8. REDATO: An archaeological database system with geographical analysis

Kazumasa Ozawa
Department of Management Engineering, Osaka Electro-Communication University, Neyagawa-shi, Osaka 572, Japan.

8.1 Introduction

Many types of research support systems have been developed (Duong *et al.* 1978; Van Nuen & Benders 1981; Matsuka & Sugimoto 1982). A common emphasis, as suggested in any overview of existing systems, seems to be placed on links with special databases. This implies that a research support system should provide not only data processing tools but also stimulate information effective for problem solving. There would appear to be many problems in archaeology, for which such systems could act as powerful research tools.

This paper describes a research support system for archaeological studies on a special type of Japanese ancient monument. Between the 4th and 6th centuries AD, a large number of huge ancient tombs, termed Keyhole (shaped) Tombs, were built all over Japan (see Fig. 8.1). Even today more than three thousand Keyhole Tombs still remain in spite of excessive urbanization in modern Japan. Keyhole Tombs have played a most important role in understanding the so-called Ancient Tomb Period, in which the Japanese regime originally started. REDATO, REsearch support system with DAtabases of ancient TOmbs, has been implemented to support archaeological studies on the Keyhole Tombs using a variety of real-time analyses including statistical, pictorial and geographic. In this paper, special emphasis is placed on the geographic analyses.

8.2 Aims and logical system structure

An iterative cycle of thinking seems to underlie most studies in archaeology, with one cycle consisting of modelling and its verification by observation. Imperfections may be detected in a model which is then modified into a new model for verification in the next cycle. Usually, such verification is done by drawing distribution maps, statistical computing or other processing of the data. REDATO performs such verification in real-time.

A Keyhole Tomb involves many kinds of archaeological information with much of the attribute information being described symbolically and then classified. Non-symbolic aspects such as tomb shape and the geographical position where it was found are also important. These can be called pictorial and geographic information respectively. The following three databases play key roles in the sophisticated performance of REDATO, providing necessary information for real-time analysis; the Attribute Database (AD), Pictorial Database (PD) and Geographic Database (GD). AD is relational in structure, storing 47 descriptive fields, PD stores encoded strings of contours of the tomb mounds and GD has detailed line primitives to draw every part of the Japanese islands.

The system also includes a number of real-time analytical modules to manipulate the scientific information stored in the three databases. Among them, the following four modules are fundamental in the performance of the system:

Attribute Processing Module (APM)
Pictorial Processing Module (PPM)
Geographic Processing Module (GPM)
Statistical Computing Module (SCM)

Fig. 8.2 shows a diagrammatic representation of the logical structure of REDATO. The first three modules APM, PPM and GPM manage different types of data in AD, PD and GD, respectively. SCM handles numerical data sets. Inter-module communication is carried out through commonly accessible data sets defined in the working file.

8.3 Databases

As previously mentioned, the archaeological information describing the Japanese tombs has been classified into three types, i.e. symbolic (attribute), pictorial and geographic. This classification appears to be effective for many other problems in archaeology for in most cases an archaeological report consists of text, sketches or plates of artifacts and distribution maps. This leads to the basic idea that underlies REDATO and links the following three databases:

8.3a Attribute database (AD)

Attribute data includes character strings such as *name*, *address* or *type* and numeric strings such as *latitude*, *longitude*, *elevation* or *length*; 47 different attributes have been defined in AD. The conceptual schema is based on Codd's relational model (Codd 1970) with eight tables each including a group of attributes. The tables are hierarchically structured with one mother table and seven child tables. A number of attributes in the mother table are associated with PD and GD.

8.3b Pictorial database (PD)

PD stores pictorial information of shapes of the tomb mounds (see Figs. 8.1 and 8.3). *Freeman's chain coding* (Fu 1976) has been used to describe the contour lines of a Keyhole Tomb mound. The original digitised data obtained from a contour plan of a tomb is a set of many point coordinates, with usually more than ten contour lines per tomb. This requires a huge amount of memory space and Freeman's chain coding is a method for reducing the required memory. As shown in Fig. 8.3, a given line can be coded approximately to a string of numbers 1,...,8, each of which signifies an octal primitive as defined in Fig. 8.3a. A contour line

