Validating Clusters using the Hopkins Statistic

Amit Banerjee
Department of Mechanical Engineering
New Jersey Institute of Technology
Newark, NJ 07102
E-mail: ab2@njit.edu

Rajesh N. Davé
Department of Mechanical Engineering
New Jersey Institute of Technology
Newark, NJ 07102
E-mail: dave@njit.edu

Abstract – **A novel scheme for cluster validity using a test for random position hypothesis is proposed. The random position hypothesis is tested against an alternative *clustered* hypothesis on every cluster produced by a partitioning algorithm. A test statistic such as the well-known Hopkins statistic could be used as a basis to accept or reject the random position hypothesis, which is also the null hypothesis in this case. The Hopkins statistic is known to be a fair estimator of randomness in a data set. The concept is borrowed from the clustering tendency domain and its applicability to validating clusters is shown here using two artificially constructed test data sets.**

I. Introduction

This paper examines a vital issue closely related to the problem of clustering, that of validating results of clustering procedures. Clustering attempts to partition a data set into self similar groups, called clusters. The literature is abound with references to clustering procedures and algorithms on a wide variety of data sets, on the other hand not much attention has been paid to related issues such as clustering tendency and cluster validity. Before subjecting a data set to partitioning, one needs to be confident whether the data set exhibits a predisposition to cluster into natural groups without identifying the groups themselves [1] and this is what constitutes the clustering tendency domain. A data set without inherent natural clusters could be thought of as a random collection of feature vectors and such random data sets should not be subject to partitioning. Clustering tendency studies could also give information about the general nature of the data and hence indicate which clustering procedure to use, if any, on the data set. Another related issue is the validation of clustering results, in a quantitative and objective manner. Clustering tendency studies do not give any information about the number of natural clusters in the data; it only provides (almost always subjective) information about the presence of natural groupings. Furthermore any partitioning algorithm would find any pre-specified number of clusters (ranging from 2 to n-1, where n is the number of patterns to be classified), irrespective of the actual number of natural groups present. The best and the most natural partition is the one that captures the inherent natural grouping in the data, and cluster validity aims at finding the most natural partition among the many partitions generated by the clustering scheme. The issue of *unknown number of clusters* also falls under the purview of cluster validity studies.

Statistical measures used to validate clustering results are based on the premise that problems of cluster validity are inherently statistical [1]. A result is usually tested by building an alternative hypothesis and comparing it against a null hypothesis, H_0. The null hypothesis is a statement of randomness and could be based on a random graph, a random label or a random position hypothesis. An alterative hypothesis, H_a is then a statement of *orderliness* and captures the intent of the phrase – "the data are clustered". The test is then one of comparing H_0 with H_a based on the value of some test statistic T and deciding whether to accept or reject H_0 with a certain degree of certainty. Hubert's Γ statistic [2] and the Goodman-Kruskal γ statistic [3] are well known examples of test statistics used for cluster validity studies. The Γ statistic compares a clustering structure which is the result of a clustering scheme, to an *a-priori* structure, such as one generated by using pre-defined category labels. The γ statistic measures the rank correlation between two ordinal sequences of numbers, one of which might be derived from the clustering structure and the other from an *a-priori* labeling scheme. For a detailed discussion on statistical cluster validity statistics, tests and indices, the reader is referred to [1].

Several non-statistical cluster validity indices and procedures were independently developed within the fuzzy clustering community to validate partitions obtained using fuzzy clustering algorithms. Prominent among these are the partition coefficient [4], classification entropy [5], proportion exponent [6], the uniform data functional [7], nonfuzziness index [8], information ratio [9], separation ratio [10] and perhaps the best known of the validity indices, the Xie-Beni index [11]. These criteria measure the fuzzy overlap between clusters, with or without regard to the geometrical properties of the data set. Some indices work directly on the fuzzy clustering outputs while the others first convert the results to a hard c-partition before evaluating it. Specialized partitioning schemes such as shell partitioning require the use of specialized validity indices such as the partition density and the shell thickness measure [12]. More recently advances have been made in visual assessment of clustering using intensity displays [13,14]. Some indices are known to function poorly across a wide range of data sets while others are specifically suited to a particular type of data set. In other words their dependence on data sets and on the type of clustering scheme employed, seriously hinder the practical usage of fuzzy validity indices.

In this paper, we borrow a statistical concept from the field of clustering tendency and show its applicability to validate clustering results generated in a partitioned data.

