

the text for  
this paper is  
at the front -  
and the  
figures are at  
the back

## **SADIE: software to measure and model spatial pattern**

By J N PERRY<sup>1</sup>, E D BELL<sup>2</sup>, R H SMITH<sup>2</sup> and I P WOIWOD<sup>1</sup>

<sup>1</sup> *Department of Entomology & Nematology, IACR Rothamsted Experimental Station,  
 Harpenden, Herts. AL5 2JQ, UK*

<sup>2</sup> *Department of Zoology, University of Leicester, University Road, Leicester LE1 7RH, UK*

### **Summary**

The current status of the SADIE system for measuring and testing for spatial pattern in data from a single species and for spatial association in data for two species is outlined. SADIE (Spatial Analysis by Distance IndicEs) is summarized, and examples are given of its use in analysis and modelling. Fortran software is available free from the first-named author.

**Key words:** Spatial pattern, regularity, crowding, aggregation, distance indices, scale, cellular automata, dispersal, heterogeneity

### **Introduction**

Data for invertebrates, weeds and diseases in agriculture are gathered at a variety of levels of spatial information. At the highest level are maps, where the location of each individual is known. These are less common than counts of individuals of a certain species at a particular location. At the lowest level are summary statistics such as the sample mean and variance of a frequency distribution. For counts, previous approaches considered only the relationship between variance and mean. However, although the set of counts of cyst-nematodes in six soil cores: {0, 1, 4, 56, 484, 4095}, may be highly-skewed and obviously non-Poisson, their spatial arrangement may be completely random. Conversely, a set of counts of carabid beetles in pitfall traps: {0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 5}, may conform closely to a Poisson distribution, but if sampled in that order along a line transect show an obvious linear trend departing strongly from randomness. To overcome such problems Perry (1995a) introduced a new method to detect and measure spatial pattern, termed Spatial Analysis by Distance IndicEs (SADIE). Consider the three different arrangements of 36 individuals in a 3x3 grid shown in Table 1. The observed arrangement in (b) is clearly clustered towards the top-left of the grid. Its degree of aggregation may be measured by calculating the minimum amount of effort, equated to distance moved, that the sampled individuals would need to expend in order to achieve an extreme pattern, such as

Table 1. *Three arrangements of 36 individuals in a 3x3 grid*

(a) : Crowded	(b): Observed	(c) : Regular
36    0    0	13    7    3	4      4      4
0      0    0	5      4    1	4      4      4
0      0    0	1      1    1	4      4      4

complete crowding (a), when all individuals occur in a single sample unit, or complete regularity (c), when each unit has the same number of individuals. The concept has also been extended to data in the form of maps (Perry, 1995b). The two advantages of SADIE for counts are its improved intuitive basis, compared with traditional more abstract, mathematical approaches, and its use of all the spatial information in the sample. The purpose of this paper is to summarise progress to date in the development of methodology and software for analysis and modelling.

## Analysis of Spatial Pattern

### *Maps, single species data*

For mapped data, the SADIE technique requires two-dimensional coordinates of each individual to be specified, within a given rectangular area. The movement of these observed locations to a final regular arrangement is effected by means of an iterative algorithm (Perry, 1995b) that relocates each point towards an arrangement in which points occupy a triangular lattice. The algorithm operates by simple rules based on the construction of Voronoi polygons. The distance between the corresponding positions of each individual in the observed and final arrangement is noted and the sum, over all the  $n$  observed individuals gives the distance to regularity,  $D$ . A test of randomness and an index of non-randomness may be found as follows.

Firstly,  $n$  random points are generated independently of one another within the sample area.