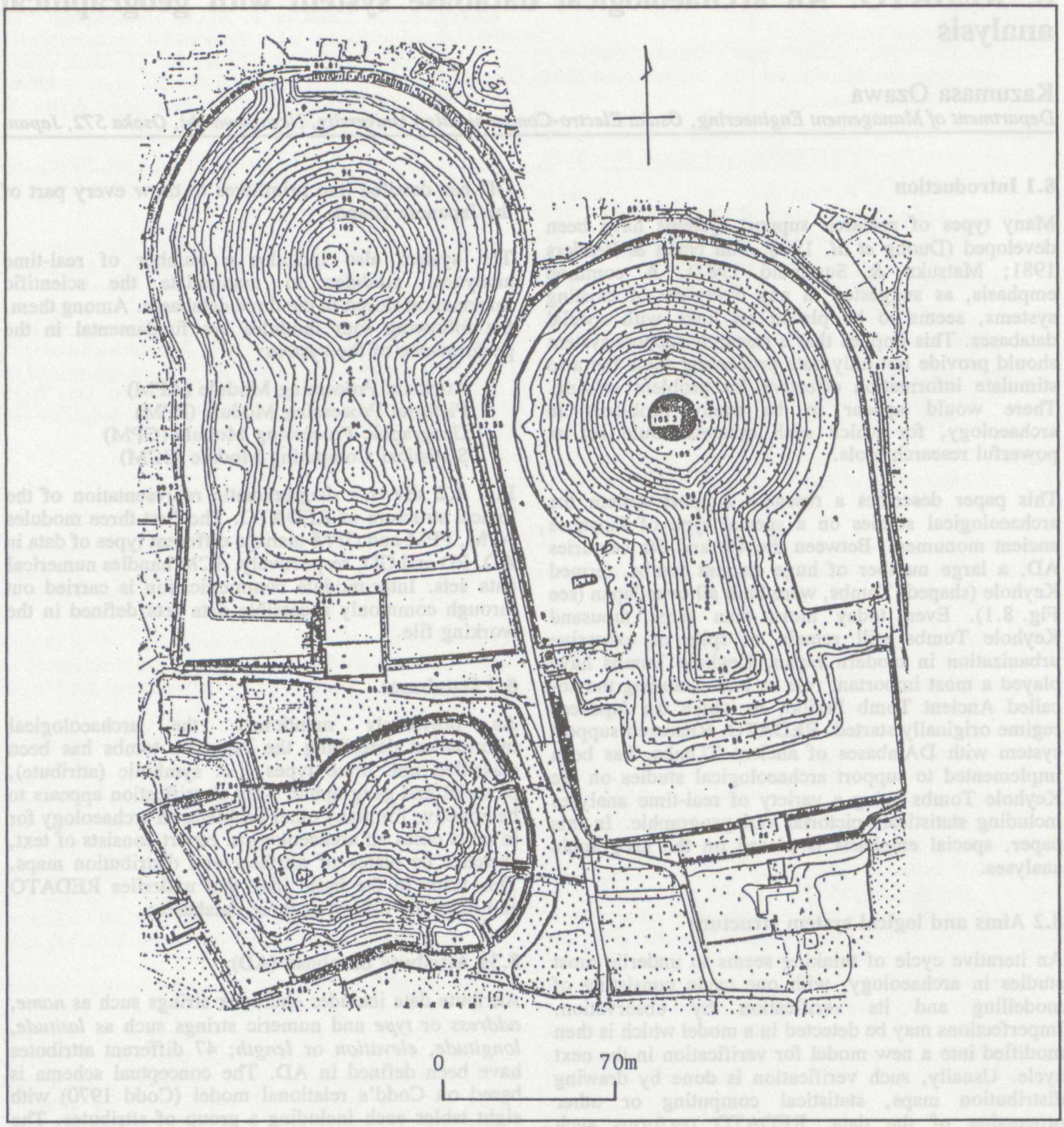


Figure 8.1: Contour maps of the Keyhole tomb mounds near Nara City.

can thus be described by single point coordinates from a starting point consisting of a coded string. Octal primitives of a reasonable size can save memory space while providing a good approximation to the original contour lines.

8.3c Geographic database (GD)

In archaeology, distribution maps have been very important to illustrate geographic relations between sites or points. Land maps are essential for drawing such distribution maps. Many computer techniques to store geographic maps have been presented (Olson 1983; Matsuyama *et al.* 1984). The data structure of GD has been designed specially by referring to these existing techniques: Partition of the Japanese islands can be described as a hierarchy of three levels; region, prefecture and city level, respectively. A city, the smallest element of partition, can be drawn by a number of line elements. Since every prefecture is composed of a collection of cities, it can also be drawn by line elements; a region or the whole Japanese islands can be handled in the same way. GD stores all the line elements each of which is described as a series of points given in terms of latitude and longitude.

8.4 Data processing modules

REDATO includes as its main analytical components the four data processing modules shown in Fig. 8.2,

each of which plays a role in the creation of information for the iterative process. Inter-module communication is carried out through commonly accessible data sets, termed *C-sets*, defined in the working file. For example, a *C-set* could be a set of tombs with a specified common property defined within a module. A couple of existing *C-sets* can be reduced to a new *C-set* using the set operation within APM. Every module works with other modules, reorganizing data stored in the databases to establish useful sets for study. Emphasis has also been placed on how to represent graphically the results for visual impact. In the following, the basic ideas underlying the design of the modules and their capabilities are presented.

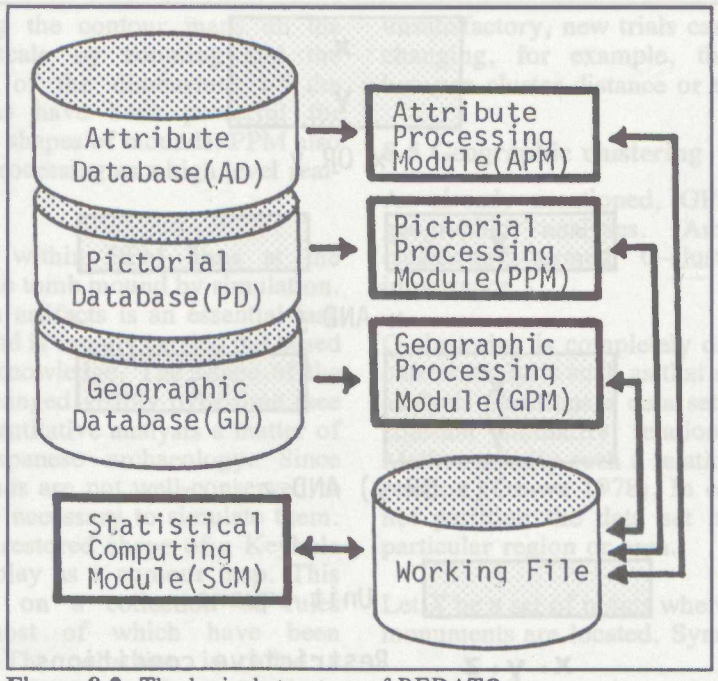


Figure 8.2: The logical structure of REDATO.

