Red-blue plots for detecting clusters in count data

Joe N. Perry[1], L. Winder[2], J.M. Holland[3] & R.D. Alston[4]

[1]*Department of Entomology & Nematology, IACR - Rothamsted Experimental Station, Harpenden, Herts. AL5 2JQ, U.K.*

E-mail: joe.perry@bbsrc.ac.uk

[2]*Department of Land Use and Rural Management, Seale-Hayne Faculty, University of Plymouth, Newton Abbot, Devon TQ12 6NQ, U.K.*

[3]*The Game Conservancy Trust, Fordingbridge, Hampshire, SP6 1EF, U.K.*

[4]*Cancer Research Group, Royal Manchester Children's Hospital, Manchester, U.K.*

SHORT RUNNING TITLE: Red-blue plots to detect clusters

KEYWORDS: Counts, Spatial pattern, Clusters, Aggregation, Patchiness, Heterogeneity, Biometry, Aphids, Population density, SADIE

CORRESPONDENCE TO: Professor Joe N. Perry, Dept. Entomology & Nematology, Rothamsted Experimental Station, Harpenden, Herts. AL5 2JQ, UK
email: joe.perry@bbsrc.ac.uk
Fax: +44 1582 760981
Phone: +44 1582 763133 (extension 2375)

BIOSKETCH: Joe Perry has worked at Rothamsted Experimental Station for 22 years, firstly in the Department of Statistics, then in Entomology & Nematology. His research interests cross these disciplines; he is particularly interested in spatial pattern and population modelling.

ABSTRACT

A new index and four new graphical displays, termed ‘red-blue’ plots, are presented to study and measure clustering in spatially-referenced count data. The index can detect clusters in the form of patches, comprising several nearby large counts, and in the form of gaps, comprising several nearby small counts. The new methods quantify the degree to which the count for each sample unit contributes towards the overall degree of clustering, either as part of a patch or a gap; provide tests of non-randomness to detect clustering; and facilitate a comprehensive definition of the size and dimension of a cluster. The methods are illustrated using aphid field data.

KEYWORDS: Counts, Spatial pattern, Clusters, Aggregation, Patchiness, Heterogeneity, Biometry, Aphids, Population density, SADIE

INTRODUCTION

This paper concerns spatial pattern in count data and introduces new methods to detect and measure clusters. These methods are linked to graphical displays that benefit from two-colour presentation, and are thus termed 'red-blue plots'; however, only black-and-white versions are presented here. The term 'cluster' is used to mean a region of either relatively large counts close to one another in two-dimensional space (i.e. a patch), or of relatively small counts (i.e. a gap). Much ecological data concerns the number of plants within specified areas, or of animals caught in traps; in such samples data are recorded in the form of integer counts. The importance of heterogeneity, or non-randomness, in ecological systems that generate such data has long been recognised in models (Hassell & May 1973) and empirical descriptions (Taylor *et al.* 1978). Traditionally, this has been quantified by methods such as the relationship between the variance and mean of the counts (Lloyd 1967; Pacala & Hassell 1991; Clark *et al.* 1996). These operate solely on the list of counts and relate only to the numeric properties of the underlying frequency distribution. Since no utilization is made of any additional spatial information in the sample, their capacity to describe spatial pattern is very limited and they can only infer non-randomness at some unknown spatial scale below that at which the data was collected. The method of Greig-Smith (1952) is an improvement, but even this do not explicitly use the coordinates of the sample units, and is restricted to data sampled on a regular grid.

However, it is realized increasingly that the location of individuals in relation to one another may be critical for the outcome of ecological interactions (Hassell *et al.* 1991; Silvertown et al. 1992; Powell *et al.* 1995; Grenfell & Bolker 1998). Count data may be spatially referenced, when the position of each sample unit is recorded in two dimensions. Recently, the SADIE system was introduced (Perry *et al.* 1996; Perry 1998a) to provide a means of describing the spatial features of such a set of counts, independently of their numeric properties. The method works through equating the degree of spatial pattern in an observed arrangement of counts to the minimum effort that the individuals in the population would need to expend to move to a completely regular arrangement in which abundance was equal in each sample unit. In practice, this effort is equated with the minimum distance, D, required to move to complete regularity; the value of D is provided by the Transportation algorithm from the linear programming literature (Kennington & Helgason 1980). This algorithm provides a unique solution based on notional 'flows' of individuals and fractions of individuals, from 'donor' sample units, with greater than average abundance, to 'receiver' units, with less than average abundance. The flows may be depicted graphically in a red-blue 'initial-and-final' plot that aids visual interpretation of the major spatial features of the data (Perry 1998b). The spatial pattern is quantified by permuting the observed set of counts amongst the sample units. This provides data for the null-hypothesis that the counts are arranged randomly with respect to one another, while maintaining precisely the numeric properties of the data, which may be investigated separately. In particular, division of the observed value of D by the mean value from several hundred such randomizations gives an index of aggregation, I_a; values of $I_a = 1$ indicate randomly arranged counts, while $I_a > 1$ indicates aggregation of observed counts into clusters.

Furthermore, comparison of the observed value of D, with the tails of the distribution of corresponding values from these permutations, leads to the usual formal randomization test (Besag & Diggle 1977) of the null hypothesis that the observed counts are arranged randomly. The example in this paper relates to a rectangular grid of sample units, but this is not a requirement of any of the SADIE methods. Sample units may be located anywhere, without restriction, in two dimensions.

Although the SADIE technique currently provides an overall measure of spatial aggregation, the single index, I_a, does not encompass all the facets of spatial pattern in an arrangement. In particular, it does not provide evidence to infer whether a particular, large count represents some isolated, small subpopulation, or alternatively, whether it is part of a patch to which neighbouring relatively large counts also contribute. Further, it does not answer whether, and to what degree, a group of nearby zero or small integer catches form a true gap that might indicate poorer than usual habitat, or whether their proximity to one another is no more than expected by chance, given the list of counts observed. This is the problem addressed by this paper. Previous work recognised the importance of the initial-and-final plot visually, but failed to utilise all its available information (Perry 1998b). This information is used here to: quantify the degree to which the count for each sample unit contributes towards the overall degree of clustering, either as part of a patch or a gap; provide tests to detect the occurrence of clustering; and to facilitate a more formal definition of the size and dimension of clusters than that given previously. The methods are illustrated using aphid field data. The methods lead to four new 'red-blue' plots; together with the existing initial-and-final plot these five red-blue plots are intended to provide a comprehensive suite of visual tools to study clustering.

METHODS

Data

The data are counts of 554 individuals of the cereal aphid, *Sitobion avenae* (F.), sampled on 28 June 1996 in a 250m x 180m field of winter wheat near Wimborne, Dorset, UK (Winder *et al.* 1998). Total aphid abundance was recorded at each of 63 sample units by the inspection of five tillers; the units were located on a 9x7 rectangular grid at intervals of 30m (Fig. 1). The sample mean and variance were 8.8 and 286.4, respectively. Counts were greatest in regions of relatively large values of both x and y.

The initial-and-final plot

The initial-and-final plot for these data (Fig. 2) shows the outflows as directed lines, originating from donor units denoted by filled circles, and terminating at receiver units denoted by open circles. The impression given is of a trend, from larger counts (donor units) in the north-east towards smaller counts (receiver units) in the south-west of the sampled area. To maintain visual clarity, the strength of each flow is not marked on the plot.

