DETECTING REGIONS OF ABRUPT CHANGE IN MAPS OF BIOLOGICAL VARIABLES

GUIDO BARBUJANI,1,3 NEAL L. ODEN,2 AND ROBERT R. SOKAL3.

1Dipartimento di Biologia, Universita di Padova, via Trieste 75, I-35123 Padova, Italy; 2Department of Preventive Medicine, Health Sciences Center, State University of New York at Stony Brook, Stony Brook, New York 11794-8036; and 3Department of Ecology and Evolution, State University of New York at Stony Brook, Stony Brook, New York 11794-5245

Abstract.—The areas where the rates of change of biological variables across space are particularly high may correspond to either steep ecological gradients or regions of limited admixture among demes. A method for detecting such biological boundaries was proposed by Womble (1951), who suggested averaging the absolute values of the derivatives of the functions describing biological variation in space at various locations. We present here algorithms that quantify both the mean magnitude and the mean direction of change in surfaces representing distributions of biological measures (such as gene frequencies, measures of quantitative traits, etc.). Inferences on the microevolutionary processes affecting the populations can be made by comparing the boundaries detected with the distribution of environmental characteristics, or with the location of factors that may have prevented population admixture. Examples of the application of this method to both simulated data and gene frequencies of two natural populations are given. [Boundary detection; gene frequencies; Wombling]

Consider a set of observations of the frequencies of a single allele made contemporaneously at various localities in an area. It is natural to consider these data as if they were values, observed at the localities, of a continuous function defined over the entire study area. We call such a function a "surface." It is common to approximately represent a surface by a matrix of (estimated or observed) surface values at the nodes of a geographic grid. Of course, surfaces may represent biological data other than gene frequencies. Examples include population density, biomass, etc.

Descriptive studies have traditionally focused on detection of areas in which the biological characteristics of the populations are homogeneous. This is often referred to as the regionalization problem (e.g., Matheron, 1970), and involves the use of statistical techniques for grouping areas considered as operational taxonomic units (OTUs). Some examples for such an approach can be found in Edwards and Cavalli-Sforza (1963), Bunge (1966), Ray and Berry (1966), Spence and Taylor (1970), Sneath and Sokal (1973:201-244), Ohno et al. (1979), Symons et al. (1983), and Huel et al. (1986). A recent ecological review paper is by Legendre and Fortin (1989). However, detecting areas of difference or change is probably equally important since regions of rapid transition may indicate zones of contact between unlike biological entities or the effect of rapid change in the underlying environment. Monmonier (1973) and Sokal et al. (1988) developed several methods for ascertaining the existence of such areas, or boundaries, when their potential location is known from other evidence. But when no prior information about the location of boundaries is available, their discovery becomes more difficult. A time-honored approach is to plot contour maps of the variable and look for regions where the contour lines, equidistant to each other with respect to the variable of interest, run closely together spatially. Alternatively, we may resort to a method proposed by Womble (1951).

In their simplest form, boundaries may be visualized as lines separating two different regions, each one displaying comparatively little variation; we contrast boundaries to wide areas of gradual biological change, referred to as gradients or clines, even though the difference between boundaries and clines cannot be defined.
unambiguously. In natural populations boundaries may be due to various causes, the most obvious being a sharp environmental transition, or ecotone (Endler, 1977: 80). A low rate of exchange of individuals or gametes between neighboring populations, whether separated by physical barriers or not, may be another important factor (Wright, 1969:290–345, 1982).

We can locate boundaries on a single surface by searching for regions in which the absolute value of the surface slope is large. A drastic change in an underlying causal variable, such as an impediment to gene flow, will presumably induce boundaries simultaneously in several genetic surfaces. We recognize these by locating regions with high average values of the absolute slopes across the genetic surfaces. The average absolute slope, itself continuous over the study area, has been termed the “systemic function” by Womble (1951). Below, we extend Womble’s concept to consider not only slope magnitude, but also direction. Use of the systemic function need not be limited to gene frequencies, but seems applicable to the analysis of any continuous variable, including morphological measurements, frequencies of polygenic and environmentally influenced traits, and even environmental variables. We have adopted the term “wombling” for the procedures detailed below.

Womble’s suggestions did not lead to development of a descriptive method for detection of boundaries. One reason for this may have been the absence of suitable sets of data; another, the cumbersome calculations required, prohibitive in the pre-computer years. To our knowledge, Womble’s approach has so far been employed only for the analysis of existing contour maps (Hagmeier, 1958; Adams, 1970). Adams was the first to put forward suggestions on how to deal with matrices of character measures, rather than isopleth maps as in Womble’s original formulation.

