Statistical Science
1989, Vol. 4, No. 2, 87–120

A Survey of the Statistical Theory of Shape

David G. Kendall

Abstract. This is a review of the current state of the "theory of shape" introduced by the author in 1977. It starts with a definition of "shape" for a set of k points in m dimensions. The first task is to identify the shape spaces in which such objects naturally live, and then to examine the probability structures induced on such a shape space by corresponding structures in $\mathbf{R}^m$. Against this theoretical background one formulates and solves statistical problems concerned with shape characteristics of empirical sets of points. Some applications (briefly sketched here) are to archeology, astronomy, geography and physical chemistry. We also outline more recent work on "size-and-shape," on shapes of sets of points in riemannian spaces, and on shape-theoretic aspects of random Delaunay tessellations.

Key words and phrases: Central place theory, convex polygon, Delaunay tessellation, galaxy, quasar, riemannian submersion, singular tessellation, spherical triangle, stochastic physical chemistry, void.

1. THE ORIGINS OF STATISTICAL SHAPE THEORY

First a few words about terminology. When I was working in Princeton in 1952–1953 someone (I think it was Hassler Whitney) posted a notice in Fine Hall listing a large number of four- and five-letter words not yet used as technical terms in pure mathematics. I do not remember whether "shape" was one of these, but about 1968, according to Borsuk (1975), it was duly appropriated for such a purpose by topologists, so now when we wish to write about the mathematics and statistics of *real* shapes we are required to add an explanatory adjective in order to make it clear that we do mean shape as ordinarily understood and not an arcane concept in topology. Oddly enough, as will shortly become apparent, some other branches of topology turn out to play an important role in our shape theory, but this has nothing to do with Whitney's list.

There are several different approaches to the *statistical analysis* of (real) shapes (see for example Kendall (1984) and the recent reviews of Bookstein (1986) and Small (1988)). There is an equal diversity of approaches to the *geometric description* of shape, but here I will only describe one that I have developed in a series of papers starting in 1977 (Kendall, 1977) and associated with what now seems a very premature attempt to study shape-valued stochastic processes. W. S. Kendall's most recent work on shape-diffusions (W. S. Kendall, 1988) substantiates and considerably generalizes that exploratory essay and links it with research in stochastic physical chemistry (Clifford, Green and Pilling, 1987).

My interest in shape theory was prompted by a statistical topic on the fringes of archeology. When one looks at Stonehenge one accepts the underlying circular structure without asking for statistical authentication, and the same is true of the underlying linear structures in the monuments of Carnac. Statistical tests here would be quite out of place. But there are other archeological situations in which a linear structure is accepted by some and dismissed by others. Thus the set of 52 standing stones near Land's End, Cornwall, studied by Broadbent (1980) yields $\binom{52}{3}$ = 22,100 triplets of stones, and there are those who say vaguely that "too many" of these are "too nearly" collinear, and who attribute this to deliberate planning, whereas others dismiss such claims as ridiculous. Who is right?

We can quantify "too nearly collinear" by interpreting this to mean "the obtuse angle of the triangle defined by the triplet differs from two right angles by less than (say) $\varepsilon = 0.5$ degrees." Figure 1 shows a map of the plan positions of the 52 stones. There are 81 such "nearly collinear" triplets. Figure 2 shows these by means of line segments drawn to join the extreme members of each such triplet.

David G. Kendall is a Fellow of Churchill College and Emeritus Professor of Mathematical Statistics, Department of Pure Mathematics & Mathematical Statistics, University of Cambridge. Now retired, he continues to work fulltime in an honorary capacity. His mailing address is 37 Barrow Road, Cambridge CB2 2AR, England.

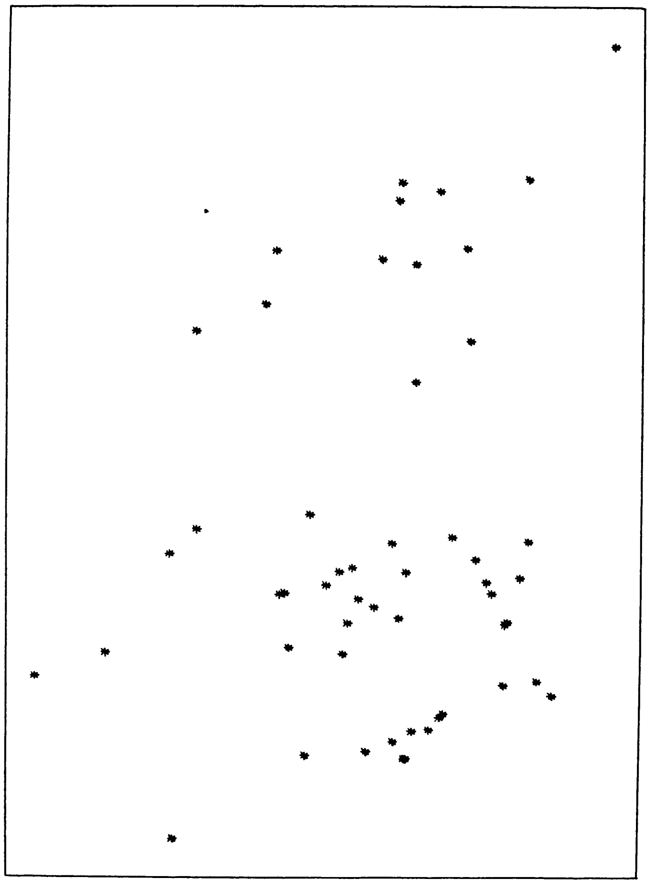

FIG. 1. *The plan positions for the 52 standing stones. (Data provided by S. R. Broadbent.)*

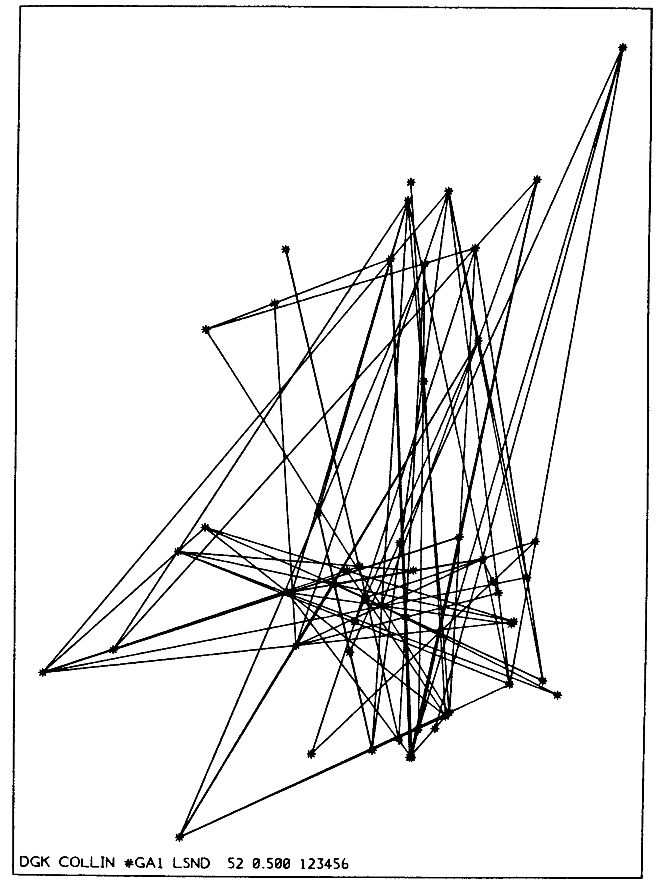

FIG. 2. *The 81 (half-degree) collinearities.*

Is 81 "too many"? A model-based or data-based interpretation of "too many" is evidently called for.

One's first attempt at answering that question might be to pretend that the stones are independently uniformly distributed inside a rectangular frame whose length to breadth ratio is equal to the ratio of the component standard deviations of the configuration, and on that basis one finds that the expected number of triplets meeting the half-degree standard of near-collinearity is about 73, so that relative to approximately Poisson variations (here reasonable) the comment "too many" is unjustified. There is a small excess, but it could quite well be due to chance. I should add that it is preferable to avoid such artificial models and instead to devise a data-based simulation test *of the whole set of 52 sites* employing random lateral perturbations, as was done by W. S. Kendall and myself (Kendall and Kendall, 1980) in a study complementary to that of Broadbent and leading to the same conclusions. Our approach there also avoids the objectionable feature of fixing ε in advance; instead we set a rather broad tolerance region for ε. A detailed account of that and of the method of lateral perturbations would however take us too far from our present theme.

