MICROCOMPUTER GRAPHICS CLUSTERING TECHNIQUES FOR URBAN ENVIRONMENTAL EVALUATIONS

Noboru Ohsumi and Kinji Mizuno

Institute of Statistical Mathematics
4-6-7, Minami-Azabu, Minato-ku, Tokyo, JAPAN

SUMMARY

The relationship between environment and dwellers' attitudes was visualized with colored maps presented on a color graphics display, based on survey data of urban dwellers' attitudes toward environmental concerns. This paper presents the areal clustering techniques and the associated computer-aided processing system which together seem to provide an effective aid to understanding the characteristics of area environments. The paper first provides an outline of the areal clustering techniques and the features of the microcomputer system and color graphics terminal which were developed to support the techniques. The statistical data processing techniques required to use the system are then discussed. Some examples of the application of areal clustering to actual attitude survey data are presented to demonstrate the effectiveness of the microcomputer system and color graphics for this type of study.

1. Microcomputer-Aided Areal Clustering

With the increasing accumulation of many kinds of statistical information on districts, there is a significant progress in computer graphics techniques for exploratory analysis of these types of information. Included in these techniques are the analysis of data from remotely-sensed measurements or the mesh system and computer-aided cartography.

Another type of important geographical information is district information comprising various social indicators or characteristics of urban dwellers' attitudes, as presented on maps used for the analyses of urban structures or environmental problems. In general, a clearer understanding of the characteristics of an environment or the dwellers' attitudes is obtained when the standpoint is selected at a level which provides a broad view of the features of interest. To do this, it seems effective to partition the objective region into several similar areas and to visualize the characteristics of each area. The conventional approach to such analyses are the multivariate analysis techniques such as automatic classification or discriminant analysis. These approaches alone, however, seem to be insufficient to fully represent the characteristics at each point located within the objective region as the overall area information.

In this paper, we propose the areal clustering techniques as an exploratory technique for drawing out characteristics which are hidden in data as latent information, prior to detailed quantitative analysis of urban environments or urban dwellers' attitudes. The most prominent characteristic of these techniques is the use of a microcomputer system with a raster color graphics display as an aid to observing and interpreting the results of analyses. Most microcomputer-based color graphics are now used in many fields for visualizing data description in fairly conventional ways. For example, these methods include graphical representations (e.g., bar graphs, pie charts, radar charts), computer tomography, pseudo-color presentation of remotely-sensed data (Beatty, 1983).

Keywords: Areal clustering techniques; Colored attitude maps; Urban dwellers' attitude survey; Microcomputer-aided color graphics; Urban environmental evaluations

While the techniques and system proposed in this paper are an extension of these applications, they provide greater effects than conventional techniques of drawing or painting statistical information, in the sense that they create "colored attitude maps" based on the results of urban environment evaluations using dwellers' attitude survey data. The areal clustering techniques proposed here consist of sequences of procedures and a supporting computer system which convert multivariate characteristics at locations distributed irregularly and discretely over an objective region of interest into a map of areal information spreading over the entire objective region. The areal clustering techniques basically use a combination of the automatic classification and distance weighted interpolation techniques to produce areal partitions. The features of the areal partitions created by areal clustering are then represented as three degrees of continuous color information: hue, saturation, and brightness. Finally, the areal partitions data is converted to appropriate color-intensity information for display or a specific computer color graphics system. The result is an "colored attitude map", which presents an easy-to understand, visual representation of the relationship between dwellers' attitudes and features of interest as continuous "cloud-like" color gradations on a map of the region. We have applied the term areal to these clustering techniques because they use multivariate characteristics observed at locations on a map to generate areas of similar features.

Clearly presenting differences and similarities in dwellers' attitudes in different areas as colored attitude maps also requires processing before and after areal clustering, for example, editing of data acquired from individual respondents into a form acceptable for areal clustering or re-editing an attitude map for a specific application. Also important for the effective application of areal clustering, is an overall system design policy for functional allocation of computer system resources.