The purpose of this paper is to present an updated method for the detection of boundaries based on Womble’s approach, while incorporating the useful suggestions of both Womble and Adams. Both the magnitude of the maximum slope of gene frequencies and its orientation will be calculated. Measures of dispersion for these statistics can be calculated according to Batschelet (1965, 1981). The performance of the method will be assessed by analyzing computer-generated surfaces, and examples will be given of applications to real data.

**METHODS**

Let the \( n \) surfaces of interest be represented by \( n \) sets of equally spaced points, each set forming a grid. Each set has \( h \) rows and \( k \) columns of points, and will ideally be the result of a regular sampling of the region. The grid tessellates the plane into \((h - 1)(k - 1)\) square or rectangular sub-regions we term “pixels.” In real life, most surfaces will probably be obtained from irregularly scattered observations through an interpolation procedure, such as SYMAP (Doughenik and Sheehan, 1979) or kriging (Matheron, 1970). As is widely recognized there are drawbacks to the use of interpolation procedures in studying spatial variation. We shall consider them below in the last paragraph of this paper.

As a first approximation, the systemic function can be envisaged as a matrix of vectors in the plane representing the geographic area studied. Each vector has a magnitude, which is a measure of the average absolute slope of the surfaces at that point, and a direction (in the original Womble formulation the direction of the slope was disregarded). In the procedure here proposed, each surface of gene frequencies (or other variables) may be transformed into a matrix of measures of magnitude of the slope and a matrix of indices of orientation of this slope, both calculated at the center of each pixel. The magnitudes and the directions of the slopes are then averaged across surfaces (detailed below), yielding two matrices of size \((h - 1)\) by \((k - 1)\). The matrix of average values of the absolute magnitude (matrix AVMA) corresponds strictly to Womble’s systemic function. The matrix of average directions (matrix AVAN) contains angles as entries. Each angle shows the deviation of the av-
average gradient direction from a reference axis, such as the parallels on coordinate axes of a map.

Figure 1 shows a section of a matrix of surface values, say gene frequencies. Let us establish the convention (illustrated in Fig. 1) that \(X\) and \(Y\) denote respectively horizontal and vertical axes in the plane of the paper, and that gene frequency constitutes a third dimension, coming up out of the plane of the paper. Let \(p_A, p_B, p_C\) and \(p_D\) be the gene frequencies at the vertices of the square ABCD. A Cartesian coordinate system with origin in A is superimposed; if the distances AB and AD are assumed to be unity, the coordinates of the vertices will be, respectively: A(0; 0), B(1; 0), C(1; 1) and D(0; 1). We choose a bilinear function (Rogers and Adams, 1976) to approximate the surface within ABCD because it is simple, continuous, interpolates between the corners of the surface of ABCD, and has simple derivatives. If the square ABCD is small enough, the bilinear function will not deviate far from the surface.

The bilinear function \(f(X, Y)\) is

\[
f(X, Y) = p_A(1 - X)(1 - Y) + p_BX(1 - Y) + p_CXY + p_D(1 - X)Y. \quad (1)
\]

If one thinks of the boundary of the pixel as if it were the four edges of a square, warped tennis racket, then the bilinear function gives the surface described by the strings of the racket.

The partial derivatives of the function \(f\) are

\[
\frac{\partial f}{\partial X} = p_B - p_A + Y(p_A - p_B + p_C - p_D) \quad (2)
\]

and

\[
\frac{\partial f}{\partial Y} = p_D - p_A + X(p_A - p_B + p_C - p_D). \quad (3)
\]

The maximum slope of \(f\) will then be the magnitude \(m\) of the vector \(v = (\frac{\partial f}{\partial X}, \frac{\partial f}{\partial Y})\) (Thomas, 1972)

\[
m = |v| \sqrt{\left(\frac{\partial f}{\partial X}\right)^2 + \left(\frac{\partial f}{\partial Y}\right)^2}. \quad (4)
\]

This is evaluated at the point \(Q = (1/2; 1/2)\) to obtain a representative slope for the square. As observed above, if the square ABCD is small, this magnitude will not deviate far from the magnitude of the gradient of the actual surface. Adams (1970) assumed \(AB = AD = 1/\sqrt{2}\), and obtained the same result, but the magnitudes computed using his algorithm will all be \(\sqrt{2}\) times the corresponding \(m\)-values, based on the formulas outlined above.