The scheme of our survey will be as follows. In Section 2 we introduce the shape space associated with k labeled points in $\mathbf{R}^m$ and discuss its local metric characteristics, and we illustrate the general discussion by fully identifying some of the simpler shape spaces. In Section 3 we organize the shape spaces in a two-dimensional array and use this to find weak but now global (homology) characteristics of all the shape spaces, with a precision sharp enough to distinguish any one space from all the others. In Section 4 we introduce probability distributions and densities for shape and illustrate this by a brief account of Huiling Le's recent determination of all shape densities for random triangles with vertices uniformly iid in arbitrary compact convex polygons. In Section 5 we turn to a brief account of size-and-shape spaces, and to the more general shape spaces associated with k labeled points in a riemannian manifold M relative to a nicely transitive group of symmetries $\mathscr{G}$. The latter very general situation is illustrated by a discussion of random spherical triangles (important in quasar astronomy). In Section 6 we discuss the size-and-shape problems associated with a random Delaunay tessellation. Finally in Section 7 we outline a few applications.

Much of the work covered by the present survey is still unpublished. It is intended to give a comprehensive account in the book by Carne, Kendall and Le now in preparation.

2. FINDING A NATURAL HOME FOR SHAPES

It was Broadbent's work on the 52 Land's End stones that made me ask the question: what is the natural mathematical home for the shape of a labeled set of k not totally coincident points in m dimensions? (We say labeled points because labels always exist explicitly or implicitly, for example in the form of reference numbers in the archeologist's notebook.) The idea is to filter out effects resulting from translations, changes of scale and rotations and to declare that shape is "what is left."

It is natural first to move the origin to the centroid G of the k points, and then to eliminate size we can compute $L = \sqrt{\sum_{j=1}^{k} GP_j^2}$ (where GP_j denotes the distance from G to P_j) and change the scale by making $L = 1$. This, to a statistician, is the most natural way to standardize for size, but it is not the only possible one, and as we shall see there are contexts in which a different standardization is worth consideration.

This leaves us with an $m \times k$ matrix of rank at most $k - 1$, and to clarify the rank situation we multiply the matrix on the right by a *fixed* element T of the orthogonal group $O(k)$ that maps the column vector $(0, 0, \ldots, 0, 1)$ to a column vector all of whose elements are equal to $1/\sqrt{k}$. The new matrix will then have a final column of zeros. We omit that column, so that we are left with what is now an $m \times (k - 1)$ matrix the squares of whose elements sum to unity, and obviously we can identify this with a point on a sphere of unit radius and $m(k - 1) - 1$ dimensions. That sphere we shall call the sphere of preshapes. Each of its points is identified with an $m \times (k - 1)$ matrix on which the special orthogonal (rotation) group $SO(m)$ acts from the left, and we define the shape space Σ_m^k to be the quotient of the preshape sphere by $SO(m)$. Thus each $SO(m)$ equivalence class in the preshape sphere is now viewed as a single point (by definition the shape of the original configuration) in this new space. Further details are given in Kendall (1984, 1985, 1986).

Notice that while the construction of the shape space depends on an arbitrary choice of T, the effect of varying that choice does no more than replace the first shape space by another isometric with it. Thus any such T in $O(k)$ can be used, but must not thereafter be altered. A once for all choice is suggested in Kendall (1984).

It will be observed that it is in the process of standardization for size that we lose the opportunity to include the totally degenerate k-ad all of whose points are coincident. This does, of course, have a distinct shape, and we can if we wish adjoin it to the shape-space as a non-Hausdorff point. Normally the totally degenerate situation is ignored, but an exception is made when discussing the diffusion of size and shape; the non-Hausdorff point representing total degeneracy is then of importance as an entrance boundary. The true significance of this will become apparent when we extend our definition to cover size-and-shape spaces (for which see below).

Geometrically a maximal set of preshapes equivalent modulo $SO(m)$ forms what is called a *fiber* in the sphere of preshapes, and two k-ads will be said to have the same shape if and only if they determine preshapes lying on the same fiber. It is customary to think of the preshape sphere as lying above the shape space, with the quotient-projection acting downward, so that the whole of each fiber can be thought of as lying above the shape(-point) to which it corresponds. Notice that these fibers do not intersect one another, so that we have a decomposition of the preshape sphere into nonoverlapping fibers. Thus we get the shape space by using the quotient operation that maps fibers *down* onto points (= shapes), and we then throw as much as we can of the natural structure of the preshape sphere down the projection into the shape space. Figure 3 gives a (drastically!) oversimplified sketch of the relationship between (i) the set of k points (here a triangle) in the ambient space (here $\mathbf{R}^2$), (ii) the preshapes and fibers in the preshape space (here a sphere $\mathbf{S}^3(1)$), and (iii) the shapes in the shape space (here a sphere $\mathbf{S}^2(\frac{1}{2})$).

That we can metrize a quotient space via the projection is well known, but we can do better. This is because (away from certain singularities when $m \geq 3$, to be discussed below) the projection that maps fibers to points is here what is called a *riemannian submersion* endowing the shape space with a natural smooth riemannian structure inherited from the ordinary riemannian structure of the preshape sphere in such a way that

> each pair of tangent vectors at any preshape on a given fiber, with the property of being orthogonal to the fiber, will map to a pair of tangent vectors at the image of the fiber in the shape space, *these last tangent vectors having the same inner product as the original "horizontal" pair.*

It is usual and helpful to think of the tangent space at a point on the preshape sphere as decomposed into two orthogonal complements: one contains the "vertical" tangents that are tangent to the fiber, whereas the other contains the "horizontal" tangents that are orthogonal to the fiber, so that the above statement is an assertion about pairs of horizontal tangent vectors at a point in the preshape space.

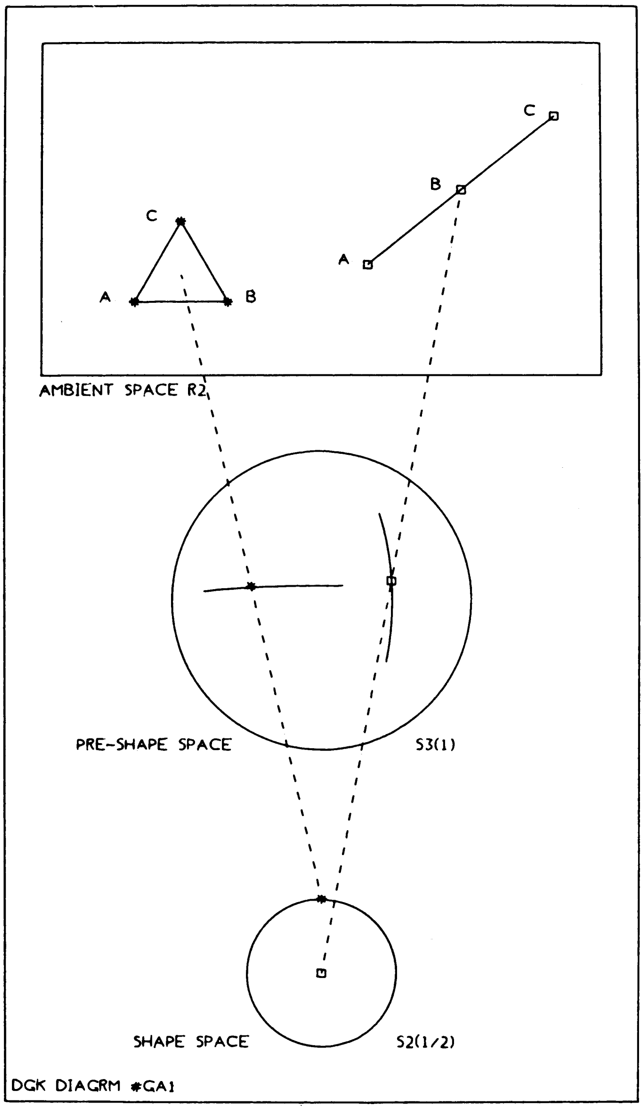

FIG. 3. *The ambient space, the preshape space and the shape space: an impressionistic sketch of the situation when $k = 3$ and $m = 2$.*

This inner product property holds everywhere outside the singular set and imposes a natural riemannian metric on Σ_m^k that has now been determined explicitly; it will be reported on in detail elsewhere. It is relevant that this method when fully implemented also tells us exactly which shapes are located in the singularity set.

