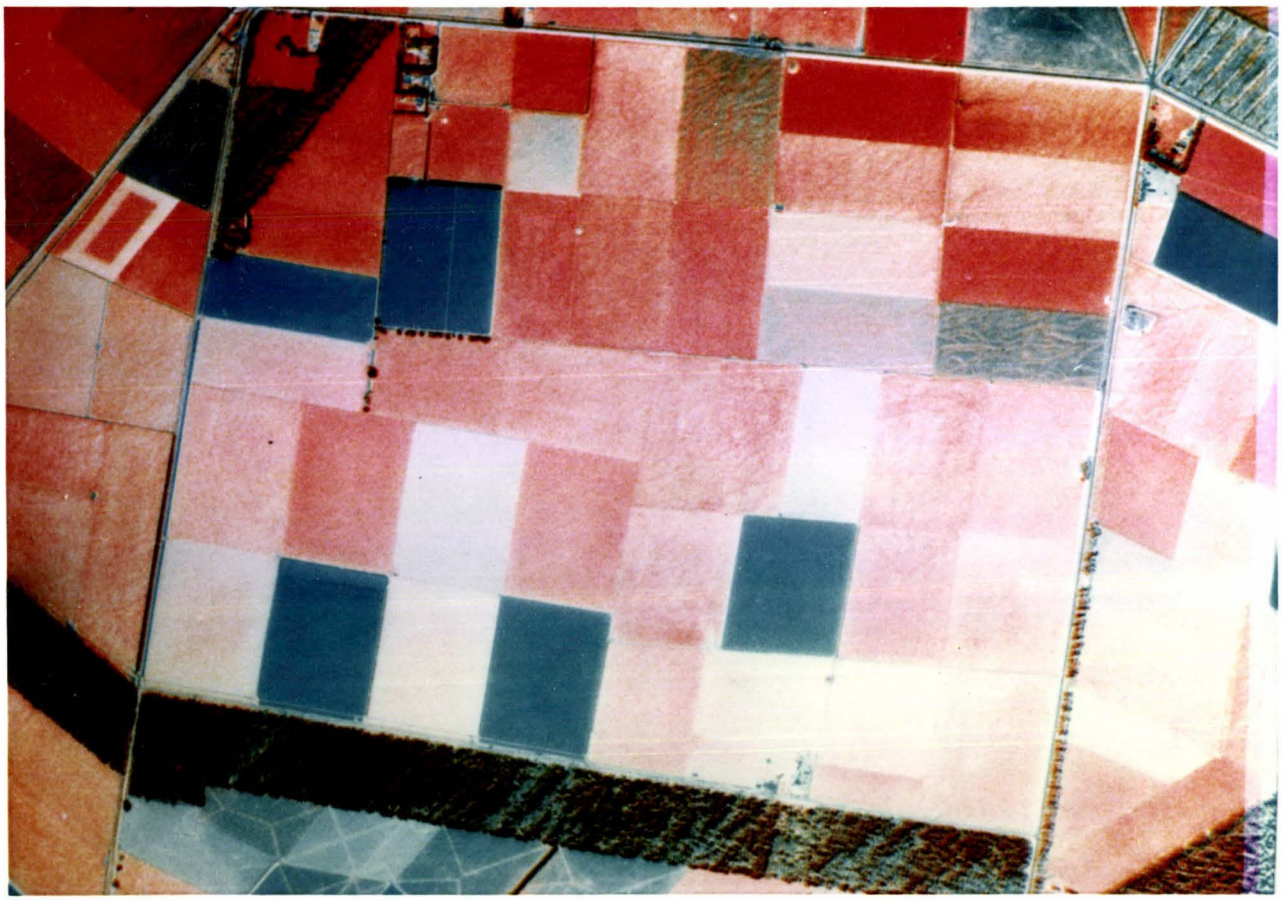
UNSUPERVISED CLUSTERING OF SPECTRAL SIGNATURES IN LANDSAT IMAGERY

by Brian C. Clement B.E. (Hons)

December 1977

A Thesis presented in partial fulfilment
of the requirements for the degree of
Master of Arts in Computer Science
at Massey University



ABSTRACT

This thesis describes an investigation into automatic recognition of satellite imagery from the LANDSAT Project. Clustering techniques are shown to be the most suitable; of the three clustering algorithms investigated the k-means is shown to be the most effective. The need to perform edge detection on the images prior to clustering is also demonstrated. A suitable algorithm for edge detection is described.

Indexing terms: clustering, LANDSAT Satellite project, pattern recognition, Satellite data

ACKNOWLEDGEMENTS

I wish to express my thanks to my project supervisor, Mr K.J. Hopper, for his encouraging assistance during the preparation of this thesis, and to Dr. B.E. Carpenter for introducing me to this area of Artificial Intelligence.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

1 INTRODUCTION

As man seeks to automate increasingly more complex tasks, he demands a higher degree of pseudo-intelligence from the controlling systems. This pseudo-intelligence ensures that the system is indeed automated, requiring less human intervention in either setting up or operation. In addition it may assist people in making the 'best use' of such a system - since the system is to be used why not make it do as much as possible of the work itself?

The increased demand for machine intelligence has given rise to much research effort in areas of Artificial Intelligence, one of the most prominent being Pattern Recognition - the automatic recognition and classification of objects into sets or classes.

1.1 Data Requirements

The type of Pattern Recognition techniques applicable in any particular case is determined largely by the nature of the data to be analysed. Consequently the first stage of the investigation described in this thesis was to establish a source of a suitable set of data. Suitable data should satisfy the following conditions:

a. It should be available with as little requirement as possible for pre-processing. In the worst case, it should be available at all, so that there would be no need for data generation.

b. It should be sufficiently large to demonstrate the inherent structure of the data and it should be a representative sample from a larger population.

c. There should also be available sufficient information to determine the accuracy of the recognition system output, i.e. there should be a corresponding set of 'answers'.

d. If possible there should also be some practical applications for the solution of a recognition system for the data.

Of the above, b is clearly the most important, since the data used will determine the generality of the resulting system. Requirement c is essential for this research project since there is a need to measure the performance of the system being developed to determine how nearly the original goals have been met.

Contact was made with the Remote Sensing Section of the Physics and Engineering Laboratory of the DSIR who were beginning work on Pattern Recognition in satellite images to produce thematic maps. These include land use maps, and maps for crop census work. They were able to supply data from the Earth Resources Technology Satellite (ERTS) project which met the requirements for this research.

1.2 Remote Sensing

In an effort to aid his understanding of his environment, man has developed many diverse and specialised techniques to collect information about it. The term Remote Sensing is used to refer to methods of obtaining information by:

a. Gaining views which were previously impossible e.g. a satellite photograph of a sub-continent in a single frame.

b. Gaining "unfamiliar views of familiar" [a], e.g. views at infra-red or ultra-violet wavelengths.

Information is sometimes collected by non-photographic sensors - in non-visible wavelengths - but stored in a photo like form, e.g. the output from a radar scanner may be used to expose light sensitive film in synchronised sweeps; since this has not strictly produced a photograph the terms 'image' and 'imagery' are often used.

Such imagery may be used in the discovery of previously unknown features or in the grouping of features which are in some way alike. This grouping may be done by detection of spectral signatures i.e. visual characteristics of an object which define it and distinguish it from other similar objects. This may be applied in a number of areas:

Meterology:-	Forecasting, storm tracking.
Agriculture:-	Plant disease, insect infestation.
Forestry:-	Forest fire and disease detection.
Ecology:-	Pollution monitoring e.g. oil spills, thermal pollution.
Medicine:-	Diagnosis of human ailments from thermal infra-red body scans.
Natural Resource Management:-	Mineral prospecting, mapping.

Because of the nature of such data, processing is most often necessary to enable human observers to interpret it, certain applications may even require on-line processing, thus high-speed computing resources have facilitated or even made possible for the first time the handling of such data Further, some data retrieval systems provide a continous stream of high-speed data e.g. satellite projects may return 10^{14} bits per year, confirming the need for high speed data processing capabilities and large volume storage.

1.3 LANDSAT

The primary objective of the ERTS project, which launched the LANDSAT satellites, was to assist in experiments concerning utilization of the earth's resources. This is done by monitoring and recording earth observation data from space. The first LANDSAT satellite was launched in July 1972 into a sun-synchronous orbit with a period of 103 minutes. A second similar satellite was launched in January 1975. Each satellite orbits the earth 14 times daily viewing a 185km wide swathe, covering the globe's entire surface in 18 days. The term sun-synchronous means that any ground point is re-visited at the same time of day thereby minimising the effect of shadows.

There are two separate sensor systems aboard LANDSAT. One system has three television cameras in a Return Beam Vidicon (RBV) system which returns a television image of the same ground scene in each of three wavebands between 0·48 and 0·83 μm. Each image is of an area 185 km × 185 km. The second system, which is of more immediate interest for this research project, is a Multi-Spectral Scanner (MSS) in which an oscillating mirror scans the field of view across the line of flight reflecting the radiation on to solid state detectors. The detectors are sensitive to four wavebands, two in the visible and two in the near infra-red spectrum. Each has a field of view of a ground area of 79m × 79m, see fig 1. The data received by ground tracking stations needs then to be corrected for aberrations in satellite platform attitude, height and speed, and for distortion due to to forward motion while scanning, and finally for the earth's rotation.

One of the most powerful aspects of the LANDSAT project is the fact that not only spatial and spectral information is collected, but since each ground point is revisited every 18 days, it is also possible to analyse the data temporally i.e. to monitor not only what ground features are present but also how they change with time.

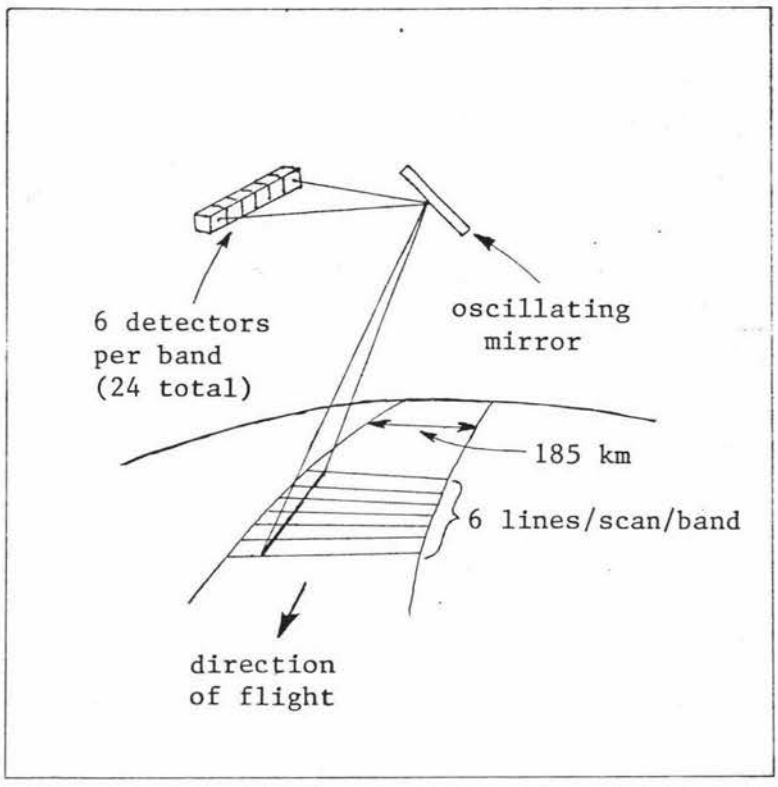


Fig 1 The LANDSAT Multispectral Scanner

1.4 The Research Objective

The aim of this research is to investigate and compare Pattern Recognition techniques most suitable for recognition of LANDSAT MSS imagery.

2 PATTERN RECOGNITION

Before discussing the details of this research project it is necessary to outline the principles involved in recognising and classifying objects or patterns.

2.1 General

Given a set of data it may be required to separate its elements into groups or classes according to some criteria determined by the characteristics of those elements. Each element in the set is a number of measurements of an object in the physical world, p. It is called a pattern and may be represented by a vector in a p-dimensional Euclidian Space called a Pattern Space. If p is very large, in some cases it may be greater than 1000, a reduction in the number of dimensions will need to be made to enable any further processing to be done. This is known as feature extraction or feature selection and should ideally retain all the independent variables, or features, so that no information is lost. A transformation function is then applied to the pattern in order to determine to which of a number of classes it belongs, being similar to the other members of that class according to some criteria. There may also be an 'undecided' class for those patterns which do not appear to be closely associated with any of the other classes already defined.

This final classification or recognition phase forms the largest part of most Pattern Recognition systems and may be one of several types:-

a. Template Matching. Stored within the Recogniser is a set of patterns whose classes are known. Each pattern to be classified is compared with those stored, one by one,

and is assigned to the class of the pattern it matches or most nearly matches. This is a simple-minded but intolerant approach, however it performs satisfactorily for near perfect patterns, i.e. those which vary little within each class, such as the characters of a known type font.

b. Similar Features. Each pattern's features are used to determine its similarity to sample patterns stored in the system, and it is classified according to the most similar pattern class. For example, a character recogniser may search for vertical or horizontal straight lines, or for curves or line segment intersections, any particular letter corresponding to a certain combination of such features. This relies on the assumption that patterns belonging to a particular class will probably display similar features. A difficulty may arise in determining a complete of discrminating features for each class.

c. Clustering. The structure of the data should produce localised regions of high density in the pattern space. By searching for and isolating these density maxima or clusters, which should correspond to the classes, a pattern may then be classified according to its membership of a cluster. A drawback with this technique is in the measure of success likely to be achieved since there exists no easily definable criteria for optimal clusters. Possible measures are:

(1) Maximisation of correct classification, although this requires an independent check on the data, i.e. knowledge of the 'answers', and may thus not be possible.

(2) Optimisation of some other criterion, such as maximisation of inter-cluster separation, or minimisation of intra-cluster distances.

The preceding three types of classification may be considered to be arranged in order of suitability for processing data with fewer dimensions although this may not hold true in all cases.

2.2 Choice of a Technique for LANDSAT Data

The LANDSAT data consists of a large number of points, or patterns a standard image of 185 km square ground scene has 7.6×10^6 points each of which has 4 values ($p = 4$) corresponding to the measured intensity levels in the four wavebands. Sample patterns are generally not available and the data is of few dimensions removing the need for feature extraction. Clustering techniques are therefore the most suitable. Thus the remainder of this discussion concerns clustering methods.

2.3 Supervised and Unsupervised Recognition

Pattern recognisers or classifiers may have available to them representative or sample patterns from each class to be recognised. The techniques used are then called supervised. In such systems the recogniser is taught or trained to recognise patterns from each class through exposure to the sets of training patterns using some adaptive heuristics; they are 'taught' the representative characteristics which identify each class. This teaching takes place in a separate training phase prior to the recognition of unknown patterns, although further modification or adaptation may be carried out during the actual recognition phase.

In certain applications there may be no training patterns available because either they are difficult to collect or, possibly, the number of classes expected may be unknown. In these cases the techniques used are called unsupervised. The system may be given no information at all concerning the representative characteristics of each class or of the number of classes.

The nature of the LANDSAT project and its resultant data generally precludes any form of training or supervision since:

a. The ground area covered may include features, corresponding to pattern classes, which were previously unknown and hence for which there are no training patterns available. In effect this means that the number of classes to be expected may be unknown.

b. Some features have a time varying spectral signature, e.g. the sesonal variations in agricultural crops, thus current or relevant training patterns may be difficult to correlate with available satellite imagery.

2.4 Clustering Principles

Each data point, or pattern, may be represented by a p-dimensional vector in a Euclidian Space:

$$x = [x_1, x_2, \ldots x_p]^T$$

where x_i is the ith measurement of the pattern

Clustering procedures seek to group the data according to any naturally occurring structure within it, either as a means of data compression or in order to determine differentiating characteristics to classify data

according to natural grouping tendencies.

These procedures may be separated into several types as follows:-

a. A *bottom-up* approach in which the clusters are built up by combining data points which are in some way measurably similar. A hierarchical tree structure may be built in which sub-clusters which are found to be similar enough are combined to form fewer clusters at the next higher level and so on. One of the simplest measures of similarity between two points is that of proximity in the Pattern Space using Nearest Neighbour considerations, see [b]. A point may be grouped together with its nearest neighbour, or classified according to the most populous class of its k nearest neighbours. Patrick and Jarvis [c] have proposed a method of combining two points which have in common more than a certain number of nearest neighbours.

b. A *top-down* approach in which the pattern space is partitioned into regions according to population density considerations. The space may be searched for local density maxima and their surrounding minima which define the cluster boundaries. Points are then classified according to the cluster boundaries which include them.

c. A compromise between top-down and bottom-up approaches. Using this procedure the clusters are created by using several given cluster centres to group the points initially on a nearest-is-best basis. These clusters are then refined in an iterative fashion by using their member points to modify the

cluster centres and then reclassifying all of the points etc. This is the original *K*-means algorithm discussed in ref. [d]. A refinement of this method, the ISODATA algorithm also discussed in ref. [d], includes possible fusing or lumping of two neighbouring clusters, and the splitting of large clusters. The parameters affecting such alterations may be changed dynamically. Some method of selecting meaningful initial cluster centres, by for instance choosing points from the data set itself, will speed the iterative process.

3 CURRENT SATELLITE IMAGERY RESEARCH

The advent of wide-spread availability of data from satellite projects such as LANDSAT has instigated an increased amount of effort into automated classification methods.

3.1 Traditional Approaches

A traditional statistical approach involving the computation of covariance matrices has been suggested by Haidar [e]. This relies on the assumption that the data points may be described by Gaussian probability density functions. It also requires training patterns to be used in estimating parameters for those density functions. The availability of sufficient training patterns can not always be assumed, and in any case such methods as are developed ought not be dependent on this. Also, because of the large data sets being processed, the computational requirements for the matrices would be excessive, e.g. to process an image of only 1000 points an array of 500,000 elements is required.

Schell [f] proposes a 'spatial-spectral clustering'classification philosophy. In this a fairly straightforward nearest-neighbour clustering method is enhanced by use of spatial information. If there is any doubt as to a point belonging to a particular cluster, e.g. if it is further than a threshold distance from the cluster centre, the classes of the points adjacent in the image are considered. This makes use of the intuitive principle that adjacent points in the image may belong to the same ground feature and thus to the same pattern class. Spatial considerations do indeed add some information to a recogniser system, though it may be regarded as only complementary. However, the extra effort required to extract this information, e.g. by re-reading the file to find previous data

points, may not be justified by a small improvement in overall system performance.

Some use of spatial information is also made by Jayroe et al [g]. An attempt is made to identify areas of a minimum size which are spectrally homogeneous, i.e. which probably correspond to large ground features. To do this the areas must be isolated by detection of feature boundaries which should enclose the areas. Spectral information for each area is then used to determine whether any areas may be considered as the same class or cluster, and the data set is classified around the determined clusters. A second search is then made of the unclassified points to find any remaining (smaller) homogeneous areas. Cluster statistics for these are calculated as before and the data set reclassified.

The consideration of boundary points between ground features is an important step since that may be, in part, how a human observer intuitively distinguishes individual areas which are then compared with each other.

Other workers, such as Weeden et al. [h] and Borden [i], have developed comprehensive interactive systems which may use operator interpretation of preliminary images e.g. to identify homogeneous areas and choose them as training patterns. However, since the objectives of this research did not include the development of such sophistication, the criteria for the selection of suitable methods will probably not fully coincide with theirs. Indeed it seems unreasonable to expect a completely automated system of this type to perform as well as those which also incorporate human intuition.

3.2 Important Aspects

It was apparent from the work described above that in the development of an effective recogniser system no single clustering technique is sufficient. Either two techniques are integrated, their combined effectiveness being greater than that of either individually, or various methods of enhancing the clustering results are used. It was decided, therefore, to first investigate several basic clustering algorithms to determine which performs the most accurately with the least amount of resources. Refinements or other enhancements could then be incorporated to improve the performance.

Initially no use of spatial information would be made. This was to avoid any over-complications of the system so that areas of strength or weakness might be more easily identified. It was recognised, however, that consideration of spatial information would provide confirmation or 'greater certainty' when classifying.

Boundary information was thought to be important. Note that although the boundaries do have relevance in a spatial context, they are detected primarily from the spectral characteristics of the data and may therefore be regarded as spectral in origin for the purposes intended here.

4 PROBLEM APPROACH

The research to be undertaken could now be divided into three main areas. An outline of each is given below, more or less in the order in which they need to be dealt with, and successive chapters describe them in greater detail. All programming was carried out on Massey University's Burroughs B6700 system. The programming language used throughout was Burroughs extended Algol since:-

a. Although portability was not seen to be a major consideration all Universities in New Zealand have Burroughs systems.

b. The author finds it to be the most pleasant and effective of the high level languages available in which to program, but mainly

c. On this Computer system, this language is the most versatile.

4.1 General Data Manipulation

A need existed for a number of general utility programs. This includes such things as restructuring the original data files for reasons of efficiency, and routines to display and analyse the data in various ways. Chapter 5 is concerned with these.

4.2 Edge Detection

Following from 3.2 an attempt to determine the edges or boundaries between ground features in the images was to be made. This is to assist the subsequent classification of the data. Chapter 6 explains in some depth the reasons for this and the manner in which it was done.

4.3 Classification

In 2.4 three types of clustering methods were outlined. An attempt was made to implement an algorithm to perform each of these. The principles are described below and Chapter 7 gives details of the actual implementation.

4.3.1 Shared Near Neighbour

This is a bottom-up approach which relies on the characteristics of the data to effect the clustering. In the absence of any other information the data points must simply be compared with one another in a meaningful way. Jarvis and Patrick [c] prpose that a suitable similarity measure between two points can be found using their nearest neighbours in the pattern space. In this algorithm each point's nearest neighbours are tabulated, this involves a pass through the entire data set for each point calculating the point pair separations and retaining the k smallest of these. A comparison is then made for all pairs of points between their sets of nearest neighbours If there are more than a threshold number, k_t, of nearest neighbours common to the two points, then the points are regarded as belonging to the same class. The two points themselves must also be included in each other's list of nearest neighbours. In this way the groups or clusters are built up as more are determined to be similar enough, and representative values for clusters may be calculated. By suitable selsction of k and k_t the tolerance in regarding points as similar enough may be varied.

