New Index for Clustering Tendency and Its Application to Chemical Problems

RICHARD G. LAWSON and PETER C. JURS*

152 Davey Laboratory, Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received August 18, 1989

The tendency of a multivariate data set to cluster can be quantified by numerical indices. A new clustering index, related to the Hopkins' statistic, is presented that calculates the degree to which a data set is clustered compared to variables with the same univariate distributions. This new index is shown to provide a more realistic assessment of clustering tendency for real data sets than the standard Hopkins' statistic. The new index is tested with both artificial and real chemical data sets to investigate its properties.

INTRODUCTION

Cluster analysis is a valuable tool used to explore chemical problems ranging from elemental distributions in coal, to environmental sample analysis, to monitoring malt in process analytical chemistry, to seeking structural relationships among sets of pharmaceutical drugs.[1-4] Exploration in this sense refers to looking for organization in a set of measured or calculated physical and chemical properties for a set of compounds. For example, all low molecular weight insect juvenile hormone mimetics with low partition coefficients, high σ charge on a specified oxygen, and a relatively large degree of branching may fall into a group distinct from all other compounds. Each juvenile hormone mimetic in this example would be represented by a data point in four-dimensional space, in which the four axes represented molecular weight, partition coefficient, σ charge, and branching index. Two-dimensional plots of properties such as these may not contain sufficient information to separate the compounds into distinct groups or clusters, but computer-based clustering programs may be able to detect sets of data points that are close to each other and relatively distant from the remaining points, even in high-dimensional space. This tendency for data points to group together on the basis of chemical properties may provide new insight into chemical relationships. However, before these relationships in a set of molecules are looked for, it is important to be sure that the data set is amenable to clustering. Much of the software used for clustering studies will report some kind of clustering regardless of the organization in the data. Forcing unstructured data into clusters would not only waste time and effort, but could lead to erroneous conclusions about data organization.

Part of the solution to this problem is to examine the data set of interest for its tendency to cluster as the first step of a study.[5] If a set of compounds and their associated descriptors tend to aggregate, clustering programs could be appropriate in investigating structure. If there were no significant tendency to aggregate, then cluster analysis would not be appropriate. There is no absolute measure of clustering tendency, so to determine if there is sufficient clustering tendency to proceed with cluster analysis, the data set being investigated is compared to a data set with a known amount of structure. In statistical terms this involves forming a null hypothesis, H_0, calculating a test statistic, and then comparing the test statistic to standard values to determine the degree of confidence in rejecting the null hypothesis. A natural and commonly used standard of structure is a uniform distribution of random numbers. The null hypothesis corresponding to this standard would be that the data set in question has no more tendency to cluster than uniformly distributed random numbers.

One method to determine clustering tendency in chemical systems has been described by Willett, and applied to several data sets from the chemical literature.[4,6] This graph theoretic method, based on work by Ling and Killough, provides a degree of confidence in the clustering tendency of a data set.[7,8] There are several limitations to this method that restrict its application, however. The most restrictive limitation is an upper limit of 100 objects due to the time-intensive nature of the necessary calculations. Many data sets in practice contain several hundred or even several thousand compounds. In addition, the method is sensitive to the presence of outliers. Finally, the only favorable conclusion that can be drawn in the best case is that the data are more clustered than are uniformly distributed random numbers.

Researchers outside chemistry have explored other ways to determine cluster tendency. Zeng and Dubes compared five commonly used methods on a wide variety of data sets with different dimensionality, different degrees of cluster separation, and different sampling windows.[9,10] After an extensive series of Monte Carlo experiments comparing the Hopkins' statistic, the Cox-Lewis statistic, the Eberhardt statistic, and two T-square statistics, they concluded that in virtually every case the Hopkins' statistic had the greatest power (i.e., the highest probability of accepting the alternative hypothesis when that hypothesis was true). This is obviously a desirable feature in a test statistic.

In these cases, as with the method of Ling and Killough, the alternative hypothesis is that the data are more clustered than a set of uniformly distributed random numbers. This conclusion seems to be rather weak. One would expect that virtually any measured or calculated data would be more clustered than uniformly distributed random numbers. After Hopkins' statistic is described in more detail and applied to chemical data, a modification of the null hypothesis will be examined that may allow researchers to draw more meaningful conclusions from these tests.

