ELSEVIER

Analytica Chimica Acta 446 (2001) 59–70

ANALYTICA
CHIMICA
ACTA

www.elsevier.com/locate/aca

New index for clustering tendency

M. Forina[a,*], S. Lanteri[a], I. Esteban Díez[b]

[a] *Department of Chemistry and Technology of Drugs and Foods, University of Genova, Viale Brigata Salerno (s/n), I-16147 Genova, Italy*
[b] *Department of Chemistry, University of La Rioja, C/Madre de Dios 51, E-26006 Logroño, Spain*

Received 7 November 2000; received in revised form 23 February 2001; accepted 22 March 2001

Abstract

A new index for clustering tendency is described. The index is based on the frequency distribution of the lengths of the edges in the minimum spanning tree connecting the objects, compared with the probability distribution of the lengths of edges of the minimum spanning tree connecting the same number of objects described by variables extracted from the uniform distribution. The here suggested index shows some advantages when compared with the Hopkins original index and with its modification suggested by Fernández Pierna and Massart. It can be used both to detect clusters, to measure the degree of non-uniformity of a data set (as required in many cases of multivariate calibration and QSAR studies), and to detect outliers.

Keywords: Clustering; Clustering tendency; Minimum spanning tree; Multivariate calibration

1. Introduction

Both in multivariate calibration and in QSAR, the quality of the training set used in the calibration phase is of fundamental importance in that regard the quality and the performances of the regression model. It is generally accepted that the training set must have an uniform distribution, and for this reason a reasonable number of objects is extracted from the set of available objects by means of a design made with the use of space-filling algorithms, as the Kennard–Stone algorithm. The design is applied to the significant principal components of the predictors (X-block, frequently spectral variables in multivariate calibration, molecular descriptors in QSAR). When the value of the response(s) is available for all the available objects, the selection of a suitable training set can be made on the response(s). Sometimes the available objects cluster in groups, and in this case separate calibration models are almost always computed.

Clusters can be identified or by means of visualization techniques or by using one of the well-known techniques of hierarchical clustering.

A clustering index should be an indicator of the degree of non-uniformity of the distribution of the objects. It should be used both as an automatic warning or, better, as a quantitative measure of the quality of the data set, original or extracted.

Hopkins clustering index [1,2] represents the first approach to obtain an index with the above qualities.

In the case of a data set with N objects in the V-dimensional space, it is obtained by means of the following algorithm:

1. M objects are randomly selected among the N objects; the Euclidean distances of each of these objects, m, from its nearest neighbor (one, o, of

[*] Corresponding author. Tel.: +39-10-353-2630; fax: +39-10-353-2684.
E-mail address: forina@dictfa.unige.it (M. Forina).

0003-2670/01/$ – see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0003-2670(01)01033-9

the other $N-1$ real objects) is computed, as

$$d_m = \sqrt{\sum_{v=1}^{V}(x_{mv}-x_{ov})^2} \quad (1)$$

2. M artificial objects are generated. The Euclidean distance of each of these artificial objects, a, from its real nearest neighbor (one, o, of the N real objects) is computed as

$$D_a = \sqrt{\sum_{v=1}^{V}(x_{av}-x_{ov})^2} \quad (2)$$

3. Hopkins index is computed as

$$H_i = \frac{\sum_{a=1}^{M} D_a}{\sum_{m=1}^{M} d_m + \sum_{a=1}^{M} D_a} \quad (3)$$

4. The procedure is repeated many times with different randomization seeds and the mean of the H_i is the final value of the index. The number of repetitions necessary to obtain a stable value of the index depends on M and N.

Each coordinate v of the artificial objects is extracted from random uniform distribution $U(0, R_v)$, where R_v is the range of the variable v in the real objects. So, with reference to Fig. 1, the artificial objects can fall in one of the subspaces indicated with A, B and C in the figure.

When, the real objects come also from an uniform distribution the sum of distances d and that of distances D must be about equal, so that, H must approach 0.5.

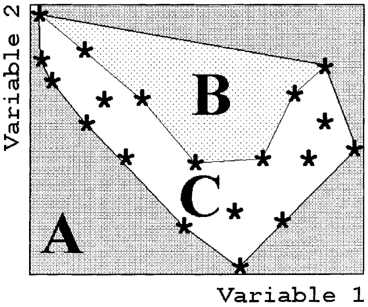

Fig. 1. Subspaces in Hopkins algorithm. A + B + C: total space; B + C: space within the dispersion polygon of Fernández Pierna and Massart; C: subjective evaluation of the multivariate boundary of the data set.

When on the contrary the N real objects are separated in clusters, the distances d tend to be small compared with distances D. In the extreme case, the N objects are separated in two or more groups of identical objects, so that all the distances d are null and H is 1.

The hypothesis that the N objects are homogeneous, extracted from a uniform distribution, is rejected with significance <10% when $H > 0.75$.

We suggested (as reported in reference [3]) a modification of Hopkins equation. The modified index is computed as

$$H^* = \lim_{M\to\infty} \frac{\left(\sum_M D_a/M\right) - \left(\sum_N d_n/N\right)}{\left(\sum_M D_a/M\right) + \left(\sum_N d_n/N\right)} \quad (4)$$

Here, d_n is computed for all the N objects (d_n is obtained from Eq. (1)). The subscript has been modified because now n is one of the N objects. M indicates only the number of artificial objects. The procedure is performed once, instead of many times as in point 4 of the original algorithm.

The modified index ranges from 0 (no clustering) to 1 (extreme clustering).

Fig. 2 shows as H^* is almost stable when $M > 20{,}000$ (few minutes of computer time with $N = 100$). When $M \to \infty$, it must be

$$H^* = 2H - 1; \qquad H = \tfrac{1}{2}(H^* + 1) \quad (5)$$

Afterwards we will use only the modified index H^* for the Hopkins parameter, and the symbol H will be used for a characteristic of simulated data.

