The distance method for estimating densities

*M. Keuls**), *H. J. Over***), and *C. T. de Wit****)

U.D.C. 57:311.15

Samenvatting.

De afstandsmethode voor het schatten van een dichtheid, bijv. het stamtal per ha in bossen, werd voorgesteld en uitgewerkt door F. E. Essed [1]. Men zou deze ook kunnen interpreteren als een „wachttijdmethode" ter bepaling van de frequentie per tijdseenheid van een (in de tijd) poisson verdeelde gebeurtenis als het binnenkomen van gesprekken op een telefooncentrale. In dit artikel wordt de methode verder ontwikkeld, waarbij een toepassing, nl. het tellen van slakken (van de soort Galba truncculata, gastheer van de leverbot) in greppels – dit als onderdeel van een onderzoek naar de biologie van de leverbot –, nader wordt beschreven (par. 2).

Na definities (par 1.0) (in bosbouwkundige termen) van achtereenvolgens: een bos, een toevallig bos, een (homogeen) poisson-bos een locaal poisson-bos, worden enige nieuwe schatters gedefinieerd ((1) en (2) in par. 1.1.), en hun eigenschappen voor een homogeen poisson-bos besproken. Voor een locaal poisson-bos wordt een afzonderlijke schatter besproken (par. 1.2). In par. 1.3 wordt een homogeniteitstoets besproken, alsook een dichtheidsschatter, die al naar de omstandigheden een telling van bomen of een afstandsmeting is.

In par. 1 worden de statistische technieken met enkele hulptabellen besproken. Par. 2 beschrijft het ontwikkelen van een veldmethode voor het tellen van slakken bij het leverbotonderzoek. Par. 3 geeft de wiskundige achtergrond voor de methoden beschreven in par. 1.

0. Introduction

In the Netherlands, as in many other West-European countries, the pulmonate snail Galba (Limnaea) truncatula acts as the intermediate host of the liverfluke (Fasciola hepatica), a common parasite in cattle and sheep. As a part of an investigation on the epizootiology of liverfluke disease it was necessary to develop a method for estimating the density of the snailhost in small ditches (the so-called „greppels", the principal places where liverfluke snails occur).

The method to be developed had to give reliable estimates. On the other

*) Mathematical Department, Agricultural University, Wageningen.

**) Central Organisation for Applied Scientific Research (T.N.O.), Den Haag.

***) Institute for Biological and Chemical Research on Field Crops and Herbage (I.B.S.), Wageningen.

hand the need was for a technique in the field, easy enough to handle.

First some remarks about estimating the density of organisms are made. When organisms are distributed in their domain according to some unknown pattern, an unbiased estimate of their density can always be obtained by choosing sample plots (all of the same size) at random, counting the number of organisms in each, averaging and multiplying by the appropriate scale factor. The total sample area being given, better estimates are obtained with a greater number of smaller plots. The size of the sample area of these small plots should not be made too small, because of extra labour and because it may be difficult to decide whether an organism is inside or outside the sample area. A good standard seems to be the use of plots of such a size that the expected number within the plot is about twenty (Essed, [1]).

Essed [1] worked out a method for estimating the density of trees in a forest which is based on the measurement of the distance from an arbitrary point to the n^{th} tree in order of distance, on the supposition that the trees in the forest are distributed in such a way that the expectation value of the density is independent of the place.

He arrived at a formula which relates a consistent estimate of the density with the inverse of the square of the average distance to the fourth tree. The estimate of the density by means of distance measurements has the useful property that the size of the sample area is adjusted automatically to the density of the organisms, which results in a constant coefficient of variation of the estimate.

In this paper Essed's method is adopted to practically any kind of unknown pattern with a stochastic element. The paper consists of three parts. Section 1 gives the model and the statistical techniques developed. It gives the formulae and the tables that are useful in practical applications. Section 2 gives the statistical and practical arguments for the field technique developed in the liver-fluke project. The scheme followed may be used in other fields of application as forestry; even for estimating the density of moving organisms a technique is suggested in section 1.1. The last section gives the mathematical derivations for the formulae and tables of section 1, and some generalizations. However the non-mathematical reader who reads only the first and second parts loses no new aspects of practical use.

1. Description of statistical techniques

1.0 Definitions and model

A *forest* will be a domain in which trees occur. The terms forest and trees will be used in a generalized sense. Depending on the problem one could also speak of snails and a ditch, of telephone calls and a time interval.

A *random forest* (see Essed [1]) is a set of rules for generating a forest, i.e. a law which assigns to each subdomain G of a given domain A a random variable $\underline{k}(G)$ [1] being the number of trees to occur in G.

A random forest will be called a *poisson forest* if for any pair of subdomains G_1 and G_2 that do not overlap:

(1) the numbers $\underline{k}(G_1)$ and $\underline{k}(G_2)$ of trees are independent random variables.

(2) for any G, $\underline{k}(G)$ is a *poisson* random variable $\underline{k}(\lambda)$.

i.e. the probability that $\underline{k}(G)$ takes a value k equals:

$$P[\underline{k}(\lambda) = k] = e^{-\lambda}\, \lambda^k/k!, \text{ where}$$
$$\lambda = \text{the expected number of trees in } G.$$

The *expected density* of trees $N(x)$ (i.e. the expected number of trees per unit of area) at points x of the domain need not be a constant, not even for a *poisson* forest. In fact, taking an arbitrary continuous density function $N(x)$ over the domain; let rules (1) and (2) hold for infinitesimal small regions. As the sum of a set of independent *poisson* variables is itself a *poisson* variable, it is clear that rules (1) and (2) hold for *any* pair of non-overlapping subdomains.

If moreover the expected density $N(x)$ is a constant over the domain, we speak of a *homogeneous poisson forest.* Essed derived his radius-method for estimating the density on the assumption of this model, but showed that even for some other kinds of forest, e.g. a more or less "systematic" forest, the distance to the fourth tree can be applied satisfactorily.

In this paper we are interested in the (general) poisson forest, where $N(x)$ is not a constant. We assume however, that in certain small subdomains $N(x)$ remains rather close to the mean density in its subdomain, so that methods of estimation for homogeneous poisson forests can be applied for the subdomains. We express this by saying that the forest is *locally (homogeneous) poisson.*

An example of a forest, that will be taken as generated according to the general poisson law, is given in fig. 3, concerning snails and a ditch. It is seen that within small areas the points are apparently scattered at random, but that due to causes which do not matter in this section, there are large and significant density differences within large areas. The density here has been estimated according to the techniques developed for the local poisson forest, as is discussed in sect. 1.2 from the statistical viewpoint, and in sect. 2.0 for the practical aspects.

