Computers & Geosciences 35 (2009) 1205–1213

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

GraphClus, a MATLAB program for cluster analysis using graph theory ☆

Clifford S. Todd [a,*], Tivadar M Toth [b], Róbert Busa-Fekete [c]

[a] Department of Geology & Geophysics, University of Hawaii, 1680 East-West Road, Honolulu, HI 96822, USA
[b] Department of Mineralogy, Geochemistry and Petrology, University of Szeged, P.O. Box 651, H-6721 Szeged, Hungary
[c] Research Group on Artificial Intelligence, University of Szeged, P.O. Box 652, H-6721 Szeged, Hungary

ARTICLE INFO

Article history:
Received 4 February 2008
Received in revised form
9 May 2008
Accepted 14 May 2008

Keywords:
Cluster analysis
Graph theory
Classification

ABSTRACT

Cluster analysis is used in numerous scientific disciplines. A method of cluster analysis based on graph theory is discussed and a MATLAB^{TM} code for its implementation is presented. The algorithm is based on the number of variables that are similar between samples. By changing the similarity criterion in a stepwise fashion, a hierarchical group structure develops, and can be displayed by a dendrogram. Three indexes describe the homogeneity of a given variable in a group, the heterogeneity of that variable between two groups, and the usefulness of that variable in distinguishing two groups. The algorithm is applied to both a synthetic dataset and a set of trace element analyses of lavas from Mount Etna in order to compare GraphClus to other cluster analysis algorithms.

1. Introduction

Results of classification procedures are generally believed to be better established if a larger number of variables are used. In these cases multivariate techniques, including cluster analysis, are used. Several cluster methods are available in software packages, using one of several classification algorithms based on various mathematical frameworks (e.g. Anderberg, 1973; Everitt et al., 2001; Gordon, 1999; Romesburg, 2004). The aim of this paper is to present a MATLAB^{TM} computer code for the cluster analysis algorithm introduced by M Tóth and Engi (1997). Two example applications of the method are also included.

The new algorithm presented here can be understood using the concept of graph theory. Graph theory is not commonly used in cluster analysis or in geosciences, although there are some examples (Botafogo, 1993; van Groenewoud and Ihm, 1974; Hartuv and Shamir, 2000; Pacheco, 1998; Pacheco and Van der Wejjden, 1996). At first sight a graph is simply a graphical representation of a connection network among points (samples). Any graph consists of an X set $(x_1, \ldots, x_n)$ of finite elements and a set U of element pairs of X. The basic types of graphs are directed or not directed depending on whether the (x_i, x_j) pairs are ordered or not. Graphs that are not directed can be represented by a symmetric matrix, whereas directed graphs cannot (one can obtain a representation of a directed graph using an asymmetric incidence matrix). Straightforward binary matrix representations of graphs exist. In the case of the A matrix, $A(x_i, x_j) = 1$ if x_i and x_j are connected, otherwise $A(x_i, x_j) = 0$. Matrix representation of a graph makes mathematically well-established evaluation of connections possible. In the case of the present algorithm (M Tóth and Engi, 1997) vertices of a graph represent samples (Fig. 1). They are joined with a tie-line if

☆ Code available from server at http://www.iamg.org/CGEditor/index.htm.

* Corresponding author. Present Address: The Dow Chemical Company, 1897 Building, E69, Midland, MI 48667, USA. Tel.: +1 989 636 0392; fax: +1 989 638 6443.
E-mail addresses: ctodd2@dow.com (C.S. Todd), mtoth@geo.u-szeged.hu (T. M Toth), busarobi@inf.u-szeged.hu (R. Busa-Fekete).

0098-3004/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2008.05.007

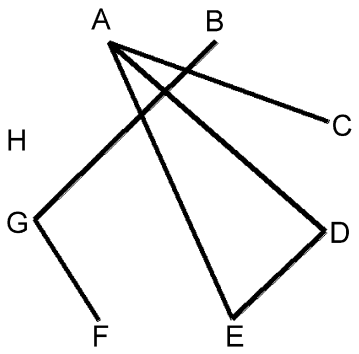

Fig. 1. Representation of graph structure. Letters represent samples. Lines show which samples are connected. Three groups are indicated: {ACDE}{BFG}{H}.

they are similar to each other in at least a certain number of variables (connection level). Two samples are defined as similar in a variable if their difference is smaller than a chosen threshold (the similarity level). A group is defined by collecting samples that have tie-lines to one another. By increasing the number of similar variables necessary to establish a tie-line (the connection level) in a stepwise manner a hierarchical group structure develops. In the case of the algorithm presented here, a path of connections joining a subset of samples defines a group. Therefore, not every sample needs to have a connection to every other sample in the group. A more stringent criterion for group definition is the equivalence class, where a sample must be connected to every other sample in the group (i.e., Pacheco, 1998). More background on graph theory can be found in textbooks (i.e., Diestel, 2005; Marshal, 1971; Trudeau, 1993; Wilson, 1996).

Cluster analysis is a group of mathematical methods for recognizing natural and meaningful groups within a set of samples. The importance of these techniques is that they can divide data into groups with similar characteristics without any *a priori* knowledge about the data under question. That is why this field of statistics is also called unsupervised classification. There are two basic approaches to construct a group structure of data, namely partitioning and hierarchical methods (Kaufman and Rousseeuw, 1990). Most widely used classification algorithms are based on the calculation of distances between pairs of samples that are used to rearrange samples into a hierarchical structure of clusters. To do so, various similarity (e.g. correlation coefficient) or dissimilarity functions (e.g. Euclidean distance) can be used to define distance (Everitt, 1980). Hierarchical methods are either agglomerative or divisive depending on whether they build from individual samples or start with the set of samples as a whole, respectively. Because of their recursive logic, agglomerative methods meet with difficulty when calculating the distance between groups. The results can differ depending on exactly how the group differences are defined, and it is difficult to know which method is appropriate for a given dataset. The family of partitioning clustering approaches (*k*-means, MacQueen, 1967; fuzzy *k*-means, Bezdek, 1974; among others) requires the specification of initial cluster seeds, which are extended by the procedure. Here *a priori* knowledge of the number of natural clusters is essential, and may be estimated by several potential algorithms (e.g. Pacheco, 1998). Graph theory clustering methods resolve this problem, because they do not need *a priori* knowledge of the number of clusters. The most widely used graph clustering methods are the Markov clustering process (MCP) (Van Dongen, 2000) and the CFinder algorithm (Palla et al., 2005). The MCP approach forms clusters in the dataset using random walks in the full weighted graph that represents the similarities among the objects to be classified. CFinder is also based on graph theory, and works especially well on networks, like protein interaction networks and various social networks.