Fernández Pierna and Massart [3] noticed that the extraction of artificial objects using the univariate ranges of the training set R_v "leads to selection of

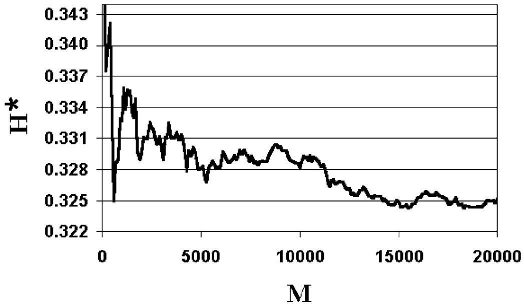

Fig. 2. Modified Hopkins index H^* as a function of the number M of artificial objects. Data set H10G00.

points outside the multivariate limits of experimental points". So, they draw a boundary around the experimental objects, in the form of a polygon, and the extraction of artificial objects used to compute the modified Hopkins statistics is limited to points within the polygon. With reference to Fig. 1, the modification of Fernández Pierna–Massart eliminates the subspace A, but the artificial objects can fall within one of the subspaces B and C. The reason is that it is not possible to develop a satisfactory algorithm to draw a convex polygon, as the boundary of subspace C in Fig. 1. So, the improvement of Fernández Pierna–Massart reduces only partially the original problem, of the artificial objects outside the multivariate boundary of the experimental objects. Moreover, the polygon can be drawn only in a bidimensional space. The clustering study is usually made in the inner space, that of the significant principal components. The outer space, the space of the noise, is almost empty and must be not considered in the computation of the clustering index, as in the case of subspace A in the bidimensional example of Fig. 1. The inner space can have more than two significant components, and its study by means of all the combinations of two components seems complicated and with results difficult to be interpreted.

For these reasons we developed a completely different approach to the problem of finding a suitable clustering index. This index is based on the distribution of the lengths of the edges in the minimum spanning tree connecting the objects in the training set, generally considered in the space of the significant components.

The first results seem to indicate that the suggested index has some advantages.

2. Theory

A graph is a set of vertices and edges, which connect them. In our case the vertices are the objects. A path or tree p through a graph is a sequence of connected objects: $p = \langle o_0, o_1, \dots, o_k \rangle$. The length of a path is the number k of edges. A graph contains no cycles if there is no path of non-zero length through the graph, $p = \langle o_0, o_1, \dots, o_k \rangle$ such that $o_0 = o_k$. A spanning tree of a graph, G, is a set of $N - 1$ edges that connect all the N objects of the graph.

If a cost, c_{ij}, is associated with each edge, $e_{ij} = (o_i, o_j)$, then the minimum spanning tree (MST) is the set of edges such that the sum of the costs over the $N - 1$ edges is a minimum.

In our case the cost associated with each edge is the Euclidean distance between the two connected objects.

Two algorithms are usually used to find the minimum spanning tree connecting N objects. The Prim [4] algorithm begins from whatever object, that constitutes a zero-length tree, with only one object connected. Then in each step of the algorithm an object is connected to the tree. The object is the object (previously non-connected) with the minimum distance from one of the connected objects.

The Kruskal [5] algorithm begins with the connection of the two nearest objects. In each step the two nearest objects are connected, provided that they are not in the same tree (in this case a cycle would be formed). When both the two objects are not connected they constituted a new path with non-zero length. So a number of separated trees, a forest, can be formed. When the two connected objects are in different trees, the two trees merge. The algorithm continues until the complete link in a unique tree.

In this paper the two algorithms were used, with identical results. Because of the nature of the variables (continuous) describing the objects, the MST solution was always unique. So it is unique also the frequency distribution of the distances between the connected objects (real sample distribution). Large distances are probably related with the existence of clusters, and consequently also of singletons, outliers.

In the case of a sample of N objects described by V variables (with range R_v), extracted from a uniform distribution $U(0, R_v)$, MST distances have some variability and the associated frequency distribution of the distances. When the extraction of the sample is repeated many times the frequency distribution approximates the probability distribution of the edge's length. From this approximation of the probability distribution it is possible to compute confidence limits at a selected probability level.

The test based on the MST distances distribution considers the largest distance in the tree obtained with the N objects in the training set. The null hypothesis H_0, that the objects are a sample extracted from a uniform distribution, is rejected when the significance level (unilateral right) of this largest distance is less than a selected critical value.

In the definition of the MST clustering index we used as a critical significance value 5%. Since, in the study of the real set of N objects there are $N-1$ edges, the critical value of the distance d_{crit} obtained from the probability distribution of the distances must correspond to the $5/(N-1)$% right significance of the probability distribution of the distances.

Alternatively, it is possible to repeat many times the extraction of a sample of N objects described by V variables from an uniform distribution $(0, R_v)$, and each time to consider only the maximum distance in MST to obtain an estimate of the probability distribution of the maximum distances. The 5% right critical value of this distribution is equal to the $5/(N-1)$% value of the distribution of the distances.

The MST clustering index is defined as

$$\mathrm{Index}_{\mathrm{MST}} = \sum_{d>d_{\mathrm{crit}}} \left(\frac{d}{d_{\mathrm{crit}}} - 1\right) \tag{6}$$

When, the maximum tree distance is equal to the critical value, the index is zero: the null hypothesis is accepted. Non-null values of the index indicate that the data set is not homogeneous. In the case of extreme clustering the maximum distance in the tree is the maximum possible distance in the V-dimensional space, D_V. It would be possible to take into account this case to modify Eq. (6) to obtain an index with range 0–1

$$\mathrm{Index}_{\mathrm{MST}}^{\mathrm{normalized}} = \sum_{d>d_{\mathrm{crit}}} \left(\frac{(d/d_{\mathrm{crit}}) - 1}{(D_V/d_{\mathrm{crit}}) - 1}\right) \tag{7}$$

However, we think that extreme clustering is a very rare case, corresponding to a heavy outlier, easily identifiable and eliminable. The use of Eq. (7) would give very small values in the case of real non-uniform data, with an instinctive underestimate of the degree of clustering.