4.3.2 The Divisive Approach

Intuitively an observer will group or cluster data in the pattern space by identifying occurrences of comparative high density. Following this line a top-down approach attempts to search the pattern space for local density maxima which should correspond to clusters, or more correctly to cluster centres. These centres may then be used to cluster the remainder of the data

or indeed the entire data set. Two methods of searching the pattern space were considered:-

a. The points are projected on to two-dimensional planes which are then searched for local maxima. This reduction in dimensions allows the use of normal hill-climbing methods to find the maxima. Then by correlation of the coordinates of the maxima between all (6) possible planes the coordinates of the maxima in the pattern space can be found. Since this technique was being investigated by M^{c}Donnell et al [j] no attempt was made to investigate this any further.

b. The pattern space is divided into equivolume hypercubes whose populations are measured by counting the number of contained points. This may be seen as another form of data compression since the pattern space is divided into progressively fewer cells, each of larger volume. As a result there are fewer comparison operations necessary to determine local maxima. There is also a corresponding lessening of precision since the integrating effect of summing cell populations obscures some of the detail.

4.3.3 The *K*-means Algorithm

This method is conceptually very simple, consisting of only three steps, last two of which are repeated until a stopping criterion is satisfied, fig 2. This algorithm essentially minimises the squared distances of all points from their cluster centres, see e.g. Tou and Gonzalez [d]. The initialisations are arbitrary, as is k, but understandably they will affect the performance of the algorithm. One variant chooses the first k data points from the data set as the initial values. During classification, which

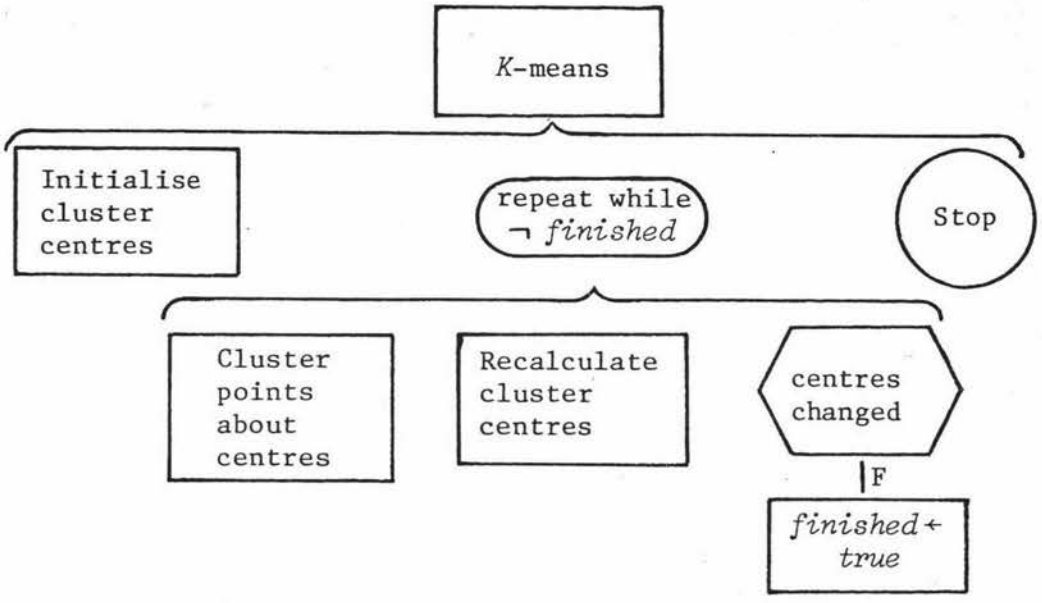


Fig 2. The *K*-means Algorithm

should perhaps be more correctly called 'allocation', the points are assigned to one of the clusters using a minimum distance criterion e.g. in the nth iteration a point $\bar{x}$ is assigned to the jth cluster if:

$$d(\bar{x}, \bar{z}_j(n)) < d(\bar{x}, \bar{z}_i(n))$$

for all $i = 1,2,\ldots,k \quad i \neq j$

where: $\bar{z}_i(n)$ is the centre of the ith cluster during the nth iteration, and

$d(\bar{x}, \bar{y})$ is some metric between vectors $\bar{x}$ and $\bar{y}$ such as the Euclidian distance:

$$d(\bar{x}, \bar{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

The cluster centres are then updated by averaging the values of the members of each cluster. Thus if there are $N_i(n)$ points belonging to the ith cluster after the nth iteration and they are represented by the set $S_i(n)$, then:

$$z_i(n+1) = \frac{1}{N_i(n)} \sum_{x \in S_i(n)} \bar{x} \qquad i = 1.2\ldots,k$$

The stopping criterion could be that there be no change between $\bar{z}_i(n)$ and $\bar{z}_i(n+1)$ for $i = 1,2,\ldots,k$. Or, similarly, that there be no change between $N_i(n)$ and $N_i(n+1)$ for $i = 1,2,\ldots,k$.

A more comprehensive version of the algorithm, known as ISODATA [d], includes heuristics to enhance its performance. It allows clusters to be lumped or merged together if they are considered to be sufficiently close i.e. sufficiently similar. It also allows the splitting of large clusters (those which are large in spread - particularly in one dimension - not merely populous) since this may indicate that two clusters are present but are confused as one. ISODATA requires several parameters besides the centres to be given to determine allowable cluster characteristics.

5 GENERAL DATA MANIPULATION

The data made available by the DSIR was an image 128 picture elements (pixels) square of an area of the Canterbury Plains in the South Island of New Zealand. This area was chosen because it is flat and features agricultural crops with a high visual contrast, some ground truth was available for the area. The frontispiece is an Infra Red aircraft photograph of part of the test area.

A number of general routines were required since:

a. The format of the data as received was not optimal.

b. There needs to be some method of displaying the data pictorially.

c. To aid in preliminary numerical analysis statistical information of the data needs to be displayed.

These tasks are now described in greater detail.

5.1 File Handling

Repacking

The data was contained in one file - all four bands - and some initial restructuring was necessary to facilitate efficient I/O operations. Since each pixel represents a radiation level of between 0 and 127, an 8-bit byte or character may be used to store its value, each scan or row of 128 pixels thus requires 128 bytes. The B6700 system incorporates 6-byte words, so that each data row would use 22 words, though to enable more efficient blocking 24-word records were used. A routine, FILEMAKER, described in ANNEX A, was written to perform this conversion.

Flipping

Pattern Recognition usually requires the comparison of adjacent pixels. Rather than reading all the data into arrays which could then be scanned either row-wise or column-wise, it was decided to create another set of four files corresponding to the four already described but with rows and columns transposed. That is to say each record of the second set of files contains the data corresponding to a column of the image and not a row as before, see ANNEX B.

5.2 Display Routines

The Raw Data

In order to have available an easily accessible visual record of the 'raw' data values, a procedure ALLLEVELS, see ANNEX C,was written to print the intensity values for the entire image or any section of it for any one of the wavebands. The procedure produces a matrix of points each with a symbol representing the measured data value for the corresponding pixel. The B6700 lineprinter has available 64 distinct printable character symbols, therefore to cover the entire range of 128 possible levels overprinting was used. Where the value (i) recorded was greater than 63 the printed symbol used is the same as that used for $i-64$ with the addition of an underscore bar: '_'. The underscore was found to be a suitable compromise between distinctiveness and obscuring of the overprinted symbol. Fig 3 shows the output produced by ALLLEVELS for Band 7; in this particular case there were no pixels with a value greater than 63.

Shaded Output

For ease of visual interpretation of the printed images SHADES produces a grey tone version of the image by overprinting suitable combinations of characters to give a shading effect. The number of grey levels may be varied though 8 or 10 was found to produce the best picture since with any more the

small increments between levels becomes difficult to distinguish, and with fewer some effect is lost. In any case the lineprinter is, at best, a poor graphics output device [k].

Thought was given to the use of a Tektronix 4010 display terminal as a more effective graphics device but no further action was taken because:

a. Coding the character matrix itself was not possible, thus only those character forms produced by the terminal are available.

b. The screen is a storage type and although the terminal has hardcopy facilities available the quality is questionable, particularly since it is so variable.

c. This is of minor importance in this research - the shaded images described will be used only as an aid, assisting in identifying the presence and shape of ground features and their similarity to other features in an effort to roughly determine the recogniser's accuracy and consistency.

For similar reasons no effort was made to use the available drum plotter as a display device. Fig 4 shows the shaded output produced by SHADES corresponding to fig 3, note the distinctive forest and worked fields (appearing as blue in the frontispiece). See ANNEX D for details of SHADES.

5.3 Analysis Routines

Two routines were written to display some aspects of the data characteristics. They enable the distribution of the data values to be observed, this is helpful in determining initial parameters.

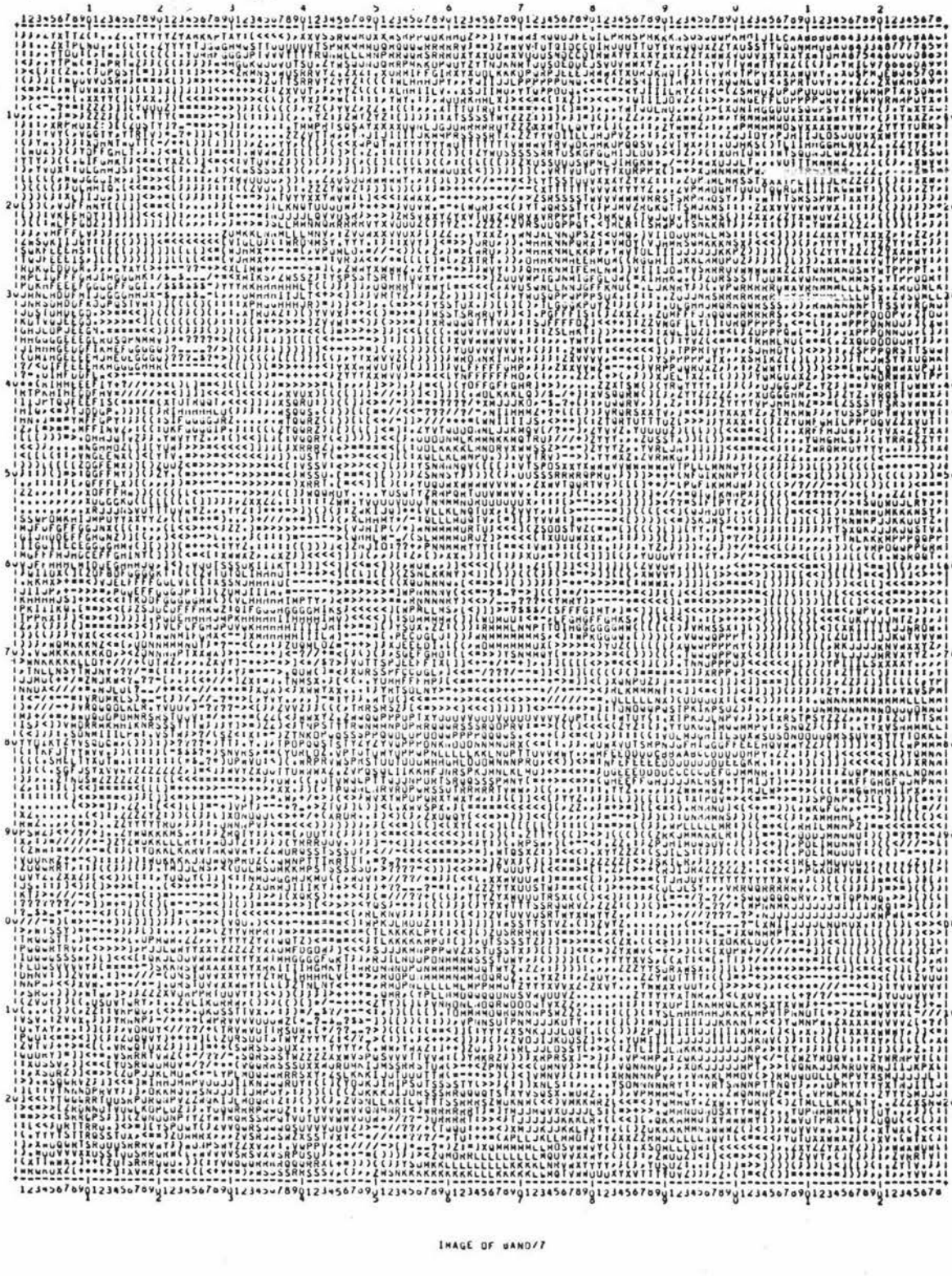


IMAGE OF BAND/7

KEY OF INTENSITY LEVELS AND THEIR REPRESENTATIVE SYMBOLS:

Fig 3a. The output produced by ALLLEVELS for Band 7 showing the intensity levels recorded for the entire image.

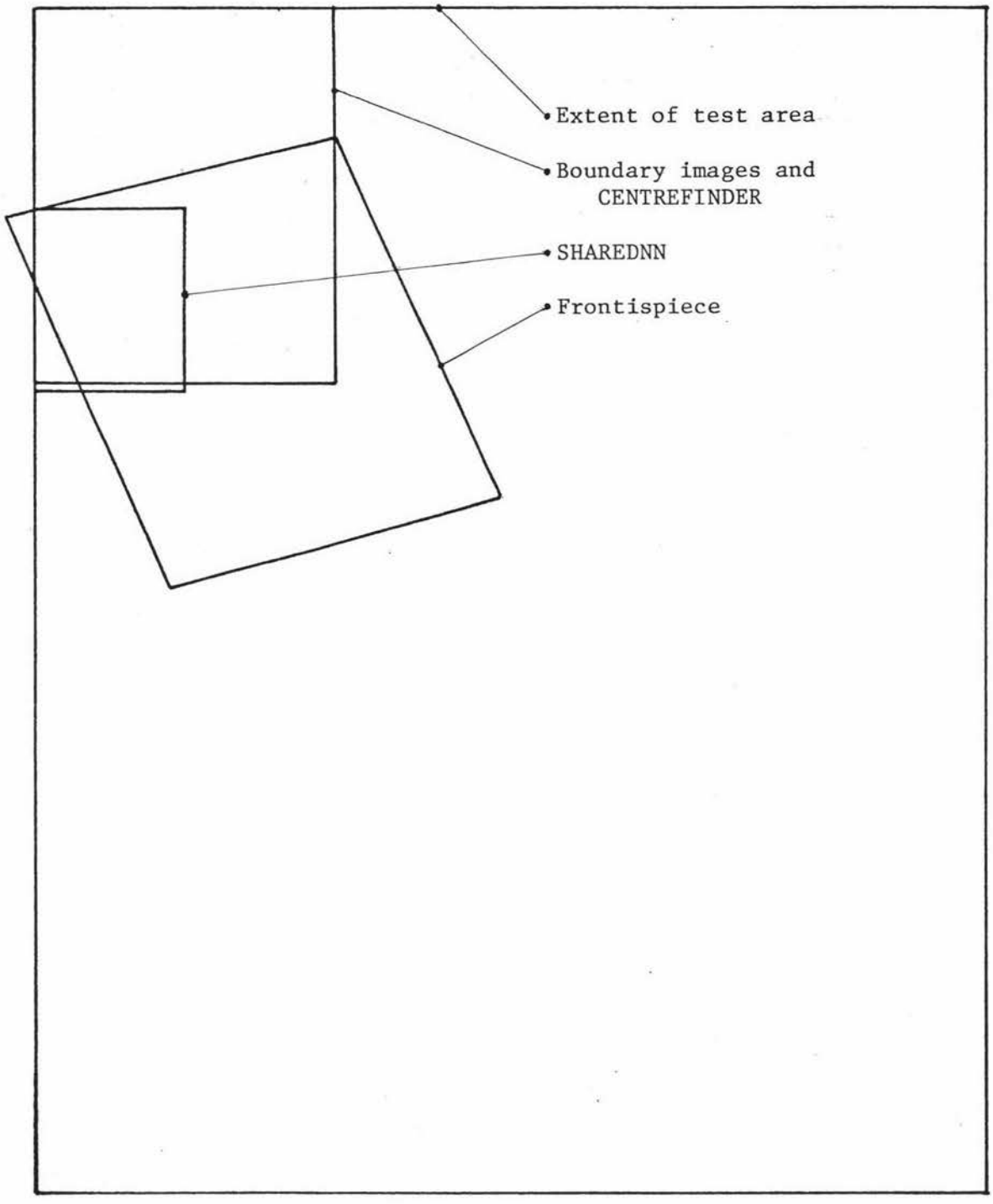


Fig 3b. Comparison of image registration.

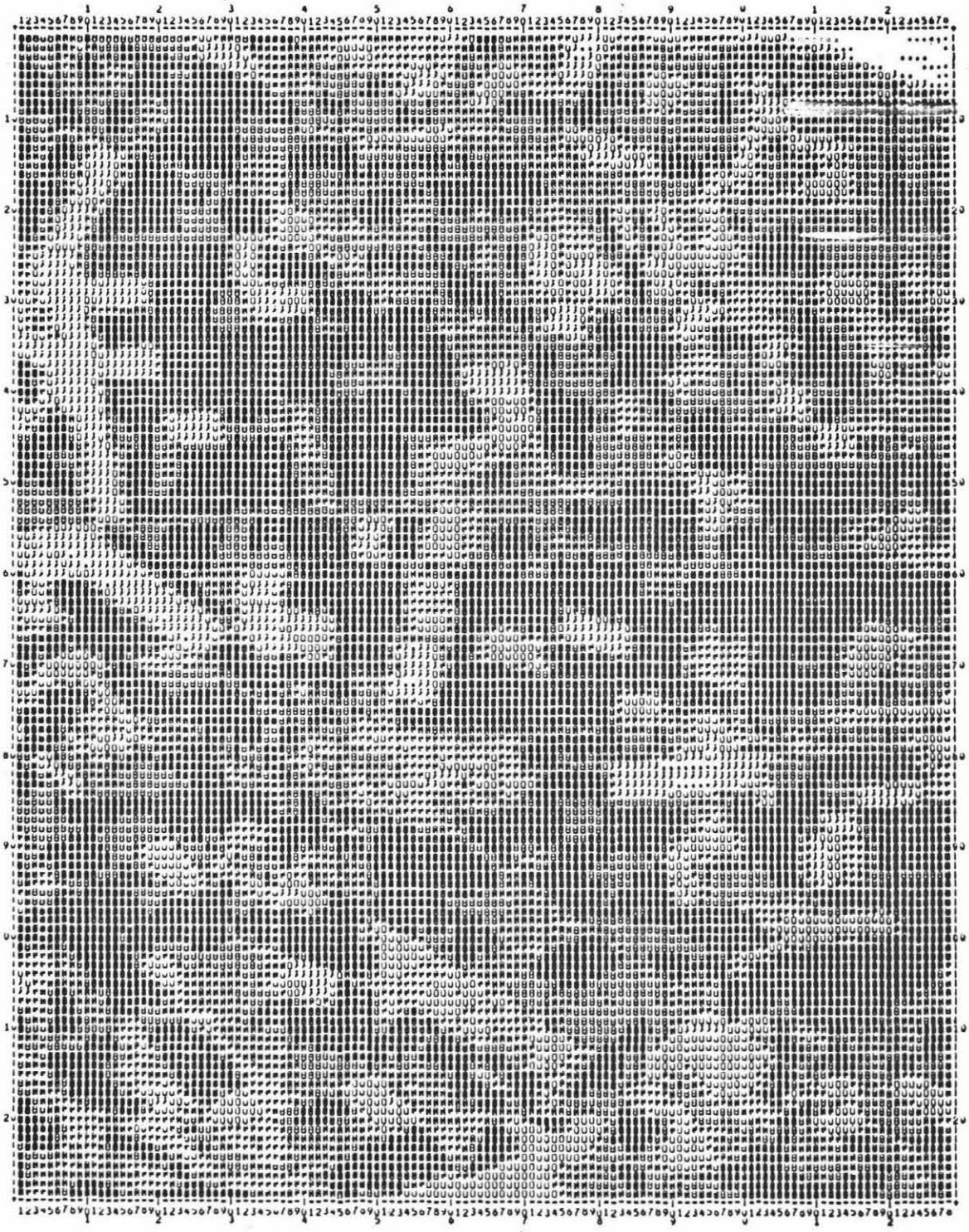


Fig 4. The output for routine SHADES for Band 7. A shaded version of fig 3a.

Histograms

HISTOGRAMMER, see ANNEX E, produces a histogram of the intensity levels for any one file. Of particular interest are the maximum and minimum intensity levels recorded for each file. Fig 5 shows the histogram produced for Band 7.

Distribution within the Pattern Space

The provision of any visual display of the distribution of the data within the pattern space is limited by the fact that the space is four-dimensional. The best that can be done is to use projections of the space on to two-dimensional planes. A routine, DISTRIBUTION/ANALYSER (see ANNEX F), was written to do this. As output from this routine shows only two dimensions or files, it must be run against every possible combination of pairs of files, i.e. 6, to display all the information. The number of possible intensity levels conveniently corresponds to the number of pixels along an edge of the images, i.e. 128. So either ALLLEVELS or SHADES may be used to print out the output of DISTRIBUTION/ANALYSER. In fig. 6 the correlation between Bands 4 and 6 has been displayed using SHADES, regions of high density are confused because of the reduction in dimensionality.