[1] The underlined variable $\underline{k}(G)$ indicates the random variable; k is a realization of $\underline{k}(G)$.

1.1 Density estimates for the homogeneous poisson forest

The distance from an arbitrary point P to the 1^{st}. 2^{nd} ... or n^{th} tree is represented by the random variables $\underline{a}_1$, $\underline{a}_2$ ... $\underline{a}_n$, of which fig. 1 presents a sample. We assume that the trees in the area around the point P are approximately distributed in the poisson way. Then it follows from the geometry of the situation that an unbiased estimator of the density at the point P is given by $c(\underline{a}_n)^{-2}$ for some constant c. The factor of proportionality c can be calculated.

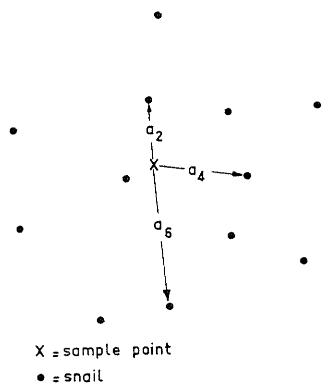

Fig. 1. The distance from a sample point to the 2^{nd}, 4^{th} and 6^{th} tree (or snail).

An unbiased estimator of the density N of the trees at the point P is given by

$$\underline{S}(N) = \frac{n-1}{\pi}\,\frac{1}{\underline{a}_n^{\,2}}. \tag{1}$$

The value of the coefficient $(n - 1)/\pi$ is given in the third column of table 1.

Starting from s sample points in a homogeneous poisson forest, the corresponding measurements ${}_i\underline{a}_n$ with $i = 1, 2, \dots s$ can be obtained and combined in the unbiased estimator

$$\underline{S}(N;\, s,\, n) = \frac{sn-1}{\pi}\,\frac{1}{\underline{a}_{sn}^{\,2}} \tag{2}$$

in which $\underline{a}_{sn}^{\,2}$ stands for ${}_1\underline{a}_n^{\,2} + {}_2\underline{a}_n^{\,2} + \dots {}_i\underline{a}_n^{\,2} + \dots {}_s\underline{a}_n^{\,2}$

By substituting s=1 in (2), the unbiased estimator (1) for a single distance measurement is obtained.

The random variable $\underline{S}(N;\, s,n)$ obtained from s measurements to the n^{th} tree (2) is shown (sect 3.2) to be identically distributed with $S(N;\, 1, sn)$ obtained from one measurement to the $s.n^{th}$ tree:

$$\underline{S}(N;s,n) \cong \underline{S}\,(N;1,s.n) \tag{3}$$

Hence in the following only the properties of the estimator (1), i.e. $\underline{S}\;(N) = \underline{S}\;(N;1,n)$ are discussed, it being understood that in a homogeneous poisson forest n may be read as a product $s.n$.

The variance of $\underline{S}(N)$ equals:

$$\text{Var }[\underline{S}(N)] = \frac{N^2}{n-2}. \tag{4}$$

An unbiased estimator of this variance is obtained by substituting $\underline{S}(N)$ for N and by replacing $n - 2$ by $n - 1$. The variation coefficient of $\underline{S}(N)$ equals

$$\frac{\sigma[\underline{S}(N)]}{N} = \frac{1}{\sqrt{n-2}} \tag{5}$$

and is therefore a known constant only depending on n. It is given in the fifth column of table 1.

TABLE 1

The multiplication factors of $1/\underline{a}_n^{\,2}$ to obtain: 1) *the lower 95% confidence limit of N,* 2) $\underline{S}(N)$, 3) *the higher 95% confidence limit (90% confidence interval). The coefficient of variation of* $\underline{S}(N)$.

n	multiplicationfactors of $\underline{a}_n^{-2}$ for:			variation coefficient
	95% lower limit	$\underline{S}(N)$	95% higher limit	
1	0.016	—	0.95	—
2	0.113	0.32	1.51	—
3	0.26	0.64	2.01	1.00
4	0.44	0.95	2.47	0.71
5	0.63	1.27	2.91	0.58
6	0.83	1.59	3.35	0.50
7	1.05	1.91	3.77	0.45
8	1.27	2.23	4.19	0.41
9	1.50	2.55	4.60	0.38
10	1.73	2.86	5.00	0.35
12	2.20	3.50	5.80	0.32
15	2.94	4.46	6.97	0.28
20	4.22	6.05	8.87	0.24
25	5.53	7.64	10.74	0.21
30	6.87	9.23	12.59	0.19
40	9.61	12.41	16.22	0.16
60	15.19	18.78	23.28	0.13
100	26.74	31.51	37.20	0.10

Confidence limits for N are derived in sect. 3.3. In table 1 the factors are given by which the value of an estimate $\underline{a}_n^{-2}$ should be multiplied. (The variance of $\underline{S}(N)$ cannot be used here, as $\underline{S}(N)$ is not distributed normally.)

For example: there is .90 confidence that the value of N is between $0.44 \dfrac{1}{a_4^2}$ and $2.47 \dfrac{1}{a_4^2}$ if the distance to the fourth tree measured is a_4.

Practical considerations are the following:

There are several reasons, why the distance to the third, fourth or fifth tree is measured:

(1) The amount of work done to find the appropriate tree increases rapidly with increasing n, even for fixed ns, where s is the number of measurements to the n^{th} tree to be combined in a single distance estimator.

(2) The surveyed area around a sample point has to be small if one considers a local poisson forest.

(3) Measurements to the first and second tree are to be avoided if one takes $s = 1$, because the estimator (1) is non-existent for $n = 1$ and the coefficient of variation of this estimator is undefined for $n = 1$ or 2. However these measurements may be combined as in (2).

Essed [1] proved (for $s = 1$) that for estimating standing timber, measurements to the fourth tree are to be preferred as they minimize the cost of the information to be obtained. The fourth organism is also preferred in the case of snails in ditches (sect. 2).

For rapidly moving organisms one can measure in a small more or less homogeneous sample area at s moments the distance to the first organism from *one* sample point, and combine these distance measurements by means of (2).