On the contrary we prefer add to the above index two auxiliary parameters

$$\mathrm{Index}_{\mathrm{MST}}^{98} = \sum_{d>d_{98}} \left(\frac{d}{d_{98}} - 1\right) \tag{8}$$

$$\mathrm{Index}_{\mathrm{MST}}^{95} = \sum_{d>d_{95}} \left(\frac{d}{d_{95}} - 1\right) \tag{9}$$

The two distances d_{98} and d_{95} are respectively the 98 and the 95% values of the probability distribution of the distances.

The MST index considers only the largest distance in the tree of the real data set. It aims to detect a special type of non-uniformity, a large empty space between spaces where some objects (the clusters) are distributed. The lack of uniformity within these clusters does not affect the index, provided that the interpoint distances are smaller than the critical value. We will call this non-uniformity “first-type clustering”. A second-type of clustering is present, when there are parts of the space with high density of points, and part with low density, without a sharp boundary.

The comparison between the real sample distribution and the probability distribution of the distances can be more detailed.

For example, it is possible to compute the discrepancy between the two distributions.

The discrepancy has been largely used in the tests of goodness-of-fit, e.g. in the normality test. The test based on the discrepancy is not very efficient, and at present it has been substituted by tests of the family of Kolmogorov–Smirnov test. However, in our case the discrepancy can be used to study the distribution of the objects in the data set in some details.

The discrepancy is defined by

$$D = \sum_{a=1}^{A} \frac{(f_a - e_a)^2}{e_a} \tag{10}$$

The interval of the distances in the probability distribution, from 0 to ∞, is divided in A parts, with about the same value of the probability, p_a. The expected frequency in the interval a is $e_a = p_a N$. Here f_a is the frequency in the same interval, obtained from the real sample distribution.

The terms of the discrepancy are related with the binomial distribution. When, the expected frequency in each interval is ≥ 10, so that the binomial distribution can be approximated by the normal distribution, the discrepancy tends to the sum of $A-1$ independent Z^2 variables, i.e. to a function χ^2.

The study of the terms of the discrepancy can be of great interest, to evaluate the presence of local abnormal densities of objects, as in second-type clustering.

In this paper we present a very limited study of the discrepancy, and all the reported examples are based on the discrepancy measured with $A = 5$ intervals, and on the values of the ratios between frequency and expected frequency

$$r_a = \frac{f_a}{e_a} \tag{11}$$

3. Data

Several sets of artificial two-dimensional data and some real data sets were used to evaluate MST clustering index.

All artificial data sets were created to have objects only in the subspaces B and C (see Fig. 1), since, the effect of subspace A can be eliminated by using the procedure of Fernández Pierna–Massart.

Some *U*-type data sets are shown in Figs. 3 and 4.

The height *H* of *U* sides ranges from 0 (Fig. 3) to 10. Experimental points in the *U* basis and in both *U* sides are drawn from a uniform distribution. In *U*-types data with first-type clustering a gap *G* (as in the example in Fig. 4) divides the objects in two clusters. *G* ranges from 10 to 40. In *U*-type data the dispersion polygon of Fernández Pierna–Massart approximately coincides with the total space, so that it can not take into account the convexity of the real multivariate boundary. The B-space shown in Fig. 1 can be very large.

The details in Fig. 5 show that for a family of data sets with the same nominal gap the real separation (because of the random generation) can be different. Fig. 5 refers to *U*-files with $G = 10$: the real gap ranges from 10.5 to 14.8.

U-type data are indicated with names as H20G10, where the value of the height *H* and of the gap *G* is reported.

Fig. 6 shows an example of a second-type of simulated data (*V*-type data), where there is no B-space. The data in Fig. 6 are indicated as V2G40 (2 variables, gap 40).

In *V*-type data, variable 1 is responsible of clustering (first-type); the distribution of the other variables is uniform. The number of objects is 200.

Finally, Fig. 7 shows an example (data set N2) of second-type clustering. Sixty-four objects are placed at the nodes of an 8 × 8 grid (with a small amount of noise added), and 16 objects come from a bivariate normal distribution with center in point 25, 50; the other 16 objects come from a bivariate normal distribution with center in point 75, 50.

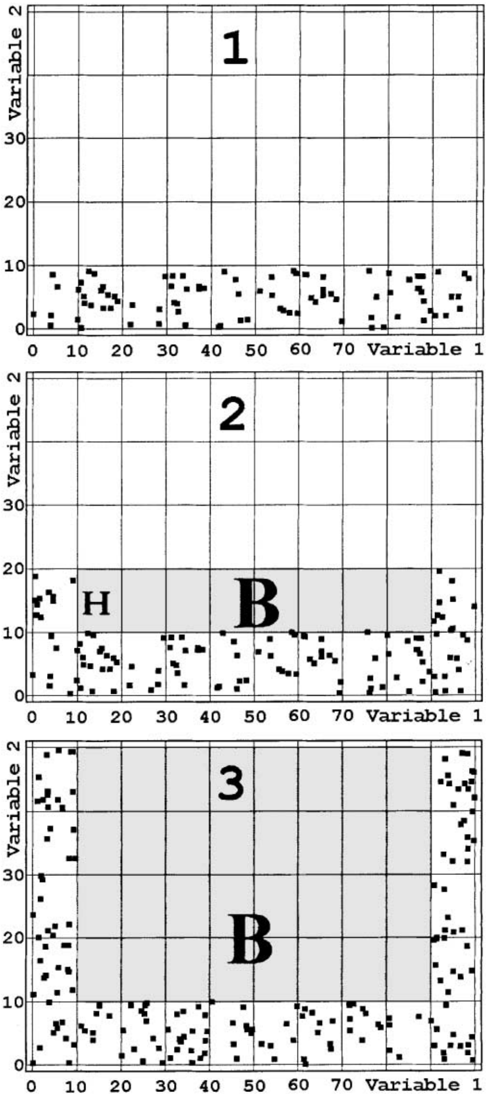

Fig. 3. Variable-by-variable plot of some *U*-type data sets. (1) H00G00; (2) H10G00; (3) H40G00. B: empty space considered in the dispersion polygon (see Fig. 1).