8.4a Attribute processing module (APM)

This module is equivalent to the database management system (DBMS) which facilitates conventional retrieval from AD. One specific feature of APM can be seen in its manner of query where every query should be given in a special graphical expression instead of an inquiry language such as CODASYL or SEQUEL (Codd 1970). A query is, in most cases, written by a logical expression composed of several conditions and logical operators such as AND and OR. As a simple example, suppose x and y be the following conditions:

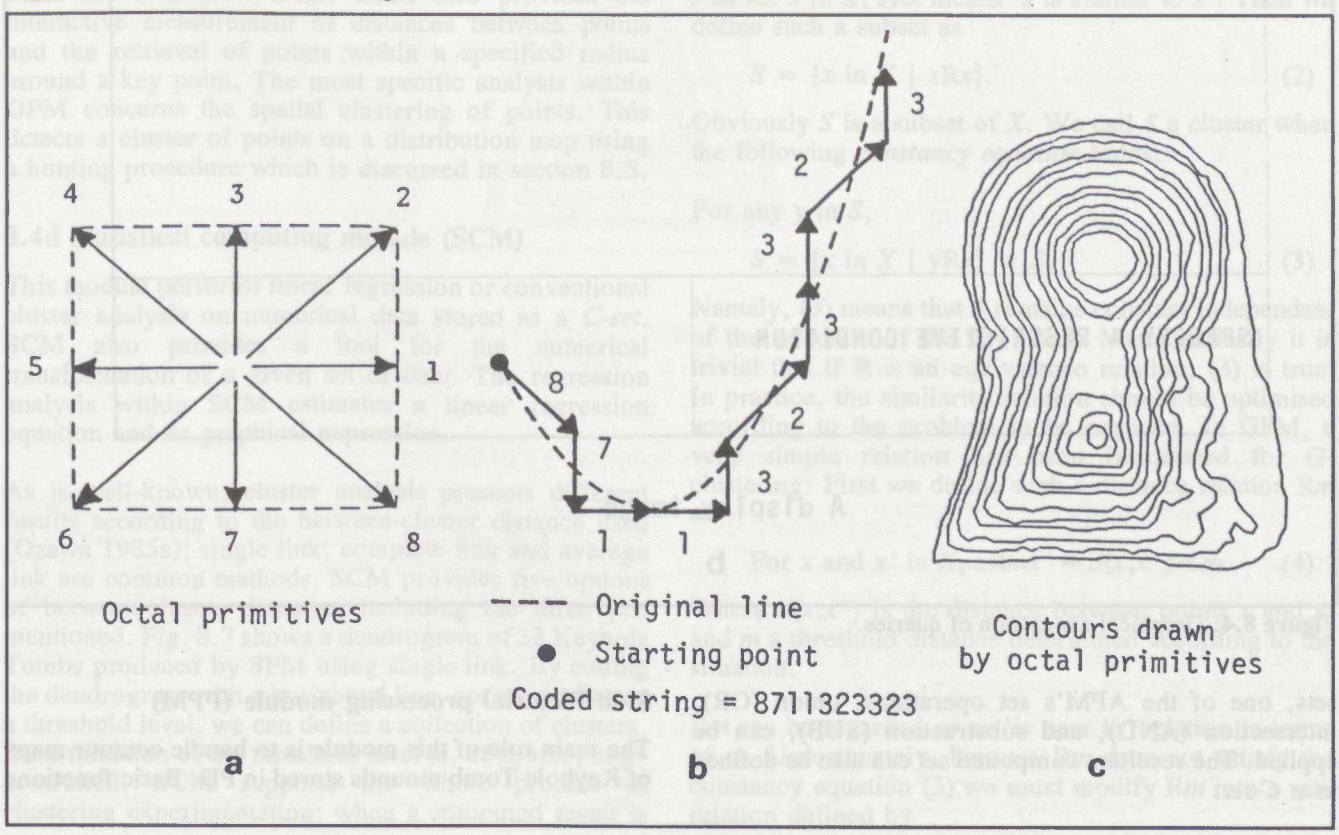
x = It is in Osaka Prefecture

y = It is longer than 100m

Fig. 8.4a shows APM's graphical expressions including xANDy and xORy where a unit square island indicates a restrictive condition. For example, when a query includes xORy the two unit islands x and y are joined to a bigger island. On the other hand, for a query involving xANDy, two unit islands are represented separately. Such graphical expression can be made on the monitor by specifying conditions and logical operators using the keyboard. A hardcopy of the display image is shown in Fig. 8.4b.

A collection of the items that agree with a query can be defined as a *C-set* in the working file. For a pair of

Figure 8.3: Freeman's chain coding.



