IC/90/228
INTERNAL REPORT
(Limited Distribution)

International Atomic Energy Agency
and
United Nations Educational Scientific and Cultural Organization
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS

ON THE ESTIMATION OF THE SPHERICAL CONTACT DISTRIBUTION $H_s(y)$ FOR SPATIAL POINT PROCESSES *

Sani I. Doguwa **
International Centre for Theoretical Physics, Trieste, Italy.

ABSTRACT

RIPLEY (1977, **Journal of the Royal Statistical Society, B39 172–212**) proposed an estimator for the spherical contact distribution $H_s(s)$, of a spatial point process observed in a bounded planar region. However, this estimator is not defined for some distances of interest, in this bounded region. A new estimator for $H_s(y)$, is proposed for use with regular grid of sampling locations. This new estimator is defined for all distances of interest. It also appears to have a smaller bias and a smaller mean squared error than the previously suggested alternative.

MIRAMARE – TRIESTE
August 1990

* Submitted for publication.
** Permanent address: Department of Mathematics, Ahmadu Bello University, Zaria, Nigeria.

1 Introduction

The second order methods (RIPLEY,1977), represent a natural and valuable starting point for the description of a spatial point process, but do not give a complete picture, since these methods cannot distinguish processes which have identical second order properties. RIPLEY(1977) has suggested looking at the spherical contact distribution - the distribution function of the distance from an arbitrary location to the nearest point of the process. This function is defined by

$$H_s(y) = \Pr(Y \leq y) \tag{1}$$

where the random variable Y denotes the distance between an arbitrary location and the nearest point of the process. An equivalent definition of $H_s(y)$ is given by

$$H_s(y) = 1 - \Pr(N[b(o,y)] = 0) \tag{2}$$

where $N[b(o,y)]$ is the number of points in the disc $b(o,y)$, with center the origin and radius y (see STOYAN et al, 1986, pp 43).

The estimation of $H_s(y)$ is complicated by the bounded nature of the pattern being studied. For example, the disc of radius y surrounding a location is certain to overlap one or more of the boundaries of the sampling window as y increases. Also the position of the nearest neighbour to a location will only be known with certainty for those locations lying within the interior of the study region. Thus edge effects play a significant role in the estimation of $H_s(y)$, and as would be seen in the sequel, estimators without edge correction have great downward biases and cannot be recommended for use in a bounded region.

In this paper another edge-corrected estimator of $H_s(y)$ is proposed and compared with the existing estimators using simulation technique. We illustrate the results for the new estimator in the analysis of a set of data relating to the positions of Biological cells.

2 The sampling locations

In this paper, we assume a rectangular window W, and our edge-corrections relate only to a regular lattice of sampling locations as favoured by DIGGLE(1979). Our results are based on a lattice obtained by partitioning a rectangular window into $(k+1)^2$ similar subwindows and then using the $m = k^2$ inner corners of these subwindows. That is, for a given window W of dimensions a by b, the position of the $(i,j)^{th}$ sampling location P_h say, where i,j = 1,2,3,...,k; and $h = i + k(j-1)$ is

$$P_h = (\frac{ia}{k+1}, \frac{jb}{k+1}). \tag{3}$$

When collecting information for the refined distance analysis using the estimators of $H_s(y)$, opinions have varied over a suitable choice of m - the number of sampling locations(see UPTON and FINGLETON 1985, pp81). Given that the sampling window contains n points, DOGUWA and UPTON(1989a) have chosen the integer k so that

$$k = \begin{cases} \sqrt{n} & \text{if } \sqrt{n} \in I^+ \\ Int(\sqrt{n}+1) & \text{otherwise} \end{cases} \tag{4}$$

where $Int(x)$ means the integer part of x.

Suppose y_i is the value of the random variable Y in the i^{th} sampling location. Then the m locations yield $y_1, y_2, \ldots, y_m$ observations, with mean $\bar{y}$. DE VOS(1973) suggested the use of the statistic

$$M = \frac{\bar{y} - E(\bar{y})}{\sqrt{Var(\bar{y})}} \tag{5}$$

which follows an $N(0,1)$ distribution in an infinite plane, to test the null hypothesis of spatial randomness. For a bounded finite region, DOGUWA and UPTON(1988) use the regular lattice of locations to provide revise expressions for $E(\bar{y})$ and $Var(\bar{y})$ as:

$$E(\bar{y}) = \frac{0.5A^{0.5}}{\sqrt{n}} - \frac{0.5219A^{0.5}}{n^{1.1500}} + \frac{0.1044P}{n^{1.1156}} \tag{6}$$

and

$$Var(\bar{y}) = \frac{0.04445PA^{0.5}}{n^{1.9585}} - \frac{0.1265A}{n^{1.9387}} \tag{7}$$

where A is the area of the region W, P is its perimeter and n is the number of points observed in it. These revised expressions can then be used to test

the randomness hypothesis. The rejection of the null hypothesis of spatial randomness may not be an end in itself, but should be seen as an aid to a more ambitious analysis involving tests based on, say some functional statistics of $H_s(y)$. However such tests are normally based on unbiased estimates of $H_s(y)$.