IRIS is the data set (150 objects, 4 variables, and 3 categories, three varieties of flowers) used by Fisher [6] in the presentation of linear discriminant analysis. WINES [7,8] is a data set of 178 objects (wine samples) described by 27 variables. Objects are divided in three categories, three red WINES of Piedmont,

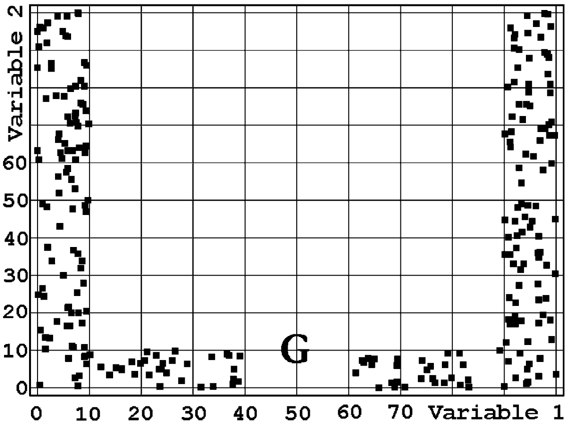

Fig. 4. Variable-by-variable plot of a *U*-type broken data set, H90G20.

Italy. KAL-Y (100 objects, 2 variables) and KAL-X (100 objects, 701 variables) indicate, respectively the Y-block (moisture and protein) and the X-block (spectra) of Kalivas data set [9] used also by Fernández Pierna and Massart.

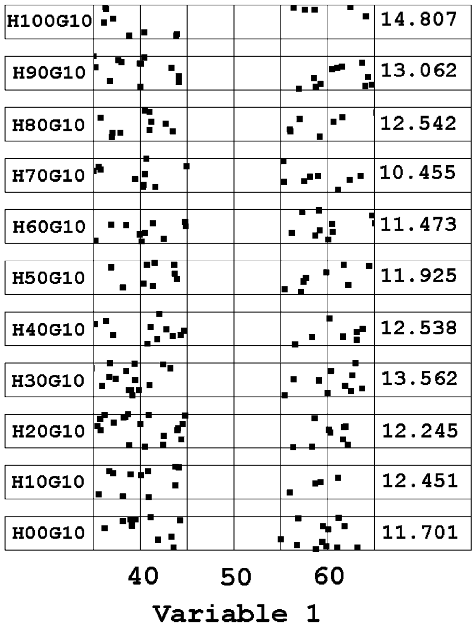

Fig. 5. Magnified plot (variable 1 from 35 to 65, variable 2 from 0 to 10) of the 11 *U*-type data sets with G10. On the right the true distance between the two clusters.

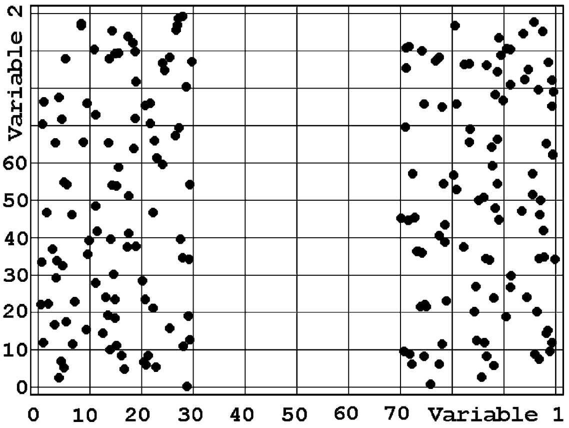

Fig. 6. Variable-by-variable plot of data set V2G40.

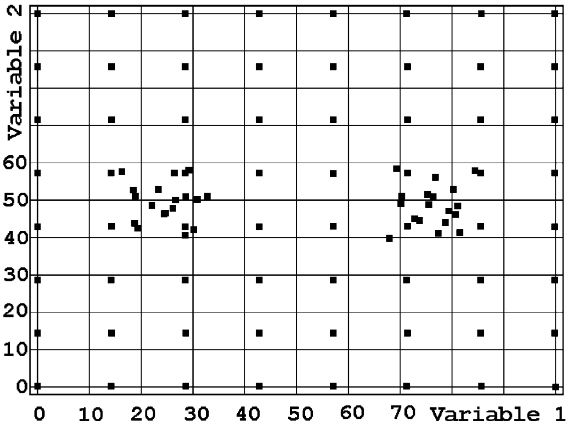

Fig. 7. Variable-by-variable plot of data set N2.

4. Results and discussion

Some results are reported in Tables 1–3.

Table 1 refers to *U*-type data sets. No pretreatment as range scaling was used, to avoid the elimination of the characteristic of data from uniform distribution in the multivariate space of the data (subspace C of Fig. 1).

In the case of data sets of Fig. 1 the generation of artificial objects used to evaluate the probability distribution of MST distances was repeated with the constraint that the artificial points are in the subspace C (see Fig. 1) with the form of *U* where the “experimental” points were generated.