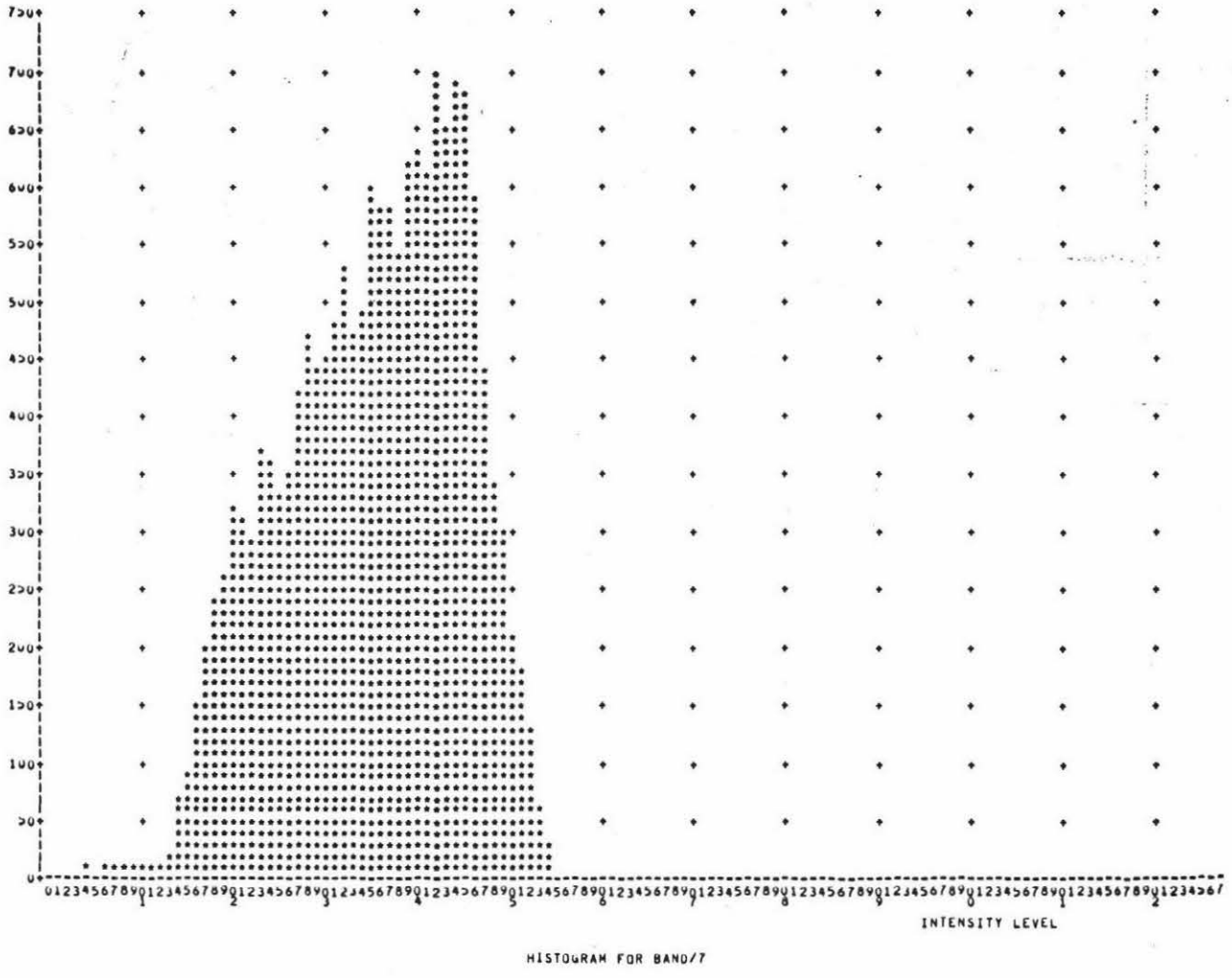


Fig 5. The output produced by HISTOGRAMMER for Band 7.

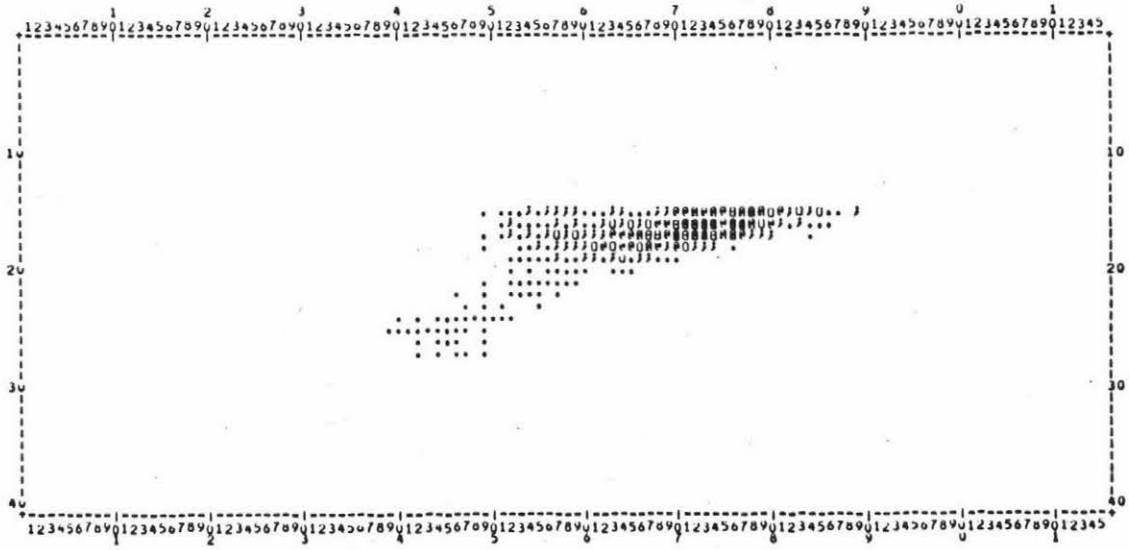


Fig 6. The output produced by DISTRIBUTION/ANALYSER for Bands 4 and 6 showing the correlation between the two wavebands.

6 EDGE DETECTION

6.1 Fuzziness

Since each ground resolution element, or pixel, covers a square area of side approximately 79m, the integrating effect of the sensors will merge adjacent ground features at their boundaries producing indistinct or fuzzy edges on the images. These may be seen in fig 4. In the pattern space these will appear as points between the clusters, forming 'bridges' and as a result the clusters' separations may be confused. Consequently it was decided to investigate the removal or ignoring of such boundary points from consideration in determining the clusters and cluster centres, so that only reliable data is used. An attempt may be subsequently made to classify these boundary points according to the calculated cluster characteristics. It is important to note that this was not an attempt to define enclosed areas in the image which may then be regarded as homogeneous features. Jayroe et al [g] use information from such bounded areas as starting points for cluster centres.

The principle used in edge detection is simply to detect any sudden changes in intensity level in the image, [1]. A derivative operator will produce high values where edges of this sort exist. Since the images are made up of discrete values the differentiating is carried out by differencing between adjacent values.

6.2 Differentiating

A routine called DIFFERENTIATOR (see ANNEX G) scans first rows and then columns of an image, $x_{i,j}$, to produce a further image in which a point $y_{i,j}$ is defined by:

$$y_{i,j} = |x_{i,j} - x_{i,j+1}| + |x_{i,j} - x_{i+1,j}|$$

where: i is the row number and

j is the column number of the point in the image

i.e. the sum of the differences across a row and down a column. The image may then be printed out using either ALLLEVELS or SHADES. Fig 7 shows the results of differentiating the image for Band 7 as printed by SHADES, and may be compared with fig 4.

6.3 Finding Boundaries

Some decision needs to be made concerning the differential's value below which it may be regarded as insignificant and therefore ignored, and above which it indicates the presence of an edge or boundary. For this, routine BOUNDARYFINDER (ANNEX H) was written to perform the following algorithm. Each row of the image is scanned, if any pair of adjacent points differ by more than $T\%$ (of one of their values) then part of a vertical edge is considered found and one of the two points, say the leftmost, is 'marked' as a boundary point. The threshold value, T, is a parameter supplied to the routine. Next each column is scanned and, once again, where adjacent points are found to differ significantly a boundary point is considered found, this time corresponding to part of a horizontal edge. The rows and columns are in fact scanned in both directions to average out any directional bias.

Up to this stage the difference was taken to be the absolute value of the aritmetic difference between the two values; if it was greater than $T\%$ of the larger value (and therefore both values) then it was regarded as 'significant'. However, this takes no account of the range of values encountered in the image. If high values are being compared, the difference is much less likely to be significant than if low values are being compared. A

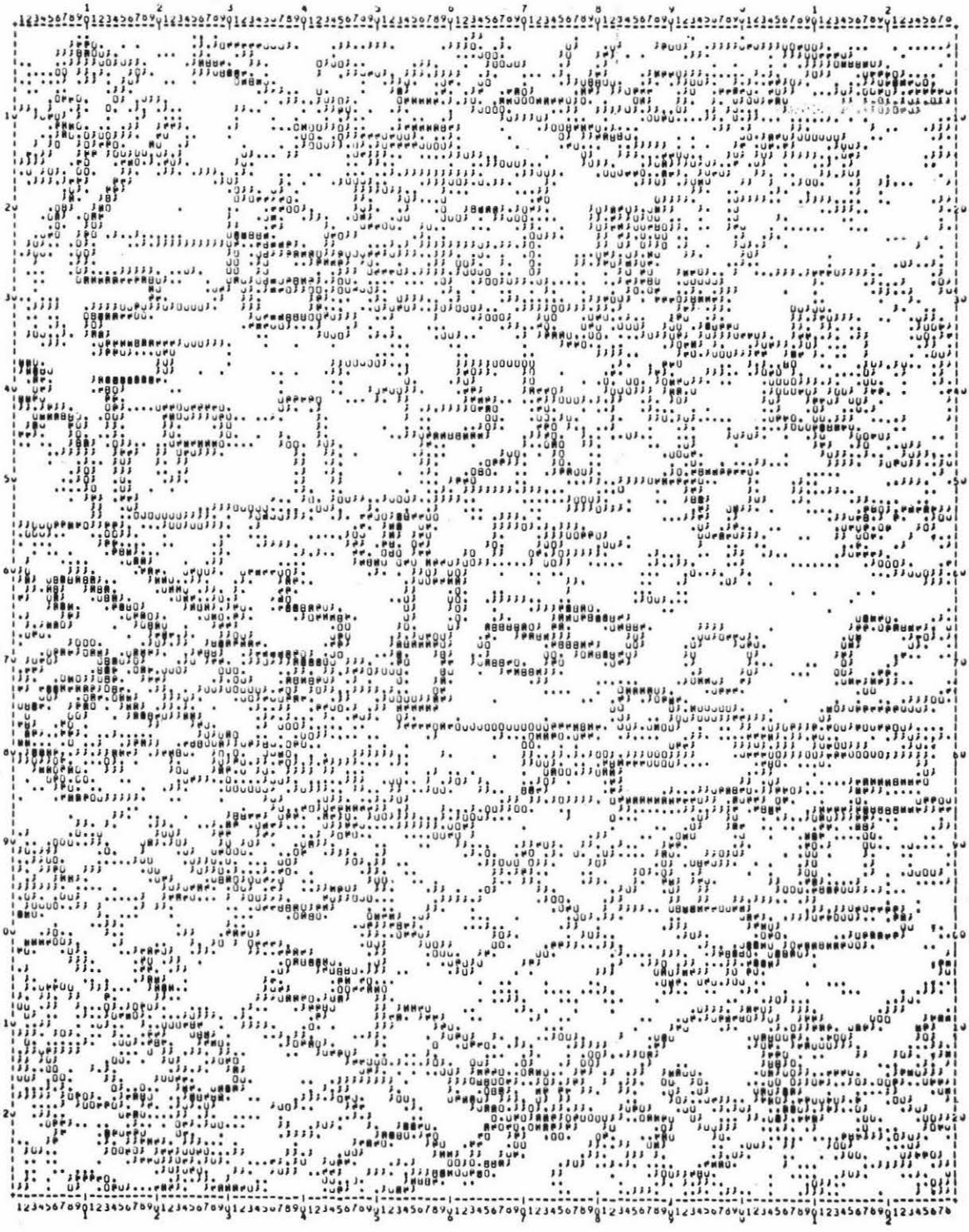


Fig 7. The output produced by DIFFERENTIATOR for Band 7, showing the edges detected.

modified test was tried in which the difference of each point pair was compared with *T%* of the median value for that entire image, i.e.:

$$(maxvalue + minvalue)/2$$

Various values of *T* were tried in an effort to find an acceptable balance between:

a. *Too many points* being identified as boundaries thus removing information from being used in the later stages of clustering and classification; and

b. *Too few*, thus not removing all of the 'unreliable' information, the overall goal of this edge detection.

Fig 8 shows the edges detected in the top left hand corner of the test area for the four files for $T = 10\%$. Fig 9, similarly, shows the results for $T = 15\%$.

It may have been noted that these methods are most sensitive to edges which lie parallel to the image axes, i.e. those previously referred to as 'horizontal' and 'vertical' edges. More complex algorithms exist which are also sensitive to edges in other orientations. These are intended for applications where the edges themselves are of importance. In this project, however, the concern is only to remove data which is probably 'unreliable'. It is considered, therefore, that the algorithms used are sufficiently comprehensive for this purpose.

6.4 Correlating Boundary Information

There is one image of boundary points from each of the four wavebands, and since each ground feature's boundaries will be more distinct in one particular image - depending on its own signature and that of its neighbour-

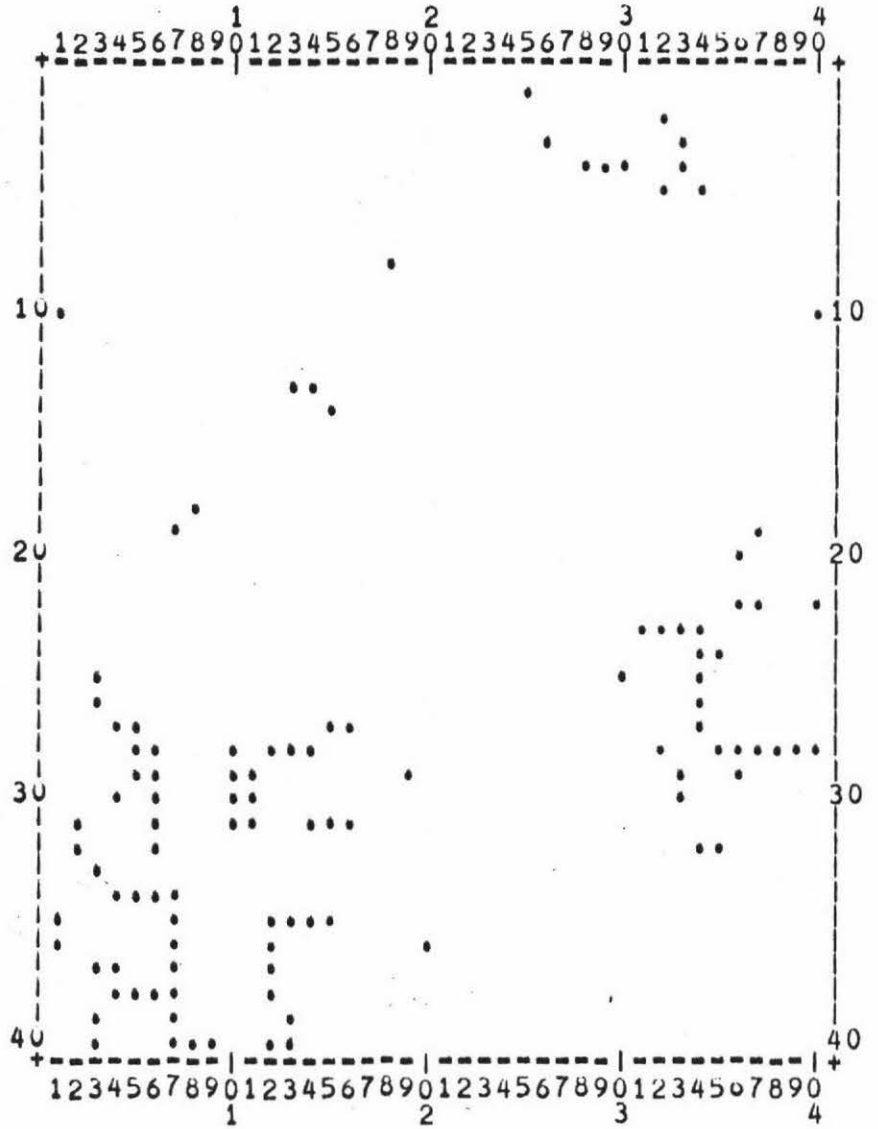


Fig 8a. Boundary points detected in Band 4 for $T = 10\%$.

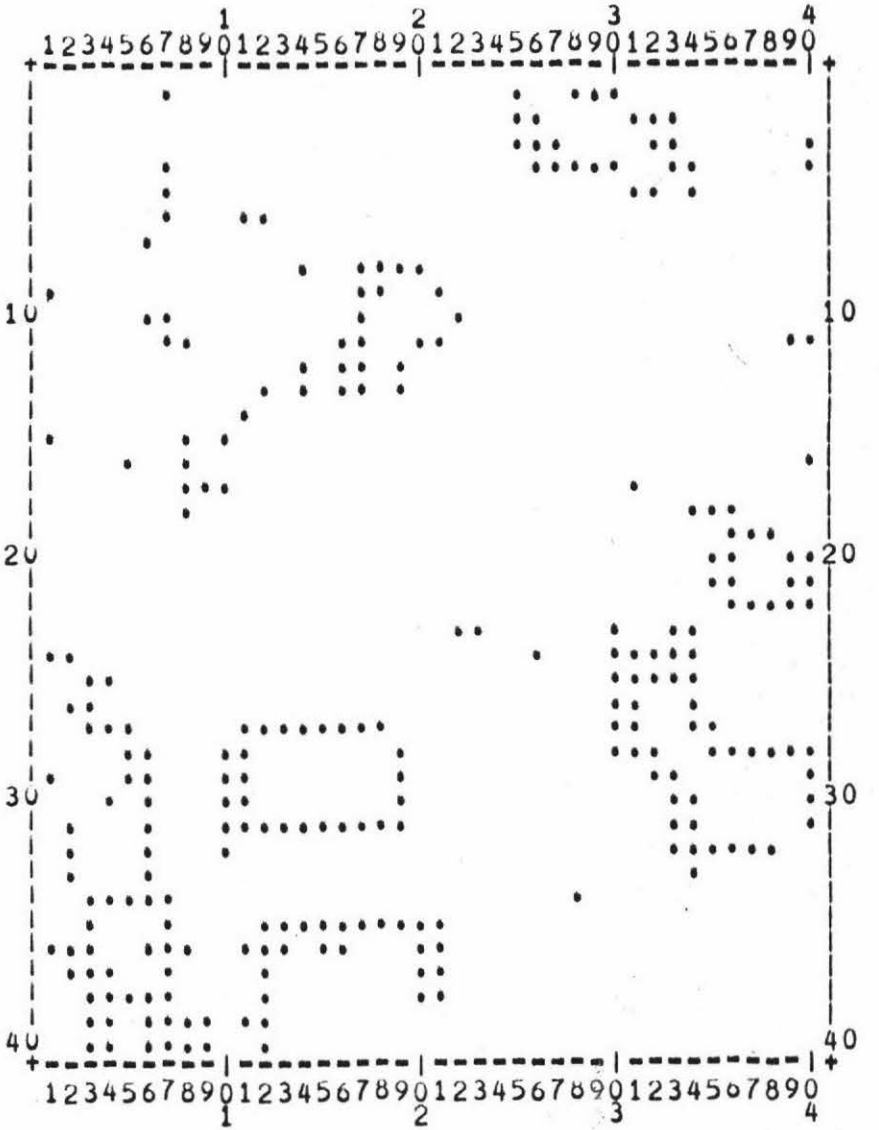


Fig 8b. Boundary points detected in Band 5 for $T = 10\%$.

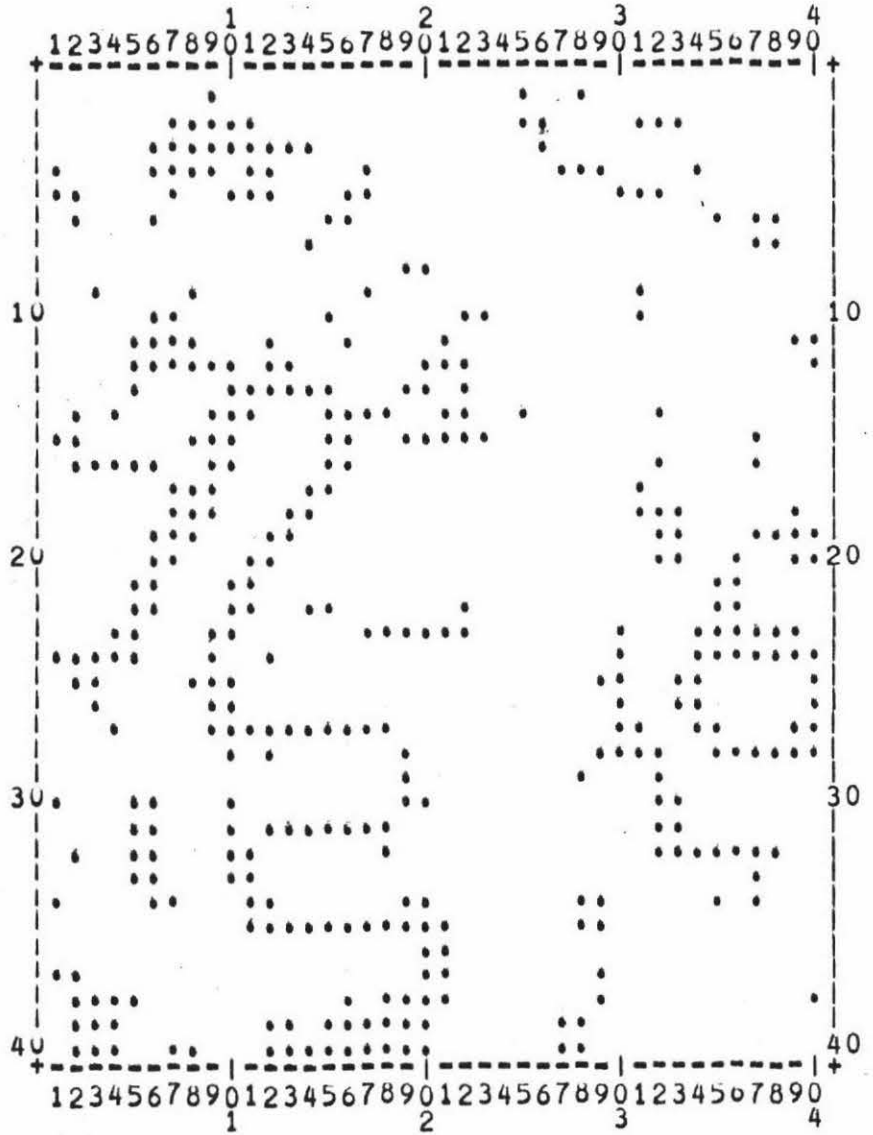


Fig 8c. Boundary points detected in Band 6 for $T = 10\%$.

