AN ASSESSMENT OF NUMERICAL CLASSIFICATORY METHODS FOR GEOGRAPHY

SHARIFAH MASTURA SYED ABDULLAH
Universiti Kebangsaan Malaysia

SINOPSIS

Perkembangan teknik pelbagai angkubah adalah selari dengan perkembangan dan kemudahan alat komputer. Kertas ini cuba menilai kegunaan-kegunaan salah satu daripada teknik-teknik pelbagai angkubah ini, iaitu, analisa perkelompokan, juga dikenali sebagai kelasifikasi numerikal. Dalam membincangkan kertas ini prosedur-prosedur teknik perkelompokan ini diterangkan. Sebagai satu contoh untuk menjelaskan teknik ini beberapa sedimen yang diambil daripada 30 sungai digunakan. 'Coefficient Dissimilarity iaitu 'Euclidean Distance' digunakan sebagai ukuran persamaan dan lima strategi penyatuan dipilih untuk menunjukkan berbagai-bagai keputusan.

SYNOPSIS

The development of the multivariate technique has grown parallel with the development and ready availability of digital computers. This paper attempts to assess the usefulness of one of these multivariate techniques, that is, the cluster analysis also commonly known as numerical classification. In this paper, procedures in developing clustering technique are described. As an illustration of this technique, continuous data of various sediments taken from 30 streams are used. Dissimilarity coefficient (the Euclidean distance) is used as measure of similarity and five fusion strategies are selected to produce various results.

Introduction:

The main surge of computer-based studies began in 1961 with the widespread use of machines produced in the IBM series. The demand on computer time is now doubling every two years. The main uses of computers in systems analysis appear to be studies in multivariate statistics, trend-surface decomposition, computer graphics and simulation (Hagget, 1969: 497-520).

Workers in the 1950's, using conventional statistics such as mean test, variance analysis, correlation and regression analysis, were the first to respond to the facilities and resources opened up by the computer. Studies were quickly extended to include multivariate methods such as principal

component analysis, factor analysis and multiple descriminant functions (King, 1969).

Researchers struggling through desk calculators to an exhausting five variables equation were suddenly confronted with powerful 'package' programmes which could not only compute coefficients for dozens of variables but provided means of selecting optimal sets from the rack (Hagget, 1969).

Something of the power of multivariate analysis may be seen in the geographical use of principal component analysis and factor analysis. These techniques have the capability of reducing immense arrays of data to a series of coherent and interpretable dimensions of factor. The amount of mathematics involved is normally outside the range of mechanical calculators, hence computer programmes have to be used.

It must be stressed that multivariate methods, including numerical taxonomy, have been applied far more in the field of biology and ecology when compared to the earth sciences, especially geomorphology.

In the last decades or so there has been a swift growth of taxonomic methods in ecology. Since the pioneering work of Sorenson in 1948 who employed both coefficient of association and cluster analysis, others have followed suit: Goodall (1953: 39-63), Williams and Lamberts (1959: 427-445), Greig-Smith (1964), and Whittaker (1967: 207-264).

Multivariate analysis in Geography is picking up in popularity. Geographical investigations are often of a multivariate nature in that they seek to analyse many attributes measured at different localities. Three of these techniques, namely, factor analysis, cluster analysis and multiple discriminant analysis are widely used.

This paper attempts to discuss how one of the multivariate techniques, that is, the cluster analysis, can be applied in geography. In this discussion procedures in developing clustering techniques are described. Continuous data of various sediments taken from 30 streams are used as an illustration. Dissimilarity Coefficient, the Euclidean distance, is used as measure of similarity and five fusion strategies are selected to produce various results.[1] It should be emphasised, however, that this paper is mainly concerned with the methodology rather than the problem solving work of this technique.

NUMERICAL CLASSIFICATION

Classification can be derived subjectively or objectively or by varying degrees of objectivity. This paper attempts to assess the usefulness of a clustering technique which is actually a numerical classification.

Clustering techniques have been developed in response to the following problem: given a sample of 'N' objects or individuals, each of which is measured on each of 'P' variables, devise a classification scheme for grouping the objects into 'g' classes (Everit, 1974). In clustering methods,

1. The analysis of the data is by the use of ICL 1907 computer of Sheffield University. The package programme used is clustan 1A which is a comprehensive suite of Fortran IV programme developed by Wishart 1969-1970.