Accordingly we immediately obtain local metrical information about the shape space, and obviously to gain a full geometric understanding we shall need to supplement this by other information of a more global character.

One way of meeting that need is to study the geodesics on the shape space. In such a situation as this it is known (O'Neill, 1966, 1967) that the geodesics in the shape space are exactly the projections of the horizontal geodesics (here horizontal great circles in the preshape sphere); moreover, these project with local preservation of arc length, so that the geodesic geometry of the shape space can be read off.

O'Neill has shown that for riemannian submersions one can greatly shorten the computation of sectional curvatures. Another important fact is that the relation between the riemannian structures associated with preshapes and shapes is such that it can also be used to relate diffusions in the two spaces. Here to avoid wearying the statistical reader with technicalities we omit the details.

We now summarize a few specific results for small values of k and m for which the corresponding shape spaces can be fully identified (that is, where they are known up to isometry). Many of these examples are of considerable practical importance.

For $m = 1$ the shape space is the sphere $\mathbf{S}^{k-2}(1)$, where 1 denotes the radius, and for $m = 2$ it is what is known as the complex projective space $\mathrm{CP}^{k-2}(4)$, where 4 denotes the (constant) holomorphic sectional curvature. For $m = 2$ and $k = 3$ we thus find that $\Sigma_2^3 = \mathbf{S}^2(\frac{1}{2})$ (here again the $\frac{1}{2}$ denotes the radius). (This follows because $\mathrm{CP}^1(4) = \mathbf{S}^2(\frac{1}{2})$.) Accordingly $\mathbf{S}^2(\frac{1}{2})$ is the shape space for labeled triangles. It will be observed that the (constant) curvature of Σ_2^3 is equal to 4, although that of the preshape sphere was equal to 1. These facts illustrate a general principle established by O'Neill: this kind of mapping never decreases the curvature.

The singularities that arise when $m \geq 3$ correspond to the k-ads that lie in an $(m - 2)$-dimensional subspace. Thus the singular set is a projective image of Σ_{m-2}^k in Σ_m^k.

The natural generalization of the set of "collinear" labeled triangles in the plane is the set of $(m + 1)$-ads in $\mathbf{R}^m$ that happen to lie in an $(m - 1)$-dimensional subspace, and the corresponding set of shapes is a projective image Eq_m of Σ_{m-1}^{m+1} in Σ_m^{m+1}. In particular when $m = 2$ this tells us that the collinearity set Eq_2 is the projective image $\mathbf{S}^1(\frac{1}{2})$ of $\mathbf{S}^1(1)$ in $\mathbf{S}^2(\frac{1}{2})$, so that it is a special great circle in $\mathbf{S}^2(\frac{1}{2})$. We call this the *equator*, and it is useful to employ that terminology even when m is not equal to 2. It should by now be obvious to the reader that the study of near-collinearities for labeled triplets of points in two dimensions reduces to a study of the shape data in the vicinity of the equator on the shape space $\Sigma_2^3 = \mathbf{S}^2(\frac{1}{2})$. If we use a circularly symmetric gaussian model to describe the random distribution of the original points in $\mathbf{R}^2$ then it turns out that the corresponding distribution of the shape point is uniform on the surface of the spherical shape space. Thus the collinearity studies with which we started have been converted into an elementary exercise in spherical trigonometry.

The projective nestings of shape spaces in other shape spaces briefly illustrated above are very useful in other ways, and are possible because we have chosen to standardize the size measure L to a value (unity) that does not depend on k or m.

We have emphasized that a knowledge of the riemannian geometry of the shape space does not answer all the questions that one needs to ask, because some of those are essentially linked to the global rather than merely local geometrical situation. A complete elaboration of this point would be very technical, but we shall give a sketch of the possibilities in the next section. In preparation for this the reader might like to be reminded of the sort of tools that can be employed.

When one is confronted with a space (such as a shape space) and asks about its geometric structure, there is a hierarchy of levels at which one can operate. At the crudest level one might merely ask for the value of the Euler-Poincaré characteristic χ, probably familiar to all readers in the context of polyhedra in three dimensions. A useful fact to bear in mind is that for spheres this characteristic is equal to 2 when the dimension is even, and is zero when the dimension is odd, these facts holding for any space that is topologically equivalent to a sphere. (For polyhedra in 3 dimensions the dimension is 2, and as this is even the characteristic is also equal to 2.)

Now if one requires slightly more detailed information one can ask for the sequence of homology groups for the space, and here one can operate at three levels, using coefficients drawn from the additive group Z_2 of residues modulo 2, or the additive group Q of rational reals or the additive group Z of signed integers. The last of these three options yields the most detailed information.

After this one could proceed to cohomology, which is a theory dual to homology but with a more detailed structure related to naturally defined product operations, or beyond that to homotopy.

Even when all these possibilities are explored one may find that there is still some lack of detail; one will not necessarily have identified the space up to topological equivalence.

The metrical level at which we were working earlier in this section lies way beyond all the other approaches we have just mentioned, but if it happens to be accessible only in a local form then the additional global information supplied by the "weaker" methods can sometimes provide us with a much more satisfactory result. To illustrate this fact we recall that a complete n-dimensional riemannian manifold with $n \geq 2$ and with positive curvatures K such that $\frac{1}{4} + \varepsilon \leq K \leq 1$ must be a topological sphere $\mathbf{S}^n$ if it is known to be simply connected.

3. THE ARRAY OF SHAPE SPACES Σ_m^k

If we arrange the shape spaces in an array labeled by (k, m), where the number of points $k \geq 2$ increases down the columns and the dimension $m \geq 1$ increases along the rows, then all the spaces along the diagonal $m = k - 1$ are topological spheres (this important result was discovered by A. J. Casson). Clearly however they cannot be metric spheres when $m \geq 3$ because there are then singularities in the differentiable structure. Notice that these "diagonal" shape spaces are the ones needed for the discussion of the shapes of labeled simplexes. Their statistical importance is therefore considerable.

Beyond this diagonal, that is when $m > k - 1$, the shape spaces in the kth row are all metrically the same and topologically they can be identified with a "hemisphere" of the topological sphere Σ_{k-1}^k. (There are two such "hemispheres" in this topological "sphere" that are metrically congruent under a reflection operation and intersect in what we have called Eq_{k-1}.) The reflection referred to is that induced by a reflection of the original configuration in one of the coordinate planes of $\mathbf{R}^m$.

These "hemispheres" and "equators" play an important role when one studies the shape diffusion induced via a time change by a given k point diffusion in the ambient space $\mathbf{R}^m$. The paper by W. S. Kendall already mentioned starts with a brownian or an Ornstein-Uhlenbeck diffusion for a set of $k = 3$ points in $\mathbf{R}^m$, where $m = k - 1, k, k + 1, \ldots,$ and then examines the corresponding time-changed diffusions in Σ_m^k as was done in a very tentative way in (Kendall, 1977). The first of these shape spaces is a topological sphere, and all the others are metric copies of one and the same "hemisphere" whose boundary is the "equator" defined above. But the successive diffusions in this "hemisphere" are not the same, and that fact leads to interesting and indeed surprising conclusions about the nature of the stochastic motion when the ambient dimension m tends to infinity.

There remains a triangular region of the array defined by the inequalities $m \geq 3$ and $m < k - 1$ about which we have so far said nothing, but I can now make a fairly complete statement concerning the homology-properties of the spaces in question. (i) None of these spaces is a sphere even in the crudest sense; i.e., none has the homology of a sphere. (ii) None is a topological manifold. (iii) All have torsion in homology. (iv) Any two such spaces are homologically distinct.

The details are long and complicated, but depend chiefly on the interesting fact that if we write

$$\Sigma_m^k \leq \Sigma_n^l$$

when both $k \leq l$ and $m \leq n$, then there is an elegant topological three-term recurrence that constructs Σ_m^k up to homeomorphy out of the topological spaces Σ_{m-1}^{k-1} and Σ_m^{k-1}, and so up to homeomorphy we can (in principle) construct all the shape spaces inductively, following the partial order and starting with the spaces at the margins of the array (for the nature of these is already known).

For the Euler-Poincaré characteristics χ^k_m we obtain a related two-dimensional numerical three-term recurrence that can be solved explicitly. We then find that the triangular region mentioned above contains some even-dimensioned spaces with $\chi = 2$ (as for even-dimensioned spheres) and some odd-dimensioned spaces with $\chi = 0$ (as for odd-dimensioned spheres). Nevertheless these are not topological spheres; for the proof of that we require the corresponding recurrence in homology.

