PREFACE

Rapid developments in computer technology and computational techniques, advances in a wide spectrum of technologies, and other advances have resulted in cross-disciplinary pursuits between technology and its applications to human body processes and systems. This set of four volumes with the overall title, "Computational Methods in Biophysics, Biomaterials, Biotechnology and Medical Systems", represents, to my knowledge, the first multi-volume treatment of this broadly significant subject on the international scene.

The subtitles of the four volumes are:

Volume 1: Algorithm Techniques
Volume 2: Computational Methods
Volume 3: Mathematical Analysis Methods
Volume 4: Diagnostic Methods

The great breadth and significance of this field on the international scene requires multiple volumes for substantive treatment. Indeed, the great significance of this subject internationally is very amply testified to by the fact that the coauthors come from eighteen countries in addition to the USA. Moreover, the authors or coauthors in a number of cases have the degree combinations of M.D. and Ph.D. or the equivalent thereof, which is, of course, only to be expected in a work of such magnitude.

Readers will find a logical flow in the treatment of this broad subject through the respective volumes. The utilization of computers in medical systems starts with a

"road map", i.e., algorithms. The first volume, "Algorithm Techniques", is densely rich with significant algorithmic methods in a wide variety of areas, as a review of the contents of this volume will make apparent. Once algorithms are developed, there remains what is generally a rather formidable task in medical systems, that is, developing a solution for a given complex problem. Volume 2, "Computational Methods," presents numerous significant computational methods in a substantive array of areas. Next, in order to develop algorithms and then solve them, a given medical system must be analyzed to a degree sufficient to produce an adequately descriptive model or set of equations that describes it. Volume 3, "Mathematical Analysis Methods", presents numerous significant techniques for the analysis of medical systems. Finally, the knowledge and insight demonstrated in the first three volumes are applied to the solving of non-trivial diagnostic problems. Volume 4, "Diagnostic Methods", presents a rich variety of diagnostic techniques that reflect a broad spectrum of approaches.

The contributions to these volumes clearly reveal the effectiveness and significance of the techniques presented and, with further development, the essential role that they will play in the future. I hope that students, research workers, practitioners, computer scientists and others on the international scene will find this set of volumes to be a uniquely valuable and significant reference source for years to come.

Cornelius T. Leondes
University of California, Los Angeles
May, 2002

10. TECHNIQUES FOR THE ACCURATE RECOVERY OF TIME-VARYING 3D SHAPES IN MEDICAL IMAGING

YOSHINOBU SATO, MASAMITSU MORIYAMA,
TAKASHI UEGUCHI, MASAYUKI HANAYAMA,
HIROAKI NAITO AND SHINICH TAMURA

1 INTRODUCTION

Acquisition of 3D shape models is one of the most important topics in the computer vision field because of increasing demands for graphic displays in a virtual space as well as quantitative shape analysis in medical diagnosis and industrial inspection. For the purpose of acquiring 3D shape models directly from images, the use of occluding contours has received considerable attention. There have been two main approaches to 3D model acquisition using occluding contours: one approach integrates apparent contours from continuously varying viewpoints [1–3], and the other approach uses deformable models in which model constraints are incorporated such as symmetries and some other regularities [4–6]. The former has the advantage that accurate shape recovery is possible. Although several results have been reported for acquiring 3D models of rigid objects, it seems difficult to recover 3D models from fragmented contours. On the other hand, the latter has the advantage that 3D models can be recovered even using contours from one viewpoint or fragmented contours. Several recovery results including non-rigid moving objects have been demonstrated from one viewpoint [4,5]. More recently, rigid object recovery from multiple viewpoints was reported [6]. However, recovered models were imposed to have rotational or mirror symmetries, and then they were not regarded as accurate but only as plausible.

In this paper, we propose a method for acquiring accurate 3D shape models of non-rigid moving objects directly from images [7]. Especially, we aim at the recovery

of the left ventricular (LV) shapes, one of the most important types of non-rigid objects [8]. We use X-ray cineangiocardiograms, that is, the 2D projections of LV, as an image data source. If we use ultra-fast CT or MRI synchronized with an electrocardiogram, we can directly obtain 3D cross-sectional information to recover 3D models [9,10]. However, X-ray cineangiocardiography still has the advantages on temporal resolution, and spatial resolution along an axial direction of tomography as compared with ultra-fast CT and gated MRI. Furthermore, LV imaging by biplane or single-plane cineangiocardiography is a procedure commonly performed in cardiac catheterization, which is a routine examination regarded as the most reliable and accurate method for cardiac diagnosis by physicians. From the clinical aspect, there is a need for a more accurate LV recovery method by improving conventionally used cineangiography without introducing any special examination such as ultra-fast CT and gated MRI. One of such efforts is to use the density profiles [11] as well as the apparent contours of LV. The problem of this approach is the difficulty of keeping the uniform density of contrast media.

Our recovery method is based on the integration of apparent contours from various viewpoints in order to acquire not only plausible but also accurate 3D models for quantitative shape analysis in cardiac diagnosis. We perform direct fitting to a closed surface model similar to a deformable model in order to deal with fragmented contours such as extracted from X-ray cineangiocardiograms. While LV images are taken from one or two fixed viewing directions in conventional cineangiocardiography, we vary viewing directions continuously when LV images are taken, which can be easily realized using conventional devises. There is previous work on such an image acquisition method motivated from purely clinical concerns [12]. However, the previous recovery method was very primitive, which reconstructed cross section shapes one by one and placed side by side along axial and temporal directions. Therefore, its formulation was quite insufficient as concerns the use of spatiotemporal smoothness of cardiac motion and shape and the evaluation of matches between the extracted contours and the projections of recovered shapes. Also, it was difficult to deal with fragmented contours. In this paper, in order to overcome those problems, we use a time-varying closed surface represented using B-spline functions having three variables (two as surface and one as time) to fit directly and simultaneously to all contour data extracted from time and viewpoint varying images.

The organization of the paper is as follows: In Section 2, we describe the assumption on non-rigid motion, and clarify the advantages of time and viewpoint varying images for non-rigid object recovery. In Section 3, we describe the representation of a time-varying closed surface and clarify the constraints relating the model surface to contour data extracted from images. An iterative recovery method is formulated in order to fit the time-varying model surface to contour data. In Section 4, we present experimental results using synthesized and real image sequences using balloon phantoms. In Section 5, we give discussion on the representation issue of time-varying surface, and the problems toward clinical application.

2 OBTAINING TIME AND VIEWPOINT VARYING IMAGES

In general, non-rigid motion includes translation and rotation as well as the time-variations of object shape. In this paper, however, we assume that object motion originates from only shape variations, but does not include translational and rotational components. When we observe such non-rigid moving 3D objects in order to obtain their time-varying shape information, it is natural that we should vary viewpoints while observing them. We try to give a consideration to this intuition in this section.

