unavoidable because, particularly in strongly aggregated populations, distance sampling is severely distorted when sample points fall closer to the edge of the population map than to the nearest population member or its neighbours. "True densities" were therefore estimated by counting in plots located inside a line approximately half-way between the edges and the nearest members.

Similar but minor problems arose in estimating "true density" of some paper dot populations. In the case of the natural populations, most "true densities" were subject to an estimate of sampling error because it was not feasible to count the total population. Typically, these errors were less than $\pm$10 percent in uniform populations such as pine plantations, and ranged up to $\pm$42 percent for estimates of aggregated animal faecal pellets (Batcheler, 1973 pp. 140–141).

RESULTS

Simulated populations.

The results of tests where limits are not imposed on the distance searched from sample points (R) or from one neighbour to another are shown in Figure 2. They illustrate the main features of the pattern found throughout this study.

Referring first to the confidence limits ($P = 0.95$) of the plot estimates (thick vertical bars, Fig. 2), it is clear that the probable limit of error of "true density" ranged from $\pm$5 to 10 percent for the relatively uniform—random populations (left side) up to $\pm$20 to 25 percent for the more severely aggregated populations (right side). Similarly, the proposed PLE's broaden significantly across the spectrum from uniform ($\overline{A} < 1$) to severely aggregated ($\overline{A} > 3$) populations.

Among the 15 samples which were uniform ($\overline{A} = 0.4$–0.6), or tended from uniform towards randomness ($\overline{A} = 0.95$) the PLE's of 14 encompass "true density". Of 19 samples which ranged from slightly aggregated to severely aggregated ($\overline{A}$ ranging from 1 to 3) 18 PLEs'

