DATA ANALYSIS AND INFORMATICS 3

NORTH-HOLLAND

edited by

E. Diday

M. Jambu

L. Lebart

J. Pagès

R. Tomassone

INRIA

DATA ANALYSIS AND INFORMATICS, III

Proceedings of the Third International Symposium on
Data Analysis and Informatics,
organised by the Institut National de Recherche en Informatique
et en Automatique,
Versailles, October 4-7, 1983

edited by

E. DIDAY
Université Paris IX - INRIA

M. JAMBU
Centre National d'Etudes des Télécommunications -
Centre National de la Recherche Scientifique

L. LEBART
Centre National de la Recherche Scientifique - CREDOC

J. PAGES
Commissariat à l'Energie Atomique

R. TOMASSONE
Institut National de la Recherche Agronomique - INA P-G

1984

NORTH-HOLLAND
AMSTERDAM • NEW YORK • OXFORD

Third International Symposium on
Data Analysis and Informatics
Versailles, October 4-7, 1983

Sponsored by
AFCET, ASU, CNET, CNRS,
CHAPITRE FRANÇAIS ACM,
CEA, IERESM, INRA, ISI, SFC

Organised by
Institut National de Recherche en Informatique
et en Automatique (INRIA)

Scientific Organisation Committee
E. Diday, M. Jambu, L. Lebart,
J. Pagès, R. Tomassone

Scientific Secretariat
J. J. Daudin, A. Morineau,
J. Quinqueton, A. Schroeder

NORTH-HOLLAND
AMSTERDAM • NEW YORK • OXFORD

ISBN: 0 444 87555 7

Published by:
ELSEVIER SCIENCE PUBLISHERS B.V.
P.O. Box 1991
1000 BZ Amsterdam
The Netherlands

Sole distributors for the U.S.A. and Canada:
ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
52 Vanderbilt Avenue
New York, N.Y. 10017
U.S.A.

Library of Congress Cataloging in Publication Data

International Symposium on Data Analysis and Informatics
(3rd : 1981 : Versailles, France)
Data analysis and informatics, III.

1. Multivariate analysis--Data processing--Congresses.
2. Factor analysis--Data processing--Congresses.
I. Diday, E. II. Institut national de recherche en informatique et en automatique (France) III. Title.
QA278.I56 1981 519.5'35'02854 84-10246
ISBN 0-444-87555-7 (Elsevier)

PRINTED IN THE NETHERLANDS

FOREWORD

Along with the development of **Computer Science**, the collection storage of data arrays is extending in all fields. The aim of **Data Analysis** is to extract from such arrays useful information that can be easily read by the user. For that purpose, it relies on computing techniques and develops methods (using from the simplest to the most sophisticated mathematical theories). Its current dynamism results from the large variety of stored data and of issues raised by the users.

Further, the advent of new technology (micro, mini and parallel computers) and new software tools (relational data-bases, graphic systems, expert systems ...) raises new and interesting problems and increases the importance of the discipline. Because of the potential applications and research issues in the area, **Data Analysis** is taught in an increasing number of schools and universities.

The organizers wish to extend their thanks to all the authors, to the referees who have selected the best contributions, to all the scientific specialists for their participation in the Conference, and to the Public Relations Department of INRIA who actively contributed by its efforts to the organizational success of this meeting.

REFEREES

AGUILAR-MARTIN	J.	(FRANCE)
BADIA		(FRANCE)
BARTKOWIAK	A.	(POLAND)
BELLACICCO	A.	(ITALY)
BERTHOD	M.	(FRANCE)
BOCK	H.H.	(GERMANY)
BOUROCHE	J.-M.	(FRANCE)
BRENOT	J.	(FRANCE)
BURTSCHY	B.	(FRANCE)
CAILLIEZ	F.	(FRANCE)
CARROLL	J.D.	(U.S.A.)
CAUSSINUS	H.	(FRANCE)
CAZES	P.	(FRANCE)
CELEUX	G.	(FRANCE)
CHANDON	J.L.	(FRANCE)
CHARLES	C.	(FRANCE)
CHIFFLET	R.	(FRANCE)
COLLOMB	G.	(FRANCE)
CORMACK	R.M.	(G.B.)
DAUDIN	J.-J.	(FRANCE)
DAUXOIS	J.	(FRANCE)
DELATTRE	M.	(BELGIUM)
DELBOS	M.	(FRANCE)
DELLA RICCIA		(ITALY)
DEPAIX		(FRANCE)
DER MEGREDITCHIAN	G.	(FRANCE)
DEVIJVER		(BELGIUM)
DEVILLE	J.-C.	(FRANCE)
DIDAY	E.	(FRANCE)
DIEBOLT	J.	(FRANCE)
DROUET D´AUBIGNY	G.	(FRANCE)
DUBUISSON	B.	(FRANCE)
ESCOFIER	B.	(FRANCE)
ESCOUFIER	Y.	(FRANCE)
FACY	F.	(FRANCE)
FALGUEROLLES	A. DE	(FRANCE)
FENELON	J.-P.	(FRANCE)
FICHET	B.	(FRANCE)
FLORY	A.	(FRANCE)
FRIEDMAN	H.P.	(U.S.A.)
FRIEDMAN	J.H.	(U.S.A.)
FU	K.S.	(U.S.A.)
GABRIEL	K.R.	(ISRAEL)
GNANADESIKAN	R.	(U.S.A.)
GONDRAN	M.	(FRANCE)
GORDON	A.D.	(G.B.)
GOVAERT	G.	(FRANCE)
GOWER	J.C.	(G.B.)
GRAF-JACCOTTET	M.	(SWITZERLAND)
GUTTMAN		(ISRAEL)
HANANI	U.	(ISRAEL)
HARTIGAN	J.A.	(U.S.A.)
HAYASHI		(JAPAN)
HILL	M.O.	(G.B.)