A viewpoint can be represented as a point on a spherical surface, which can be parameterized using latitude ϕ and longitude θ. In the case of LV recovery using a biplane X-ray system, we plan to vary LAO and RAO (left- and right-anterior-oblique view) angles from 0° to 90° and from 90° to 0°, respectively. When a patient body is aligned to the polar direction of the spherical coordinate system, the variations of LAO and RAO angles correspond to the variation of longitude θ. Thus, we consider the variation of only longitude θ here. We consider a time–viewpoint space whose axes are time t and viewpoint θ. We assume that an object is observed by two cameras (a biplane X-ray system) whose viewing directions are orthogonal and given by $(\theta_1(t), \theta_2(t))$, where $|\theta_1(t) - \theta_2(t)| = \pi/2$. The sample points for image acquisition are taken along $\theta_1(t)$ and $\theta_2(t)$ in t–θ space. If fixed viewpoints are assumed, sample points in t–θ space are represented as shown in Figure 10.1(a). If time-varying viewpoints are assumed, more uniform sampling can be realized in t–θ space as shown in Figure 10.1(b). If we can assume both shape and motion are smooth, the recovery of more accurate 3D shapes can be expected using the combination of uniform sampling in t–θ space and an appropriate spatiotemporal interpolation method as compared with the combination of images obtained by dense sampling along either t or θ and strong constraints on object shape or motion, except for the case of rigid objects (i.e. no motion) or rotationally symmetric shapes.

In general, cardiac motion can be approximated as periodic motion. When object motion can be assumed to be periodic, time-varying viewpoints are more advantageous. If observable time is long enough compared with one cycle, we can obtain dense uniform image sampling in t–θ space. In LV imaging by cineangiocardiography, observable time is two or three seconds (that is, from three to five cardiac cycles) during one injection of contrast medium. The sampling pattern shown in Figure 10.1(c) is realized using viewpoint variations given by $(\theta_1(t), \theta_2(t)) = (\pi t/8T_0, \pi t/8T_0 + \pi/2)$ (where T_0 is one cycle of periodic motion) and image acquisition by sampling interval $T_0/9$. Different sampling patterns can be realized by changing viewpoint variations $(\theta_1(t), \theta_2(t))$ and the sampling interval (Figure 10.1(d)).

3 RECOVERY OF TIME-VARYING 3D SHAPE MODELS

3.1 Representation of time-varying 3D shape models

We represent time-varying 3D shapes using uniform B-spline functions. In order to parameterize a closed surface, we use spherical coordinates. We specify 3D position using latitude u and longitude v, and distance r from the origin to the direction

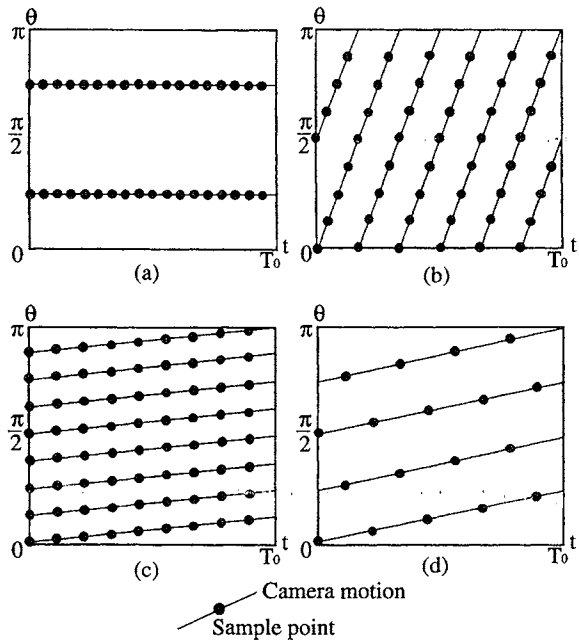

Figure 10.1. Sampling patterns in time–viewpoint space (t–θ space) assuming the use of two cameras whose viewing directions are orthogonal. (a) Sampling pattern for time varying but viewpoint fixed images. Viewpoints are given by $(\theta_1(t), \theta_2(t)) = (\pi/4, 3\pi/4)$. (b) Sampling pattern for time- and viewpoint-varying images of shapes with non-periodic motion. The velocity of viewpoint variation should be fast enough in order to obtain relatively uniform sampling. (c) Sampling pattern for periodic motion with cycle T_0. The velocity of viewpoint variation can be slow enough in order to obtain dense uniform sampling. This pattern is realized by viewpoint variations $(\theta_1(t), \theta_2(t)) = (\pi t/8T_0, \pi t/8T_0 + \pi/2)$, and sampling interval $T_0/9$. (d) Sampling pattern for periodic motion with cycle T_0 realized by viewpoint variations $(\theta_1(t), \theta_2(t)) = (\pi t/4T_0, \pi t/4T_0 + \pi/2)$, and sampling interval $2T_0/9$.

specified by latitude u and longitude v. This means that recovered closed surfaces are limited to star-shaped surfaces with respect to the origin of the spherical coordinate system. Nevertheless, we believe that this class of surface is useful in many domains, especially in LV shape representation. Also, we assume that motion is periodic and its cycle is T_0. Therefore, time-varying 3D shapes are represented by

$$r(u, v, t) = \sum_{i=-3}^{i_0-1} \sum_{j=0}^{j_0-1} \sum_{k=0}^{k_0-1} R_{ijk} U_i(u) V_j(v) T_k(t), \tag{1}$$

where $u \in [0, \pi]$, $v \in [0, 2\pi]$, $t \in [0, T_0]$, R_{ijk} is a coefficient, $U_i(u)$ is the basis function of uniform cubic B-spline for non-periodic functions, and $V_j(v)$ and $T_k(t)$ are the basis functions for periodic functions.

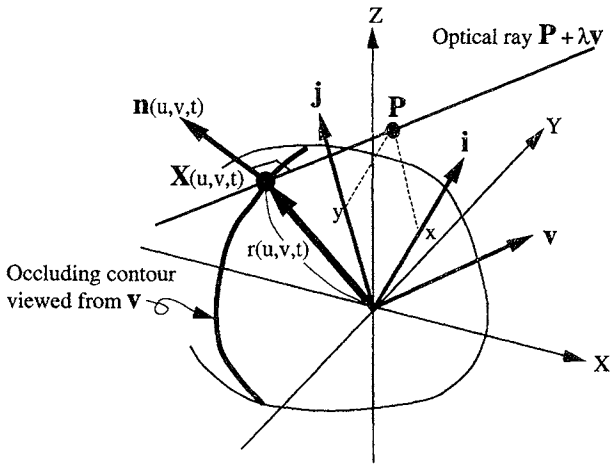

Figure 10.2. Constraints relating 2D coordinates $\mathbf{I} = (x, y)$ of contour points in an image to 3D position $\mathbf{X}(u, v, t)$ and normal $\mathbf{n}(u, v, t)$ on time-varying surface $r(u, v, t)$.