The new algorithm presented here inherits several advantages from commonly used hierarchical cluster methods, and also eliminates some disadvantages. Some important points of the new approach are

- Although the method is neither agglomerative nor divisive, it builds a hierarchical group structure. Similar to other hierarchical cluster analyses, the result can be represented by a dendrogram.
- The method uses a dichotomized (indicator) dataset and so possesses the basic advantages of the divisive methods (Gill and Tipper, 1978).
- The algorithm is not a recursive one. It uses the original similarity matrix during the entire procedure, and so does not require recalculation of similarities between groups.
- There is a natural way to calculate indexes that measure the homogeneity within each group as well as the heterogeneity between groups, variable by variable.
- For different settings of similarity level, slightly different group structures form. By examining a series of these groupings one can separate samples that consistently belong to the same group from those that change their group. In this way not only the discrete "end member" groups may be determined, but also samples transitional between them.

2. Algorithm

If variable k is sufficiently similar between two samples, the samples may be from the same genetic group. For a given dataset, the matrix A_{ij} can be defined as $A_{ij} = \sum_k x_{kij}$. For variable k and samples i and j, $x_{kij} = 1$ if $|v_{ki} - v_{kj}| \leqslant Sig_k$, and $x_{kij} = 0$ if $|v_{ki} - v_{kj}| > Sig_k$, where v_{ki} is the value of variable k in sample i. Sig_k, the significance cutoff, represents the criterion used to establish whether the values of a variable in two samples are sufficiently close enough to indicate they may be from the same group. A_{ij} represents the number of variables that are within their significance cutoff between two samples, and can range from zero to the number of variables used to describe a sample. In order to be meaningful, the value of Sig_k must be a function of the variability of k in the dataset. Sig_k can be defined in many ways (for example, Pacheco, 1998). The one put forth in Tóth and Engi (1997) is $Sig_k = S\sigma_k/n$, where σ_k is the standard deviation of

variable k in the dataset, n is the number of samples in the dataset, and S is a new parameter called the similarity level. Similarity level can vary from one to a large number in order to examine different group structures under more stringent or more lenient significance cutoff criteria, respectively. In cases where the univariant distributions of the variables are far from normal, standard deviation in the definition of Sig_k can easily be changed to a more robust, non-parametric indicator of dispersion, for example the interquartile range.

Samples i and j are deemed to be in the same group if $A_{ij} \geqslant C$, where C is a new parameter called the connection level. On a graph, it is considered that i and j are connected, and share the same group. The results are completely symmetric. If i is connected to j, then j is connected to i. Connection level can be set between one and the number of variables used to define the dataset in order to examine different group structures under more stringent or more lenient connection criteria. For a given choice of connection and similarity levels, a network of connections between samples is developed. Groups are formed by choosing a sample and collecting all other samples that are connected to it, and all others connected to these new samples, iteratively until no new samples join the group. It is noteworthy that the variables used to define group formation may differ for each pair of samples in a group. Also, a sample may not be connected to every other sample in the group; it need only be connected to one other in order to become part of that group.

At low connection level, not many variables need to be similar for two samples to be connected. Therefore groups have many members. At a higher Connection Level, samples must share many variables to be connected. Therefore, large groups split into smaller and smaller subgroups as connection level increases. For a given choice of similarity level, the group structure as a function of connection level can be displayed graphically in a dendrogram. The group structure is robust, and not dependent on the order in which samples or variables are listed in the data file. However, the resulting group structure is not transitive; if the pair of samples 1 and 2 meets the criteria for S and C as well as the pair 2 and 3, it is not guaranteed that samples 1 and 3 meet the S and C criteria even though they will be in the same group.

The fact that a different group structure is produced for different choices of similarity level lends flexibility to the method and allows the user to investigate aspects of group formation. One can choose a low similarity level to force groups to be made of very similar samples, or a high S can be chosen to allow groups to be more inclusive. Samples that change from one group to another at different similarity levels are likely to be transitional between the two groups. Identification in this way can be useful. An additional strength of this algorithm is that the criterion for similarity depends on all variables taken individually.

Once groups are identified, several parameters can be calculated that give information about the characteristics of a particular group and the differences between groups. If group p has n_p members, then the homogeneity index of variable k in group p is defined as $I_{kp}^{\hom} = (2\sum_{i=1}^{n}\sum_{j=1}^{n} x_{kij}/n_p(n_p-1))$ for both i and j members of group p, and $i \neq j$. This number represents the fraction of possible sample pairs within the group for which variable k surpasses the significance cutoff criterion, since there are $n_p(n_p-1)/2$ different pairs of samples in the group. A homogeneity index close to one indicates that variable k is similar in many samples in the group, and therefore may be characteristic of that group. Variable k is said to be homogeneous within group p. If $I_{kp}^{\hom}$ is close to zero, it indicates that variable k is not similar between most pairs of samples in the group. It is said that variable k is not homogeneous within group p.

It is also instructive to examine the role of different variables in defining differences between groups. The heterogeneity index of variable k between groups p and q is defined as $I_{kpq}^{\text{het}} = 1 - (\sum_{i=1}^{n}\sum_{j=1}^{n} x_{kij}/n_p n_q)$, for i member of group p and j member of group q. This number represents the fraction of possible sample pairs between the two different groups for which variable k does not surpass the significance cutoff criterion. A heterogeneity index close to one indicates that variable k is not similar between most pairs of samples across the two groups and therefore may be useful in characterizing the difference between groups p and q. Variable k is said to be heterogeneous between the groups. If I_{kpq}^{het} is close to zero, it indicates that variable k is similar between most pairs of samples across the two groups, and therefore is not useful in characterizing the difference between groups p and q. Variable k is said to be not heterogeneous between the groups.