HOPKINS' STATISTIC

Hopkins' statistic is a simple and intuitively appealing measure of clustering proposed in basic form by the botanist Hopkins in 1954.[11] This statistic is based on the difference between the distance from a real point to its nearest neighbor, U, and the distance from a randomly chosen point within the data space to the nearest real data point, W. Figure 1 shows these relationships. To calculate the Hopkins statistic, the first step is to mark a small percentage of the real data points. In Figure 1 the four circled squares represent marked data points. Dubes and Jain suggest choosing 5% of the data points so nearest-neighbor distances will be independent and thus approximate a Beta distribution.[4] The distance from each marked point to its nearest neighbor is calculated, W_i. Next, the artifical points are spread uniformly throughout the data space. These are the open circles in Figure 1. The distance from each artifical point to the nearest real data point is calculated, U_i. These distances are inserted into

$$\mathbf{H} = \sum U_i/(\sum U_i + \sum W_i)$$

0095-2338/90/1630-0036$02.50/0

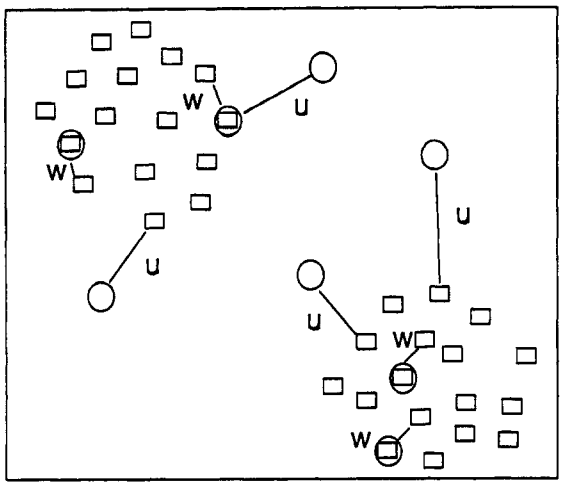

Figure 1. Square symbols are actual data points. Circled square symbols are marked data points. Open circles are random locations assigned as artificial points. *W* represents the distance from a real data point to its nearest neighbor. *U* represents the distance from an artificial data point to the nearest real data point.

where **H** is the value of Hopkins' statistic, U_i represents a distance from an artificial point to a real data point, and W_i represents a distance from a real data point to its nearest neighbor. If the real data contain little structure, then the distance from one real point to another real point will be approximately the same, on average, as the distance from a uniformly distributed random point to one of the real points, so the value of Hopkins' statistic will be approximately 0.5. If the data are arranged in tight clusters, then the distances W_i will be very small relative to U_i, so the value of Hopkin's statistic will be approximately 1.0. Assuming that few enough points have been chosen so that the nearest-neighbor distances approximate a Beta distribution, it is possible to assign some degree of confidence to the rejection of the null hypothesis as a function of the value of *H*.

This method of determining clustering tendency was implemented within the ADAPT system to allow its application to chemical problems. ADAPT has been described elsewhere, but briefly it is an integrated set of FORTRAN 77 programs that runs on a Sun 4/110 Unix-based workstation that allows the user to sketch in compounds, calculate descriptors on the basis of their structure, and search for relationships within the data using multiple linear regression, pattern recognition, and cluster analysis.[13,14]

Before Hopkins' statistic was applied to real data sets, it was necessary to determine the number of sampling points required to provide reproducible results. If too few points are chosen, then the nearest-neighbor distances chosen will not be representative of the entire distribution of distances. If too many points are chosen, Dubes and Zeng warn[10] that the assumptions about the Beta distribution will be invalid.

To experiment with the number of sampling points, we generated somewhat idealized artificial data sets at the endpoints of the **H** values. One data set consisted of a set of uniformly distributed random variables. This served as the unclustered extreme which should lead to an **H** value of 0.50. Figure 2 shows a scatterplot of two of the eight dimensions of this data set. The second artificial data set consisted of three well-separated clusters with 50, 100, and 150 data points, respectively, arranged in a 10-dimensional space. Figure 3 shows a plot of the first two principal components of these data. This data set should lead to an **H** value close to 1.0. Since these two artifical data sets had known distributions, no outliers, little correlation, known structure and contained only interval data, we were able to isolate changes in Hopkins' statistic to the varying number of sampling points.