0-7803-8353-2/04/$20.00 © 2004 IEEE

II. SPARSE SAMPLING TESTS AND THE HOPKINS STATISTIC

The problem of testing for clustering tendency can also be described as problem of testing for spatial randomness. Unlike statistic based cluster validity measures, a test for clustering tendency is stated in terms of an internal criterion and no *a-priori* information is brought into the analysis [1]. The null hypothesis in most cases is a random position hypothesis, such as,

H_0: The patterns are generated by a Poisson process with an intensity of L patterns per unit volume.

Under H_0, the number of patterns falling in a region of volume V has a Poisson distribution with mean LV and since L is constant and the numbers of patterns falling in disjoint regions of V are independent random variables, the Poisson process is a reasonable model for randomness (absence of structure). Sparse sampling tests have been shown to have high power against clustered alternative hypotheses. On the other hand, tests based on small interpattern distances (such as nearest neighbor distance tests) have low power against clustered alternatives primarily because such tests depend heavily on the intensity L of the Poisson process assumed under H_0. Other tests for spatial randomness include Scan tests [15], Quadrat analysis [16] and Second moment structure tests [17].

Sparse sampling tests are based on sampling origins randomly identified in a sampling window. Several tests involving sampling origins have been proposed in literature, based on a multitude of test statistics such as the Hopkins [18], Holgate [19,20], T-square [21], Eberhardt [22] and the Cox-Lewis statistic [23]. These statistics have been compared but little is known about any one of them outperforming the others. The Hopkins statistic is easy to use and comprehend and has been shown to be as good as the Holgate statistic [24].

Let $X = \{x_i \mid i=1 \text{ to } n\}$ is a collection of n patterns in a d-dimensional space such that $x_i = \{x_{i1}, x_{i2}, .., x_{id}\}$. Also, let $Y = \{y_j \mid j=1 \text{ to } m\}$ be m sampling origins placed at random in the d-dimensioned sampling window, $m << n$. A sampling window can be thought of as a subspace of the entire d-dimensioned sample space. Two types of distances are defined: u_j as the minimum distance from y_j to its nearest pattern in X and w_j as the minimum distance from a randomly selected pattern in X to its nearest neighbor (m out of the available n patterns are marked at random for this purpose). The Hopkins statistic in d-dimensions is defined as,

$$H = \frac{\sum_{j=1}^{m} u_j^d}{\sum_{j=1}^{m} u_j^d + \sum_{j=1}^{m} w_j^d} \qquad (1)$$

This statistic compares the nearest-neighbor distribution of randomly selected locations to that for the randomly selected patterns. Under the null hypothesis, H_0, the distances from the sampling origins to their nearest patterns should, on the average, be the same as the interpattern nearest neighbor distances, implying randomness and hence H should be about 0.5. However when the patterns are aggregated or clustered into clouds, the sampling origin to pattern nearest neighbor distances should, on the average, be larger than the randomly selected interpattern nearest neighbor distances, so H should be larger than 0.5, almost equal to 1.0 for very well defined clustered data. By the same reasoning, H is supposed to be much less than 0.5 for regularly spaced data, data that are neither clustered nor random. To ensure that no pattern is the neighbor of more than one sampling origin, m is chosen to be substantially less than n; it has been suggested that $m < 0.1n$ [17]. With such a condition, it can be ensured that all $2m$ nearest neighbor distances are statistically independent and H has a beta distribution with parameters (m, m), independent of both the intensity L and the dimensionality of the data set d. The distribution and the density function of each of the terms in (1) are also known, the individual sums each have a gamma distribution (assuming again that the nearest neighbor distances are all independent random variables). Our studies on random data sets, clustered data sets and regularly spaced data sets show that Hopkins statistic H consistently has a value of around 0.5, 0.7-0.99 and 0.01-0.3 respectively and is hence a powerful estimator of randomness.