In the remainder of this paper, we would like to describe the implementation of a system using the areal clustering techniques and examine its practical use in analyzing surveys conducted in 1983 on urban dwellers' attitudes toward environmental concerns. Appendix 1 is a summary of the surveys. Appendix 2 is a summary of previously published studies on the areal clustering techniques using experimental data (Ohsumi and Sibuya, 1978; Ohsumi, 1983).

2. Areal Clustering System

2.1 Hardware Configuration

In analyzing attitude survey data using the areal clustering techniques, the following capabilities must be available as needed:

(i) Aggregating the survey data,
(ii) Executing areal clustering,
(iii) Mapping of the objective region and integrating geographical data for the region,
(iv) Editing of color images.

One possible approach meeting these requirements would be to link a microcomputer to a host system and use the microcomputer as an intelligent terminal. For example, data aggregation and areal clustering handle large volumes of data and require considerable processing time and are performed by the host computer. Whereas observation or editing of mapped or district color images require real-time response and are allocated to the microcomputer, color graphics terminal, and other image processing peripherals. Thus, the microcomputer is sometimes used as a terminal for the host system and sometimes as a stand-alone processor for graphics data processing. Figure 1 shows a block diagram of the system hardware. The handlers and associated software linking these devices and processors were developed exclusively for areal clustering.

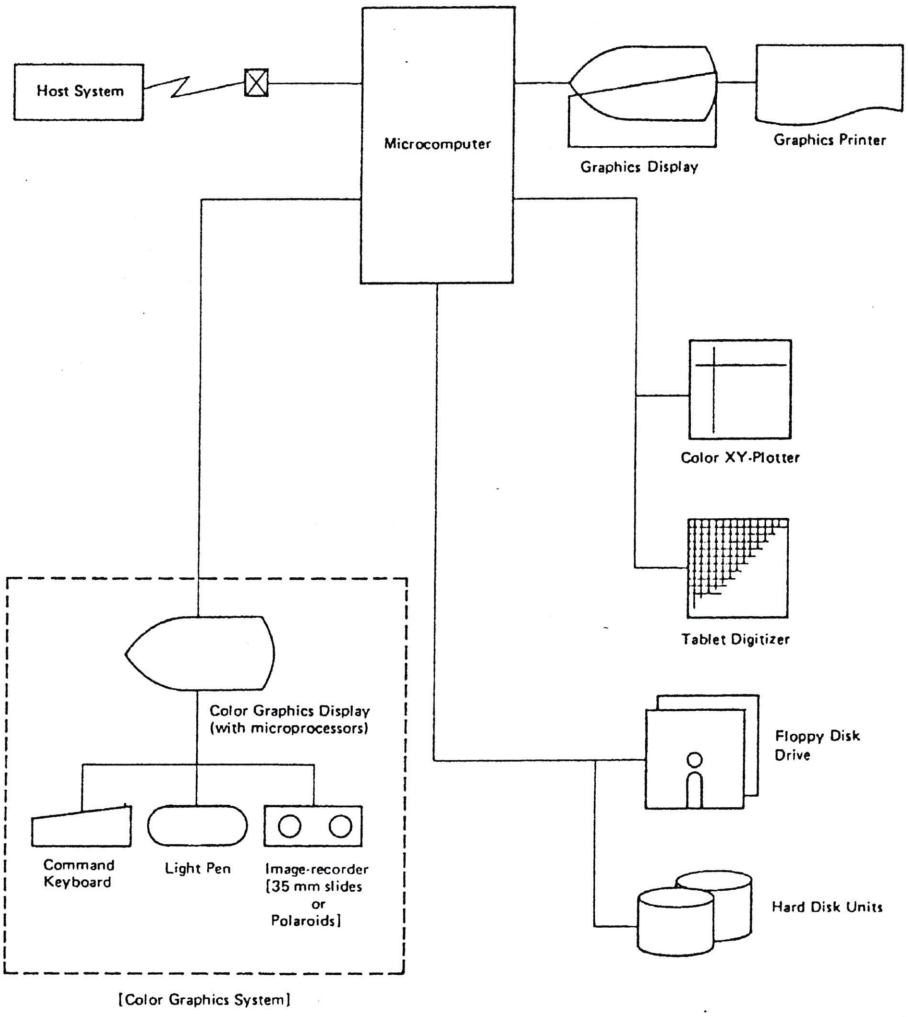

Figure 1 Microcomputer System for Areal Clustering

2.2 Software Features

To supplement the areal clustering techniques, the following software utilities were developed.