1.2 Density estimates for the local poisson forest

In a local poisson forest distance measurements $\underline{a}_n$ can be made from k sample points chosen at equidistant intervals. An unbiased estimator of the average density in the sample area of the forest is

$$\underline{S}(\bar{N}) = \frac{1}{k} \sum_{j=1}^{k} \underline{S}(N_j) \tag{1}$$

The variance of this average density $\underline{S}(\bar{N})$ is

$$\text{Var } \underline{S}(\bar{N}) = \frac{1}{k^2(n-2)} \left\{ \bar{N}^2 + \frac{1}{k} \sum_{j=1}^{k} (N_j - \bar{N})^2 \right\}. \tag{2}$$

The unbiased estimator of this variance is again obtained by substituting the estimates for N_j (according to the foregoing section) and by replacing $(n - 2)$ by $(n - 1)$.

A .95 confidence interval is approximated by

$$\underline{S}(\bar{N}) - 1{,}96\, \sigma_{\underline{S}(\bar{N})} < \bar{N} < \underline{S}(\bar{N}) + 1{,}96\, \sigma_{\underline{S}(\bar{N})}$$

where

$$\sigma_{\underline{S}(\bar{N})} = \sqrt{\text{Var } [\underline{S}(\bar{N})]}.$$

(1) and (2) are used if stratification in more or less homogeneous sub-forests of equal size is applied.

More generally let the local poisson forest be partitioned in k more or less homogeneous sub-forests: $1, 2, \ldots, j, \ldots, k$. Throughout the whole forest equidistant sample points are chosen. The number s_j of sample points in the sub-forest j, is chosen proportional to the size of the subforest. (2) of the foregoing section will be the estimator of the density in each subforest. The average of these estimates, weighted proportionally to the numbers s_j is an unbiased estimator for the average density in the forest:

$$\underline{S}(\bar{N}) = \Big(\sum_{j=1}^{k} s_j\, \underline{S}(N_j) \Big) \Big/ \sum_{j=1}^{k} s_j \tag{3}$$

The variance of this average density is

$$\text{Var } [\underline{S}(\bar{N})] = \frac{1}{\{\sum_{j=1}^{k} s_j\}^2} \sum_{j=1}^{k} s_j^2 \frac{N_j^2}{s_j n - 2}. \tag{4}$$

An unbiased estimator of (4) is again obtained by substituting $\underline{S}(N_j)$ for N_j and $(s_j n - 1)$ for $(s_j n - 2)$.

1.3 Additional techniques and remarks.

Justified stratification improves the efficiency of the estimates considerably. Conventional *homogeneity tests* may be used to determime whether a series of values of $\underline{a}_n$ may be considered to be a sample from a homogeneous (sub-) forest.

One test is as follows. One considers the function U on the values $a_n^{(1)}, \ldots, a_n^{(k)}$ of distance measurements to the n^{th} tree:

$$U\, (a_n^{(1)}, \ldots, a_n^{(k)}) = \left\{ \max_{i,j} \left[\frac{a_n^{(i)}}{a_n^{(j)}} \right] \right\}^2 \tag{1}$$

i.e. the maximal ratio U of density estimates. Upper 5 percent points of the corresponding random variable $\underline{U} = U(\underline{a}_n^{(1)}, \ldots, \underline{a}_n^{(k)})$ are given in table 2.

TABLE 2

Upper 5-percent points of the maximal ratio of density estimates
$U\,(\underline{a}_n(^1), \ldots, \underline{a}_n(^k))$

	$k = 2$	3	4	5	6	8	10	12
$2n = 8$	4,43	6,00	7,18	8,12	9,03	10,5	11,7	12,7
$2n = 10$	3,72	4,85	5,67	6,34	6,92	7,87	8,66	9,34
$2n = 12$	3,28	4,16	4,79	5,30	5,72	6,42	7,00	7,48
$2n = 20$	2,46	2,95	3,29	3,54	3,76	4,10	4,37	4,59
$2n = 30$	2,07	2,40	2,61	2.78	2,91	3,12	3,29	3,39
$2n = 60$	1,67	1,85	1,96	2,04	2,11	2,22	2,30	2,30

k = numbeı of strata = number of independent estimates of $\underline{a}_n{}^2$.
n = (or ns) = "combined sample size" for the radius estimate of the density in each of the k strata.

If it is decided to measure the distance to e.g. the fourth tree, it will be found that around certain points the density is so low that it takes a long time to find the fourth tree. It is temping therefore to fix a distance a beyond which the distance method is abandoned, and one counts instead the number of trees within this distance; the proportion of this number (i.e. 0, 1, 2, or 3) to the area concerned is then taken as an estimate of N.

It is proved in section 3.3 that this procedure leads to an unbiased estimator of the expected density N. Table 3 gives the variance of this new "truncation estimate".

TABLE 3

Variances of the truncation estimate $\underline{z}$ of N.
$x = N\pi a^2$ is an area in terms of expected number of trees N for given a.
If $\underline{a}_n \leqslant a$, the distance estimator $(n—1)/\pi\underline{a}_n{}^2$ is applied,
if $\underline{a}_n > a$, the trees in the circle with radius a are counted.

	$n = 4$	$n = 5$	$n = 6$
$x = 0{,}5$	2,016	2,001	2,000
$= 1$	1,052	1,008	1,001
$= 2$	0,635	0,536	0,509
$= 4$	0,514	0,362	0,299
$= \infty$	0,5	0,333	0,25

To arrive at the present methods we accepted the supposition that the expectation value of the density varies continuously with place. However, the errors which are made if discontinuities occur are small as is illustrated in figure 2.

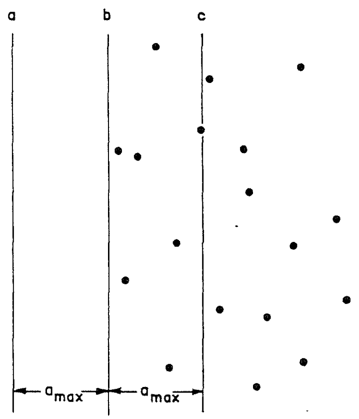

Fig. 2. For explanation see text.

Here it is supposed that to the left side of the line b there are no trees and to the right side the trees are scattered at random and that the maximum area which is surveyed to find n^{th} organism has a radius a_{max}.

The estimate of density is correct so long as the arbitrarily chosen point P is to the left of line a or to the right of line c or on the line b. The density is underestimated with P within the area a-b and overestimated with P within the area b-c.