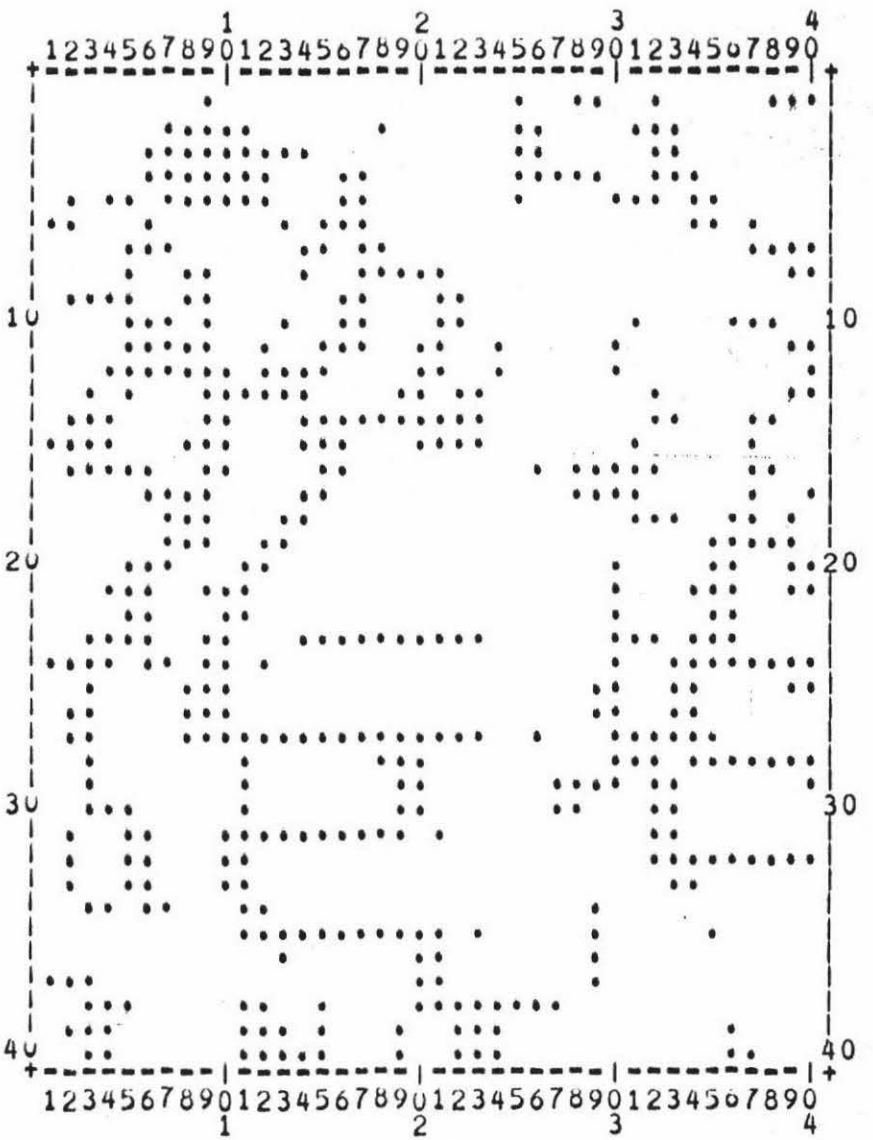


Fig 8d. Boundary points detected in Band 7 for $T = 10\%$.

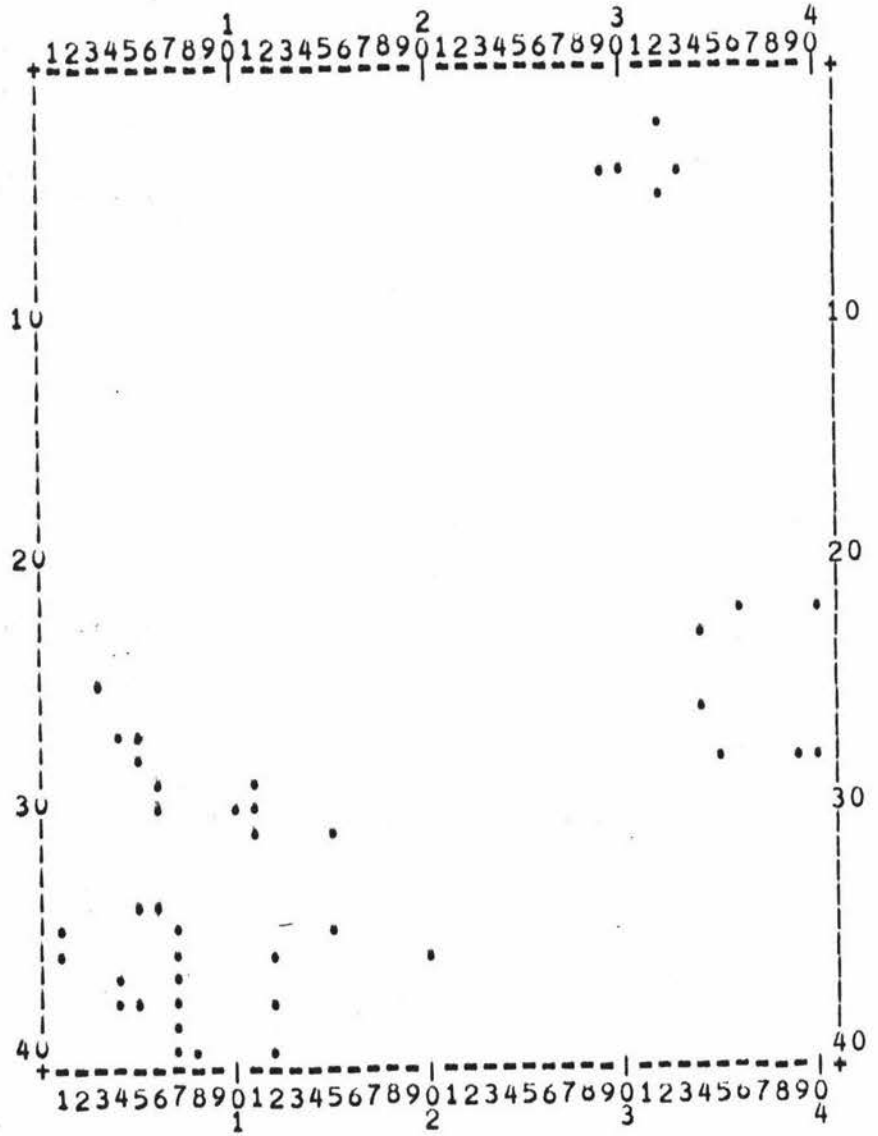


Fig 9a. Boundary points detected in Band 4 for $T = 15\%$.

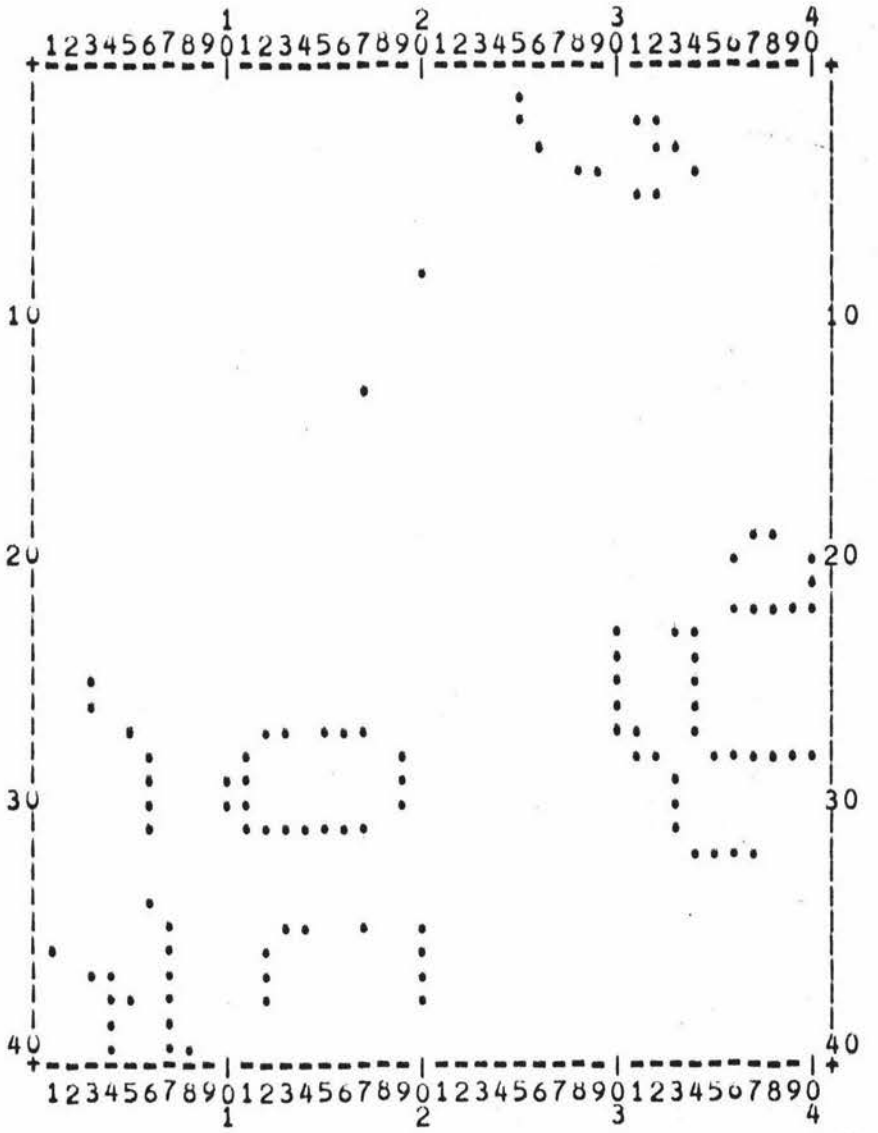


Fig 9b. Boundary points detected in Band 5 for $T = 15\%$.

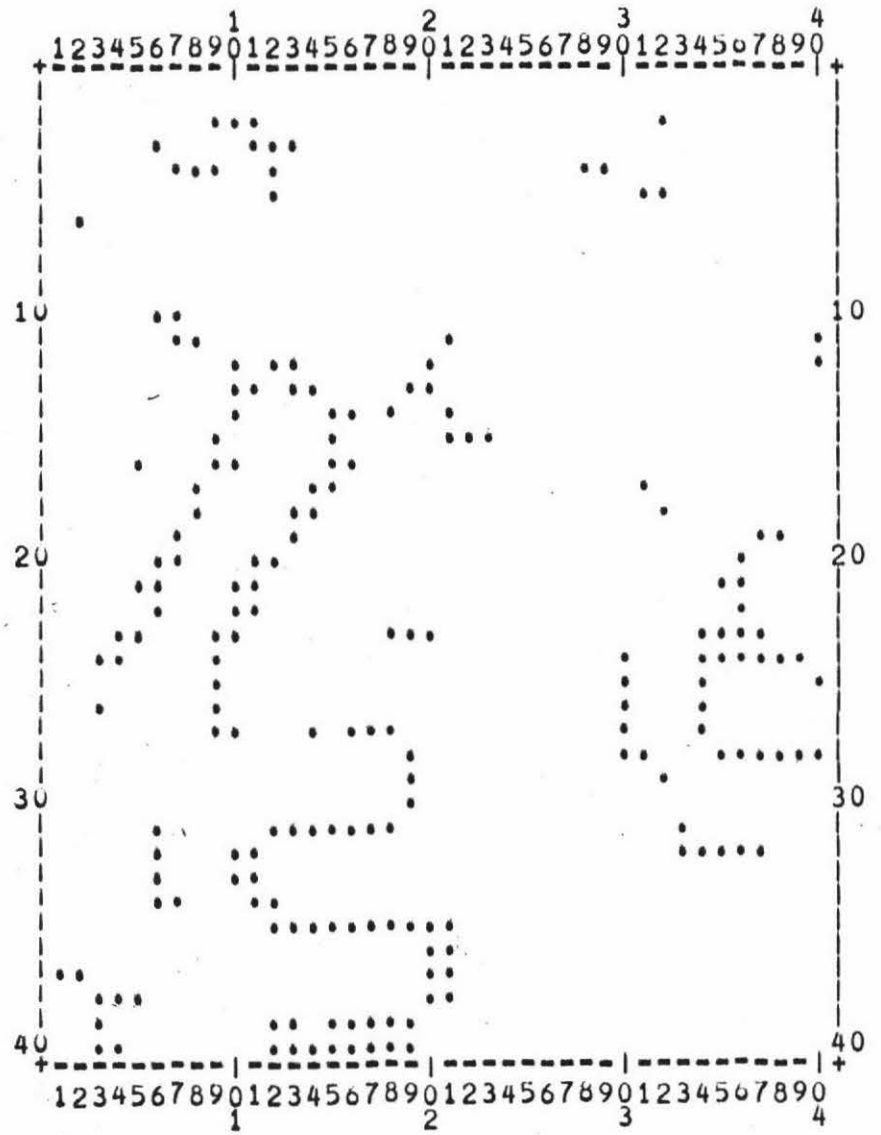


Fig 9c. Boundary points detected in Band 6 for $T = 15\%$.

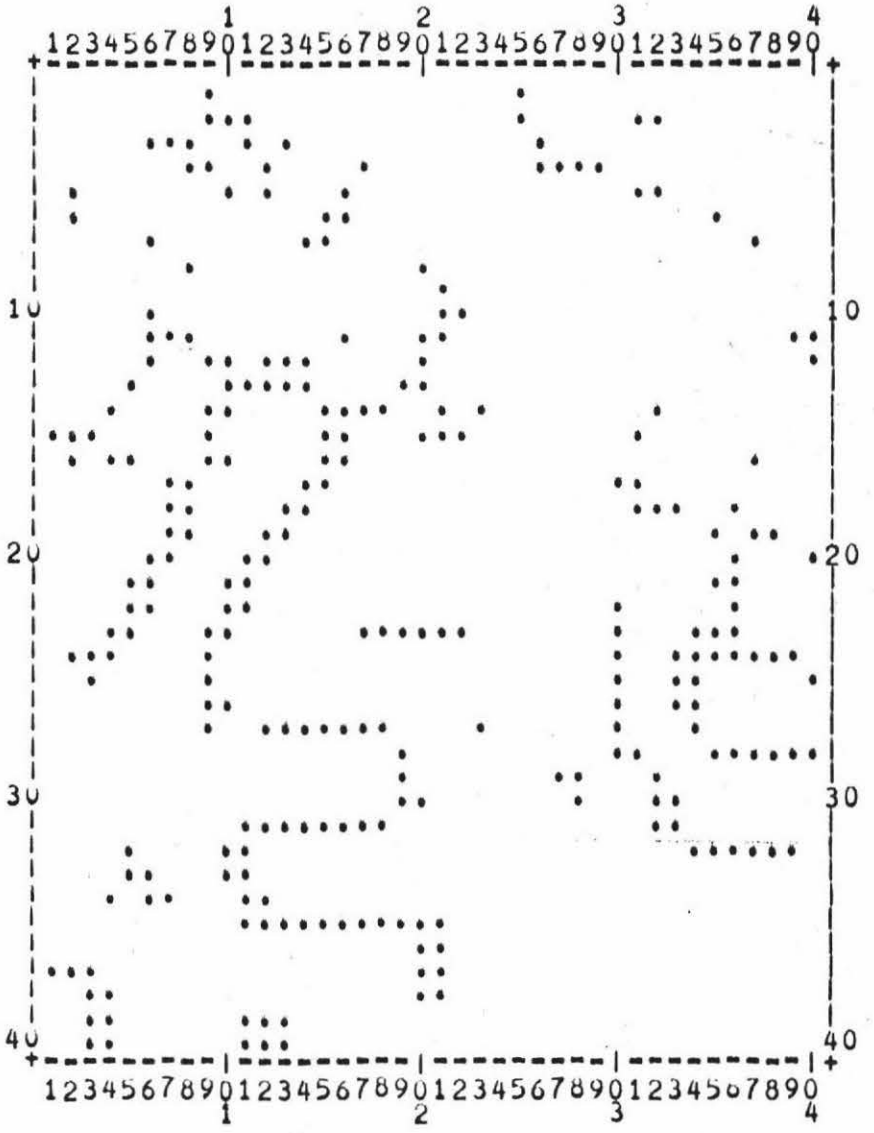


Fig 9d. Boundary points detected in Band 7 for $T = 15\%$.

ing features - the complementary information from all four boundary images needs to be combined in some way.

The routine BOUNDARYMERGER (ANNEX I) was written to experiment with various methods of combining the four boundary images into one. Some of the combinations tried were to:

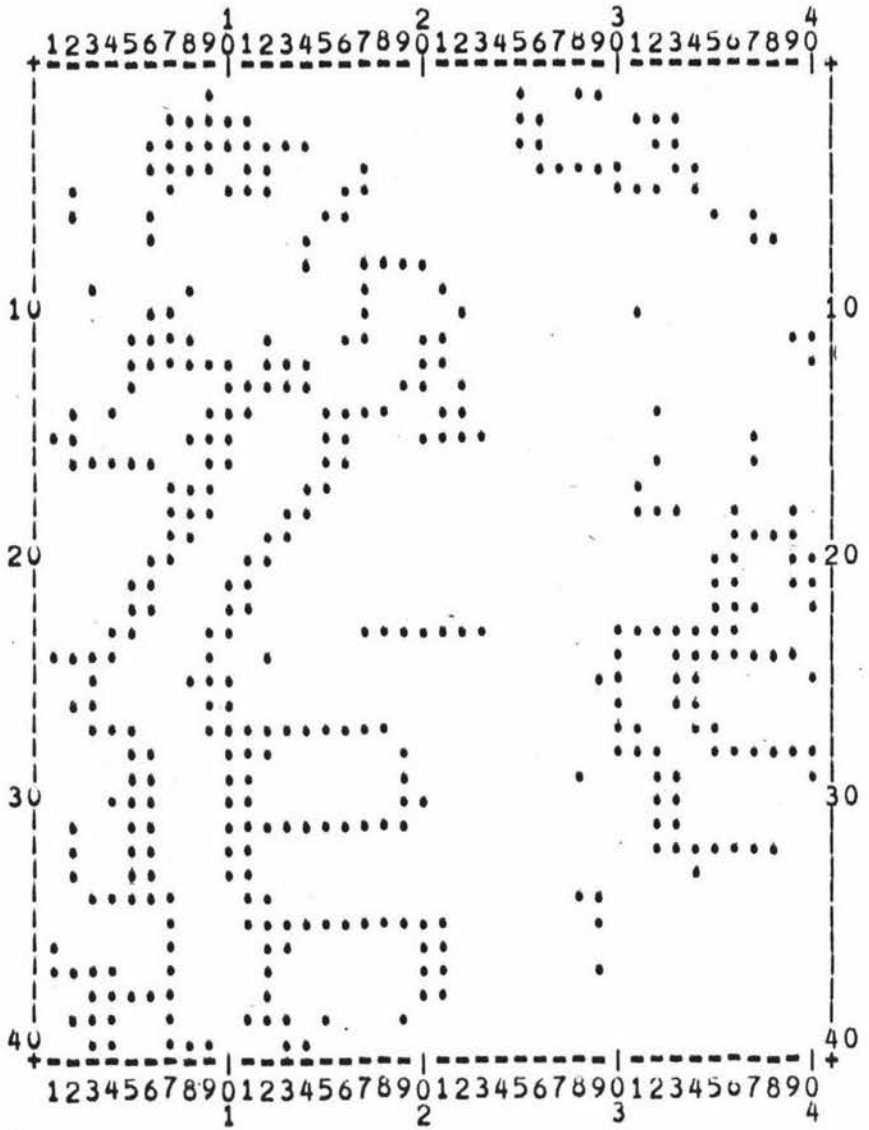
a. Include those points which occurred in any of the four images.

b. Include those points which occurred in at least two of the images.

c. Include those points which occurred in at least three of the images.