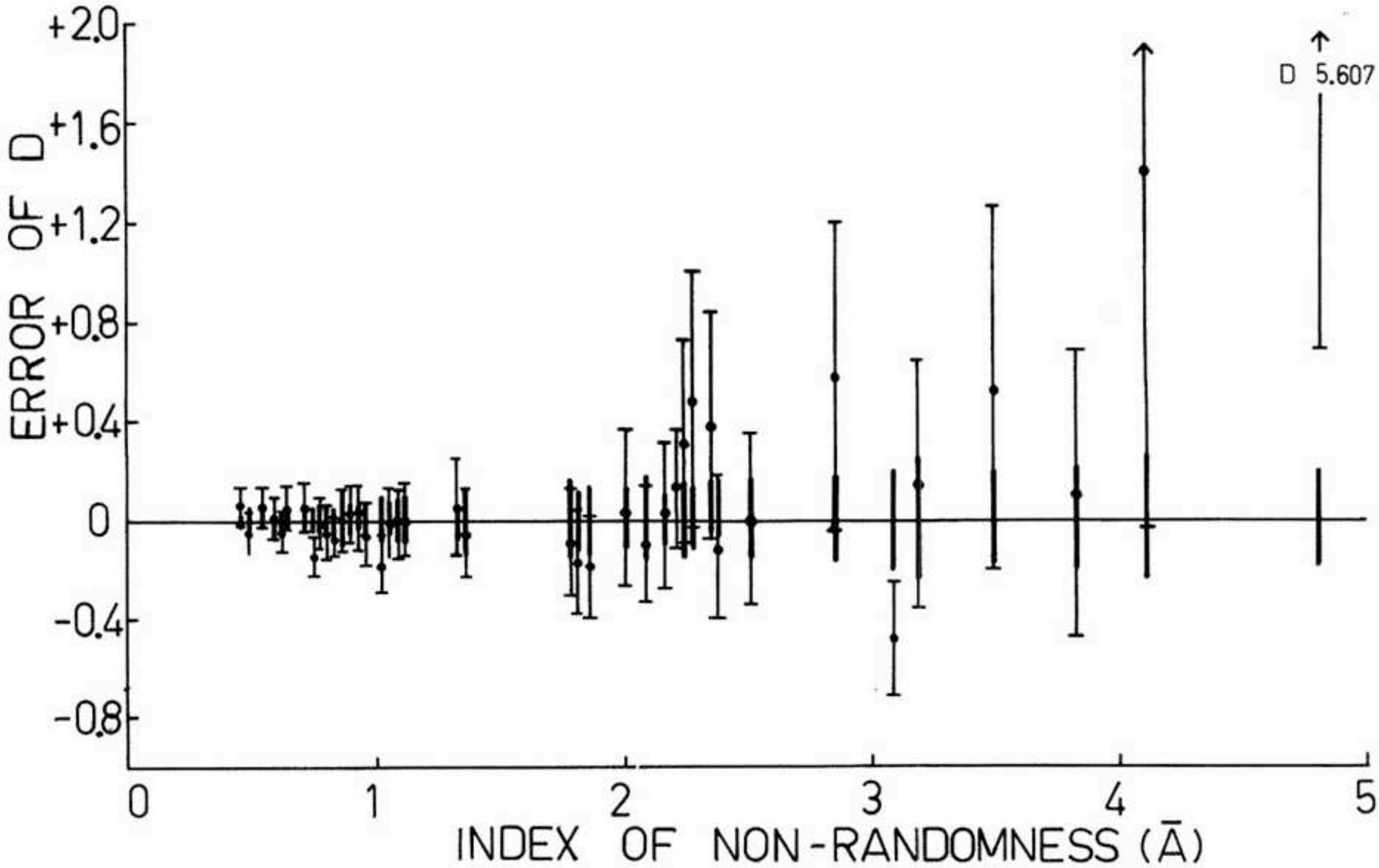

FIGURE 2. *The pattern for 40 simulated populations without limit on R, of the proposed PLE (vertical thin bars) plotted on the deviation of D from "true density" (ordinate) and the index of non-randomness (Ā, abscissa). The probable limits of "true density" are indicated by the shorter, thick bars. Generally, D is less accurate, and PLE is wider, as the index of aggregation increases.*

encompass "true density". Beyond this, in extremely aggregated populations ($\bar{A} > 3$) four PLE's encompass "true density", but two did not. Over all 40 populations, 90 percent of the PLE's overlapped with "true density".

Paper-dot and natural populations.

The results of 129 estimates of PLE for truncated and unrestricted measurements are given in Figure 3. Truncation of the measurements was achieved by listing the r_p (and associated r_n and r_m) in order from smallest to largest, cutting them into blocks at progressively larger R and accumulating the sums up to the maximum measured. Samples were rejected if the imposition of a small R resulted in exclusion of all the nearest neighbour measurements (i.e. n, or more frequently m, were zero), because in these situations estimates of A_1 and A_2 cannot be made.

The left section of Figure 3 is comprised of 38 values calculated for frequencies (formula 1) less than 0.5; the centre section gives 55 results corresponding to frequencies between 0.5 and 0.99; the right section gives the 36 estimates with no R limit imposed upon measurements. Those for paper-dot populations are shown as open circles and those for natural populations are solid dots and crosses (see below).

Ninety-two percent of the density estimates for the paper-dot populations were within $\pm$PLE (eight of 10 at $f < 0.5$, 15 of 15 at $f = 0.5$–0.99, and 10 of 11 at $f = 1.0$). However, in the natural population samples, only 76 percent of 93 estimates were within the range of PLE. Successful prediction of these errors declined from 87 percent (of 28 estimates) when $f < 0.5$, to 75 percent (of 40 estimates) with f between 0.5 and 0.99, and to 72 percent at $f = 1.0$.

The excessive proportion of failures of PLE to embrace "true density" of the natural populations was largely attributable to the results from three estimates of hare (*Lepus europaeus*) faecal pellet density, and one small experiment in beech (*Nothofagus*) forest. These are indicated in Figure 3 by crosses. The problem with the