By the direction of a vector lying in the \(X, Y\)-plane we mean the angle made by that vector with the \(X\)-axis. From the definition of \(v\) its direction is

\[
\theta = \arctan \frac{\frac{\partial f}{\partial Y}}{\frac{\partial f}{\partial X}} + \Delta \quad (5)
\]

where

\[
\Delta = 0^\circ \quad \text{if } \frac{\partial f}{\partial X} > 0
\]

\[
= 180^\circ \quad \text{if } \frac{\partial f}{\partial X} < 0
\]

This is illustrated in Figure 2.

In this way it is possible to compute a matrix of magnitudes and a matrix of angles for each surface. In some instances the simple inspection of such matrices will allow inferences on the location of boundaries. However, in most cases the information present in various surfaces will have to be pooled.

We now address the problem of obtaining average magnitudes and directions over a set of surfaces at a particular geographic locality. Suppose that there are \(i = 1 \ldots n\) surfaces, so that, at each locality, we need consider \(n\) vectors, \((a_i, b_i)\).
The squared magnitudes are
\[ m^2_i = a^2_i + b^2_i. \]

The average magnitude is
\[ \bar{m} = \frac{1}{n} \sum_{i=1}^{n} m_i. \]

To obtain the average direction, we modify a procedure given by Batschelet (1965, 1981). This procedure involves scaling each vector to unit length, and then obtaining the vector that is the average of the scaled vectors:
\[ (A, B) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{m_i} (a_i, b_i) \right) \]

The angle of \((A, B)\), calculated as in expression (5) above with \(A\) substituted for \(\partial f/\partial X\) and \(B\) substituted for \(\partial f/\partial Y\), is the average angle of the vectors \((a_i, b_i)\).

(We note parenthetically that information on the variance of the angles is available from the length of \((A, B)\). See Batschelet for elaboration of this issue.)

Our modification of Batschelet's procedure stems from the fact that vectors pointing in opposite directions will cancel each other in the average direction calculation. This issue was raised by Womble (1951) and stressed again by Adams (1970): when working with genetic data, we wish two genes with opposite slopes to reinforce each other in direction, rather than cancelling each other. This is because opposite slopes may be caused either by the same migrational wave (one migrant allele frequency being low, and the other high, with respect to the residents) or by two migrational waves moving in opposite directions (both migrant allele frequencies being simultaneously high, or low, with respect to the residents). In the same way, similar slopes may either be caused by one migrational wave (both frequencies simultaneously higher, or lower, than the residents) or two opposite migrational waves (one migrant frequency high, the other low). On the basis of slope alone, we cannot logically distinguish between uni- and bi-directional migration. We therefore require our method to treat them the same.

The solution, suggested by Batschelet in a different context, involves doubling the angles before averaging, and then taking one half the average as our result. This is done because two vectors pointing in opposite directions (\(\theta\) and \(180^\circ + \theta\)) will be transformed to vectors pointing in the same direction when their angles are doubled \((2\theta, 2(180^\circ + \theta) = 360^\circ + 2\theta = 2\theta)\). By using this approach, directions at right angles to each other will now cancel out. This is a desirable attribute in view of the rationale developed earlier.

To multiply the angle of a vector by a constant \(K\), we first calculate \(\theta\) for the vector as specified in (5), and then take
\[ x = m \cos K\theta \]
\[ y = m \sin K\theta \]

The motivation of this procedure is given in Batschelet (1981). See Figure 3 for an illustration of the entire angle doubling and averaging procedure. In Figure 3a, vectors of unit length \(a, b,\) and \(c\) depict the directions of the slopes of three biological surfaces at \(Q\). They are defined unambiguously by the angles they form with the abscissa: \(\alpha, \beta\) and \(\gamma\), respectively. In Figure 3b, to calculate a mean direction the angles are doubled obtaining the vectors \((a', b',\) and \(c',\) respectively, with their new coor-
ordinates on the $X$- and $Y$-axes. In Figure 3c, the average coordinates $(\bar{X}; \bar{Y})$ will define the average vector $\bar{\mathbf{v}}'$, whose direction is given by the angle $2\phi$ that can be computed as in expression (5). By halving this angle we obtain the angle defining the direction of the average vector $\bar{\mathbf{v}}$, which, accordingly, will be defined in the range $0-180^\circ$.

In some cases a weighted average of the magnitude values of each individual surface may be appropriate. For example, the researcher may wish to give equal weight to each locus studied, rather than to each allele. The $m$-values (equation (4)) observed for each allele should then be divided by the number of allele frequency surfaces considered for the respective locus, and the final average will require division by the number of loci (and not alleles) considered. For gene frequencies the magnitudes could be standardized according to the potential variability of the original surface, by dividing the magnitudes by $[\bar{p}(1 - \bar{p})]^{1/2}$, where $\bar{p}$ is the mean frequency of the surface mapped. This is analogous to the procedure currently employed by comparing variograms computed for different loci (see Piazza and Menozzi, 1983; Cavalli-Sforza, 1984; Barbujani, 1988). Other types of variables can be standardized to a common mean and variance.