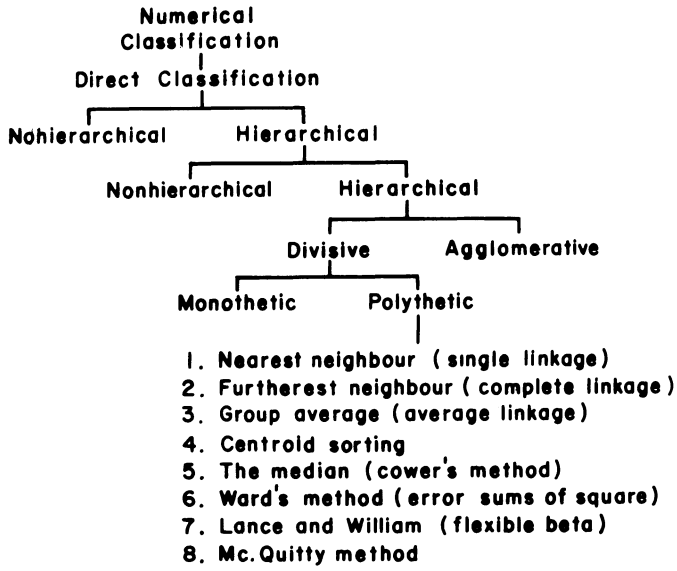

Figure 1. A choice of classification strategies.

it is conventional to arrange data in the form of a N x P matrix in which N columns represent the individuals and P rows represent variables. The data matrix is then used to estimate the resemblance between pairs of individuals. The scores in a data matrix may be expressed in many ways and they depend on the nature of the variables. They may be presence/absence data, rank data, meristic data or continous data.

Cluster analysis or numerical classification is divided into two distinct types: the hierarchical or non-hierarchical. The former is equated with the production of a dendrogram while the latter is often known as recticulate.

The hierarchical type is divided into two major groups or methods. The first one is the divisive hierarchical method which begins with a complete set of entities. These entities will then be classified and divided progressively into smaller groups.

The second group is the agglomerative hierarchical method which starts with a single entity It then considers which other entity is most similar to the chosen one in some defined sense, followed by a third entity which is most similar to the first and second entities and this proceeds to build up a cluster of entities. At some point it may be decided that none of the remaining entities is similar enough to be associated with the cluster being formed and a new cluster is initiated.

MONOTHETIC AND POLYTHETIC METHODS:

The terms "monothetic" and 'polythetic" have been associated with hierarchical classification. In monothetic procedures every group at every stage is being defined by the presence or lack of special attributes, or in other words, the union is based on one attribute. In polythetic procedures