A single number that combines information from I^{hom} and I^{het} between two groups is the discriminant index, defined as $I_{kpq}^{disc} = \sqrt[3]{I_{kp}^{hom} I_{kq}^{hom} I_{kpq}^{het}}$. A discriminant index close to one indicates that variable k is homogeneous within each group p and q and heterogeneous between the two groups. This indicates that k may be useful in distinguishing members of group p from group q.

3. GraphClus

The algorithm discussed above was implemented in MATLAB™, a widely used math and statistics package. To run the developed standalone program it is necessary to install MATLAB Component Runtime, which is freely available for most computer operating systems. The GraphClus program itself is comprised of three modules: the visualization module, clustering module, and dendrogram visualization module. The visualization module allows the user to visualize their dataset by various dimension reduction methods. Dimension reduction methods are borrowed from the field of data mining, like multi dimensional scaling, locally linear embedding and principal component analysis. In this way the user can better understand the spatial structure of their database. The second module performs the clustering based on the graph theory approach presented in this paper. The only adjustable parameter of the algorithm is the similarity level, which is set in an edit box. Because the algorithm is computed in polynomial time, this module works well on

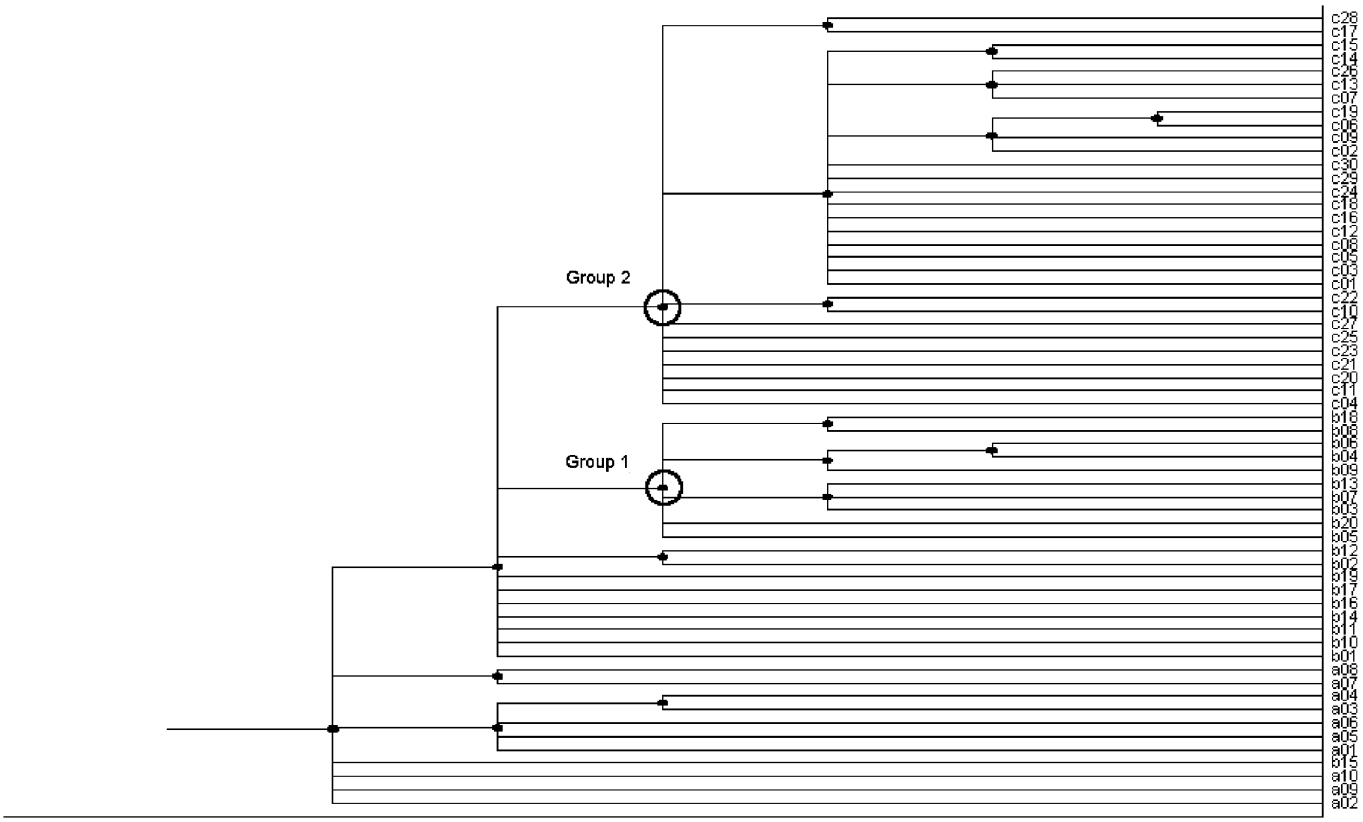

Fig. 2. Dendrogram for synthetic dataset (Table 2) using GraphClus, connection level 12. Homogeneity, heterogeneity and discriminant indexes for labeled groups are in Table 3. Consistency indexes for this dendrogram are Group A, 0.767; Group B, 0.850; Group C, 1.000.

large datasets as well as small ones. The output data of this module is then processed by a dendrogram visualization engine, which makes finding connections in the dataset easier. A dendrogram is a diagram that represents the relationships between samples. In dendrograms such as Fig. 2 each sample is listed along the right side of the diagram. The *X*-axis of the diagram represents the degree of similarity with other samples, decreasing to the left. When the lines for two or more samples join together, it means that the samples are sufficiently alike to be in the same group at that degree of similarity (connection level, in the case of GraphClus). The dendrogram visualization module has two functions: to visualize the dendrogram of the database and to calculate the mean and variance values of the variables and the indexes introduced above (homogeneity, heterogeneity and the discriminant indexes). The computer code can be downloaded from this journal's web site,[1] as well as a site maintained by one of the authors.[2]

The input data of GraphClus must be in comma separated values or in XLS, both of which are widely used data storage formats. With both formats, the first row of the input file contains the names of the variables. Each remaining row contains the data for a sample, beginning with the name of the sample and then the values of the variables in the order given in the first row of the file. The output can be saved as an image (e.g. PDF, BMP, JPEG, PS), or it can be saved into tabulated data formats such as CSV or XLS.