The information contained in the magnitude and direction matrices must eventually be synthesized to show the areas of the region in which biological change is particularly sharp. One approach is to plot a series of rods on the geographic map. The rod length is proportional to the AVMA value (average magnitude) at that location, and its orientation is given by the value in the AVAN matrix (average direction).

An alternative approach involves automatic, computerized recognition of boundaries, based on the principle that a boundary is characterized by slopes that are not only high, but consistently so across at least some neighboring points of the map. We have defined two criteria for linking pixels into boundaries. Under criterion 1, two points are "connected" if two conditions hold:

1) They are both in the highest decile of the AVMA values.
2) They are linked

2.1) by a king’s move (by analogy to chess) directly and/or
2.2) by a king’s move to a point in the second-highest decile of the AVMA values. The point serves as a "bridge" between them.

Once the entire AVMA matrix has been scanned, the connected points can be plotted. This eliminates from the final map the points at which average slope magnitude is high by chance, and those which are not part of a region of consistent biological change, or boundary. Obviously the choice of deciles is arbitrary. In other studies with "busier" maps of the systemic function we have used the highest 5 percent only.

Criterion 2 is identical to criterion 1, with
the additional restriction that connections are made only when the difference between the values of AVAN for the two points to be connected under criterion 1 is smaller than some arbitrary value, say, 30 degrees.

The two criteria will locate areas of rapid change plus isolated spots of high biological variation. As the difference between a biological boundary and a steep cline is not clear-cut, the researcher must decide whether an observed string of connected points should be called a boundary. The criteria for this choice include evaluation of the length of the string with respect to the entire area studied and assessment of biological homogeneity within the regions that such a string separates.

**ANALYSIS OF COMPUTER-GENERATED SURFACES**

In this section the systemic function will be employed to describe distributions of gene frequencies simulated by computer, and representing the outcome of three basic processes: isolation by distance (IBD); differentiation of subpopulations along a cline (CD); and additional differentiation of subpopulations in two patches (PD). The program employed (Rohlf and Schnell, 1971; Sokal and Wartenberg, 1983) simulates a population of 10,000 individuals, arranged on a 100 × 100 grid. For each individual the initial genotype at a biallelic locus is determined randomly, according to Hardy-Weinberg probabilities corresponding to predetermined initial allele frequencies. At each generation, each individual is replaced by an offspring of a cross between two parents chosen at random in a 5 × 5 neighborhood (25 individuals) centering on it. The genotype of the offspring depends only on Mendelian probabilities in the simulations of the IBD surfaces. The effects of selection against one allele, with the selection coefficient varying from 0 to 0.05 along a horizontal axis, is additionally simulated for the CD surfaces. The entire procedure is repeated for 200 generations. The frequency of one allele is then measured in sublattices of size 5 × 5. Therefore, every surface is eventually represented by a 20 × 20 matrix of allele frequencies calculated for 400 subpopulations of 25 individuals each. The simulation ends at this stage for the IBD and CD surfaces (see Sokal, Jacquez and Wooten, 1989). Seventeen PD surfaces were generated by adding a constant quantity (0.2) to the gene frequencies of 14 IBD surfaces and 3 CD surfaces yielding surfaces IBD + PD and CD + PD, respectively. The location of the two patches is the same in all replicate PD surfaces, (see Fig. 4c, f).

Table 1 summarizes some characteristics of the 37 surfaces employed in this study (15 IBD, 5 CD, 17 PD). Note that in sets 1.6 and 1.7 the average allele frequency is lower at the end of the simulation than at its beginning. The quantity added to the allele frequencies in the PD surfaces to define the patches is insufficient to overcome random downward fluctuations.

Ideally, the method described above should detect no boundary in the IBD surfaces, and two boundaries (and only two) in the IBD + PD surfaces. In the CD surfaces it should indicate the area(s) where the gradient is steepest, or no area at all if the slope is uniform. In addition, it should be able to recognize the edges of the simulated patches when PD surfaces are analyzed together with other surfaces that do not include boundaries. When the surfaces are analyzed individually, and the resulting matrices are plotted according to criterion 1 or 2, no clear picture emerges (see Fig. 4a, b, c for examples). Areas of abrupt