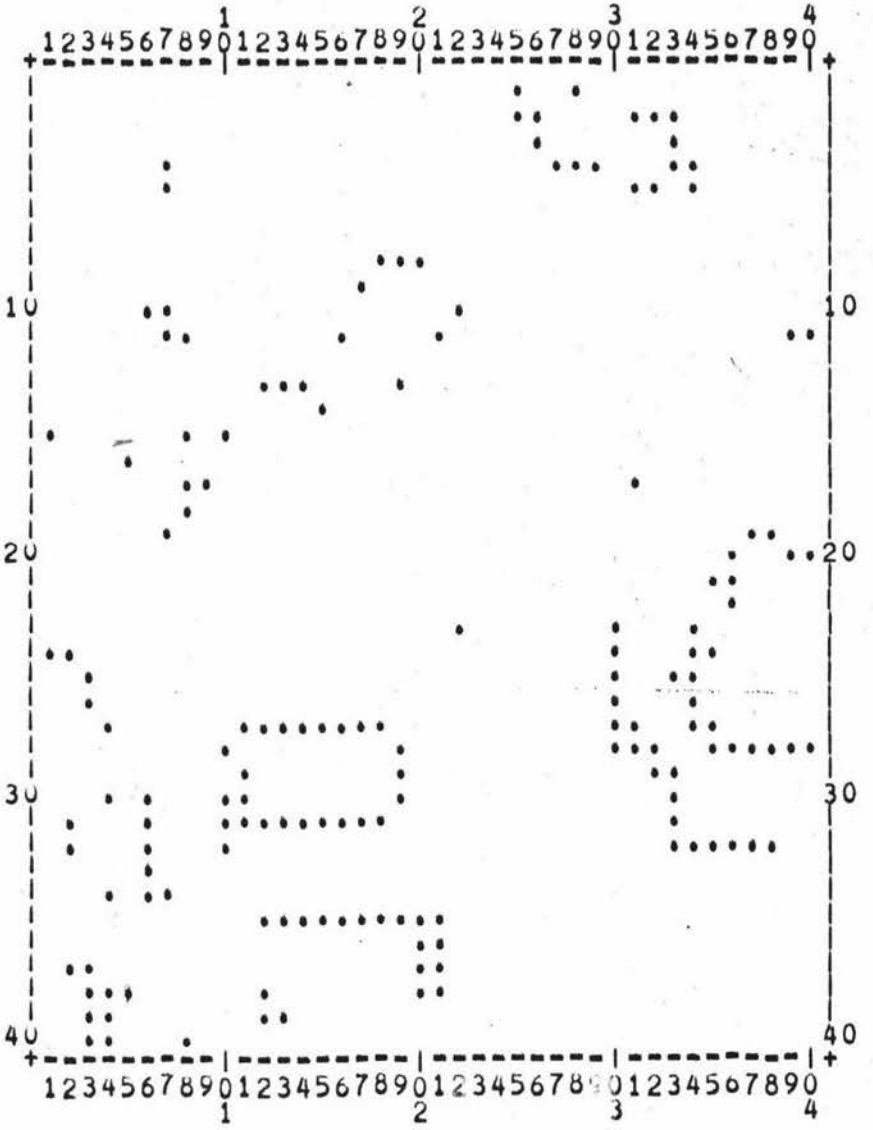
d. Include only those points which occurred in all of the four images.

Fig 10 shows some of the combinations for boundary points detected using $T = 10\%$, and fig 11 shows the corresponding images for $T = 15\%$. Comparisons with fig 4 are now less meaningful since these images also include information from the other three bands. Once again a balance is needed between including too many boundary points and too few. With this in mind the combination chosen was that which included points common to at least two images for $T = 15\%$ i.e. fig 11a.



24% ARE BOUNDARY POINTS

Fig 10a. Merged boundaries for $T = 10\%$. Boundary points appearing in at least two of the four files (fig 8) are included.



10% ARE BOUNDARY POINTS

Fig 10b. Merged boundaries for $T = 10\%$. Boundary points appearing in at least three of the four files (fig 8) are included.

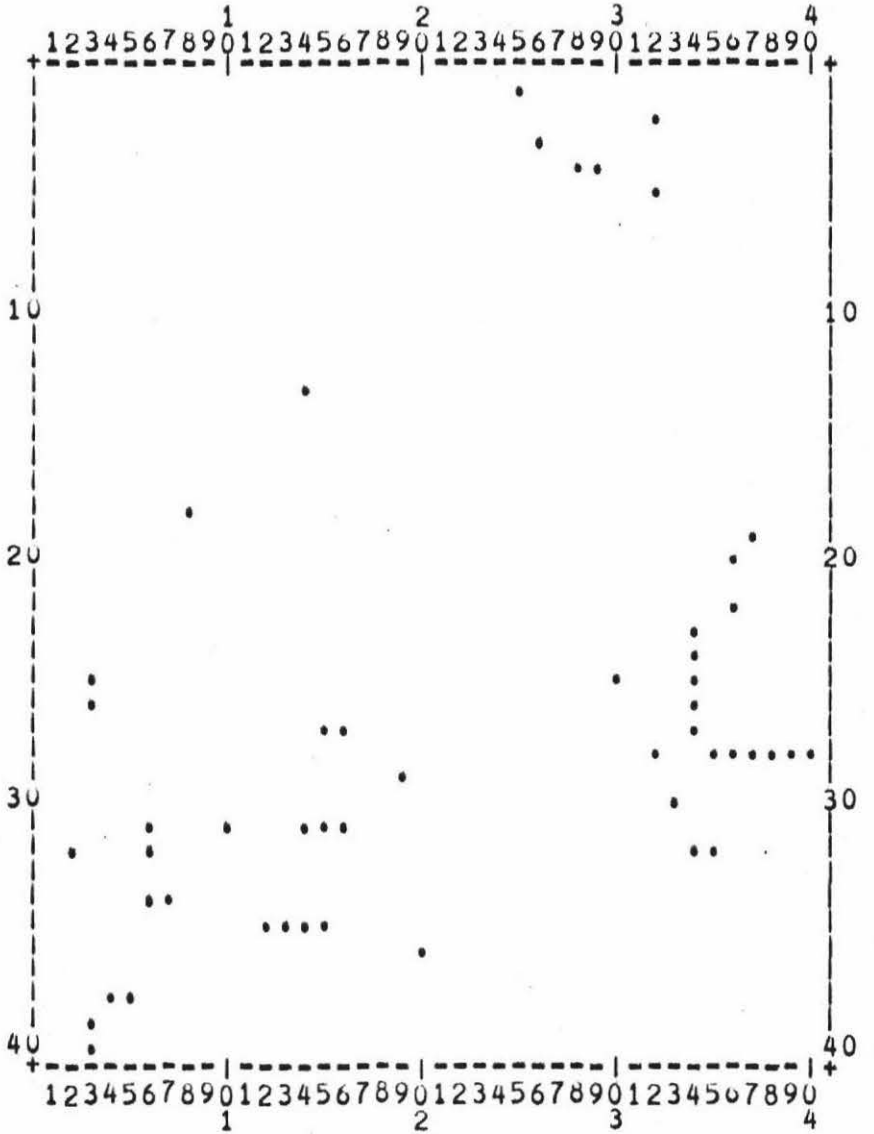


Fig 10c. Merged boundaries for $T = 10\%$. Only boundary points appearing in all four files (fig 8) are included.

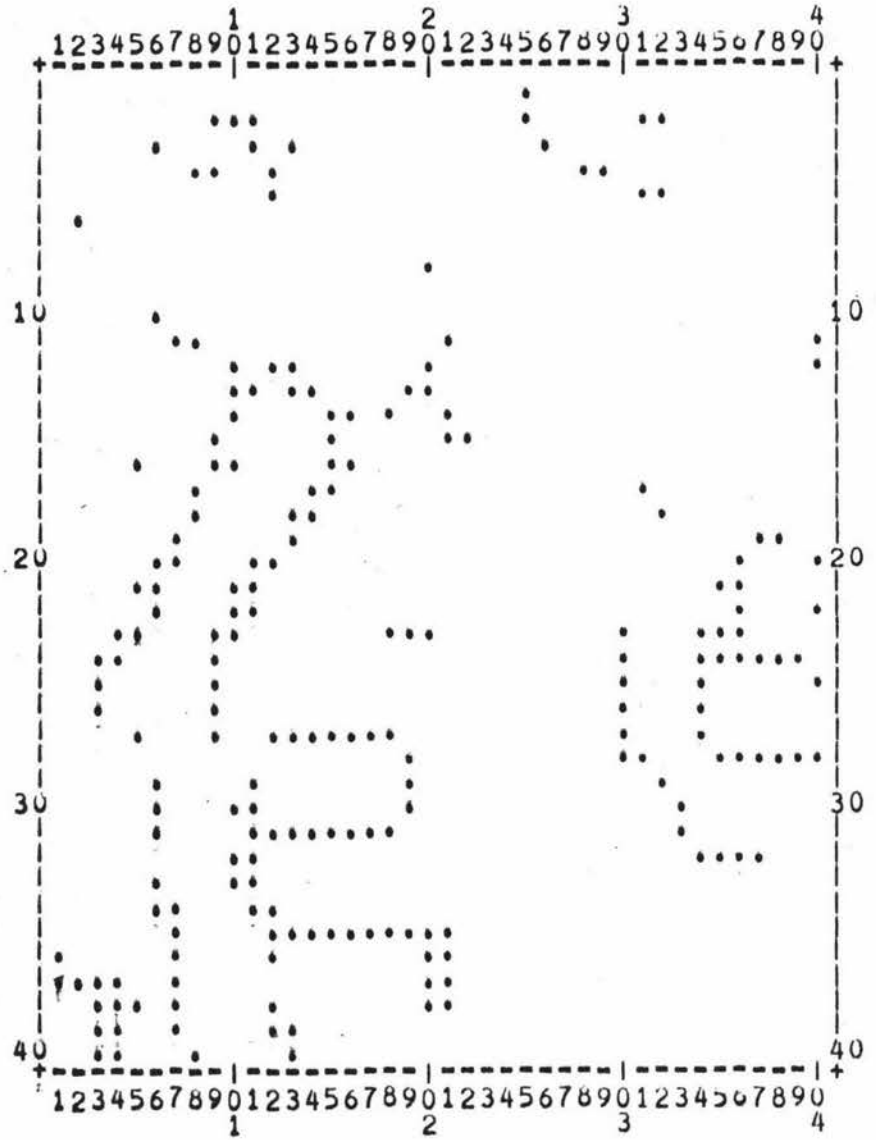


Fig 11a. Merged boundaries for $T = 15\%$. Boundary points appearing in at least two of the four files (fig 9) are included.

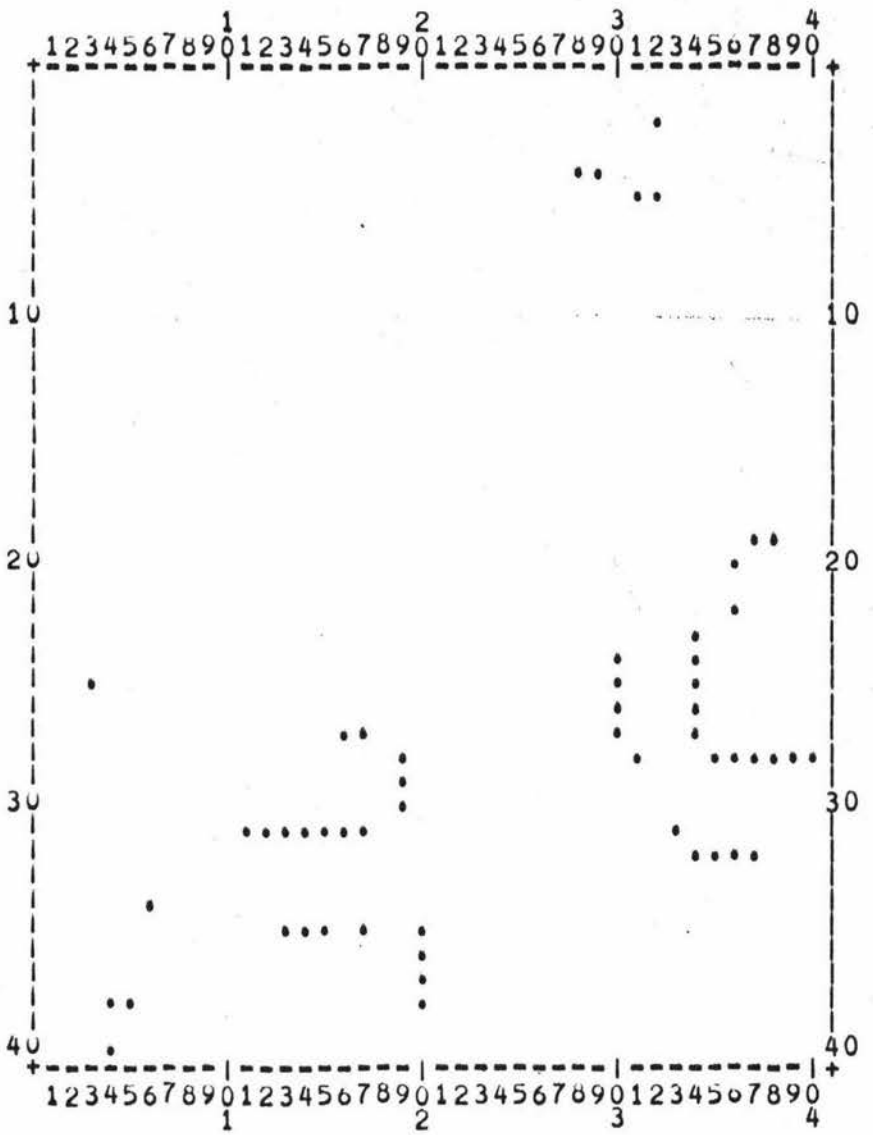


Fig 11b. Merged boundaries for $T = 15\%$. Boundary points appearing in at least three of the four files (fig 9) are included.

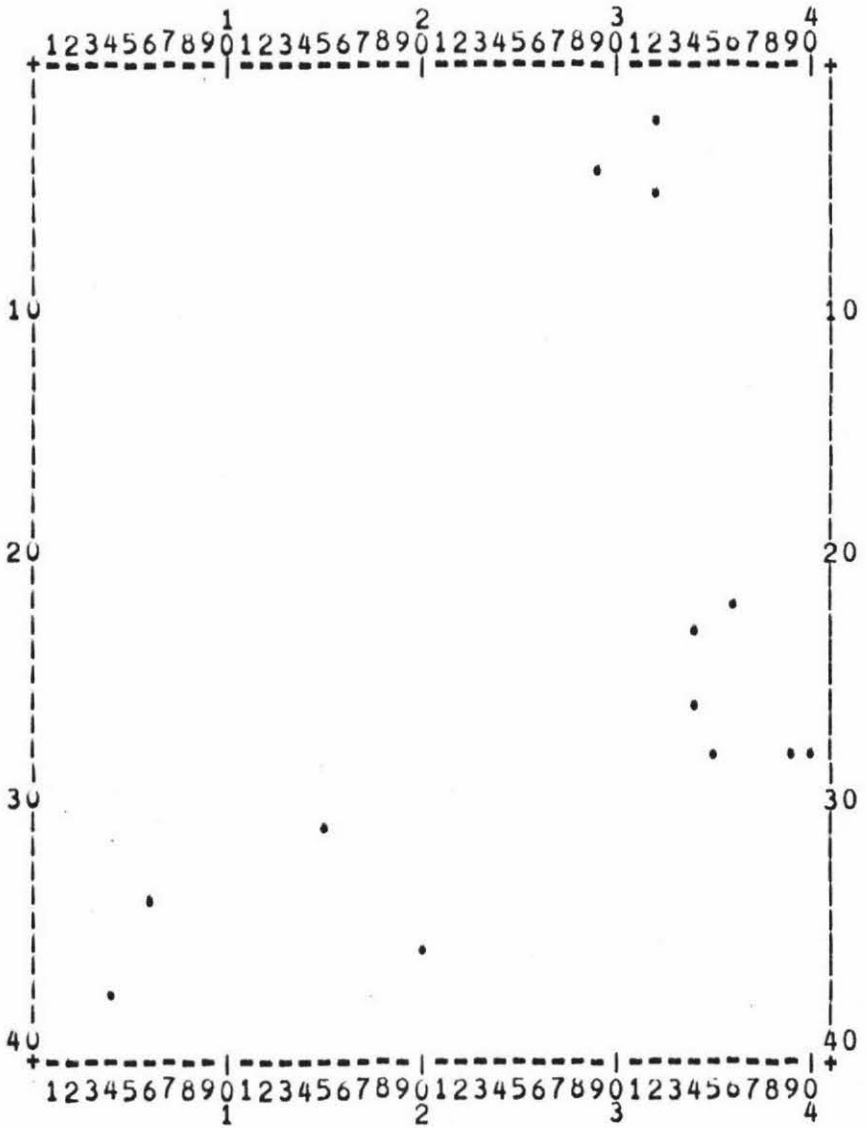


Fig 11c. Merged boundaries for $T = 15\%$. Only boundary points appearing in all four files (fig 9) are included.

7 THE ALGORITHMS USED

Details of the implementation of the three clustering algorithms outlined in section 4.3 are now described.

7.1 Shared Near Neighbour

A program to perform the Shared Near Neighbour algorithm, see ANNEX J was written. The unsuitability of the algorithm for use on large data sets such as this one became apparent. If there are two neighbouring clusters it is possible - in fact it is most probable - that some points will lie between them, forming a 'bridge', the reasons for this were given in 6.1. As k_t is made smaller these points will become similar enough to points near to the edge of one of the clusters, and thus are classified as belonging to that cluster. If k_t is too small they will, in a similar way, be also classified as belonging to the other cluster. The effect is that both clusters are then regarded as being in the same class, i.e. the system's discernment is impaired. This is a result of both the presence of such bridges, and the system's sensitivity to the parameters k and k_t. It is suggested that these parameters may be altered according to information concerning the separation between the point and its k nearest neighbours, e.g. k_t could be determined from the number of nearest neighbours lying within a certain distance.

In fig. 12 the classified image is shown corresponding to a small area on the left side and slightly below the top of the test area, in all cases k, the number of nearest neighbours being considered, is 20. Note in particular the sensitivity of the number of clusters to k_t.

The most restricting feature of this method arises from the computation requirements. Since for n points there are n^2 distance calculations and up

to $n^2 \times k^2$ comparisons. It is true, though, that the calculation and tabulation of nearest neighbours need only be performed once, subsequent to which various values of k_t may be used as required. Further, nearest neighbour calculations may be optimised by use of algorithms such as the Branch and Bound algorithm described in ref. [m]. However, the restriction imposed by resource requirements proved to be serious, e.g. the CPU time for the 20 × 20 pixel images in fig. 12 was about 75 secs, and increasing the image to only 24 × 24 pixels would double the CPU time.

7.2 MAXFINDER - the Divisive Approach

Procedure MAXFINDER subdivides the pattern space into cells of a specified size, see ANNEX K.It then searches for and records those cells whose population is greater than that of all immediately adjacent cells in all directions. These are the required maxima.

Unfortunately no suitable compromise could be found between the cell size being:

a. *Too large*, in which case the detail is obscured, i.e. more than one cluster may be contained within one cell. This will be particularly troublesome where clusters with markedly different populations are adjacent in the pattern space. And

b. *Too small*, in which case little progress has been made - in the limit the cell edge (and therefore the volume) is unity, and the pattern space is unaltered. In these cases, e.g. for cells of edge 2 or 3, several dozen local maxima were found.

Overall, the inflexiblity of the cell's shape and size proved to be too limiting.

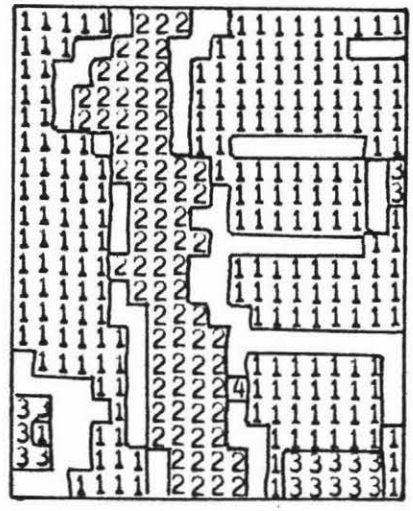


Fig 12a. Clustered output from SHAREDNN for $k = 20$, $k_t = 12$.

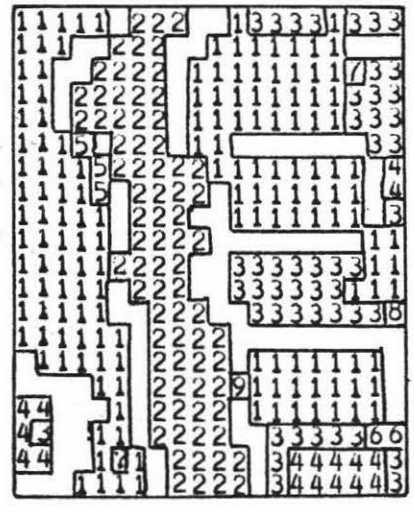


Fig 12b. Clustered output from SHAREDNN for $k = 20$, $k_t = 13$.

7.3 CENTREFINDER - the *K*-means Approach

A procedure, CENTREFINDER, was written to implement initially only the basic k-means algorithm, see 4.3.3 and ANNEX L. Values for k and the initial cluster centres were chosen largely from experience gained with the data during trials of other algorithms. Some simple modifications were added, first a procedure to remove any clusters found to have zero population. This occurred occasionally during the second iteration but most often during the first, indicating an inappropriate initial value for a cluster centre.

The 'clean' data set produced by BOUNDARYFINDER and BOUNDARYMERGER (see 6.3 and 6.4) was used in subsequent versions of CENTREFINDER, thus the boundary points were not used in determining the cluster centres. Values for the spread of each cluster were also calculated during the iterative phase, these are simply the distances from each cluster centre to its furthest member point. Thus:

$$m_j = d(\bar{z}_j, \bar{x}_i)$$

where: m_j is the value of spread for the jth cluster, and

$$d(\bar{z}_j, \bar{x}_i) > d(\bar{z}_j, \bar{x}) \quad \text{for all } \bar{x} \in S_j,\ \bar{x} \neq \bar{x}_i$$

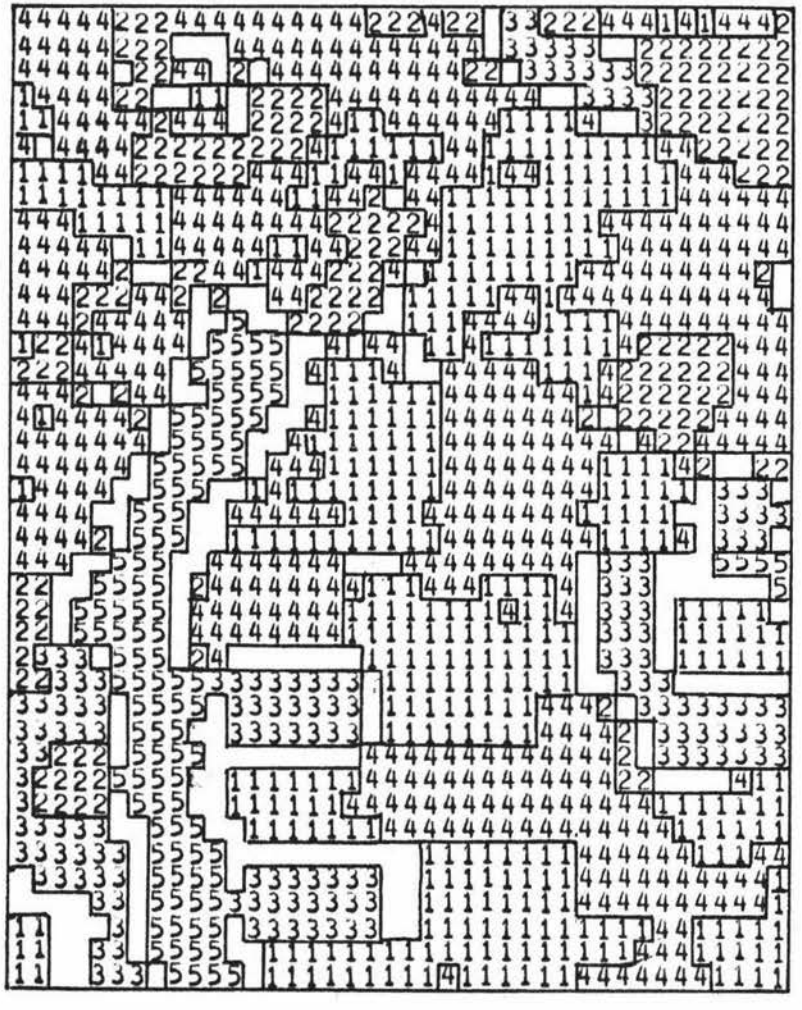
After the iterations are completed a pass is made through the set of boundary points and an attempt made to classify them according to the final cluster centres. First the nearest centre to a boundary point is found, then if the point is within the spread for that cluster it is classified with that cluster, otherwise it is rejected as unclassified.