In homology we get a short exact sequence that extends to a long exact sequence with homomorphisms whose action can be fully identified. In this way I have found the Z_2 homology explicitly for all the shape spaces Σ^k_m and a corresponding determination of the Q homology is in progress. Fitting these together via what is called the general coefficient theorem should yield the integer homology, but for our present purposes the Z_2 results suffice; in particular they immediately prove (i) and (iv) above.

One might expect that the shape spaces possessing singularities would prove to be of no practical interest, but that is not so. In fact the spaces Σ^k_3 are precisely those that are important in the work of Clifford, Green and Pilling (1987) on stochastic problems in physical chemistry.

4. RANDOM SHAPES: CONVEX-POLYGONALLY GENERATED SHAPE DENSITIES

It might seem that the concentration on the geometry of shape spaces is excessive, but recent events have justified it in a striking way. When this program began in the 1970s, C. G. Small and I knew that a diffuse probability law $\mathscr{L}$ in $\mathbf{R}^m$ must determine in a natural way an induced law $\mathscr{L}^*$ on the shape space Σ^k_m, this being the law of distribution of the shape of a labeled k-ad of points each one of which is independently distributed with law $\mathscr{L}$, but we only knew one or two examples of such situations that we were able to study in explicit detail. It therefore seemed desirable to find a wide range of such explicit shape distributions, at least in the basically important case $k = 3$, $m = 2$. In particular we thought it would substantially remedy the situation if we could find the shape distribution for three points independently uniform in (i) a square and perhaps (ii) an equilateral triangle.

Unfortunately, what we thought a modest objective proved for 10 years unattainable, even in the "simple" cases just mentioned. To assist the mechanics of such calculations we used a stereographic projection of the sphere Σ^3_2, projecting it from the shape point where "$A = B \neq C$" onto the tangent plane at the shape point where "C is the midpoint of AB." This is a plane projection, and in the plane we took cartesian coordinates (x, y) such that the shape point where "$A = C \neq B$" had the coordinates $(-1/\sqrt{3}, 0)$, and the shape point where "$A \neq C = B$" had the coordinates $(+1/\sqrt{3}, 0)$. The occurrence here of $\sqrt{3}$ may seem peculiar, and some writers (e.g., Small, 1981, 1988) avoid it, but it is the kind of notational wrinkle that remains however much one pushes it under the carpet, and I prefer to accept it at this point for the sake of getting cleaner formulas elsewhere. With these coordinates we find that $y = 0$ is the locus of all collinearities apart from the one corresponding to the shape projected to the point at infinity. Thus, $y = 0$ together with the point at infinity is the stereographic version of Eq_2 and near-collinearity studies focus on its immediate neighborhood.

With these conventions it is natural to seek an explicit form for the shape density $m(x, y)$, that is the Radon-Nikodym derivative of the shape measure relative to the σ-finite measure $dxdy$. After many years of unsuccessful attempts by myself to find $m(x, y)$ in the two "simple" cases, the situation has changed dramatically with the work of the young Chinese mathematician Huiling Le, who succeeded in obtaining explicit formulas for the function $m(x, y)$ whenever the probability model is

three points A, B, and C are independently and uniformly distributed inside a compact convex polygon K.

Her solution (1987a, b; see also Kendall and Le, 1986, 1987a) is perfectly general, and covers all compact convex polygons K whatsoever.

What led to this remarkable achievement was the observation that for given K the function $m(x, y)$ is real-analytic inside each tile of a K-dependent "singular tessellation" $\mathscr{T}$ of the (x, y)-plane and jumps abruptly in analytic form when any edge of the tessellation is crossed. As there is a continuum of possible shapes for K, and so a continuum of possible tessellations, this seemed at first to make a complete solution even more unattainable, but the tradition of doing the geometry first, and then tackling the probability calculations when that was fully understood, turned out to be the key to the situation. The geometric dependence of $\mathscr{T}$ on the shape of K was fully investigated (Kendall and Le, 1987a, b), and once that was done the outline of what might be a possible way to a solution came into view, although many difficulties had to be overcome before this intuition was shown to be correct.

Next, the jump suffered on crossing any tile-edge of the tessellation was shown via an analytic continuation argument to be completely characterized by a real analytic *jump function* attached to that tile-edge and free of singularities in a two-dimensional open set containing the edge after removal of its end points. Thus, if one were given the function $m(x, y)$ in some "basic tile" and if also one knew all the jump functions, then it could in principle be extended by analytic continuation to an arbitrary "target" tile by a "stepping-stone" procedure following a sequence of pairwise contiguous tiles. Moreover, it was clear that if the argument could be pushed through then all such stepping-stone routes would provide equally good ways of arriving at a solution, although in practice one would expect some routes to be more convenient than others.

An equally vital and surprising step was Huiling Le's discovery that the problem can in fact be reduced to finite form. She followed this by finding explicitly (a) the function $m(x, y)$ in a particular so-called "basic" tile for all polygons K, and (b) the jump functions for all tile-edges of all tiles in the tessellations associated with all polygons K. Her solution has now been implemented in a computer-algebra language, although in practice the choice of the stepping stone route itself is best performed by eye after inspecting the tessellation. The successive contributions to the (finite) stepping-stone expansion of $m(x, y)$ can contain algebraic terms having singularities in the target tile, but the theory guarantees that in a computer-algebra implementation all such apparent singularities in the target tile will automatically cancel out when their sum is formed, to give a version of $m(x, y)$ that is real analytic inside that tile.

Figure 4 shows a typical tessellation; here K is an irregular pentagon. The shape density $m(x, y)$ is C^2 smooth save at the three shapes corresponding to coincidences among the triangle vertices, and the presence of the jump functions at the tile-edges is betrayed by jumps in the 3rd or 4th normal derivatives. It was indeed the detection of these (using central differences) in earlier numerical studies (Kendall and Le, 1986) that first gave us the necessity insights into the geometrical and analytical structure underlying the tessellations.

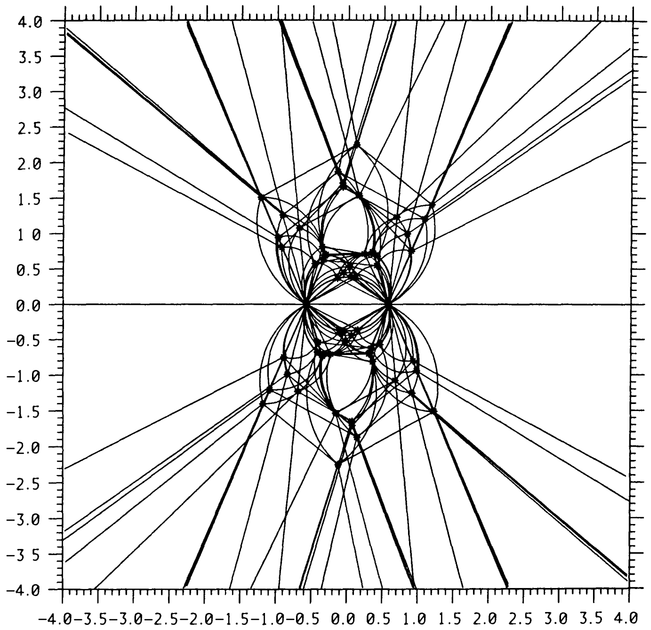

FIG. 4. *A singular tessellation for an irregular pentagon. (Reproduced by permission from Kendall and Le,* 1987a.)