**Table 1. Some characteristics of the simulated gene-frequency surfaces employed.**

<table>
<thead>
<tr>
<th>Set</th>
<th>Process</th>
<th>Initial allele frequency</th>
<th>Number of replicates</th>
<th>Avg. final allele freq.</th>
<th>Avg. final $F_{st}$ ($\times 100$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>IBD</td>
<td>0.50</td>
<td>5</td>
<td>0.53</td>
<td>9.43</td>
</tr>
<tr>
<td>1.2</td>
<td>IBD</td>
<td>0.25</td>
<td>5</td>
<td>0.21</td>
<td>8.19</td>
</tr>
<tr>
<td>1.3</td>
<td>IBD</td>
<td>0.10</td>
<td>5</td>
<td>0.10</td>
<td>7.86</td>
</tr>
<tr>
<td>1.4</td>
<td>CD</td>
<td>0.50</td>
<td>5</td>
<td>0.89</td>
<td>14.22</td>
</tr>
<tr>
<td>1.5</td>
<td>IBD + PD</td>
<td>0.50</td>
<td>5</td>
<td>0.57</td>
<td>11.23</td>
</tr>
<tr>
<td>1.6</td>
<td>IBD + PD</td>
<td>0.25</td>
<td>5</td>
<td>0.21</td>
<td>8.75</td>
</tr>
<tr>
<td>1.7</td>
<td>IBD + PD</td>
<td>0.10</td>
<td>4</td>
<td>0.09</td>
<td>11.85</td>
</tr>
<tr>
<td>1.8</td>
<td>CD + PD</td>
<td>0.50</td>
<td>3</td>
<td>0.86</td>
<td>14.09</td>
</tr>
</tbody>
</table>

$F_{st}$ is the standardized allele frequency variance (Wright, 1969). "Final" parameters are measured after 200 simulated generations. CD—clinal differentiation, IBD—olation by distance, PD—differentiation in patches.
Variation occur occasionally in the IBD and CD surfaces. In the latter, they appear as a network, rather than a string, of contiguous high values of the systemic function. In the PD surfaces, however, the boundaries are usually detected, but they sometimes extend over areas that do not correspond to the simulated patches, or occur elsewhere. When surfaces representing independent realizations of the same generating process are analyzed together, the systemic function is computed, and vectors for the highest connected decile are plotted, we obtain the results shown in Figure 4d–f for five IBD surfaces, five CD surfaces, and six IBD + PD surfaces, respectively. The falsely positive boundaries of some realizations (not shown) tend to disappear, as is true also for false branches of the “real” boundaries in the PD surfaces. Some new significant pixels appear in the systemic functions (Fig. 4d–f) but these do not meet the criteria of length and parallelism that justify calling them boundaries. The area of sharp gene-frequency change at the left end of the CD surfaces is a result of fixation of the “adaptive” gene in the region (at the right) where selection against the alternative allele was strongest (s = 0.05). As a consequence, the systemic function attains its maxima at the left, where almost all gene-frequency variation can be found.

To assess the accuracy of the method we employed 9 sets of 8 simulated surfaces each. These sets differed from each other in the number of PD-surfaces (from 2 to 4) and CD-surfaces (from 1 to 3) included. The exact number is shown in the column and row headings of Table 2. Each set also contains one CD + PD surface, included in the total CD and PD counts in the margins. The conventions for summarizing the results are as follows: (i) simulated boundaries are considered recognized whenever more than half of their length is apparent. For example, when in the matrix of the systematic function, at the appropriate locations, at least 8 out of the 14 values forming boundary 1, and 6 out of 10 forming boundary 2 are connected according to criterion 1, the method is considered successful. False positives include (ii) extensions of boundaries recognized by a string of at least 4 values, connected according to criterion 1, in areas where no abrupt change of gene frequencies had been simulated; and (iii) random boundaries recognized by strings of at least 4 connected values that are not in contact with the “real” boundaries.

The boundaries of the patches escaped detection in only 3 out of 18 possible cases. When the number of PD-surfaces in the analysis was low, the values of the systemic
function did not reach significance in all the pixels forming the simulated boundaries. As the number of patchy surfaces increases there is a clear trend toward successful recognition of the simulated patch boundaries. The occurrence of false positives scattered without any pattern across Table 2, depends on occasional wide gene-frequency fluctuations in the IBD-surfaces, which in turn are due to the small size of the sampling units (25 individuals each). The probability of detection of a boundary is a function of the number of PD-surfaces in the set analyzed; no obvious factor seems to be associated with the false positive results.

### ANALYSIS OF ACTUAL GENE-FREQUENCY SURFACES

The methodology described and tested above was applied to two groups of data from natural populations. The first group is composed of frequencies of 9 alleles in 56 eastern Australian populations of *Drosophila buzzatii* (Sokal et al., 1987). (In the original publication 12 alleles had been analyzed. We left out 3 that varied little over much of the area and therefore would have little impact.) The second dataset includes frequencies of 15 alleles, measured in 50 villages inhabited by the Yanomama Indians of Venezuela and Brazil (Sokal et al., 1986). To decrease statistical dependence between data values, we considered only \( n - 1 \) alleles at each n-allelic locus.