In fig 13a and 13b the clustered output from CENTREFINDER is shown for a

In fig 13a and 13b the clustered output from CENTREFINDER is shown for a 40 × 40 pixel image. The basic algorithm produced the results in fig 13a after 17 iterations, and the attempt to classify the boundary points produced fig 13b. Total CPU time was less than 28 secs. Fig 11a has been repeated for easier comparison and fig 13c shows the available ground truth for the corresponding area, compare also with the frontispiece.

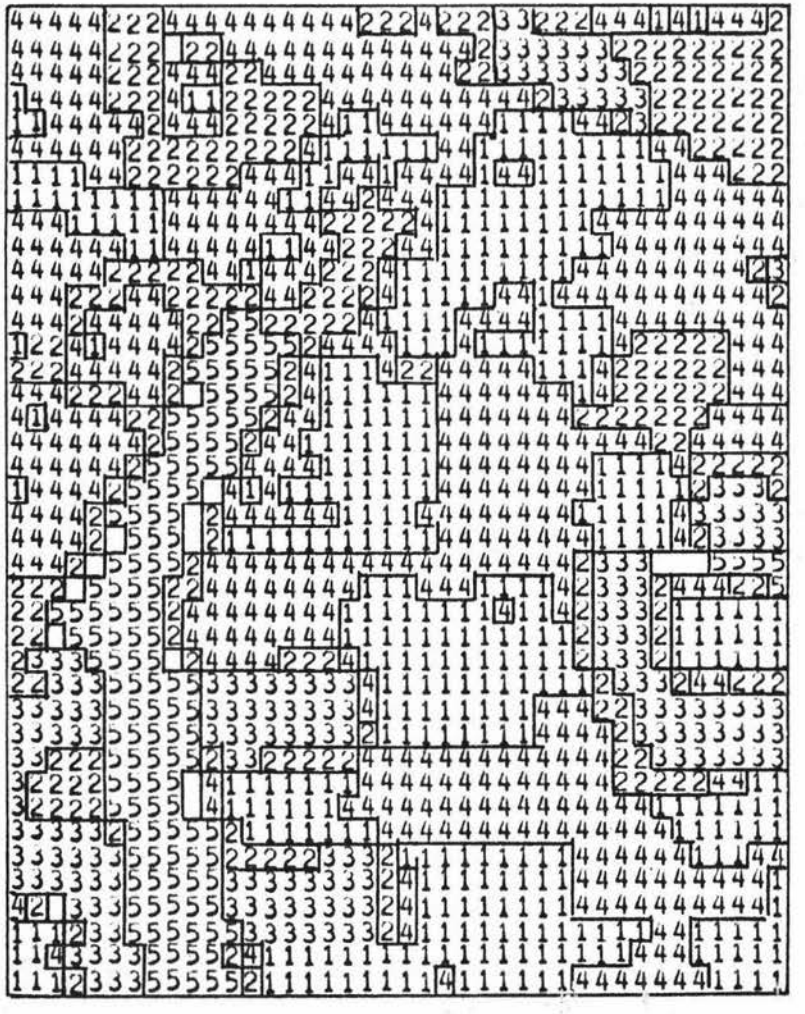
The distinction of forest and worked fields is very clear in both the satellite images and the Infra-red aircraft image (frontispiece). However, the distinction between various types of pasture is not as clear. Since this is true for both types of imagery it may indicate a limit to the recognition possible using this type of data.

The attempt to classify boundary points has met with partial success - about half the points being classified meaningfully i.e. in the same class as one of their neighbouring points. Considering the simple-minded approach used this is regarded as quite acceptable.



12.6% ARE BOUNDARY POINTS

Fig 13a. Clustered output from CENTREFINDER.



1.0% ARE BOUNDARY POINTS

Fig 13b. Clustered output from CENTREFINDER, with boundary points classified.

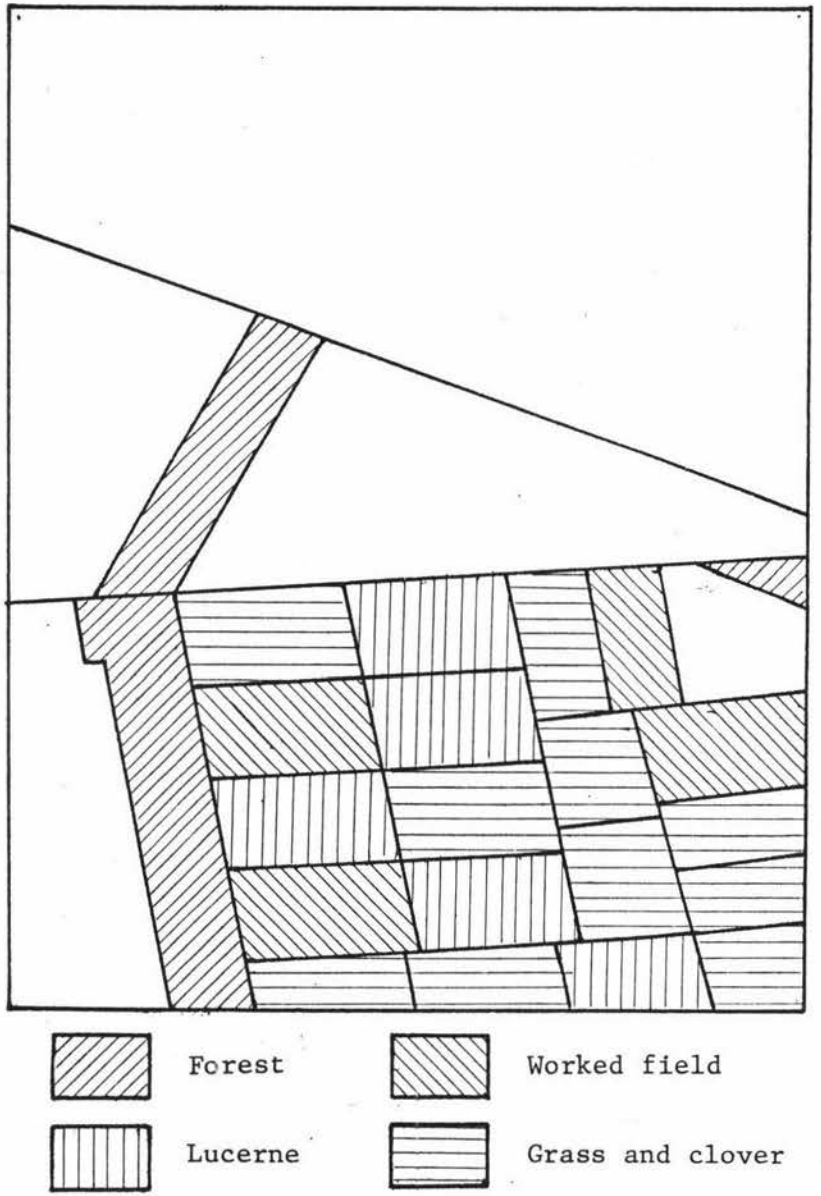


Fig 13c. Ground truth for the test area.

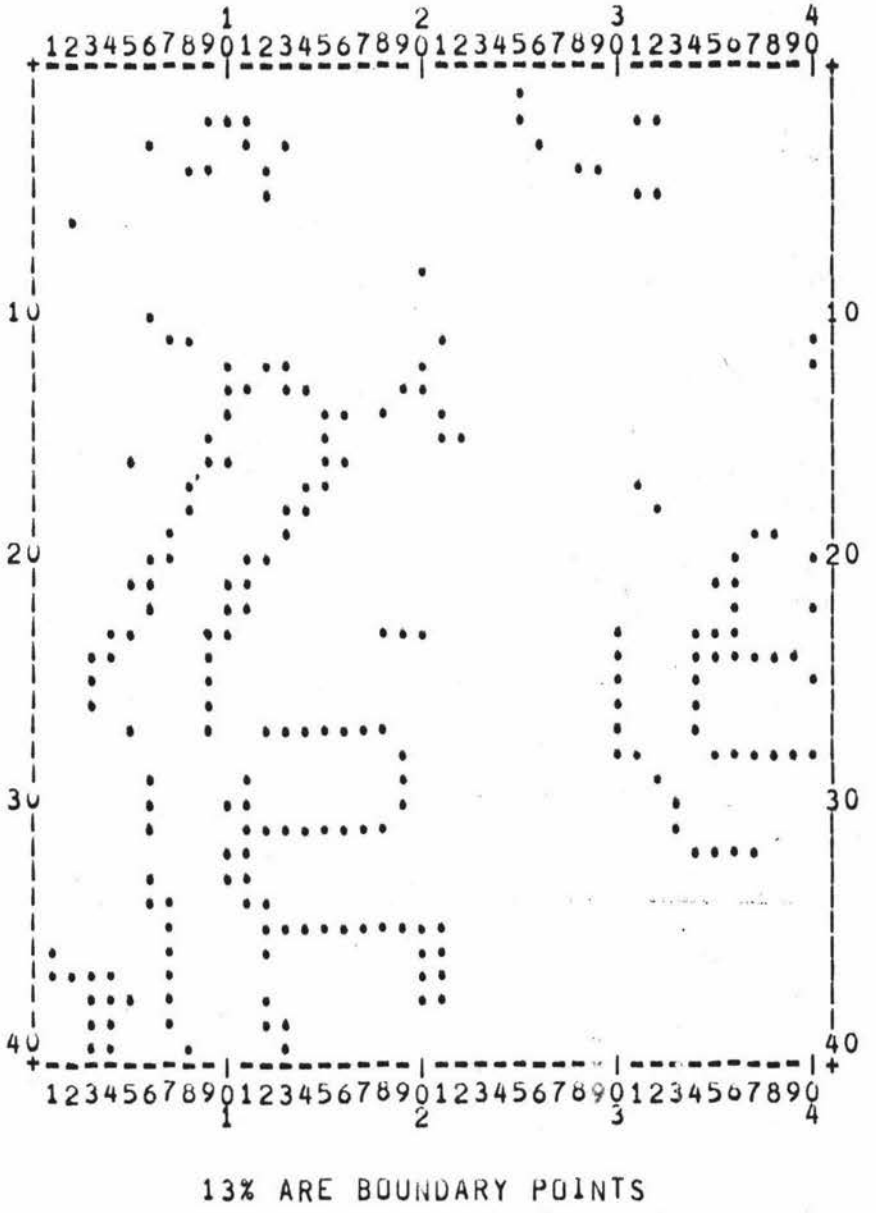


Fig 11a. Merged boundaries for $T = 15\%$. Boundary points appearing in at least two of the four files (fig 9) are included.

8 PROJECT ASSESSMENT

8.1 Practical Problems

When the data sets under consideration are so large there will almost inevitably be some difficulties due to resource limitations. Thus although the rate of correct classification remains the foremost criterion for the effectiveness of the method, the 'practical' considerations, e.g. storage requirements and computation times, may prove limiting. There are compromises available between one resource and another, e.g. instead of reading the data into arrays and storing it there throughout the computation, the files could be re-read. This is a trade-off between storage and I/O requirements.

The realities of a practical situation, however, are that in order to process such a large amount of information a correspondingly large amount of computing resources are indeed required.

8.2 Boundary Points

The incidence of unreliable data points referred to as boundary points has been found to be about 13% for this data set. These boundary points lie outside the hypersurfaces bounding the clusters in the pattern space. An agglomerative method of clustering, such as the Shared Near Neighbour method, will find these points troublesome since, through the chaining effect (see 7.1) all the clusters become grouped into one single class. The prior removal of boundary points will lessen the chance of such difficulties. In the *K*-means method, however, their effect is quite different. Cluster centres are calculated by averaging the member points' values, and since:

a. The boundary points are relatively few, and

b. They are likely to 'surround the cluster,

then it is expected that they will not have a marked effect on the cluster centress' coordinates.

8.3 Classification

Of the methods used, the shortcomings were apparent, and in two cases proved the unsuitability of those methods. The k-means method has shown to be effective, although its simplicity is a limiting factor to final recognition accuracy.

8.4 Overall Success

Comparison of fig 13 with the frontispiece gives an idea of the considerable recogniser accuracy achieved by the method investigated. There is, unfortunately, no direct absolute measure of success since the ground truth available was not sufficiently comprehensive. It is considered that the CPU time needed for recognition is likely to be acceptable even though each iteration involves a pass through the entire data set, although with further development work this could be reduced by e.g. some form of pre-processing to enable more suitable initial values to be chosen.

8.5 Suggestions for Further Work

The area of greatest weakness in current research is probably that of knowledge of the clusters' sizes and shapes. The assumption has been made that all are spherical and roughly equal in size, and in some cases this may be in error. Some method of determining more accurately the cluster statistics ought to be investigated, possibly by following and mapping cluster boundaries by detecting density minima in the pattern space.

REFERENCES

a Rudd, R.D.
Remote Sensing: a Better View. Belmont California, Duxbury Press, 1974.

b Patrick, E.A.
Fundamentals of Pattern Recognition. New Jersey, Prentice-Hall, 1972.

c Jarvis, R.A. and E.A. Patrick
Clustering Using a Similarity Measure Based on Shared Near Neighbours. *IEEE Transactions on Computers*, Vol. C-22, p 1025-1034, Nov 1973.

d Tou, J.T. and R.C. Gonzalez
Pattern Recognition Principles. 1st ed. Reading, Mass., Addison-Wesley, 1974.

e Haidar, M.
Procedures for Analyzing Data Obtained from LANDSAT Multispectral Imagery. Thesis, M.Sc., University of Illinois, 1976, 67p.

f Schell, J.A.
A Comparison of two Approaches for Category Identification and Classification Analysis from an Agricultural Scene. *In* Shahrokhi,F. ed. *Remote Sensing of the Earth's Resources Vol 1*. The University of Tenessee, Tullahoma, Tenessee. p 374-393. 1972.

g Jayroe, R.R. et al.
Computer and Photogrammetric General Land Use Study of Central North Alabama. NASA Technical Report TR R-431, October 1974.

h Weeden, H.A. et al.
Investigation of an urban area and its locale using ERTS-1 data supported by U-'photography.*In Symposium on Significant Results Obtained from the Earth Resources Technology Satellite-1 Vol 1.* NASA/Goddard Space Flight Centre, Greenbelt, Md. p 1015-1022.

i Borden, F.Y.

A Digital Processing and Analysis System for Multispectral Scanner and Similar Data. *In* Shahrokhi, F. ed. *Remote Sensing of Earth's Resources Vol 1.* The University of Tenessee, Tullahoma Tenessee. p 481-507. 1972.

j McDonnell, M.J.

Personal Communication.

k Henderson, P. and S. Tanimoto

Considerations for Efficient Picture Output via Lineprinter. *Computer Graphics and Image Processing,* Vol 4, p 327-335, Dec 1974.

l Rosenfeld, A. and J.S. Weszka

Picture Recognition. p 135-166 *In* Fu, K.S. ed. *Digital Pattern Recognition* Berlin, Springer Verlag, 1976.

m Kootnz, W.L.G. et al.

A Branch and Bound Clustering Algorithm. *IEEE Transactions on Computers,* Vol. C-24, p 908-914, 1975.

BIBLIOGRAPHY

Sources which have been read but are not explicitly referenced in the text

Bernstein, R

Digital Image Processing of Earth Observation Sensor Data. *IBM Journal of research and development*, Vol. 20, p 40-57, Jan 1976.

Davis, L.S.

A Survey of Edge Detection Techniques. *Computer Graphics and Image Processing*, Vol. 4, p 248-270, 1975.

Duda, P.E. and R.O. Hart

Pattern Classification and Scene Analysis. New York, J. Wiley and Sons, 1973.

Everitt, B.

Cluster Analysis. London, Heinemann, 1974.

Fukunaga, K.

Introduction to Statistical Pattern Recognition. New York, Academic Press, 1972.

Hord, R.M. and N. Gramenopoulos

Edge Detection and Regionalised Terrain Classification from Satellite Photography. *Computer Graphics and Image Processing*, Vol. 4, p 184-197, June 1975.

Lawrence, R.D. and J.H. Herzog

Geology and Forestry Classification from ERTS-1 Digital Data. *Photogrammetric Engineering*, Vol. 41, p 1241-1251, 1975.

Meisel, W.S.

Computer-Oriented Approaches to Pattern Recognition. New York, Academic Press, 1972.

ANNEX A

Routine: FILEMAKER

The original form of the data was one file containing all four wavebands. To convert this into an acceptable format for the B6700 FILEMAKER reads through this file once and produces four files (one per waveband) of 128 records,each of 144 8-bit characters - only 128 of which are used.

Input: The original data file - read from magnetic tape on to disk.

Output: Four optimised disk files.

Runnning Instructions:

```
RUN FILEMAKER;
FILE IN (TITLE=RAWDATA);
FILE ONE(TITLE=BAND/4);
FILE TWO(TITLE=BAND/5);
FILE THREE(TITLE=BAND/6);
FILE FOUR (TITLE=BAND/7);
```

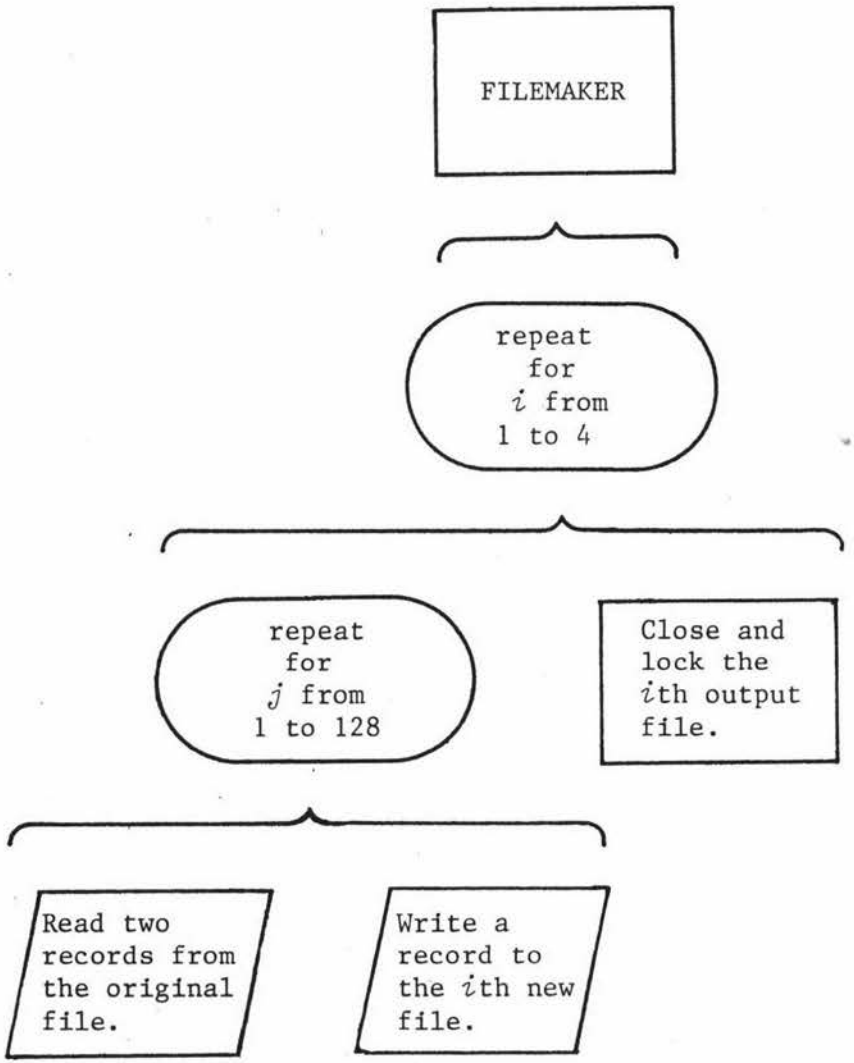


Fig A.1 Structure diagram of FILEMAKER.

ANNEX B

Routine: FLIPPER

FLIPPER inverts a square image about its leading diagonal, i.e. transposes rows and columns.

Input: Any one of the (four) data files stored on disk.

Output: A disk file containing the inverted version of the input file.