Since the errors more or less cancel and the frequency of P falling within suspected boundary areas is in general small, it is admissable to apply the proposed method even if discontinuities occur.

2. The estimation of the number of liver-fluke-snails in ditches.

2.0 Development of the method.

As a part of the investigation on the epizootiology of liverfluke disease 15 kilometers of greppels (see the introduction) had to be surveyed regularly for determining the density of snails.

To develop a useful method for surveying the density of snails, the snails in a ditch were mapped over 10 meters (fig. 3). The figure represents the bottom of the ditch (b), which is on the average 10 cm wide, and both sides (s) with an average height of about 20 cm. The sides were folded down into the horizontal plane of the diagram.

The snails are scattered at random around any arbitrary point, but density difference are so large that the distribution is as in a local poisson forest. Estimating density by counting the snails within quadrats appears to be too time consuming and not feasible. The distance method as discussed in this paper was therefore chosen.

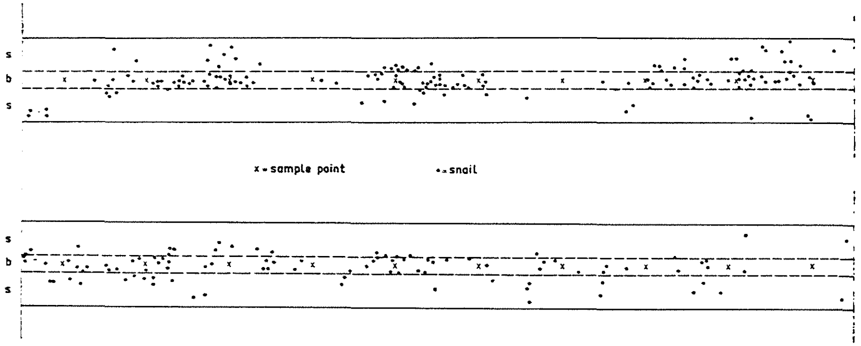

Fig. 3. An observed distribution of snails in a ditch. The bottom (b) is 10 cm wide and the sides (s) with a height of 20 cm are represented in the horizontal plane.

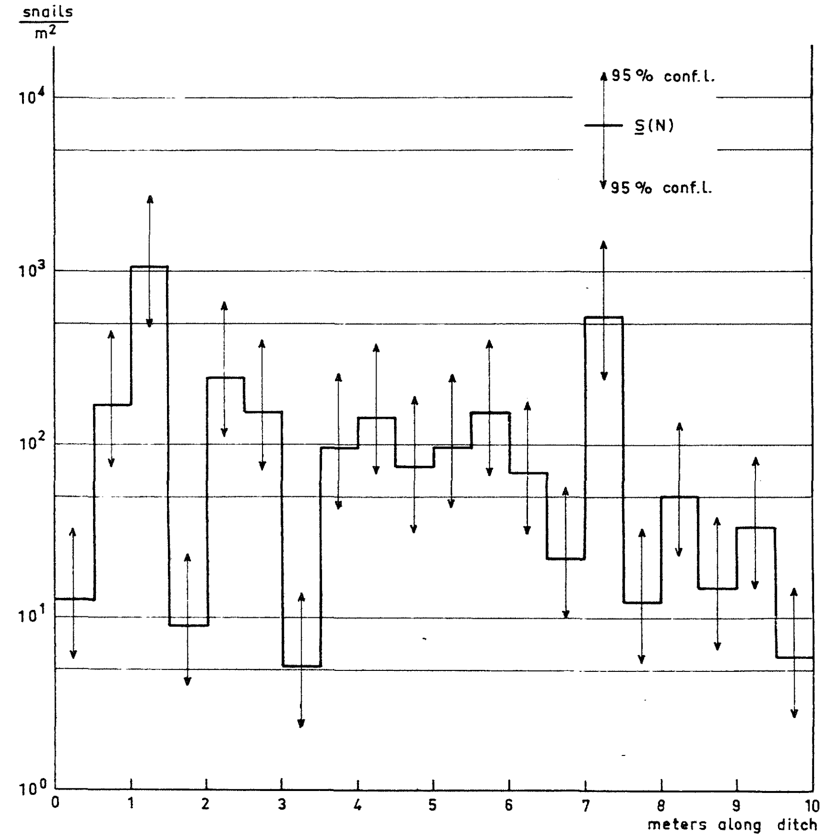

Fig. 4. Densities and their confidence intervals obtained by measuring the distance to the fourth snail at 50 cm intervals in fig. 3.

As the snails in general are restricted to the bottom of the ditch, it appears sufficient to take sample points along the bottom of the ditch only, for estimating the number of snails per m² of ditch[1]).

Twenty sample points, 50 cm apart, are marked by crosses in figure 3. Around each sample point the distance to the fourth snail is measured. Estimates of the density around each point and of the .95 confidence limits are obtained by multiplying the inverse of the square of these distances by the factors given in table 1 for $n = 4$. The resulting density estimates and the confidence intervals (at a confidence 0.90) are represented on a logarithmic scale against the distance along the ditch in fig. 4. A logarithmic scale is chosen here to accommodate the wide range of densities and because the length of the confidence interval is independent of the density.

The graphical representation of fig. 4 appears to give a good picture of the actual densities in fig. 3. The confidence limits seem to be rather wide apart in table 1, but this is not unduly so compared with the large density differences along the ditch. Maximum densities around 1, 2, 5 and 7.25 meters and minimum densities around 0.5, 1.75, 3.25 and 9.75 meters appear to be significantly different from the densities in the neighbourhood. The estimates of $\overline{N}$ and $\sigma_{\underline{S}(\overline{N})}$ ((1) and (2) of sect. 1.2) are 148 and 364, respectively. The large estimate of σ is due to the large density differences along this 10 meter of ditch, and is therefore acceptable.

The low densities (it concerns the rather dry weeks of May 1960) appear to occur where the ditch is dry, and filled up or trampled, the high densities where it is filled with some water and open. Because of the good correlation of density with the condition of the ditch it is advantageous to make a stratification by noting down for each sample point, whether the ditch is open and wet (o.n.), open and dry (o.d.), filled up and wet (d.n.) or filled up and dry (d.d.).