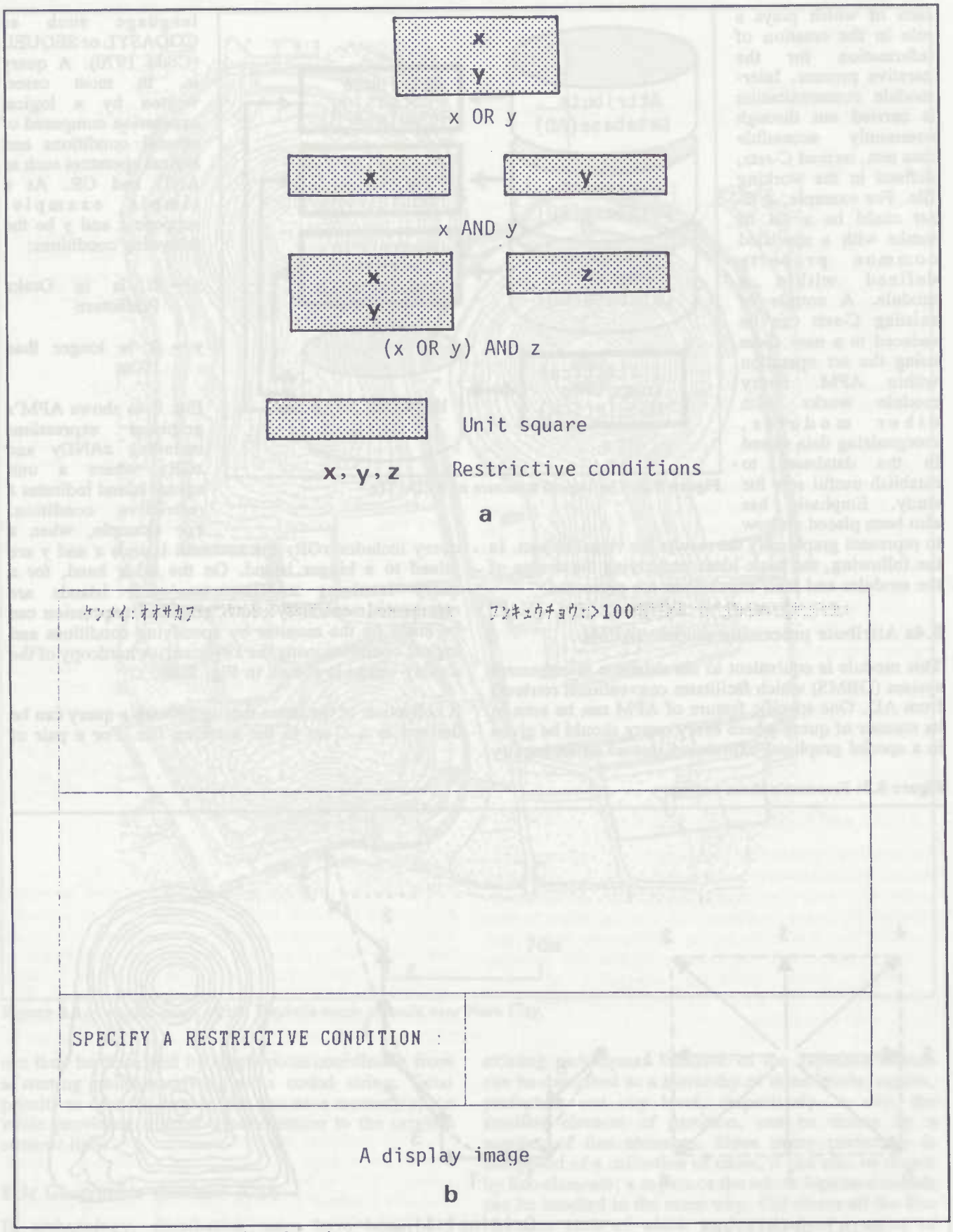


Figure 8.4: Graphical expression of queries.

sets, one of the APM's set operations, union (OR), intersection (AND), and substruction (SUB), can be applied. The resultant compound set can also be defined as a *C-set*.

8.4b Pictorial processing module (PPM)

The main role of this module is to handle contour maps of Keyhole Tomb mounds stored in PD. Basic functions

of PPM include drawing the contour maps on the display, changing the scale by zooming and the interactive measurement of the dimensions of the mounds. These functions have been powerful for comparative studies of the shapes of mounds. PPM also provides pictorial query processing as a high level real-time process.

Another useful process within PPM aims at the automatic restoration of the tomb mound by simulation. The restoration of broken artifacts is an essential task for many archaeologists and is usually carried out based on empirically acquired knowledge. The shape of the Keyhole Tomb mound changed visibly over time (see Fig. 8.5a), making its quantitative analysis a matter of some significance in Japanese archaeology. Since almost all the tomb mounds are not well-conserved to serve such purposes, it is necessary to simulate them. PPM rapidly depicts the restored shape of a Keyhole Tomb mound on the display as a contour map. This process depends mainly on a collection of rules embedded in PPM, most of which have been determined statistically. The collection is called a knowledge-base, and consists of rules related to taxonomical studies. An example of such automatic restoration can be seen in Fig. 8.5b; the contour map on the left shows an actual mound and that on the right its restored contours.

8.4c Geographic processing module (GPM)

A variety of distribution maps can be quickly obtained by GPM. The principle to draw a distribution map is very simple, it is completed by plotting a given set of items, according to their latitudes and longitudes, on the background map concerned. Fig. 8.6 shows an example of a distribution map drawn by GPM. In general, a set of items to be plotted is produced by the other modules as a *C-set*. GPM also provides the interactive measurement of distances between points and the retrieval of points within a specified radius around a key point. The most specific analysis within GPM concerns the spatial clustering of points. This detects a cluster of points on a distribution map using a hunting procedure which is discussed in section 8.5.

8.4d Statistical computing module (SCM)

This module performs linear regression or conventional cluster analysis on numerical data stored as a *C-set*. SCM also provides a tool for the numerical transformation of a given set of data. The regression analysis within SCM estimates a linear regression equation and its graphical expression.

As is well-known, cluster analysis presents different results according to the between-cluster distance used (Ozawa 1985a); single link, complete link and average link are common methods. SCM provides five options of between-cluster distance, including the three just mentioned. Fig. 8.7 shows a dendrogram of 33 Keyhole Tombs produced by SPM using single link. By cutting the dendrogram with a horizontal line, corresponding to a threshold level, we can define a collection of clusters. Determination of the threshold level is, of course, user-controlled. SCM supports the whole process of clustering experimentation; when a computed result is unsatisfactory, new trials can be quickly performed by changing, for example, the distance measure, the between-cluster distance or the threshold level.

8.5 Geographic clustering

As already mentioned, GPM provides a variety of geographic analyses. Among them, geographic clustering, termed G-clustering, is of particular importance.

G-clustering is completely different from conventional cluster analysis such as that shown in Fig. 8.7. Cluster analysis partitions a data set, which is equivalent to a constant similarity relation defined over the set. Mathematically such a relation is called an equivalence relation (Ozawa 1978). In contrast, G-clustering does not partition the data set but detects a cluster in a particular region or area.

Let X be a set of points where Keyhole Tombs or other monuments are located. Symbolically we have

$$X = \{x_1,\ldots, x_n\}. \quad (1)$$

Since all points $x_1,\ldots, x_n$ usually extend over a wide area and involve different cultures, it is unreasonable to assume a similarity measure uniformly constant over X. It follows that the similarity measure between points depends on localised clustering producing subsets of points (see Fig. 8.8). Strictly, therefore, a similarity measure should be applied to a small region, for example, the neighbourhood of every point.