The algorithm is run as above for this set of random points, the distance to regularity computed, and this value,  $D_{\text{rand}}$ , is stored. This is repeated, usually about 100 times. The resulting set of values of  $D_{\text{rand}}$ , is then ordered and a one-sided randomization test of complete spatial randomness follows (e.g. Perry & Hewitt 1991) against the alternative that the observed arrangement is either aggregated or regular. If the average value of  $D_{\text{rand}}$  over the randomizations is denoted as  $E_p$ , then an index of aggregation,  $I_p$ , is calculated from  $D/E_p$ . Values of  $I_p > 1$  indicate aggregation,  $I_p < 1$  indicates regularity and  $I_p \approx 1$  suggests randomness.

It is visually useful to plot the observed location (as a numeral) linked to the final position with a straight line, for each point, in an 'initial and final' (IAF) plot. This aids the identification of clusters and of areas of relatively low density; Fig. 1(a) shows an example for 105 of Bliss' Japanese beetle larvae (Perry, 1995b), for which  $I_p = 2.32$  ( $P_p = 0.0063$ ). Another useful diagnostic plot is analogous to the EDF plot of Diggle (1983), in which the individual distances shown in the IAF plot are ranked and cumulated. Fig. 1(b) shows the observed data (dotted line) greatly exceeds the upper 97.5th centile (top bold line) of the randomizations.

Fig. 1. Data for 105 beetle larvae (Perry 1995b): (a) an IAF plot and (b) an EDF plot.

### Counts, single species data

For count data, the SADIE technique requires only the two-dimensional coordinates of each sample unit and its associated count to be specified, but places no restriction on the arrangement of the sample units themselves. The distance to regularity,  $D$ , is found by means of the transportation algorithm from the operational research literature (Perry, 1996a). Now, the randomizations are made by permuting the observed counts among the sample units, not the total number of individuals. This conditioning on the counts allows inferences to be made about the spatial pattern of the observed counts, after allowance for the heterogeneity of their frequency distribution. (This heterogeneity is a function of the spatial pattern at a smaller scale than that to which the sample unit count relates and is therefore ignored.) As for maps,  $D$  is compared with  $E_a$ , the average value from the random permutations, to yield an index of aggregation,  $I_a$ , formed from  $D/E_a$ , and a randomization test of randomness is based on the probability,  $P_a$ . Alternatively, the distance to crowding,  $C$ , may be found, by direct search over all the sample units; the permuted randomizations yield an index,  $J_a$  and probability,  $Q_a$ , in similar fashion (Perry, 1996a). Consider, as an example, counts of *Heterodera avenae*, the cereal cyst-nematode, collected in soil cores by B. Boag (Perry, 1996a), on a 15x15 grid, in Fig. 2. The presence of at least two clusters is clear and may be demonstrated by the equivalent plot, for counts, of the IAF plot for mapped data ( $I_a=1.46$ ,  $P_a=0.005$ ). A little thought will show that

Fig.2. Counts of nematodes at 7.14m spacing, collected by B. Boag (Perry 1996a)

the index based on distance to crowding,  $J_a$ , cannot detect non-randomness with any power if the data contain more than a single cluster. Indeed, here,  $J_a=1.03$  and  $Q_a=0.265$ . When sample units form large contiguous rectangular grids, Perry (1996a) showed how changes in the values of both  $I_a$  and  $J_a$  with sub-grid size may be considered simultaneously, to yield useful information regarding cluster-size and inter-cluster distance. For the nematode data, sub-grids of size  $r \times r$  units yielded median values of  $J_a$  that achieved a maximum of about 1.3 between  $r=5$  and  $r=7$ , whilst  $I_a$  stabilized around 1.5 for  $r>9$ . These results implied that the main contribution to the spatial pattern came from clusters with an approximate diameter of five to six units, separated by a distance of about four units. The IAF plot confirmed this conclusion. One of us (EDB) has proposed a potentially more powerful approach to utilising information from the distance to crowding that should prove simpler for data analysis and that does not require a rectangular grid. This involves an extension of the single focus concept for the distance to crowding,  $C_{(1)}$ , to several, say  $n$ , foci, and the derivation of information concerning the clusters from the relationship between  $C_{(k)}$  and  $k$ ,  $k=1, \dots, n$ .