3 The estimators of $H_s(y)$

We are concerned with the m locations whose positions are known in W. We denote the distance from location i to the nearest boundary of W and to the nearest point within W by r_i and y_i respectively. For a given value of y, there are six situations of interest, which are summarised in Table 1.

3.1 The existing estimators

For an infinite plane, $H_s(y)$ can be estimated by

$$\hat{H_s^0}(y) = \frac{\sum_{i=1}^{m} f(y_i, y)}{m} \tag{8}$$

where

$$f(t,u) = \begin{cases} 1 & \text{if } t \le u \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

However for a finite region W, the estimator $\hat{H_s^0}(y)$, is biased downwards because of edge-effects, and cannot therefore be used to analyse mapped spatial point patterns.

RIPLEY(1977) suggested an estimator for $H_s(y)$ which we denoted as $\hat{H_s^1}(y)$ defined by

$$\hat{H_s^1}(y) = \frac{\sum_{i=1}^{m} f(y_i, y) f(y, r_i)}{\sum_{i=1}^{m} f(y, r_i)} \tag{10}$$

The estimator $\hat{H_s^1}(y)$ restricts attention to the sampling locations belonging to the sets 1, 2 and 3; which are locations that can be surrounded within W by a complete circle center the given location and radius y. As y increases, the number of informative locations contained in these sets diminishes, and

the estimator is undefined whenever $f(y, r_i) = 0 \forall i$. In the case of a rectangular window of dimensions a by b, with $a \leq b$, the maximum value of y for which an estimate of $H_s(y)$ is possible when using the RIPLEY's estimator is $\frac{a}{2}$. This estimator cannot therefore give us any information about the observed pattern for values of $y > \frac{a}{2}$. Furthermore, as y approaches $\frac{a}{2}$ the variance of this estimator will increase considerably.

3.2 The new estimator

Whilst $H_s^{\hat{1}}(y)$ is based on a rational approach to the problem of estimating $H_s(y)$, it will be seen that this estimator uses only the information in the sets 1,2 and 3 of Table 1. In the spirit of DOGUWA and UPTON(1989b) we shall provide another estimator $H_s^{\hat{2}}(y)$, which uses all the information in the six sets of sampling locations given in Table 1.

Consider a realization of a Poisson process of intensity λ, in W. Also consider location i say, within the sampling window. Suppose that $y > r_i$, so that b(i,y), the disc of radius y and centred on location i cuts the boundary of W. Denote the unobservable portion of b(i,y) as u(i,y) and the observable part as o(i,y). Also let the area of u(i,y) be $A_i(y)$. In order to obtain an idea of the value of $H_s^{\hat{2}}(y)$, we must consider the conditional probability $\theta_i(y)$, relating to the observable region o(i,y).

Specifically for sampling locations in set 6 we define $\theta_i(y)$ as

$$\theta_i(y) = \Pr(N[b(i,y)] > 0 \setminus N[o(i,y)] = 0) \tag{11}$$

and for a Poisson process, the above expression reduces to,

$$\theta_i(y) = 1 - e^{-\lambda A_i(y)} \tag{12}$$

From this result, combined with the clear-cut results for the remaining 5 sets of sampling locations, we obtain the new estimator defined by

$$H_s^{\hat{2}}(y) = \frac{\sum_{i=1}^{m}[f(y_i, y) + P_y(r_i, y_i)\hat{\theta_i}(y)]}{m} \tag{13}$$

where

$$P_y(r_i, y_i) = [1 - f(y_i, y)][1 - f(y, r_i)] \tag{14}$$

$$\hat{\theta_i}(y) = 1 - e^{-\frac{\pi A_i(y)}{A}}. \tag{15}$$

Simply stated $P_y(r_i, y_i) = 1$ if location i belongs to set 6 and is zero, otherwise. Thus, the new estimator is the mean score of all the m sampling

locations, where the score is 1 if b(i,y) contains a point, and the score is $\hat{\theta_i}(y)$ if b(i,y) is empty and cuts the boundary of W. We can therefore expect that the extra stability that results from using all the six sets of sampling locations will amply offset any minor bias that has been induced.