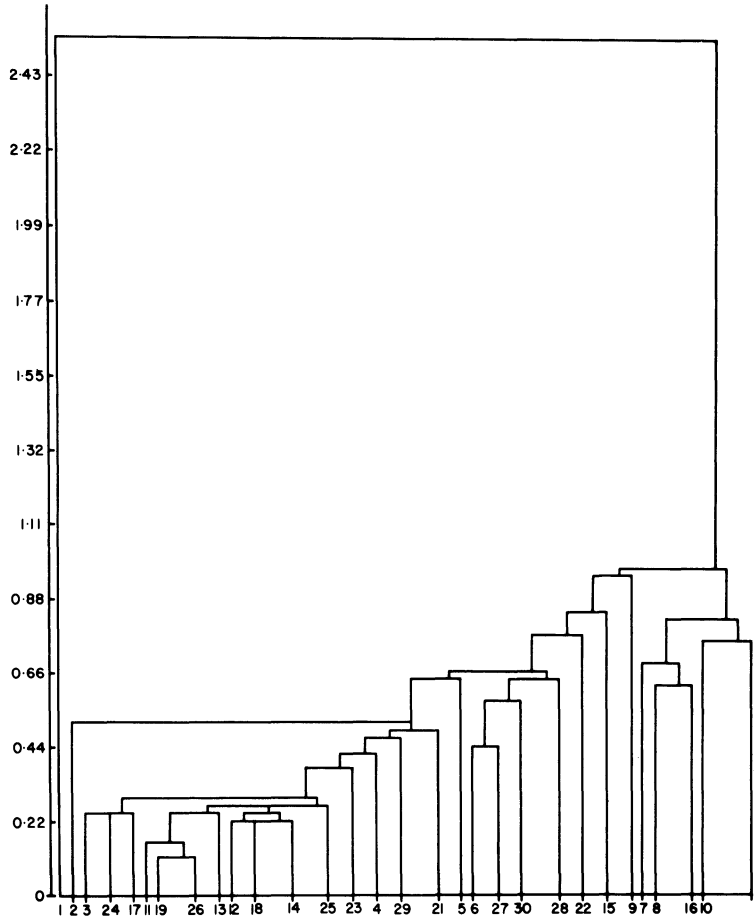

Figure 2: Polythetic agglomerative classification employing Euclidean distance similarity coefficient and displaying nearest neighbour sorting strategy

the union is based on all the attributes.

The agglomerative monothetic cannot exist but in a trivial sence. At the other extreme the divisive polythetic classification is computationally out of the question for most researchers and thus it has not been sufficiently developed. The most popular classifications in use at present are the divisive monothetic and agglomerative polythetic. Agglomerative polythetic method is, however, historically older than the divisive monothetic, deriving at least from the work of Kulczynski in 1972. The agglomerative polythetic will be dealt with in greater detail below as this method is comparatively more superior and popular than the other methods.

AGGLOMERATIVE POLYTHETIC METHOD:

The first stage in agglomerative polythetic method is the calculation of

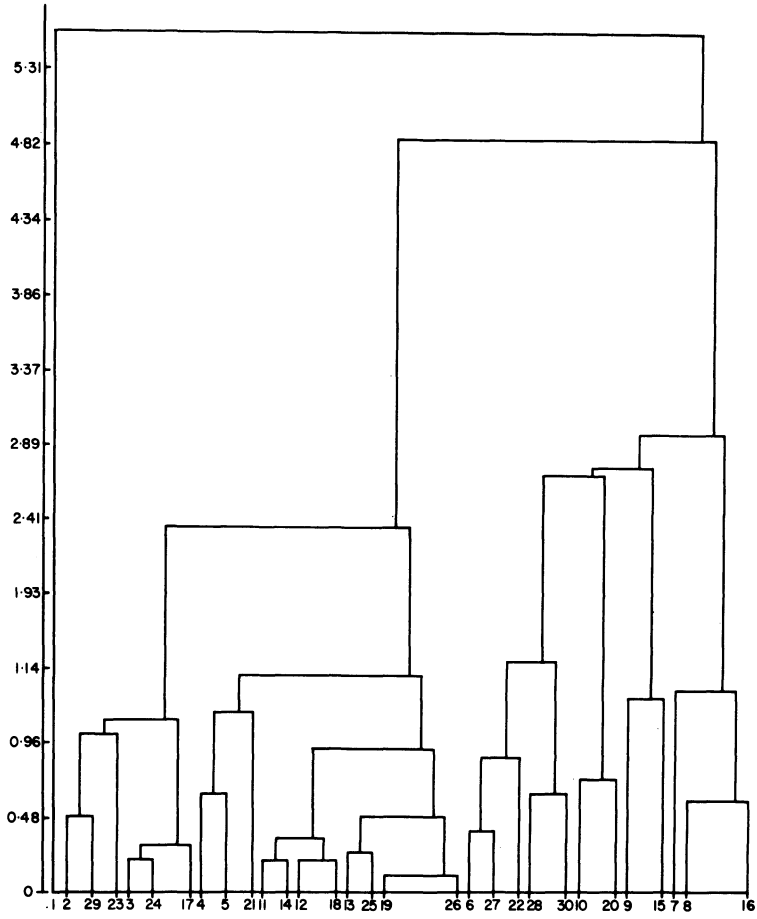

Figure 3: Polythetic agglomerative classification employing Euclidean distance similarity coefficient and displaying furtherest neighbour sorting strategy

similarity matrix between two individuals using some kind of index of coefficient. Index of similarity coefficient can be broadly divided into four groups: distance coefficient; association coefficient: correlation coefficient; and probablistic similarity coefficient.