The distance from the middle of the ditch to the edge, measured along the surface, is so small that it is necessary to restrict distance measurements to a maximum of 26 cm and to count the number of snails (0, 1, 2, 3) within this radius otherwise. In this way much time is saved at spots where there are few snails.

Some additional trial runs showed that the method is feasible enough if about three minutes at most are spent on one estimate. To achieve quick results a ruler, 30 cm (fig. 6*b*) long, was made, in which the chart of fig. 6*a*

[1]) During the investigation it appeared that the differences between the three zones fade away during the grazing season due to a high ground water level and trampling down by cattle.

Fig. 5. A fourth snail (length: 7 mm).

is fixed. This chart contains four columns marked o.n., o.d., d.n., d.d., with a length of 26 cm and a section to add remarks.

2.1 Description of the method chosen for field work

The observer walks with a tentpeg and the ruler along the ditch and drops the tentpeg at intervals of 10 or 20 meters in the middle of the ditch. He then measures the distance from the tentpeg to the projected position of the fourth snail. (This measurement is sometimes a rather rough one, because of plant growth and because the snails may be found on plants or in the water above the soil surface.)

He marks the distance on the chart in the ruler with a ball point. The column on the paper chosen depends on the observed condition of the ditch. The number of snails within a radius of 26 centimeters is noted down on the right side of the paper in the appropriate row, if the fourth snail is outside this radius.

The presence of brood of snails within a radius of 26 centimeters is also noted down here.

To tabulate the results later on, the paper is replaced in the ruler, and then fixed with a transparant cover on which a density scale is engraved on logarithmic progression (fig. 6*c*).

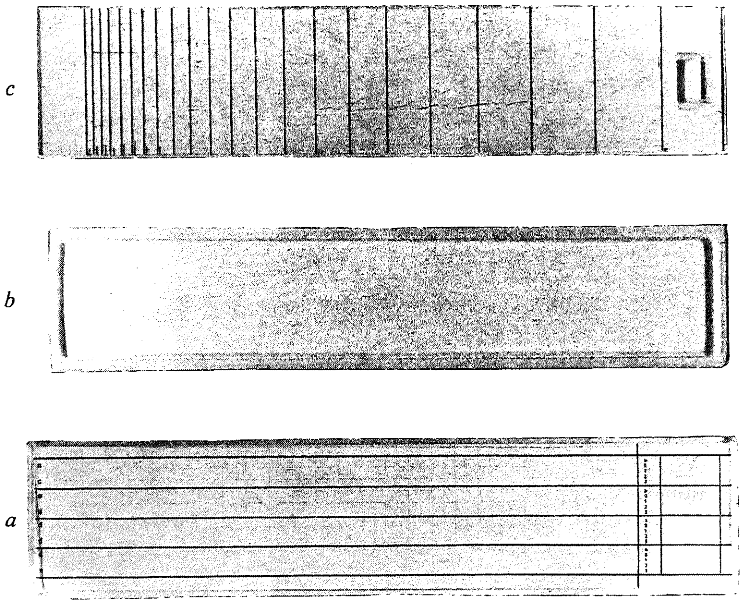

Fig. 6. Chart (a) for noting distances. Ruler (b). Transparant cover with density scale (c).

3. Derivation of formulae for the homogeneous Poisson forest

3.0 The homogeneous poisson forest

Some definitions have been given in sect. 1.0.

By Essed [1] a random forest is shown to be a homogeneous poisson forest, if the following conditions are satisfied[1]).

1. For any subdomain S_x of size x, the probability $P[\underline{k}(S_x) = k]$ that the number of trees $\underline{k}(S_x)$ in the subdomains S_x takes the value k only depends on x.
2. If the intersection of the subdomains S and S' is void, then the random variables $\underline{k}(S)$ and $\underline{k}(S')$ are independent.
3. $P[\underline{k}(S_x) = 0]$ is a continuous function of x.
4. "Trees are not tied". Bij "tied" is meant the following. If two or more trees occur in a subdomain S_x of size x, x sufficiently near to zero, then the probability is almost 1 that these trees occupy the same point of S_x.

[1]) Another way of defining a poisson forest has been given by M. S. Bartlett [3].

In the random forest so defined, the number of trees $\underline{k}(S_x)$ in a subdomain S_x is [1] a poisson random variable $\underline{k}(\lambda)$, where λ is proportional to x:

$$\underline{k}(S_x) \cong \underline{k}(\lambda), \text{ where } P[\underline{k}(\lambda) = k] = e^{-\lambda}\,\lambda^k/k!$$
$$\lambda = E[\underline{k}(S_x)] = \alpha x \qquad (1)$$

The property $\lambda = \alpha x$ allows to choose as a standard unit of size, that size for which $\lambda = 1$. Then the size x of S_x expressed in standard units equals λ:

$$\text{size } x \text{ of } S_x = E[\underline{k}(S_x)] = \lambda \qquad (2)$$

The size y in conventional units (e.g. square meters in the twodimensional case) is related to x by

$$x = Ny \qquad (3)$$

where the constant N is the expected density, i.e. the expected number of trees per conventional unit of size.

3.1 The random size $\underline{x}_n$ up to the n^{th} tree

Essed considers the distance $\underline{a}_n$ from a chosen point P in the random forest up to the n^{th} tree. The distance $\underline{a}_n$ is the radius (in meters) of a closed circular random subdomain $S_{\underline{y}_n}$, with P as a centre and with random size $\underline{y}_n = \pi\underline{a}_n^2$ (square meters). The subdomain $S_{\underline{y}_n}$ is completely determined by the condition that it contains exactly $n - 1$ trees as inner points, and at least one tree on the border. By way of speaking $\underline{y}_n$ is obtained by inflating a closed circular subdomain with centre P and initial radius O until the border meets the n^{th} point. We call $\underline{y}_n$ the random size (in square meters) up to the n^{th} point.

In more general terms a random variable $\underline{x}_n$ (in standard units) is obtained as follows. First a point P is fixed. Then a class S of closed (measurable) subdomains S_x, where $x \geqslant 0$ denotes the size of S_x, is chosen satisfying the conditions:

a. for every pair S_x, S_{x^1}: $S_x \subset S_{x^1}$ for $x^1 \geqslant x$
i.e. the elements S_x are "nested".
b. to every $x \geqslant 0$ there corresponds one element S_x.
c. S_0 is one point, the fixed point P, called the centre of S[1]).

Such a class will be called an *inflation*. Examples are given in fig. 7. Example VI may give considerable trouble with border trees in practice.