Note also that there may be flows to several receiver units from each donor unit as well as flows from several donor units to each receiver unit (Perry 1998a,b). For example, the sample unit labelled A (Figs. 1,2) at (150,120) has a count of 27, so is a donor unit; also, since the mean abundance for the set is 8.79, the excess flow for unit A is (27 - 8.79) = 18.21. The magnitude of the excess flow contains no spatial information of itself; it is a function purely of the magnitude of the count, so is a numeric property of the data. However, the Transportation algorithm partitions this excess flow spatially, between four receiver units (Figs. 1,2) as follows: a flow of 5.70 to the unit B at (0,0), 2.92 to C at (30,30), 0.80 to D at (90,60) and 8.79 to E at (120,90). For each unit partitioned there are two items of unused spatial information; these form the basis of the new red-blue techniques.

Average flow distance

The first item of information concerns the average distance of flow associated with each unit. This is of interest because, for units within a cluster, average distances of flow tend to be greater than for units in regions where similar counts are randomly distributed. The methodology works towards using this information to provide a dimensionless index of clustering for each unit.

Limited notation follows. For donor unit i, at position (x_i , y_i), the outflow to the jth of n_i receiver units, $j = 1 ,\ldots, n_i$, at position (x_j , y_j), is denoted as v_{ij} . The distance of this flow, $[(x_i - x_j)^2 + (y_i - y_j)^2]^{1/2}$, is denoted as d_{ij}. The average distance of outflow from unit i, weighted by the magnitude of each individual flow, is denoted as Y_i, where $Y_i = \Sigma_j d_{ij} v_{ij} / \Sigma_j v_{ij}$. As an example, distances d_{ij} from the donor unit A at (150,120) in Fig. 2, to its four receiver units B,C,D,E, were, respectively, 192.1, 150.0, 84.9 and 42.4. The weighted average flow distance is then computed as $Y_i = 108.4$, using the magnitude of each of the four flows given above. In exactly similar fashion, there is an average distance of inflow associated with each receiver unit j; by convention the values of d_{ji}, and hence of Y_j, are taken to be negative. For units within groups of relatively large counts close to one another in two-dimensional space, i.e. for units within a patch, values of Y_i tend to be relatively large, compared to a unit with an isolated large count. Furthermore, for units within groups of relatively small counts close to one another, i.e. within 'gaps', values of Y_j tend to be relatively large and negative. Hence, clustering is indicated by unexpectedly large values of Y_i or of $|Y_j|$. However, because an average flow distance for a unit depends on the measurement units of the co-ordinate system adopted, on the count at that sample unit, and on the distances from that unit to others nearby, Y requires some scaling before it may be used profitably. This scaling aims to produce a dimensionless index of clustering, ν, that may be used to compare between different units and between different sets of data.

Consider the randomizations described above, in the kth of which the coordinates of the count 27 are altered from their observed values, (150,120), to values permuted from the set of sample unit locations. The value of the average distance of outflow for that count under that randomization, $Y_{k(i)}$ may be computed and stored. The mean value of $Y_{k(i)}$ may then be computed for that same count over all the randomizations; for the

count *c* this value is denoted as ${}_{c}Y$. Essentially, ${}_{c}Y$ is the expected value of the average flow distance for the observed count, *c*, under the assumption of a random arrangement of the observed counts among the observed sample units. As an example, ${}_{27}Y$, the average distance of outflow under randomizations for the count 27 observed originally in sample unit A, was 50.87, considerably less than the observed value of 108.4. In a similar fashion, instead of following the count through its randomizations to other units, the absolute values of the average flow distances of all the counts that were randomized to the sample unit A at (150,120) could be followed and averaged; for the unit *i* this is denoted as ${}_{i}Y$. Hence, ${}_{i}Y$ is the expected value of the average absolute flow distance for the *i*th unit, again under the assumption of a random arrangement of the observed counts among the observed sample units. As an example, ${}_{A}Y$, the average distance of outflow under randomizations for the sample unit A at (150,120), was 62.23, again considerably less than the observed value of 108.4. Furthermore, the average absolute value of ${}_{c}Y$ over all the counts, *c*, which is equal to the average value of ${}_{i}Y$ over all the units, *i*, is denoted ${}_{o}Y$. Thus, ${}_{o}Y$ is the expected value of the overall average absolute distance of flow for all points and counts in the randomizations. For these data the value of ${}_{o}Y$ was 60.44.

For outflows, a standardized and dimensionless index of clustering, ν_i, is then given by:

$$\nu_i = Y_i \, {}_{o}Y / {}_{i}Y \, {}_{c}Y$$

For inflows, ν_j is defined similarly, again with the convention that it is negative in sign. The index, ν_i, measures the degree to which the unit contributes to clustering, as a member of a group of donor units that constitute a patch. For the count of 27 in sample unit A, ν_i was computed as: (108.4×60.44)/(50.87×62.23) = 2.07, so the degree of clustering for this unit was about double that expected by chance. As a general rule, large values of ν_i, > *c*.1.5, indicate patchiness; large negative values of ν_j, < *c*. –1.5, indicate membership of a gap; values close to unity indicate a random placement of that unit in relation to others nearby.

In fact, the value of ν_i = 2.07 for unit A was the largest of the 17 donor units, which ranged from ν_i = 0.64, the value for the unit with count of 23 at (150,240). In the observed arrangement, clustering in gaps was much more apparent than in patches. Hence, seventeen of the 46 receiver units had values of ν_j negatively greater than -1.5; values ranged from ν_j = -0.44, for the zero count at (180,240), up to ν_j = -3.48, for the count of 5 at unit G (Fig. 1) at (0,30). Note that the allowance for the properties of both count and unit in the construction of the index implies that the largest absolute values of Y do not always correspond to the largest absolute values of ν. Thus, the largest observed value of Y_i was 126.9 for the count of 102 at (180,150), but the standardization decreased the value of ν_i for this unit by a relatively large amount to 1.54. When plotted on a map of the sample units, the values of the indices ν_i and ν_j indicate visually the location and extent of clusters in the data (Fig. 3). Since they are correlated with each other and are on a continuous scale, the values of the indices may be contoured. The first of the new red-blue plots is a contour map showing both the positive and negative indices of clustering (Fig. 3). Clearly, for these data patchiness is at best moderate and the observed arrangement is dominated by a large gap extending down almost the entire left-hand portion of the field around *x*=30 and *x*=60. The area for which ν_j < -1.5 totals 11,172m^2; the four separate areas for which ν_i > 1.5 total

less than only 100m^2. As required, ν_i and ν_j encapsulate spatial and not numeric information; the required lack of relationship between the magnitude of a count and the degree of clustering may be seen by, for example, comparison of the counts 27 at unit A and 20 at unit F (Figs. 1,3). The former's proximity to the other large counts 102, 21, 14 and 9 ensure that its index value indicates membership of a small patch; the latter has a small index of clustering of $\nu_i = 0.74$, reflecting its relative isolation amongst nearby small counts. Alternatively, compare the count of 5 located at unit G (Fig. 1), the heart of the large gap referred to above (Fig. 3), and which has the most extreme negative value recorded, of $\nu_j = -3.48$, with the count of 0 at unit H (Fig. 1). Although a smaller count, the latter has a small value of $\nu_j = -0.51$, reflecting its proximity to several units with relatively large counts.