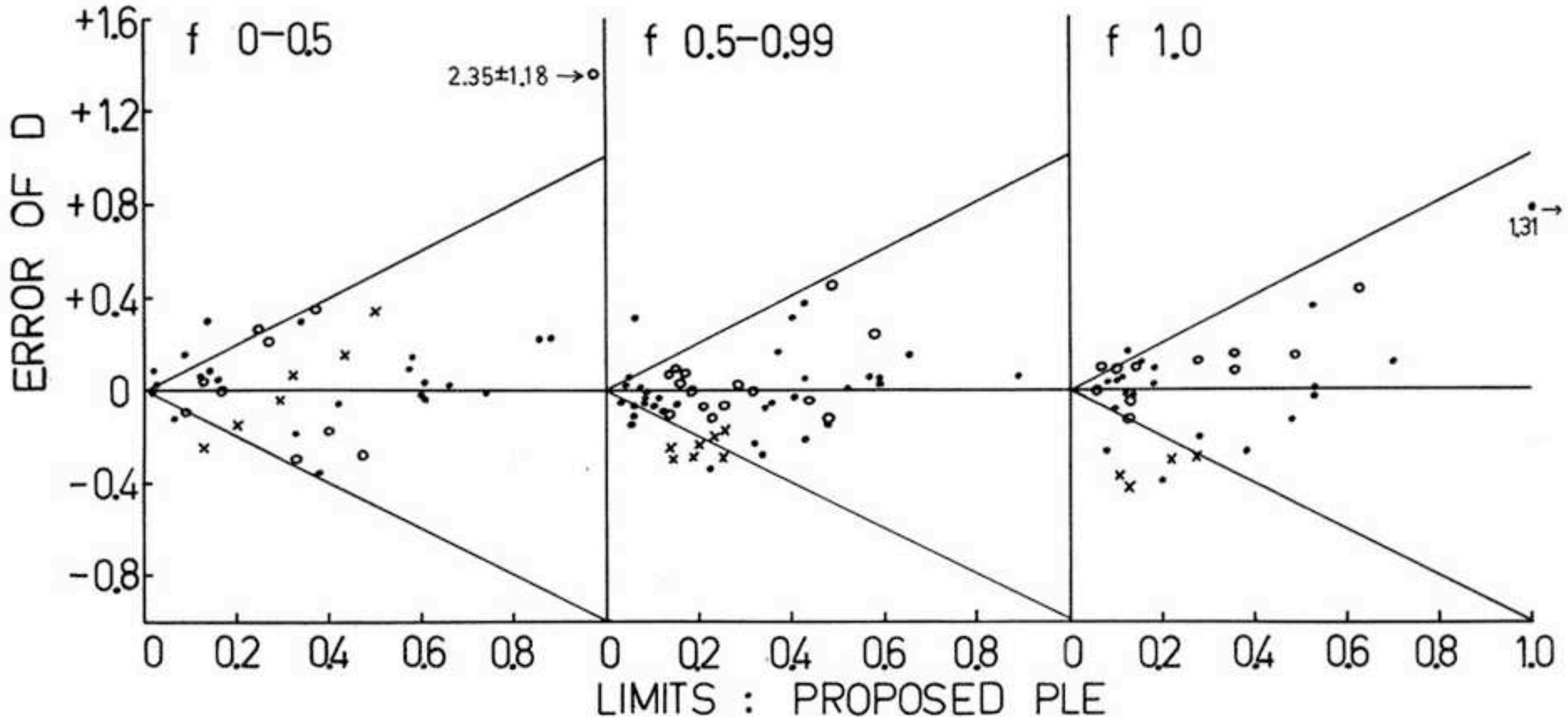

FIGURE 3. *PLE's and differences between "true density" and D of the paper-dot and natural populations, grouped in three ranges of f. Points outside the wedge bounded by the diagonal lines are those where the differences between D and "true density" exceeds PLE. Open circles refer to paper-dot populations. Dots denote the natural field populations, except those of hare faecal pellets and one beech forest which gave unacceptable results (see text).*

hare pellet experiments arose because A_2 was not estimated, and this is now known to invariably generate unacceptably low estimates at large f in populations in which clumps are distributed within clumps (Fig. 3) see also comments p. 145, Batcheler, 1973). In the beech forest experiment, the total population consisted of only 261 trees. Many measurements were made from successive sample points to the same trees and their neighbours. This gives poor estimates of d and $\bar{A}$. Deleting these four sets of data leaves 84 per cent of the remaining PLE's overlapping with "true density".

DISCUSSION

The proposed formula failed to give estimates which embraced the differences between D and "true density" in 17 per cent of the tests rather than the five percent implicit in the use of Student's t for the 95 percent level of probability.

Experimental problems were probably the major cause of this discrepancy. The populations were severely restricted in size, which must influence the normal distribution theory of sampling infinitely large populations. "True density", the value against which D and PLE are evaluated, cannot be determined exactly because of edge effects and sampling errors. Third, D and PLE must be presumed to contain some element of error or bias arising from the nature of the data used in the development of estimating formulae.

Obviously, the tests with data from the simulated populations are prone to all these errors, and suspect as circular evidence because they are tests against the same information from which the estimating formulae are derived.