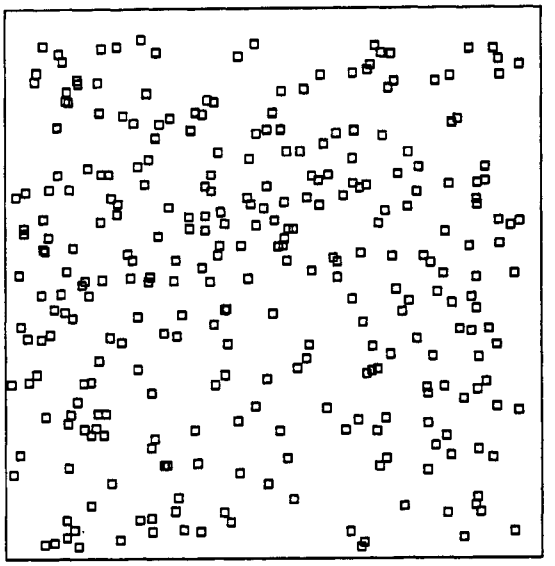

Figure 2. Two dimensions of the eight total with 300 uniformly distributed random numbers. This served as the unclustered extreme.

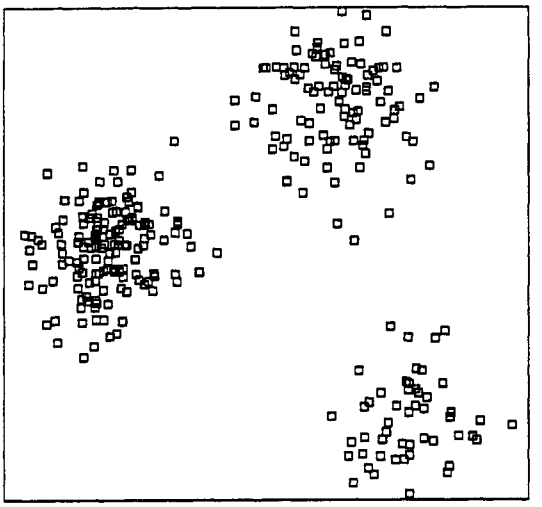

Figure 3. First two principal components of artificial 10-dimensional space with 300 points arranged in three clusters of 50, 100, and 150 points. This served as the clustered extreme.

Table I. **H** Values for Random Data Set, 300 Points in 8 Dimensions

trial	points	iterations	av	max	min	range
1	5	60	0.50	0.55	0.41	0.14
2	5	60	0.50	0.57	0.44	0.13
1	10	30	0.50	0.54	0.45	0.09
2	10	30	0.49	0.54	0.44	0.10
1	15	20	0.50	0.52	0.47	0.05
2	15	20	0.49	0.52	0.46	0.06
1	30	10	0.49	0.50	0.48	0.02
2	30	10	0.49	0.51	0.47	0.04

The first experiments were performed by using the data set consisting of uniformly distributed random variables. Initially, 300 data points were used for each of eight variables so that a relatively large number of points could be sampled without risking distance dependence. For the large random data set we selected 5 sample points, calculated Hopkins' statistic, selected 5 more points (with replacement), calculated another value, and so on for a total of 60 iterations so that theoretically every point could be chosen once. This procedure was repeated with 10 points for 30 iterations, 15 points for 20 iterations, and 30 points for 10 iterations. Each combination of sample points and iterations was repeated twice. The results of these experiments are shown in Table I.

Table II. H Values for Random Data Set, 143 Points in 8 Dimensions

trial	points	iterations	av	max	min	range
1	5	28	0.49	0.54	0.45	0.09
2	5	28	0.49	0.52	0.39	0.13
1	7	20	0.50	0.54	0.46	0.08
2	7	20	0.49	0.53	0.44	0.09
1	10	14	0.50	0.54	0.47	0.07
2	10	14	0.49	0.54	0.45	0.09
1	14	10	0.50	0.53	0.47	0.06
2	14	10	0.49	0.51	0.46	0.05
1	30	5	0.49	0.49	0.48	0.01
2	30	5	0.50	0.51	0.48	0.03

Hopkins' statistic averaged 0.5 for every number of sample points. Values of 0.5 give us no confidence in rejecting the null hypothesis that the data are no more clustered than uniformly distributed random numbers. Since the data *are* uniformly distributed random numbers, they can not be more clustered than uniformly distributed random numbers. The range of these **H** values decreased to a minimum when 5–10% of the data points were sampled. This is not surprising, since with more points being sampled the results should be more representative and less influenced by extreme values. It was possible to take reasonable sample sizes with this data set because there were so many total points. Thirty points is generally accepted as an adequate number to define normal distributions, and here since it only represented 10% of the data, there was little concern about distance independence.