HUBERT	L.	(U.S.A.)
JAMBU	M.	(FRANCE)
JORESKOG	K.G.	(SWEDEN)
KAMINUMA	T.	(JAPAN)
KERBAOL	M.	(FRANCE)
KOBILINSKI		(FRANCE)
KOULOUMDJIAN	J.	(FRANCE)
KRUSKAL	J.	(U.S.A.)
LAFAYE DE MICHAUX		(FRANCE)
LAURO	N.	(ITALY)
LE CALVE	G.	(FRANCE)
LEBART	L.	(FRANCE)
LEBEAUX	M.-O.	(FRANCE)
LECHEVALLIER	Y.	(FRANCE)
LEFEVRE		(FRANCE)
LEMAIRE	J.	(FRANCE)
LEMOINE	Y.	(FRANCE)
LERMAN	I.C.	(FRANCE)
LEY	J.P.	(FRANCE)
LINGOES	J.C.	(U.S.A.)
LOPEZ DE MANTARAS	R.	(SPAIN)
MACDONALD		(AUSTRALIA)
MAILLES	J.P.	(FRANCE)
MASSON	J.P.	(FRANCE)
MONGET	J.-M.	(FRANCE)
MONJARDET	B.	(FRANCE)
MORINEAU	A.	(FRANCE)
NAKACHE	J.-P.	(FRANCE)
NELDER	J.A.	(G.B.)
NORA	C.	(FRANCE)
OHSUMI	N.	(JAPAN)
OK	Y.	(FRANCE)
PAGES	J.-P.	(FRANCE)
PHILOCHE		(FRANCE)
PICARD	J.	(FRANCE)
POUGET	J.	(FRANCE)
POUSSE		(FRANCE)
RAO	C.R.	(U.S.A.)
RASSON	J.-P.	(BELGIUM)
RIZZI	A.	(ITALY)
ROMIER	G.	(FRANCE)
ROSS	G.J.S.	(G.B.)
ROUSSEAU	P.	(CANADA)
ROUX	M.	(FRANCE)
SANTINI	G.	(FRANCE)
SAPORTA	G.	(FRANCE)
SCHEKTMAN	Y.	(FRANCE)
SIBSON		(G.B.)
SIMON	J.-C.	(FRANCE)
SNEATH	P.H.A.	(G.B.)
SPATH	H.	(GERMANY)
STUETZLE	W.	(U.S.A.)
TALLUR	B.	(ALGERIA)
TENENHAUS		(FRANCE)
TERRENOIRE		(FRANCE)
TESTU	F.	(FRANCE)
TOMASSONE	R.	(FRANCE)
TOUSSAINT	G.T.	(CANADA)
TRECOURT	PH.	(FRANCE)
TUKEY	J.W.	(U.S.A.)
WOLD		(SWEDEN)

TABLE OF CONTENTS

CHAPTER 2 - CLUSTERING

CHAPTER 3 - FACTOR ANALYSIS AND DISTANCE TABLES

CHAPTER 4 - SOFTWARE

CHAPTER 5 - LINEAR METHODS AND DISCRIMINATION

DATA ANALYSIS AND INFORMATICS, III
E. Diday et al. (editors)

PRACTICAL TECHNIQUES FOR AREAL CLUSTERING

Noboru Ohsumi

The Institute of Statistical Mathematics
Tokyo, Japan

The procedures of areal clustering for data analysis using multivariate observations at digitized pixels on the plane are proposed here. These are broadly classified into NTAP and IMAGE, according to the characteristics of given data, the coordinates of the pixels and by how such data are used. Several different ways are provided for user's choice, according to the complexity of the object region. The results obtained from areal clustering can be exhibited on a color display, making it easy to explore and observe the multidimensional knowledge contained in the object region. To evaluate the validity and practicality of procedures, artificial data similar to actual one are generated. The performance and achievement of areal clustering procedures are investigated by the experimental results of processing these data.