### Modelling Spatial Pattern and Association

#### *Counts, single species data*

There are now many software packages that allow the simulation of discrete statistical frequency distributions, such as the Poisson, negative binomial or beta-binomial, that are used commonly to describe counts in agriculture and ecology. However, the lack of a methodology to study the spatial pattern of counts, now provided by the SADIE system, has prevented hitherto the development of techniques to simulate spatial arrangements of counts with given levels of aggregation or regularity. The automatic generation of such arrangements will prove useful in simulation models and in evaluations to compare the efficiencies of different sampling plans for pest monitoring. Perry (1996b) described an algorithm to generate an arrangement of a given set of counts of a single population over predefined locations that has any desired level of aggregation. In particular, this may be used to permute a set of observed counts between sample units, to form a different arrangement, but one with a very similar degree of aggregation (defined through the distance to regularity) as the original. Furthermore, because of the multiplicity of possible permutations, for most sets of data it is possible to find hundreds of such different arrangements. The algorithm works by selecting pairs of counts at random and exchanging them if certain criteria are met, starting from a random permutation of the counts.

Fig.3. *C. assimilis* counts (Perry 1996b): (a) observed, (b) permutation with similar  $D$  value

In addition to aggregation, as measured by distance to regularity, an important feature of any observed arrangement concerns the degree to which the counts occupy units towards the 'edge' of the sampled area. This may be formally measured by  $\bar{\delta}$ , the distance from the centroid of the counts to the centroid of the sample units. Because the distance to regularity is inflated by relatively large values of  $\bar{\delta}$ , it is important to select the new arrangement to have a very similar value of  $\bar{\delta}$  to that of the observed; fortunately this restriction is relatively easy to impose. As an example, consider counts of *Ceutorhynchus assimilis*, a weevil, collected in a field of oil seed rape by A.K. Murchie (Perry, 1996b), in Fig. 3a ( $D=716.3$ ;  $\bar{\delta}=0.45$ ). One realization of the algorithm to produce a permutation with a similar degree of aggregation is in Fig. 3b ( $D=716.8$ ;  $\bar{\delta}=0.34$ ); note the lack of correlation between the observed and permuted sets.

### *Counts, data from two species*

Another common source of data occurs when two populations are studied, with a count from both being available at each sample unit. The two populations may be spatially disassociated, as for an insect host in refuge from its parasitoid attacker, where the counts are negatively correlated; or they may be positively associated, as for diseased plants and their pathogen; or they may occur at random with respect to one another. Perry (1996b) described another algorithm to generate an arrangement of a given set of counts over predefined locations, at which there are known counts of a second set, with any desired level of association between the two. Such an algorithm is useful in simulation models that study species interactions or to generate data with a known structure to test competing indices and tests of association. Briefly, let  $K_{1i}$ ,  $i = 1, \dots, n$ , represent those counts of set one to be assigned;  $K_{2j}$ ,  $j = 1, \dots, n$ , those (ranked) counts from set two at known locations; and  $k$  ( $>0$ ) be an association parameter with values smaller (larger) than unity implying association (disassociation), and  $k=1$  indicating random placement of one set with respect to the other. Then the algorithm associates labels  $j$  with  $i$  so that, at any stage, the probability that the  $r$ th ranked count of the labels as yet unassigned of set two is associated with the highest ranked unassigned label of set one is  $k^{r-s}$  times that of the  $s$ th ranked count of the labels as yet unassigned of set two. The algorithm, in brief, is as follows. Step 1: rank  $K_{1i}$  and  $K_{2j}$ ; step 2: put  $t$  equal to the number of counts yet to be assigned; step 3: draw a uniform random number on (0,1); step 4: select the minimum integer  $w$  for which  $u < \sum_{j=1}^w k^{t-j} (k-1)/(k^t-1)$  for  $k \neq 1$ , or for which  $u < w/t$  if  $k=1$ , and associate the  $w$ th as yet unassigned count of set two with the highest ranked as yet unassigned count of set one. Decrease  $t$  by unity and, unless  $t$  is zero, return to step 2. For example, we may wish to simulate arrangements of the  $n=19$  observed counts of the parasitoid *Trichomalus perfectus* (set one) caught by A.K. Murchie (Perry, 1996b):  $\{0, 1^2, 2^3, 3, 4, 7, 8^2, 10^2, 12, 14, 22, 24, 32, 35\}$ , in the same traps as its weevil host *C. assimilis* (set two, Fig. 3a), to