Data aggregation

Data handled in areal clustering is characterized by multivariate characteristics that include location information. However, dwellers' attitude data obtained through a sampling survey consists of data from individual respondents sampled by two-stage sampling. Such data must be converted into a form usable in areal clustering without losing any tendencies in the data. The optimum design of the preliminary data processing for this purpose is quite important.

(i) District representative points

Respondents are more or less randomly distributed in sampling small areas. While it is possible to include location information on each respondent and use each as a separate representative point, the method is not very practical because it requires too much labor. In addition, the purpose of the analysis is to determine dwellers' attitudes toward the region in question, not toward their immediate neighbors. Thus, in the present study, we determined the geometrical centroid of the district (e.g., town or block) that includes the sampled small areas, and defined the centroid as the "district representative points".

(ii) Data aggregation procedures

These procedures aggregate the response data from individual respondents into a

form that represents the district representative point. It is preferable that similarities or differences in dwellers' attitudes in different regions or sampling small areas and geographical separation between two or more districts be taken into account at this time. Changes in processing conditions may also be needed due to the type of questions used or the objective regions. To meet this need, we designed the procedures so that they can be interactively used from the terminal.

Preliminary processing: The questions are classified into several groups of related questions. Next, they are scaled for each group by one of several scaling methods. Also recoding of categories is done as needed.

Aggregation: Averages as classified by questions and sampling small areas are calculated by using the set of individual data. Other statistics such as standard deviation are calculated to observe differences between and within districts, and to clarify the characteristics of each sampling small area and each question.

Examing the relationship between questions: Coefficients of associations between questions are calculated by using all of the individual data. Next, the correlation coefficients among questions are calculated by using the average of each sampling small area. The two sets of coefficients are compared to examine the similarity of inter-relationship among the small sampling areas before and after aggregation. At this time, automatic classification techniques are used as an aid to classifying the set of questions into groups.

Coordinate assignment to averages: To established the correspondence between the average of each sampling small area, the geographical centroid of the objective district is defined as the representative point, and both are joined.

Reduction of correlation among questions: When a set of similar questions are used, the correlation between questions is large. This may cause the loss of discernible color gradations in the display image, leaving only extreme intensity values. To avoid this, the composite scores of the averages are calculated using the correspondence analysis or principal component analysis. This also contributes to improved reliability of the scores of similar questions. In actual application, similar questions may be composed, or sometimes, the averages are used to observe the characteristics of individual questions. In either case, the procedures must be capable of responding quickly.

Map creation and display

In general, the colored attitude maps of the regions obtained by areal clustering are overlaid with a map of the major geographical features of the regions (traffic network, major buildings) in order to compare the relationships more fully. To do this, the development of map creation and display techniques are needed.

(i) Physical map entry

A tablet digitizer is used to directly read boundaries between districts from a physical map of the objective region.

(ii) Editing of boundary values and calculation of district representative points. The data from the previous step is edited into a boundary values file. This is transferred to the host system, and is subjected to adjustment for the color display screen size and the real clustering images. The combined data is stored in a map file and transferred to the microcomputer for storage on disk. Other district information (traffic network, land use conditions, geography, etc.) is input in similar way when required and stored as a district data file.