Table 1
Results with U-type data sets. Ratio (column 5) is the ratio between the mean MST distance between random objects and the mean MST distance between real objects

	H	G	H^*	Ratio	Total space				Only subspace C			
					Critical distance	MST index	MST 98%	MST 95%	Critical distance	MST index	MST 98%	MST 95%
1	00	00	0.003	0.985	8.897	0.000	0.298	0.877	8.682	0.000	0.273	0.876
2	10	00	0.327	1.233	9.125	0.000	0.099	0.248	8.731	0.000	0.280	1.117
3	20	00	0.592	1.416	10.540	0.000	0.000	0.000	8.288	0.000	0.084	0.745
4	30	00	0.653	1.459	10.528	0.000	0.000	0.000	8.571	0.000	0.137	0.513
5	40	00	0.712	1.577	11.203	0.000	0.000	0.010	8.948	0.000	0.540	1.336
6	50	00	0.739	1.613	11.368	0.000	0.125	0.284	9.832	0.000	1.586	3.387
	50[a]	00	0.739	1.615	11.637	0.000	0.132	0.293	8.546	0.002	1.634	3.278
	50[b]	00	0.737	1.654	11.698	0.000	0.000	0.000	9.206	0.000	0.161	0.846
	50[b]	00	0.751	1.697	11.857	0.000	0.000	0.123	9.009	0.000	0.997	2.449
7	60	00	0.758	1.637	11.734	0.000	0.000	0.000	8.499	0.000	0.445	2.200
8	70	00	0.751	1.662	12.130	0.000	0.360	0.518	8.962	0.211	1.485	2.570
9	80	00	0.774	1.709	13.017	0.000	0.000	0.000	8.572	0.000	0.556	2.119
10	90	00	0.779	1.782	13.396	0.000	0.187	0.326	9.436	0.016	2.026	3.891
11	100	00	0.787	1.780	12.418	0.000	0.000	0.000	10.047	0.000	1.136	2.664
12	00	10	0.015	0.959	8.564	0.366	1.086	1.898	8.479	0.380	1.131	1.947
13	10	10	0.362	1.243	9.234	0.348	1.123	1.531	8.965	0.389	1.592	2.309
14	20	10	0.560	1.413	9.234	0.326	0.760	0.964	8.940	0.370	1.241	2.086
15	30	10	0.651	1.465	7.412	0.241	0.830	1.115	8.452	0.605	1.795	3.140
16	40	10	0.705	1.605	11.039	0.136	0.657	0.857	9.009	0.392	1.663	2.484
17	50	10	0.730	1.569	11.603	0.028	0.533	0.713	9.230	0.292	1.497	2.678
18	60	10	0.757	1.660	12.658	0.000	0.453	0.610	10.048	0.142	1.178	2.577
19	70	10	0.755	1.710	12.922	0.000	0.290	0.559	9.226	0.133	1.675	3.296
20	80	10	0.777	1.708	13.164	0.000	0.517	0.711	9.118	0.375	1.715	3.199
21	90	10	0.784	1.814	12.956	0.008	0.700	1.031	9.148	0.475	3.170	5.203
22	100	10	0.790	1.809	12.528	0.182	0.775	0.964	9.621	0.539	1.991	3.608
23	00	20	0.158	0.937	9.006	1.380	2.680	3.416	9.759	1.196	2.644	3.386
24	10	20	0.360	1.158	9.443	1.270	2.192	2.672	9.003	1.381	2.702	3.499
25	20	20	0.603	1.379	10.284	1.372	2.611	3.209	8.833	1.762	3.938	5.535
26	30	20	0.663	1.500	12.080	0.761	1.791	2.133	9.228	1.305	2.729	3.727
27	40	20	0.717	1.568	11.369	0.798	1.615	1.923	9.013	1.268	2.669	3.681
28	50	20	0.738	1.593	11.533	1.155	2.125	2.481	9.067	1.742	3.408	4.644
29	100	20	0.788	1.788	12.834	0.800	1.712	2.006	9.864	1.342	3.474	4.727
30	200	20	0.799	1.891	14.274	0.555	1.500	1.786	11.204	0.982	2.913	4.518

[a] Same data set, different randomization for both H^* and MST.

[b] Different data set, with the same U structure, data originated by different randomization seed.

Moreover in Table 1, for H50G00, it is shown the effect of the repetitions on the same U-files with different randomization for both H^* and MST. The results obtained with other two data sets, with the same U structure H50G00 but with data originated by different randomization seed, are listed too.

Results in Table 2 refer to real data and to some other simulated data.

Fig. 8 shows the minimum spanning tree computed for data set V2G40. For the same data set, Fig. 9 shows the frequency distribution of the MST distances and the estimate of the probability distribution for the uniform parent population. Table 3 reports the difference between the frequency and the probability (the probability distribution was estimated two times, to show the uncertainty associated with the estimate).

Fig. 10 shows the difference between the probability distribution of the edge distances in the trees and of only the maximum tree distance. The 95% right critical values (obtained from the data used to draw the plot) were 38 for distances, 60 for maximum distances (corresponding to 99.90% critical value for distances).

Table 2
Results with some simulated data sets and with real data sets[a]

Data set	Data	Treatment	H^*	Ratio	MST index	r_1	Discrepancy
H00G00	Original 2 variables	None	0.003	0.985	0.000	0.91	3
H00G20	Original 2 variables	None	0.159	0.937	1.380	1.11	2
H10G00	Original 2 variables	None	0.327	1.233	0.000	1.66	19
H10G20	Original 2 variables	None	0.360	1.158	1.270	1.50	12
V2G00	Original 2 variables	Range scaling	−0.008	0.991	0.000	1.13	2
V2G10	Original 2 variables	Range scaling	0.020	1.005	0.000	1.16	4
V2G20	Original 2 variables	Range scaling	0.118	1.055	0.478	1.24	5
V2G40	Original 2 variables	Range scaling	0.400	1.190	1.496	1.39	28
V3G40	Original 2 variables	Range scaling	0.142	1.096	0.458	1.27	20
V4G40	Original 2 variables	Range scaling	0.071	1.070	0.072	1.20	37
V5G40	Original 2 variables	Range scaling	0.021	1.031	0.000	1.05	6
V10G40	Original 2 variables	Range scaling	−0.002	0.972	0.000	0.87	13
N2	Original 2 variables	None	−0.261	0.739	0.000	1.28	102
WINES	First 3 PCs	None	0.436	1.663	0.242	3.54	268
WINES	First 2 PCs	None	0.463	1.416	0.233	2.41	103
IRIS	Original 4 variables	Range scaling	0.662	2.493	0.000	4.77	540
IRIS	First 2 PCs	None	0.551	1.579	0.808	2.79	143
KAL-X	First 3 PCs	None	0.358	1.399	0.687	2.92	100
KAL-X	First 2 PCs	None	0.363	1.338	0.743	2.38	63
KAL-X	PCs 1 and 3	None	0.325	1.294	0.543	2.38	60
KAL-X	PCs 1 and 3	Range scaling	0.540	1.614	0.558	2.73	86
KAL-Y	Original 2 variables	Range scaling	0.524	1.538	0.556	2.68	82
KAL-Y	Original variables, 99 objects	Range scaling	0.534	1.540	0.570	2.54	74

[a] Ratio (column 5) is the ratio between the mean MST distance between random objects and the mean MST distance between real objects.