[1]) One could even think of S_0 being part of a line or some other point set of zero measure.

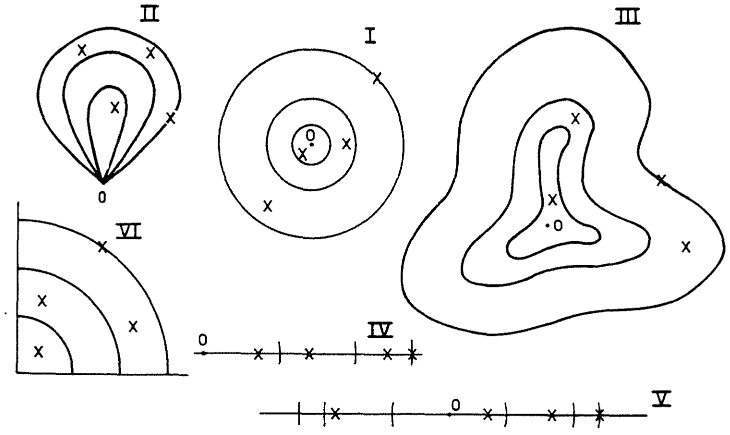

Fig. 7. Examples of inflations.

To any inflation S at a given point P and a given realization F of the random forest $\underline{F}$ there exists a unique x_n for positive integer n, so that

$$S_x \text{ contains } \begin{cases} \leqslant n-1 & \text{trees for } x < x_n \\ \geqslant n & \text{trees for } x \geqslant x_n \end{cases} \qquad (1)$$

S_{x_n} now contains exactly n-1 trees in the interior and at least one tree on the border. x_n is called the size of the inflation S up to the n^{th} tree. $\underline{x}_n(S)$ is the random size up to the n^{th} tree in the random forest $\underline{F}$ for a given inflation S.

To derive the distribution of $\underline{x}_n(S)$, we note that by (1) we have for a given inflation and a value F of $\underline{F}$, the equivalence of the following statements on x and n

$$k(S_x) < n \text{ and } x_n(S) > x \qquad (2)$$

They concern respectively the number of trees in the subdomain S_x of size x, belonging to the class S, and the "size x_n up to the n^{th} tree" of S. In probability terms (2) implies an equivalence of events. Thus:

$$P[\underline{k}(S_x) < n] = P[\underline{x}_n(S) > x]$$

Considering a homogeneous poisson forest, it follows that (λ = size x in standard units):

$$P[\underline{x}_n(S) > x] = P[\underline{k}(S_x) < n] = P[\underline{k}(\lambda) < n] =$$
$$= \sum_{k=0}^{n-1} P[\underline{k}(x) = k] = \sum_{k=0}^{n-1} e^{-x}\, x^k/k!$$

Therefore the cumulative distribution (c.d. function) of $\underline{x}_n$

$$F(x) = P[\underline{x}_n(S) \leqslant x] = 1 - P[\underline{x}_n(S) > x] = \sum_{k=n}^{\infty} e^{-x}\,\frac{x^k}{k!}$$

is independent of the class S and is a continuous function of x with probability density (p.d. function)

$$f(x) = \frac{dF(x)}{dx} = e^{-x}\, x^{n-1}/(n-1)! = e^{-x}\, x^{n-1}/\Gamma(n) \tag{3}$$

This is the p.d. function of the gamma random variable $\underline{\gamma}_n$ with parameter n:

As $\underline{\gamma}_n \cong \frac{1}{2}\, \underline{\chi}_{2n}{}^2$ (well known) we conclude

$$\underline{x}_n(S) \cong \underline{\gamma}_n \cong \tfrac{1}{2}\underline{\chi}_{2n}{}^2 \tag{4}$$

(The symbol $\cong$ expresses that the two random variables concerned have the same c.d. function, one says $\underline{x}_n$ is isomorous to $\underline{\gamma}_n$).

Moments and other properties of chisquare or gamma variables are found in textbooks. (See [5] page 370.) In particular, we have

$$E[(\underline{\gamma}_\alpha)^t] = \Gamma(\alpha + t)/\Gamma(\alpha) \text{ for any } \alpha > 0,\ t + \alpha > 0 \tag{5}$$

Thus moments of $\underline{x}_n$ are:

$$\begin{aligned} &E[\underline{x}_n] = n,\ E[\underline{x}_n{}^2] = (n+1).n \\ &E[\underline{x}_n{}^{-1}] = (n-1)^{-1},\ E[\underline{x}_n{}^{-2}] = \{(n-1)(n-2)\}^{-1},\ n > 2 \end{aligned} \tag{6}$$

and for t a positive integer:

$$\begin{aligned} E[\underline{x}_n{}^t] &= \Gamma(n+t)/\Gamma(n) = (n+t-1)!/(n-1)! = \\ &= n(n+1)\dots(n+t-1) \end{aligned}$$

3.2 The distance-method for estimating the density

Let $\underline{y}^{(1)}, \underline{y}^{(2)}, \dots \underline{y}^{(s)}$ be a random sample of independent measurements of the size $\underline{y}_n(S)$ (e.g. in square meters) up to the n^{th} tree for a chosen inflation. For the two-dimensional case and an inflation of concentric circular discs we have $\underline{y}_n = \pi \underline{a}_n{}^2$, where $\underline{a}_n$ is the distance up to the n^{th} tree from a centre P.

The size in standard units $\underline{x}_n$ is related to $\underline{y}_n$ by

$$\underline{x}_n = N\underline{y}_n,$$

where the constant N is the density (number of trees per conventional unit of size) in the homogeneous poisson forest.