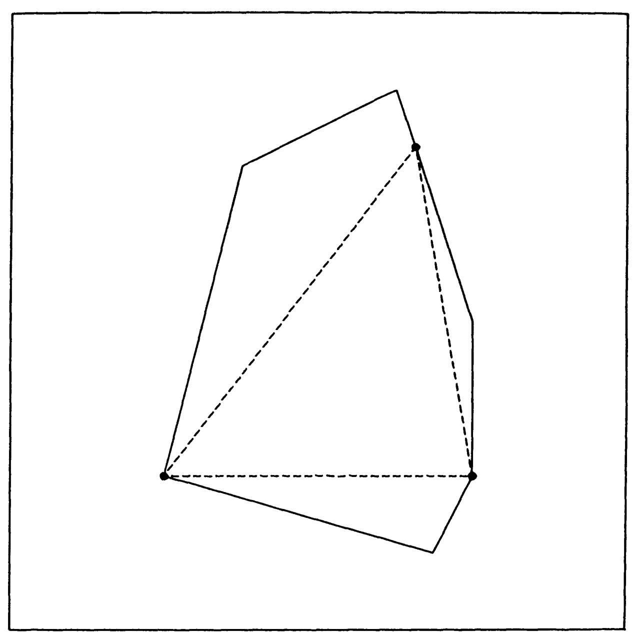

FIG. 5. *A configuration in the ambient space that projects to a shape lying on an edge of the singular tessellation.*

The reader may be curious to know whether it is possible to discern any systematic structure at all in Figure 4, and the following remarks are intended to reveal at least some of this; the whole story is a long one. On each edge of $\mathscr{T}$ each point is the shape of a triangle ABC such that two of A, B, and C lie at vertices of the polygon K, while the third lies on an edge of the polygon K, and conversely all such situations, with the triangle ABC arbitrarily labeled, will generate a shape on some linear or quadratic edge of $\mathscr{T}$. Figure 5 illustrates one possible situation, but of course there are many combinatorially distinct cases that have to be considered, and their classification yields further useful structural information concerning both $\mathscr{T}$ itself and the jump functions associated with it.

An interesting subsidiary question is: can we find the shape of K when the associated shape-density $m(x, y)$ is known? We can prove that the answer to this is affirmative if we exclude a nowhere-dense subset of the shape space $\Sigma_2^n = CP^{n-2}(4)$, where n is the number of vertices of K. The proof of this depends on the geometrical fact that the tessellation $\mathscr{T}$ is made up of three components: finite segments, semi-infinite half-lines and arcs of circles. If the semi-infinite half-lines and arcs of circles are removed, then what remains is the superposition of $n(n-1)$ scaled, rotated and shifted copies of K itself (built out of the finite segment components). The residual ambiguity is associated with the peeling apart of these $n(n-1)$ copies of K and can always be resolved if it is given in advance that the ratios of interpoint distances between all pairs of vertices of K are distinct.

This result contrasts surprisingly with a general result of Small (1981) telling us that the solution to the corresponding inversion problem can be nonunique when the common law $\mathscr{L}$ of the three triangle vertices is arbitrary (see also Small, 1988).

In this discussion of shape densities it has been supposed that the k original points $P_1, P_2, \ldots, P_k$ determining the shape are given a definite labeling, which could be either intrinsic or assigned at one's convenience. When as often happens the labeling is arbitrary, or of subsidiary importance, one is free to quotient out the permutation group on k letters to obtain a reduced shape space, but that is usually less well behaved, and such additional quotienting is normally avoided except in special circumstances. One such is the situation in which information about an unknown shape density $m(x, y)$ is to be obtained by simulating k-point configurations in m dimensions and then recording their shapes as a preliminary to plotting scatter diagrams, contour plots, etc. in the shape space or some transform of it. (Many examples of such plots will be found in Kendall, 1984.) In such circumstances an extra factor $k!$ in the effective size of the simulation can be gained by exploiting the relabeling group. Having established the result on the reduced shape space it will then often be convenient to construct its equivalents in the other $k! - 1$ permutation transforms in order to bring out more clearly the global structure that is being studied.

5. SIZE AND SHAPE SPACES AND THE SHAPES OF SPHERICAL TRIANGLES

Some of the above is now in a sense rather old work, because during the last 2 years interest in the Cambridge group has gradually shifted away from the shape spaces Σ_m^k toward the associated size and shape spaces here provisionally denoted by $\mathbf{S}\Sigma_m^k$, and also toward the shape spaces $\Sigma(M, \mathscr{G}, k)$ derived from an ambient space that is a riemannian manifold M (instead of $\mathbf{R}^m$) and an appropriate group $\mathscr{G}$ (instead of $SO(m)$) with respect to which the quotient operations take place. There is evidently a connection with the moduli spaces of the algebraic geometers, but this does not seem to lead to any further insights.

We note in passing that the size and shape space $\mathbf{S}\Sigma_m^k$ corresponding to Σ_m^k contains a point $*$ corresponding to the totally degenerate k-ad that was omitted from Σ_m^k itself; this is the point corresponding to size $L = 0$. In fact $\mathbf{S}\Sigma_m^k$ turns out to be a cone with a warped-product metric; the vertex of the cone is the point $*$, and each section of the cone is a scaled version of the shape space Σ_m^k. For m greater than unity $*$ is itself a singularity, and the remaining singularities are all the points on the rays meeting each section in a singularity of the shape space. The metrical theory for $\mathbf{S}\Sigma_m^k$ now follows immediately from these structural remarks, which have important implications for the applications to physical chemistry mentioned earlier.

An instructive example with a noneuclidean ambient space is $\Sigma(\mathbf{S}^2(1), SO(3), 3)$. This is the space of spherical triangles with labeled vertices, but it is not the space of spherical triangles that was studied by Grace Chisholm Young in her celebrated Göttingen doctoral thesis under Felix Klein in 1895. (She considered a triangle with sides that need not be minimal geodesics; also its sides were allowed to intersect at points other than the vertices.) Topologically our space is $\mathbf{S}^3$; this is easily demonstrated by a technique using the properties of identification topologies. But viewed within the differentiable category it possesses four point-singularities; one of these is the shape of total coincidence, whereas the other three correspond to the situations in which two of the vertices are coincident and the third is antipodal to them. Another interesting feature of this space is that it is in effect a size and shape space, because size for spherical triangles is just an aspect of shape. Moreover "location" is now irrelevant, because a change of location can be effected by using the group SO(3). A determined attack on this problem has been made during the last year by Carne, Huiling Le, and myself. We have now obtained (by three different methods, two involving computer-algebra) the riemannian structure, the sectional curvatures, the brownian differential generator, the geodesic geometry and the different but related metric that arises from a parallel procrustean study. We therefore now know almost as much about this shape space as we do about Σ_2^3. Still more recent investigations by Carne and by Huiling Le have extended many of these results to the shape space for k points in $\mathbf{S}^m$.

It is thus appropriate to turn to the statistical problem of finding interesting shape measures on $\Sigma(\mathbf{S}^2(1), SO(3), 3)$, and in the last few months Huiling Le has found the probability law for shapes of spherical triangles whose (labeled) vertices are independently uniform inside a spherical cap of angular radius α $(0 < \alpha \le \frac{1}{2}\pi)$. (The upper bound on α is needed because otherwise it could happen that A and B lie in the cap although some part of the geodesic arc AB does not.) The whole bundle of these calculations taken together puts us in a position to resolve a problem of interest in quasar astronomy; this is, how should one analyze the claims that "too many" triplets of quasars lie on or suspiciously close to arcs of great circles on the celestial sphere. Or, to put it more roughly, how should one analyze the evidence for there being "too many" triplets of "nearly collinear" quasars. It is known that the effect of the curvature of the celestial sphere is not negligible here. We are now in a position to make an accurate assessment of it, by examining the above results for small α and comparing them with comparable results (again due to Huiling Le) for $\mathbf{S}\Sigma_2^3$. Note that size must come into this comparison, on the one hand because it is not separable from shape in the spherical triangle context, and on the other hand because in the collection of the astronomical data a selection for size will have been exercised.

6. RANDOM DELAUNAY TRIANGLES

Another problem in which size plays a significant role is that in which one studies the shapes of the simplexes that are the tiles of the Delaunay tessellation of a realization of an m-dimensional Poisson point process. We shall call these simplexes PDLY tiles, for short. The present account summarizes (Kendall, 1983, 1989) and adds some more recent results.

We first recall that the Delaunay tessellation of a (suitable) infinite set of distinct isolated points in $\mathbf{R}^m$ was introduced by the Soviet number theorist Boris Nikolaevitch Delone (1890–1980). The construction goes as follows. We look at each $(m+1)$-ad of points in turn. If its circumsphere contains a point of the set in its interior we do nothing, but if its circumsphere is "empty" in that sense then we draw in the simplex determined by these $m+1$ points. When all $(m+1)$-ads have been examined in this way we obtain a nonoverlapping covering of $\mathbf{R}^m$, and that is the Delaunay tessellation, the component simplexes being the tiles thereof. We omit the necessary restrictions on the original set of points but remark that with probability one they will all be satisfied if we apply the construction to a realization of an m-dimensional Poisson process.