Table 3
Frequency distribution and parent probability distribution[a]

Index	From	To	Frequency	Probability	Difference	Probability	Difference
1	0	1	0.070	0.029	0.0412	0.030	0.0398
2	1	2	0.126	0.088	0.0376	0.086	0.0398
3	2	3	0.141	0.127	0.0140	0.126	0.0146
4	3	4	0.201	0.152	0.0488	0.153	0.0479
5	4	5	0.176	0.160	0.0162	0.158	0.0182
6	5	6	0.136	0.145	−0.0089	0.144	−0.0082
7	6	7	0.080	0.119	−0.0384	0.122	−0.0417
8	7	8	0.050	0.087	−0.0372	0.088	−0.0377
9	8	9	0.010	0.054	−0.0438	0.053	−0.0427
10	9	10	0.005	0.024	−0.0189	0.024	−0.0189
11	10	11	0.000	0.009	−0.0088	0.009	−0.0092
12	11	12	0.000	0.004	−0.0038	0.004	−0.0043
13	12	13	0.000	0.002	−0.0016	0.001	−0.0013
14	13	14	0.000	0.001	−0.0007	0.001	−0.0007
15	14	15	0.000	0.001	−0.0005	0.000	−0.0003
16	15	16	0.000	0.000	−0.0002	0.000	−0.0001
17	16	17	0.000	0.000	0.0000	0.000	−0.0001
18	17	18	0.000	0.000	0.0000	0.000	0.0000
19	18	19	0.000	0.000	−0.0000	0.000	−0.0000
⋮							
42	41	42	1.000	0.000	1.0000	0.000	1.0000

[a] Distances of the histogram classes in columns 2 and 3. Probability distribution estimated two times (columns 5–6 and 7–8) from frequency in 200 repetitions, i.e. about 40,000 distances. Data set V2G40.

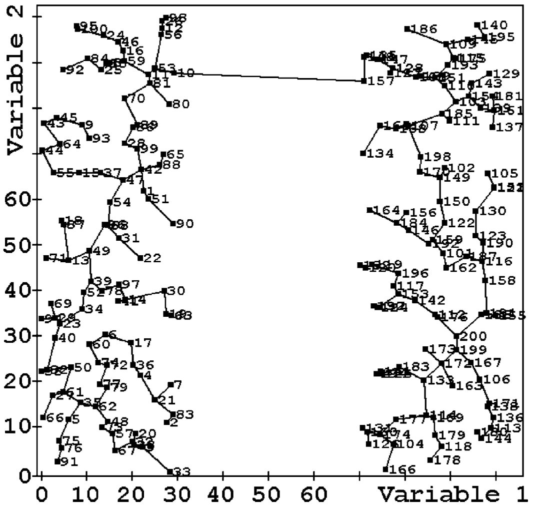

Fig. 8. Minimum spanning tree for data set V2G40.

4.1. *U*-type artificial data sets with uniform distribution

Results in the first 11 rows of Table 1 refer to *U*-type data with gap 0, and different *H*. The distribution of the objects in the C subspace (multivariate space of the data set) is uniform, so that an acceptable clustering index must be about 0. Because of the generation of the artificial objects in both subspaces C and B, Hopkins index cannot be considered a measure of the clustering tendency, but on the contrary a non-linear measure of the fraction of the space occupied by the data (C/(A + B + C) or C/(B + C)) with the correction of Fernández Pierna–Massart. MST index was always 0, when the objects were generated in the subspaces B + C. In one only case, the index was >0 with the objects generated only in subspace C, in agreement with the dispersion of the maximum tree distance and the 95% confidence level. The generation in the only subspace C (not possible with real data sets) was used to evaluate the effect of the empty space B on the MST index. The ratio between the mean tree distances with artificial and real objects increases very much from H00 to H100, from 1 about to 1.78, since, obviously, the artificial objects are dispersed in a wider space. However, the increase of the critical distance used in the evaluation of the MST index in Eq. (6) is not so large. When the artificial objects are generated in the only subspace C the ratio between the mean tree distances with artificial and real objects is always about 1. The critical distance is almost constant (within the large dispersion measured by the results with data sets

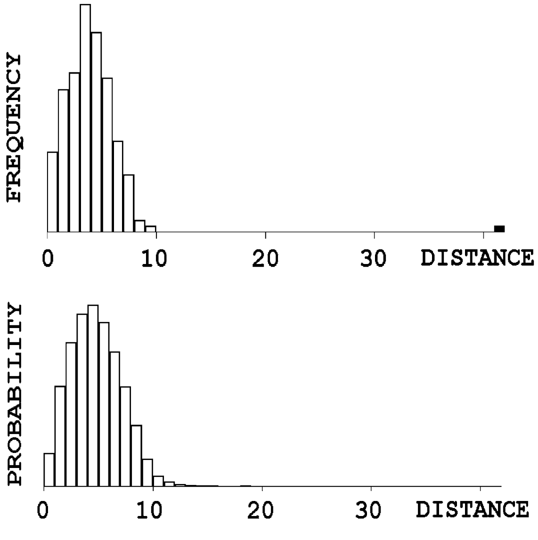

Fig. 9. Top: distribution of frequencies for the distances (lengths of the edges) in the minimum spanning tree of data set V2G40. Bottom: probability distribution of the distances in minimum spanning trees obtained in the two-dimensional space for 200 objects randomly drawn from uniform distribution (uncorrelated variables, $U(0, 100)$). Probability distribution estimated from frequency in 200 repetitions, i.e. about 40,000 distances.