Table 1
Average and standard deviation of consistency indexes of clustering methods on model datasets

Algorithm	Method	Consistency index	Standard deviation
GraphClus	Connection level 12	0.884	0.076
GraphClus	Connection level 9	0.809	0.027
GraphClus	Connection level 5	0.743	0.017
Complete linkage	Euclidean distance	0.733	0.059
Complete linkage	Cosine Theta	0.460	0.123
Single linkage	Euclidean distance	0.617	0.047
Single linkage	Cosine Theta	0.327	0.197
Average linkage	Euclidean distance	0.723	0.041
Average linkage	Cosine Theta	0.352	0.264

4. Testing on a synthetic dataset

In order to assess the robustness of the GraphClus method and to compare it to traditional hierarchical agglomerative cluster methods, synthetic datasets were generated with the following characteristics:

- Three groups were created, having 10 (Group A), 20 (Group B) and 30 (Group C) samples.

[1] http://www.iamg.org/CGEditor/index.htm.
[2] http://www.inf.u-szeged.hu/~busarobi/GraphClus/GraphClus.html.

Table 2
Example synthetic data set used to generate dendrograms in Figs. 2–4

	v0	v1	v2	v3	v4	v5	v6	v7	v8	v9
a01	90.870	63.224	88.663	88.283	96.668	93.394	88.157	88.738	91.424	87.560
a02	96.453	88.977	81.864	91.815	93.050	92.675	71.970	89.351	93.507	90.917
a03	95.933	88.784	86.092	87.526	97.077	88.882	88.939	89.132	88.516	87.251
a04	94.012	89.650	86.974	86.770	96.356	87.172	117.890	76.860	91.518	86.793
a05	137.710	91.086	86.546	93.688	93.707	86.602	89.544	89.908	83.491	84.463
a06	87.278	90.048	88.169	91.896	108.980	85.114	88.641	51.938	90.262	89.529
a07	98.936	89.670	96.689	91.603	91.420	91.555	93.751	94.121	90.135	99.293
a08	91.021	81.242	81.223	92.583	91.714	89.807	87.973	97.045	89.447	88.382
a09	93.798	93.090	93.458	89.130	93.796	88.312	91.850	94.903	93.782	87.885
a10	87.569	87.376	70.265	103.220	94.718	91.793	98.920	63.020	96.495	91.336
b01	110.650	106.850	104.020	92.898	102.530	93.968	98.185	94.012	99.966	88.802
b02	122.580	106.550	100.800	104.790	103.130	97.788	33.865	103.320	106.990	95.066
b03	101.840	103.240	96.349	209.050	99.687	97.443	99.312	101.900	111.630	94.806
b04	109.460	110.390	95.309	105.190	103.680	94.789	100.270	96.233	107.500	99.661
b05	106.260	107.700	96.997	102.830	103.530	98.258	107.200	96.867	101.710	94.288
b06	108.840	110.380	96.619	107.980	94.283	94.328	98.564	96.513	107.120	94.907
b07	100.640	102.630	92.122	103.840	100.190	94.742	98.487	101.470	85.600	94.242
b08	104.800	101.160	97.590	104.630	98.558	88.811	98.160	99.116	106.790	97.548
b09	107.330	40.067	96.346	104.120	102.430	100.910	97.992	98.293	107.350	99.494
b10	102.390	106.390	100.200	117.920	98.688	9.053	92.825	100.200	103.870	107.960
b11	94.869	112.000	96.400	93.371	98.232	94.154	91.600	100.670	102.620	93.467
b12	102.810	105.190	101.180	91.244	101.840	97.113	91.002	104.760	97.568	94.574
b13	102.880	104.970	96.736	106.160	104.360	94.172	95.672	99.634	107.250	93.294
b14	104.180	103.630	101.280	105.920	99.899	93.628	100.040	123.160	102.930	99.690
b15	97.653	111.230	47.939	102.100	96.350	84.443	95.656	99.728	96.139	97.482
b16	93.542	107.950	93.032	102.930	108.880	98.108	106.840	102.120	102.180	97.804
b17	106.020	114.980	94.597	110.050	89.768	90.703	95.566	96.112	104.920	94.049
b18	104.430	99.367	85.406	103.030	99.763	97.163	99.556	97.400	105.940	93.157
b19	106.180	100.720	128.100	153.270	96.968	93.849	95.213	96.383	102.260	92.647
b20	102.150	156.750	90.580	103.170	94.459	97.718	95.162	108.400	104.850	94.348
c01	109.500	100.340	114.690	110.640	104.870	113.070	108.670	108.130	111.810	105.110
c02	113.460	111.950	103.730	107.890	105.120	113.290	106.480	108.700	109.170	105.000
c03	109.350	108.200	107.640	104.700	113.030	119.480	108.940	109.500	111.170	105.730
c04	108.380	112.920	110.930	108.960	108.710	109.370	108.670	100.440	104.720	104.530
c05	111.770	113.420	107.710	101.770	109.930	115.520	106.640	108.670	106.830	109.270
c06	111.050	110.370	103.920	107.650	105.740	109.390	160.350	106.210	107.400	105.590
c07	112.390	108.430	107.010	108.050	109.530	116.940	103.810	109.470	108.020	104.540
c08	109.350	101.940	120.050	104.230	104.750	116.290	110.040	107.020	110.220	107.820
c09	112.760	108.910	104.380	104.970	105.770	112.330	109.220	101.770	103.570	103.780
c10	107.260	116.200	101.520	108.730	102.370	114.080	107.000	102.580	110.320	108.220
c11	110.750	110.320	107.110	102.290	91.187	109.440	110.250	100.390	108.890	18.803
c12	111.690	110.480	104.350	107.880	102.850	116.510	106.900	104.460	107.160	105.870
c13	110.780	114.490	119.710	107.630	105.140	113.790	104.940	109.890	106.500	105.370
c14	109.400	112.730	102.360	107.680	110.530	112.420	112.190	106.220	111.480	104.730
c15	110.090	111.250	103.320	104.550	109.950	110.660	103.700	109.830	111.060	105.330
c16	108.690	111.550	107.970	107.800	110.540	114.240	91.931	105.600	110.030	107.950
c17	102.520	110.480	97.421	161.390	108.270	115.790	103.700	106.610	97.607	103.920
c18	117.110	112.450	103.840	105.560	105.130	115.470	109.370	92.754	95.246	104.160
c19	112.090	109.850	104.410	105.330	106.050	112.310	111.680	105.650	107.740	104.750
c20	132.870	107.340	102.200	105.600	105.640	113.000	104.810	123.250	111.990	104.210
c21	110.810	112.580	94.275	108.080	100.650	122.900	105.350	121.620	112.070	107.840
c22	110.750	115.870	100.020	107.310	98.012	116.400	109.720	113.220	109.610	107.790
c23	108.470	113.350	101.790	105.270	106.280	79.379	110.840	166.890	110.230	111.880
c24	105.740	109.170	107.670	107.780	112.840	118.880	109.420	127.930	94.448	103.950
c25	106.250	111.690	104.320	103.610	108.760	106.380	105.830	113.790	106.820	108.960
c26	111.600	113.580	108.820	105.520	124.120	116.100	105.420	108.400	107.290	103.930
c27	118.310	112.950	92.469	107.700	104.830	116.380	112.180	106.740	105.390	109.000
c28	107.640	110.850	99.430	104.210	108.890	115.100	104.870	109.470	115.450	102.170
c29	110.670	121.400	107.600	108.970	101.170	117.530	109.860	107.590	107.470	109.310
c30	113.090	113.680	104.530	104.730	101.280	120.570	103.070	107.200	109.790	108.550