The next set of experiments was done with 143 data points distributed uniformly along each of eight dimensions to duplicate the size constraints of a real data set. In this case there were samples of 5 points chosen 28 times, 7 points chosen 20 times, 10 points chosen 14 times, 14 points chosen 10 times, and 30 points chosen 5 times. Again, every combination of sample points and iterations was done twice. The results are shown in Table II.

These results are somewhat surprising in that the **H** values with only seven sample points (5% of the total) had a reasonably low range. All sample levels, even just five points, had average **H** values of 0.5. Once again, 30 sample points provided the minimum range, even though here 30 points represented roughly 20% of the total. It is possible that 20% sampling does not lead to loss of distance independence, but that may be true only for such a highly unstructured data set. More importantly, it appears that the number of iterations at each level of sampling averages out extreme values, so although low sample sizes had larger ranges, the average values were the same. Since these were ideal, unstructured data, these conclusions were tentative.

More structured data were investigated next. The data set at this extreme, as mentioned earlier, was arranged into three clusters with 50, 100, and 150 points. The points in each cluster had normal distributions with means and standard deviations specified along each of the 10 dimensions. These data are obviously not as simple as the unstructured data in the previous experiments, but the increased complexity is more likely to resemble data sets examined in practice. The experiments were performed just as before, with 5 points chosen 60 times, 10 points chosen 30 times, 15 points chosen 20 times, 30 points chosen 10 times, and 60 points chosen 5 times. Each combination was evaluated twice. The results are shown in Table III. Sixty points were examined in this case to see if 20% of the data set was too large a sample size for distance independence to be preserved. The **H** values at every sampling level are essentially the same. The range in **H** values decreases substantially going from 5 sample points to 10, but after that, the decrease in range is more modest until 60 sampling points are selected. Once again, it appears that repeating the distance

Table III. H Values for Three Artificial Clusters, 300 Points in 10 Dimensions

trial	points	iterations	av	max	min	range
1	5	60	0.77	0.84	0.70	0.14
2	5	60	0.77	0.82	0.72	0.10
1	10	30	0.77	0.82	0.75	0.07
2	10	30	0.78	0.82	0.75	0.07
1	15	20	0.77	0.80	0.74	0.06
2	15	20	0.78	0.81	0.76	0.05
1	30	10	0.78	0.80	0.76	0.04
2	30	10	0.77	0.79	0.75	0.04
1	60	5	0.77	0.78	0.76	0.02
2	60	5	0.78	0.78	0.77	0.01

Table IV. Set of Eight Structural Descriptors

no.	mean	SD	description
1	7.8	4.3	$^3\kappa$, topological index measuring midchain branching
2	2.2	1.4	sum of absolute values of all σ charges
3	3.6	2.2	$^2\chi^v$, path-2 valence-corrected molecular connectivity
4	0.6	0.8	count of acrylate moiety
5	320	370	number of molecular paths
6	30	47	path environment for acrylate
7	3.0	2.0	calculated log P
8	0.7	0.6	$^3\chi_C$, cluster-3 molecular connectivity index

sampling enough will average out extreme values. The magnitude of the values also indicates a strong tendency to cluster. An **H** value of 0.77 provides a good deal of confidence in rejecting the null hypothesis that the data are no more clustered than uniformly distributed random numbers. Although in general Hopkins' statistic can range from 0.5 for the unclustered extreme to 1.0 for the clustered extreme, the shape of the Beta distribution with this number of sample points is such that there is better than 90% confidence in rejecting H_0 for values of **H** greater than 0.75.

APPLICATION TO CHEMICAL DATA

The results with idealized data sets at the structured and unstructured extremes demonstrated the robustness of the average value of Hopkins' statistic with a wide range in the number of sampling points. The next step was to experiment with a chemical data set. The data in question were a set of 143 acrylate monomers selected from the Toxic Substances Control Act Inventory that were to be investigated for toxicity. The general approach was to cluster them into natural groups on the basis of physical and chemical properties to simplify the problem of sampling for expensive and time-consuming tests.[15] The premise here, as in work by Hodes[16] and Willett,[17] was that compounds sufficiently similar in chemical properties would be similar in biological activity as well. If representative elements of the different clusters selected for bioassays and in vivo studies had comparable toxicity levels, the results could possibly be extrapolated to other members of the clusters, thereby avoiding the need for exhaustive testing.