Let s in X be a key point which defines a localised cultural trait. Practically, s will provide the starting point around which to detect a cluster. Here we assume a constant similarity relation R in the neighbourhood of s as for x in X, $s\mathrm{R}x$ means 's is similar to x'. Then we define such a subset as

$$S = \{x \text{ in } X \mid s\mathrm{R}x\}. \quad (2)$$

Obviously S is a subset of X. We call S a cluster when the following *constancy equation* holds:

For any y in S,

$$S = \{x \text{ in } X \mid y\mathrm{R}x\} = S. \quad (3)$$

Namely, (3) means that S remains constant independent of the selection of the key point. Mathematically it is trivial that if R is an equivalence relation, (3) is true. In practice, the similarity relation should be optimised according to the problem to be engaged. In GPM, a very simple relation has been introduced for G-clustering: First we define such a distance relation Rm as

$$\text{For } x \text{ and } x' \text{ in } X,\ x\mathrm{R}mx' \leftrightharpoons d(x,x') < m. \quad (4)$$

Where $d(x,x')$ is the distance between points x and x' and m a threshold distance determined according to the situation.

Rm can be regarded as an 'is near to' relation in terms of m. Unfortunately, because Rm does not satisfy the constancy equation (3) we must modify Rm into a new relation defined by

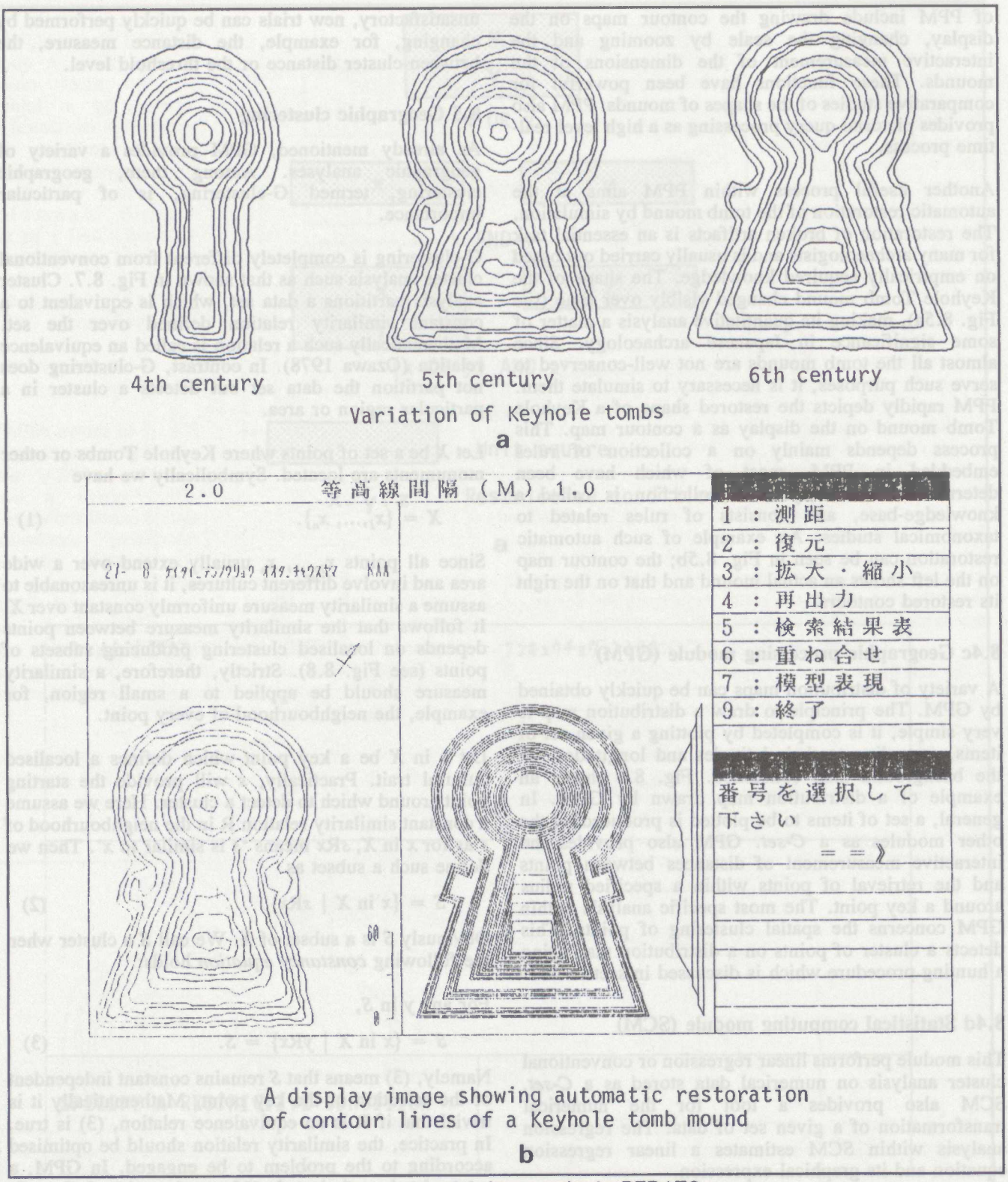


Figure 8.5: Contour lines of Keyhole tomb mounds and their restoration by REDATO.

$$\hat{R}m = Rm \cup RmRm \cup RmRmRm \cup \ldots \quad (5)$$

Where $RmRm$ means a binary relation with an intermediate point y such that for x,x',y in X, $xRmy$ and $yRmx'$. The succeeding terms after $RmRm$ are given in a similar way so that $\hat{R}m$ is nothing but a transitive closure of Rm and, concurrently, is an equivalence relation satisfying (3).

$\hat{R}m$ has been employed in GPM as the similarity relation for G-clustering Keyhole Tombs. The procedure to detect a cluster is very simple starting with the drawing of a distribution map to be shown on the monitor. Secondly, using the cursor, we point at a key point from the selection shown on the map. Finally we specify the threshold distance using the keyboard. The system then automatically detects a cluster around the key point.