Table 2. *Three simulated arrangements of nineteen T. perfectus counts showing varying degrees of association with observed counts of C. assimilis, in the same units (Perry, 1996b)*

X-coordinate of trap	77	75	69	75	73	71	81	73	67	79	77	81	71	79	73	75	77	79	70
Y-coordinate of trap	44	46	52	54	56	50	48	48	54	50	52	52	54	42	52	50	48	46	55
<i>C. assimilis</i>	102	90	73	70	64	58	53	51	45	39	33	29	25	22	16	15	15	12	1
<i>T. perfectus</i> ( $k=0.2$ )	32	35	24	12	22	14	10	8	4	10	8	7	3	2	1	2	2	1	0
<i>T. perfectus</i> ( $k=1.0$ )	3	2	14	2	7	1	12	8	0	24	32	35	22	10	8	2	4	10	1
<i>T. perfectus</i> ( $k=2.0$ )	1	2	2	0	1	3	2	8	10	8	4	10	7	22	32	14	12	24	35

demonstrate different degrees of association. The algorithm generated arrangements with the

desired degrees of association for the three chosen values of  $k$ , 0.2 (strong association), 1.0 (random pairing), and 2.0 (moderate disassociation), see Table 2. This algorithm is non-parametric, and non-spatial in the sense that it does not utilise the spatial information concerning the location of the counts of set two.

*Counts, single-species cellular automaton, movement dependent on scale of spatial pattern*

We now describe work done by one of us (EDB) towards his PhD thesis. Cellular automata models, in which the environment is tessellated into contiguous cells, each of which has one of several qualitative states that alters according to the interaction between itself and the states of its near neighbours, are now ubiquitous in ecology. Their recent use reflects the acknowledgement amongst theoretical ecologists that spatial effects may be important and their desire to model such effects within a truly spatial framework. Simultaneously, it has been realized that processes that operate at a given spatial scale may not operate similarly, or at all, at other scales. The processes of aggregation towards increased states of crowding, and of dispersal towards increased states of regularity, are fundamental in the modulation of spatial effects, and these too may be subject to differences according to spatial scale.

For example, suppose we begin with a widespread and reasonably dense population. When individuals aggregate they may initially do so by pairing with very close neighbours and then coalescing into groups of perhaps  $n < 6$ . In the next phase, these groups of five or fewer may combine predominantly as units with a correspondingly greater purview, rather than solely by the attraction of lone individuals, to yield assemblages of possibly  $5 < n < 20$ . It is easy to extend this hierarchical concept, and to propose that these larger-sized assemblages may attract and merge with others of similar and smaller sizes, detected over a still wider landscape, and that the resulting throngs themselves may unite at a yet grander spatial scale. This may be a plausible model for several possible agglomerative mechanisms within the animal kingdom, such as flocking in birds, schooling in fish, leks in various animals and rioting in humans. In practice, there may be several scales operating simultaneously, according to the size of the units locally. Indeed, the hierarchical definition of military units into platoons, companies, battalions, brigades, divisions and regiments, with the implied need for each level to move independently at its appropriate scale within the scale appropriate to the next-highest level, gives an example of such a process. Clearly, a similar concept, with splitting of groups by repulsion replacing their formation by attraction, might be proposed to model dispersal from a point source. We are not aware of any model which seeks to explicitly relate the degree of aggregation to the current scale of the spatial pattern for discrete counts. However, such a model may be easily structured as the following cellular automaton, where the state of a unit is defined by the discrete integer representing the count of the individuals within it.