Distance coefficient measures distances between individuals in a space defined in various ways, and the most familiar measurement is simple Euclidean distance. Distance coefficient is the converse of similarity; it is in fact a measure of dissimilarity. It has great intellectual appeal to taxonomists as it is easiest to visualise (Sneath & Sokal, 1973). Association coefficient may involve quantitative data. It can be applied to rank and continuous data by sacrificing information. Correlation coefficient measures proportionality and independence between pairs of individual sectors. The product moment correlation coefficient is frequently applied

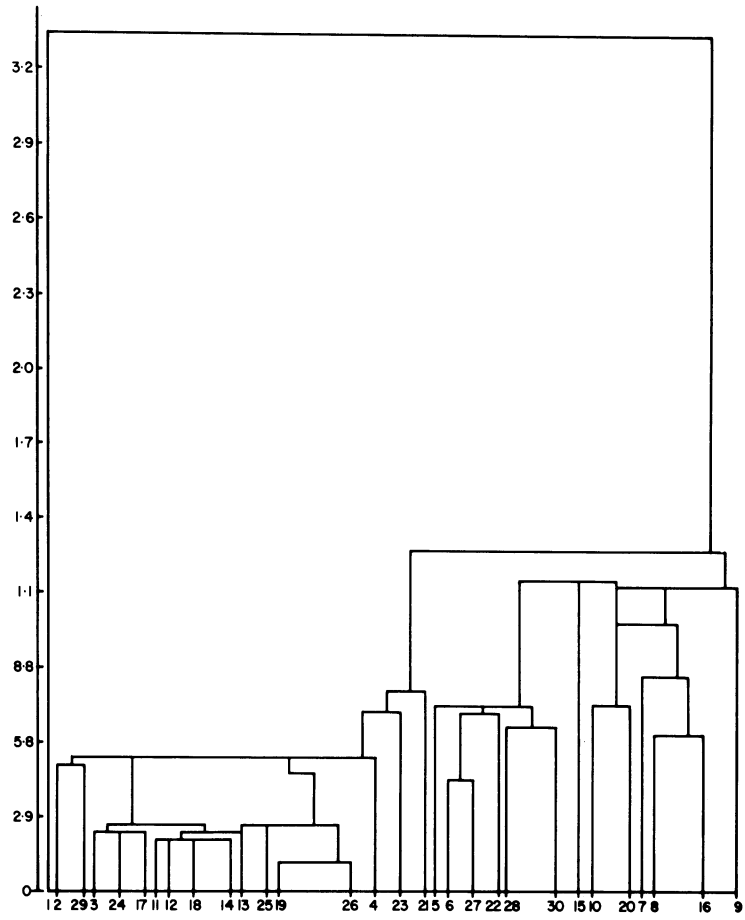

Figure 4: Polythetic agglomerative classification employing Euclidean distance similarity coefficient and displaying centroid method sorting strategy

to continuous data. Probablistic coefficient is more recent and it includes information-type statistics which measure the homogeneity of the data by partitioning or sub-partitioning sets of individuals (William & Dale, 1965: 35-68).

The next stage is the sorting strategy of fusion. The sorting strategy involves grouping together of individuals to form a dendrogram. Eight types of sorting strategies are recognized. They are nearest neighbour or single linkage, furtherest neighbour or complete linkage, centroid, medium, group average, Ward's, Mcquitty and the Lance and William flexible Beta method.

The result of fusion is expressed in a dendrogram form. The distance away from the base line at which two individuals or groups joined is a direct expression of their degree of similarity. Hence the most similar