A detected cluster of points is quickly indicated in a different colour from the other points but as this is impossible to illustrate by monochrome hardcopy, in

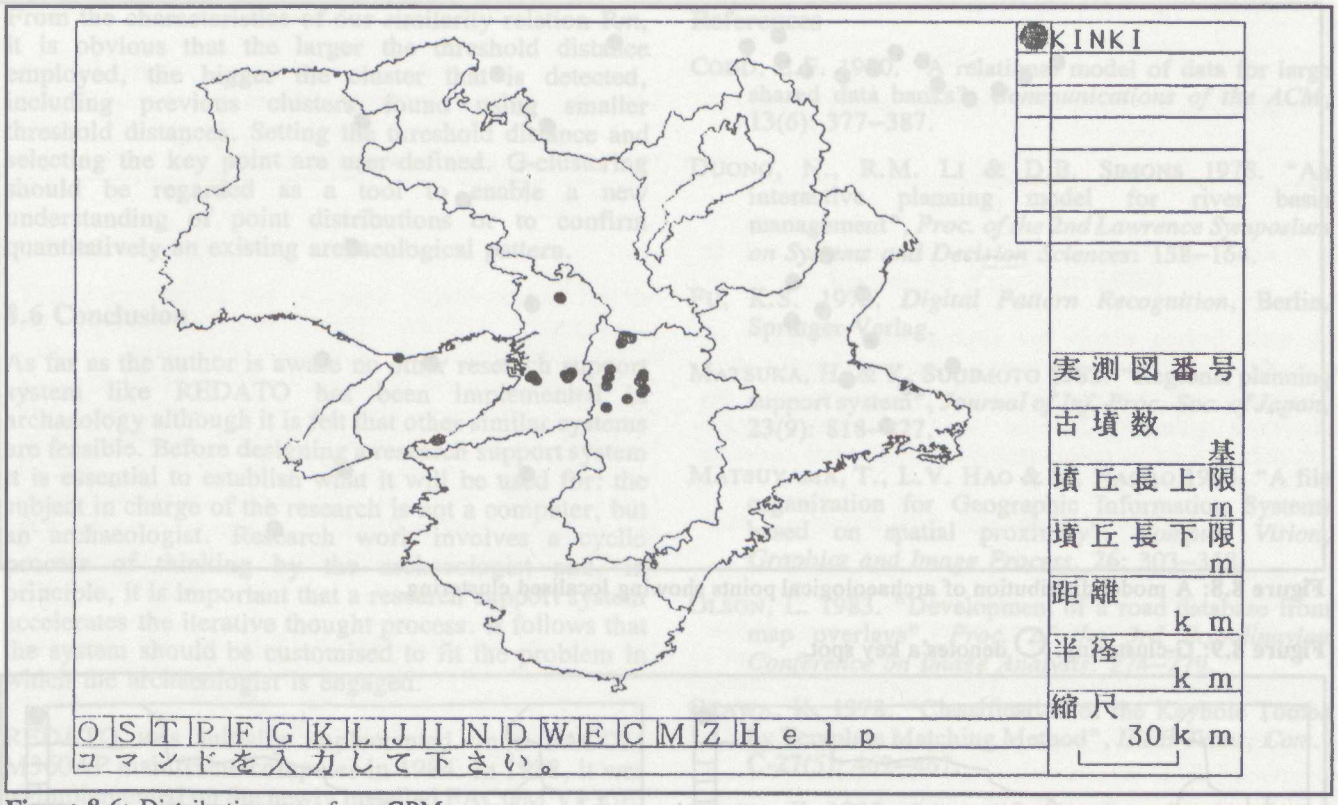
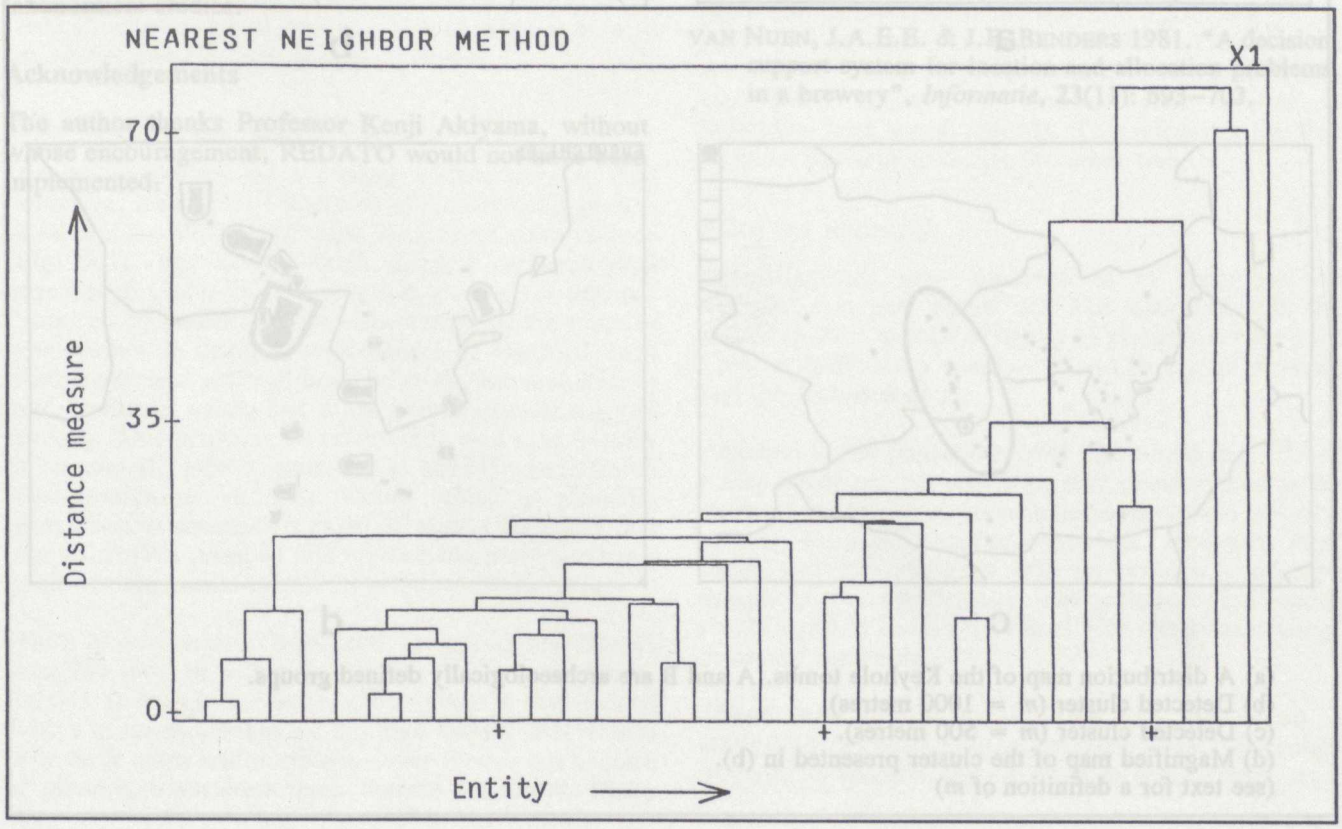