The heuristic threshold of 1.5 referred to above cannot be used for a formal test of the degree of clustering. There are two complications. Firstly, all the values of ν_i and of ν_j are correlated with each other, and a test should not focus on a single individual value; secondly, the expected magnitude of the extreme values of ν_i and of ν_j depend on the number of sample units in the dataset. However, a rough visual guide to significance that addresses both problems is provided by means of the so-called 'empirical distribution function' plot (Diggle 1983; Perry 1995). In this, the second of the new red-blue plots, each observed value of Y_i is ranked and then plotted against the expected value of Y_i for that same rank, i.e. the mean obtained from the randomizations, as well as against the 95th centile of the distribution of Y_i for that rank, again obtained from the randomizations (Fig. 4). Diggle (1983) emphasises the use of such a plot as a visual aid to interpretation. Firstly, comparison of the observed line with that expected from a random arrangement of the counts allows an assessment of the overall degree of patchiness. Secondly, exceedence by the observed line of the envelope of 95th centiles at any value is an initial and informal indication of the presence of patchiness, beyond what would be expected from random variation. Thirdly, a plot of all the units guards against excessive concentration on the highest-ranked values, which often approach less closely to the envelope than do middle-ranking values. A similar plot is used for the values of Y_j to obtain the third of the new red-blue plots, but here the observed line must be negatively greater than the 5th centile envelope to demonstrate strongly the presence of gaps (Fig. 5). For these observed data, the line of Y_i values for the 17 observed donor units is well above the expected values for most of its length (Fig. 4), but never exceeds the envelope of 95th centiles, so the data indicates moderate patchiness but the evidence is not overwhelming. Again, compare (Fig. 4) the fifth-smallest value of Y_i, for unit F, slightly below the average for randomizations, with the second-largest value, for unit A, one of the closest to the upper 5th centile envelope. By contrast, several of the 46 observed Y_j values negatively exceed their corresponding 5th centiles (for example, unit B, Fig. 5), and the presence of the large gap is demonstrated more convincingly. Furthermore, as stated above, it is mainly the values spanning the inter-quartile range that exceed the envelope, indicating that the gap is not manifested in a very few small counts, but is considerable in extent.

A more formal test of overall clustering may be provided by comparing $\bar{\nu}_i$, the average value of ν_i over all 17 outflows, and $\bar{\nu}_j$, the average value for all 46 inflows, with their corresponding values from randomizations. For example, $\bar{\nu}_i$ for the observed data was 1.15 and for the average of 1950 randomizations was 0.996, close to its expected value of unity. Of the 1950 randomizations, 296 had total patchiness, as measured by $\bar{\nu}_i$, that exceeded the observed value of 1.15 (P = 0.15); this confirms the diagnosis of moderate patchiness, not significantly greater than expected by chance. By contrast, the observed value of $\bar{\nu}_j$ was -1.34, compared to the average of the randomizations which was -0.998, again close to its expected value of -1. Only 82 randomizations negatively exceeded the observed value of $\bar{\nu}_j$ = -1.34 (P = 0.042), formally confirming the presence of strong and significant clustering into gaps. This separation of clustering inferences into those relating to patches and those relating to gaps provides more information than I_a, the overall index of aggregation derived previously (Perry, 1998b), which for these data was 1.28 (P = 0.0575).

Vector flow direction

The second item of previously unused information from the initial-and-final plot is the resultant of the vector direction of flow for each unit. Since there may be several outflows from or inflows to a unit, it is possible, for each unit, to define a vector flow, the resultant direction of which is the average direction weighted by the magnitude of the individual flow components. The vector strength of flow has little relevance here, being a numeric property at least partly dependent on the magnitude of the count in a particular unit. Instead, the resultant direction of the vector is the component that expresses spatial information. The vector concept has already been used in the definition of an index of spatial association (Perry 1998a), where the resultant direction for unit i was denoted as θ_i, and measured clockwise from the vertical direction (y increasing). As examples, the four outflows from unit A with the count of 27 (Fig. 2) all go roughly in the south-west direction (y and x both decreasing), and the resultant direction is θ_i = 230.3. For the receiver unit, G, with count of 5 (Fig. 2) there is just one inflow from the unit at (180,150) roughly north-east of G (i.e. a flow with direction towards approximately the south-west) and the resultant direction is θ_j = 236.3. The fourth new red-blue plot results when these directions are plotted (Fig. 6) for each unit, for vectors of identical length, where unit locations are denoted by circles, the direction of an outflow is indicated by a filled arrowed line originating from a filled circle, and the direction of an inflow by an open arrowed line terminating at an open circle. Such a vector flow plot gives a good visual impression of the major directions of flow and the relationships between outflow and inflow. Spatial pattern is demonstrated by systematic correlation between the directions of nearby units. For example, a single patch in an area of otherwise homogeneous density would show filled arrowed lines radiating out from the cluster in successively incremented angles, resembling the spokes of a wheel. When nearby units within the same cluster are alike in direction, then a local trend may be indicated, especially if units of opposite type show a similar direction, as for these data (Fig. 6). For example, notice the similarity in vector flow direction between donor unit A and receiver units B,C,D,E, and G, and neighbouring receiver units (Fig. 6). In

these plots, absence of spatial pattern is indicated by haphazard directions, and areas where inflow and outflow directions appear unmatched. Obviously, no outflow on the edge of the sampled area can possibly point away from O, the centroid (Fig. 1) of the sampled area, and likewise no inflow on the edge can point towards it; hence caution is required in interpreting directions of units close to the boundary of the sampled area. One advantage of the vector plot over the initial-and-final plot lies in its avoidance of a multiplicity of lines.

DISCUSSION

Using the techniques in this paper clusters may be defined, using various levels of threshold index values, and contour methods may be used to define their exact dimensions and area covered. A key improvement on previous methods is the ability to detect gaps of below-average density, as well as patches of above-average density. Also, the dimensionless indices of clustering, ν_i and ν_j, may be used to compare informally different units within the same set of data, while their overall average values, $\bar{\nu}_i$ and $\bar{\nu}_j$, may be used to formally test for clustering within one set, and compare informally different sets of data. Under certain circumstances, this measurement of patchiness can be much more powerful at detecting non-randomness than the overall index I_a; this is certainly the case for data with edge effects when relatively large counts range around all the boundaries of the sampled area. By contrast with the formal tests, the visual aids, especially the contour map of clustering (Fig. 3) and the vector flow plot (Fig. 6), present visual information that must be interpreted more intuitively.

The formulation of the unit clustering indices, ν_i and ν_j, given here is not the only one possible, although it remains the best found to date. Ideally, for each unit the observed value of Y_i, or of Y_j, should be compared with the average value from randomizations in which all counts except that being considered are permuted. However, for samples of size n this would entail of the order of n^3 simulations, which requires excessive computing time at present, even for small n. Essentially, the index proposed here incorporates a sensible estimate of what that average value would be, and despite very occasional minor anomalies seems to work well. Crucially, it gives expected values around unity under randomizations, as required. Likewise, use of $\bar{\nu}_i$, or of $\bar{\nu}_j$, as an overall clustering index is not the only possible formulation. For example, instead of averaging over all donor units to compute $\bar{\nu}_i$, a restriction could be imposed to include only those values with $\nu_i > 1$, or only those with $\nu_i > 1.5$. Note also that the choice of 1.5 as a absolute value indicating clustering and for contour intervals is entirely arbitrary; other smaller or larger values could equally well be used.