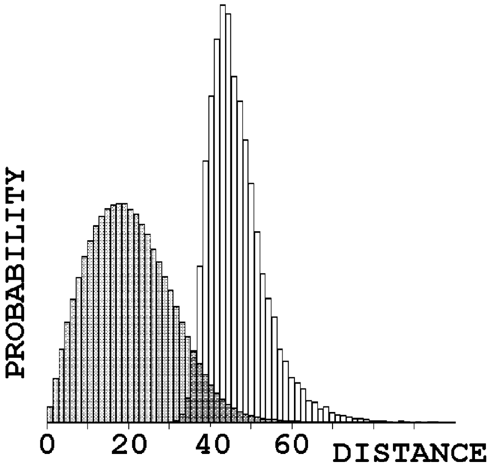

Fig. 10. Distributions of MST distances and maximum distances. About 50 objects, two-dimensional space. Range (0–100) scaling. Probability distributions estimated from frequency distributions in 20,000 repetitions (20,000 maximum distances, about 1 million distances).

H50G00), not less 0.7 times the critical distance measured with artificial objects generated in subspaces $\mathrm{B}+\mathrm{C}$.

MST 98% and MST 95% should be used as "warning" when the MST index is zero. Their direct use as a measure of clustering tendency would overestimate the tendency. In rows 1–11, column 9, of Table 1 the 95% MST index is significantly larger than 0 in 7 cases (about the 65%). So the evaluation of the critical distance d_{crit} by the $5/(N-1)\%$ right significance of the probability distribution of the distances, as described in the theoretical part, seems correct and necessary.

4.2. *U-type artificial data sets with clusters*

Results in rows 12–30 of Table 1 refer to *U*-type data sets with a small gap (10–20) between two clusters, as shown in Fig. 4.

H^* is not able to detect the break: its values depend almost only on the extension of the subspace B.

MST index is able to detect the presence of the two clusters with the small gap 10 when the empty subspace B is not too large. The largest gap 20 is detected also with very large subspace B, with increasing difficulty: the index decreases from 1.38 to 0.555 with H increasing from 0 to 200. Taking into account that the gap 20 cannot be considered as an unusual break between two real clusters, and that the extreme values of H correspond to extent of subspace B surely wider than those encountered in real problems, the performances of the index seems satisfactory.

The ideal, but unreal, cases where artificial points were generated only in the subspace C would always detect the presence of the clusters, and the index would be a good measure of the distance between the clusters, as shown by column 11 in Table 1, in spite of the large variability due to the random origin of the data sets (Fig. 5 shows the different distance between the two clusters in data sets with the same nominal gap).

4.3. *Second-type clustering and discrepancy*

Results in Table 2 were obtained without scaling in the case of the use of principal components and when (*U*-type data set) the scaling would destroy the uniform distribution in the subspace C.

In the case of artificial data *V*-type data (as that in Fig. 6) the ratio between the distances between artificial data and "real" data is always very close to 1, since the empty space is always caused by the separation between the two clusters. In the case of *U*-type data the ratio is not very large, because only the results with data sets with $H \leq 10$ are reported. So, the values of H^* and MST indexes are both a good measure of the separation between the clusters. When the number of variables in *V*-type data set increases, the distance between objects in the subspace C, where their distribution is uniform, increases too. Gradually, the separation between the clusters becomes less significant as compared to the mean distance between the objects, e.g. in the case of separation G40, when the number of variables increases from 2 to 3, 4, 5, 10 and H^* decreases from 0.40 to 0.142, 0.071, 0.021, -0.002 and MST index decreases from 1.496 to 0.458, 0.072, 0.000, 0.000.

In the case of data set V2G40 (Fig. 6) the clustering is determined by the value of the first variable; the maximum MST distance is 42, the mean MST distance is 4, the tree crosses only one time the space between the two clusters. In the case of data set V10G40 the distribution is uniform in the $V-1$ dimensions other than the first variable. The mean distance of the edges is 66, larger than the separation between the two clusters. The maximum distance, 91, was between two objects in the same cluster, and the space between the two "clusters" was crossed seven times by the minimum spanning tree. Consequently, the clusters are not detected because they do not exist in the *V*-dimensional space. Only marginal evaluation can easily detect that data are clustered on one of the variables.

In all these cases (*U*-type and *V*-type data sets, Table 2) the ratio r_1 between frequency and expected frequency in the first fifth part of the probability distribution is relatively small, 1–1.7, and the discrepancy is rarely more than the critical 95% value (about 10) of the Chi-square distribution with 4 degrees of freedom corresponding to the five intervals used to compute the discrepancy.

In the case of data set N2 shown in Fig. 7 the discrepancy is very large (Table 2). The value of H^* is negative, due to the fact that the "real" objects in N2 are generally placed artificially at the corner of a square, and artificial points fall within the square with distance from the "real" points less than the side