Nevertheless, the results from the simulated populations adequately show the general behaviour of the proposed formula. When R is not limited, the PLE range indicated for uniform populations is less than ± 10 percent, and failed to encompass the "true density", by a trivial percentage, in only a few cases (Fig. 2). Substantial anomalies arose only in extremely aggregated populations.

Similarly, too many estimates were outside the expected range in the calculations for the paper-dot and natural populations to meet the implied 95 percent specification, but the discrepancy is minor. Ninety-two percent of the errors of the paper-dot populations were within the proposed PLE. Seventy-six percent of natural population estimates or, deleting those from the three hare pellet and the one beech forest experiments, 84 percent of PLE's enclosed "true density". These results seem quite reasonable, particularly when taken in conjunction with the probable errors implicit in the "true density" values.

It is therefore concluded that PLE is a realistic esti-

mate of the probable error of density of populations, provided estimates of total aggregation (A_1 and A_2) are made, and provided few samples (say, 5 percent) are comprised of repeated distance measurements to the same members and their sequential neighbours.

Assuming this, a general table of PLE for different N and $\bar{A}$ can be given to indicate the behaviour of the formula when no limit (R) is imposed on distances (Table 1). Errors will be rather larger when imposition of a small R results in few measurements being made. The table indicates that PLE less than $\pm$10 percent of D requires 100 to 400 samples as dispersion tends from uniform towards random, and rises to 1,000 or more as $\bar{A}$ tends towards extreme values which have already been encountered in natural populations such as clumped *Celmisia* and rabbit faecal pellets.

TABLE 1. *Table for $t\bar{A}/\sqrt{N}$ for selected values of N and $\bar{A}$, and for the error of random population samples calculated from distribution of χ^2 ($P = 0.95$). The values under $\bar{A} = 1$ are for (nearly) random populations. Compare with those from χ^2 (right column).*

N	Index of non-randomness ($\bar{A}$)							PLE for random population from χ^2	
	.5	.75	1.0	1.5	2.0	2.5	3.0		
10	.36	.54	.72	1.07	1.43	1.79	2.15	−.52	+.71
20	.23	.35	.47	.70	.94	1.17	1.40	−.39	+.48
40	.16	.24	.32	.48	.64	.80	.96	−.29	+.33
80	.11	.17	.22	.33	.45	.56	.67	−.21	+.23
120	.09	.14	.18	.27	.36	.45	.54	−.17	+.18
200	.07	.10	.14	.21	.28	.25	.42	±.14	
300	.06	.08	.11	.17	.23	.28	.35	±.11	
400	.05	.07	.10	.15	.20	.25	.29	±.10	
500	.04	.07	.09	.13	.18	.22	.26	±.09	
1000	.03	.05	.06	.09	.12	.15	.19	±.06	
2000	.02	.03	.04	.07	.09	.11	.13	±.04	

Though such numbers appear daunting, they compare favourably with the necessary magnitude of plot sampling in extremely aggregated populations. The formula $N = S^2t^2/(\text{specified error})^2$ indicates that in the *Celmisia* population mentioned, variance of the counts in the 100 3.3 m² plots was such that if the result is applied to sampling an infinitely large population, 840 plots would be necessary to estimate the mean within $\pm$10 percent at 95 percent probability. For counting rabbit pellets, the data from the two experiments listed in Batcheler (1973) indicate that approximately 430 plots of 0.09 m² would have been necessary to obtain the above-mentioned precision.

The proposed formula gives very similar estimates of variance and probable error of random populations to those suggested by other authors for shortest distance sampling. Morisita (1957) has shown that variance = density²/(N − 2) from which it follows that, because density per sample unit is invariably 1, $S/\sqrt{N-2}$ is the standard error and $tS/\sqrt{N-2}$ is the corresponding limit of probable error. Therefore, since $\bar{A}$ is very nearly 1 in the CPD method, $t\bar{A}D/\sqrt{N}$ is very nearly the same expression as given by Morisita if N is large (say, >50).

Skellam (1952), Thompson (1956) and Kendal and Moran (1963) have shown that r_p is distributed as χ^2, and for random populations Kendal and Moran give the formula for the upper (u) and lower (l) limits of D at given probability as

$$D_{u \text{ or } l} = \chi^2_{2N \text{ deg. freedom}}/2\pi\Sigma r^2,$$

or

$$= \chi^2_{2N_{df}}D/2N \text{ (remembering that for given probability } \chi^2 \text{ is one-tailed).}$$

Estimates of PLE given by this method are almost identical with those given by the proposed formula, even for N as small as 60 (Table 1, last column). Evidently, the proposed formula is effectively the same as those derived from formal treatment of random populations, and it has the important advantage of giving an estimate of probable error for non-random populations at about the probability implied by choice of t.