3.2 Constraints for 3D recovery from contours

Given an image with known viewpoint (ϕ, θ) and time t, we want to derive the constraints which relate 2D coordinates of contour points in an image to 3D position and normal on time-varying surface $r(u, v, t)$ (see Figure 10.2). For simplicity, we assume that the spherical coordinate system for representing a viewpoint is coincident with the one for representing a closed surface without loss of generality. Also, we assume orthography as an image projection model. A viewing direction can be given by

$$\mathbf{v} = (\cos\phi\cos\theta, \cos\phi\sin\theta, \sin\phi). \tag{2}$$

We define two orthogonal directions of image axes as

$$\mathbf{i} = (-\sin\theta, \cos\theta, 0), \tag{3}$$

and,

$$\mathbf{j} = (-\sin\phi\cos\theta, -\sin\phi\sin\theta, \cos\phi). \tag{4}$$

The optical ray corresponding to image coordinates $\mathbf{I} = (x, y)$ is given by

$$\mathbf{P} + \lambda\mathbf{v}, \tag{5}$$

where $\mathbf{P} = x\mathbf{i} + y\mathbf{j}$, and λ is a scalar value. Here, we suppose that the optical ray passing through the origin of the spherical coordinate system is defined as $\lambda\mathbf{v}$.

Now, we derive the constraints on a surface represented by $r(u, v, t)$, given an image contour point $\mathbf{I} = (x, y)$ at viewpoint $\mathbf{v}$ and time t. 3D position $\mathbf{X}(u, v, t)$ on

its time-varying surface is obtained by the transformation from spherical coordinates to Cartesian coordinates. Because $r(u, v, t)$ is the distance from the origin along 3D direction $(\cos v \cos u, \cos v \sin u, \sin v)$ at time t, 3D position $\mathbf{X}(u, v, t)$ is given by

$$\mathbf{X}(u, v, t) = r(u, v, t)(\cos v \cos u, \cos v \sin u, \sin v). \tag{6}$$

Surface normal $\mathbf{n}(u, v, t)$ at $\mathbf{X}(u, v, t)$ is given by

$$\mathbf{n}(u, v, t) = N\left[\frac{\partial \mathbf{X}(u, v, t)}{\partial u} \times \frac{\partial \mathbf{X}(u, v, t)}{\partial v}\right], \tag{7}$$

where $N[\mathbf{x}] = \mathbf{x}/|\mathbf{x}|$. If there is an image contour point $\mathbf{I} = (x, y)$ with viewpoint $\mathbf{v}$ and time t which is a projection of an occluding contour of a surface, the constraints given by

$$\mathbf{X}(u, v, t) = \mathbf{P} + \lambda \mathbf{v}, \tag{8}$$

and

$$\mathbf{n}(u, v, t) \cdot \mathbf{v} = 0 \tag{9}$$

must be satisfied at the corresponding surface coordinates (u, v), where $\lambda = \mathbf{X}(u, v, t) \cdot \mathbf{v}$, and $\mathbf{P} = x\mathbf{i} + y\mathbf{j}$.

3.3 Iterative method for time-varying 3D recovery

Based on the constraints given by Eqs (8) and (9), we formulate a method for estimating $r(u, v, t)$. Equations (8) and (9) are the basic constraints for 3D recovery from occluding contours, which are also described in [3]. In our problems, however, it is difficult to directly obtain $r(u, v, t)$ satisfying these constraints because it is unknown what coordinates (u_ℓ, v_ℓ) correspond to the optical ray $\mathbf{P}_\ell + \lambda \mathbf{v}_\ell$ determined by given image coordinates (x_ℓ, y_ℓ).

In order to obtain an approximate solution, we decompose the problem into two stages: First, we use Eq. (9) to find the correspondence between surface coordinates (u_ℓ, v_ℓ) and each optical ray determined by image coordinates (x_ℓ, y_ℓ) of a given contour point. Second, we estimate $r(u, v, t)$ by solving a linear equation system obtained from Eq. (8). We iterate these two stages to finally obtain a solution satisfying both Eqs (8) and (9).

The recovery algorithm is described as follows. We start the algorithm by setting initial shape $r^{(0)}(u, v, t)$, and computing $\mathbf{X}^{(0)}(u, v, t)$ and $\mathbf{n}^{(0)}(u, v, t)$. (In the experiments, we used a sphere as initial shape $r^{(0)}(u, v, t)$.) Let m be an iteration count. We set $m = 0$ initially.

During the first stage, we find tentative correspondence between surface coordinates (u_ℓ, v_ℓ, t_ℓ) and an optical ray determined by image coordinates (x_ℓ, y_ℓ) at time t_ℓ (see Figure 10.3). We find this correspondence for every image contour point at every viewpoint and time. Given $\mathbf{X}^{(m)}(u, v, t)$, $\mathbf{n}^{(m)}(u, v, t)$, and the optical ray $\mathbf{P}_\ell + \lambda \mathbf{v}_\ell$

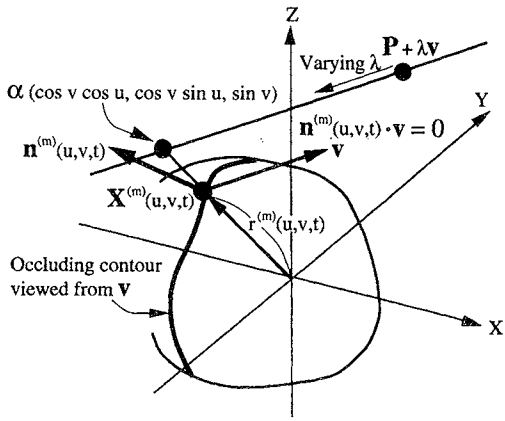

Figure 10.3. Tentative correspondence between surface coordinates and the optical ray determined by image coordinates.

which corresponds to image contour point (x_ℓ, y_ℓ), we find u_ℓ, v_ℓ, and λ_ℓ satisfying the constraints

$$\mathbf{n}^{(m)}(u_\ell, v_\ell, t_\ell) \cdot \mathbf{v}_\ell = 0, \tag{10}$$

and

$$\mathbf{P}_\ell + \lambda_\ell \mathbf{v}_\ell = \alpha_\ell(\cos v_\ell \cos u_\ell, \cos v_\ell \sin u_\ell, \sin v_\ell), \tag{11}$$

where α_ℓ is a scalar coefficient. If we suppose that λ_ℓ is given, the 3D position of $\mathbf{P}_\ell + \lambda_\ell \mathbf{v}_\ell$ is determined. By representing this 3D position using spherical coordinates, u_ℓ, v_ℓ, and α_ℓ are uniquely determined using Eq. (11). We can check whether Eq. (10) is satisfied for determined u_ℓ and v_ℓ. In order to find u_ℓ, v_ℓ, and α_ℓ satisfying the constraints, we continuously vary λ_ℓ and search the value of λ_ℓ which satisfies Eq. (10). If there are multiple values of λ_ℓ satisfying Eq. (10), we select λ_ℓ where $|\alpha_\ell - r^{(m)}(u_\ell, v_\ell, t_\ell)|$ is the minimum.