The p.d. functions $f(t)$ and $g(t)$ of $\underline{x}_n$ and $\underline{y}_n$ respectively are related by

$$g(t) = N.f(Nt)$$

Substituting the function f from (3) of sect. 3.1 we get, in terms of a variable y:

$$g(y) = N\, e^{-yN}\, (yN)^{n-1}/(n-1)! \tag{1}$$

Now the likelihood function for a given sample is

$$\begin{aligned} L &= \log \{g(y^{(1)}).g(y^{(2)})\dots g(y^{(s)})\} = \\ &= s \log N - N\, \Sigma_1{}^s\, y^{(i)} + (n-1)\, \Sigma_1{}^s \log\, [N.y^{(i)}] - s \log\, [(n-1)!] = \\ &= ns \log N - N\, \Sigma_1{}^s\, y^{(i)} + \{\text{terms independent of } N\} \end{aligned}$$

We denote by $\hat{N}$ the solution of

$$\frac{dL}{dN} = 0, \text{ i.e. } \frac{ns}{N} - \Sigma y^{(i)} = 0$$

Therefore: $\hat{N} = ns/ \sum_{i=1}^{s} y^{(i)}$

In random samples $\underline{y}^{(1)}, \dots, \underline{y}^{(s)}$ the maximum likelihood (M.L.) estimator $\underline{\hat{N}}$ of N satisfies:

$$\underline{\hat{N}} = ns/ \sum_{i=1}^{s} \underline{y}^{(i)}$$

The M.L. estimator $\underline{\hat{N}}$ of N is a *sufficient* estimator of N as is clear from the fact, that terms in $\underline{L}$ depending on N, contain no further functions of the measurements $y^{(i)}$ than the one occurring in the expression for $\underline{\hat{N}}$.

The distribution of the estimator $\underline{\hat{N}}$ will now be obtained. According to a well-known theorem a sum of independent chisquare (or gamma) random variables is itself a chisquare (or gamma) random variable with the sum of the parameters (number of degrees of freedom) as a parameter. In view of (4) of sect. 3.1 we obtain

$$\sum_{i=1}^{s} \underline{y}^{(i)} = \frac{1}{N} \sum_{i=1}^{s} \underline{x}_n{}^{(i)} \cong \frac{1}{N} \underline{x}_{ns} = \underline{y}_{ns} \tag{3}$$

Thus the sum of s independent "sizes up to the n^{th} tree" is equivalent to one "size-measurement up to the sn^{th} tree".

The random variable $\underline{\hat{N}}$ is therefore:

$$\underline{\hat{N}} = \frac{ns}{\sum_{i=1}^{s} \underline{y}^{(i)}} \cong \frac{ns}{\underline{y}_{ns}} = \frac{N\, ns}{\underline{x}_{ns}} \cong \frac{2N\, ns}{\underline{\chi}_{2ns}{}^2}. \tag{4}$$

By (4) and (5) of sect. 3.1:

$$E(\underline{\hat{N}}) = Nns\,.\,E[\underline{x}_{ns}{}^{-1}] = N\, ns\, (ns-1)^{-1}$$

Thus an unbiased and sufficient estimator $\underline{S}(N)$ of N is given by

$$\underline{S}(N) = \frac{ns-1}{ns} \underline{\hat{N}} = \frac{ns-1}{ns} \cdot \frac{ns}{\underline{y}_{ns}} = \frac{ns-1}{\underline{y}_{ns}}. \tag{5}$$

The variance[1] of the estimator $\underline{S}(N)$ follows from (6) of sect. 3.1:

$$\text{var } [\underline{S}(N)] = E\, [\{\underline{S}(N)\}^2] - N^2 = E\left[\left(\frac{N(ns-1)}{\underline{x}_{ns}}\right)^2\right] - N^2 =$$

$$= N^2\{(ns-1)^2 E[\underline{x}_{ns}^{-2}] - 1\} = N^2\left\{\frac{(ns-1)^2}{(ns-1)(ns-2)} - 1\right\} = \frac{N^2}{ns-2}. \quad (6)$$

Concerning the radius-method of estimating density and its generalization, the following theorem may now be formulated.

Theorem: Given a random sample of independent measurements $\underline{y}^{(1)}$, $\underline{y}^{(2)}$ $\ldots \underline{y}^{(s)}$ of the size $\underline{y}_n(S)$ up to the n^{th} tree ($\underline{y}_n = \pi \underline{a}_n^2$; $\underline{a}_n$ is the distance from P up to the n^{th} tree for an inflation S of concentric circular discs with centre P) in a homogeneous poisson forest, there exists an unbiased estimator $\underline{S}(N)$ of the density N (number of trees per unit of size) with minimal variance among all unbiased estimators:

(1) $\underline{S}(N) = (ns-1)/\underline{y}_{sn}$, where $\underline{y}_{sn} = \underline{y}^{(1)} + \underline{y}^{(2)} + \ldots \underline{y}^{(s)}$
(2) $\underline{S}(N) \cong 2N(ns-1) \,.\, \{\underline{\chi}_{2ns}^2\}^{-1}$
(3) $E[\underline{S}(N)] = N$, Var $[\underline{S}(N)] = N^2/(ns-2)$.

Essed used for his radius estimate an average of distance measurements instead of an average of their squares. His estimate is easier to compute but not so efficient; also the distribution of his estimate was only approximated.

It should be stressed that according to the theorem in the *homogeneous poisson* forest the two ways of measurement
1. s independent measurements up to the n^{th} tree
2. 1 measurement up to the sn^{th} tree
give equivalent estimates $\underline{S}(N)$, which justifies the notation

$$\underline{y}_{sn} = \underline{y}_n^{(1)} + \underline{y}_n^{(2)} + \ldots \underline{y}_n^{(s)}.$$

An exact confidence interval of N is now easily obtained by means of a chisquare table.

The statement with confidence $1 - 2\alpha$ at a given estimate $S(N)$

$$\frac{2N(n-1)}{\chi^2_{2n;\alpha}} < S(N) < \frac{2N(n-1)}{\chi^2_{2n;1-\alpha}}$$

where $\chi^2_{2n;\alpha}$ is the upper α-percent point of the chisquare random variable for $2n$ degrees of freedom, gives by conversion

$$\frac{S(N)\chi^2_{2n;1-\alpha}}{2(n-1)} < N < \frac{S(N)\chi^2_{2n;\alpha}}{2(n-1)}.$$

Table 1 gives upper and lower limits for a 0,90 confidence interval.

Next suppose we want to examine whether a local poisson forest is homogeneous. We may then subdivide the domain into k subdomains (stratification). The hypothesis under test will be $N_1 = N_2 = \ldots N_k$ for the k strata. We consider the case $s_1 = s_2 = s_k = s$, i.e. the estimates $S(N_j)$ depend on equal numbers s_j of size measurements up to the n^{th} tree. The estimates $\underline{S}(N_j)$ correspond to independent samples from $\underline{\chi}^2_{2ns}$. Table 31 of Pearson and Hartley [2] gives upper 5% points for the quotient of maximum and minimum in random samples of size k. In table 2 some values have been taken from this table.