INTRODUCTION

The objective of this report is to propose a number of procedures, called "areal clustering techniques", which are effective in "partitioning the object region of interest on the map into several similar areas" using statistical data on that region. It is also difficult to develop a system generally applicable to statistical data analysis because areal partitioning is intended to serve a variety of purposes and data given for these purposes are very versatile. Therefore, main efforts were directed toward developing practical areal clustering techniques with as wide a range of applications as possible by assuming hypothetical conditions for the purpose of areal partitioning and the construction of given data.

CONCEPTUAL BACKGROUND

Actual data are applied to the object region on the map. For the purpose of analysis, however, these data can be condsidered image data on a color graphic display. In other words, the basic data used in areal clustering described here are consisted of multivariate observations at grid-squared or digitized pixels on the two-dimensional plane, although consideration of altitude are occasionally required. Here, the multivariate observations of this kind are considered in two broad categories, described below.

In one category fall the data regularly or consecutively observed at all pixels, as exemplified by the data of remotely sensed measurements and data obtained by mesh or grid systems. The amount of these kinds of data is very large, and even the standard image involves pixels ranging from several tens to hundreds of thousands in number, depending on the size of the object region and the required accuracy of measurement.

In the other category fall the multivariate observations obtained not from irregularly sampled points (though these points are not totally random), but from specific points. Data of this kind are extremely small in amount in comparison with the total number of pixels in the image.

For example, data falling in this category include survey data compiled and aggregated for each spot selected at random by two-stage sampling from the whole area covered by a dwellers' attitude survey, and the data of statistical area, such as census tract, which are applied to the geometrical centroid of the area of concern.

A factor common to these two categories is that multivariate observations at pixels are used. How the pixel coordinate data are handled bears closely on the data analysis, so careful attention should be paid to the basic approach to areal linkage and proximity.

For rough areal partitioning, it is possible to integrate neighbor areas into a single area, and the coordinates of each pixel can be included in the multivariate observations. When any specific, small objectives or small regions are to be selected, coordinates are not necessary. If observation points are given irregularly, then the problem arises as to how to define and delineate the sphere influenced by the information obtained from each observation point.

Data of multivariate observations cannot be utilized to the full by simply delineating the object region or segmenting it by coloring. For more effective utilization of such data, the multidimensional information obtained by areal partitioning should be converted to color image by a suitable coloring system.

This calls for the availability of an easy-to-use color display unit which permits the user to examine a new partitioning method in quick response to the need which he notes while watching the partitioning data of the object region.

The above fact nearly eliminates the possibility that computer performs completely automatically the whole process of areal partitioning, especially because coloring must proceed step by step on the basis of some prior knowledge given about the region. Considering the diversity of purposes of data analysis and the complexity of characteristics of the object region, a number of different cases can be perceived for providing the said prior knowledge, as enumerated below.

(1) Regions relatively uniform or homogeneous (called "typical regions") and a plural number of typical subregions representing each of them can be specified from various statistical data.

(2) Designation of typical regions is difficult, and selection of typical subregions require certain pre-processing of data.

(3) Designation of a large area like typical region is difficult, and observation can be made only at limited points suitably sampled from the object region or suitably arranged in it.

In cases (2) and (3), selection of typical subregions can be made easier by automatic partitioning, even if it is not performed perfectly. If automatic classification is applied in the accumulation of area knowledge and areal partitioning is conducted using such area knowledge, then it can no longer be called simple automatic classification.

For this reason, some procedures proposed in this report are slightly different from the methods of spatical statistics, spatial pattern analysis, and statistical pattern models which have come to be discussed recently in many related fields. Nor are the proposed procedures identical to the methods of usual automatic classification or cluster analysis. This is why the author named these procedures "areal clustering techniques" (abbreviated to AC- techniques).

AC-techniques has the following two phases, which are intended to maintain compatibility with the aforementioned characteristics of the object region.

Phase I: NTAP (Numerical Techniques for Areal Partitions)
NTAP is an areal partitioning system using primarily the data given regularly at pixels of the object region, such as mesh data and remotely sensed data.

Phase II: IMAGE (Image Generating Techniques)
IMAGE is a sort of image generating system using data obtained from irregular sampling points, such as dwellers' attitude survey data and environmental data.

OBJECTIVE OF NTAP

The data covered by NTAP are the multivariate observations given regularly at all pixels on the plane, and have performed suitable image processing. Data analysis by this system is made easy or difficult according to the complexity presented by the object region and the extent of mixing of its characteristics. Application of NTAP may be considered for the following cases:

(1) When the object region is classified into a number of areas which are not very intricate in shape and are expected to present relatively uniform characteristics. For example, if an extensive and flat farmland area is classified by the kind of crops, each sub-area will be relatively large in size and homogeneous.