For both *Drosophila* and Yanomama, the sampling localities were irregularly scattered in the area studied. The greatest interpoint distances are 2,274 and 579 km for *Drosophila* and Yanomama, respectively. The 24 sets of data were therefore transformed by interpolation using the SYMAP algorithm (Dougenik and Sheehan, 1979) onto the nodes of 9 regular lattices of 63 rows \( \times \) 79 columns (*Drosophila*), and 15 regular lattices of 51 rows \( \times \) 65 columns (Yanomama).

The interpolated surfaces were then subjected to the Womble method. The Australian area studied ranged from 21° south to 37° south; as a consequence, the horizontal (longitudinal) distances between pairs of neighboring pixels in the interpolated surfaces varied substantially, from 21.2 km near the Equator to 18.2 km in the southernmost part of the map (the vertical or latitudinal distances were constant: 28.7 km). Equations (2) and (3) assume that each pixel is a square; in this way, the length of its edges can be disregarded. Conversely, if it is a rectangle, or approximately so, as for Australian *Drosophila*, the partial derivatives along the X- and Y-axes refer to different distances in space, and should be modified to compensate for that. We achieved that by multiplying the X-derivative (equation 3) by a correction factor \( c(i) = EDY/EDX(i) \), where EDY and EDX(i) are the distances (in km, not in degrees) between the edges of that pixel along the Y- and X-axes, respectively (EDY is constant in the map, EDX changes as a function of latitude). Equal weight was given to each allele studied. The maps summarizing magnitude and orientation of the overall gene-frequency change in the two study areas are given in Figures 5 and 6 based on the highest 5% vector magnitudes. In these maps, boundaries were identified by connecting pixels according to criterion 2. The maximum difference allowed between directions of the slope at connected sites was 45 degrees.

*D. buzzatii* was apparently introduced into Australia from Argentina between 1931 and 1936 (Barker et al., 1985). It inhabits rotten cladodes of cactus *Opuntia* infested with

### Table 2. Tests of the systemic function with varying numbers of patchy and clinal surfaces.

<table>
<thead>
<tr>
<th>Number of CD-surfaces</th>
<th>Number of PD-surfaces</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/8</td>
<td>(+)</td>
</tr>
<tr>
<td>2/8</td>
<td>(+)</td>
</tr>
<tr>
<td>3/8</td>
<td>(+)</td>
</tr>
</tbody>
</table>

The PD-surfaces include 2 boundaries indicated by 2 plus or minus signs. +: boundary detected; (+): boundary detected in part; −: boundary not detected; *: false positive, i.e., boundary recognized that had not been simulated.
the lepidopteran *Cactoblastis cactorum*, a biological control agent for the cactus which had spread over wide areas of eastern Australia. By 1940, the *Opuntia* was controlled and its distribution was reduced more or less to the islands found today. The *Drosophila* are correspondingly isolated.

For the *Drosophila* data the variation pattern (Fig. 5) can be characterized by a central area of homogeneity ranging from locality 54 in the north through locality 2 in the south. This area of homogeneity is included within the area of the main *Opuntia* infestations (see Sokal et al., 1987:fig. 1a). There are 3 single outlying localities: 36 in the north in Queensland, 25 in western Queensland, and 49 in South Australia. Each of these is separated from the main body of locality samples both by geographic distance and also by high allele-frequency gradients revealed by the Womble method. They occur on local, isolated *Opuntia* patches. Additionally there are 4 clusters of locality samples each lying in a zone of rapid change. These clusters are localities 8, 19, 30, 31, 34, 35, and 56 in southeast Queensland extending from the coast across the coastal plain to the eastern edge of the tablelands; localities 23, 24, 26, 27, and 50 in southwest Queensland; 3, 4, and 58 in New South Wales on the coast and coastal plain; and 39–41 along the western part of the border of New South Wales and Victoria. Note that all of these clusters are at the periphery of the distribution of localities, not near its center. The two eastern clusters may be differentiated not only because they are peripheral and hence possibly exhibit founder effects, but also because they occur on or near the coast and hence may be physiographically and ecologically separated from each other and from other populations. The New South Wales–Victoria border cluster is quite isolated from the other localities, and the *Drosophila* in this cluster occur on very local, isolated patches of *Opuntia*. Localities 42–48 just northwest of 39–41 are differentiated from all other populations as shown by the significant gene-frequency boundaries surrounding them. Thus the overall pattern of differentiation appears to agree well with the distribution of the obligate host plant, and the detailed differentiation appears to relate to additional microgeographic and microclimatic factors.