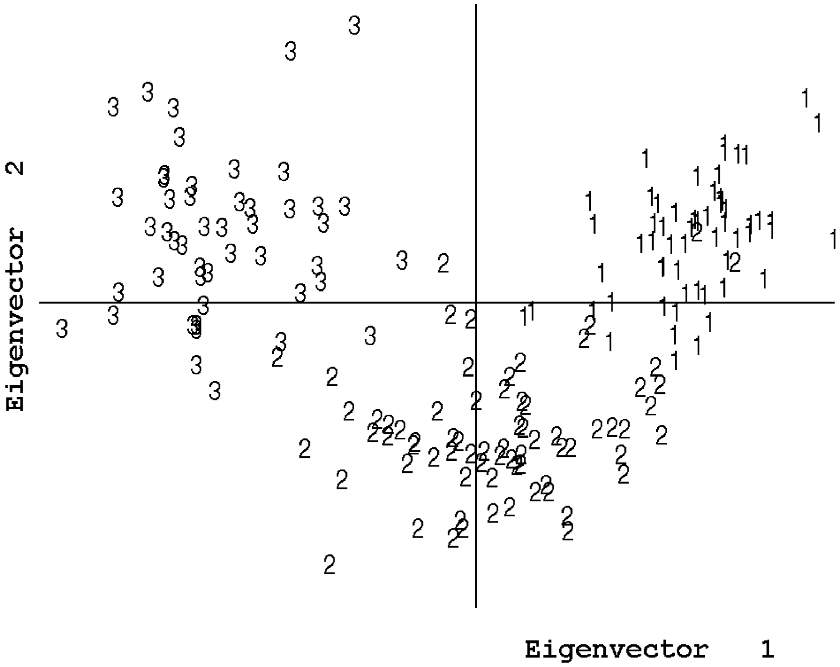

Fig. 11. Data set WINES — projection on the two first PCs (autoscaled data). Index of category reported: (1) Barolo; (2) Grignolino; (3) Barbera.

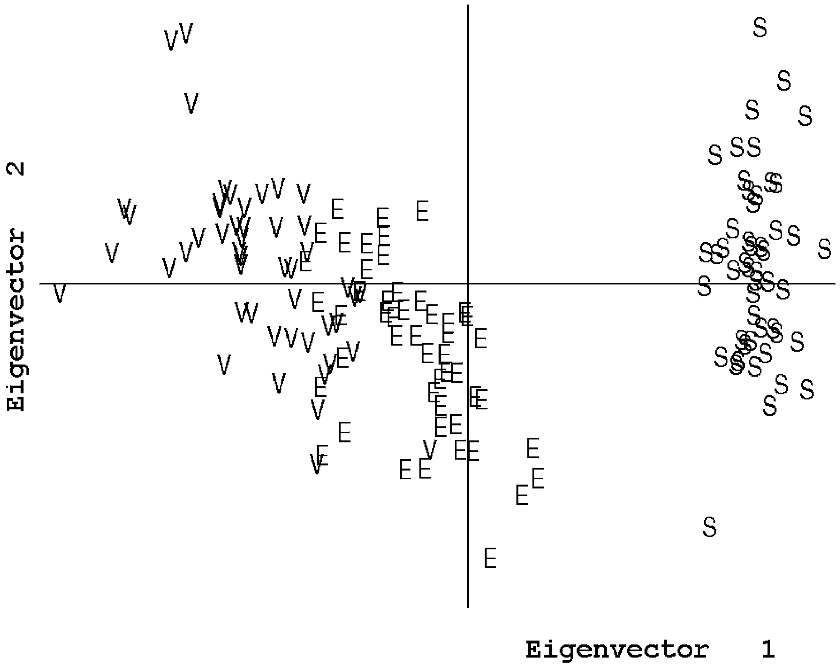

Fig. 12. Data set IRIS — projection on the two first PCs (autoscaled data). Categories: S: Iris setosa, E: Iris Versicolor, V: Iris Virginica.

of the square. For the same reason, the fifth term of the discrepancy is very large, and the discrepancy is much larger than the critical value of the Chi-square distribution. MST index is zero, because there is not a first-type clustering. Discrepancy detects the second-type clustering.

More detailed study of the difference between the frequency and the probability parent distributions, as that shown in Table 3, can help to understand the structure of data.

In the case of real data almost always the discrepancy is very large, due to the fact that the objects in the categories crowd round the center of the category spaces. MST index is not so large with the data set WINES, because the classes are not neatly separated (Fig. 11). In the other cases, the category spaces are separated by a significantly large empty space (MST large) and within the category spaces the distribution is not uniform, but shows the element of second-type clustering (large discrepancy). Only with the data set IRIS with all the 4 variables it seems that only the second-type clustering can be detected. Really, the three categories are separated by variable 3 (petal length). Variable 4 improves the separation. The other two variables have low separation power, so that the structure is similar to that of *V*-type data sets with many variables. When the first two principal components are used MST index is very large, because the first component uses the synergy of variables 3 and 4 to separate with a large break category 1 from the other two categories (Fig. 12).

5. Conclusions

MST index is a measure of clustering tendency with some advantages as compared with Hopkins clustering index. However, also MST index is sensitive (but less than H^*) to the empty space outside the multivariate limits of the cloud of experimental points. Moreover, both indexes are not able to identify second-type clustering, very important in many real cases. The discrepancy between the frequency distribution of the distances in the minimum spanning tree and the estimated probability distribution of the distances in the case of the uniform parent distribution seems offer a way to compute a "density" index, useful to characterize within the clustering index the quality of a data set.

Acknowledgements

Study developed with funds from the project UE STM4-CT98-7521 "European Network for the Intercomparison of Chemometric Software and Methods", from the University of Genova, and from CNR.

References

[1] B. Hopkins, Ann. Bot. 18 (1954) 213.
[2] R.G. Lawson, P. Jurs, J. Chem. Inf. Comput. Sci. 30 (1990) 137.
[3] J.A. Fernández Pierna, D.L. Massart, Anal. Chim. Acta 408 (2000) 13.
[4] R.C. Prim, Bell Syst. Technol. J. 36 (1957) 1389.
[5] J.B. Kruskal, Proc. Am. Math. Soc. 7 (1956) 48.
[6] R.A. Fisher, Ann. Eugen. Lond. 7 (1936) 179.
[7] M. Forina, S. Lanteri, Data analysis in food chemistry, in: B.R. Kowalski (Ed.), Chemometrics: Mathematics and Statistics in Chemistry, NATO ASI Series, Ser.C, Vol. 138, Reidel, Dordrecht, 1984, p. 439.
[8] M. Forina, C. Armanino, M. Castino, M. Ubigli, Vitis 25 (1986) 189.
[9] J. Kalivas, Chemom. Intell. Lab. Syst. 37 (1977) 255.