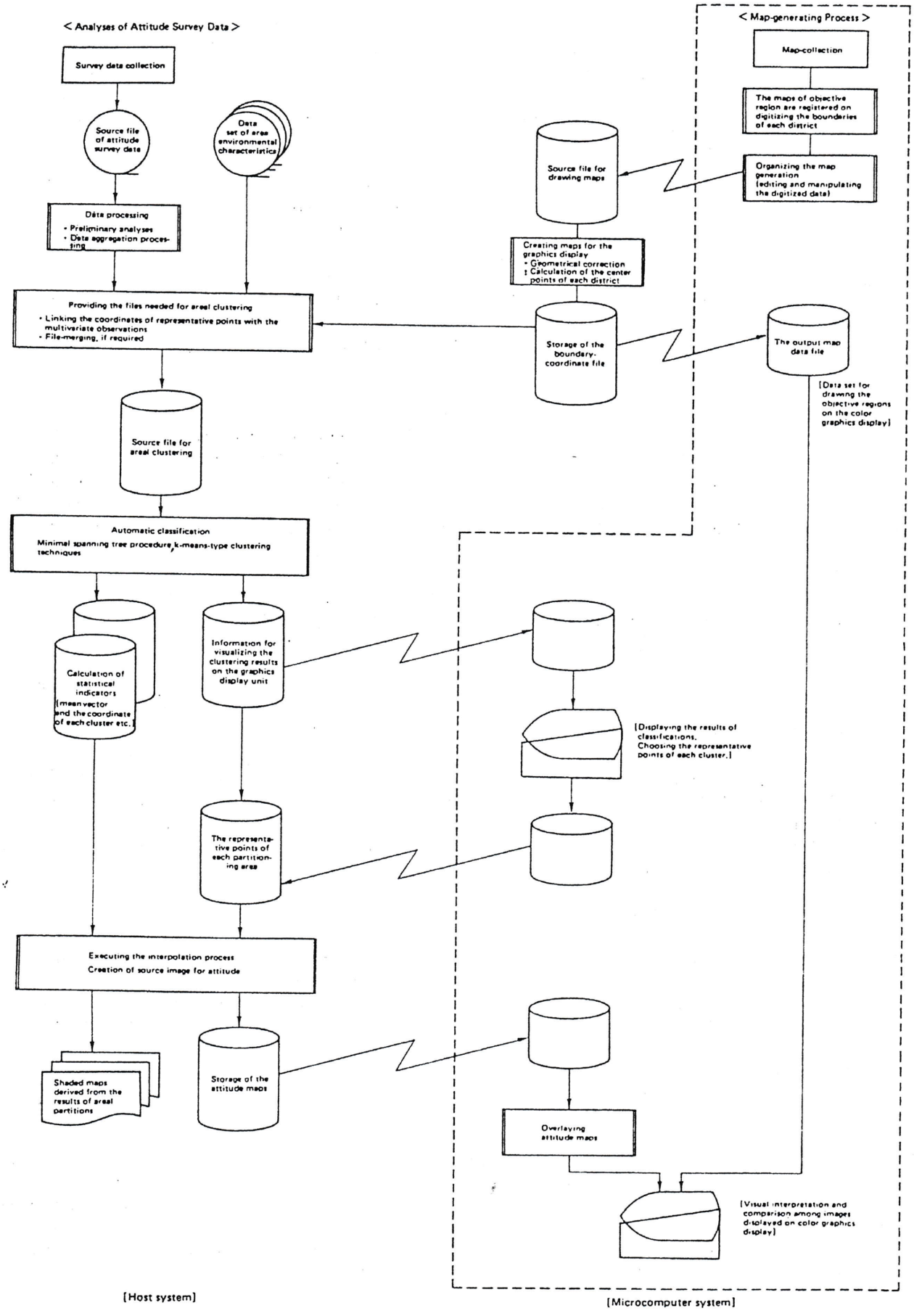

Figure 2 Areal Clustering System

2.3 Execution of Areal Clustering

Areal clustering is applicable to data created by the aggregation procedure described above. Figure 2 shows the major process flowcharts for areal clustering. As shown in these figures, processing which requires considerable processing time, such as automatic classification or image generation, is performed by the mainframe computer, whereas the manipulation of classification results is performed by the microcomputer. Thus, processing functions are distributed for higher efficiency. To execute automatic classification or areal partitioning, the user enters and receives necessary commands in an interactive manner. To observe a clustering map for classification results, the user uses the display unit as a graphics terminal of the host system.