At first sight this looks like a problem involving Σ_m^{m+1}, but that turns out to be a partly misleading clue. Some years ago Miles (1970, 1974) proved that for PDLY tiles the circumradius R (which has a scaled χ^2 distribution) is statistically independent of the complete set of shape variables, so that *if* we use R as the measure of size, then size and shape will be independent. This suggests that the connection with Σ_m^{m+1} should be abandoned altogether, but that too would be a mistake. A more fruitful procedure is to compare

(i) the shape distribution for PDLY tiles, and
(ii) the shape distribution for a simplex with independent Gaussian vertices.

This gives us two shape measures μ_1 and μ_2, say. I have proved that the Radon-Nikodym density $d\mu_1/d\mu_2$ is of the form

$$c_m/\rho^{m^2},$$

where c_m is a known function of m only, and where $\rho = R/L$ (note that this is a shape variable with $1/\sqrt{m+1}$ as its minimum value). It follows that on constructing by simulation an independent sequence of simplexes with independent Gaussian vertices, and at each step computing ρ and using the obvious acceptance-rejection rule based on

$$(\rho_{\min}/\rho)^{m^2},$$

then the resulting sequence of accepted simplexes will be a sequence of independent PDLY tiles.

In other words, we have found a way of creating "lone" PDLY tiles without doing any tessellating, with consequent immense gains in speed! This work is still in progress, but a few comments will illustrate what has so far been done.

First, the procedure just outlined works spectacularly well for dimensions m from 1 up to about 6 or 7. After that the random sample size obtained falls off drastically because the chance of acceptance (which is known exactly) tends rapidly to zero as m tends to infinity, and so eventually nearly all the Gaussian simulations are rejected. Recently Miles has pointed out to me that such "lone tile simulations" could also be carried out in another way; there μ_2 is to be replaced by μ_3, the shape measure for a simplex whose vertices are independently uniform on a unit sphere. A decomposition formula given by Miles (1974) then yields the Radon-Nikodym derivative, and thereafter one proceeds as before. It will be interesting to see how these two methods compare. I have also looked at the effect of replacing the acceptance/rejection rule by an importance-sampling procedure, the (normed) weights being proportional to the Radon-Nikodym derivatives. Here every simulation is retained, but the weights vary wildly because of their dependence on ρ, and so a suitable smoothing procedure has to be used. My experiments so far show that useful information on the distribution of shape can be obtained by this technique even when $m = 10$. This is remarkable when we consider that, from yet another formula established by Miles (1974), at dimension 10 the expected number of PDLY tiles having a given point as vertex is about

100 million. Thus at $m = 10$ we are working with a stochastic tessellation of PDLY tiles of fantastic complexity, even when viewed from such a narrowly local standpoint. It will be interesting to see how much more insight we can gain when trying new ways of studying this formidable stochastic object.

The main conclusions arrived at so far are that a typical PDLY tile is (1) more likely to be nearly regular (i.e., equilateral) and (2) less likely to be nearly degenerate than is the case for a typical Gaussian simplex. Figures 6 and 7 illustrate a few of the results that have been obtained in this way for $m = 2$, 3, and 6. I hope, by pushing the simulation technique out to higher dimensions, and by supplementary asymptotic calculations, to get some idea of what happens as $m \longrightarrow \infty$.

One would very much like to be able to make comparable statements about groups of "adjacent" tiles in the tessellation, but although some limited progress is possible, this problem seems to be exceedingly difficult. It will be observed that the whole tessellation could be thought of as a somewhat novel form of stochastic field. This remark does not appear to be very helpful, however.

A similar comment can be made about the analysis of the Land's End data. The locations of the 52 stones should strictly be regarded as identifying a single point in the enormous space Σ_2^{52}, and the associated

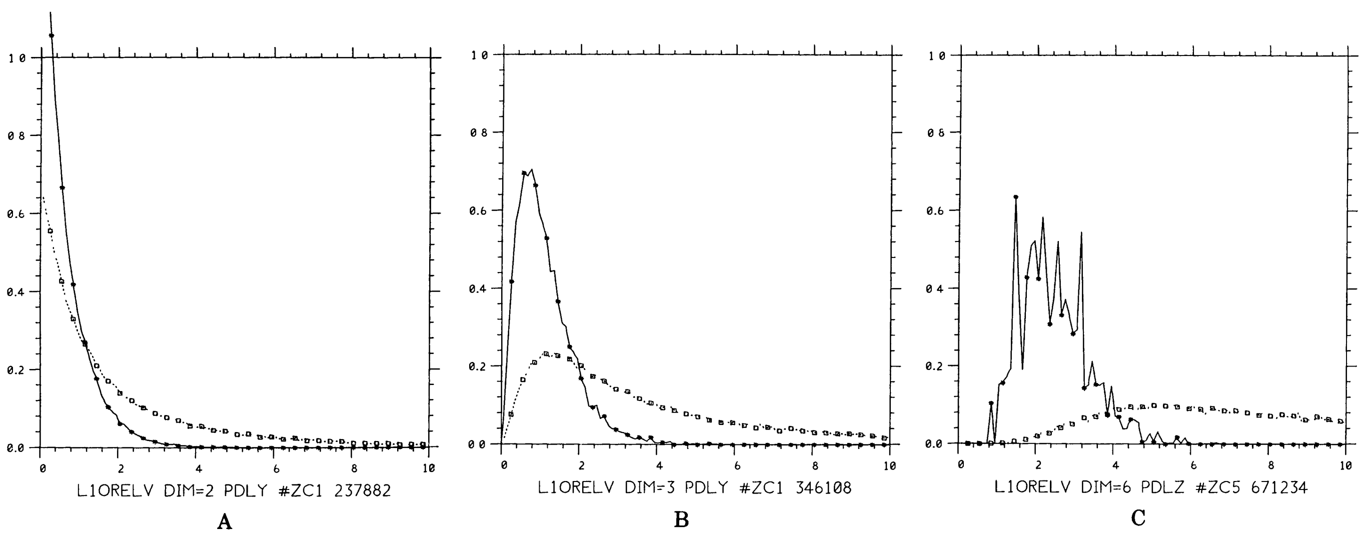

FIG. 6. *Distribution of* $\ln(V_{\text{reg}}/V)$. *Here V is the volume of the PDLY simplex and* V_{reg} *is the volume of a regular simplex of equal circumradius. Dotted line, Gaussian simplex; full line, PDLY simplex; m* = 2, 3, 6. *Equilaterals to the left of diagram, splinters to the right.*

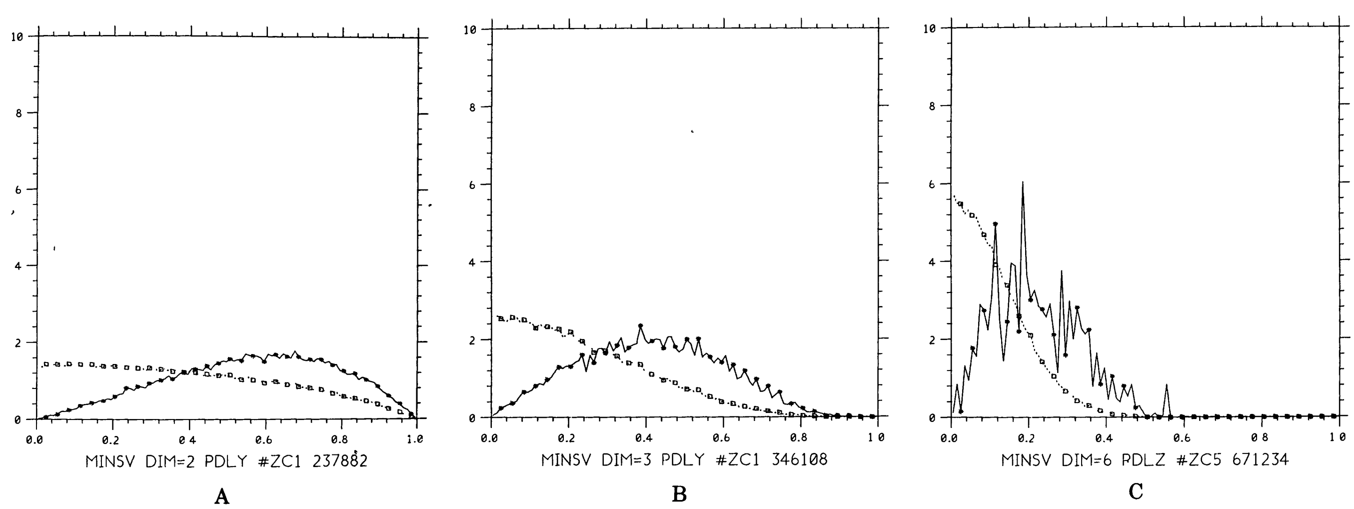

FIG. 7. *Distribution of* $\lambda\sqrt{m}$ (λ = *least singular value*). *Dotted line, Gaussian simplex; full line, PDLY simplex; m* = 2, 3, 6. *Equilaterals to the right of diagram, splinters to the left.*

stochastic field is then governed by the appropriate shape measure (given for a Gaussian model in Kendall, 1984) on that shape space.