The heterogeneous environment was modelled as an  $(2r+1) \times (2r+1)$  grid of squares, with each edge mapped to its opposite, giving a two-dimensional array on the surface of a torus. There were  $N = (2r+1)^2$  individuals, at an average density of unity, with no births or deaths. Initially, the individuals were distributed randomly between the cells. For the agglomerative phase, in each iteration each cell was considered sequentially, to determine whether any of its individuals would be attracted to a cell within a certain area around it, and, if so, to which. For iteration  $t$ , the area,  $A$ , surrounding the cell  $(a,b)$ , comprising cells to one of which some of the  $s(t)$  individuals within cell  $(a,b)$  could potentially be attracted, was a square grid of size  $(2m+1) \times (2m+1)$ , centred on that cell and denoted by  $A(a,b,t)$ . The scale over which the attraction was possible, measured by this local value of  $m = m(a,b,t)$ , was assumed to be dependent on  $N$ ; on  $r$ ; on  $M(A(a,b,t-1))$ , the maximum count in any cell within the corresponding area for the previous iteration; and on  $k$ , the only parameter of the model, through the following equation:  $m = \text{int}[q+0.5]$ , i.e.  $m$  is the nearest integer to  $q$ , where

Fig.4. Output from the cellular automaton model: (a) agglomerative phase, (b) dispersal phase

$q = \{(r+1)\exp(\theta)/[1+\exp(\theta)]\} - 0.5$ , and where  $\theta = k\log_e[(M+0.5)/(N+0.5-M)]$ . The parameter  $k$  controlled the rate of change of local scale with increase of local density. For most runs,  $r=50$ , and  $k$  was chosen to be 0.352, which yields, for example,  $m=3$  for a value of  $M=6$ . Having determined the local scale and, thereby,  $A(a,b,t)$ , the cell with the maximum count within  $A$  was identified, say  $(c,d)$ , and  $u$ , a randomly chosen integer with  $0 \leq u \leq \min[m(a,b,t),s]$  of the  $s$  individuals in  $(a,b)$  migrated one unit 'towards' that cell. By 'towards' we mean either towards it in the  $x$ -direction, with probability  $|a-c|/(|a-c| + |b-d|)$ , or towards it in the  $y$ -direction, with probability  $|b-d|/(|a-c| + |b-d|)$ , with the choice determined stochastically, and with due allowance in the formula for any cells on the edge of the co-ordinate system. Movement of all individuals took place simultaneously, at the end of an iteration. After sufficient iterations, between three and five clusters remained, the clusters occupied only single cells and the inter-cluster distances exceeded their range of attraction.

In the dispersive phase of the model, the process was similar, but operated in reverse, so that now, the cell with the minimum count within  $A$  was identified as  $(c,d)$ , and some of the individuals within  $(a,b)$  moved towards it. This resulted in the splitting of large clusters and the dispersive process continued until the maximum count over all the cells fell below some specified value, after which the system began agglomerating once again.

The model generated reasonably realistic maps. Clusters had approximately bivariate normal density contours on logarithmic scales. Agglomeration preceded dispersal in a continuing process, with at least three clusters drifting ceaselessly around the arena. Fig. 4a gives an example of a typical presence/absence map for  $r=25$ , towards the end of the agglomeration phase; Fig. 4b from towards the end of the dispersal phase. This work could be extended to vary the environmental conditions in each cell, and link the phase to the quality of the local environment. Unfortunately, there appear to be few data sets for model validation.