III. RANDOM POSITION TESTS FOR CLUSTER VALIDITY

A cluster is characterized by two main properties – compactness and isolation [1]. A natural cluster is *unusually* compact and *unusually* isolated. A clustered data set is ordered because of the presence of natural clusters; in the absence of natural groups, it is a random collection of data points, approximating a Poisson process distribution. In this section we show the applicability of the random position hypothesis and use the Hopkins statistic as a measure for cluster validity. Suppose a data set with 3 compact and isolated clusters as shown in fig. 1, is subject to partitioning. At $c = 2$, where c is the number of clusters to be found, most partitioning schemes would club clusters II and III together as one cluster, say cluster A and identify cluster I as an independent cluster, B. A random position hypothesis test would lead to the rejection of H_0 for cluster A because it yet is a collection of clustered data points. However it would be difficult to reject H_0 in case of cluster B because it is the natural cluster, cluster I. *A natural cluster hence apart from being isolated and compact is also random within itself.* Intracluster data might also exhibit some kind of mutual repulsion as in fig. 1 and in such a case the null hypothesis H_0 should include a statement conforming to random position as well as a non-random non-clustered (regularly spaced) distribution. The null hypothesis then can not be rejected if the data is either random or regularly spaced. However in real world clustering problems, this is rarely the case and so a random position hypothesis alone should suffice. At $c = 3$ however all the three natural clusters are most likely to be identified during partition and hence it would be difficult to reject H_0 for all the

three clusters identified. The rejection or acceptance of H_0 depends on the value of the Hopkins statistic. At any value of $c > 3$, any partitioning algorithm would either subdivide or recombine the clusters that were produced by partitioning at $c = 3$ and hence one might not have reason to reject H_0 for any of the clusters (in case clusters are subdivided further) or in some cases just enough evidence to reject H_0 (in case clusters are recombined) for some of the generated clusters. The test for cluster validity based on the random position hypothesis can hence be stated as follows – accept the lowest value of c at which it is impossible to reject the null hypothesis H_0 for all the clusters, the test applied one cluster at a time. Let H_i be the value of the Hopkins statistic for the i^{th} cluster at a particular level of clustering, the average value of the statistic is then,

$$H_{av} = \frac{1}{c}\sum_{i=1}^{c} H_i \text{ , for fixed } c \qquad (2)$$

A non-rejection of H_0, as given in section II, would mean that on an average, the value of H_{av} is very close to 0.5. Proceeding from $c = 2$ to $c = n - 1$, the lowest value of c that generates $H_{av} \approx 0.5$, most likely generates a partition that identifies the natural clusters in the data.

IV. Simulations and Results

In this section we present some cluster validity studies done on two artificially produced two-dimensional data sets, (1) 4-cloud data comprising of 160 patterns and (2) 7-cloud data comprising of 1400 patterns. The patterns are generated within a pre-specified window using the C++ *rand* function, which produces pseudo-random numbers using a seed initialized by the CPU clock time. The data sets are partitioned using FCM [25], ranging from $c = 2$ to $c = 8$ for the 4-cloud data set and from $c = 2$ to $c = 11$ for the 7-cloud set.

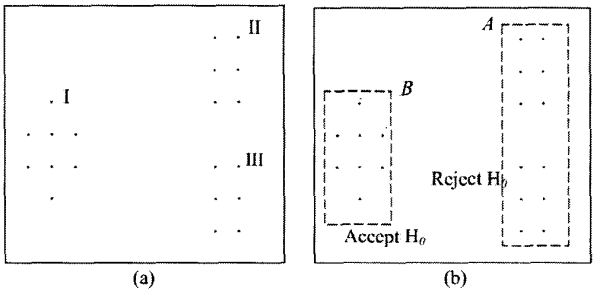


Fig.1. A three cluster data set. (a) the three natural clusters, (b) clusters identified at $c = 2$.

The fuzzy partitions are first converted into hard partitions and then all the generated clusters are subject to the random position hypothesis test. The sampling window in all cases encompasses the entire cluster set and the number of sampling origins; m is chosen to be $n/10$ (or the closest integer value), where n is the number of patterns in the cluster being investigated. In case there were less than 10 patterns assigned to a cluster, we choose $m = 1$. The value of the Hopkins statistic, H for each of the clusters is then calculated and the process repeated over 5000 runs. The hard-partition is retained if H does not fluctuate more than 0.05 over these 5000 runs and in case the fluctuation is more than 0.05, the FCM partitioning is repeated to get a better resultant hard-partition. The average of these 5000 values constitutes a consolidated value of the Hopkins statistic for that particular cluster. The average partition Hopkins statistic is then calculated using (2) and the results plotted against the number of clusters, c.

The 4-cloud data is shown in fig. 2. The resultant average partition Hopkins statistic, H_{av} is plotted against the number of clusters, c and shown in fig. 3. As can be seen, the null hypothesis cannot be accepted for $c = 2$ ($H_{av} = 0.76$) and $c = 3$ ($H_{av} = 0.64$). However it can be accepted with a fair degree of confidence for $c = 4$ ($H_{av} = 0.47$) and thenceforth. Hence according to the cluster validity criterion enunciated in section III, one could accept $c = 4$ as the partition identifying the natural groupings in the data, which is indeed the case.