7. SOME APPLICATIONS

I conclude with a few notes on applications.

It is well known that no classical test for two-dimensional stochastic point processes can match the performance of the human eye and brain in detecting the presence of improbably large holes in the realized pattern of points. This fact has generated a great deal of research in the last few years, especially in connection with the large "voids" and long "strings" that the eye sees (or declares that it sees) in maps of the Shane and Wirtanen catalogue of positions of galaxies (see for example Moody, Turner and Gott, 1983). Astronomers are interested in (i) whether these phenomena are sufficiently extreme to require explanation, and if so (ii) whether any of the various "model" universes now in vogue can be said to display them to just the same degree. Recently Icke and van de Weijgaert (1987) have suggested that useful progress might be made by studying the two- and three-dimensional Delaunay tessellations generated by the galaxy positions, and in particular by examining the observed distributions of various size and shape characteristics for the Delaunay triangles and tetrahedra. The investigation summarized in Section 6 was planned as a contribution to this enquiry.

There is another interesting application of the Poisson-Delaunay theory to geography. Geographers studying the spatial distribution of human settlements claim to see an underlying quasihexagonal structure and speak of "central-place theory." Some years ago Mardia, Edwards and Puri (1977) pointed out that this effect, if it exists, should increase the proportion of nearly equilateral Delaunay tiles. Now we have seen in Figures 6A and 7A that the Delaunay tessellation of a two-dimensional Poisson distribution will in any case contain a high proportion of nearly equilateral tiles, so that a small excess of this as the result of other causes might not be easy to detect.

In fact (Figures 8 and 9) there is indeed a striking number of nearly equilateral tiles in the Delaunay tessellation of 234 towns, villages and hamlets in Wisconsin, the other noticeable feature of that data set being a high proportion of thin splinter-shaped tiles round the edges of the region being tessellated. Of course these latter tiles are not true Delaunay tiles at all; they arise solely because their circumcircles lie mostly outside the region, and so they are "empty" in virtue of the cut-off at the edges.

Now central place theory also has something to say about distances. Thus one mechanism that has been invoked to explain central place effects in East Anglia is the tendency for neighboring market towns to be separated by the maximum distance over which one can drive sheep in a day. (I owe this remark to Dr. G. P. Hirsch.)

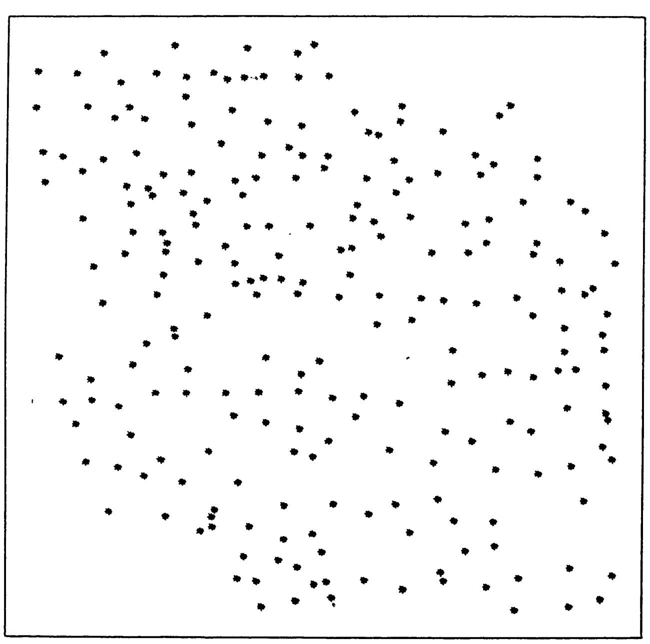

FIG. 8. *Locations of* 234 *settlements in Wisconsin.* (*Data provided by A. D. Cliff and based on Brush,* 1953.)

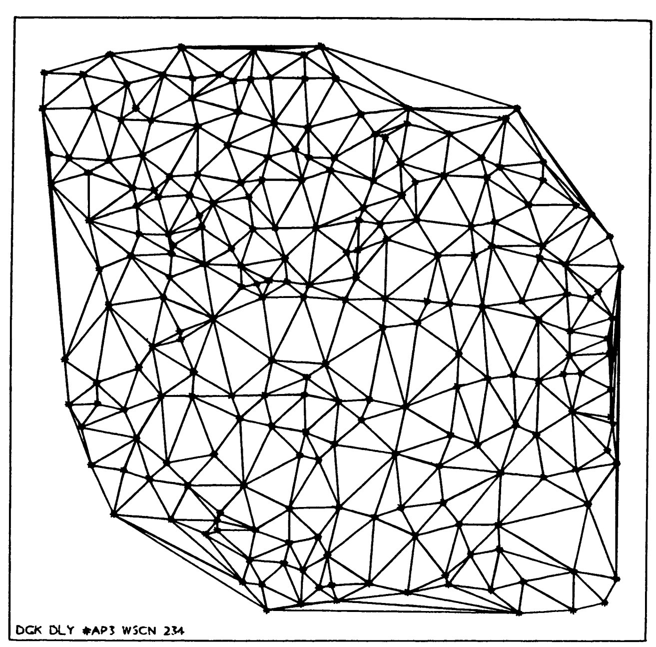

FIG. 9. *The Delaunay tessellation for the Wisconsin sites.*

Accordingly we ought to treat this as a problem that belongs to size and shape theory rather than to shape theory, and this presents no difficulties because for PDLY tiles in two dimensions the size (measured by the circumradius R) has a simple distribution; in fact R^2 has the law of a scaled χ_4^2, the scaling constant being known. The shape distribution in this situation is known from the work of Miles, and as remarked earlier size and shape are here statistically independent.

This suggests a new approach to such data, as follows. (i) Sort the tiles according to their shape, and select (a) those that by some convenient angular criterion are "nearly equilateral," then (b) those that are highly splinter-shaped and (c) the remainder. Then (ii) look at the distribution of circumradial size within the sets (a) and (c) and examine the departures from independence and from the theoretical χ^2 law.

This procedure has the attractive feature that it will not be seriously corrupted by the excess of splinter-shaped tiles that are associated with the edge effects. Normally in spatial statistics, edge effects are very difficult to deal with. Here we may be lucky!

Finally in Figure 10 we show the Delaunay tessellation for the 52 locations of the Land's End stones; the reader will perhaps find the comparison with Figure 2 instructive.

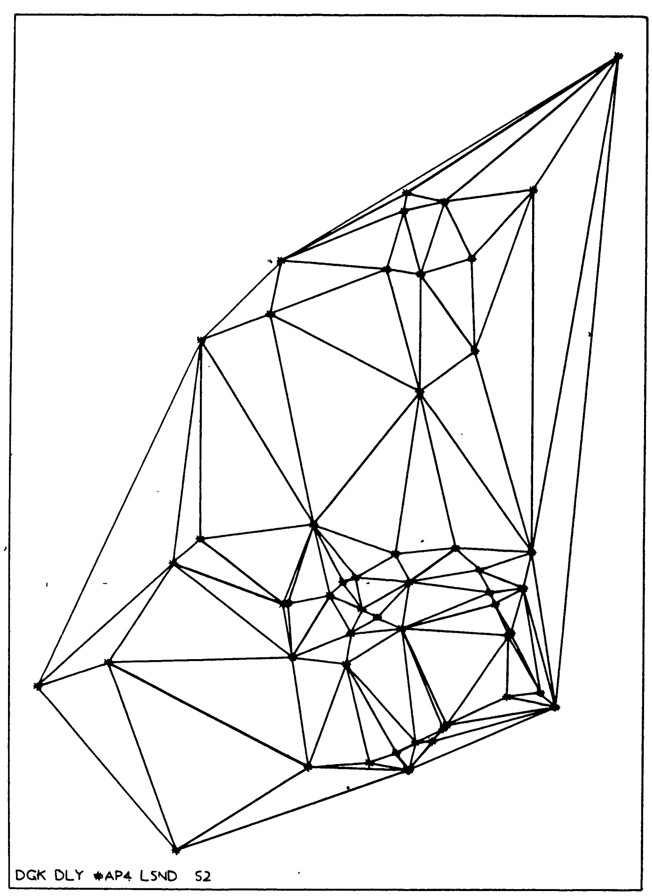

FIG. 10. *The Delaunay tessellation for the 52 Land's End sites.*

ACKNOWLEDGMENTS

This article is based on a lecture delivered to the Deutsche Mathematische Vereinigung in Berlin in September 1987, and it has been adapted, with permission, from the original version published in the *DMV Jahrbuch.*

The research was carried out with support from the Royal Society of London (T. K. C.), Darwin College, Cambridge (H. L. L.) and the Science and Engineering Research Council (H. L. L. and D. G. K.). I wish to thank Simon Broadbent and Andrew Cliff for supplying me with data, and Keith Carne, Wilfrid Kendall and Huiling Le for many discussions covering the whole field of work presented here.