- Ten variables were created (v0 through v9).
- The values for each variable within a group followed a normal distribution with the mean (M) of a randomly determined integer between 1 and 10 and a standard deviation defined such that $\sigma/M = 0.2$.
- Noise was added via a Pareto distribution with shape and scale parameters of one.

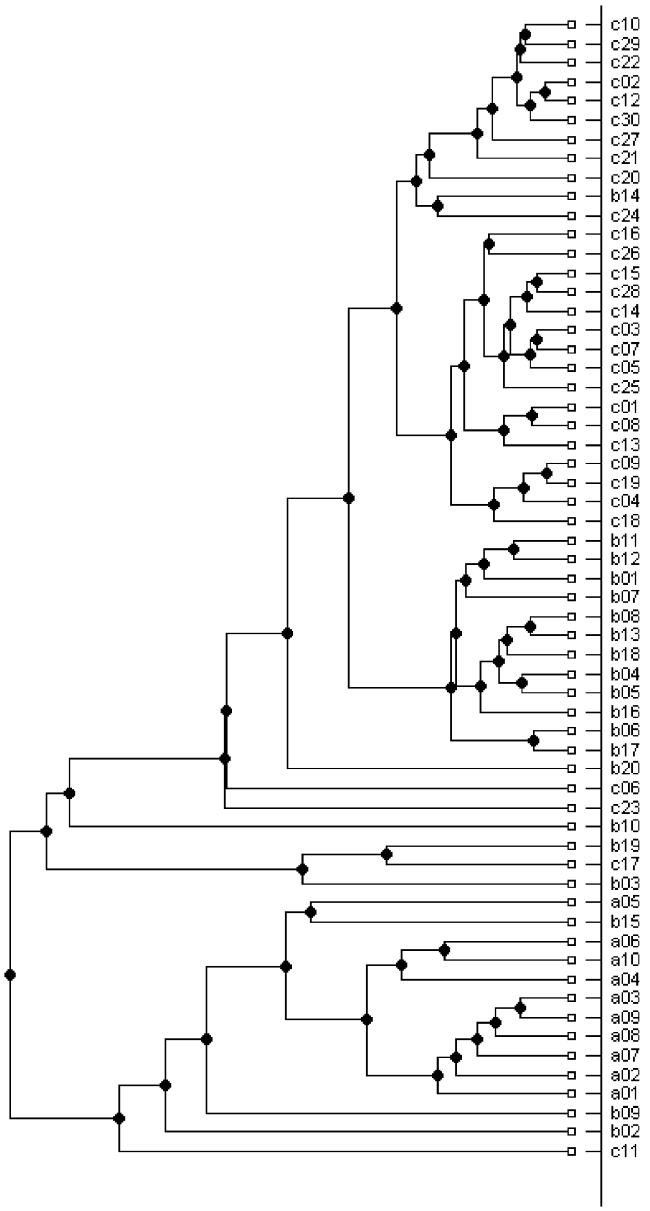

Fig. 3. Dendrogram for synthetic dataset using complete linkage Euclidian distance cluster analysis. Consistency indexes for this dendrogram are Group A, 0.95; Group B, 0.60; Group C, 0.79.