During the second stage, based on u_ℓ, v_ℓ, and α_ℓ found at the first stage, we estimate $r^{(m+1)}(u, v, t)$ by solving a set of linear equations derived from

$$r^{(m+1)}(u_\ell, v_\ell, t_\ell) = \alpha_\ell. \tag{12}$$

More precisely, combining with the smoothness constraint, we find $R_{ijk}^{(m+1)}$ minimizing

$$E^{(m+1)} = \frac{1}{\ell_0} \sum_{\ell=1}^{\ell_0} \left\{ r^{(m+1)}(u_\ell, v_\ell, t_\ell) - \alpha_\ell \right\}^2 + w_s \cdot \frac{1}{2T_0\pi^2} \int\int\int \left(\pi \cdot \frac{\partial r^{(m+1)}(u, v, t)}{\partial u} \right)^2$$
$$+ \left(2\pi \cdot \frac{1}{\cos u} \cdot \frac{\partial r^{(m+1)}(u, v, t)}{\partial v} \right)^2 + \left(T_0 \cdot \frac{\partial r^{(m+1)}(u, v, t)}{\partial t} \right)^2 \mathrm{d}u\, \mathrm{d}v\, \mathrm{d}t \tag{13}$$

where

$$r^{(m+1)}(u, v, t) = \sum_{i=-3}^{i_0-1} \sum_{j=0}^{j_0-1} \sum_{k=0}^{k_0-1} R_{ijk}^{(m+1)} U_i(u) V_j(v) T_k(t), \tag{14}$$

and w_s is a weight parameter for the smoothness constraint. $1/\ell_0$ and $1/(2T_0\pi^2)$ are factors for obtaining average values from the summation and the integral, which normalizes the smoothness constraint and the data constraint based on Eq. (12). π, 2π, and T_0 (by which the partial derivatives are multiplied) are factors for the normalization of each partial derivative. The partial derivative with respect to v is multiplied by $1/\cos u$ because of the reduction of length along v with approaching the poles. (In the experiments, the normalized partial derivatives were estimated using discrete approximations such as $(R_{ijk} - R_{i+1,jk})/(1/i_0)$). If error $E^{(m+1)}$ for newly estimated $r^{(m+1)}(u, v, t)$ is almost the same as previous error $E^{(m)}$, then we stop the algorithm, else we set $m = m + 1$ and go back to the first stage. (Empirically, four iterations were sufficient for an appropriate weight value of the smoothness constraint.)

4 EXPERIMENTAL RESULTS

We evaluated the method using synthesized and real image sequences. The method was implemented on a *SPARC Station 20*.

The synthesized image sequences were generated based on the stationary 3D shape recovered from viewpoint-varying X-ray images of a stationary balloon filled with contrast media. The recovered 3D shape was deformed using three different time-varying periodic scale functions along three orthogonal directions so that the time variation of its volume was similar to the one of LV. The periodic scale functions were described by sinusoidal functions. In a normal LV, its shape is roughly rotational symmetric, and its contraction is relatively uniform everywhere on heart wall. In a diseased LV, however, it is often that its shape is not rotational symmetric and/or the contraction is not uniform, which causes non-symmetric shapes. We generated time-varying shape so as not to be rotationally symmetric. The viewpoint-varying image sequences were generated under the assumption of orthographic projection.

The real image sequences were obtained by taking X-ray images of a balloon filled with contrast media using a biplane X-ray system (Siemens BICOR). Although the shape of a balloon is commonly rotationally symmetric, we deformed the balloon shape by covering carton frames. We controlled the volume of the balloon using a pump so that the time variation of the balloon volume was periodic and similar to the one of LV. Also, we took CT images of the balloon at several time phases and used the 3D models reconstructed from the CT images as the golden standard. In this case, perspective projection is more appropriate as an image projection model than orthography. However, we applied the recovery method assuming orthography.

Before the recovery of time-varying surfaces, the detection of image edges is necessary. We took the zero-crossings of the $\nabla^2 G$ whose gradient magnitude values are large as image edges. Our method currently cannot discriminate "spurious" edges which should be regarded as outliers. So, we manually removed "spurious" edges

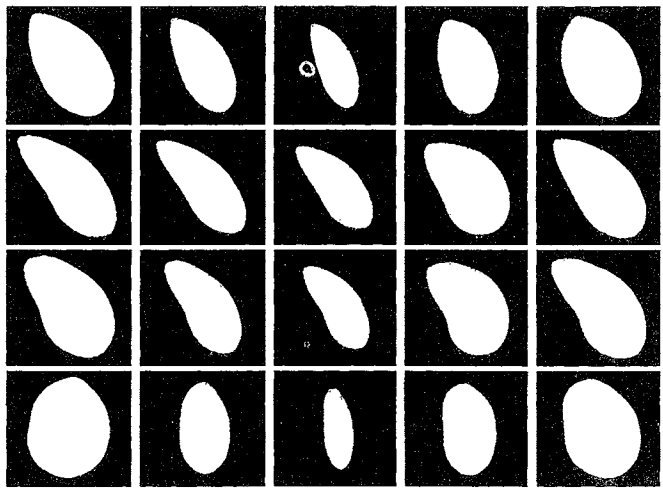

Figure 10.4. Time- and viewpoint-varying images of the synthesized image sequence.

which did not originate from the occluding contours of the balloon. We randomly selected 25% of all the extracted edges in each image and used them for the 3D recovery.

In the following experiments, we used a uniform function, that is a sphere, as initial shape $r^{(0)}(u, v, t)$. The spherical coordinate system for surface representation was selected using the following method. The axis of elongation in the projected shape was manually specified in two images taken from orthogonal viewpoints at the systolic phase. A 3D line segment was determined as z-axis of the coordinate system so that its projections were coincident with the specified two axes in these images. The origin was set at the center of the 3D line segment. The directions of x-axis and y-axis were set to the two orthogonal directions from which the two images had been taken.

4.1 Synthesized image sequence

The image sequences were synthesized assuming the use of a biplane X-ray system by which projections from two orthogonal views can be obtained simultaneously. Three image sequences were generated using sampling patterns in the time–viewpoint space as shown in Figure 10.1. One of these sampling patterns was the time and viewpoint varying sequence generated using the sampling pattern shown in Figure 10.1(c). The other two sequences were generated using fixed viewpoints $(0, \pi/2)$ and $(\pi/4, 3\pi/4)$ with time interval $T_0/36$.