4 $H_s(y)$ for a Poisson process.

For a Poisson process, the theoretical values of D(r) - the nearest neighbour distribution and $H_s(y)$ are identical, since the existence of a point at a particular location a_o say, has no bearing on the distribution of the remaining number of points in the disc $b(a_o, y)$. Thus for this process,

$$H_s(y) = 1 - e^{-\lambda \pi y^2} \tag{16}$$

Figure 1a summarises the results obtained from 500 realizations of a Poisson process of intensity $\lambda = 25$ in a rectangular region of sides $10\sqrt{0.1}$ by $\sqrt{0.1}$. This Figure shows the dependence of the bias of the estimators of $H_s(y)$ upon y for the 500 trials (Note that the bias for the three estimators have all been multiplied by 10). It appears that the uncorrected estimator defined in (8) clearly underestimate $H_s(y)$, and should therefore not be used to analyse bounded patterns.

The new estimator defined in (13) is much better than the other estimators in terms of the bias and is therefore preferable. Note that the RIPLEY's estimator defined in (10) cannot be used to estimate $H_s(y)$ for values of y greater than 0.16. While the RIPLEY's estimator has reported a much smaller bias than the uncorrected estimator, it has also recorded a much larger mean squared error than the uncorrected estimator (see Figure 1b). However, the new estimator has in addition to recording a small bias, has also reported a much smaller mean squared error than all the other two estimators. This suggests that the new estimator is preferable.

5 $H_s(y)$ for a Parent-daughter process

In this section we consider a Parent-daughter process in which both parents and mothers are present. Parents are distributed according to a Poisson process with intensity ρ per unit area. To each parent is attached a number of daughters, the number being realised independently for each parent from a Poisson distribution with mean μ.

In order to obtain an explicit formula, DIGGLE(1975) proposed that each of the daughters should be independently distributed uniformly on the circumference of the circle of radius σ centred on the corresponding parent. He further gave an explicit formula for the case when $\sigma = 1$. For any value of $\sigma > 0$, an explicit formula for $H_s(y)$ (see DOGUWA 1988), for the Parent-daughter clustered process is given by

$$H_s(y) = 1 - e^{-\pi\rho(y^2+2\phi(y))} \tag{17}$$

where

$$\phi(y) = \int_{\max(y,\sigma-y)}^{\sigma+y} t[1 - e^{\frac{-\pi\cos^{-1}\beta}{\pi}}]\,dt \tag{18}$$

and

$$\beta = \frac{(t^2+\sigma^2-y^2)}{2t\sigma} \tag{19}$$

For our simulations we used the window $10\sqrt{0.1}$ and $\sqrt{0.1}$, with $\rho = 5$, $\mu = 4$ and $\sigma = 0.1$. In order to obtain a correct representation within the window, the simulation was performed over a larger rectangle and points outside the central rectangle, W were then ignored.

Figure 2a shows the dependence of the bias of the three estimators upon y obtained for the 500 realisations. The most obvious feature of this Figure is the increased bias associated with the uncorrected estimator. RIPLEY's estimator cannot be used to estimate $H_s(y)$ for larger values of y. The new estimator has the least bias for almost all the distances considered.

Figure 2b also shows the dependence of the *mse* of the $H_s(y)$ estimators upon y, obtained for the 500 simulated realisations. The new estimator has recorded a much smaller *mse* than all the other two estimators. This suggests that the new estimator is preferable for clustered patterns. Note also that both the values for the bias and *mse* have been multiplied by 10.

6 $H_s(y)$ for a Lattice process

DIGGLE(1975) suggests superimposing a Poisson process with intensity ρ per unit area upon a square lattice of side d to provide a continous range of patterns from extreme regularity when ρ is zero, towards complete randomness as ρ tends to infinity. For simplicity in the model, we consider the extreme regularity. From the results in DIGGLE, it is easy to see that when $\rho = 0$, the distribution function $H_s(y)$ is given as,

$$H_s(y) = \begin{cases} \frac{\pi y^2}{d^2} & \text{if } 0 \le y < \frac{d}{2} \\ \frac{2y^2\sin^{-1}\gamma + d\sqrt{[4y^2-d^2]}}{d^2} & \text{if } \frac{d}{2} \le y < \frac{d}{\sqrt{2}} \\ 1 & \text{if } y > \frac{d}{\sqrt{2}} \end{cases} \tag{20}$$

where

$$\gamma = \sin^{-1}\left(\frac{d^2-2y^2}{2y^2}\right) \tag{21}$$

In order to generate this realization, a starting point for the Lattice was chosen at random in the sampling window, and the angle of inclination of the Lattice was also chosen at random in $(0, 2\pi)$. All points of the lattice that fall outside the window W were ignored.