Alternative treatments of the raw data may also be possible. For example, a non-parametric approach, in which information about the actual value of the count is discarded and its rank used instead might prove useful. Although the Transportation algorithm demands that input to the software be in the form of integer counts, if the raw data are abundances on a continuous scale they could be scaled up to integer format for an approximate but informative analysis. In this way, as long as sample sizes were sufficiently large, proportions might be analysed, without the need for complex restricted randomization software to be written for each case.

Then, epidemiological data, for example, concerning incidences of disease per head of population, might be tested for clustering, prior to seeking relationships with environmental hazards.

It must be reiterated that while the example in this paper relates to a rectangular grid of sample units, this is not a requirement of the method. Sample units may be located anywhere in two dimensions, so data concerning, say, light-traps in Great Britain or cases of infectious diseases in UK cities may be analysed easily.

Count data are ubiquitous, but contain less information than comparable mapped data where the coordinates of each individual are known. Specifically, there is no information concerning spatial pattern at any scale (Schneider 1994) below that of the area represented by the count at each sample unit. The existence of non-randomness at lower scales may be inferred from numeric techniques such as variance-to-mean ratios but the spatial form of this non-randomness cannot be elucidated. Spatial inferences must relate to scales at or above those of the sample units, and this is why SADIE conditions on the observed counts for its randomizations.

Geostatistical approaches (Liebhold *et al.* 1993), developed for variables measured on continuous scales that display a stationary, stable covariance structure over a wide area may not apply well to spatial pattern analysis for animal counts, unless the data represent very large population densities. By contrast, ecological count data usually represent discrete variables distributed exceedingly patchily, frequently comprise a majority of zero values, are often found in isolated clusters and relate to organisms that have usually evolved to shift ceaselessly in space and time, often within dynamic metapopulations. Also, geostatistical techniques are usually directed towards the estimation of interpolated values through the construction of an abundance surface. By contrast, the techniques described here are more concerned with measuring and detecting the degree of non-randomness in spatial patterns; they deliberately distinguish between density and clustering, and therefore discount the isolated large data value that might be crucial in a geostatistical analysis. The purpose of a search for clustering in ecology and biology is usually as a preliminary exercise to uncover areas worthy of further study.

Software to analyse data using the techniques in this paper is freely available from the first-named author.

ACKNOWLEDGEMENTS

IACR-Rothamsted receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom. The authors thank Mr A. Sanders of Crichel Farms Ltd for generously providing the field site and the many staff and students of The Game Conservancy Trust and Bournemouth University who helped with data collection.

REFERENCES

Besag, J. & Diggle, P.J. (1977). Simple Monte Carlo tests for spatial pattern. *Applied Statistics*, 26, 327-333.

Clark, S.J., Perry, J.N. & Marshall, E.J.P. (1996). Estimating Taylor's power law for weed species and the effect of spatial scale. *Weed Research*, 36, 405-417.

Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns.* Academic Press, London.

Greig-Smith, P. (1952). The use of random and contiguous quadrats in the study of the structure of plant communities. *Annals of Botany*, 16**,** 293-316.

Grenfell B.T. & Bolker, B.M. (1998). Cities and villages: infection hierarchies in a measles metapopulation. *Ecology Letters*, 1, 63-70.

Hassell, M.P, Comins, H. & May, R.M. (1991). Spatial structure and chaos in insect population dynamics. *Nature*, 353, 255-258.

Hassell, M.P & May, R.M. (1973). Stability in insect host-parasite models. *J. Anim. Ecol.*, 42, 693-726.

Kennington, J.L., & R.V. Helgason, R.V. (1980). *Algorithms for Network Programming.* Wiley, New York.

Liebhold, A.M., Rossi, R.E. & Kemp, W.P. (1993). Geostatistics and geographic information systems in applied insect ecology. *Ann. Rev. Entomol.*, 38, 303-327.

Lloyd, M. (1967). 'Mean crowding'. *J. Anim. Ecol.*, 36, 1-30.

Pacala, S.W. & Hassell, M.P. (1991). The persistence of host-parasitoid associations in patchy environments. II. Evaluation of field data. *Am. Nat*., 138, 584-605.

Perry, J.N. (1995). Spatial analysis by distance indices. *J. Anim. Ecol.*, 64, 303-314.

Perry, J.N. (1998a). Measures of spatial pattern and spatial association for counts of insects. In *Population and Community Ecology for Insect Management and Conservation.* ed. J. Baumgartner, P. Brandmayr & B.F.J. Manly: Balkema, Rotterdam, pp. 21-33

Perry, J.N. (1998b). Measures of spatial pattern for counts. *Ecology*, 79, 1008-1017.

Perry, J.N., Bell, E.D., Smith, R.H. & Woiwod, I.P. (1996). SADIE: software to measure and model spatial pattern. *Aspects of Applied Biology*, 46, 95 - 102.

Powell, W., Hawthorne, A., Hemptinne, J.-L. *et al.* (1995). Within-field spatial heterogeneity of arthropod predators and parasitoids. *Acta Jutlandica*, 70, 235-242.

Schneider, D.C. (1994) Quantitative Ecology - Spatial and Temporal Scaling. Academic Press, New York.

Silvertown, J., Holtier, S., Johnson, J. & Dale, P. (1992). Cellular automaton models of interspecific competition for space - the effect of pattern on process. *J. Ecol.*, 80, 527-534.

Taylor, L.R., Woiwod, I.P. & Perry, J.N. (1978). The density-dependence of spatial behaviour and the rarity of randomness. *J. Anim. Ecol.*, 47, 383-406.

Winder, L., Holland, J.M. & Perry, J.N. (1998). The within-field spatial and temporal distribution of the grain aphid (*Sitobion avenae*) in winter wheat. In: *Proceedings of the 1998 Brighton Conference: Pests.* BCPC, Farnham, Surrey (in press).

LEGENDS TO FIGURES

Figure 1 Counts of *Sitobion avenae* sampled over 63 sample units in a 9x7 rectangular grid at intervals of 30m in a field of winter wheat. The centroid of the counts at (x=126.6, y=137.0), denoted by an asterisk, was displaced about 40.4 m from unit O, the centroid of the sample units, at (90,120). Axes show distances in metres. Other labelled units referred to in text.