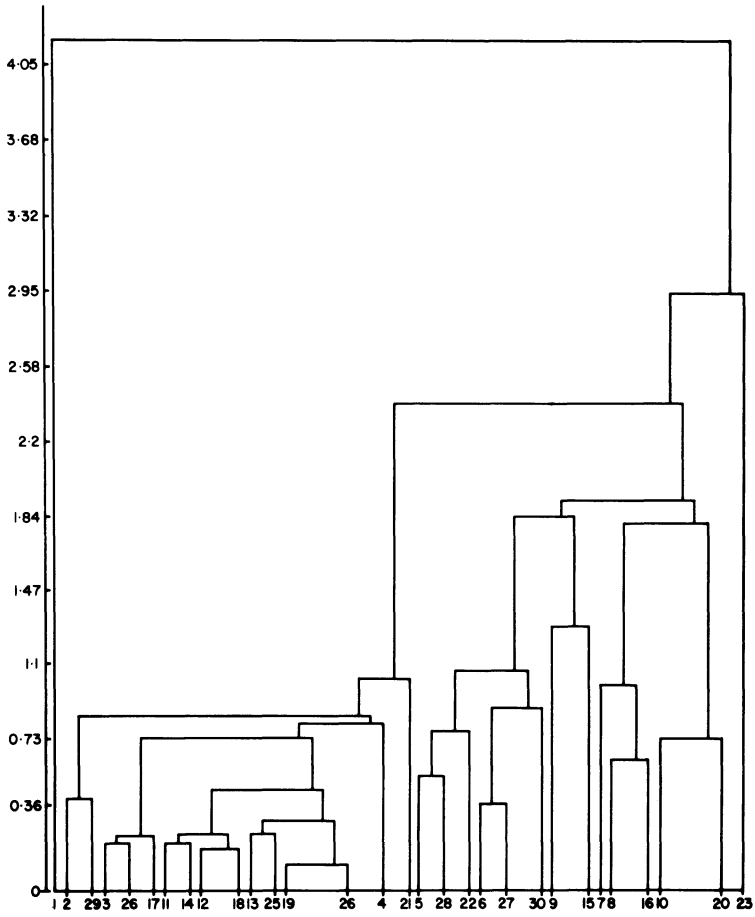

Figure 5: Polythetic agglomerative classification employing Euclidean distance similarity coefficient and displaying group average sorting strategy

groups are first joined. The best fusion will depend on its ability to form a good group structure. Figure 1 shows the stages of numerical classification techniques already described above. The differences between these arise essentially because of the different ways of defining distance (or similarity) between an individual and group containing several variables or between groups of individuals.

Test runs on agglomerative polythetic techniques of numerical classification are given in Figures 2, 3, 4, 5 and 6. The variables consist of 13 different sediment samples taken from 30 different streams. Euclidean distance coefficient is used for similarity measures and fusion strategies used are the nearest neighbout, furtherest neighbour, centroid, group average and Ward's method.

Some methods do not produce satisfactory clusters. Usually, this is due to chaining effects as seen in Figure 2. Chaining is a tendency of clustering