Figure 10.4 shows a part of the viewpoint-varying images. The size of each image was 220 × 220 (pixels). Figure 10.5 shows the shaded displays of time-varying 3D shapes recovered from the viewpoint-varying sequence. We used $i_0 = 12$, $j_0 = 12$, and $k_0 = 12$ as the number of knots of B-spline functions in Eq. (14), that is, the

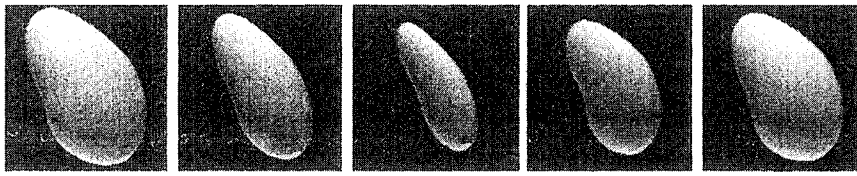

Figure 10.5. Shaded displays of time-varying 3D shapes of one cycle recovered from time and viewpoint varying synthesized image sequence.

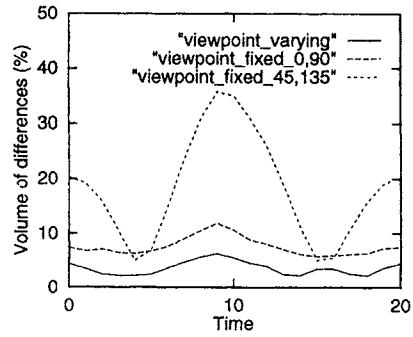

Figure 10.6. Time variations of the normalized volume of differences of the recovered shapes from viewpoint-varying sequence and viewpoint-fixed sequences.

grid was $12 \times 12 \times 12$. We used $w_s = 0.5 \times 10^{-3}$ as the weight parameter of the smoothness constraint in Eq. (13). Four iterations were needed for the recovery algorithm to almost converge. Figure 10.6 shows the time variations of error in the recovered shapes. We used the volume of the differences between the true shape and the recovered shape as a measure of error. We further divided the volume of differences by the true volume at each time phase to normalize the error. It should be noticed that the volume of differences is not the difference between the true volume and the estimated volume. That is, we did not use the difference of volumes, but used the difference of shapes. In Figure 10.6, the normalized volume of differences in the results recovered from the viewpoint-fixed images highly depended on the selection of two viewpoints. We synthesized the time-varying shapes by deforming a roughly rotational symmetric shape using time-varying scale functions along three orthogonal directions. Because two orthogonal viewpoints happened to be close to two of these three orthogonal directions when $(\theta_1(t), \theta_2(t)) = (0, \pi/2)$, the recovered shapes were relatively accurate (Figure 10.7). Nevertheless, the accuracy of the result recovered from the viewpoint-varying images was considerably higher through one cycle. Figure 10.8 shows the projection images of the recovered 3D shapes superimposed on the projection images of the true shapes. The projection images of the shape recovered from the viewpoint-varying images and the true shape were almost the same at not only

2. COMPUTER TECHNIQUES FOR SPATIAL PATTERN ANALYSIS OF OBJECTS IN BIOMEDICAL IMAGES

G. CEVENINI, M.R. MASSAI AND P. BARBINI

INTRODUCTION

The many definitions of spatial pattern (SP) have subtle differences according to the context of the application, but all refer to the disposition of a set of objects in a spatial region of interest [1–5]. Here we are concerned exclusively with object locations and their spatial interplay due to stochastic processes which sometimes are unknown or inexplicable, sometimes are the subject of study. Any other qualitative or quantitative property of objects, such as for example color or size, is disregarded. Hence we are not concerned with the problem of spatial autocorrelation, since this property occurs whenever there is a systematic change in values across a map and can therefore be investigated only if the patterns are analyzed by values recorded at locations. For a detailed treatise on this topic, see [4].

As a direct consequence of the above definition, the objects have to be conceived without dimension, i.e., as numerable points. Nevertheless we always call them objects to distinguish them from arbitrary points in the region of study.

A SP is completely identified by the angles between the lines connecting the objects (Figure 2.1(a)) and does not change with translations, rotations (Figure 2.1(b)) or enlargements equal in all directions from a reference point (Figure 2.1(c)). Any other displacement of objects with respect to the others causes pattern distortion (Figure 2.1(d)).

Approximation of the objects as points depends on the scale on which they are observed and analyzed. In most cases, the real physical dimension of objects becomes

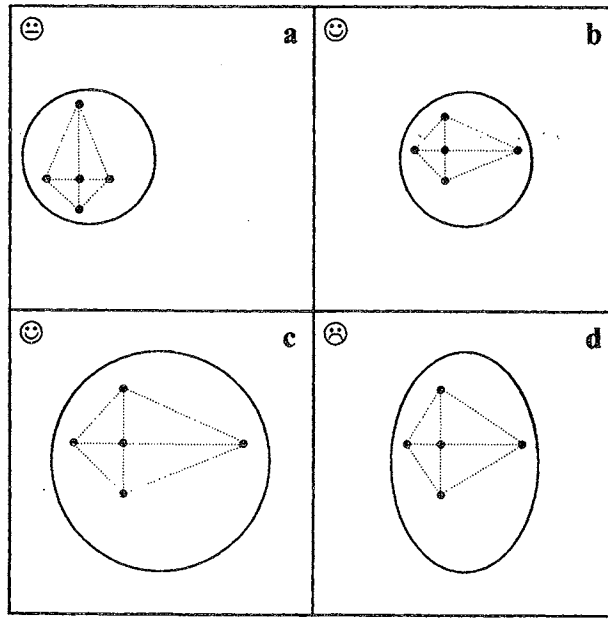

Figure 2.1. Three transformations of a spatial pattern (a) leaving it unaltered (b)–(c) and changing it (d): (a) original spatial pattern; (b) translation of (a) to the right and subsequent rotation through 90°; (c) enlargement of (b) in all directions; (d) contraction of (c) only in a horizontal direction.

negligible with a suitable choice of scale. For example, even stars on a cosmic scale can be taken as points without any error worthy of mention.

Rigorously, SPs can be studied in one, two or three dimensions. However, since we apply SP analysis to biomedical images, we refer to two-dimensional space unless otherwise specified. Most of the ideas can be extended to other spaces by introducing only slight mathematical changes.

The analysis of SPs must begin with a test of complete spatial randomness (CSR) since rejection of CSR is a prerequisite for even thinking about modeling an observed SP [2,3]. Moreover it operates as a dividing hypothesis between aggregated and regular patterns which are the two main alternatives to random patterns.