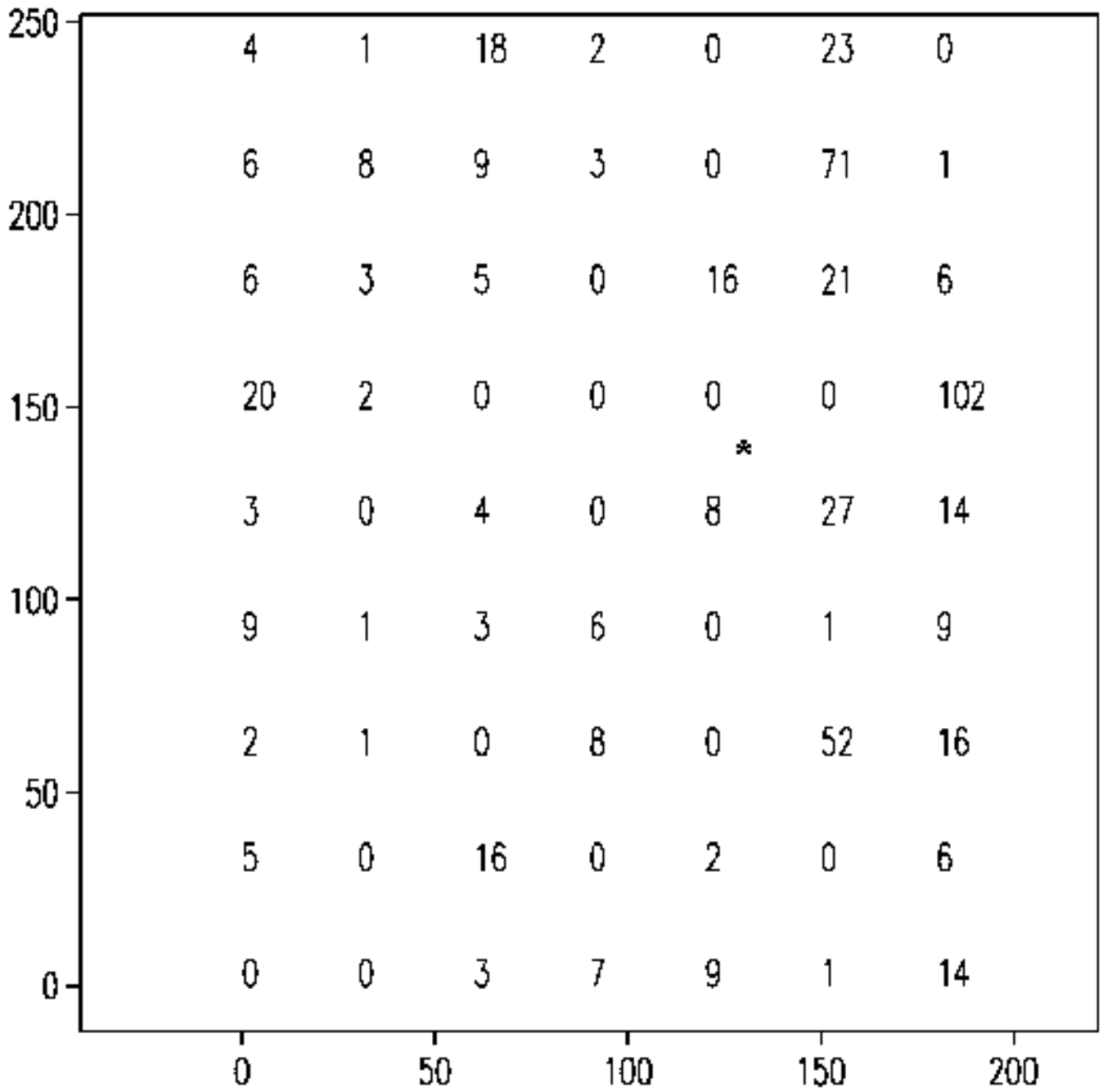
Figure 2 Initial-and-final plot for the data in Fig. 1 showing notional outflows, as directed lines, originating from donor units with above-average counts, denoted by filled circles, and terminating at receiver units, with below-average counts, denoted by open circles. Flows were computed by the Transportation algorithm, as implemented by the SADIE technique. Four outflows from unit A go in similar directions, at decreasing distances, to units B,C,D,E, respectively. The longest distance shown, 216.3m, is the inflow to unit G; the shortest is exactly 30m, for example each of the three outflows from unit F, or one of the two inflows to unit H. Axes show distances in metres.

Figure 3 Map of clustering in the data of Fig. 1. Values of clustering indices $\nu_i<1$ (below expectation), $1<\nu_i<1.5$ (slightly above expectation), and $1.5<\nu_i$ (well above expectation), shown as small, medium and large filled circles, respectively; values of $|\nu_j|$ are categorized similarly, but as open circles. Bold lines are contours enclosing patches (small dark-shaded areas) of $1.5<\nu_i$, or gaps (large lighter-shaded area) of $\nu_j<-1.5$. Contours were interpolated between sample units by the package SURFER. Areas within contours with absolute value greater than 1.5 indicate strong clustering; for example, the large gap on the left of the map contains 17 of the 63 sample units. By contrast, only four donor units on the right of the map have an index of clustering that exceeds 1.5, and patchiness is not extensive. Dashed lines are zero-value contours, representing the interpolated boundary between a patch and a gap, i.e. regions where the interpolated count is close to the sample mean, here 8.8. The long dashed line towards the right of the map confirms the impression from this plot and from Figs. 1, 2 and 6 that the dominant pattern is a trend from north-east to south-west. Axes show distances in metres.

Figure 4 Empirical distribution function plot of ranked average outflow distances (in metres), Y_i, plotted against the rank of each sample unit, for the 17 donor units with above-average counts from the data of Fig. 1. Observed values denoted as $\times$; mean value of Y_i from the 1950 randomizations as $+$; envelope of upper 95th centiles as ∇. Observed values for the two sample units labelled as A and F, that show very different flow distances in Fig. 2, are arrowed, as examples. Observed values generally exceed means from randomizations; they approach but never exceed envelope, indicating moderate patchiness, not inconsistent with random variation.

Figure 5 Empirical distribution function plot of ranked average inflow distances (negative by convention, and in metres), Y_j, plotted against the rank of each sample unit, for the 46 receiver units with below-average counts from the data of Fig. 1. Observed values denoted as ×; mean value of Y_i from the 1950 randomizations as +; envelope of lower 5th centiles as Δ. Observed values for the three sample units labelled as B, G and H, that show very different flow distances in Fig. 2, are arrowed, as examples. Almost all observed values negatively exceed means from randomizations. Several in the inter-quartile range also negatively exceed the envelope, providing strong evidence of a large gap.

Figure 6. Vector flow plot shows the resultant vector directions of the flows represented in Fig. 2. Vectors shown scaled to identical length. Locations of sample units denoted by circles; the direction of an outflow from donor unit is indicated by a filled arrowed line originating from a filled circle; direction of an inflow to receiver unit indicated by an open arrowed line terminating at an open circle. Spatial pattern is demonstrated by systematic correlation between the directions of nearby units. The trend from north-east to south-west is confirmed by similarity in direction between nearby units of opposite type, exemplified here by donor unit A and receiver units B,C,D,E, and G. Absence of spatial pattern would be indicated by haphazard directions, and areas where inflow and outflow directions appear unmatched.