Figure 8.6: Distribution map from GPM.

Fig. 8.9 every detected cluster is shown enclosed by a line. In this Figure, (a) is a distribution map of Keyhole Tombs in the south-eastern part of Osaka, where the two groups of ancient tombs, A and B enclosed with lines, have already been designated by archaeologists on qualitative criteria; (b) shows a cluster automatically detected by GPM, which is nearly equal to A, one of the archaeologically defined groups. A point marked ⊙ is a key point and, in this case, the threshold distance *m* is set at 1000m; (c) shows another good result of G-clustering where the detected cluster is completely consistent with the pre-defined group B (in this case, the threshold distance is set at 500m); (d) shows a magnified distribution map focusing on the cluster presented in (b).

Figure 8.7: Dendrogram from SCM.



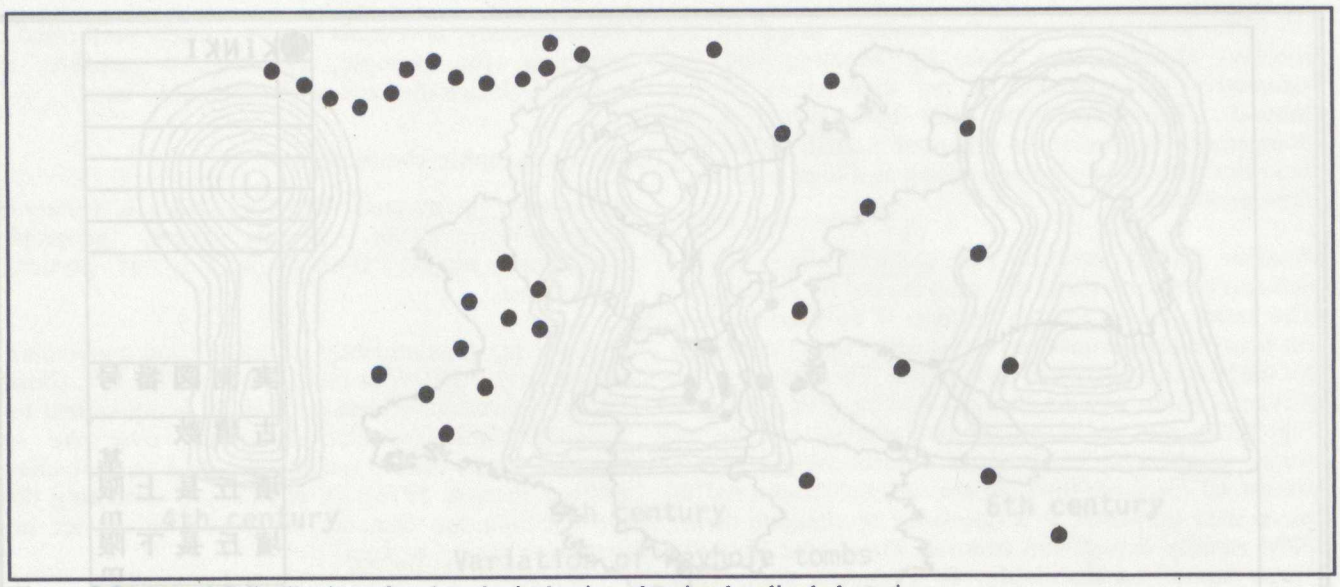
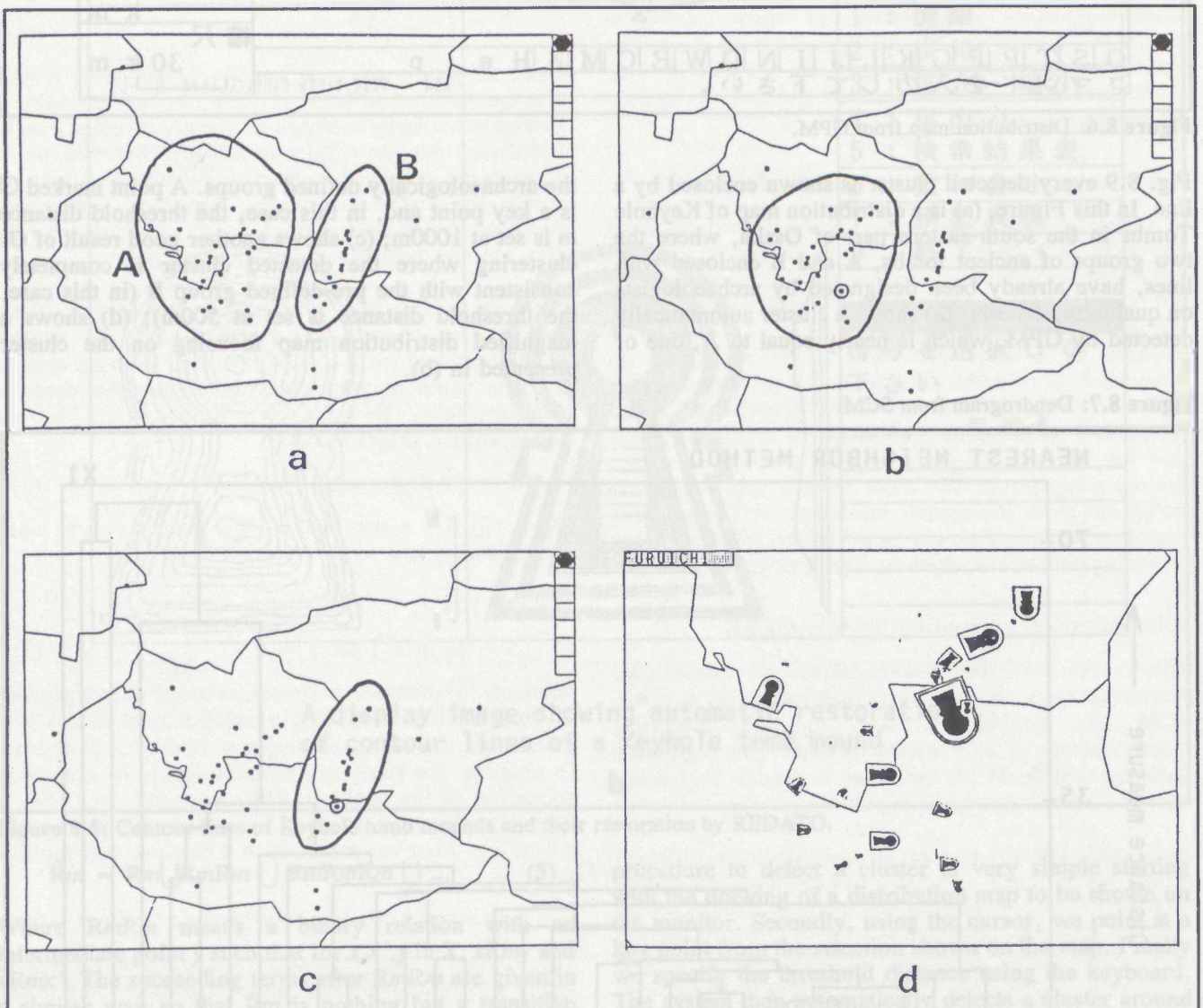