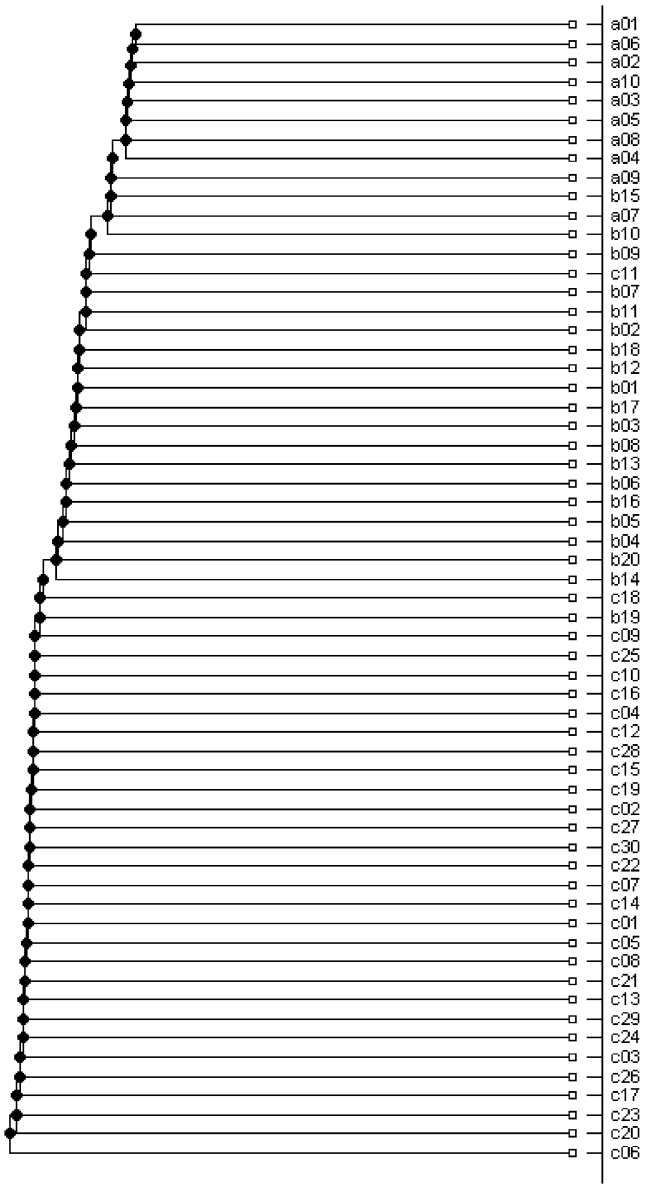

Fig. 4. Dendrogram for synthetic dataset using single linkage cosine theta cluster analysis. Consistency indexes for this dendrogram are Group A, 0.95; Group B, 0.005; Group C, 0.00.

- One hundred was added to each value in order to avoid negative numbers.

The philosophy for adding exponential-like noise to the dataset was to better represent suites of geological samples. The variables measured on a geological sample may be affected by various processes, altering their values. However, both the extent of alteration and the specific variables affected may not be identical among all members of a group, especially if the samples come from different geographic locations. As a result there may not be one "bad" variable, making simple exclusion of that variable for classification purposes impossible. By using Pareto noise, the intent was to model significant change for any variable in any sample, and that the variable most affected could be different from one sample to another. This type of data restructuring is difficult for traditional classification methods to handle because they tend to compare samples as a function of all included variables equally weighted.

Following clustering by GraphClus and other cluster methods (complete linkage, single linkage and average linkage), the ability to properly classify samples into groups was evaluated using a consistency index (P). This index was computed for every inner point T of the dendrogram as $P_i^T = (G_i^{obs}/G_i^{total})\prod_{i\neq j}(1-(G_j^{obs}/G_j^{total}))$, where i and j vary between one and the number of actual groups (three in this case). G_i^{total} denotes the total number of samples in the ith group and G_i^{obs} is the number of ith group members that can be found in the subtree below inner point T. This consistency index represents the separation capability of the clustering method; it is equal

to one for the *i*th class if and only if an inner point exists in the dendrogram for which the subtree below contains all members from the *i*th group and no samples from the other groups. Synthetic datasets with the characteristics described above were generated 100 times. Each dataset was evaluated by all the cluster methods. Each node of each dendrogram for each cluster method was then assessed to determine the maximum value of the consistency index for each of the three groups (A, B and C). The average and standard deviation of the 300 numbers (consistency index for each of the three groups in 100 datasets) for each cluster method are listed in Table 1. None of the cluster analysis methods examined separated all the groups perfectly. The GraphClus algorithm performed as good as or better than other cluster algorithms.

Table 3
Homogeneity, heterogeneity and discriminant indexes for groups chosen in Fig. 2

	1 Hom	2 Hom	Het	Disc
v6	0.67	0.42	0.93	0.64
v5	0.50	0.48	1.00	0.62
v9	0.50	0.47	1.00	0.62
v7	0.67	0.34	0.95	0.60
v3	0.67	0.74	0.38	0.57
v2	0.58	0.26	0.95	0.52
v1	0.28	0.50	0.78	0.48
v8	0.39	0.26	0.75	0.42
v0	0.25	0.34	0.85	0.42
v4	0.28	0.19	0.90	0.37

Variables listed in descending order of discriminant index.

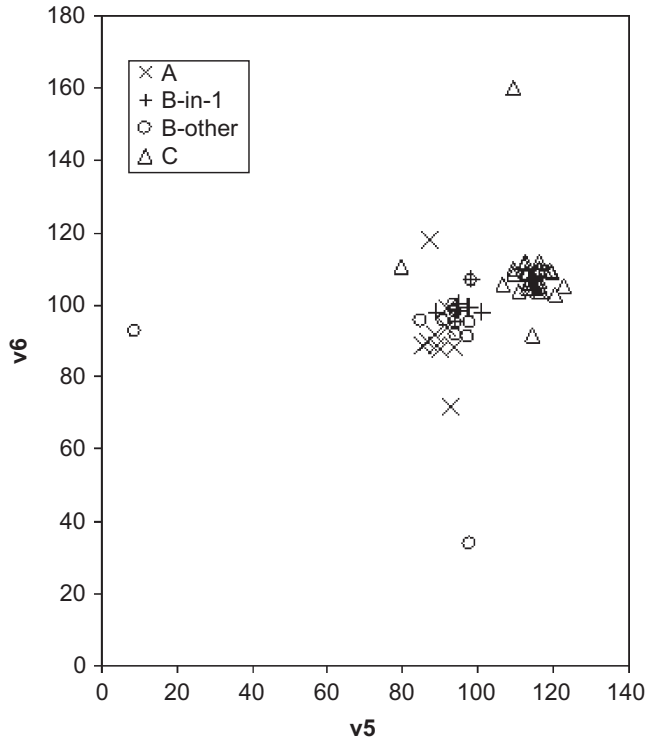

Fig. 5. Plot of v5 versus v6 from synthetic dataset. Group 2 defined in Fig. 2 contains all members of C. Group 1 contains half of the members of B (denoted B-in-1). Other members of B denoted B-other. Members of A are not in either group.