250
200
150
100
50
0
4 1 18 2 0 23 0
6 8 9 3 0 71 1
6 3 5 0 16 21 6
20 2 0 0 0 0 102
*
3 0 4 0 8 27 14
9 1 3 6 0 1 9
2 1 0 8 0 52 16
5 0 16 0 2 0 6
0 0 3 7 9 1 14
0
50
100
150
200

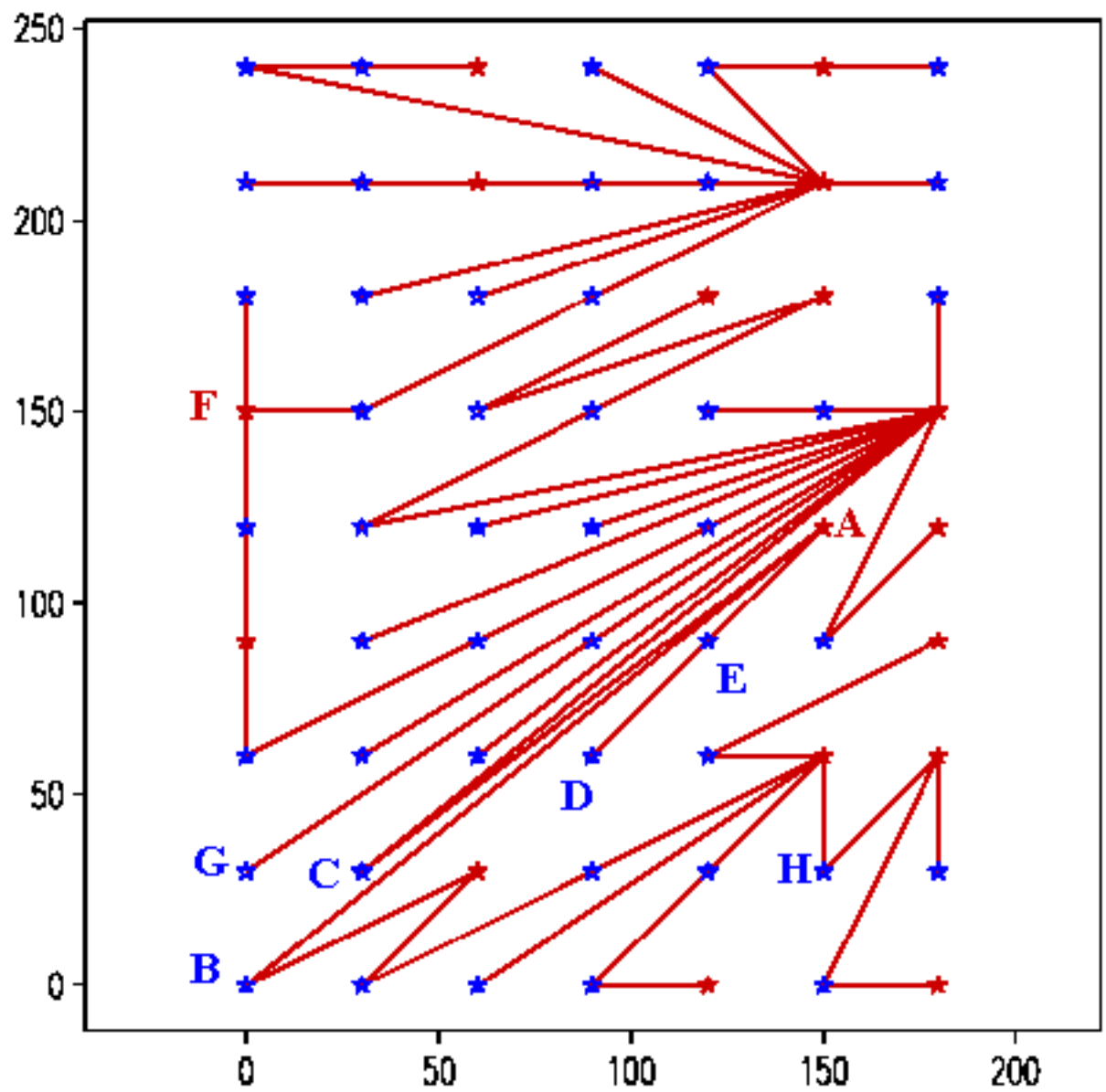
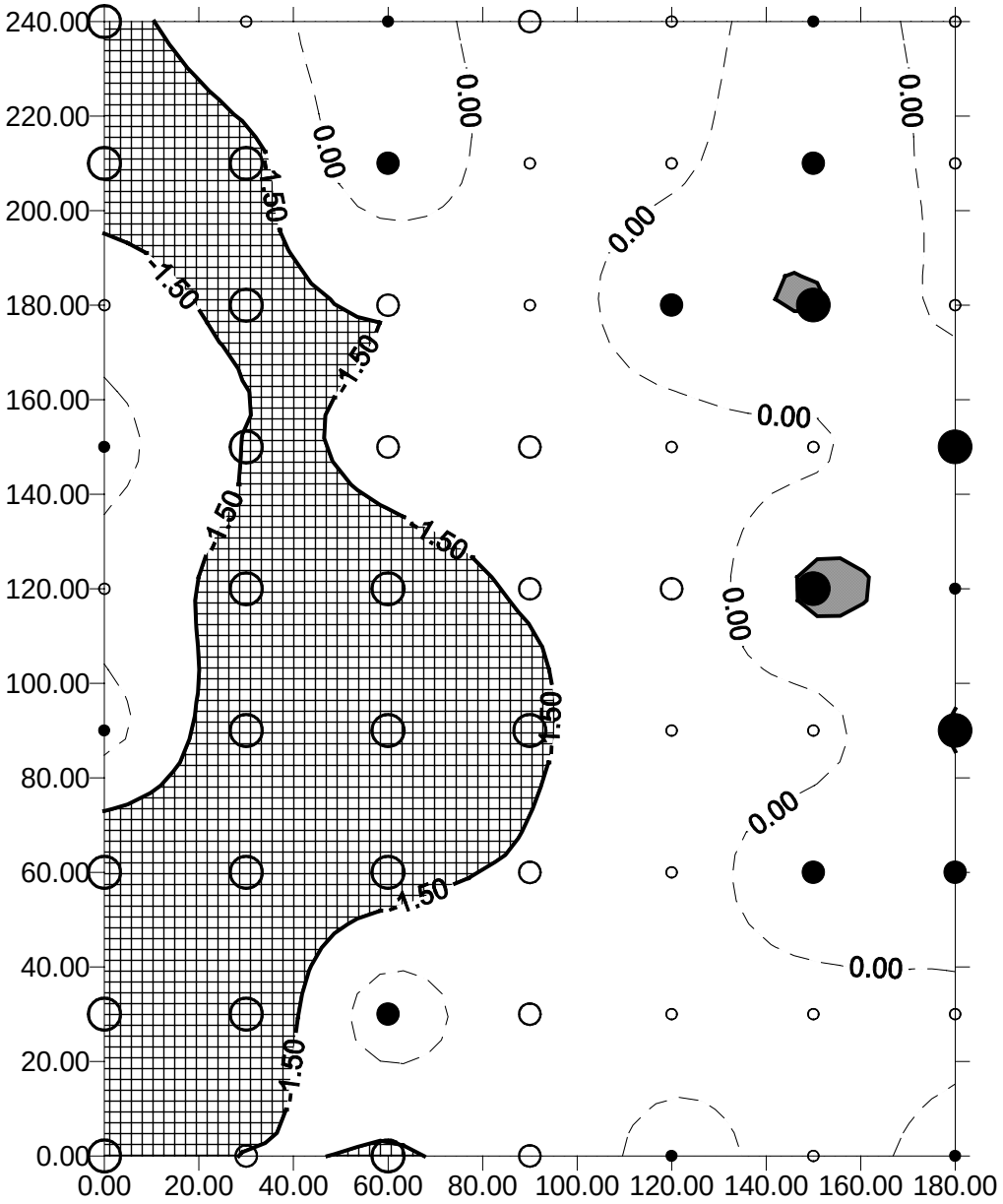


Figure 2 Initial-and-final plot for the data in Fig. 1 showing notional outflows, as red lines, originating from donor units with above-average counts, denoted by red asterisks, and terminating at receiver units, with below-average counts, denoted by blue asterisks. Flows were computed by the Transportation algorithm, as implemented by the SADIE technique. Four outflows from unit A go in similar directions, at decreasing distances, to units B,C,D,E, respectively. The longest distance shown is the inflow to unit G; the shortest is exactly 30m, for example each of the three outflows from unit F, or one of the two inflows to unit H. Axes show distances in metres.