Figures 3(a and b) summarize the results of the simulation study obtained from 500 partial realization of the Lattice process generated in the rectangular window. Figure 3a shows the dependence of the bias of the estimators of $H_s(y)$ upon y for the 500 realizations. The two corrected estimators have reported a much smaller bias than the uncorrected estimator. However, RIPLEY's estimator seems to be marginally preferable to the new estimator.

Figure 3b shows the dependence of the *mse* of the estimators upon y. It is of interest to note that even though RIPLEY's estimator has reported a smaller bias than the uncorrected estimator, it has also reported a much larger mean squared error. However the new estimator is more stable in that it has recorded a much smaller *mse* than the uncorrected estimator.

Repeated simulations using different values of these parameters for the three processes considered produce very similar results.

7 Application

As an example of the application of the estimator $\hat{H_s^I}(y)$, we consider some data explored previously by RIPLEY(1977) and DIGGLE(1983). These data illustrated in Figure 4a show the positions of 42 cell centers in the unit square.

To test for deviations from randomness, UPTON and FINGLETON(1985) suggest plotting the function c(y) against y. For the case of n points in a window W of area A, this function is defined by

$$c(y) = \hat{H_s^I}(y) - 1 + e^{-\frac{n\pi y^2}{A}} \qquad (22)$$

For a Poisson process of intensity λ, c(y) will be close to zero. For a clustered alternative, there will be a smaller number of small location to point distances than would be the case in a Poisson process, and so c(y) would be less than zero. However for a regular alternative, there will be a greater number of small location to point distances than would be the case in a random pattern and so c(y) would be much greater than zero.

In order to assess the significance of any departure from a Poisson process, we first simulate 99 Poisson processes consisting of 42 points in the same size window as the Biological cells pattern. Using the new estimator, we estimate c(y) for each simulated realisation, and Figure 4b displays the two dashed lines showing the upper and lower simulation envelopes for the realisations.

The solid line in Figure 4b indicates the value of c(y) obtained for the Biological cells pattern. This line penetrates the upper simulation envelope for some values of y, thus providing evidence at the 1 % level of significant regularity in the pattern.

8 Conclusions

A new method is proposed for estimating the spherical contact distribution $H_s(y)$ for spatial point processes. The proposed estimator is compared with the existing estimators, using three different processes whose theoretical $H_s(y)$ functions are known. The results showed that: (i) for all the values of y considered, a considerable reduction in the *mse* is associated with the proposed estimator. (ii) the bias associated with the proposed estimator was very negligible for all the three processes considered.

ACKNOWLEDGMENTS

The author would like to thank Professor Abdus Salam, the International Atomic Energy Agency and UNESCO for hospitality at the International Centre for Theoretical Physics, Trieste. He is also grateful to Ahmadu Bello University, Zaria, Nigeria, for allowing him to visit ICTP as visiting mathematician.

References

[1] DE VOS, S.(1973). The use of nearest neighbour methods. Tijdschrift voor Economische Geografie. 64, 307-319.

[2] DIGGLE, P.J.(1975). Robust density estimation using distance methods. Biometrika. 62, 39-48.

[3] DIGGLE, P.J.(1979). Statistical methods for spatial point patterns in ecology. In Spatial and Temporal Analysis in Ecology. (R.M. Cormack and J.K.Ords. eds). Finland: International Cooperative Publishing House, 95-150.

[4] DIGGLE, P.J.(1983).Statistical Analysis of Spatial Point Patterns. Academic Press: London.

[5] DOGUWA,S.I.(1988).Statistical Analysis of Mapped Spatial Point Patterns. Ph.D Thesis. University of Essex: England.

[6] DOGUWA,S.I. and UPTON,G.J.G.(1988). On Edge-corrections for the Point-Event analogue of the Clark-Evans statistic. Biom. J. 30, 957-963.

[7] DOGUWA,S.I. and UPTON,G.J.G.(1989a). Simulations to determine the mean and variance of the Point-Object analogue of the Clark-Evans Statistic. Biom. J. 31, 163-170.

[8] DOGUWA,S.I. and UPTON,G.J.G.(1989b). On the estimation of the nearest neighbour distribution, D(y) for point processes. To appear in Biom. J.

[9] RIPLEY,B.D.(1977). Modelling spatial patterns(with discussion). J. Roy. Statist. Soc. B 39, 172-212.