This classification of SPs into three categories is of course a simplification but it is a useful start and allows a glimpse of the main causal mechanisms creating the patterns. In fact the arrangement of objects in space may range from aggregated or patchy (Figure 2.2), caused for example by contagious clustering, clumping and local non homogenous rarefaction processes, through a random pattern (Figure 2.2(a)) to being, more rarely and regular (Figure 2.2(c)). For example, regular SPs can arise from inhibition or repulsion mechanisms which constrain objects to remain a certain distance from each other. Of course, in nature, and hence also in the biomedical field,

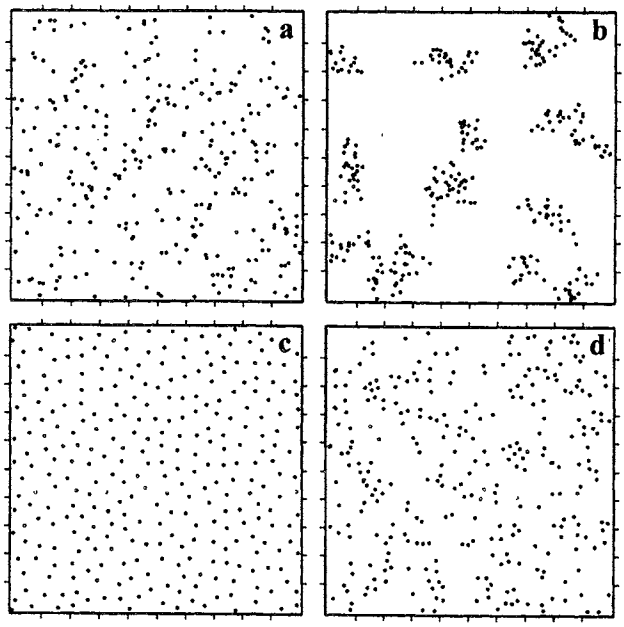

Figure 2.2. Four computer-simulated spatial patterns in a square of 10-unit sides: (a) random pattern; (b) aggregated pattern; (c) regular pattern; (d) constrained pattern.

there are objects, the spatial organization of which has more complex patterns, for example in which very small and large distances between neighboring objects are allowed, but not intermediate distances (Figure 2.2(d)). Such constrained patterns are found in nature due to growth processes.

Doing a CSR test is also important because most of these tests can be used as exploratory tools to provide complementary information about pattern characteristics, even by measuring the strength of evidence of departure from randomness through significance levels.

Before we can discuss better some specific aims of the different techniques used in SP analysis, some practical aspects about data collection have to be considered. Sometimes the region of interest is spatially well-defined and the locations of all the objects to be analyzed are completely known, as for example the cherries in a cherry tree. In most of biomedical applications, however, the region containing all the objects is too large to be measured and/or has poorly defined borders, as for example, neurons in the human brain or tumor cells in microscope images of histological sections. A sampling region is therefore chosen and studied using probability sampling schemes or representative criteria. The results are extended to the whole population of objects by statistical inference, which may also be useful to prove the consistency of the results over many replicate choices of sampling region.

In the latter situations, in which the positions of the objects in the study region can be determined exactly, we shall refer to *mapped* patterns. Besides providing a preliminary test of CSR, the mapping methods can aim at formulating a parametric stochastic model, fitting it to the data, and explaining the underlying causal processes by deterministic inferential procedures. Clearly even simple stochastic models admit many plausible interpretations.

Many different techniques for analyzing mapped SPs have been developed. They are distinguished on the basis of the type of stochastic variables used [1–5]. They can be classified in three main categories, although some combined methods do also exist.

(i) Counts of objects over squared or rectangular disjointed sub-regions of suitable size, obtained from superimposing grids on the study region [6].
(ii) Measurement of geometric quantities such as areas, perimeters or lengths, defined either by dividing the study region into some kind of polygonal cells (tessellation) with specific spatial properties [7–9], or by trees obtained by connecting all the objects in the shortest way [10,11].
(iii) Computation of distances between objects and between objects and selected points [12–14].

As an alternative to intensive mapping in a single region, information can be obtained from a large number of smaller regions, usually called *quadrats* even if they are not square. Some of these sub-regions may also partially overlap. Due to its simplicity, quadrat sampling is a very popular technique in the biological and biomedical fields, but its results are not always completely reliable [15]. Moreover, paradoxically, it does not account for the spatial nature of the patterns. Distance methods are an alternative, and give more sensitive insights into the spatial arrangement of objects and their interplay. The sampling unit is a point and the information is stored as distances to neighboring objects.

The objectives of these so-called *sparse sampling* methods based on quadrat counts or distance measurements are generally rather different. Besides testing CSR they are usually only concerned with evaluating intensity, i.e., the mean number of objects per unit area. Equivalently, if the total area of the region of interest is known, they can estimate the total number of objects simply as the product of intensity and total area. The analysis of sparsely sampled SPs can, at the most, provide a qualitative description of the underlying pattern. In fact, to obtain more and better inferential information, it is better to deal with mapped data in sub-regions and use the above-mentioned mapping methods.

Most procedures of SP analysis require intensive use of a computer. The use of sophisticated techniques requires considerable processing capacity. The availability of increasingly powerful computers has enabled techniques to be refined and new ones to be developed [16,17].

The SP is an interesting characteristic of many natural phenomena. In nearly all fields of human endeavor, point-like objects with spatial organization worthy of

study can be identified. It is almost impossible to list all the literature on SP analysis. Here we shall mention only the main fields of application, citing a few recent pertinent books and articles. Other references can be obtained from the bibliography of the principal texts on SP analysis [1–5]. Non biomedical disciplines which have made long and assiduous use of SPs are ecology [1,12,18–26], biology (especially botany [3,4,6,24,27–31], forestry [3,32,33], zoology [1,18,20,34]), geology [35], geography [4,36–41], astronomy [42–44] and archaeology [45,46]. Recently the social sciences [47,48], genetics [49] and even traffic and accidents [50] have been analyzed by SP techniques. With regard to the many branches of the biomedical fields, the interpretation of object organization in biomedical images is of great scientific interest. Spatial pattern analysis therefore has and will continue to have applications in this field. Some of the main areas of application include anatomy [51–55], histopathology [56–63] and health care [64–66].

In this chapter we consider the main SPs found in biomedical images, especially the random pattern. We describe statistical tests of CSR, particularly the Monte Carlo tests which require large processing capacity. After a brief description of the main count and geometrical methods, distance methods applied to mapped data are discussed in detail, since they generally give more refined information in biomedical applications. Sophisticated techniques for the analysis of nearest neighbor (NN) distances are discussed in relation to interpretation of SPs in biomedical images where it is not always possible to ignore object size, and where attention can be restricted to sub-areas of plausible reciprocal influence between objects. The main difficulty of distance methods, namely the influence of the boundary of the study region, i.e., the so called *edge effect* is then discussed. After this we describe a very subtle and revealing SP method combining count and distance analysis. In conclusion, we outline the meaning, utility and problems arising from extension of these techniques to multivariate patterns, i.e. patterns of different types of objects, with reference to spatial independence.

In describing the various methods of SP analysis, except those for multi-type object SPs, we frequently use the well-controlled computer-simulated data shown in Figure 2.2. These patterns were obtained by distributing 300 point-like objects in a plane in four different ways. They are useful for illustrating the properties of the different techniques. Except for the random pattern of Figure 2.2(a), obtained by computer simulation of uniformly distributed random numbers, we purposely do not specify how the other three patterns were simulated. This is to make the investigation and discussion of the main techniques of spatial data analysis more stimulating, as if we were seeking the innermost spatial characteristics of experimental data. The computer algorithms used to distribute the objects are reported in detail later, before summing up the performance of the various techniques.