240.00
220.00
200.00
180.00
160.00
140.00
120.00
100.00
80.00
60.00
40.00
20.00
0.00
0.00
20.00
40.00
60.00
80.00
100.00
120.00
140.00
160.00
180.00
-1.50
0.00

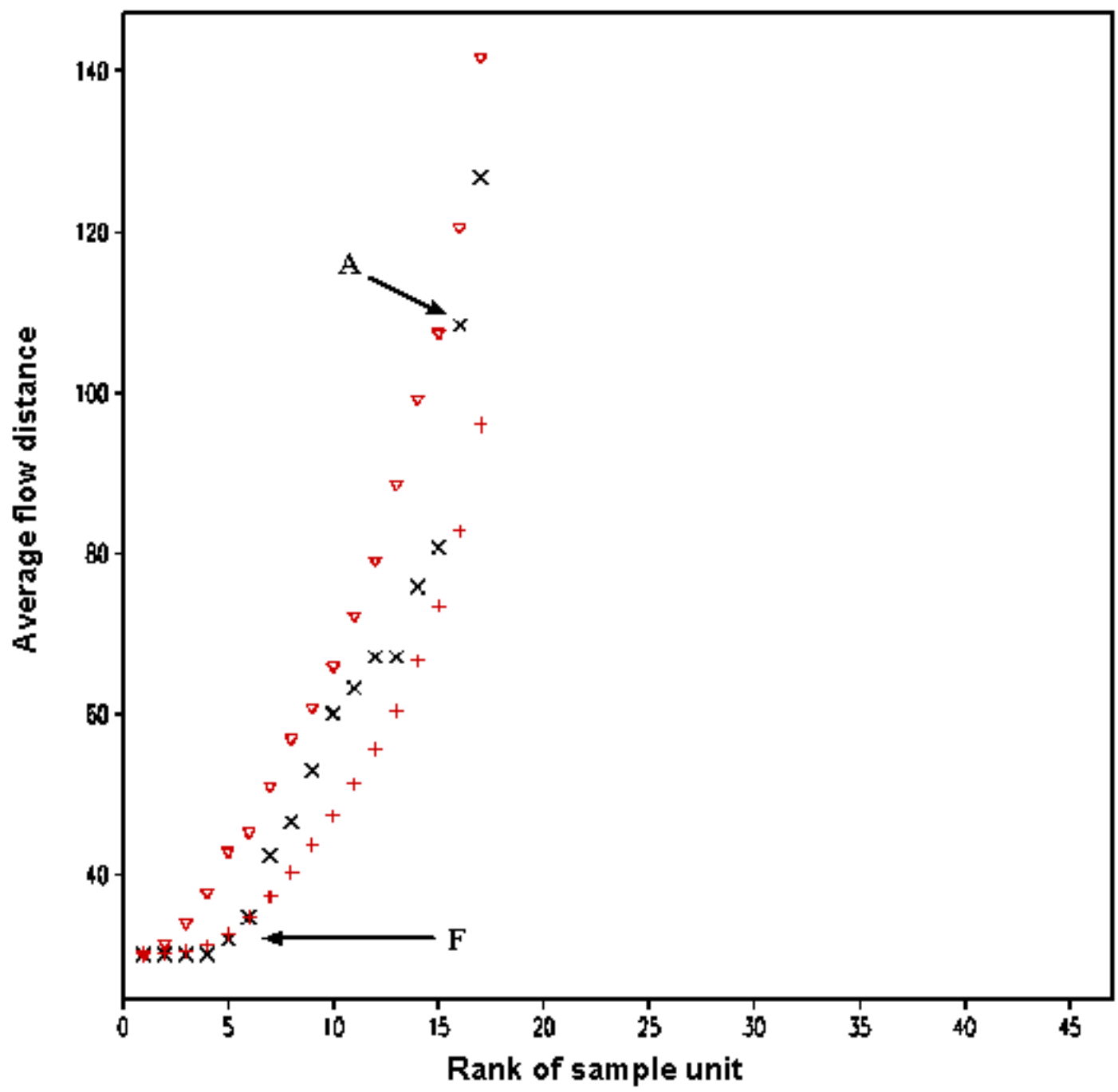


Figure 4 Empirical distribution function plot of ranked average outflow distances (in metres), Y_i, plotted against the rank of each sample unit, for the 17 donor units with above-average counts from the data of Fig. 1. Observed values, **x**, shown in black. Values from randomizations shown in red. Mean value of Y_i from the 400 randomizations denoted as **+**; envelope of upper 95th centiles denoted as **∇**. Observed values for the two sample units labelled as A and F, that show very different flow distances in Fig. 2, are arrowed, as examples. Observed values generally exceed means from randomizations, but never exceed envelope, indicating moderate patchiness, not inconsistent with random variation.