[10] STOYAN,D.,KENDALL,W.S. and MECKE,J.(1986). Stochastic Geometry and its Applications. Akademie-Verlag: Berlin.

[11] UPTON,G.J.G. and FINGLETON,B.(1985). Spatial Data Analysis by Example: Volume 1, Point Pattern and Quantitative Data. New York: Wiley.

TABLE 1

Usage of the sampling locations in W for the various methods of estimating $H_s(y)$

		Estimator defined in			
Set	Description	(8)	(10)	(13)	Is $N[b(i,y)] > 0$?
1	$y < r_i \leq y_i$	⋆	⋆	⋆	no
2	$y \leq y_i < r_i$	⋆	⋆	⋆	no
3	$y_i \leq y \leq r_i$	⋆	⋆	⋆	yes
4	$y_i \leq r_i < y$	⋆	o	⋆	yes
5	$r_i < y_i \leq y$	⋆	o	⋆	yes
6	$r_i < y < y_i$	⋆	o	⋆	maybe

key for Table: ⋆ means the sampling location is used in the estimation of $H^*(y)$. o means the sampling location is not used in the estimation.

Captions for Figures

Fig 1: The bias (a) and the mse (b) of the three estimators of $H_s(y)$ for a Poisson process of intensity $\lambda = 25$, in a rectangular region of sides $10\sqrt{(0.1)}$ by $\sqrt{(0.1)}$.

Fig 2: The bias (a) and the mse (b) of the three estimators of $H_s(y)$ for a Parent-daughter clustered process with parameters $\rho = 5$, $\mu = 4$, and $\sigma = 0.1$ in a rectangular window $10\sqrt{(0.1)}$ by $\sqrt{(0.1)}$.

Fig 3: The bias (a) and the mse (b) of the three estimators of $H_s(y)$ for the Lattice process with parameter $d = 0.25$ in a rectangular region of side $10\sqrt{(0.1)}$ by $\sqrt{(0.1)}$.

Key for Figs1, 2 and 3: $\hat{H_s^0}(y)$ is denoted by dashed line with a $\circ$ on it. The dashed line with a square on it denotes $\hat{H_s^1}(y)$. The solid line with a $\star$ on it, denotes $\hat{H_s^2}(y$.

Fig 4: (a) The positions of 42 cell centers in a unit square. (b) The values of c(y) for the cells data (solid line), together with the corresponding envelope (dashed lines), resulting from 99 simulations of a Poisson process of the same Point intensity and in the same size window.