Figure 8.8: A model distribution of archaeological points showing localised clustering.

Figure 8.9: G-clustering. ⊙ denotes a key spot.



(a) A distribution map of the Keyhole tombs. A and B are archaeologically defined groups.
(b) Detected cluster (m = 1000 metres).
(c) Detected cluster (m = 500 metres).
(d) Magnified map of the cluster presented in (b).
(see text for a definition of m)

From the characteristics of our similarity relation $\hat{R}m$, it is obvious that the larger the threshold distance employed, the bigger the cluster that is detected, including previous clusters found using smaller threshold distances. Setting the threshold distance and selecting the key point are user-defined. G-clustering should be regarded as a tool to enable a new understanding of point distributions or to confirm quantitatively an existing archaeological pattern.

8.6 Conclusion

As far as the author is aware no other research support system like REDATO has been implemented in archaeology although it is felt that other similar systems are feasible. Before designing a research support system it is essential to establish what it will be used for: the subject in charge of the research is not a computer, but an archaeologist. Research work involves a cyclic process of thinking by the archaeologist and, in principle, it is important that a research support system accelerates the iterative thought process. It follows that the system should be customised to fit the problem in which the archaeologist is engaged.

REDATO was initially implemented on a FACOM M360AP mainframe computer in 1984. In 1988, it was re-implemented on the newly installed FACOM VP30E, of which the specifications include 64 MByte of CPU memory and 10 GByte of disk memory. An end user of REDATO can access all the modules via a colour graphic display unit. From the view-point of system design, there is room for improvement in the present REDATO. One future task is to introduce an image database and its managing module, while another task is to introduce Artificial Intelligence techniques more positively. Automatic inference linked with a knowledge-base appears to act well in support of chronological studies (Ozawa 1985b, 1989) and taxonomical studies.

Acknowledgements

The author thanks Professor Kenji Akiyama, without whose encouragement, REDATO would not have been implemented.

References

CODD, E.F. 1970. "A relational model of data for large shared data banks", *Communications of the ACM*, **13**(6): 377–387.

DUONG, N., R.M. LI & D.B. SIMONS 1978. "An interactive planning model for river basin management", *Proc. of the 2nd Lawrence Symposium on Systems and Decision Sciences*: 158–164.

FU, K.S. 1976. *Digital Pattern Recognition*, Berlin, Springer-Verlag.

MATSUKA, H. & K. SUGIMOTO 1982. "Regional planning support system", *Journal of Inf. Proc. Soc. of Japan*, **23**(9): 818–827.

MATSUYAMA, T., L.V. HAO & M. NAGAO 1984. "A file organization for Geographic Information Systems based on spatial proximity", *Comput. Vision, Graphics and Image Process*, **26**: 303–318.

OLSON, L. 1983. "Development of a road database from map overlays", *Proc. of the 3rd Scandinavian Conference on Image Analysis*: 274–279.

OZAWA, K. 1978. "Classification of the Keyhole Tombs by Template Matching Method", *IEEE Trans. Com.*, **C-27**(5): 462–467.

OZAWA, K. 1985a. "A stratificational overlapping cluster scheme", *Pattern Recognition*, **18**(3/4): 279–286.

OZAWA, K. 1985b. "A rule-based archaeological periodization system for Japanese Ancient Tombs", *Trans. of Inf. Proc. Soc. of Japan*, **26**(2): 342–348.

OZAWA, K. 1989. "Rule-based dating of artefacts", in S. Rahtz & J. Richards (eds.), *Computer Applications and Quantitative Methods in Archaeology 1989*, British Archaeological Reports (International Series) **548**, Oxford, British Archaeological Reports: 375–386.

VAN NUEN, J.A.E.E. & J.F. BENDERS 1981. "A decision support system for location and allocation problems in a brewery", *Informatie*, **23**(11): 693–703.