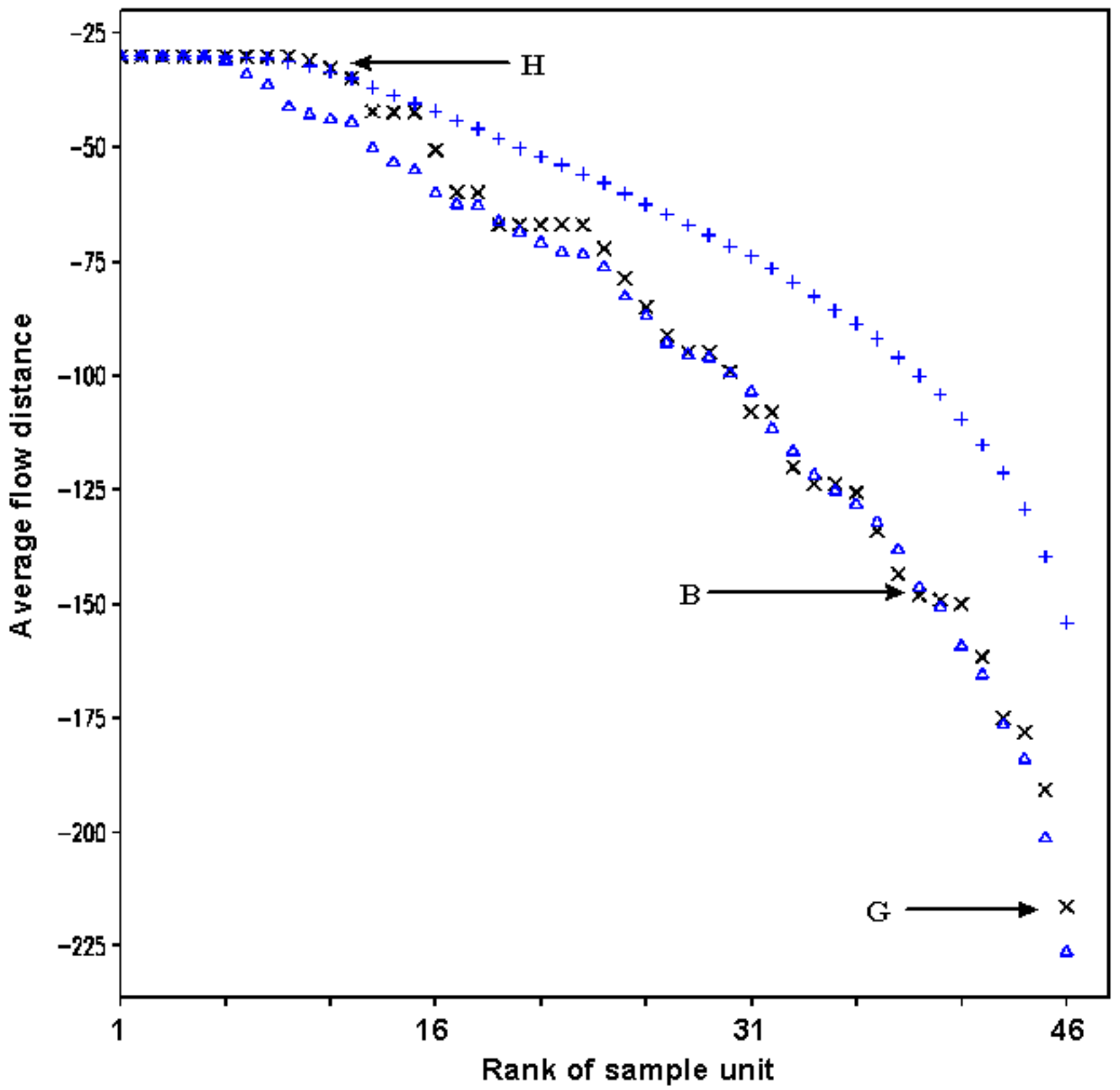


Figure 5 Empirical distribution function plot of ranked average inflow distances (negative by convention, and in metres), Y_j, plotted against the rank of each sample unit, for the 46 receiver units with below-average counts from the data of Fig. 1. Observed values, **x**, shown in black. Values from randomizations shown in blue. Mean value of Y_i from the 400 randomizations denoted as **+**; envelope of lower 5th centiles denoted as **Δ**. Observed values for the three sample units labelled as B, G and H, that show very different flow distances in Fig. 2, are arrowed, as examples. Observed values generally negatively exceed means from randomizations, and several in the inter-quartile range negatively exceed the envelope, providing strong evidence of a large gap.

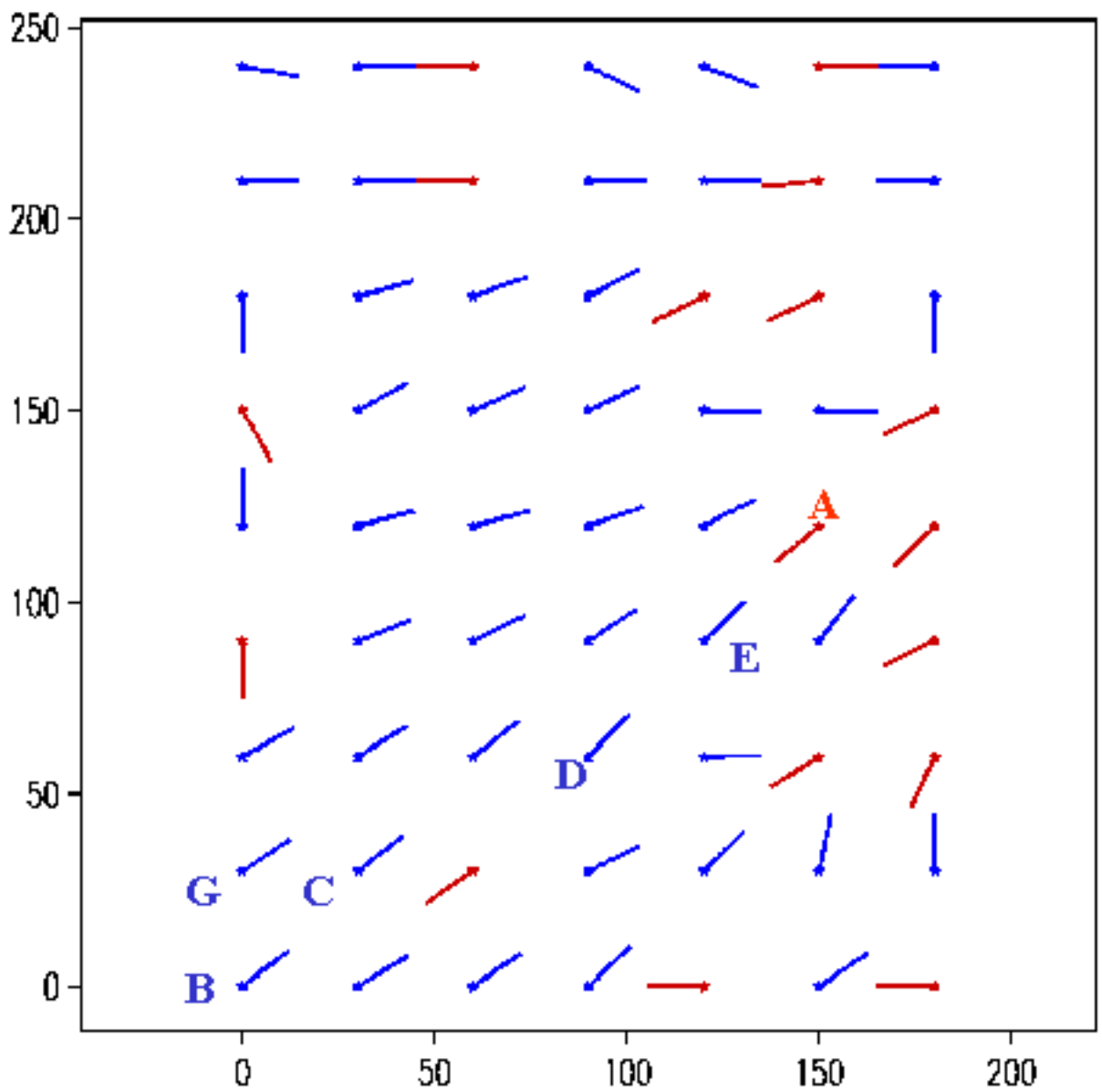


Figure 6. Vector flow plot shows the resultant vector directions of the flows represented in Fig. 2. Vectors shown scaled to identical length. Locations of sample units denoted by asterisks. Direction of outflow is indicated by a red line originating from a red asterisk; direction of an inflow by a blue line terminating at a blue asterisk. Spatial pattern is demonstrated by systematic correlation between the directions of nearby units. The trend from north-east to south-west is confirmed by similarity in direction between nearby units, especially as nearby units of the opposite colour show a similar direction, exemplified here by donor unit A and receiver units B,C,D,E, and G. Absence of spatial pattern would be indicated by haphazard directions, and areas where inflow and outflow directions appear unmatched.