One of the difficulties in image creation lies in the fact that each district must be visually discriminated in natural color gradations. In many cases, however, it is difficult to find a satisfactory color combination and easier to let the computer to automatically generate one. To solve this problem, we use the mainframe computer to do areal partitioning, assign a temporary color scheme, then transfer the areal partitioned color image obtained to the microcomputer. While observing the transferred color image on the graphics display, the user edits the image interactively. This function is ideally suited for the microcomputer and is one of the advantages of the areal clustering system.

2.4 Color Image Editing

The design of the color image edit program and color graphics system is an important factor in full realization of the advantages of the areal clustering techniques. To perform satisfactory color image processing on a microcomputer, raster color graphics built-in microprocessors are most effective.
The advantage of this type of terminal is in the fact that flexible and efficient image editing can be provided with adequate programming. In particular, the following capabilities are needed to provide color image displays which allow the selection of color combinations that are intuitivaly meaningful with respect to the results of attitude surveys.

(i) Color editing

Comparison of image data from different objective regions is possible only when the color ranges and attributes can be adjusted to the actual data distributions.

(ii) Color modification

It is possible to sample any location (pixel) on an image and obtain the color information for that point or to smooth gradations between adjacent similar districts. In addition, the centroids of regions, districts, and subdistricts can be automatically calculated as needed. The resulting data is stored in files and accessed for image editing.

(iii) Image comparison

It is possible to overlay and compare the characteristics of several colored attitude maps obtained for the same region under different conditions, or compare attitude data with areal environmental survey information such as distributions of noise or pollution claims and land use conditions.

3. Examples of Areal Clustering Application

The following shows some excerpts of the knowledge obtained from many analyses. In the following descriptions, the shaded maps that represent color intensity and gradation are used for classifications since it is not possible to include color images with this paper. The data used in the analyses were obtained through the survey listed in Appendix 1.

3.1 Example 1

Figure 3 represented as a shaded map is an example using a single question, "Do you feel there is much greenery around here?" This question is one of those

questions to which responses differ greatly depending on the district. The color graphics clearly indicates this. This example uses the average of each district obtained in Senri New Town. Green color was selected as intuitively meaningful and the brightness of green represents the dwellers' feeling to the amount of greenery: a bright section indicates an area where dwellers feel that there is much greenery around. When those results are compared with a map or aerial photos of the pertinent districts, a weak correlation is observed in terms of the degree of concentration of houses, parks, green lots, and so on. When the results are compared with land use data obtained from a survey, however, the correlation is not very significant and the results seem to reflect dwellers' sense of well-being and satisfaction rather than their physical environment.

3.2 Example 2

In this example, four questions concerning "accessibility" or "convenience" were composed by using correspondence analysis.

Questions

"Do you feel any danger while walking in your neighborhood at night?"
"Do you feel any inconvenience in daily shopping at local shops?"
"Do you feel the closest station or bus stop is too far?"
"Do you have any complaints about the availability of emergency hospital care or after-hours doctor's treatment?"

In this case, the continuity between districts is less than that in the case of Example 1. Figure 4 illustrates one of the shaded maps obtained in Senri New Town and corresponds to the 3rd factor in the three factors calculated by correspondence analysis. On the color graphics display, these three factors are alloted at any color within the three colors, respectively, and composed as a colored attitude map. In addition, observing each distribution of the composite scores as histogram, we can notice the differences in variance and the existence of modals. These pieces of information such as histograms and shaded maps can be effectively used by observing them with reference to the colored attitude maps.