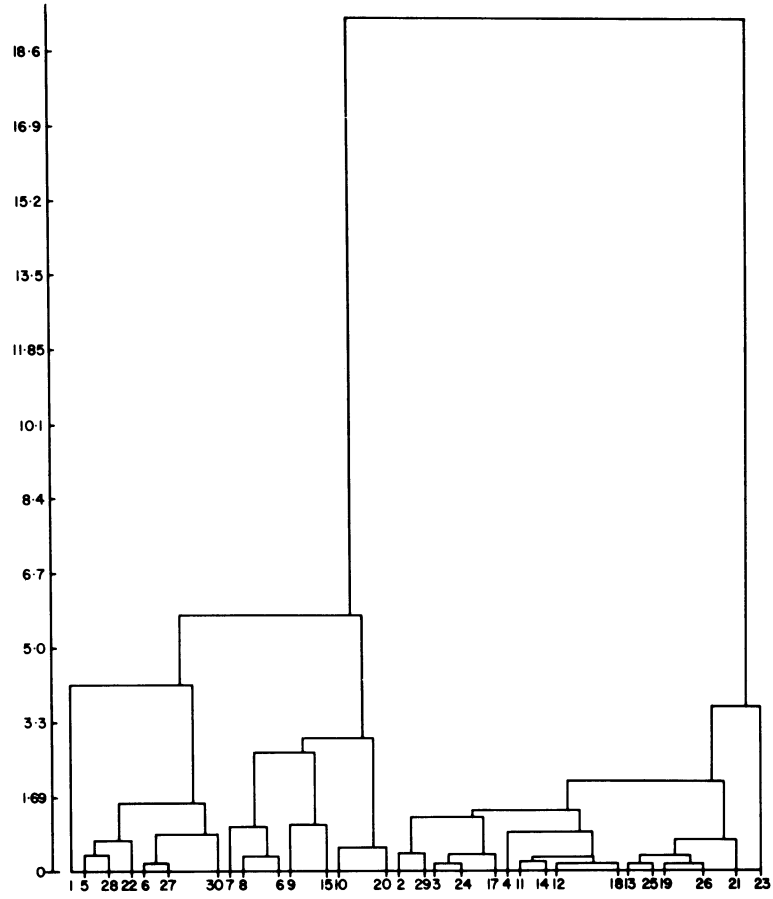

Figure 6: Polythetic agglomerative classification employing Euclidean distance similarity coefficient and displaying word's method sorting strategy

together certain individuals at a relatively low level and to incorporate individuals into existing clusters rather than to generate new clusters. Generally, chaining is considered to be a defect. However,; Jardine and Sibson (1968 177-184) argue that nearest neighbour is the best mathematical method and point out that to treat chaining as a defect is misleading. Chaining, according to them, is simply a description of what a method does and if one is looking for optimally connected clusters, and not for homogenous clusters, such a method may be useful. However, because of chaining, nearest neighbour may fail to resolve relatively distinct clusters. Many of the density reach techniques arose from attempts to correct the effect of chaining in the single linkage method (Everitt, 1974).

Figures 3, 4 and 5 produce slightly better dendrograms than the nearest neighbour method. However, with the exception of Figure 3, the imbalance in the two clusters is obvious. But Figure 4 (centroid method)

shows some reversals which reflect another weakness of a dendrogram. "Reversals effect" occurs when the computer prints one fusion which is at a lower level than has already been drawn on the diagram.

Figure 6 (Ward's method) produces dendrograms that are conspicuous for their symmetrical hierarchical structure even though the 'steps' or change in levels is average. The dendrogram identifies two large groups. Distinct smaller groups are clearly present.

In conclusion, the above discussion illustrates that geographical data can be analysed using agglomerative polythetic with five selected sorting strategies. The Ward's method shows good structured dendrogram while the others show weaker, structured dendrograms with chaining, reversals and some crowding.

Thus, suffice to say that the choice on the sorting strategy does influence the structure of the dendrogram and, therefore, to a certain extent the results of the analysis. The most important factor is that the user must be aware that the choice showed suits the data that one is dealing with so that the results would easily be grasped and analysed.

On the whole, the agglomerative polythetic technique in analysing geographical data should be useful. Firstly, most geographical data are multivariate in nature and this suits well with the numerical classification or cluster analysis technique which deals only with such data. Other uses of this technique relevant to analysing and solving geographical problems include finding a true typology, model fitting, prediction generating and data reduction (Everitt 1974).

Application of the cluster analysis technique can also and has been used in the other fields of social sciences and humanities. For example, Buechley (1967· 53-69) has made a cluster analysis of family names in different geographic localities. Wishart and Leach (1970: 90-99) have used several methods of clustering in examining the works of Plato so as to determine the probable chronology, and Grumm (1965: 350-362) has studied the voting behaviour of legislators by average linkage clustering. However, in applying the cluster analysis technique in the social science, one major problem that has to be considered is the question of quantifying variables. Some data in the social sciences are difficult to quantify and therefore would pose a problem to the data input. Such a problem is least faced by the physical sciences.

In short, in multivariate analysis, the cluster analysis method has applications in geography as well as in other fields of study

REFERENCES

Buechley, R.W "Characteristic Name Sets of Spanish Population." 1967 *Names.* 15: 53-69.

Everitt, B. *Cluster Analysis.* Social Science Research Council, London: 1974 HEB Ltd.

Godall, D.W "Objective Methods for the Classification of Vegetation. I. 1953 The Use of Positive Interspecific Correlation." *J Bot.* 1: 39-63.

Greig-Smith, P. (1964) *Quantitative Plant Ecology*. 2nd. Ed. London:
1964 Butterworth.

Grumm, J.G. "The systematic analysis of blocs in the study of legislative
1965 behaviour." *Western Political Quart.* 18: 350-362.

Hagget, P. (1969) "On Geographical Research in a Computer
1969 Environment. *Geogr. J.* 135: 497-520.