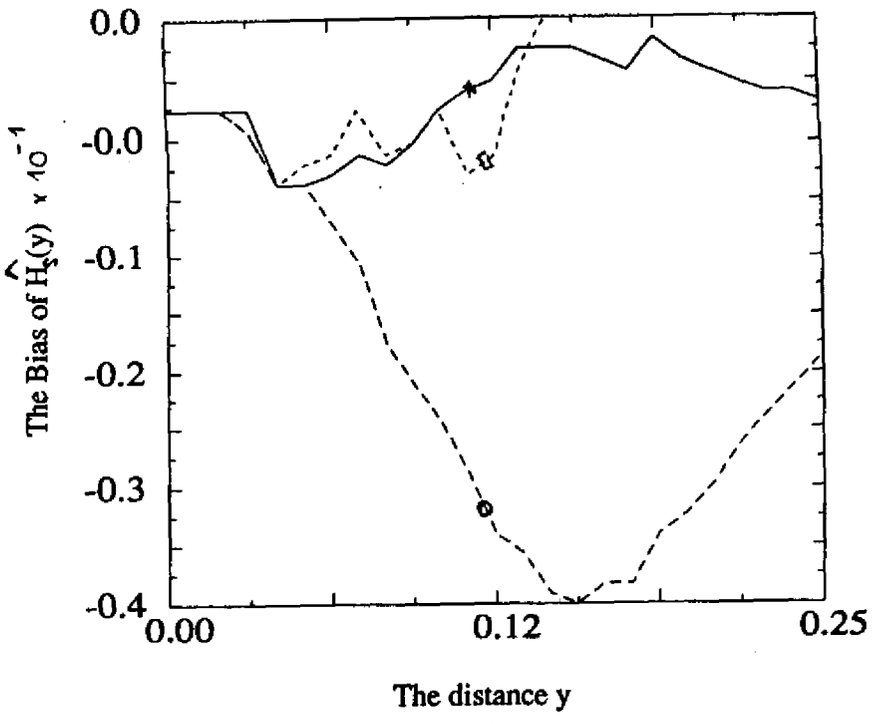

Fig.1a

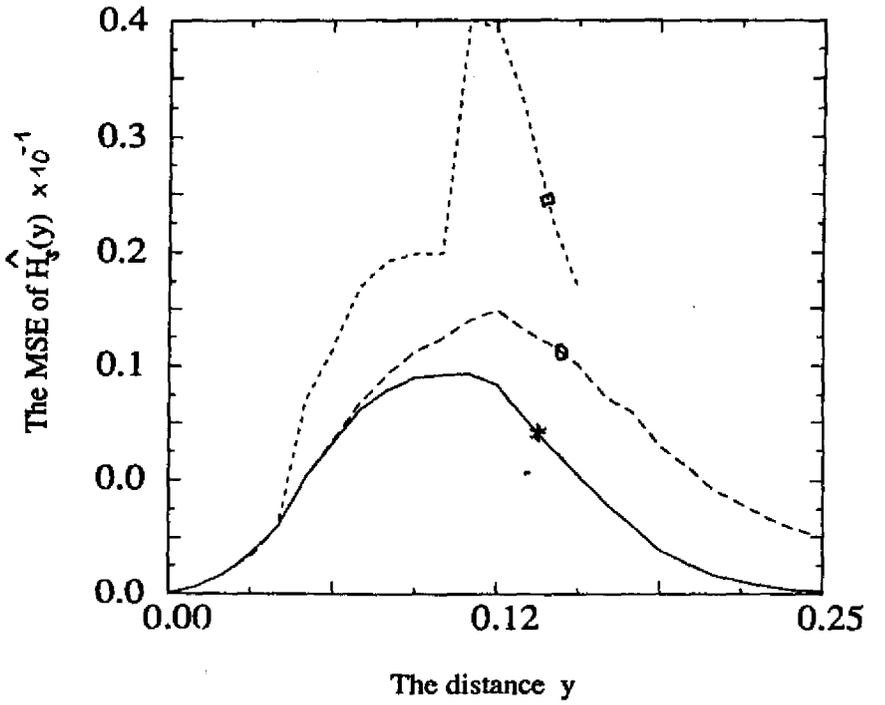

Fig.1b

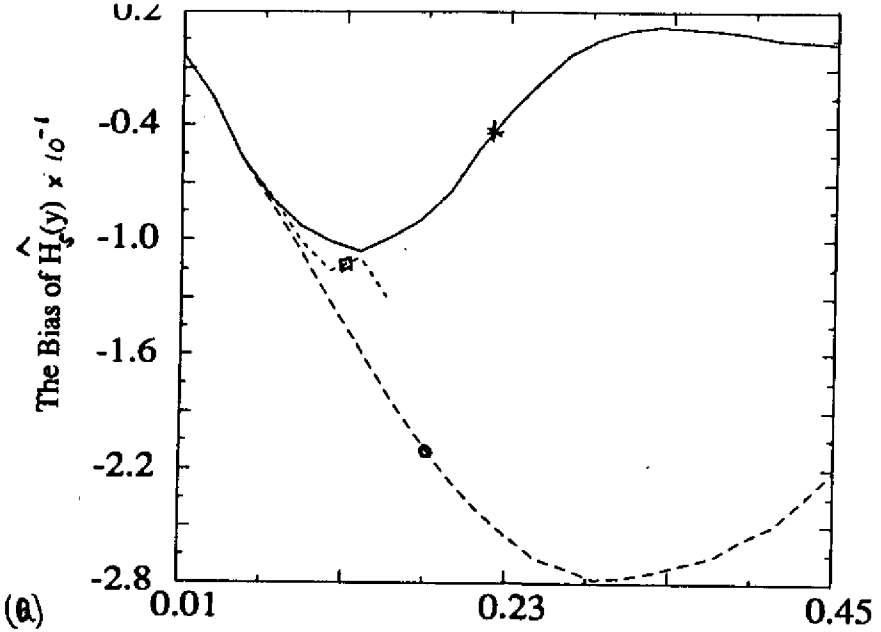

Fig.2a

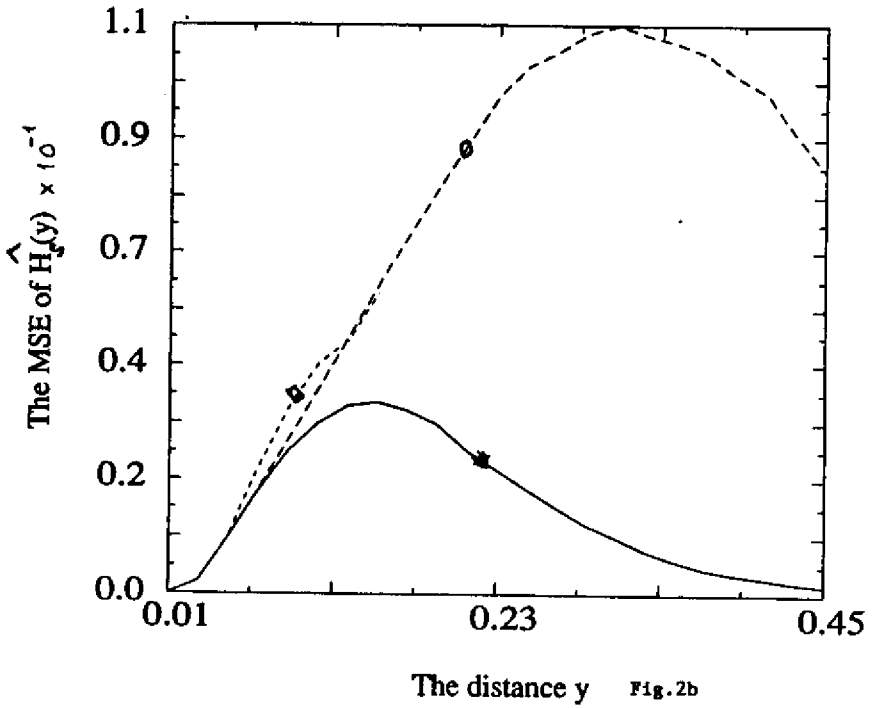