Running Instructions:

```
RUN FLIPPER;
FILE IN (TITLE=BAND/7);
FILE OUT(TITLE=FLIPPED/BAND/7);
```

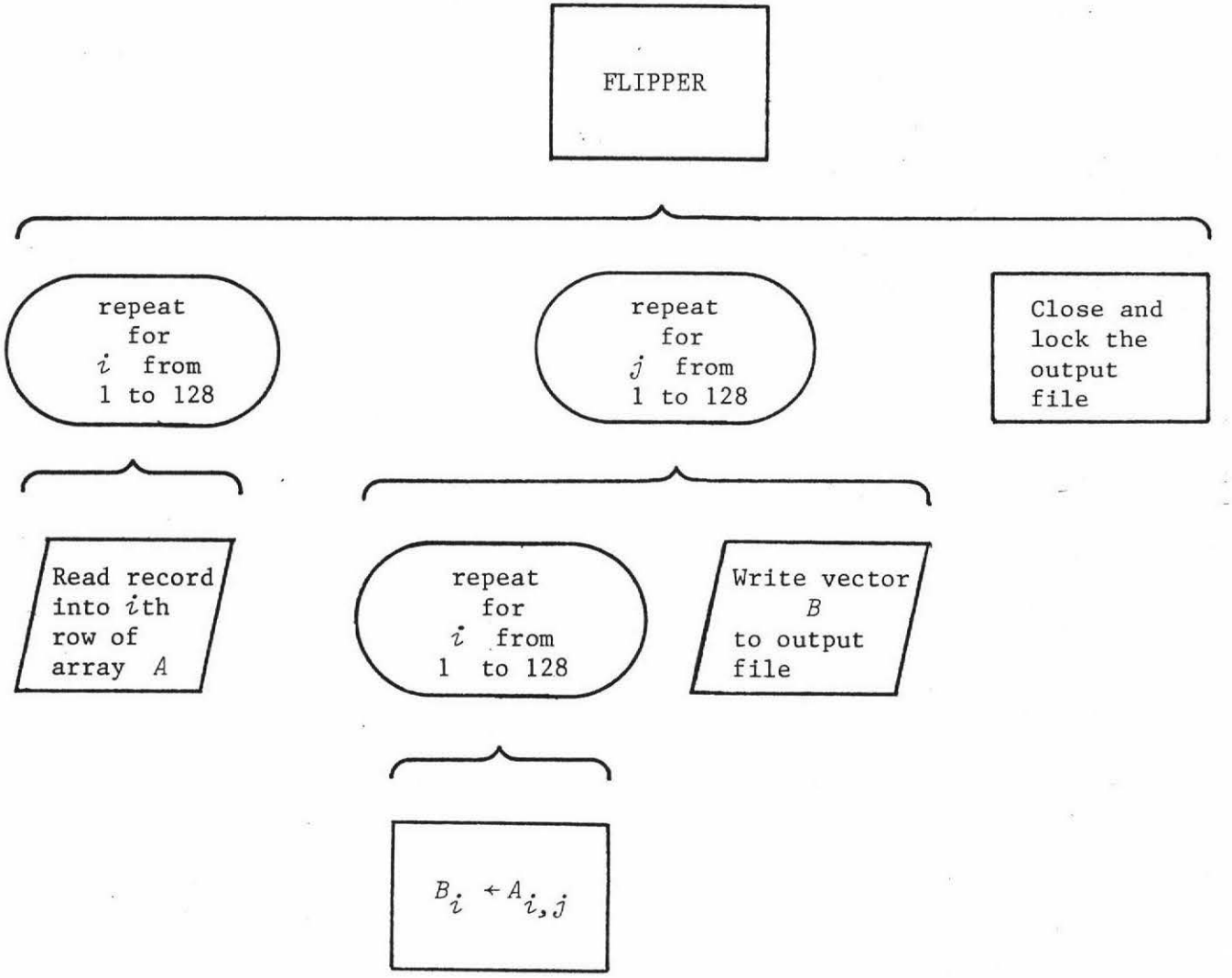


Fig B.1. Structure diagram of FLIPPER.

ANNEX C

Routine: ALLLEVELS

This routine prints out a pictorial representation of a file using a different symbol for each different intensity level recorded.

Input: One of the data files stored on disk.

Output: A printed image, 128 ×128 pixels, of the file. A key of the symbol representing each intensity level is also printed.

Running Instructions:

```
RUN ALLLEVELS;
FILE IN(TITLE=BAND/7);
```

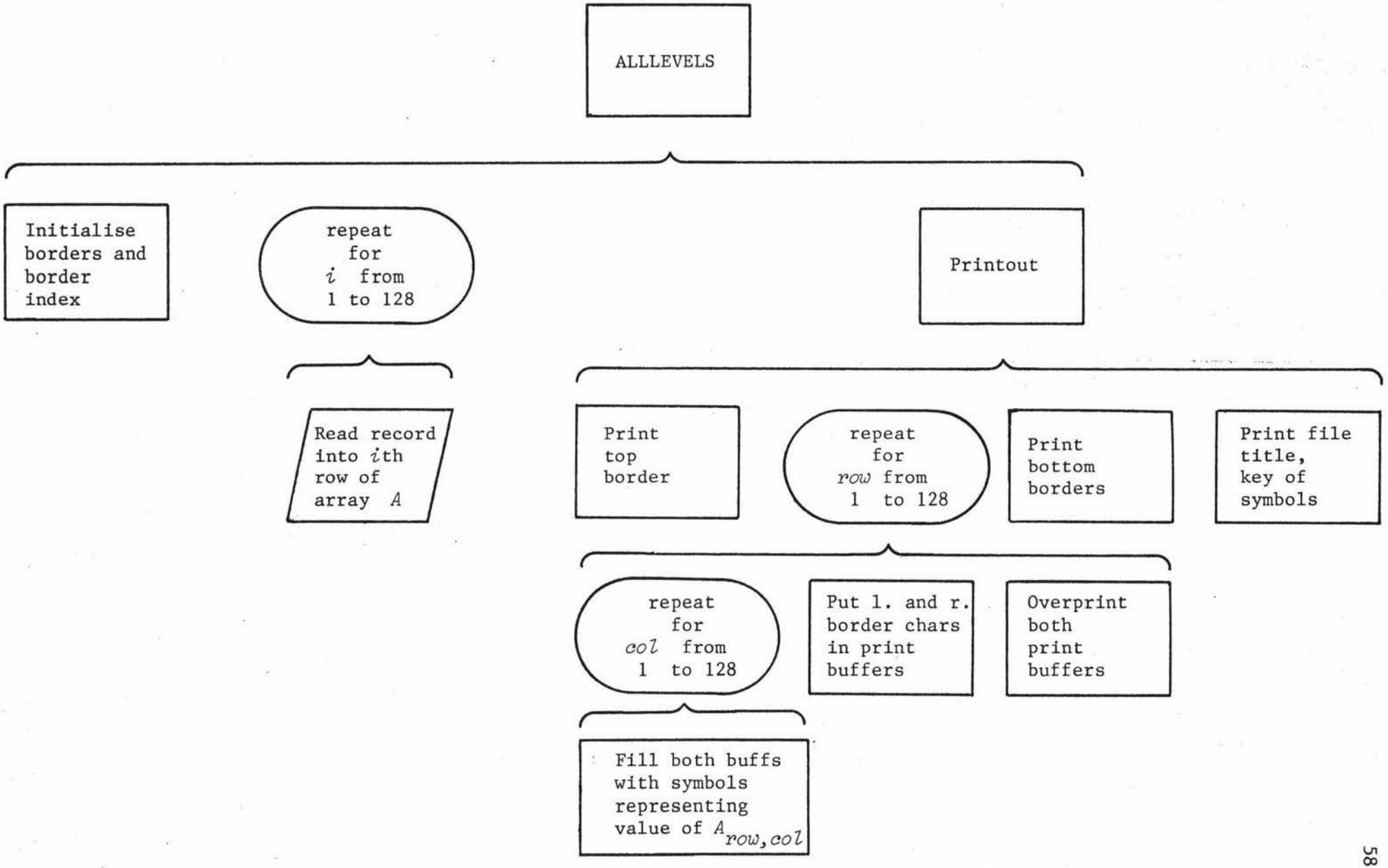


Fig C.1. Structure diagram of ALLLEVELS.

ANNEX D

Routine: SHADES

SHADES prints out a grey tone version of one of the stored images. In the standard version 10 grey levels are used and they are spread evenly throughout the range of intensity levels recorded. A pass is first made through the file to find the maximum and minimum values, and hence the range.

Input: One of the data files stored on disk.

One card containing the required height and width of the image (i.e. the number of pixels), in 2I3 Format.

Output: A printed grey tone version of the image.

Running Instructions:

```
RUN SHADES;
FILE IN(TITLE=BAND/7);
DATA
128128
```

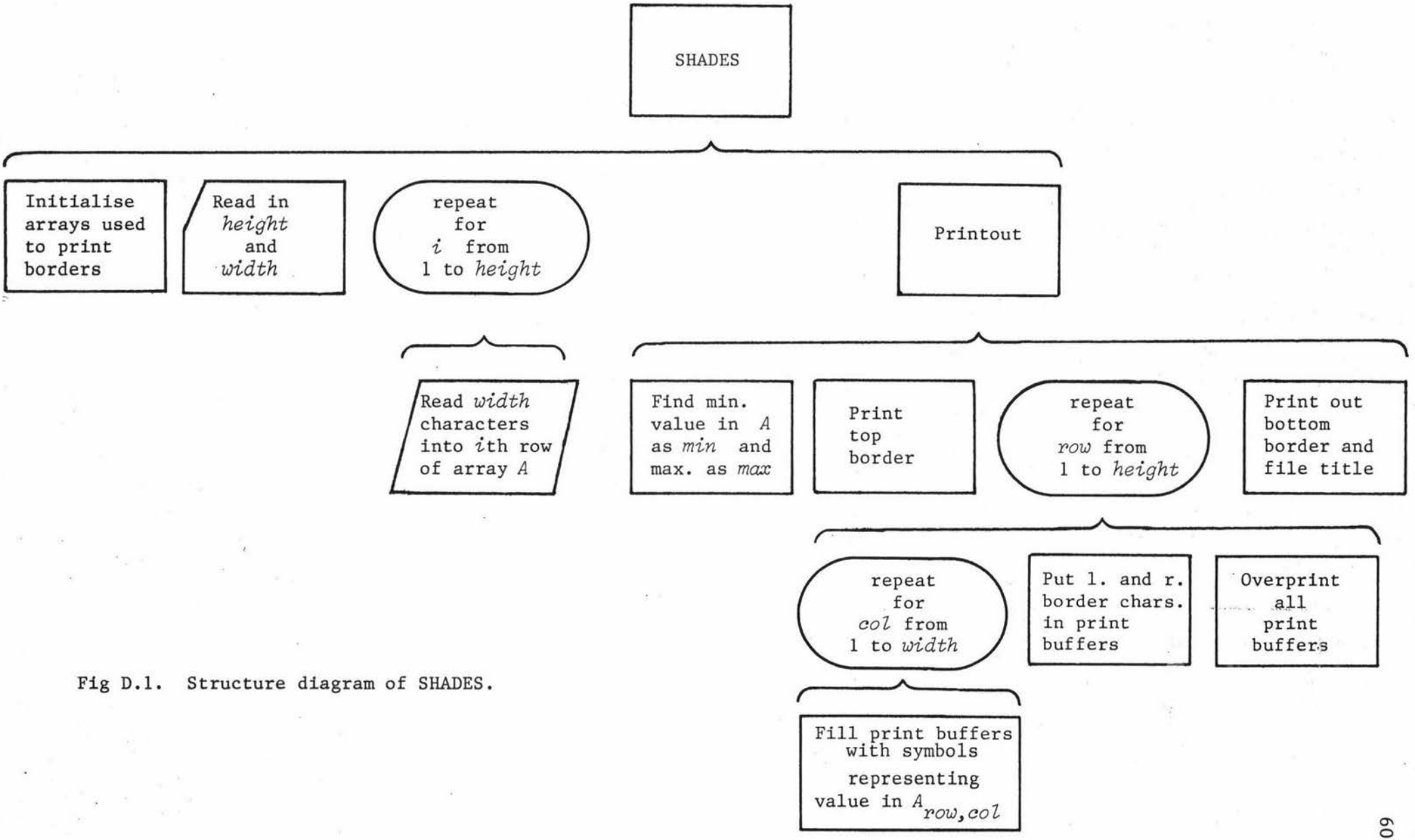


Fig D.1. Structure diagram of SHADES.

ANNEX E

Routine: HISTOGRAMMER

This routine prints out a histogram showing the frequency of occurrence of each intensity level in a particular image.

Input: The required data file (stored on disk).

Output: A printed histogram for that image.

Running Instructions:

```
RUN HISTOGRAMMER;
FILE IN(TITLE=BAND/7);
```

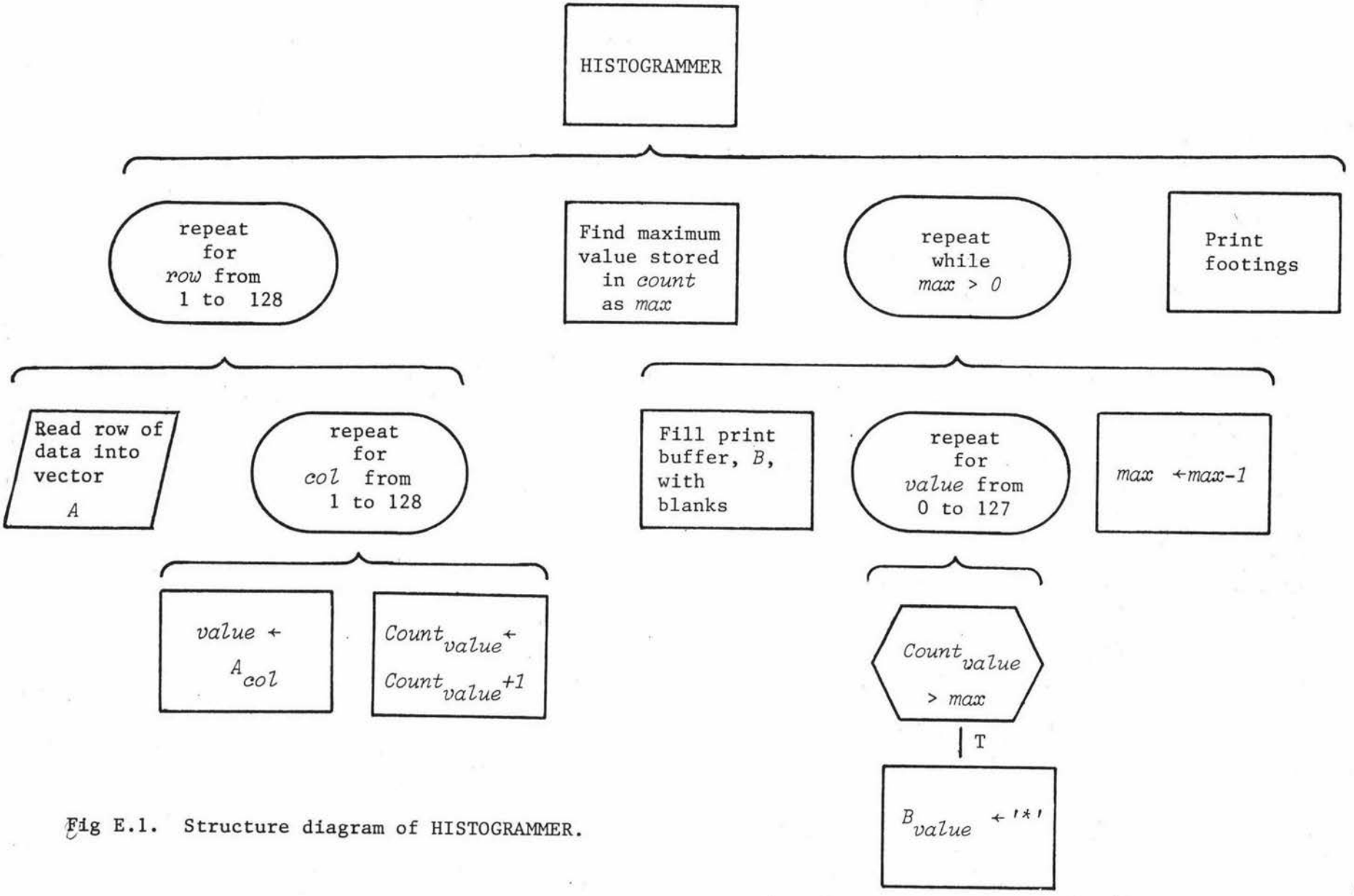


Fig E.1. Structure diagram of HISTOGRAMMER.

ANNEX F

Routine: DISTRIBUTION/ANALYSER

This routine displays the correlation of data values between any two wavebands. Each data point is plotted on a two-dimensional plane by using its two measured values (the recorded intensity levels from the two required wavebands) as coordinates. Data points with identical values correspond, increasing the value for that point in the new image. Either SHADES or ALLLEVELS may be used to display the image.

Input: The two required data files - disk files.

Output: A disk file of the distribution as described above to be printed by SHADES or ALLLEVELS.

Running Instructions:

```
RUN DISTRIBUTION/ANALYSER;
FILE IN1(TITLE=BAND/4);
FILE IN2(TITLE=BAND/6);
FILE OUT(TITLE=DISTRIBUTION/4/6);
```

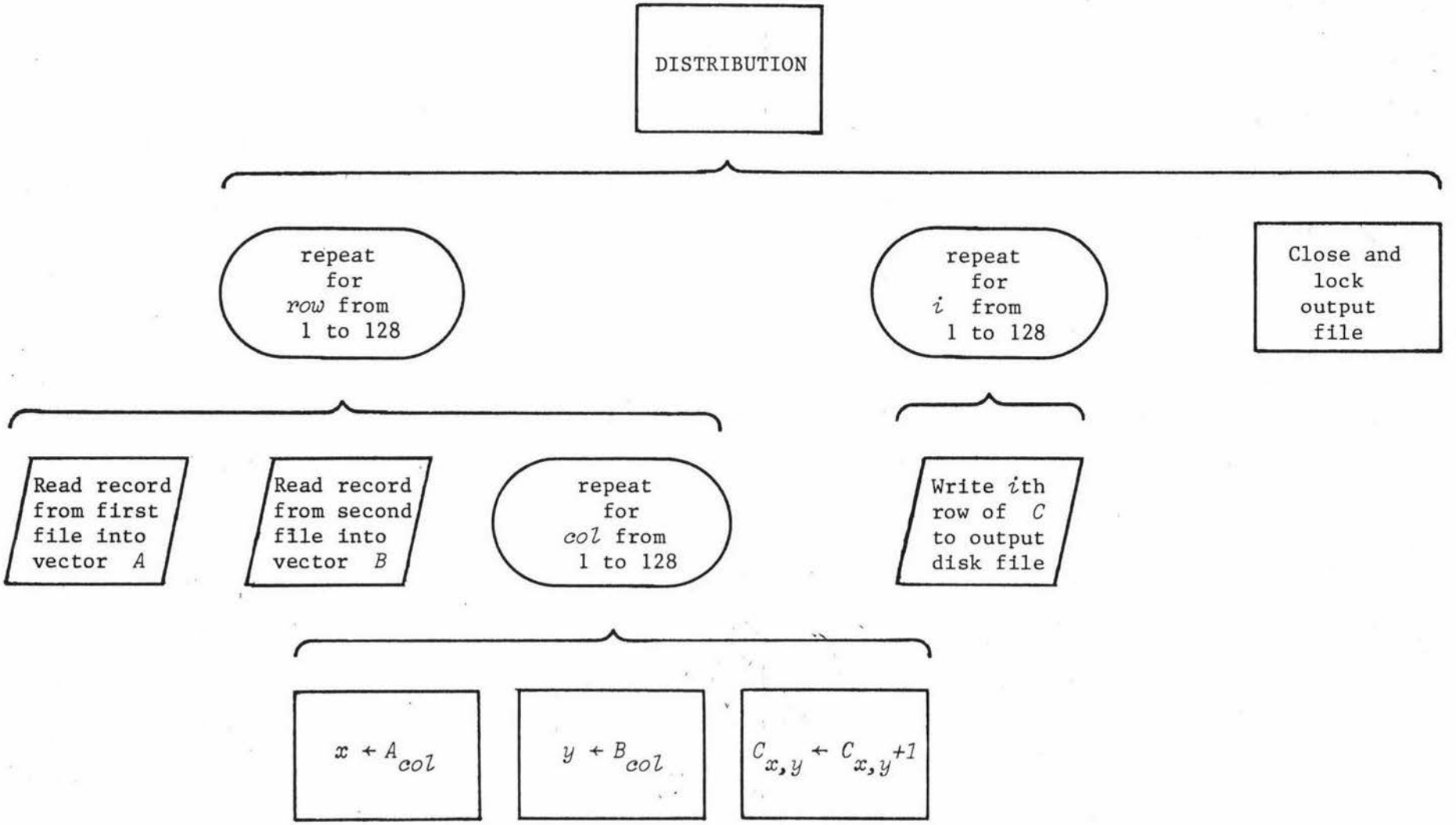


Fig F.1. Structure diagram of DISTRIBUTION/ANALYSER.

ANNEX G

Routine: DIFFERENTIATOR

An image is scanned first across rows and then down columns. The intensity level recorded for each point is compared with that of the adjacent point along the row, and also with that of the adjacent point down the column. The absolute values of these two differences are added and this is stored as the value of that point in the new image. This is a simple differencing or differentiating procedure. The new image may then be printed using SHADES or ALLLEVELS.

Input: Both (standard and inverted) versions of one of the data files.
One card containing the required height and width of the image, in 2I3 Format.

Output: A diskfile of the differentiated image, to be printed by SHADES or ALLLEVELS.