The Yanomama populations represent villages that have descended from each other by fission and have expanded within the last century in both population size and territory (Neel and Weiss, 1975). The villages can be divided into four groups representing different dialects, and there is further linguistic subdivision within these dialect groups (Spielman et al., 1974; Smouse, 1982).

The systemic function for the Yanomama gene frequencies (Fig. 6) differs from that of the *Drosophila*. There are no large

---

**FIG. 5.** Plot of the systemic function for 9 alleles in Australian *Drosophila buzzatii*. The length of the rods is proportional to the magnitude of gene frequency change, the direction of the rod is the average direction of maximum slope across gene-frequency surfaces. Only the values belonging to the highest decile of the distribution of magnitudes, and the "bridges" (average magnitude greater than 0.0225 and 0.0145, respectively) are plotted.
homogeneous areas but there are boundaries in the center of the figure as well as in the southeastern and southwestern regions, plus a large region of rapid change in varying directions in the southwestern quadrant of the figure. The boundaries indicated in the westernmost part of the area must be artifacts of the interpolation procedure, since there are no observations located there. Based on this figure one can differentiate an easternmost group of villages (15L, 3X, 15Q). This corresponds to the Ninam dialect group. A second group, separated from the Ninam by a boundary and extending west all the way into the margin of the high change region, includes localities 3W, 3KP, 3LMN, 3Q, 3T, 3RS, 11V, 11S, 11X, 8K/15M, 11ABC, 11D, 15H, and 8XY. These correspond largely to the Yanomam dialect cluster, except that village 8K/15M, included by the Womble boundaries and geographically very close, belongs to the Yanomame dialect cluster, and village 11U which belongs to the Yanomam shows up in a genetically differentiated area. A northwestern cluster of villages comprising 8D, 8E, 8F, 8I and 3U is well differentiated from both eastern and southern villages and corresponds exactly to the Sanema dialect. This leaves the largest number of villages scattered in the area of high change in the southwestern quadrant of the map. These all belong to the Yanomame dialect. Its two southernmost villages (15QR and 11YZ) actually are beyond the region of rapid change and might be separated on that account. Except for geographic contiguity, there would be no reason according to the boundary map to unite the localities that comprise the Yanomame. They are the most numerous group of villages among the Yanomama and are known to contain five (mini)clusters that are somewhat divergent from each other.
The whole dialect group may have expanded and differentiated over the last hundred years as the Yanomame expanded west and southwest into unoccupied territory accompanied by considerable fission (Neel, 1978). Figure 6 clearly indicates that on the average there are rapid changes in a variety of directions as one travels from village to village. We note that both at the level of variation among dialect groups as well as within such groups the regions of rapid genetic change agree with observable linguistic differences. These findings support conclusions from other studies on the relation of genetics and language (Sokal, 1988; Sokal et al., 1990).

**DISCUSSION**

The analysis of both simulated surfaces and actual gene frequencies in two natural populations shows that the method proposed here is sensitive enough to detect areas of abrupt biological change under various circumstances. The method recognized most simulated patches, despite the fact that PD-surfaces were never more than 50% of the surfaces considered in each cycle of analysis. Moreover, the degree of differentiation of the patches was small compared to the extent of random variation in the data. As an example, in sets 1.6 and 1.7 (Table 1) the quantity added to generate the patches did not compensate for random downward fluctuations of the simulated gene frequencies. As a consequence, the average allele frequency after 200 generations of random sampling of gametes and creation of two patches of increased frequencies was lower than at the beginning of the simulation. This means that the patches the method recognized differ from the surrounding regions by a quantity that is smaller than the random differences observed between adjacent pixels. The sensitivity of Womble’s method was, therefore, tested under conditions representing very limited genetic divergence. In nature, biological boundaries due to scarce admixture between population units are expected to be sharp, and to affect numerous loci (Ehrlich and Raven, 1969; Endler, 1973). This is confirmed by analysis of genetic variation in *Drosophila buzzatii* and the Yanomama. In these populations, many boundaries are clearly defined, and their interpretation is possible. In another application of wombling, Barbujani and Sokal (1989) were able to show that of 33 boundaries detected in 60 allele frequencies of human populations in Europe, 31 corresponded to well-known language boundaries (many of which are also montane or marine physical barriers). The remaining two boundaries are reflected in the ethnogeny of the present population but do not represent modern language boundaries or physical barriers.