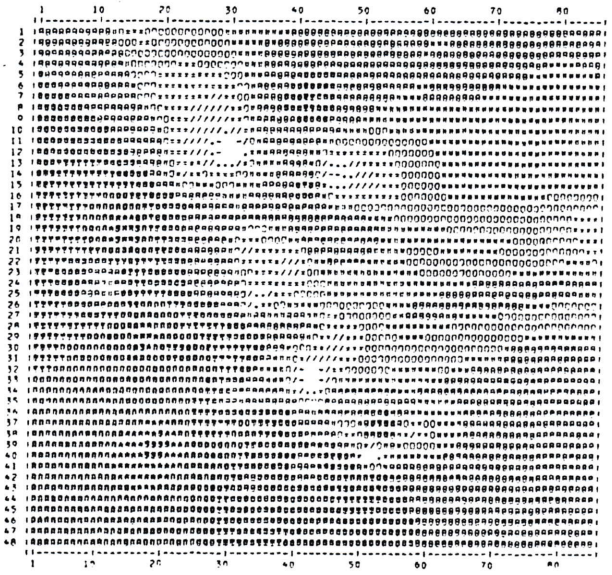

Figure 3 Shaded Map for Example 1

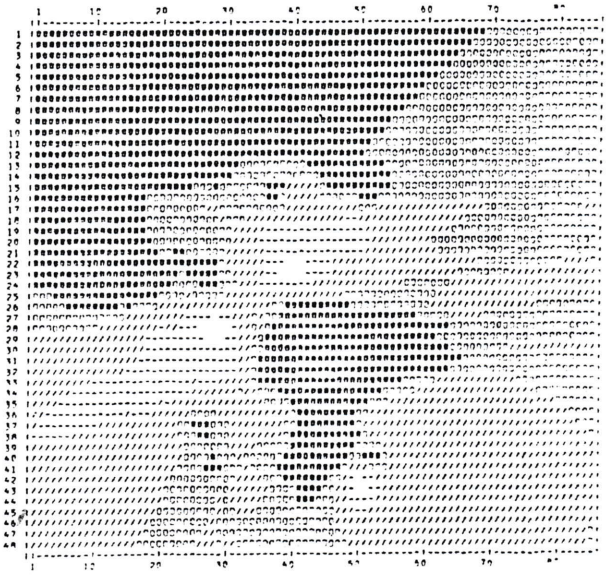

Figure 4 Shaded Map for Example 2

4. Future Improvement

As seen in application examples, the areal clustering techniques proved to be effective in comprehensive and simultaneous understanding of dwellers' attitudes or other related factors in the objective regions of the survey. However, the results also indicate some points to be improved in the future:

(i) Correction of discontinuity or singularity seen on color graphics

Geographical information or distribution of population generally have continuity or similarity between neighboring areas. Dwellers' attitude survey data does not, however, necessarily exhibit such continuity or similarity. For example, when the responses obtained at scattered locations have a tendency which departs from those obtained at other locations, the colors at those locations give us a "singular" impression. This is due to the fact that the developed algorithm does not fully take into account the sizes of the areas recognized by dwellers or the range of the "neighborhood" identified by each dweller.

(ii) Color display and use of district statistical information

The method used to edit district statistical information (e.g., mesh data, land use conditions, pollution complaints in specific areas, number of traffic accidents, etc.) into sampling small areas must be determined before correlating such data with attitudes. Directly correlating dwellers' attitudes with land shapes, traffic networks, and housing concentration does not necessarily produce useful information.

(iii) Method of setting representative points

Presently the geometrical centroid of a district to which sampling small areas belong is assumed to be the representative point of that district. In actuality, however, the dwellings of respondents do not necessarily concentrate at the representative point. In this sense, the size of the district in question in relation to the locations of the dwellings scattered in that district must be taken into account. This is needed along with sampling method development.

RESUME

En visualisant sur "les cartes coloriées des attitudes envers les intérêts environnementaux" des relations entre l'environnement et les attitudes des habitants à la base des données cueillies sur les attitudes des habitants urbains envers les intérêts environnementaux, nous exposons ici un rapport sur les techniques de grappillage territorial, qui peuvent être utiles pour éclaircir des particularités d'un environnement régional, et le système de traitement des informations par ordinateur. Nous décrivons d'abord le principe des techniques de grappillage territorial que nous proposons et les caractéristiques des cartes coloriées et du système de processus par mini-ordinateur que nous avons développés pour appuyer lesdites techniques. Ensuite nous abordons la méthode de traitement statique des informations qui est nécessaire pour les exploiter. Et nous expliquons aussi l'efficacité d'un mini-ordinateur pour de telles recherches et l'utilité des cartes coloriées en citant quelques exemples de l'application réelle des tecniques de grappillage territorial sur des données actuelles des attitudes.