(2) When each region is relatively large, as in case (1) above, but shows a certain degree of complexity or mixing of characteristics. For example, if a large area is classified roughly, with LANDSAT data analysis, the sub-regions such as urban, forest, and farmland areas present different characteristics.

(3) When small areas with specific characteristics or objectives are to be selected from the region, for example, selection of roads, rivers, or specific air-polluted areas.

(4) When homogeneous regions can hardly be found and areal clustering is desirable, according to the degree of mixing of objectives. For example, if an urban area is classified into green zone, housing district, high-rise building district, and so on. Besides these objectives are mixed with scattered, such that the state can be expressed by the term "texture."

Application of NTAP in the immediate future is considered only for cases (1) and (2) because of the difficulty in developing a system generally applicable to cases (3) and (4). Cases (3) and (4) call for very individual approaches. Case (3) demands that geometrical conditions are strictly assumed for the objectives and they should be distinguished from the peripheral areas, while case (4) requires even more troublesome processes.

STRATEGIES AND CLASSIFICATION OF NTAP

Areal partitioning by NTAP is intended not just for delineation of areas or coloring of partitioned areas. Clear delineation or coloring is of no great importance, and it is often the case that ambiguous delineation is left as it is or areas with specific characteristics (e.g., areas made up of built-up districts only) are selected for reclassification. To deal with the varying judgment and demand of users in areal partitioning, the algorithm of the NTAP is divided into the following two cases according to whether or not several typical subregions can be previously defined within the object region.

The first is a case that there is certain prior knowledge which makes it possible to designate small areas in a given area (such as area A, area B, ...). It is desirable that these small areas are scattered extensively in as wide an area as possible. If typical subregions are given, the data of their pixels are subjected to multivariate analysis and the remaining pixels are classified on the basis of such analysis. One of the key factors of this procedure, which is described later, is how the coordinates of the pixels are utilized.

It is the second case that there is no reliable prior knowledge for selecting typical subregions, and ordinary automatic classification using available data is performed. In this case, there are many different methods of automatic classification, but here we employ ISODATA (Iterative Self-Organizing Data Analysis Techniques-A) because it is suited to efficient classification of voluminous multivariate observations and is also used extensively in the other analysis of remote sensing data. k-means method and Ward's method are also capable of processing large volumes of data, but the ISODATA clustering procedure

is preferable because it exhibits the basic functions of the other methods and also displays the function of adjusting the size and number of clusters as well as repetitive cluster lumping or splitting.

One shortcoming of this procedure is that the large number of parameters to be given in advance makes it somewhat difficult to interpret the results of calculation. To compensate for this shortcoming, actual computer programming must be worked out with care taken to ensure ease of parameter change, repetitive use of the same data, and graphical presentation of the results. Areal partitioning strategies by NTAP are framed as follows, with consideration given to the points mentioned above.

Strategy I: This strategy is applicable when a plural number of typical subregions can be specified and sampled to represent each typical region regarded to be relatively unifrom or homogeneous. In this case, the typical region is classified in one of the following two ways:

(a) On the basis of statistical analysis of data within typical subregions, all pixels are allotted to typical regions most closely resembling each other to classify the regions (such as A, B, C,...).

(b) The typical regions are firstly classified into A, B, C, ... (primary classification) and one of them, A for example, is further finely reclassified into A_1, A_2, A_3, ... (detailed reclassification).

Strategy II: This strategy is employed to generate regions corresponding to typical subregions when it is difficult to choose uniform typical subregions.

(a) A plural number of regions likely to have characteristics in which the user is interested are designated, and all data contained within these regions are integrated. Provisional partitioning of selected areas is then conducted by automatic classification using the integrated data. Designation of typical subregions is made on the basis of this provisional areal partitioning. This is followed by the classification of typical regions mentioned in Strategy I above.

(b) All data within the object region are classified into some groups by automatic classification, and typical subregions are selected on the basis of results obtained from the classification.

Then all pixels are classified by the nearest neighbor rule according to the knowledge obtained by multivariate analysis using data within typical subregions selected above. One of the features of this algorithm is that the user can determine whether to use the coordinates of pixels by his own free choice, and he can also determine the distance for classifying the data set and the procedure of classifying the pixels by his own choice.

(1) Distance for classifying
Let $\underset{\sim}{z} = (u, v; x_1, x_2, \ldots, x_p)$ denote the data of a pixel, where (u, v) represents the coordinates of the pixel and $\underset{\sim}{x} = (x_1, x_2, \ldots, x_p)$ are the p-th dimensional multivariate observations at (u, v). Then we consider two kinds of distance as follows:

Mahalanobis' distance

$$d_k = (\underset{\sim}{x} - \underset{\sim}{m}_k)' W_k^{-1} (\underset{\sim}{x} - \underset{\sim}{m}_k) \qquad (1)$$

where, k = A, B, C, D, ... (symbols representing typical regions),
$\underset{\sim}{m}_k$ = mean vector of multivariate observations obtained from the data within the k-th typical region,
W_k = sample variance-covariance matrix of the data within the k-th typical region W_k^{-1} is the inverse of the matrix W_k).