Most models of the genetic structure of populations assume stationarity of the process studied. In other words, it is assumed that variation of the parameter of interest (generally, a measure of population similarity or dissimilarity) does not depend on the place where the parameter is evaluated, but only on the distance between the localities compared (Matheron, 1970). This assumption underlies the stepping-stone model of Kimura and Weiss (1964), isolation-by-distance models (Malecot, 1948; Morton et al., 1968, 1971; Morton, 1973), and in hypothesis testing of simple spatial autocorrelation (Sokal and Oden, 1978a, b; Cliff and Ord, 1981; Upton and Fingleton, 1985).

Stationarity of biological variation can reasonably be assumed for populations that occupy small regions, or for areas that are known to be ecologically uniform. Whenever the scale of a study spans wide or heterogeneous areas, that assumption likely will not hold true. On the contrary, discontinuities in the geographical distribution of biological variables become a major feature from which inferences on the microevolutionary processes affecting the population can be drawn. Apparently, biological boundaries have been seldom considered as evidence suggesting that population differentiation is not random in the area studied. Now that a descriptive method is available, the importance of their detection may be better recognized.

Testing the significance of a biological boundary, once it has been identified, is
complicated. Obviously, the test cannot be applied to the variables that have been employed to characterize the boundaries by the methods described above, since that would involve circularity of reasoning. However, significance tests can be applied to sets of variables that were not involved in determining the boundary. There are problems in applying standard statistical techniques, such as the analysis of variance, to such sets of observations. Natural populations tend to resemble each other as a function of their closeness in space, because of gene flow, spatially patterned selective processes, and other factors. Therefore, the values of biological variables tend to be spatially autocorrelated (Sokal and Oden, 1978 a, b; Barbujani, 1987); this violates the assumption of independence. For this reason, some unconventional methods have been employed to assess the extent of biological variation across potential boundaries (Sokal et al., 1988; Sokal, Oden et al., 1989; Sokal, 1988). These methods seem at present the best way for evaluating the significance of changes in a variable across boundaries detected by Womble's method, when the variable whose significance is being investigated is not a priori related to the variable used to recognize the boundary.

If the distribution of the variable under study is influenced by other variables which also have a spatial distribution, this additional information may be used in several ways. One could be by wombling a surface that is the average of a gene frequency and a physical variable. Other obvious approaches would be to carry out wombling of the physical surfaces only and test the gene frequency against the physical boundaries in the manner of Sokal et al. (1988). Another method would be to interpolate the biological variable while allowing for the physical variable, then wombling the adjusted, interpolated variable.

Finally, it may be asked to what extent the values of the systematic function and the direction of maximum slope are affected by previous interpolation of data. Obviously, the more observed points there are, the more closely the interpolated surfaces will resemble the real distribution of the variable studied. Since the ratio of observed to interpolated values is generally low, interpolated surfaces are often not very reliable. If in some areas the observations are scarce, biological boundaries located in such areas may escape detection. The opposite error, i.e., areas of gradual change recognized as boundaries, is not likely to be due to interpolation artifacts, as interpolation probably will not increase the local slope of a surface within the convex hull of the localities. Outside this hull, interpolation surfaces produced by polynomials (or other functions) may fluctuate widely, giving rise, for example, to the spurious boundaries west of the westernmost localities in Figure 6. We discount such boundaries in our interpretations. The simulations performed indicate that false positive results may be due to wide random fluctuations of gene frequencies, owing to small sample sizes. The quality of data affects the performance of Womble's method, as well as of any other method.

In the analysis of real data it seems advisable to take into account the surfaces both individually and jointly. The areas where the systemic function attains its maxima in both the pooled and the individual analysis can safely be regarded as regions of substantial population differentiation.

ACKNOWLEDGMENTS

This is contribution 729 in Ecology and Evolution from the State University of New York at Stony Brook. We are indebted to Junhyong Kim, Geoffrey Jacquez and Barbara Thomson for ideas and suggestions, and to Professors J. S. F. Barker and Peter Smouse for critical comments on a previous version of this paper. G. J. G. Upton, several anonymous reviewers, and J. W. Archie in his capacity as Associate Editor provided useful criticisms of an earlier version of this paper. Most of their suggestions have been incorporated in the present paper. Cheryl Daly typed several versions of the manuscript and Donna DiGiovanni furnished technical assistance; the artwork is by Joyce Schirmer. This research was supported by grant No. BSR 8614384 from the National Science Foundation to Robert R. Sokal.

REFERENCES


SOKAL, R. R., N. L. ODEN, P. LEGENDRE, M.-J. FORTIN,


Received 3 November 1988; accepted 25 July 1989