The first step was to determine if the eight calculated, structure-based descriptors, chosen from a larger pool on the basis of maximum variance, showed clustering tendency. These eight descriptors are listed in Table IV. No two-dimensional window on the data, not even the first two principal components, indicated any degreee of clustering. These descriptors comprise a data set that is far from ideal. None of the features follow defined distributions, some may have outliers, and some are not interval but are ordinal data. The descriptors were all autoscaled so that they had means of zero and standard deviations of unity.

The experimental procedure to assess clustering tendency was the same as for the artificial data sets. The sampling levels used were as follows: 5 points chosen 28 times, 7 points chosen 20 times, 10 points chosen 14 times, 14 points chosen 10 times, and 30 points chosen 5 times. Each experiment was done

Table V. H Values for Acrylate Data Set, 143 Points in 8 Dimensions

trial	points	iterations	av	max	min	range
1	5	28	0.80	0.97	0.47	0.50
2	5	28	0.82	0.98	0.59	0.39
1	7	20	0.78	0.96	0.55	0.41
2	7	20	0.83	0.96	0.57	0.39
1	10	14	0.83	0.93	0.68	0.25
2	10	14	0.83	0.94	0.65	0.29
1	14	10	0.77	0.87	0.64	0.23
2	14	10	0.83	0.92	0.66	0.26
1	30	5	0.84	0.90	0.71	0.19
2	30	5	0.82	0.92	0.77	0.15

Table VI. H Values for Two-Dimensional Data, 14 Points Sampled in 10 Iterations

trial	av	max	min	range
1	0.65	0.74	0.49	0.25
2	0.65	0.81	0.52	0.29
3	0.60	0.71	0.48	0.23
4	0.66	0.77	0.60	0.17
5	0.64	0.76	0.49	0.27
overall	0.64	0.76	0.52	0.24

twice. These conditions were chosen to test the tentative conclusions reached with ideal data, namely, that extreme **H** values will be averaged out with a sufficient number of iterations. The results are shown in Table V. The range of values here is much higher than with either of the ideal data sets; with 10 sample points or more the ranges are relatively constant below 0.3. Even with fewer points and wider ranges, the mean **H** values are fairly consistent at approximately 0.83. This value, even with some uncertainty, provides an extremely high degree of confidence in rejecting the null hypothesis that the data are no more clustered than uniformly distributed random numbers.

Because this **H** value was so high, especially compared to the ideal clustered data, another experiment was done to see if this value could be an artifact of the data. Outliers or noisy data can cause problems by expanding the sampling window. With the sampling window artifically expanded, the random points will be much more likely to be in regions with a few real data points, so the distance to their nearest neighbor will be further on average than for real data points.

OUTLIER RESULTS

The data set used to examine the effect of the outlier problem is shown in Figure 4. This data set is a combination of two descriptors each with 143 points, arranged in Gaussian distributions. There is only one relatively random group of points in these two dimensions, shown in the smaller sampling window. The larger sampling window is required to include all data points including the few points near the boundaries. Since uniformly distributed random points will be just as likely to fall in the sparsely populated regions as in the dense center, Hopkins' statistic can indicate more structure than is really present.

The data set was composed of 143 points, so Hopkins' static was calculated by using 14 points each for 10 iterations. This experiment was repeated five times to be sure of consistency. The results are shown in Table VI. The value of 0.64 is evidence that this data set is more likely to be clustered than uniformly distributed random numbers. In general, clustering tendency evaluations are performed to detect a tendency in the data set to form two or more clusters. In this artificial situation with only two dimensions, it is obvious from a visual inspection that there is just one cluster, which is a trivial result in clustering. Real data sets cannot be examined visually in general, so an **H** value of 0.64 could mislead researchers into concluding that more than one cluster is present.

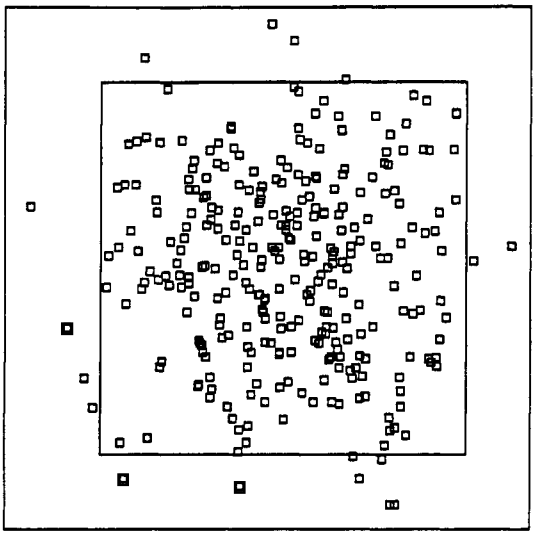

Figure 4. Two Gaussian variables with 143 points each. The inner sampling window displays the lack of overall structure. The larger sampling window is necessary to contain all points. Uniformly distributed random points could fall anywhere within the larger window.