Normalized l_1-norm
The l_1-norm, normalized by the standard deviation of each variable, is expressed as follows:

$$d_k = \sum_{i=1}^{p} |x_i - m_{ki}| / \sigma_{ki} \qquad (2)$$

where, x_i = i-th element of the vector $\underset{\sim}{x}$,
m_{ki} = i-th element of the vector $\underset{\sim}{m}_k$,
σ_{ki} = i-th element of the vector $\underset{\sim}{\sigma}_k$,
$\underset{\sim}{\sigma}_k$ = vector of the standard deviations of the data within the k-th typical region.

(2) Procedure for classifying pixels
A set of pixels is classified by the following procedure:

(Step 1) The threshold δ is specified as a criterion for determining a typical region to which pixels to be classified belong, and then the typical region k where $d^* = \min_k d_k$ is searched for.

If $d^* \leq \delta$, then it is determined that the observation vector $\underset{\sim}{x}$ belongs to the typical region k.

If $d^* > \delta$, one of the following two judgments is made according to which of the two distances shown above is used:

a) The observation vector $\underset{\sim}{x}$ is judged "unclassified" for Mahalanobis' distance.
b) For l_1-norm, the coordinates (u, v) of the pixel in question are added to the characteristic vector by Step 2 (reclassifying procedure), described below.

In the case of a), pixels not belonging to any typical region remain unclassified (or may be considered part of a region defined as unclassified). In the case of b), the vector $\underset{\sim}{x}$ is certain to be judged part of some typical region or other.

(Step 2) The reclassifying procedure can be expressed as follows:
d** satisfying the following conditions is selected, and the vector $\underset{\sim}{x}$ is determined to belong to the typical region k.

$$d^{**} = \min_j \{\min_k (d_k + \Delta_{k_j})\} \qquad (3)$$

where k = A, B, C, ... (symbols representing typical regions).
k_j = j-th subregion in the typical region k.

Δ_{kj} is the correction term for taking into account the data of coordinates, and it is defined as follows.

$$\Delta^{(1)}{}_{k_j} = \frac{d_u}{l_u} + \frac{d_v}{l_v}$$

$$\Delta^{(2)}{}_{k_j} = \frac{d}{a} = \frac{d}{\sqrt{l^2{}_u + l^2{}_v}} \qquad (4)$$

Symbols in these expressions are shown in the right Figure 1. The point P is placed arbitrarily within the small typical region k_j. If the location of P is not specified, the center of gravity of the area k_j is automatically given as a default value. The user is free to select any one of the correction terms.

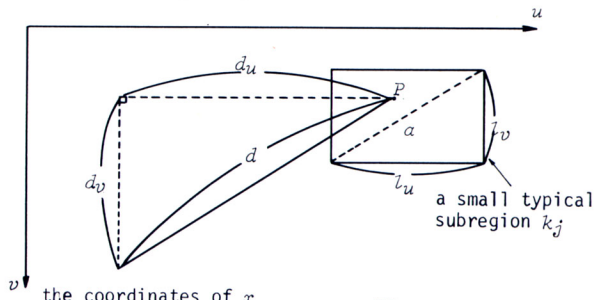


Figure 1

OBJECTIVE OF IMAGE

In case multivariate observations are obtained from points irregularly given in the object region, and their amount is small in comparison with the total number of pixels, it is difficult to give areas like typical subregions in advance or to generate such areas automatically. In this case, areal partitioning is not advisable. It is more effective to use limited multidimensional knowledge at pixels for which observations are available to generate multidimensional data at other pixels by a sort of interpolation. How to use multivariate observations and pixel coordinates is just as important with IMAGE as it is with NTAP. The principal objective of IMAGE is to generate information covering the entire plane by interpolation using limited data at a small number of pixels, and to produce color images representing the area characteristics in a very rough way. For this reason, IMAGE can be considered a sort of image generating system, and classification by IMAGE is intended to give multivariate observations to all pixels and present them as a color image on a display unit for easy observation. The IMAGE procedure is as follows:

(Step 1) Let $\underset{\sim}{z} = (u, v; \underset{\sim}{x})$ denote a data set on the plane, where (u, v) represents the coordinates of the pixel and $\underset{\sim}{x}$ are the p-th dimensional multivariate observations at (u, v). Furthermore, let (u^*, v^*) denote the coordinates of the pixels to be classified.

(Step 2) The data set excluding the coordinates is classified into L groups by using suitable automatic classification methods. Several classification methods are made available, including, in particular, the MST (Minimal Spanning Tree) and the modified k-means method. The mean vectors $\overline{\underset{\sim}{x}}_l$ $(l = 1, 2, \ldots, L)$ and other statistics are calculated for each cluster.