Jardine, N. and Sibson, R. "The Construction of Hierarchic and Non-
1968 Hierarchic Classification." *Computer J.* 11: 177-184.

King L.J. *Statistical Analysis in Geography*. New York: Englewood Cliff.
1969

Sneath, P.H.A. and Sokal, R.R. *Numerical Taxonomy*. San Francisco:
1973 Freeman.

William, W.T. and Dale, M.B "Fundamental Problems in Numerical
1965 Taxonomy." *Advance Bot. Res.* 2: 35-68.

William, W.T. and Lambert, J.M. "Multivariate Methods in Plant
1959 Ecology. V. Similarity Analysis and Information Analysis." *J.
Ecol.* 54: 427-445.

Wishart, D. and Leach, S.V. "A Multivariate Analysis of Platonic Prose
1970 rythm." *Computer Stud. Humanities Verbal Behav.* 3: 90-99.

CATATAN KEPADA PENULIS

GAYA PENULISAN

AKADEMIKA

Jernal Ilmu Sains Kemasyarakatan dan Kemanusiaan
Universiti Kebangsaan Malaysia

Makalah

Penulis hendaklah mengirimkan 3 salinan (termasuk satu salinan asal dan 2 salinan berxerox) bagı mempercepatkan prosés penilaıan makalah itu. Penulis hendaklah menuliskan nama serta tajuk makalah di atas kertas yang berasıngan. Tinggalkan ruang (margin) sekurang-kurangnya 1 inci, dan gunakan kertas berukuran quarto. Panjang makalah termasuk jadual, catatankakı, dan bibliografi mestilah tıdak melebihi 40 muka serta bertaip dengan menggunakan 3 langkau (trıple-space manuscrıpt).' Makalah boleh ditulis dalam bahasa Inggerıs atau bahasa Malaysıa. Bagı makalah berbahasa Inggeris ejaannya hendaklah disesuaikan dengan ejaan **Oxford Dictionary,** bagı makalah dalam bahasa Malaysıa ejaan **Kamus Dewan** (ejaan baru) hendaklah dirujuk.

Catatankaki

Sebaik-baiknya catatankaki janganlah memasukkan bahan-bahan rujukan atau bibliografi. Catatankaki hanya mengandungı keterangan-keterangan tambahan yang kurang sesuai untuk dimasukkan di dalam teks. Catatankaki hendaklah ditunjukkan di dalam teks dengan menggunakan angka-angka 1, 2, 3 dan seterusnya, dan hendaklah ditaıp pada muka yang berasıngan dan dilampırkan pada bahagıan akhir teks, iaitu sebelum bahan rujukan.

Penghargaan boleh dimasukkaı sekali dengan catatankakı mengikut susunan angka tadi.

Rujukan/Bibliografi

Rujukan hendaklah dinyatakan di dalam teks menurut sıstem Harvard. Misalnya: (Dahlan 1976:10), ıaitu ianya mengandungı nama pengarang, tahun dan muka. Sekıranya kita merujuk kepada seluruh makalah atau buku, tıdak perlu kita tuliskan mukanya. Sekiranya tulisan yang sama dirujukkan, akan ditulis sekali lagı seperti di atas tadi, (Dahlan 1976:15). Jika dua atau lebih makalah/buku yang dirujuk itu telah ditulis oleh pengarang yang sama dan diterbıt dalam tahun yang sama, maka rujukan dibuat seperti: (Dahlan 1976a:10) dan (Dahlan 1976b:2) dan sebagaınya. Pecahan ke 'a', 'b', dan seterusnya akan berpandu kesusunan abjad berdasarkan huruf pertama pada tajuk-tajuk makalah/buku ıtu.

Satu senaraı rujukan hendaklah disediakan di akhır makalah. Senarai

ini hendaklah ditulis berasingan serta diatur menurut susunan abjad. Contohnya seperti di bawah ini.

Jernal

Dahlan, H.M. "Malay Traditional Society and a Colonial
1976 Encounter". **Akademika,** Bil. 9, m.s. 1—20.
Keats, D.M., Keats, J.A., dan Mohammad Haji Yusuf
1976 "Attribution of Reasons for Religious Beliefs in Four Ethnic Groups". **Akademika,** Bil. 9, m.s. 21—28.