The relatively high value for **H** in this single-cluster case is due to a comparison of the structure of this data set so uniformly distributed random numbers, i.e., data intended to be completely unstructured. With unstructured data as a comparison, virtually any structure in the test data would lead to rejection of the null hypothesis whether the data are clustered or not. In the simple example shown here, it is obvious that there is more structure in the data than there is in uniformly distributed random numbers, but any calculated or measured values should have more structure than uniform random numbers almost by definition. Instead of providing evidence that data show a tendency to form at least two clusters, the statistic merely provides evidence that the data have more structure than uniformly distributed random numbers.

MODIFIED NULL HYPOTHESIS

As discussed above, in the usual formulation of Hopkins' statistic, the null hypothesis is relatively weak. One solution to this problem is to change the null hypothesis. An alternative to comparing the existing data to uniformly distributed random numbers is to compare the data to random numbers that have distributions identical with those of the real descriptors. In this way descriptors with Gaussian distributions would not lead one to suspect clustering simply because the distribution has more structure than a uniform distribution. The same holds true for any other distribution, Poisson, Chi-square, or undefined (which would be expected in many real cases). Instead of looking for more structure than in uniform random numbers, one could look for more structure than would occur simply by chance for the data set, whatever its structure happens to be. This would eliminate the problem that any degree of structure in calculated or measured descriptors will lead to **H** values greater than 0.5. This would also eliminate the problem of mixing ordinal and interval data together. Ordinal data can be thought of as a set of numbers with a discontinuous distribution. Since the new null hypothesis takes distribution into account, it also takes data type into account implicitly.

Before the implementation of this new null hypothesis is described in detail, its value can be illustrated by applying it to the same data set that was erroneously considered to be clustered when the original null hypothesis was used. The data set is exactly the same, and the experiment is exactly the same.

Table VII. H Values for Two-Dimensional Data, 14 Points Sampled in 10 Iterations with Modified H_0

trial	av	max	min	range
1	0.47	0.61	0.31	0.30
2	0.52	0.63	0.40	0.23
3	0.49	0.61	0.42	0.19
4	0.43	0.60	0.28	0.32
5	0.52	0.62	0.44	0.18
overall	0.49	0.61	0.41	0.24

Table VIII. H Values for Three Artificial Clusters, 300 Points in 10 Dimensions Using Modified H_0

trial	points	iterations	av	max	min	range
1	5	60	0.71	0.80	0.61	0.19
2	5	60	0.72	0.80	0.59	0.21
1	10	30	0.72	0.77	0.66	0.11
2	10	30	0.72	0.76	0.66	0.10
1	15	20	0.72	0.75	0.69	0.06
2	15	20	0.71	0.74	0.67	0.07
1	30	10	0.71	0.74	0.66	0.08
2	30	10	0.72	0.74	0.68	0.06
1	60	5	0.72	0.72	0.71	0.01
2	60	5	0.71	0.73	0.70	0.03

Fourteen points were sampled 10 times, and the experiment was repeated 5 times. The results are shown in Table VII. With average **H** values of essentially 0.50 there is now no confidence in rejecting the null hypothesis that there is no more clustering in this data set than there is in identically distributed random numbers. In other words, the data are not clustered. This is the expected conclusion on the basis of a visual inspection of the data set.

Since this modified null hypothesis seems to lead to more meaningful conclusions than the original null hypothesis, it is worthwhile to examine how it is implemented. As mentioned earlier, instead of uniformly distributed random numbers, this null hypothesis depends on forming a set of numbers with distributions identical with those of the original descriptors. The simplest way to duplicate the distribution of a given descriptor is to randomly sample the descriptor values themselves. For instance, if a chemical data set was comprised of a set of compounds described by a molecular weight descriptor, a molecular volume descriptor, and a σ charge descriptor, the simplest way to obtain a data set with distributions identical with those of the original data set would be to take the descriptors one by one and scramble the order of their values. The individual distributions will obviously be the same since the numbers themselves are the same but are arranged in random order. If there had been significant structure in the original data set because of a chemical relationship between the properties, then that structure would be destroyed in the scrambled data sets. If the only structure in the data set came from the nature of the distributions of the individual descriptors, however, the scrambled data sets would have the same degree of structure as the original data sets.