(Step 3) The result of clustering is exhibited on a graphic dispaly unit, and several locates (or representative pixels) are chosen in each cluster, for example, applying manually the crosshair cursor at each locate. Then, how to specify the positions and numbers of locates should be carefully considered on the basis of the distribution and density of them. The following modes of choosing locates are possible:

(a) Specify any locate within the area covered by a cluster
(b) Determine a suitable number of observation points by sampling initially given points
(c) Compute the centroid for each cluster and specify it as the locate
(d) Use the locates thus selected by suitably combining them.

The number of locates is determined in the following manners:

(a) Sample locates from each cluster in proportion to the cluster size
(b) Allot all initially given observations to the locates.

(Step 4) After adjusting the position of each selected locate, the mean vector of the cluster which includes a locate is added to the coordinates (u^*, v^*) as the multivariate characteristic vector.

(Step 5) The multivariate characteristics vectors of all pixels to be classified are generated by interpolation, and the coloring plane is constructed on the basis of the generated vector.

To explain this process in more detail,
(a) Find K locates nearest to and associated with a pixel j for smooth interpolation, and let (u_k, v_k) $(k = 1, 2, \ldots, K)$ denote their coordinates.
(b) The multivariate characteristics vector for the pixel j, generated by interpolation, is given as follows:

$$\underset{\sim}{\mu}^*_j = \alpha \left[\frac{1}{K} \sum_k^K \overline{\underset{\sim}{x}}_{(k)}\right] + (1 - \alpha)\, \underset{\sim}{\mu}_j \qquad (5)$$

where, α = weight parameter $(0 \leq \alpha \leq 1)$.
$\overline{\underset{\sim}{x}}_{(k)}$ = mean vector of the cluster which includes the locate k, and

$$\underset{\sim}{\mu}_j = \sum_k^K w_k\, \overline{\underset{\sim}{x}}_{(k)} \Big/ \sum_k^K w_k$$

where, w_k is the weight factor, usually in inverse proportion to the measures of distance which are Minkowski's metrics, that is,

$$w_k = 1 \Big/ \{(u^* - u_k)^r + (v^* - v_k)^r\}^{\frac{1}{r}} \quad (r > 0).$$

In the above expressions, α becomes the simple moving average of the mean vectors on K's nearest neighbor locates when $\alpha = 1$, and interpolation is performed with account taken of the effect of distance between the pixel and each locate when $\alpha = 0$. In other words, α is a kind of parameter for adjusting the smoothness between pixels.

(Step 6) The characteristics vectors generated by the procedure mentioned in Step 5 above are added to the coordinates (u^*, v^*) of the pixels, and we have $(u^*, v^*; \underset{\sim}{\mu}^*)$ for all of them. The vector $\underset{\sim}{\mu}^*$ thus obtained is allotted at any dimension within the RGB (red, green, blue) gamut and transformed into a color image by applying an appropriate coloring system to the vector $\underset{\sim}{\mu}^*$.

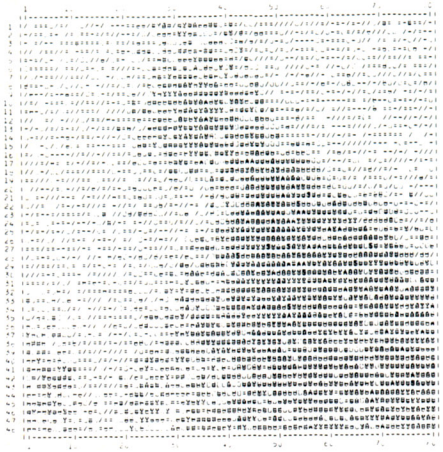
(Step 7) By the processes mentioned above, the vector $\underset{\sim}{\mu}^*$ is converted to an integer vector and transformed to color image $\underset{\sim}{m}^*$. When the $(u^*, v^*; \underset{\sim}{m}^*)$ thus obtained are input to a color display unit, the cluster is presented in smooth, "cloud-like" color gradation. The color image is suitably adjusted and the shading maps by overprinting are output according to need on the line-printer.

EXPERIMENTAL RESULTS AND BRIEF DISCUSSION

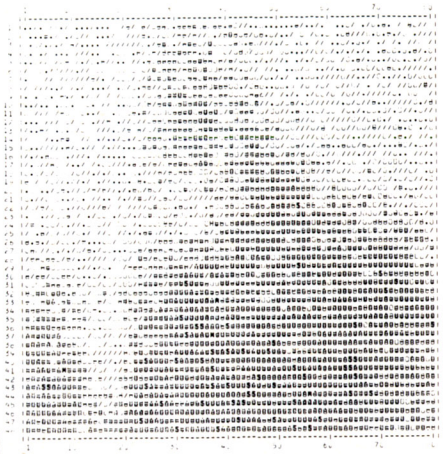
In the actual application of areal clustering techniques, the characteristics of the entire area in question are estimated on the basis of limited data of the typical regions or locates. For this reason, it is necessary to extract a part of the data with known construction to determine the reproducibility of the whole data construction by areal clustering. This process is also required to check the suitability of the computer program design and algorithm. Artificially generated data are required to be as similar to actual data as possible. Consideration must be given to the multivariate observations at pixels (i.e., correlations between the variables) as well as to the need to satisfy the actual mixed condition of a number of areas which are similar and vaguely delineated. To meet this condition by the use of pseudo-random numbers, a plural number of data sets have been constructed (see reference [6] for the method of artificial data generation).