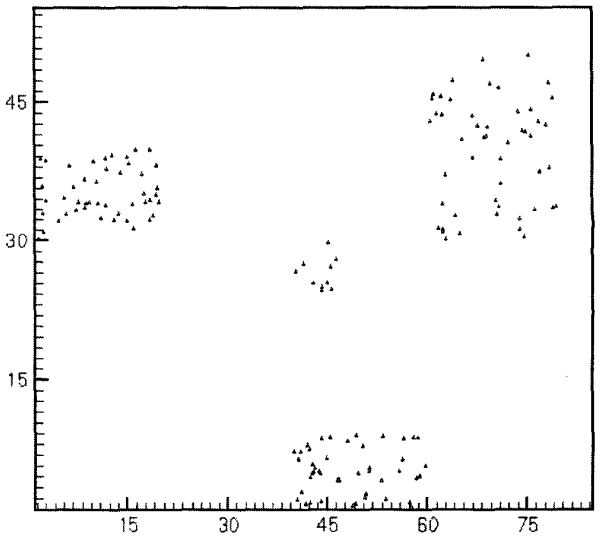


Fig. 2. The 160 pattern 4-cloud data set (3 clusters of 50 patterns each and 1 cluster of 10 patterns)

The 7-cloud data set is different from the 4-cloud data in that the former is not as well separated as the latter, as shown in fig. 4. The applicability of the Hopkins statistic as a cluster validity index and the random position hypothesis test as an appropriate test for cluster validity, are illustrated in a much broader sense in this case. The two clusters in the lower left hand corner of fig. 4 overlap each other and can be argued to be one big tilted 8-shaped cluster. This can be seen from the plot of the average partition Hopkins statistic, H_{av} versus the number of clusters in fig. 5; it is difficult to chose between $c = 6$ ($H_{av} = 0.53$) and $c = 7$ ($H_{av} = 0.48$). Other values of c can be rejected outright. Hence the Hopkins statistic does reflect the nuances and subtleties in the data set. In the absence of

the overlap, the average partition Hopkins statistic would have indicated a clear preference for $c = 7$, suggesting natural grouping at that level of partitioning.

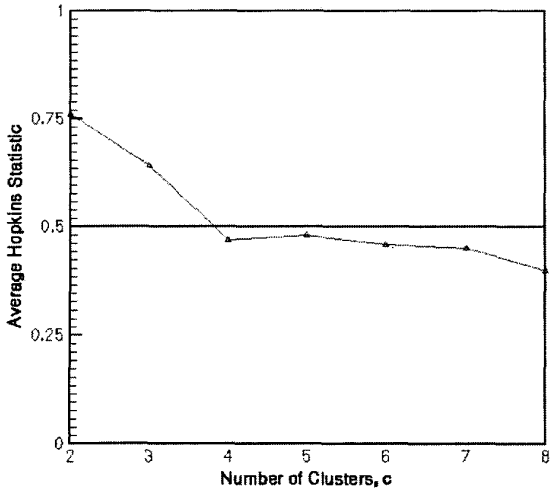


Fig. 3. H_{av} plotted against c, for the 4-cloud data set

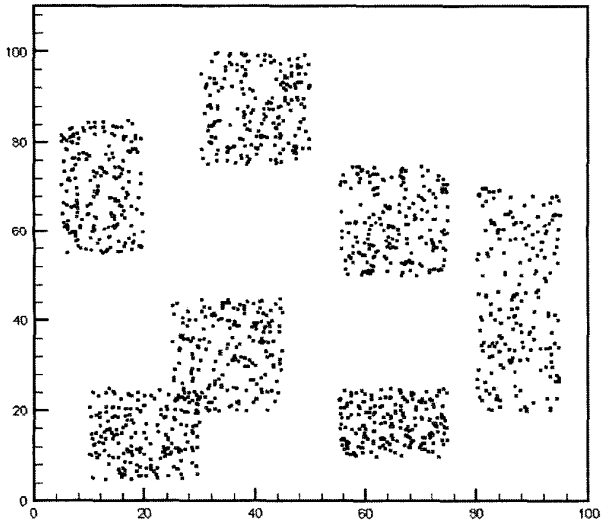


Fig. 4. The 1400 pattern 7-cloud data set (200 patterns per cluster)