One of the 100 synthetic datasets and the resulting dendrograms of cluster analysis using GraphClus (connection level 12) and complete linkage Euclidian distance and single linkage cosine theta distance are shown in Table 2 and Figs. 2–4. To illustrate the use of the GraphClus indexes, two groups were chosen as shown in Fig. 2. Group 1 contained 10 of the 20 samples in Group B. Group 2 contained all 30 members of Group C. Table 3 lists the homogeneity indexes for the two groups, as well as the heterogeneity and discriminant indexes between the two groups. The two variables with the highest discriminant index were v6 and v5. A plot of these two variables is shown in Fig. 5; the members of Group B that were in Group 1 (Group B-in-1) are shown with "+"; the members of Group B that were not in Group 1 are shown with "o". The members of Group C and B-in-1 are reasonably well separated by these two variables, with Group C having higher v5 and v6. The ability to examine the underlying reasons for groupings is an asset of GraphClus compared to other cluster methods. The members of Group A are not well separated from Group B, but this is not surprising given that the variables (v6 and v5) were not chosen to maximize the separation of Group A from other samples.

5. Testing on a geochemical dataset

It is not uncommon for lavas from volcanoes to vary in composition through time. A set of 10 trace elements (La, Nb, Rb, Zr, Sr, Ni, Co, Cr, Cu and Zn) from four samples in each of three historic eruptions of Mount Etna (1865, 1911 and 1974) (Barbieri et al., 1993) was used to test the GraphClus method and compare it to established cluster methods. The mathematical or statistical analysis of compositional data is prone to complications from closure. Closure refers to the fact that compositional data

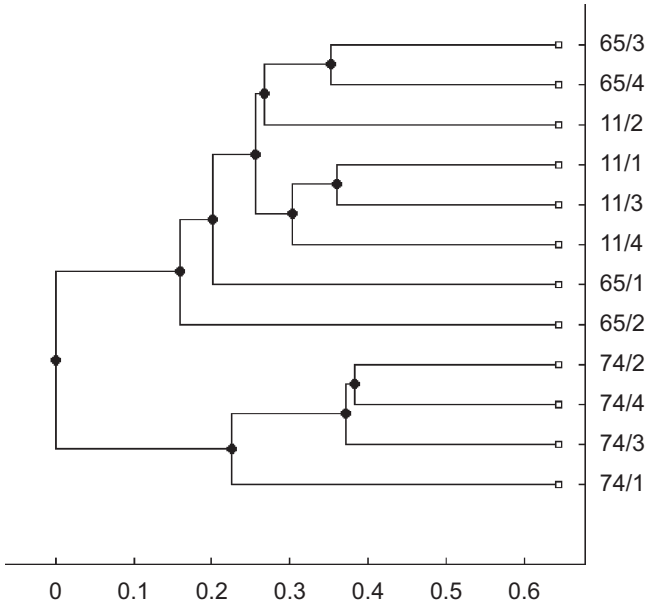

Fig. 6. Dendrogram using single linkage Euclidian distance cluster analysis for Mount Etna trace element dataset.

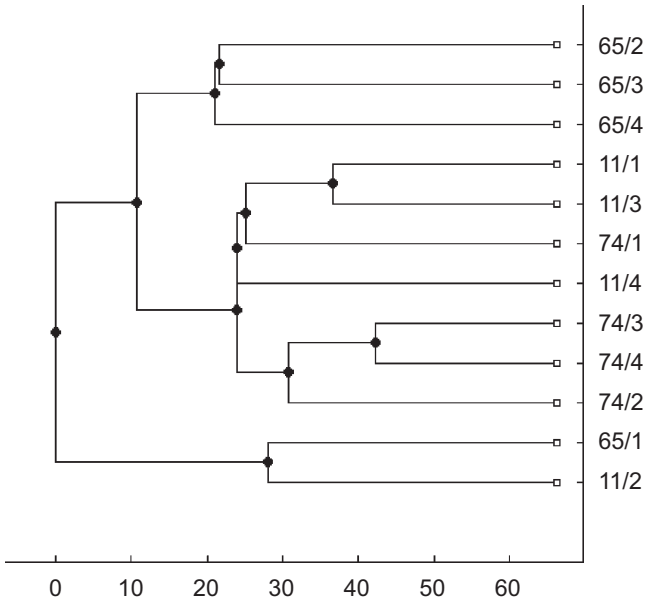

Fig. 7. Dendrogram using weighted pair average Euclidian distance cluster analysis for Mount Etna trace element dataset.

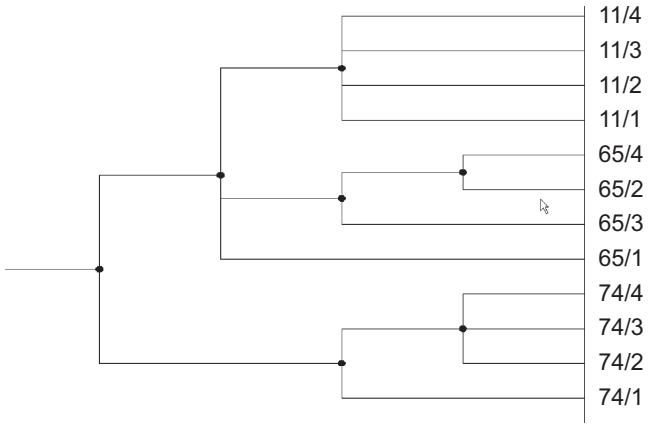

Fig. 8. Dendrogram using GraphClus (connection level 12) for Mount Etna trace element dataset.

sum to 100 and a change in the value of one variable automatically changes the values of the other variables in the composition. Therefore compositional data are not free to vary independently and may contain spurious correlations. It is the logarithms of the ratios of the compositional variables that reveal the relationships and structure in the data (Aitchison, 1986). The interested reader is directed to an excellent summary of this issue in Rollinson (1993) and methods to address this problem in Aitchison (1981, 1992, 2003). In the case of trace element data, the effect of closure is possible and as an added precaution the data were log-center transformed before clustering by single linkage, Weighted pair-group median (both with Euclidean distance) and GraphClus. The results shown in Figs. 6–8 indicate that GraphClus performed better than the other two methods at distinguishing the genetic groups. Grouping using several traditional cluster analysis methods on the non-transformed data also showed that they did not reproduce the historical structure, whereas GraphClus did (M Tóth and Engi, 1997).