These have been constructed as three-dimensional characteristics in which about four groups are ostensibly mixed in a suitable manner to make the data sets compatible with the number of pixels in currently used color graphics. Figure 2 shows a data set shading maps for each dimension. When they are implemented on a color graphic display, they can be seen visually as a color image.

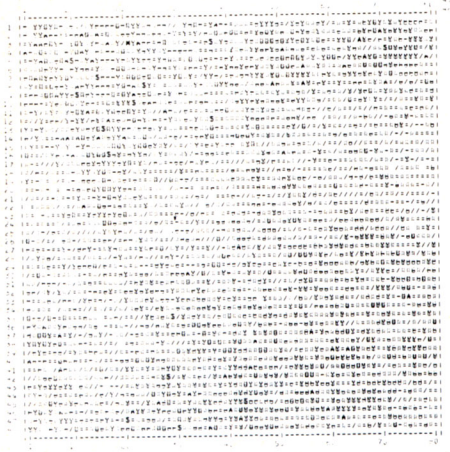
(1) NTAP Experiment
Four typical regions (A,B,C, and D) were assumed with their subregions specified as shown in Table 1. The positions of each subregion are given as illustrated in Figure 3. In addition, l_1-norm is taken as distance for classifying and using a value for δ of 3.0. Table 2 presents the classified groups obtained by NTAP. The 4 x 4 cross-classified table, which is placed in the left side of Table 2, shows the result how the whole pixels within each typical region were reallocated to four typical regions. That is, the pixels in the off-diagonal part of this table indicate that they were decided to belong to the different regions from initial ones. In columns (a) and (b) of Table 2, each value indicates the result classified into four groups. The upper value of each row indicates the number of pixels decided directly by using d^*, and the lower values are the number of pixels applying correction term to the d^* values. Furthemore, Figure 4 illustrates the classified map obtained from NTAP, using typical subregions in Table 1. It can be clearly observed that original image given in Figure 2 is recovered.



(a) 1st dimension



(b) 2nd dimension



(c) 3rd dimension

Figure 2 Artificial data set constructed by generating mixtures of three dimensional normal distributions, using random numbers.

Table 1. Typical regions provided for experiment of NTAP

Typical regions	Typical subregions 1	2	3	4	Totals
A	600	204	272	150	1266
B	128	380	252	-	760
C	187	180	110	153	630
D	156	208	228	260	852
					(3008)

Table 2. Classified table obtained by applying NTAP to the artificial data shown in Figure 2

Typical Regions	A	B	C	D	(a) Allocated Pixels	(b) Totals		(c) Totals	
A	734	1	69	4	10889	11697	0.1904	15269	0.2485
	271				3301	3572	0.0581		
B	1	480	4	126	10565	11176	0.1819	15001	0.2442
	1	163			3661	3825	0.0623		
C	217	13	423	9	12875	13537	0.2203	16500	0.2686
			128	1	2834	2963	0.0482		
D	2	103	6	534	10967	11612	0.1890	14670	0.2388
				178	2880	3058	0.0498		
Totals	954	597	502	673	45296	48022	0.7816		
	272	163	128	179	12676	13418	0.2184		
Totals	1226	760	630	852	57972	61440	1.0000		

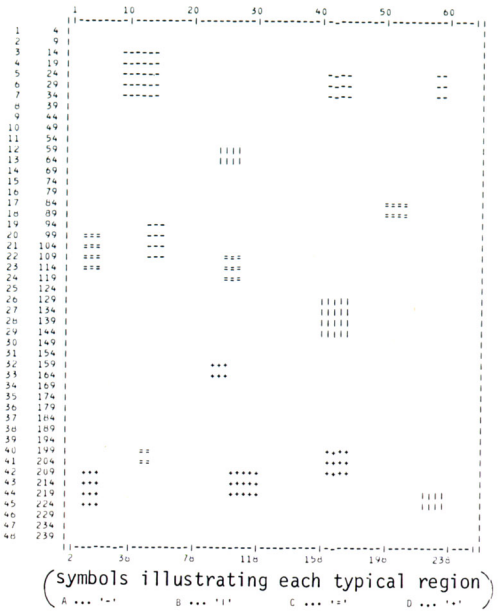


Figure 3 Four typical regions windowed from the digitized object area.