## Discussion

One of several extensions planned for the SADIE system is to develop a test and index for spatial association. Clearly, the measurement of the association between two sets of counts that share the same locations by a statistic such as a correlation coefficient, ignores the spatial information in the sample, and may therefore mislead. Consider the two sets of artificial counts of two species in the 5x5 grid shown in Table 3, where a blank entry denotes a zero count. In both sets, there is a single coincidence (row 3, column 3) of a non-zero entry for species 1 (counts in bold) with a non-zero entry for species 2 (counts in italics); hence the correlation coefficient (-0.0826,  $\log_{10}(n+1)$  scale) is identical for both sets, and non-significant. Also, in both sets, the same counts are used for species 1 and for species 2, and, for the former, in identical positions. Furthermore, the degree of aggregation for species 2 is identical for both sets, because their arrangements are identical save for a rotation about the central cell and a

Table 3. Sets of counts (a) and (b), demonstrating different degrees of association

Row 1 (a)	<b>3</b>	2	4	(b)		2	4			
Row 2		<b>4</b>	<b>2</b>							
Row 3	<b>11</b>		3	<b>3</b>	<b>11</b>		3			
Row 4				<b>4</b>						
Row 5				<b>2</b>						
	Col. 1	Col.2	Col.3	Col.4	Col.5	Col. 1	Col.2	Col.3	Col. 4	Col.5

reflection in the diagonal. However, visually, the species counts in set (a) clearly appear associated, and those in set (b) highly disassociated. Methods are required to detect such features for count data. Further, since the degree to which two species appear associated may depend on the spatial pattern of each in the absence of the other, it is clear that any method must condition on the observed aggregation of the species, separately. A possible method is as follows. The data from each set are scaled to have the same totals, and, separately, their values computed. The total count for the species combined are computed for each unit. The distance to regularity of this observed total, say  $T$ , is computed and stored. Permutations of the second set of (scaled) data are found with very similar values of  $D$  and  $\delta$  to those observed, and, for each permutation,  $k$ , the total is formed of this set and of the original set 1, as above, say  $T_{(1)2}$ . A randomization test and index are formed, as above for the single species case, from the observed  $T$  and the frequency distribution of the values of  $T_{(1)2}$ . The procedure is repeated, with sets one and two reversed, for the randomized values  $T_{(2)1}$ . This gave indications ( $P_{.05}$ ) of association for (a) and disassociation for (b). An alternative method makes use of comparison between the two species of the strength and the direction of flows such as those in the IAF plot for counts discussed above. Further details of both methods are in Perry (1996c).

### Acknowledgements

We thank B.Boag and A.K.Murchie for permission to use their data. IACR Rothamsted receives grant-aided support from BBSRC. EDB is supported by an BBSRC CASE Studentship.

### References

- Diggle P J. 1983.** *Statistical analysis of spatial point patterns*. Academic Press, London.
- Perry J N. 1995a.** Spatial aspects of animal and plant distribution in patchy farmland habitats. In *Ecology and Integrated Farming Systems*, pp. 221-242. Eds D M Glen, M P Greaves, H M Anderson. Chichester, England: Wiley.
- Perry J N. 1995b.** Spatial analysis by distance indices. *Journal of Animal Ecology* **64**: 303-314.
- Perry J N. 1996a.** Measuring the spatial pattern of animal counts with indices of crowding and regularity. *Ecological Monographs* (submitted).
- Perry J N. 1996b.** Simulating spatial patterns of counts in agriculture and ecology. *Computers and Electronics in Agriculture* (in press).
- Perry J N. 1996c.** Spatial association for counts of two species. *Acta Jutlandica* (submitted).
- Perry J N & Hewitt M. 1991.** A new index of aggregation for animal counts. *Biometrics* **47**: 1505-1518.