In case the clusters are not as well separated as they are in these two test cases, one might need to accept or reject H_0 for individual clusters one at a time, instead on making a decision based on H_{av}. As in the 4-cloud data set, one could reject H_0 for at least one generated cluster for both $c = 2$ and $c = 3$ partitions. However at $c = 4$, H_0 can not be rejected for any of the 4 generated clusters. For the 7-cloud data set, one could reject H_0 for at least one cluster in the range $2 \leq c \leq 5$. However it becomes difficult to reject H_0 for any of the clusters generated at $c = 6$ and $c = 7$. In case there is no overlap of clusters in the lower left hand corner, one could have out rightly rejected H_0 for at least one cluster in the range $2 \leq c \leq 6$ and $c = 7$ would have been the smallest value of c where one can not reject H_0 for any of the 7 clusters.

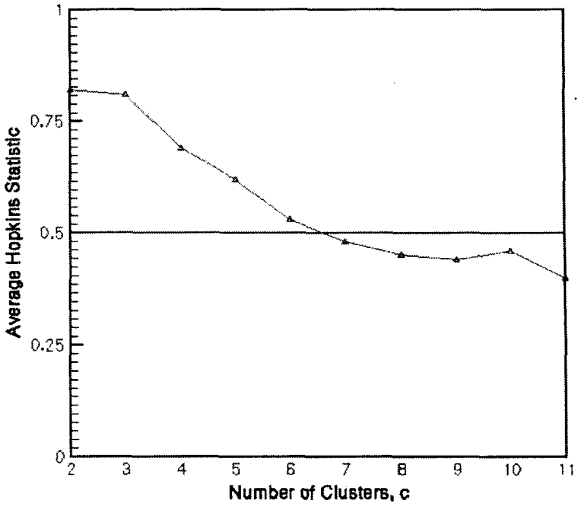


Fig. 5. H_{av} plotted against c, for the 7-cloud data set

V. Discussion and Conclusions

In this paper, we have demonstrated the applicability of the random position hypothesis test as a criterion for cluster validity and used one of the well known test statistics, the Hopkins statistic as an index for the test. The random position hypothesis and the Hopkins statistic have been used previously in the context of clustering tendency. Showing that with virtually no change in form, the same could be applied to testing for cluster validity, is a novel idea. The idea has been validated on test cases, both small (160 patterns) and big (1400 patterns). Generation of sampling windows and sampling origins, locating random patterns to generate nearest-neighbor distances and the associated calculations are computationally inexpensive on powerful desktop PCs of the day. The program to generate fuzzy partitions using FCM, defuzzify the results and then test each partition for randomness, was written in C++, compiled using a visual C++ compiler on a 32-bit windows PC environment. To generate $c = 2$ to $c = 11$ partitions on the 7-cloud data set and to get the H_{av} value (50,000 runs of the Hopkins statistic code) took approximately 2.5s CPU time.

The applicability of the Hopkins statistic and the random position hypothesis test for cluster validity to non-cloud data (such as detection of lines in a data set) needs to be investigated. The key in applying the theory to linearly clustered data might rest on an appropriate selection of the sampling window. A skewed line cluster within a rectangular sampling window might appear to be a clustered (non-random) collection of data because a line cluster is neither an isolated nor a compact cluster. It is also impossible to apply the Hopkins

statistic on very small data sets where each cluster might contain just 4-5 patterns each. At such intensities of pattern distribution, the theory of randomness does not hold and hence the random position hypothesis is meaningless. However this is a blessing in disguise because very small data sets are rarely encountered in real world situations; in fact it could be claimed that the random position test for cluster validity produces better results as the data set gets larger in size (with increasing n). The restriction however being that the data should be *isolated* in groups and *compact* within the groups. The theory can easily be extended to data in more than two-dimensions because the Hopkins statistic is essentially defined in d-dimensions. One could even use other statistics, such as the Cox-Lewis statistic [23] which extends better in d-dimensions ($d > 2$) than the Hopkins statistic. The random position hypothesis test for cluster validity using an index such as the Hopkins statistic not only tends to provide an answer to the ever-elusive question – *"how many clusters to find?"* but also provides a validation measure for individual clusters. If H_0 cannot be rejected for any of the clusters at the lowest $c = c^*$ partitioning, then one could argue that all the clusters found are in fact the true natural clusters in the data. Not intended to replace the existing cluster validity techniques and indices, the test for random hypothesis is an interesting and promising addition to the repertoire of cluster validity methodologies.