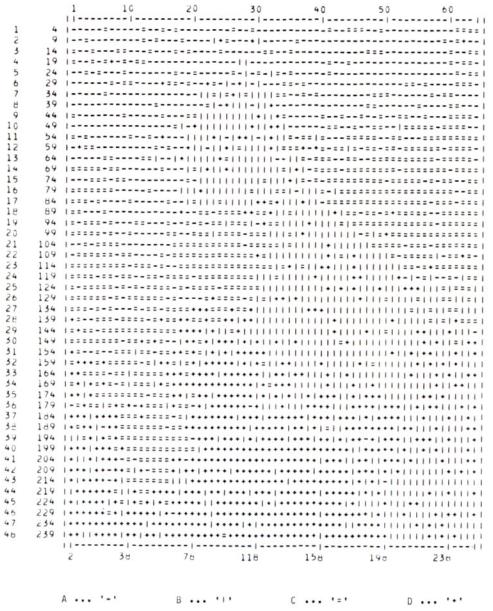
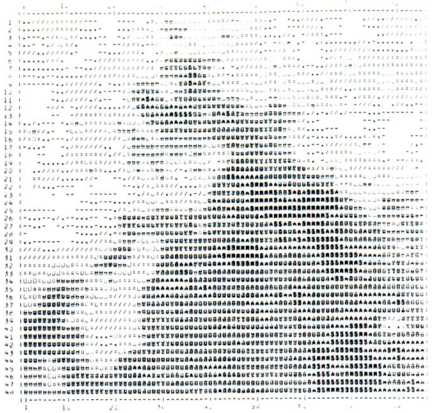
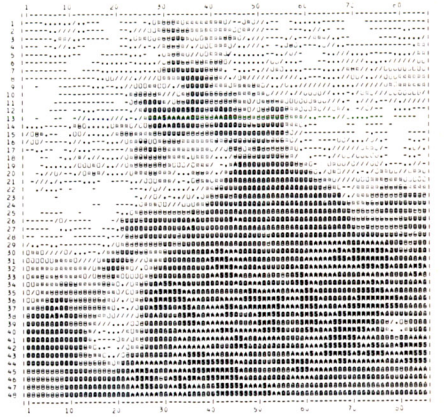


Figure 4 Showing four groups generated by NTAP using windows shown in Figure 3.



(a) 1st dimension



(b) 2nd dimension

(c) 3rd dimension

Figure 5 Shading maps reconstructed by IMAGE from irregularly sampled locates (500 locates), produced by overprinting.

(2) IMAGE Experiment

A group of 500 pixels was sampled at random from the same data set used in NTAP experiment above, and this was followed by image generation by the IMAGE procedure. Figure 4 shows one of the images generated. Supposed conditios were $K = 6$, $\alpha = 0$, weight parameter of distance $r = 2$, that is, the squared Euclidean distance. k-means method is given as clustering techniques and let $L = 10$ for the number of clusters. Figure 5 is a sort of tentative picture imaged ronghly from the small-scale observations at the limited points, however, the original feature is nicely reconstructed.

In all experiments thus far conducted, it was found that reproducibility was highly satisfactory considering the limited size of data used, indicating that practical applications of AC-techniques are very possible. Furthermore, to evaluate the performance of AC-techniques, we applied it to part of actual data sets collected in the past, and exhibited the classified results of the entire region on the display, which indicated that the initial image of the region could be faithfully reproduced. In future, we are planning to apply AC-techniques to the attitude survey data of urban dwellers in order to explore a policy for revealing the relationship between actual attitude data and areal environmental data.

ACKNOWLEDGEMENTS

The author is grateful to Professor K. Mizuno who introduced me to this project and made useful comments and, moreover, many thanks are owed to Professor M. Sibuya for his valuable ideas on this work.

This research was partially supported by the foundation for National Research Institutes of Government Ministries and Agencies coordinated by Environment Agency.

REFERENCES

[1] Ahuja, N. and Schachter, B.J., Pattern Models (John Wiley, 1983).
[2] Barnett, V. (ed.), Interpreting Multivariate Data (John Wiley, 1981).
[3] Batchelor, B.G., Practical Approach to Pattern Classification (Plenum Press, 1974).
[4] Davis, I.J. and McCullagh, M.J. (eds.), Display and Analysis of Spatial Data, Nato Advanced Study Institute (John Wiley, 1975).
[5] Foley, J.D. and Dam, A.V., Fundamentals of Interactive Computer Graphics (Addison-Wesley, 1982).
[6] Ohsumi, N. and Sibuya, M., Numerical Techniques for Areal Partitions: NTAP (in Japanese), The Proceedings of the Institute of Statistical Mathematics, 25, 1 (1978) 41 - 63.
[7] Pavlidis, T., Algorithms for Graphics and Image Processing (Computer Science Press, 1982).
[8] Ripley, B.D., Spatial Statistics (John Wiley, 1981).