In essence that is how the modified null hypothesis is used to calculate a new version of Hopkins' statistic. Calculating the distance from a selected number of real data points to their nearest neighbors is exactly the same as before. The second term in the equation for Hopkins' statistic is no longer calculated by measuring the distance from a random point to the nearest real point though. Here, the U_i terms are calculated in two steps, by first generating a pseudopoint by selecting a real descriptor value at random from each of the descriptors in turn and then measuring the distance from this point to the nearest real data point. In the example used above a pseudopoint could have the molecular weight of one molecule, the volume of another, and the σ charge of a third. Each of these values would come from one of the actual molecules but probably different molecules for each descriptor.

Table IX. H Values for Random Data Set, 143 Points in 8 Dimensions with Modified H_0

trial	points	iterations	av	max	min	range
1	5	28	0.49	0.56	0.43	0.13
2	5	28	0.49	0.53	0.43	0.10
1	7	20	0.48	0.52	0.46	0.06
2	7	20	0.49	0.55	0.45	0.10
1	10	14	0.49	0.53	0.45	0.08
2	10	14	0.49	0.52	0.47	0.05
1	14	10	0.48	0.51	0.45	0.06
2	14	10	0.49	0.52	0.46	0.06
1	30	5	0.49	0.51	0.46	0.05
2	30	5	0.50	0.51	0.49	0.02

Table X. H Values for Three Artificial Clusters, 300 Points in 10 Dimensions Using Modified H_0

trial	points	iterations	av	max	min	range
1	5	60	0.71	0.80	0.61	0.19
2	5	60	0.72	0.80	0.59	0.21
1	10	30	0.72	0.77	0.66	0.11
2	10	30	0.72	0.76	0.66	0.10
1	15	20	0.72	0.75	0.69	0.06
2	15	20	0.71	0.74	0.67	0.07
1	30	10	0.71	0.74	0.66	0.08
2	30	10	0.72	0.74	0.68	0.06
1	60	5	0.72	0.72	0.71	0.01
2	60	5	0.71	0.73	0.70	0.03

Table XI. H Values for Acrylate Data Set, 143 Points in 8 Dimensions with Modified H_0

trial	points	iterations	av	max	min	range
1	5	28	0.59	0.94	0.26	0.68
2	5	28	0.64	0.95	0.17	0.78
1	7	20	0.59	0.92	0.29	0.63
2	7	20	0.66	0.95	0.35	0.60
1	10	14	0.71	0.90	0.50	0.40
2	10	14	0.78	0.93	0.45	0.48
1	14	10	0.60	0.86	0.46	0.40
2	14	10	0.67	0.87	0.55	0.32
1	30	5	0.64	0.74	0.55	0.19
2	30	5	0.64	0.71	0.50	0.21

The application of Hopkins' statistic with the modified version of the null hypothesis to the single fuzzy cluster was already seen to be successful. The next test for this method was to check the artificially unclustered and clustered endpoints as before. First, the set of 300 points each consisting of 8 uniformly distributed random variables was tested. Here, as with the original null hypothesis, the average **H** value was 0.5 as seen in Table VIII. The maximum, minimum, and range values are also essentially the same as before. Next, the experiment was repeated by using the 143 point data set of 8 random variables, each designed to have the same size as the acrylate data. Table IX shows the results of this experiment. The average, maximum, and minimum values are all virtually the same as in the original experiment. For the extreme case when a data set has no structure, the version of the null hypothesis used makes no difference.

With the structured data, 300 points in 10 dimensions arranged in three clusters of 50, 100, and 150 points, there is a slight effect, as seen in Table X. From a comparison of these results to those from the earlier experiment, it is clear that the average **H** values have consistently dropped from approximately 0.77 to 0.72. While this difference seems to be repeatable, it is insignificant in terms of the conclusions drawn. For **H** values of 0.72 or 0.78 the conclusion is that there is strong clustering tendency. With the original value of H_0 this is relative to no structure, whereas with the new H_0 the value is relative to a data set with identical structure in the individual descriptors.