Fig.2b

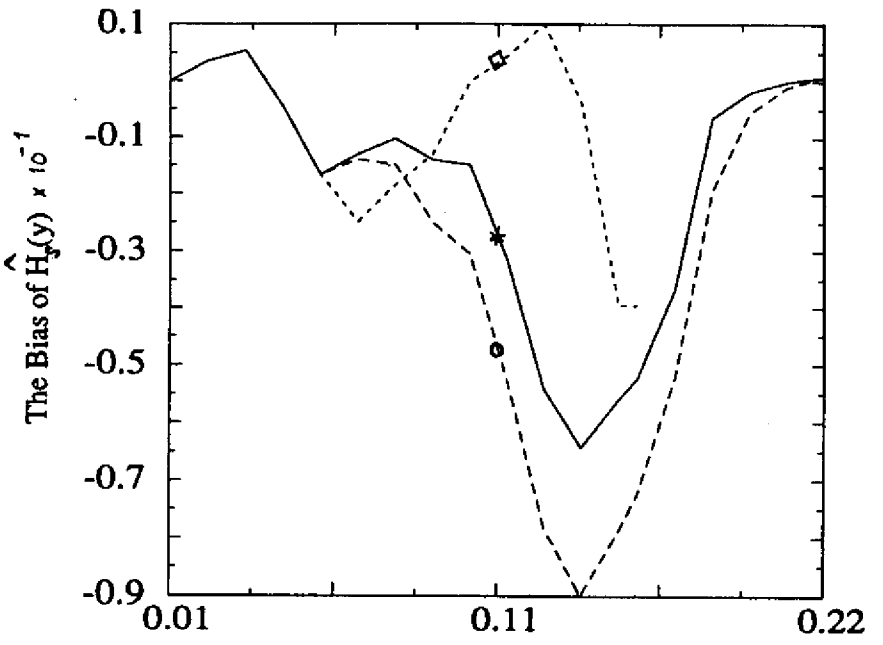

Fig.3a

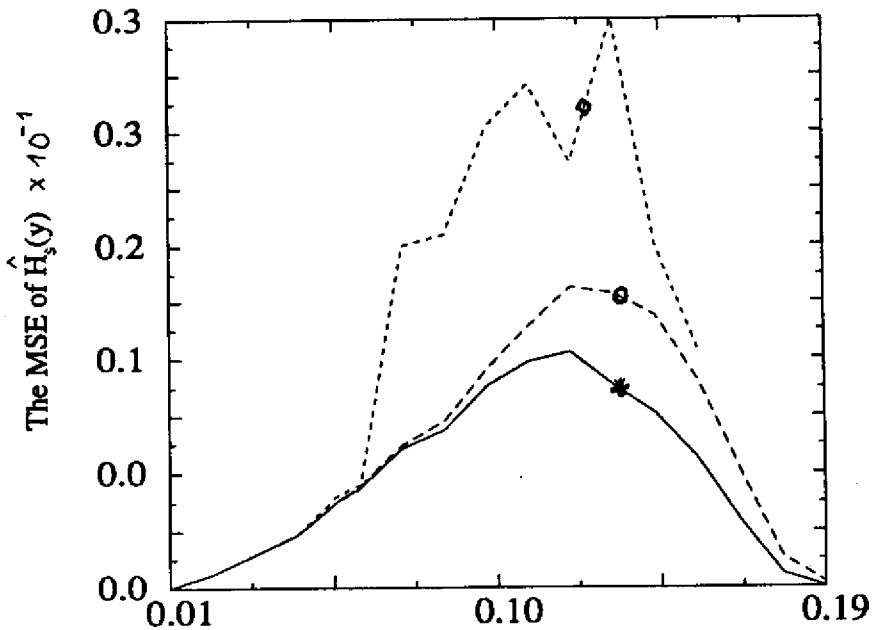

Fig.3b

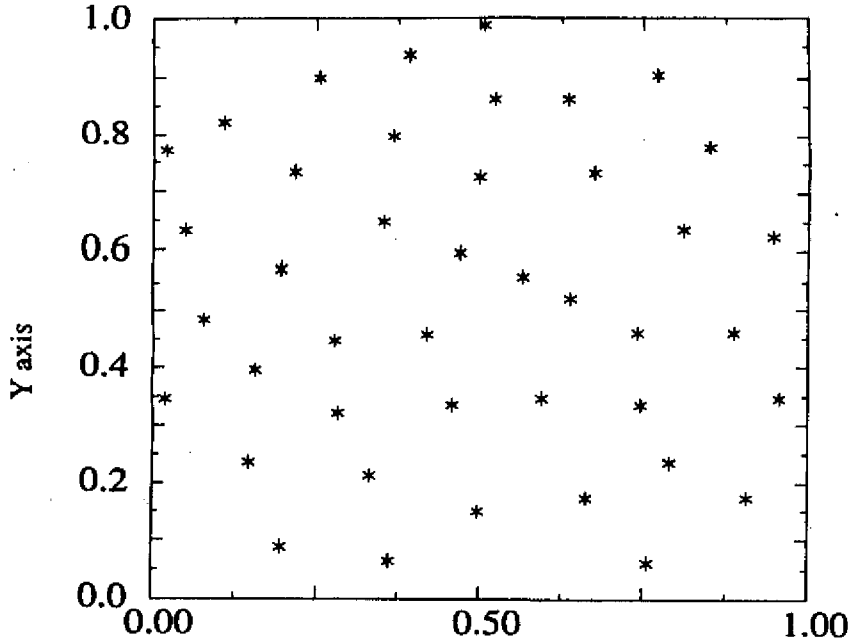
1.0
0.8