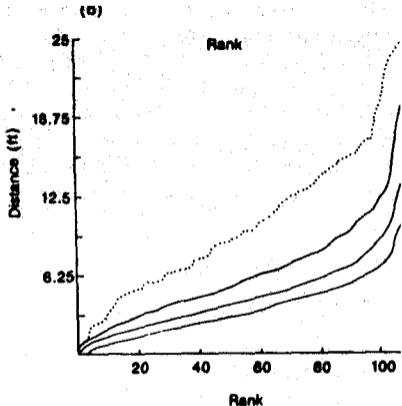
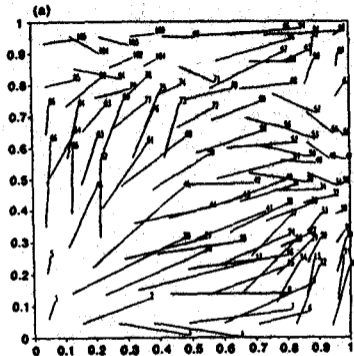


Fig. 1. Data for 105 beetle larvae (Perry 1995b): (a) an IAF plot and (b) an EDF plot.

0	0	0	1	0	0	1	1	1	0	1	0	2	0	10
0	0	0	0	1	0	2	1	7	2	0	0	1	0	0
0	0	0	3	11	3	1	0	28	16	0	0	1	31	22
0	0	0	2	9	39	24	9	24	11	1	0	0	0	16
0	0	0	0	15	3	15	12	46	7	0	1	0	0	0
0	0	0	0	7	19	18	3	2	2	0	0	0	0	5
0	0	0	0	4	1	0	1	0	0	6	0	2	3	2
0	0	0	0	13	3	3	0	2	0	0	0	0	0	0
0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
7	0	1	0	3	0	15	4	4	3	0	0	0	0	0
0	2	0	2	8	16	23	18	5	0	5	0	6	1	0
0	2	0	2	12	0	13	6	0	1	0	32	12	42	0
0	0	0	1	5	21	27	4	25	2	0	12	9	1	0
0	1	0	1	7	22	19	15	13	0	4	0	0	0	18

Fig.2. Counts of nematodes at 7.14m spacing, collected by B. Boag (Perry 1996a)

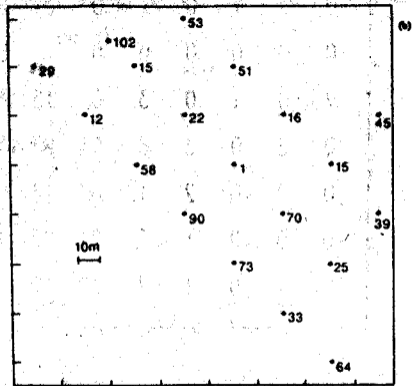
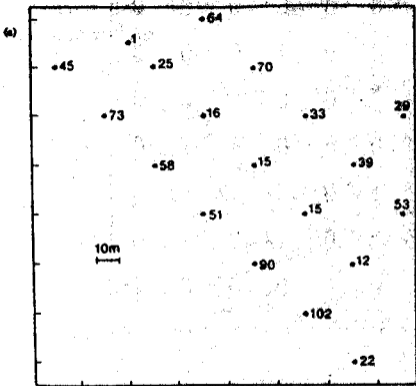


Fig.3. *C. assimilis* counts (Perry 1996b): (a) observed, (b) permutation with similar *D* value

(a)

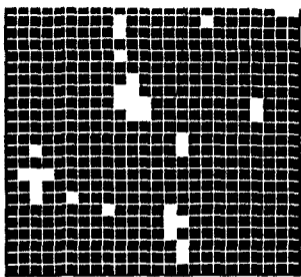
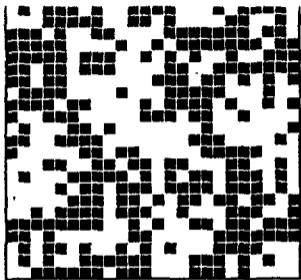


Fig. 4. Output from the cellular automaton model: (a) agglomerative phase, (b) dispersal phase.