We also explain some computer algorithm techniques of the most informative SP methods, based on distances. The efficacy of SP methods is also evaluated in relation to their capacity to provide interesting quantitative or graphic indications of causal processes underlying pattern formation.

THE RANDOM PATTERN

The SP is called random if the objects are assigned independently and at random to the available points in the study region. In other words, the objects in the study region have a random pattern or are randomly dispersed if every point is as likely as every other to be the site of an object. The Cartesian spatial coordinates of object locations are therefore independent and uniformly distributed. The random pattern is the first step in the description of SPs because an essential prerequisite for detailed study of SPs is that the objects are not randomly distributed in the study area. The absence of CSR is also useful because the CSR tests are two-tailed and unequivocally suggests one of the two possible alternatives, namely aggregated or regular pattern. If the tests cannot exclude CSR, then the only thing we can say about the SP is that it is random. On the contrary, even when the SP is clearly anything but random, tests of CSR are a useful aid in formulating hypotheses about the pattern and its origin. So, in order to define suitable tests of CSR, it is first necessary to design a stochastic model of spatial randomness.

THE POISSON DISTRIBUTION

For a random pattern, the number of objects in the study region follows a Poisson distribution, i.e., the probability that the unit area will contain n objects is

$$p_n = \frac{\lambda^n e^{-\lambda}}{n!}, \quad n = 0, 1, \ldots, \tag{1}$$

where λ is the mean number of objects per unit area, also known as intensity. The Poisson distribution, also called the distributions of rare events, is particularly used in counting methods and can be derived from the binomial distribution.

To explain this we refer to Figure 2.3. For a convenient first approximation, it is supposed that the objects cannot occupy all the infinite point locations of the whole region of interest R, but only a finite, albeit large, number of them, in each of which not more than a single object can be present. Here the finite locations are represented by tiny squares like the pixels on a digital screen and the objects are little black squares of the same size as the locations. A number, N, of objects are randomly scattered in R, so that λ is simply the ratio of N to the area of R.

Now consider any sub-region, I, of unit area, in R. The probability, p, that a location will be occupied by an object is $p = \lambda/l$, where l is the finite number of point locations in I. The probability that exactly n locations will be occupied is given by the binomial probability

$$p_n = \binom{l}{n} p^n (1-p)^{l-n}, \quad n = 0, 1, \ldots, l. \tag{2}$$

If we reduce the size of the pixels, or increase the number of finite locations l, p becomes very small, n negligible with respect to l, but lp remains equal to the constant

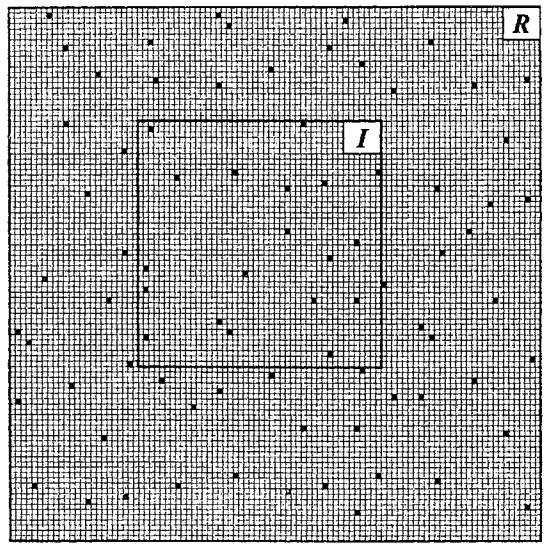

Figure 2.3. Whole region of interest R with a finite number of discrete locations occupied by at most one object, like for computer graphic pixels. Sub-region I is conventionally assumed to be of unit area.

finite value λ. Equation (2) can be rewritten and approximated:

$$p_n = \frac{l!}{n!(l-n)!} p^n (1-p)^{l-n} = \frac{l(l-1)(l-2)\cdots(l-n+1)(l-n)!}{n!(l-n)!} p^n (1-p)^{l-n}$$

$$\cong \frac{(lp)^n}{n!}(1-\lambda/l)^l = \frac{(lp)^n}{n!} e^{l\ln(1-\lambda/l)}. \tag{3}$$

Setting $m = 1/l$, letting l tend to infinity, or m to zero, and expanding the logarithm around $m = 0$ with a Taylor series, we obtain the Poisson distribution (1):

$$p_n = \frac{\lambda^n}{n!} \lim_{m\to 0} e^{\frac{1}{m}\ln(1-m\lambda)} = \frac{\lambda^n}{n!} \lim_{m\to 0} e^{-\frac{1}{m}\sum_{i=1}^{\infty} m^i\lambda^i/i!} = \frac{\lambda^n e^{-\lambda}}{n!}. \tag{4}$$

So when there are few objects in the study region (the number can theoretically be infinite for point-like objects), and all locations have the same probability of being occupied, the number of objects per unit area will follow a Poisson distribution. If we have to study a region not of unit area but of area A, we only have to substitute λA for λ in (4) to obtain the Poisson distribution of the number of objects in A. An example of a random distribution of 300 objects in a square area is also given in Figure 2.2a. Although the eye perceives a few clusters, they are an illusion.

An important property of the Poisson distribution is that it has only one parameter λ which is equal to the mean and the variance. As we shall see, this enables very simple indices for detecting non-random patterns, through count techniques, to be defined.

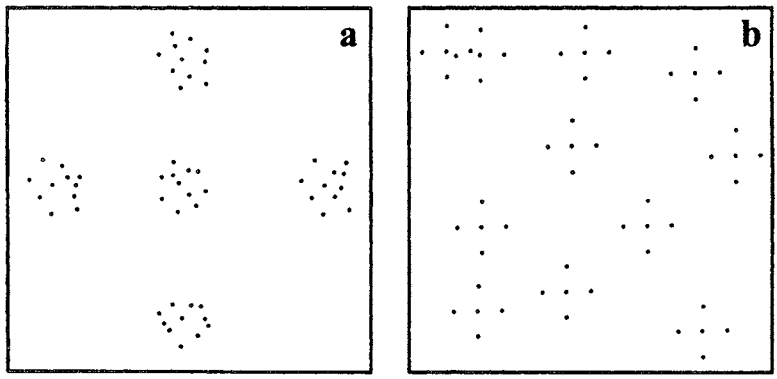

Figure 2.4. Two stylized complex patterns: (a) regular arrangement of clusters with randomly located objects; (b) randomly located clusters of regularly arranged objects.