0.6
0.4
0.2
0.0
Y axis
0.00
0.50
1.00
X axis

Fig.4a

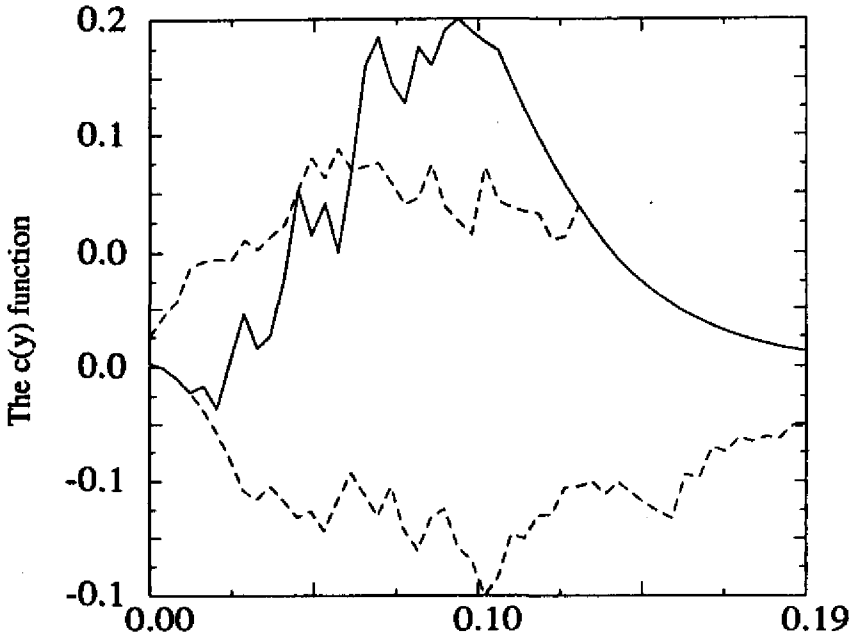
0.2
0.1
0.0
0.0
-0.1
-0.1
The c(y) function
0.00
0.10
0.19
The distance y

Fig.4b