REFERENCES

BOOKSTEIN, F. L. (1986). Size and shape spaces for landmark data in two dimensions (with discussion). *Statist. Sci.* **1** 181–242.

BORSUK, K. (1975). *Theory of Shape.* PWN, Warsaw.

BROADBENT, S. R. (1980). Simulating the ley-hunter (with discussion). *J. Roy. Statist. Soc. Ser. A* **143** 109–140.

BRUSH, J. E. (1953). The hierarchy of central places in southwestern Wisconsin. *Geog. Rev.* **43** 380–402.

CARNE, T. K., KENDALL, D. G. and LE, H.-L. *The Statistical Theory of Shape and Its Applications.* Wiley, Chichester. In preparation.

CLIFFORD, P., GREEN, N. J. B. and PILLING, M. J. (1987). Statistical models of chemical kinetics in liquids (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 266–300.

ICKE, V. and VAN DE WEIJGAERT, R. (1987). Fragmenting the Universe. I. Statistics of two-dimensional Voronoi foams. *Astronom. Astrophys.* **184** 16–32.

KENDALL, D. G. (1977). The diffusion of shape. *Adv. in Appl. Probab.* **9** 428–430.

KENDALL, D. G. (1983). The shape of Poisson-Delaunay triangles. In *Studies in Probabilities and Related Topics, etc.* (M. C. Demetrescu and M. Iosifescu, eds.) 321–330. Nagard, Montreal.

KENDALL, D. G. (1984). Shape manifolds, procrustean metrics and complex projective spaces. *Bull. London Math. Soc.* **16** 81–121.

KENDALL, D. G. (1985). Exact distributions for shapes of random triangles in convex sets. *Adv. Appl. Probab.* **17** 308–329.

KENDALL, D. G. (1986). Further developments and applications of the statistical theory of shape. *Theory Probab. Appl.* **31** 407–412.

KENDALL, D. G. (1989). Random Delaunay simplexes in $\mathbf{R}^m$. *J. Statist. Plann. Inference.* To appear.

KENDALL, D. G. and KENDALL, W. S. (1980). Alignments in two-dimensional random sets of points. *Adv. in Appl. Probab.* **12** 380–424.

KENDALL, D. G. and LE, H.-L. (1986). Exact shape-densities for random triangles in convex polygons. *Analytic and Geometric Stochastics: Adv. in Appl. Probab.* suppl. 59–72.

KENDALL, D. G. and LE, H.-L. (1987a). The structure and explicit determination of convex-polygonally generated shape-densities. *Adv. in Appl. Probab.* **19** 896–916.

KENDALL, D. G. and LE, H.-L. (1987b). A brief account of the structure and explicit determination of convex-polygonally generated shape-densities. *Proc. First Internat. Congress Bernoulli Soc.* **1** 611–613.

KENDALL, W. S. (1988). Symbolic computation and the diffusion of shapes of triads. *Adv. in Appl. Probab.* **20** 775–797.

LE, H.-L. (1987a). Explicit formulae for polygonally-generated shape-densities in the basic tile. *Math. Proc. Cambridge Philos. Soc.* **101** 313–321.

LE, H.-L. (1987b). Singularities of convex-polygonally generated shape-densities. *Math. Proc. Cambridge Philos. Soc.* **102** 587–596.

MARDIA, K. V., EDWARDS, R. and PURI, M. L. (1977). Analysis of central place theory. *Bull. Internat. Statist. Inst.* **47** (2) 93–110.

MILES, R. E. (1970). On the homogeneous planar Poisson point process. *Math. Biosci.* **6** 85–127.

MILES, R. E. (1974). A synopsis of 'Poisson flats in euclidean spaces'. In *Stochastic Geometry* (E. F. Harding and D. G. Kendall, eds.) 202–227. Wiley, London.

MOODY, J. E., TURNER, E. L. and GOTT, J. R. (1983). Filamentary galaxy clustering: A mapping algorithm. *Astrophys. J.* **273** 16–23.

O'NEILL, B. (1966). The fundamental equations of a submersion. *Michigan Math. J.* **13** 459–469.

O'NEILL, B. (1967). Submersion and geodesics. *Duke Math. J.* **34** 363–373.

SMALL, C. G. (1981). Distributions of shape and maximal invariant statistics. Ph. D. dissertation, Univ. Cambridge.

SMALL, C. G. (1982). Random uniform triangles and the alignment problem. *Math. Proc. Cambridge Philos. Soc.* **91** 315–322.

SMALL, C. G. (1988). Techniques of shape analysis on sets of points. *Internat. Statist. Rev.* **56** 243–257.

Further references will be found in the survey papers by Bookstein (1986), Kendall (1984, 1986, 1989) and Small (1988).

Comment

Fred L. Bookstein

The elegant metric geometry of David Kendall's shape spaces Σ_m^k is inherited from the Euclidean metric of the spaces $\mathbf{R}^m$ containing the original point data. In the applications he has sketched here, the points in $\mathbf{R}^m$ are independent and identically distributed (iid) and the metric in shape space, in turn, is symmetric in the points, a sort of spherical distance. Point data generated in other disciplines, however, are not always iid; different metrics may be appropriate to those applications. In this comment I justify a certain analysis of small regions of Kendall's shape space by using a metric quite different from the usual Euclidean-derived version, depict its relation to Kendall's metric and indicate the sort of inquiries it permits.

Morphometrics is the quantitative description of biological form. Its data can often be usefully modeled as sets of labeled points, or landmarks, that correspond for biological reasons from organism to organism of a sample (Bookstein, 1986). We say that these points are biologically homologous among a series of forms: they have identities—names—as well as locations in some Cartesian coordinate system. Any set of landmark locations has a "size" and a "shape" that may be construed according to Kendall's definitions. But the biological relations among different instances of such configurations partake of a feature space not effectively represented by the metric inherited from $\mathbf{R}^2$ or $\mathbf{R}^3$.

Fred L. Bookstein is Research Scientist, Center for Human Growth and Development, University of Michigan, Ann Arbor, Michigan 48109-0406.

In the biological context, my style of statistical analysis of shapes proceeds, as Kendall pointed out in 1986, within a tangent space of his Σ_2^k or Σ_3^k in the vicinity of a sample "mean form." (Small (1988) has an interesting comment on this construction.) The questions that in Kendall's applications are asked of an entire shape space—questions about concentration upon the "collinearity set," and the like—are replaced in morphometric applications by the more familiar concerns of multivariate statistical analysis: differences of mean shape, covariances involving shape or factors that may underlie shape variation.

In the linearization of Kendall's shape space that applies to this tangent structure, the natural shape metric is an algebraic transformation of the "Procrustes metric," the ordinary summed squared Euclidean distance of two-point configurations after an appropriate optimizing rotation and scaling. But the Procrustes approach is not flexible enough fairly to represent biological structure within the context of multivariate statistical analysis. If two landmarks are typically close together, like the pupil of the eye and the outer corner, then we expect them to move together in their relation to more distant structures. The half-width and the orientation of the eye are more tightly controlled by diverse biological processes of regulation than is, say, the distance from the eye to the chin. These considerations lead one naturally to search out a shape metric that weights changes in small distances more heavily than changes in larger ones. In 1985, David Ragozin of the University of Washington suggested to me that the formalism of