REFERENCES

Ahuja, N. and Schachter, B.J. (1983). Pattern Models, John Wiley.

Beatty, J.C. (1983). Raster Graphics and Color, The American Statistician, 37, 60-75.

Davis, I.J. and McCullagh, M.J. (eds.) (1975). Display and Analysis of Spatial Data, Nato Advanced Study Institute, John Wiley.

Foley, J.D. and Van Dam, A. (1982). Fundamentals of Interactive Computer Graphics, Addison-Wesley.

Ohsumi, N. and Sibuya, M. (1978). Numerical Techniques for Areal Partitions: NTAP (in Japanese), The Proceedings of the Institute of Statistical Mathematics, 25, 1, 41-63.

Ohsumi, N. (1983), Practical Techniques for Areal Clustering, Proceedings of 3rd International Symposium Data Analysis and Informations (INRIA, France).

APPENDIX 1 Outline of Survey

The questionaires used for the survey consisted of questions asking about the overall degree of satisfaction with the dwelling areas, complaints about individual environmental factors (greenery, noise pollution, etc.), the deviation of the present living environment from the ideal one, requests for environmental administration, and so on.

(i) Sampling areas and objective dwellers

The objective regions of survey and the number of dwellers were as follows:

Urban area of Chiba City: 48 sampling small areas, 1,440 dwellers
Senri New Town: 60 sampling small areas, 1,800 dwellers

Both regions are part of larger metropolitan areas and exhibit typical urban environmental problems. The urban Chiba City, however, has a long history of gradual urban development, and includes a mix of old and new buildings and facilities, where a variety of environmental factors such as air pollution and traffic problems are observed. On the other hand, Senri New Town near Osaka City is a planned community that was constructed approximately 20 years ago. Each of these cities demonstrate typical urban dwelling patterns which are significantly different from each other. They were selected as the objectives of the survey to see how the differences between the two cities would be reflected in the results of the same survey.

(ii) Sampling and survey methods

The objective dwellers were sampled randomly from polling registers using the two-stage sampling method. First, the sampling small areas were selected in proportion to population within each district, and 30 dwellers were sampled in the selected sampling small areas, respectively. This means that the geographic locations of the sampling small areas were not predetermined. The distribution of sampling small areas is affected by density of population, which resulted in an uneven distribution of sampling small areas on the map. This may have some affect on the results of areal clustering. Sampled dwellers entered their responses themselves on response sheet.

(iii) Response completion condition

The survey sheets were recovered as follows:

	Number of sampled dwellers	Number of respondents	Response rate
Urban Chiba City (the central region only)	1,440	768	53.3%
Senri New Town	1,800	1,205	67.0%

The response rate was low, as is usually the case with recent surveys conducted in large urban areas.

APPENDIX 2 Areal Clustering Techniques

The principal objective of areal clustering 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. The areal clustering 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 k-means-type methods. The mean vectors $\bar{\underset{\sim}{x}}_l$ $(l = 1, 2, \ldots, L)$ and other statistics are calculated for each cluster.

(Step 3) The result of clustering is exhibited on a graphic display unit, and several locates (or representative pixels) are chosen in each cluster, for example, applying manually the crosshair cursor at each locate. 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 a pixel j, 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:

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

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

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

where, w_k is the weight factor, usually in inverse proportion to the measures of distance that is,

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

(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) 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.

84 TOKYO IBC

THE XIIth INTERNATIONAL BIOMETRIC CONFERENCE

PROCEEDINGS

Invited Papers

September 2–8, 1984
Keio Plaza Inter•Continental Hotel
Tokyo, Japan