6. Conclusions

A cluster analysis method based on graph theory was implemented in a computer program that can run on many operating systems and is available at the journal's web site. The method is well suited to uncovering genetic groups within altered datasets where the nature of the alteration is different from sample to sample. The identity of variables that are responsible for establishing different groups can be revealed by calculating an index related to the homogeneity of a variable within a group and heterogeneity and discrimination indexes related to the differences between groups. These help the user gain insight into the underlying reasons for the groupings. Testing on both natural geochemical and synthetic datasets indicates that this method generally performs better than traditional cluster analysis methods for these altered datasets.

As with all cluster analysis methods, GraphClus can be applied in any situation when genetic relationships between the samples in a dataset are sought. GraphClus has been used to group a set of major and trace element data from amphibolites in the Szeghalom crystalline dome in Hungary to better define the boundaries between different structural units in a geologically complex polymetamorphic area (M Tóth, 1992). Other possible applications could be: to source ancient stone tools to quarries by classifying based on geochemistry (Sinton and Sinoto, 1997), to classify materials developed under open system conditions, such as soils.

Acknowledgements

We wish to thank Fernando Pacheco and two anonymous journal reviewers for reviewing much earlier, practically unrelated versions of this manuscript. The efforts of John Tipper and Hugh Rollinson drastically changed and improved the overall thrust of the manuscript presented here. Randal Pell and Eric Grunsky are thanked for helpful comments.

Appendix A. Supplementary materials

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.cageo.2008.05.007.

References

Aitchison, J., 1981. A new approach to null correlations of proportions. Mathematical Geology 13, 175–189.
Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London, 416pp.
Aitchison, J., 1992. On criteria for measures of compositional difference. Mathematical Geology 24, 365–379.

Aitchison, J., 2003. The Statistical Analysis of Compositional Data. Blackburn Press, Caldwell, NJ, 416pp.
Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press, New York, NY, 359pp.
Barbieri, M., Cristofolini, R., Delitala, M.C., Fornaseri, M., Romano, R., Taddeucci, A., Tolomeo, L., 1993. Geochemical and Sr-isotope data on historic lavas of Mount Etna. Journal of Volcanology and Geothermal Research 56, 57–69.
Bezdek, J.C., 1974. Cluster validity with fuzzy sets. Journal of Cybernetics 3, 58–72.
Botafogo, R.A., 1993. Cluster analysis for hypertext systems. In: Proceedings of the 16th Annual International Association for Computing Machinery Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval. ACM Press, Pittsburgh, PA, pp. 116–125.
Diestel, R., 2005. Graph Theory. Springer, New York, NY, 415pp.
Everitt, B., 1980. Cluster Analysis. Wiley, New York, NY, 136pp.
Everitt, B.S., Landau, S., Leese, M., 2001. Cluster Analysis. Hodder Arnold Publication, London, 238pp.
Gill, D., Tipper, J.C., 1978. The adequacy of non-metric data in geology: test using a divisive-omnithetic clustering technique. Journal of Geology 86, 241–259.
Gordon, A.D., 1999. Classification. Chapman and Hall, New York, NY, 256pp.
Hartuv, E., Shamir, R., 2000. A clustering algorithm based on graph connectivity. Information Processing Letters 76, 175–181.
Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data. Wiley, New York, 342pp.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkley Symposium on Mathematical Statistics, Berkeley, CA, pp. 281–297.
Marshal, C.W., 1971. Applied Graph Theory. Wiley, New York, NY, 322pp.
M Tóth, T., 1992. Földtani objektumok csoportosítása gráfelmélet segítségével Szeghalmi amfibolitok példáján (Classification of geological samples using graph theory, demonstrated on amphibolites from Szeghalom). Földtani Közlöny 122, 251–263.
M Tóth, T., Engi, M., 1997. A new cluster analysis method for altered rock samples. Schweizerische Mineralogische und Petrographische Mitteilungen 77, 439–447.
Pacheco, F.A.L., 1998. Finding the number of natural clusters in groundwater data sets using the concept of equivalence class. Computers & Geosciences 24, 7–15.
Pacheco, F.A.L., Van der Wejjden, C.H., 1996. Contributions of water–rock interactions to the composition of groundwater in areas with sizeable anthropogenic input. A case study of waters of the Fundão area, central Portugal. Water Resources Research 32, 3553–5370.
Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814.
Rollinson, H.R., 1993. Using Geochemical Data: Evaluation, Presentation, Interpretation. Wiley, New York, NY, 352pp.
Romesburg, C., 2004. Cluster Analysis for Researchers. Lulu Press, Napa, CA, 344pp.
Sinton, J.M., Sinoto, Y., 1997. A geochemical database for Polynesian adze studies. In: Weisler, M. (Ed.), Prehistoric Long-Distance Interactions in Oceania: An Interdisciplinary Approach. New Zealand Archeological Association Monograph 21, Auckland, pp. 194–204.
Trudeau, R.J., 1993. Introduction to Graph Theory. Dover Publications, Inc., Mineola, NY, 224pp.
Van Dongen, S., 2000. Graph clustering by flow simulation. Unpublished Ph.D. Dissertation, University of Utrecht, Utrecht, The Netherlands, 173pp.
van Groenewoud, H., Ihm, P., 1974. A cluster analysis based on graph theory. Vegetatio 29, 115–120.
Wilson, R.J., 1996. Introduction to Graph Theory, fourth ed. Addison-Wesley, Boston, MA, 184pp.