Finally, it is worth emphasising that, as discussed briefly in Batcheler (1973, p. 145), use of large R (or absence of any limit on R) often leads to a problem of dealing with a few very large "tail-end" distances in the distribution of r_p and its joint neighbours, particularly if the population is severely aggregated. This can have profound effects upon D, the severity of which can be objectively determined by ranking the r_p (and joint neighbours) from smallest to largest, and calculating D and PLE for conveniently chosen steps of (increasing) R. In most cases, PLE/D steadily diminishes as R increases —as will be intuitively expected—so precision is improved by incorporating the larger measurements. But in samples from extremely aggregated populations, A_1 and A_2 often increase rapidly as f exceeds about 0.8, and the quotient PLE/D becomes less precise (see rabbit pellet example, Figure 7, Batcheler, 1973). In these cases it_s is consistently found that the most precise estimate (i.e. minimum PLE/D) is also the most accurate. This undoubtedly reflects some basic properties of the joint distance technique which lead me in an earlier paper (Batcheler, 1971) to suppose that the smallest 50 percent of the measurements would yield the best estimate. It is probably also related to the situation in bounded plot sampling where $S^2/\bar{x}$ increases rapidly when plot size exceeds the area occupied by the average clump in a population (Greig-Smith, 1964). Further study of these attributes will probably throw better light on the theory of sampling non-random populations by distances.

ACKNOWLEDGEMENTS

H. A. I. Madgwick, I. L. James, A. B. Robson and J. Orwin of this Institute gave valued criticisms of this paper. Mrs Janice Benney typed and retyped the various drafts, and Mrs K. P. Smith drew the diagrams.

REFERENCES

BATCHELER, C. L. 1971. Estimation of density from a sample of joint point and nearest neighbour distances. *Ecology* 52: 703–709.

BATCHELER, C. L. 1973. Estimating density and dispersion from truncated or unrestricted joint point-distance nearest-neighbour distances. *Proceedings of the New Zealand Ecological Society* 20: 131–147.

GREIG-SMITH, P. 1964. *Quantitative Plant Ecology*. Butterworths, London.

KENDAL, M. G. and MORAN, P. A. P. 1963. *Geometrical probability*. Griffin, London.

MORISITA, M. 1957. A new method for the estimation of density by the spacing method applicable to non-randomly distributed populations. *Physiology and Ecology* 7: 134–144.

SKELLAM, J. G. 1952. Studies in statistical ecology, I. Spatial pattern. *Biometrika* 39: 346–362.

THOMPSON, H. R. 1956. Distribution of distance to nearest neighbour in a population of randomly distributed individuals. *Ecology* 37: 391–394.

ERRATA: BATCHELER, C. L. (1973). ESTIMATING DENSITY AND DISPERSION FROM TRUNCATED OR UNRESTRICTED JOINT POINT-DISTANCE NEAREST-NEIGHBOUR DISTANCES.

Proceedings of the New Zealand Ecological Society 20: 131–147.

In para. 2 p. 139 it is stated that the average of two estimates of density, derived from

$$D_1 = d/ab^{-A1},$$

and

$$D_2 = d/ab^{-A2},$$

consistently gave the most accurate estimate for the population. This average is given as

$$(\overline{d/D}) = \frac{(1 + 2.473f)}{2}((1 + 2.717f)^{-A1} + (1 + 2.717f)^{-A2}).$$

However, this formula (which also appears on p. 140) should be

$$\overline{D} = \frac{d}{2(1 + 2.473f)}((1 + 2.717f)^{A1} + (1 + 2.717f)^{A2})$$

or

$$\overline{D} = \frac{d}{2a}(b^{A1} + b^{A2}).$$

Also, on p. 141, errors were made in converting densities of hare and rabbit pellets from Imperial to metric measure.

The two estimates in (25–26) should be

$$239 \pm 103/m^2 \text{ and } 301 \pm 129/m^2.$$

Those listed in (27–29) should be

$$85 \pm 12,\ 66 \pm 47 \text{ and } 113 \pm 31/m^2.$$