ALTERNATIVE PATTERNS

A preliminary qualitative and graphical distinction between regular and aggregated alternatives to CSR (see also Figure 2.2) was mentioned above. In an attempt to obtain a quantitative description of the processes leading to many aggregated and regular patterns, stochastic models have been proposed in many different applied sciences. Among those for the biomedical field, an interesting recent one, for example, is based on the principle of maximum entropy and designed for the description of spatial interaction of biological entities like cells in a tissue [67]. Comprehensive foundations for spatial processes can be found in [3,4,39,68,69].

Our approach to the analysis of alternative SPs is to identify some intrinsic characteristics from supplementary information provided by appropriate tests of CSR. Once CSR has been excluded, we will try to distinguish patterns and interpret them directly from the distributions of their associated statistics, particularly NN distances.

Aggregated patterns have a higher frequency of small and large intensities, i.e., over-dispersion, with respect to the random pattern; regular patterns have a more uniform distribution of intensities about the mean, i.e., under-dispersion. Aggregated patterns have many more small NN distances between objects (inter-object NN distances) with respect to the random patterns, whereas regular patterns have many more large inter-object NN distances. The opposite is true for the distribution of NN distances between points arbitrarily selected in the study region and objects (point-object NN distances). More details about point-object NN distances are given later.

This simplistic classification as aggregated, random and regular patterns, is even useful for describing and interpreting more complex SPs. For example Figures 2.4(a) and (b) show a regular arrangement of clusters containing randomly located objects, and randomly located clusters of regularly arranged objects, respectively. In characterizing SPs, the choice of the size of the region to be explored plays a crucial role. For example, if we analyze the whole area shown in Figure 2.4(a) we find an aggregated

pattern of objects, but if we analyze the sub-areas containing a single cluster, we find random patterns of objects. If we examine the pattern of the clusters, replacing them with macro objects identified as points through their centroids, we find a regular pattern. Similar argumentation can be advanced for Figure 2.4b. The two examples of SPs given in Figure 2.4 are of course simulations of stylized patterns, unlikely to be encored in reality. However, they are useful for understanding the multiple scale aspects involved in the interpretation of complex patterns which are not random.

In the interpretation of SPs of biomedical images, it is crucial that the problem be stated exactly at the beginning. This includes the choice of object type, the definition of area size or scale, and the choice of one technique of SP analysis rather than another. The results obtained with only slightly different initial choices can be radically different.

MONTE CARLO TESTS

The Monte Carlo approach for testing models against data is very useful in SP analysis where even simple stochastic models give rise to distributions which do not lend themselves to mathematical analysis [3,70]. Because the Monte Carlo tests are based on onerous computer simulations, they have become more useful since powerful modern microprocessors reduced computation times.

The Monte Carlo approach to SP analysis can be briefly described as follows. Given a stochastic variable S, the values s of which are distributed according to a function $F(s)$, let s_0 be an observed value of S. $F(s)$ is used to generate t independent random values s_i, $i = 1, \ldots, t$ by computer simulation. The probability that s_0 be the largest, $s_{\max}$, or the smallest, $s_{\min}$, of the $t + 1$ equally probable values s_i, $i = 0, 1, \ldots, t$, is

$$P(s_0 = s_{\max}) = P(s_0 = s_{\min}) = (t + 1)^{-1}. \tag{5}$$

This enables the probability of error level p of rejecting the null hypothesis that s_0 belongs to $F(s)$ to be determined through the rank of s_0. In fact, for example, in a one-sided test about the lower side

$$p = \frac{\text{rank of } s_0}{t+1}. \tag{6}$$

The extension to two-sided tests is immediate.

In theory, the number t of computer simulated random values for S must be chosen as large as possible to obtain the finest resolution of p. In practice, $t = 99$ (so that $t + 1 = 100$) should be sufficient to ensure a resolution of 1% and to allow the conventional 1% and 5% probabilities of error to be evaluated for rejecting the null hypothesis. Unfortunately, because of power loss associated with Monte Carlo tests based on ranks [71] a value of $t = 99$ is recommended for a 5% and $t = 499$ for a 1% probability.

The main advantage of the Monte Carlo approach is that practically any informative stochastic variable can be chosen without the need to know its distribution law. However, if this freedom of choice is not used wisely, there is the risk of choosing variables S that emphasize only deviant characteristics of the experimental data.

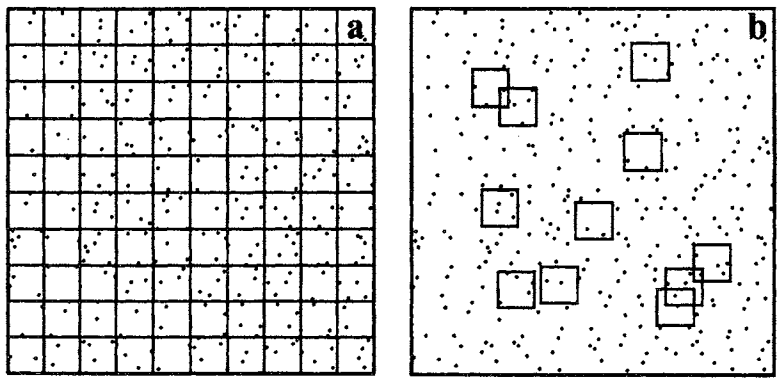

Figure 2.5. Two quadrat count methods: (a) squared grid of contiguous quadrats; (b) randomly disposed quadrats.

Furthermore, when the asymptotic distribution laws are known, Monte Carlo testing is a useful tool of inspection. Disagreements between classical and Monte Carlo approaches are usually due to improper assumptions about distribution.

A disadvantage of Monte Carlo methods is their intrinsic weakness in analyzing combined aspects and they consequently tend to be conservative, rarely rejecting null hypotheses when they should. To obviate this, it is necessary to carry out a very large number of simulations that take many composite hypotheses into account, but this is seldom practicable. It is therefore a good idea to purge the data first of combined and undesired effects, if possible.

A problem which however can arise with Monte Carlo tests is that computers cannot provide a completely random sample of S, since they usually have only a pseudo-random number generator, albeit of high quality.

QUADRAT COUNTS

One of the more widely used methods of SP analysis evaluates the number of objects in selected sub-areas of the study region A [6,15,28–30,72,73]. All the counting methods rely on the use of quadrat, i.e., sub-areas, usually but not necessarily squares, as in Figure 2.5. Rectangular and circular quadrats are also not uncommon. The use of quadrat counts as indicators of pattern is very popular because of its simplicity, but this method has some serious problems and paradoxically ignores the spatial nature of the data.

For mapped patterns one proceeds to divide A into m disjoint sub-areas $A_i, i = 1, \ldots, m$, of the same or different areas, but so as to obtain a grid of contiguous quadrats completely covering A. Again, for the sake of simplicity, and because it is the usual choice, we assume that A has unit area and is partitioned into a regular grid of m smaller squares (Figure 2.5(a)). Under the hypothesis of CSR, the numbers, n_i, $i = 1, \ldots, m$, of objects in the quadrats can be regarded as samples extracted