Future research might focus on applying the test directly on the fuzzy partition instead of having to defuzzify the c-partition first. The fuzzy membership function can be used to weigh some (or all) of the distances used in the calculation of the Hopkins statistic. The complete partition could then be evaluated at one go, instead of one cluster at a time.

REFERENCES

[1] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, NJ, 1988.

[2] L. J. Hubert and J. Schultz, "Quadratic assignment as a general data-analysis strategy," *British J. Math. Statist. Psych.*, vol. 29, pp. 191-241, 1976.

[3] L. A. Goodman and W. H. Kruskal, "Measures for association for cross-classifications," *JASA*, vol. 49, pp. 732-764, 1954.

[4] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," *J. Math. Biol.*, vol. 1, pp. 57-71, 1974.

[5] J. C. Bezdek, "Mathematical models for systematics and taxonomy," In *Proc. 8th Int. Conf. Numerical Taxonomy*, Ed G. Estabrook, San Francisco, CA: Freeman, pp. 143-166, 1975.

[6] M. P. Windham, "Cluster validity for fuzzy clustering algorithms," *Fuzzy Sets Syst.*, vol. 5(2), pp. 177-185, 1981.

[7] M. P. Windham, "Cluster validity for fuzzy c-means clustering algorithm," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-4(4), pp. 357-363, 1982.

[8] G. Libert and M. Roubens, "New experimental results in cluster validity of fuzzy clustering algorithms," In *New Trends in Data Anal. and Appl.*, Eds J. Janssen, J.-F. Macrotorchino and J.-M. Proth, Amsterdam, The Netherlands: North Holland, pp. 205-218, 1983.

[9] M. P. Windham, H. Bock and H. F. Walker, "Clustering information from convergence rate," In *Proc. 2nd Conf. Int. Federation Classification Soc.*, Washington D.C., pp. 143, 1989.

[10] R. Gunderson, "Application of fuzzy ISODATA algorithms to star tracker pointing systems," In *Proc. 7th Triennial World IFAC Cong.*, Helsinki, Finland, pp. 1319-1323, 1978.

[11] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-13(8), pp. 841-847, 1991.

[12] R. N. Davé, "Validating fuzzy partitions obtained through c-shells clustering," *Patt. Recog. Letters*, vol. 17, pp. 613-623, 1996.

[13] J. C. Bezdek and R. J. Hathaway, "VAT: A tool for visual assessment of (cluster) tendency," In *Proc. IJCNN*, IEEE Press, Piscataway, NJ, pp. 2225-2230, 2002.

[14] J. C. Bezdek and R. J. Hathaway, "Visual cluster validity displays for prototype generator clustering methods," In *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE 2003)*, pp. 875-880, 2003.

[15] J. I. Naus, "Approximations for distributions of scan statistics," *JASA*, vol. 17, pp. 177-183, 1982.

[16] R. Mead, "A test for spatial pattern at several scales using data from a grid of contiguous quadrats," *Biometrics*, vol. 30, pp. 295-308, 1974.

[17] B. D. Ripley, "Modelling spatial patterns," *J. Royal Statis. Soc.*, B39, pp. 172-212, 1977.

[18] B. Hopkins, "A new method of determining the type of distribution of plant individuals," *Annals of Botany*, vol. 18, pp. 213-226, 1954.

[19] P. Holgate, "Some new tests of randomness," *J. Ecol.*, vol. 53, pp. 261-266, 1965.

[20] P. Holgate, "Tests of randomness based on distance measures," *Biometrika*, vol. 52, pp.345-, 1965.

[21] J. E. Besag and J. T. Gleaves, "On the detection of spatial pattern in plant communities," *Bull. Int. Statist. Inst.*, vol. 45, pp. 153-158, 1973.

[22] L. L. Eberhardt, "Some developments in distance sampling," *Biometrics*, vol. 23, pp. 207-216, 1967.

[23] T. F. Cox and T. Lewis, "A conditioned distance ratio method for analyzing spatial patterns," *Biometrika*, vol. 63, pp. 483-491, 1976.

[24] E. Panayirci and R. C. Dubes, "A test for multidimensional clustering tendency," *Patt. Recog.*, vol 16(4), pp. 433-444, 1983.

[25] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, NY, 1981.