Buku

Kolko, James **The Limits of Power,** New York:
1972 Harper and Row Publishers.

Oberg, Kalervo **Culture Shock.** Indianapolis:
1974 Bobbs—Merril Company.

Makalah Dalam Buku

Lazarus, Richard S. "Cognitive and Personality
1967 Factors Underlying Threat and Coping", dalam Mortimer H. Appley and Richard Trumbull, penyunting, **In Psychological Stress,** New York: Appleton — Century Crofts.

SINOPSIS

Penulis hendaklah melengkapkan 2 sinopsis pendek, satu dalam bahasa Malaysia dan satu lagi dalam bahasa Inggeris.

Proof

Jika masa mengizinkan, penulis-penulis makalah akan menerima salinan proof untuk disemak bagi kesalahan-kesalahan percetakan **sahaja.** Sekiranya masa mendesak dan tempat tinggal penulis terlalu jauh, salinan proof akan disemak oleh sidang pengarang sendiri. Sidang pengarang mempunyai hak untuk membuat sebarang "editorial revision" tanpa menghubungi penulis.

A NOTE TO CONTRIBUTORS

AKADEMIKA

Journal of Humanities and Social Sciences
National University of Malaysia

GUIDE FOR CONTRIBUTORS:

ARTICLES

Authors should send 3 copies (the original and two photostat copies) to expedite evaluation of their articles. Authors should write their names as well as the titles of their articles on a separate sheet of paper. Leave at least one-inch margins and use quarto-sized papers. Each article, including tables, footnotes, and bibliography must not exceed 40 pages in length and should be type-written in triple spacing. Articles may be written either in English or Bahasa Malaysia. Spellings in the English-language articles should conform to those in the Oxford Dictionary and Bahasa Malaysia articles to those in Kamus Dewan (new spelling).

Footnotes

As far as possible, footnotes should not include references or bibliographies. They should only consist of additional information which are not suitable for inclusion in the text. Footnotes should be indicated in the text by the use of numbers 1, 2, 3 and so on, and should be type written on a separate sheet and attached at the end of the text immediately before the list of references. Acknowledgements may be included together in the footnotes in numerical order

Reference/Bibliography

References shown in the text should follow the Harvard System. For example: (Dahlan 1976 : 10), that is, it includes the author's name, year and page. Where reference is made to the article or book as a whole, page numbers need not be indicated. If the same article or book is subsequently referred to, it is necessary to repeat what has been written before, that is, (Dahlan 1976 : 15). If two or more articles/books referred to have been written by the same author in the same year, then the references is made thus: (Dahlan 1976a : 10) and (Dahlan 1976b 2) and so on. The subdivisions 'a', 'b', etc. should be in alphabetical order based on the first letter of the titles of the articles/books.

A list of references should be made available at the end of the article. The list should be type written separately and arranged in alphabetical order.

Some examples are shown below:

Journal

Dahlan, H.M. "Malay Traditional Society and a Colonial Encounter",
1976 Akademika, No. 9, p. 1—20.

Keats, D.M. Keats, J.A., and Mohammad Haji Yusuf "Attribution of
1976 Reasons for Religious Beliefs in Four Ethnic Groups".
Akademika, No. 9, p. 21—28.

Books

Kolko, James The Limits of Power. New York: Harper and Row Publishers.
1972

Oberg, Kalerov Culture Shock. Indianapolis: Bobbs—Merrill Company.
1974

Articles from Books

Lazarus, Richard S. "Cognitive and Personality Factors Underlying
1967 Threat and Coping". In Mortimer H. Appley and Richard Trumbull, editors, In Psychological Stress, New York: Appleton-Century Crofts.

Synopsis

Authors should complete two short synopsis, one in English and the other in Bahasa Malaysia.

Proof

Time permitting, authors shall receive a copy of their pre-publication articles to be proof read. However, if time does not permit, proof reading will be done by the Editorial Board. The Editorial Board of Akademika reserves the right of "editorial revision" without consultation with respective authors.

All editorial correspondence and articles for submission should be addressed to:—

The Editor,
Akademika,
c/o Department of Political Science,
Universiti Kebangsaan Malaysia,
Bangi, Selangor,
MALAYSIA.