3.3 An unbiased estimator of N from a truncated inflation

So far we suggested measuring the distance from a random point up to the n^{th} tree. However if the trees are thinly spread, counting the trees might be a much more practical device. Therefore the following procedure is suggested One measures the distance $\underline{a}_n$ as long as $\underline{a}_n < a$, where a is a value, fixed beforehand. If on the other hand within a distance a the n^{th} individual does not occur, then we consider the number of trees in the circle with radius a divided by πa^2 as the estimate of the density N.

In other words we consider a realization F of the random forest $\underline{F}$, an inflation S with centre $P = S_0$ and the following function $z(x; F, S, n)$ of x (where $x = Ny = N\pi a^2$) for chosen n

$$\begin{aligned} z &= \frac{N(n-1)}{x_n} \text{ for } x_n \leqslant x. \quad \left(z = \frac{n-1}{\pi a_n^2} \text{ for } a_n \leqslant a\right) \\ &= \frac{N \,.\, k[S_x]}{x} \text{ for } x_n > x \quad \left(z = \frac{k[S_a]}{\pi a^2} \text{ for } a_n > a\right), \end{aligned} \quad (1)$$

to define for given x an estimator $\underline{z} = z(\underline{F}; S, n, x)$ of the density N in the random forest $\underline{F}$. We will derive the expectation and the variance of $\underline{z}$.

First we note for a given F the equivalence of the statements concerning x, n and the inflation S:

$$x_n(S) > x \quad \text{and} \quad k(S_x) < n \quad (2)$$

Therefore the probabilities of the $n + 1$ events concerning $\underline{x}_n = x_n(\underline{F}; S)$:

$$\underline{x}_n(S) \leqslant x,\ \underline{k}(S_x) = k,\ k = 0, 1, \ldots n-1$$

add to 1.

[1] In van der Waerden [4] sect. 43, Example 29 it is shown that the sufficient unbiased estimator (5) is unique and therefore has minimal variance in the class of all unbiased estimators.

The expectation of $\underline{z}$ is given by

$$E[\underline{z}] = \int\limits_{u\leqslant x} \frac{N(n-1)}{u}\, dF(u) + \sum_{k=0}^{n-1} \frac{Nk}{x}\, P[\underline{k}(S_x) = k],$$

where $F(x)$ is the c.d. function of $\underline{x}_n(S)$.

According to sect 3.1 we find

$$\int\limits_{u\leqslant x} \frac{1}{u}\, dF(u) = \int\limits_{u\leqslant x} \frac{1}{u} \frac{e^{-u}\, u^{n-1}}{(n-1)!}\, du = \frac{1}{n-1} \int\limits_{u\leqslant x} \frac{e^{-u}\, u^{n-2}}{(n-2)!}\, du = \frac{1}{n-1} P[\underline{x}_{n-1}(S) \leqslant x]$$

$$\sum_{k=0}^{n-1} \frac{k}{x} P[\underline{k}(S_x) = k] = \sum_{0}^{n-1} \frac{k}{x} \frac{e^{-x} x^k}{k!} = \sum_{1}^{n-1} \frac{e^{-x} x^{k-1}}{(k-1)!} = \sum_{0}^{n-2} \frac{e^{-x} x^k}{k!} =$$

$$= P[\underline{k}(S_x) < n-1] = P[\underline{x}_{n-1}(S) > x]$$

Therefore substituting the two results:

$$E[\underline{z}] = N \,.\, P[\underline{x}_{n-1}(S) \leqslant x] + N \,.\, P[\underline{x}_{n-1}(S) > x] = N \qquad (3)$$

The variance of $\underline{z}$ is obtained in an analogous way:

$$\text{var}\ [\underline{z}] = E[\underline{z}^2] - N^2 = (n-1)^2 N^2 \int\limits_{u\leqslant x} \frac{1}{u^2} dF(u) + N^2 \sum_{0}^{n-1} \frac{k^2}{x^2} P[\underline{k}(S_x) = k] - N^2$$

$$= \frac{n-1}{n-2} N^2 P[\underline{x}_{n-2} \leqslant x] + N^2 \left\{ P[\underline{x}_{n-2} > x] + \frac{1}{x} P[\underline{x}_{n-1} > x] \right\} - N^2$$

$$= \frac{N^2}{n-2} + N^2 \left\{ \frac{1}{x} P[\underline{x}_{n-1}] > x] - \frac{1}{n-2} P[\underline{x}_{n-2} > x] \right\}. \qquad (4)$$

It may be verified that the second term is positive and decreases with x. Table 3 gives some values of this variance.

3.4 The local poisson forest

The formulae given in sect 1.2 for the local poisson forest are straightforward and no derivations seem necessary. From (2) of sect. 1.2 one sees that the variance of the estimator of average density increases with greater heterogeneity between the strata by a term $\Sigma(N_j - \overline{N})^2$. At its minimum the variance equals

$$\text{Var}\ [\underline{S}(\overline{N})] = \frac{1}{k(n-2)} N^2$$

for the case of homogeneity. Apparently in that case the estimator $\underline{S}(\overline{N})$ is not an efficient one. It may be asked what is the loss by stratification or in other terms what is the relative efficiency in case of homogeneity.

According to (3) of sect. 1.1. we have in the case of homogeneity the efficiency factor with respect to the "unbiased minimal variance" estimator:

$$E = \frac{1}{kn-2} : \frac{1}{k(n-2)} = \frac{n-2}{n-2/k}$$

i.e. using one distance up to the 4th tree per stratum, E tends to $\frac{1}{2}$ for large k. Taking however 10 distances up to the fourth tree per stratum, we have $n = 40$, and E tends to 0,95 for large k.

Acknowledgements

Thanks are due to prof. dr. N. H. Kuiper for criticism and encouragement.

Literature.

[1] Essed, F. E., Estimation of standing timber. Thesis (1957) (Agr. Un. Wageningen).

[2] Pearson, E. S. and H. O. Hartley, (1954). Biometrika Tables for statisticians. (Cambridge University Press).

[3] Bartlett, M. S. (1960)., Stochastic Population Models in Ecology and Epidemiology. Methuen and Co. London.

[4] Waerden, B. L. van der (1957), Mathematische Statistik. Springer-Verlag, Göttingen.

[5] Kendall, M. G. and Alan Stuart (1958), The Advanced Theory of Statistics· Ch. Griffin & Co., London.