Since the conclusions about clustering at the endpoints were identical with the conclusions at the endpoints for the original null hypothesis, the clustering tendency of the acrylate data was also measured with the new version of the null hypothesis. Table XI shows the results obtained. Although the **H** values vary more for this experiment than for the others, both within and between trials with identical conditions, the overall **H** value of approximately 0.64 clearly indicates clustering. It is not clear why there is such a tremendous range of values even with 14 and 30 sampling points. Undoubtedly, this is somehow connected to the structure of the data, and it is conceivable that this information could somehow be extracted through a careful examination of the distribution of the U_i and W_i values for each of the iterations. In any case, the conclusions about clustering are more conservative in this experiment using the new null hypothesis, but they are also more meaningful.

CONCLUSIONS

Calculation of a quantitative measure of clustering tendency for a data set can provide valuable insight into the structure of the data. Once a data set has been shown to have a high clustering tendency, then it is justifiable to pursue actual cluster analysis to identify the memberships and sizes of the clusters.

Hopkins' statistic and a modified version of this statistic have been shown to be effective measures of clustering tendency. With both versions of the statistic, it has been demonstrated that a small number of sampling points (10%) can still be representative of the data set given a sufficient number of sampling repetitions.

It has been shown that the modified version of the null hypothesis solves several problems in the calculation of clustering tendency by Hopkins' statistic. Outliers are less of a problem with this method, because by definition they occur infrequently in the data set, so they are less likely to be selected as part of the comparison data space. Comparisons between nominal or ordinal data and interval or ratio data will no longer have to be made because ordinal values will be compared to other ordinals and interval data will be compared to other interval data. Finally, the weak conclusions about clustering relative to uniformly distributed random numbers have been replaced with conclusions about clustering relative to identically distributed random numbers. Now, instead of determining that there is some structure in the data rather than none, it is possible to compare the existing structure to the amount of structure due solely to the distribution of the individual descriptors. In other words, it is now possible to determine that an association between molecular weight, molecular volume, and σ charge within a set of compounds is more significant than a chance association. This would provide evidence that compounds are grouping together for chemical reasons, not merely because of mathematical artifacts.

This method can be extended beyond simply testing an existing set of descriptors. The identity of descriptors, the number of descriptors, and the form of scaling used could all be chosen in an optimum way for clustering by maximizing the clustering tendency, **H**. Since it is now possible to measure the level of significant structure rather than simply the raw structure, optimizing this value could be very valuable. Between the improvements to the existing method and the possible extensions to feature selection and optimization, the modified version of the null hypothesis for Hopkins' statistic should prove very useful in exploratory analysis in chemical systems.

ACKNOWLEDGMENT

The National Science Foundation supported this research through Grant CHE-8815785 and also provided partial support for the computing facilities used.

REFERENCES AND NOTES

(1) Glick, D.; Davis, A. *Org. Geochem.* **1987**, *11*, 331.
(2) Ismail, S.; Grass, F.; Varmuza, K. *J. Trace Microprobe Tech.* **1988**, *6*, 563.
(3) Jacobsen, T.; Kolset, K.; Vogt, N. *Mikrochim. Acta* **1986**, *2*, 125.
(4) Willett, P. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 78.
(5) Jain, A.; Dubes, R. *Algorithms For Clustering Data*; Prentice-Hall: Englewood Cliffs, NJ, 1988; pp 136–137.
(6) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Wiley: New York, 1987; pp 138–142.
(7) Ling, R. *Ann. Prob.* **1973**, *5*, 876.
(8) Ling, R.; Killough, G. *J. Am. Stat. Assoc.* **1976**, *71*, 293.
(9) Zeng, G.; Dubes, R. *Pattern Recognit.* **1985**, *2*, 191.
(10) Dubes, R.; Zeng, G. *J. Classif.* **1988**, *4*, 33.
(11) Hopkins, B. *Ann. Bot.* **1954**, *18*, 213.
(12) Jain, A.; Dubes, R. Algorithms for Clustering Data; Prentice-Hall: Englewood Cliffs, NJ, 1988; p 218.
(13) Stuper, A.; Bruger, W.; Jurs, P. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(14) ADAPT is a commercial software package licensed through Molecular Design Limited, San Leandro, CA.
(15) Lawson, R. G.; Jurs, P. C. Clustering Studies of Acrylate Compounds for Structure-Activity Relationships Investigations (submitted for publication in *J. Chem. Inf. Comput. Sci.*).
(16) Hodes, L. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 66.
(17) Willett, P.; Winterman, V.; Bawden, D. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109.