Running Instructions:

```
RUN DIFFERENTIATOR;
FILE IN1(TITLE=BAND/7);
FILE IN2(TITLE=FLIPPED/BAND/7);
FILE OUT(TITLE=DIFFERENCE/BAND/7);
DATA
128128
```

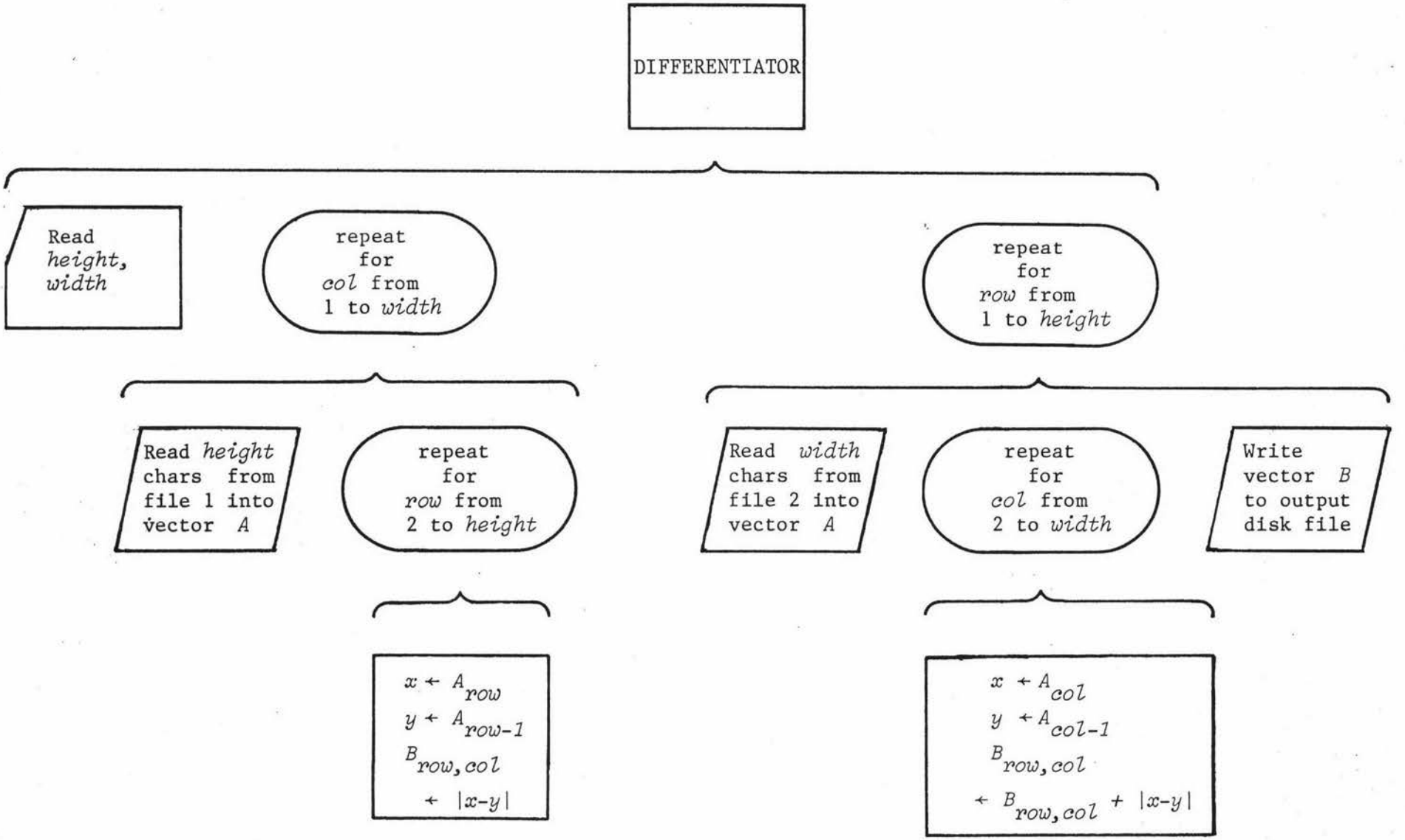


Fig G.1. Structure diagram of DIFFERENTIATOR.

ANNEX H

Routine: BOUNDARYFINDER

An image is scanned across rows and then down columns to look for significant differences between adjacent pairs of points. Where such differences are found a new image, initially all blank, is marked with a '.' . These should correspond to the edges or boundaries in the image. The level of significance, T, is input as a parameter. A printer file and an output disk file may be selected by control parameters.

Input: Both (standard and inverted) versions of one of the data files.

Three control cards:

1. Containing the required height and width of the output image, and the percentage significance threshold; in 3I3 Format.
2. Containing the minimum and maximum intensity levels recorded in the file, in 2I3 Format.
3. Determining presence of output printer file and disk file, in 2L5 Format.

Output: Where selected: a printer file of the new image, showing '.'s for boundaries, blank elsewhere.

Where selected: a disk file corresponding to the printer file.

Running Instructions:

```
RUN BOUNDARYFINDER;                    cont:  DATA
FILE IN1(TITLE=BAND/7);                       128128010
FILE IN2(TITLE=FLIPPED/BAND/7);               003057
FILE OUT(TITLE=BOUNDARY/BAND7/10);            TRUE FALSE
```

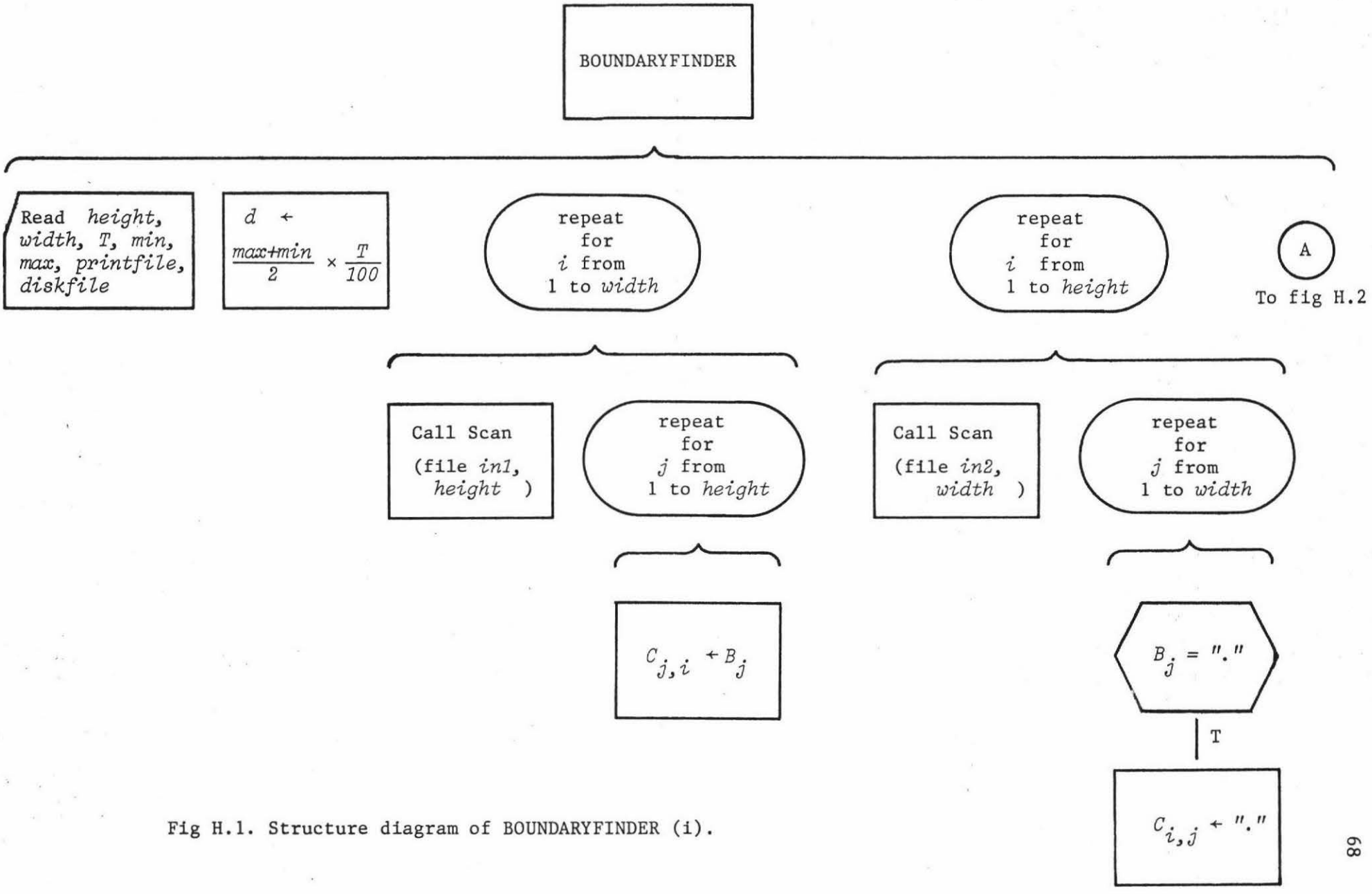


Fig H.1. Structure diagram of BOUNDARYFINDER (i).

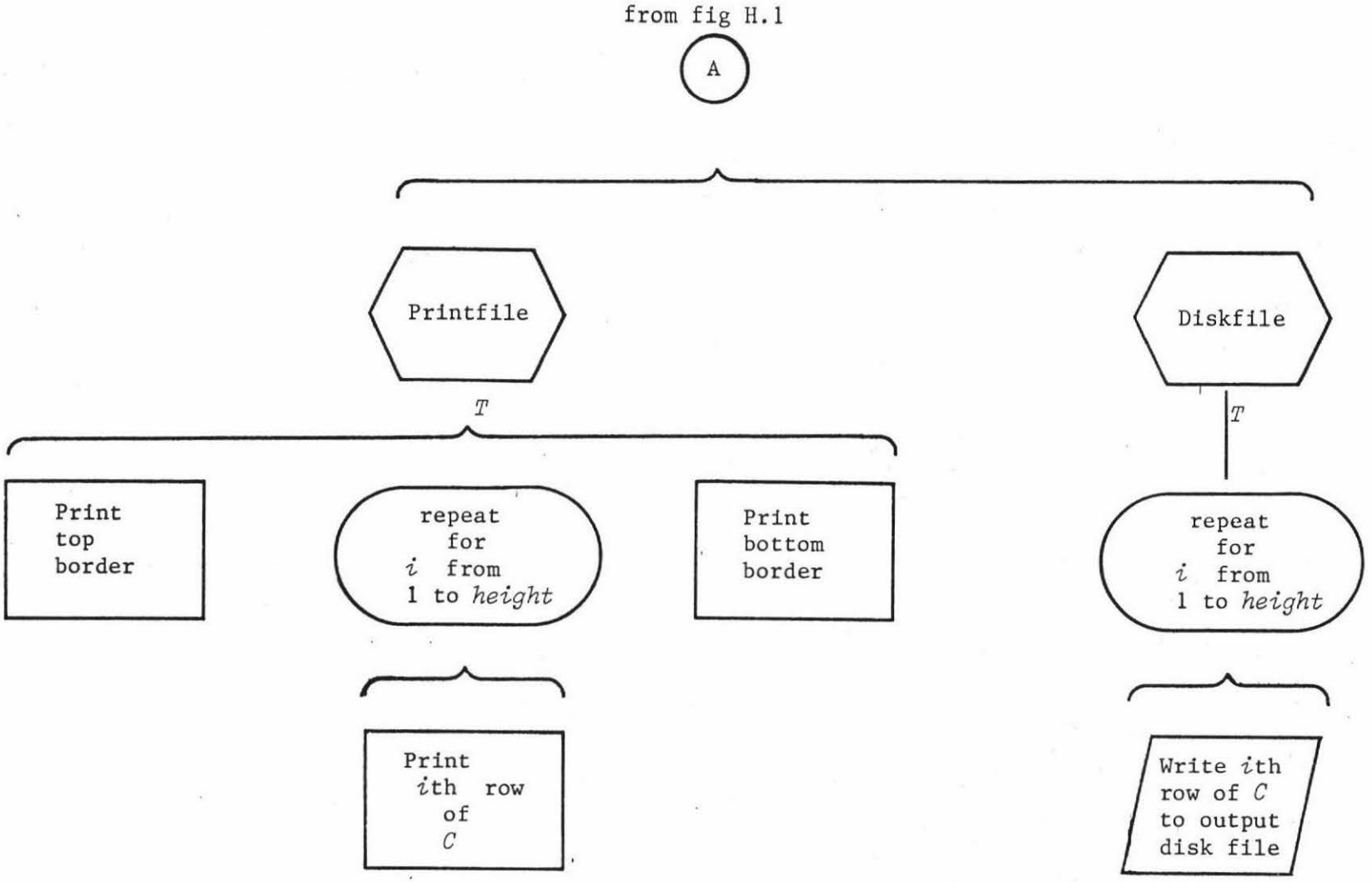


Fig H.2. Structure diagram of BOUNDARYFINDER (ii).

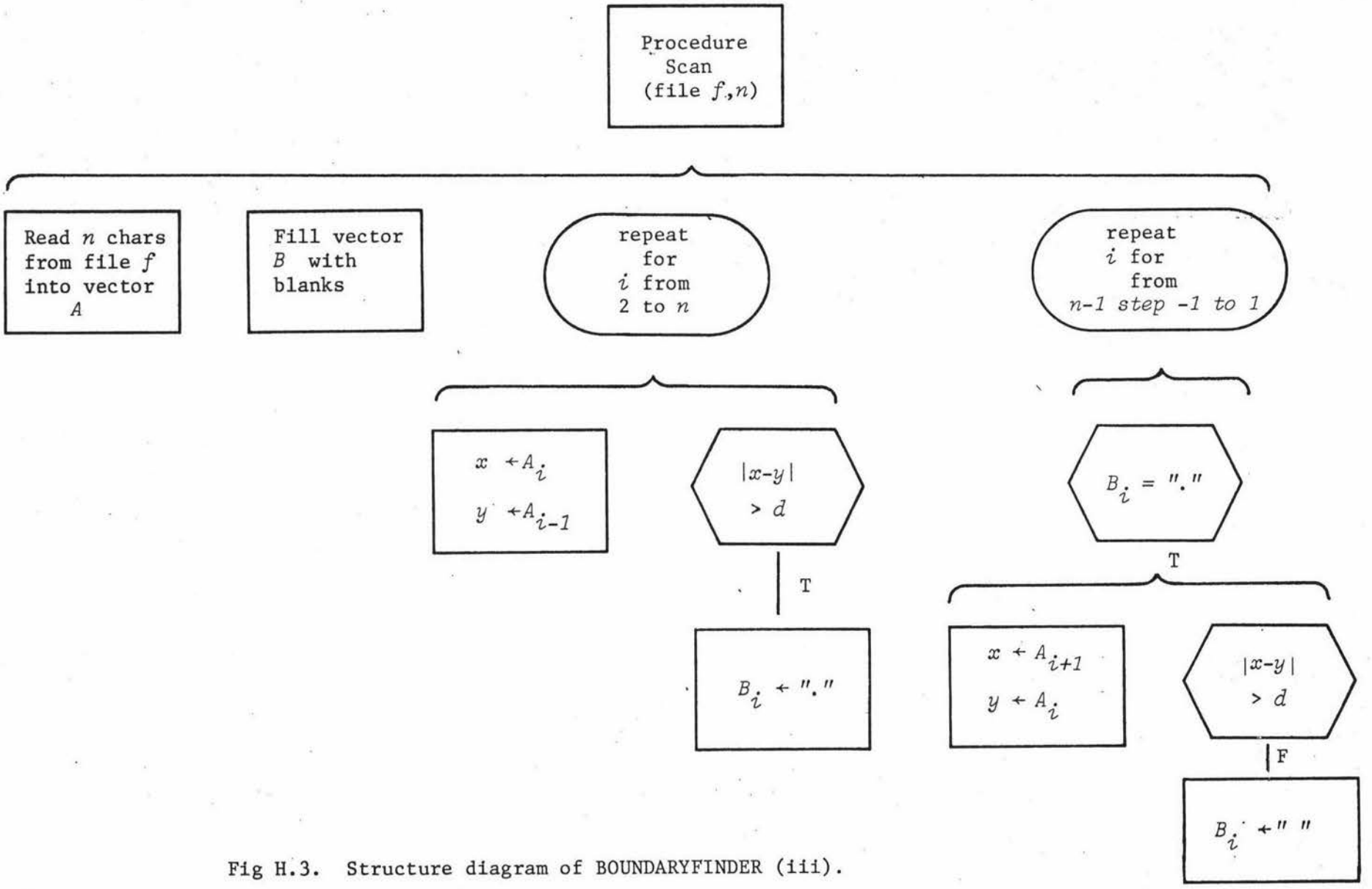


Fig H.3. Structure diagram of BOUNDARYFINDER (iii).

ANNEX I

Routine: BOUNDARYMERGER

This routine combines the boundary files on disk so that information from all four wavebands is used to produce a single boundary image. There are four different combination functions available, the output image may include:

a. Boundary points common to all four input files.

b. Boundary points common to any three input files.

c. Boundary points common to any two input files.

d. Boundary points common to any two input files plus those points which have a certain minimum number of immediate neighbours which satisfy one of a. - c. above.

The resulting image may be printed or a disk file version of it may be produced.

Input: All four disk files of boundary points - one per band.

Seven control and parameter cards:

1. Containing the required height and width of the output images, in pixels. 2I3 Format.
2 - 5. Containing headings for the four different images which may be produced.
6. Determining which of the four combination functions (a. - d. above) will be used. In 4L5 Format.
7. Determining whether a disk file is produced - the absence of this card means that no file will be produced.

Output: Where selected: printer files for the selected boundary combinations, showing '.' for boundary points and blank elsewhere.

Where selected: a disk file of one of the selected boundary combinations.

Running Instructions:

```
RUN BOUNDARYMERGER;
FILE IN1(TITLE=BOUNDARY/BAND4/10);
FILE IN2(TITLE=BOUNDARY/BAND5/10);
FILE IN3(TITLE=BOUNDARY/BAND6/10);
FILE IN4(TITLE=BOUNDARY/BAND7/10);
FILE OUT(TITLE=DARFIELD/BOUNDARY);
DATA
128128
ONE - BOUNDARY POINTS APPEARING IN ALL FOUR FILES
TWO - BOUNDARY POINTS APPEARING IN AT LEAST THREE FILES
THREE - BOUNDARY POINTS APPEARING IN AT LEAST TWO FILES
FOUR - BOUNDARY POINTS APPEARING IN AT LEAST TWO FILES PLUS EXTRAS
FALSETRUE TRUE FALSE
THIS LAST CARD CAN HAVE ANYTHING WRITTEN ON IT
```

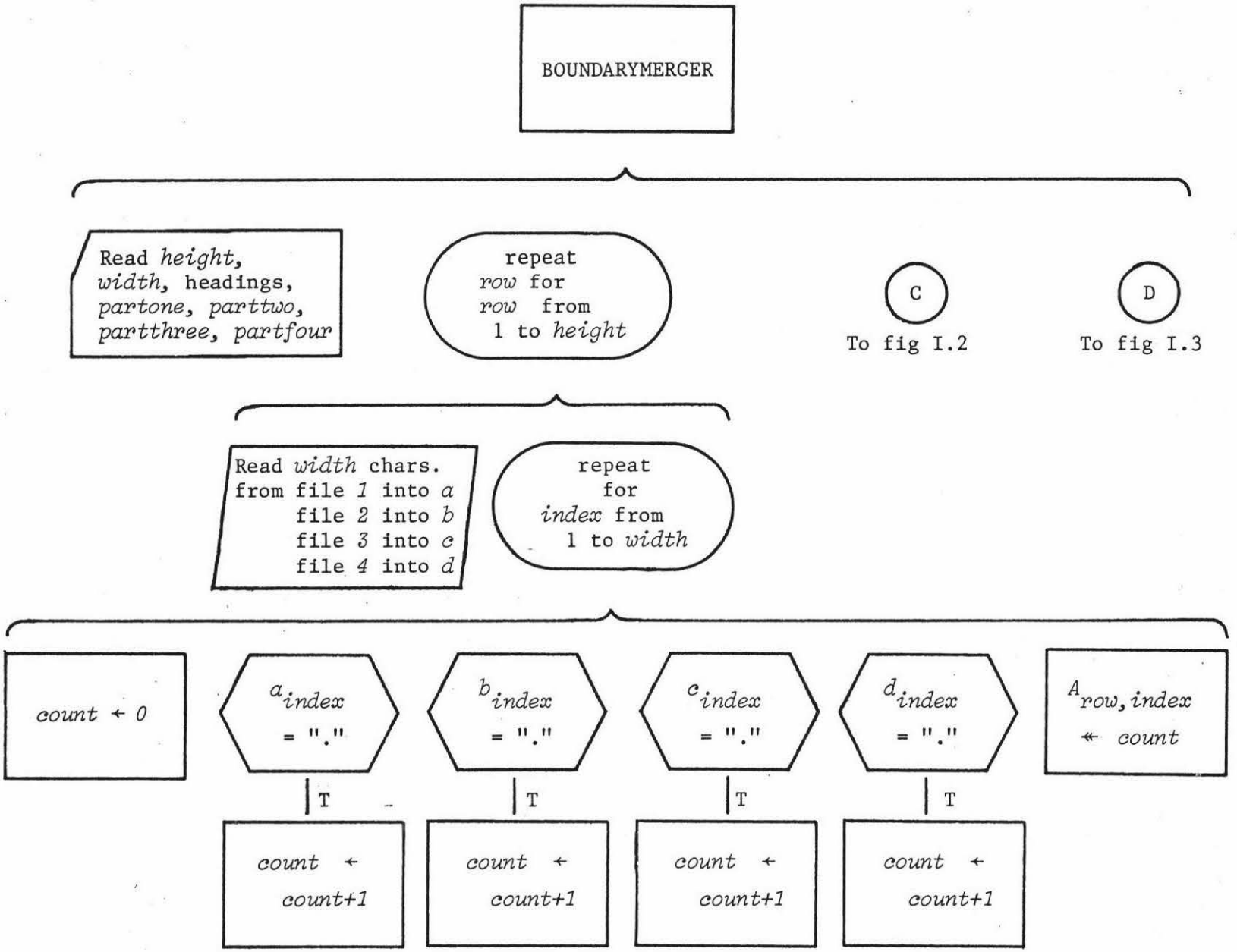


Fig I.1. Structure diagram of BOUNDARYMERGER (i).

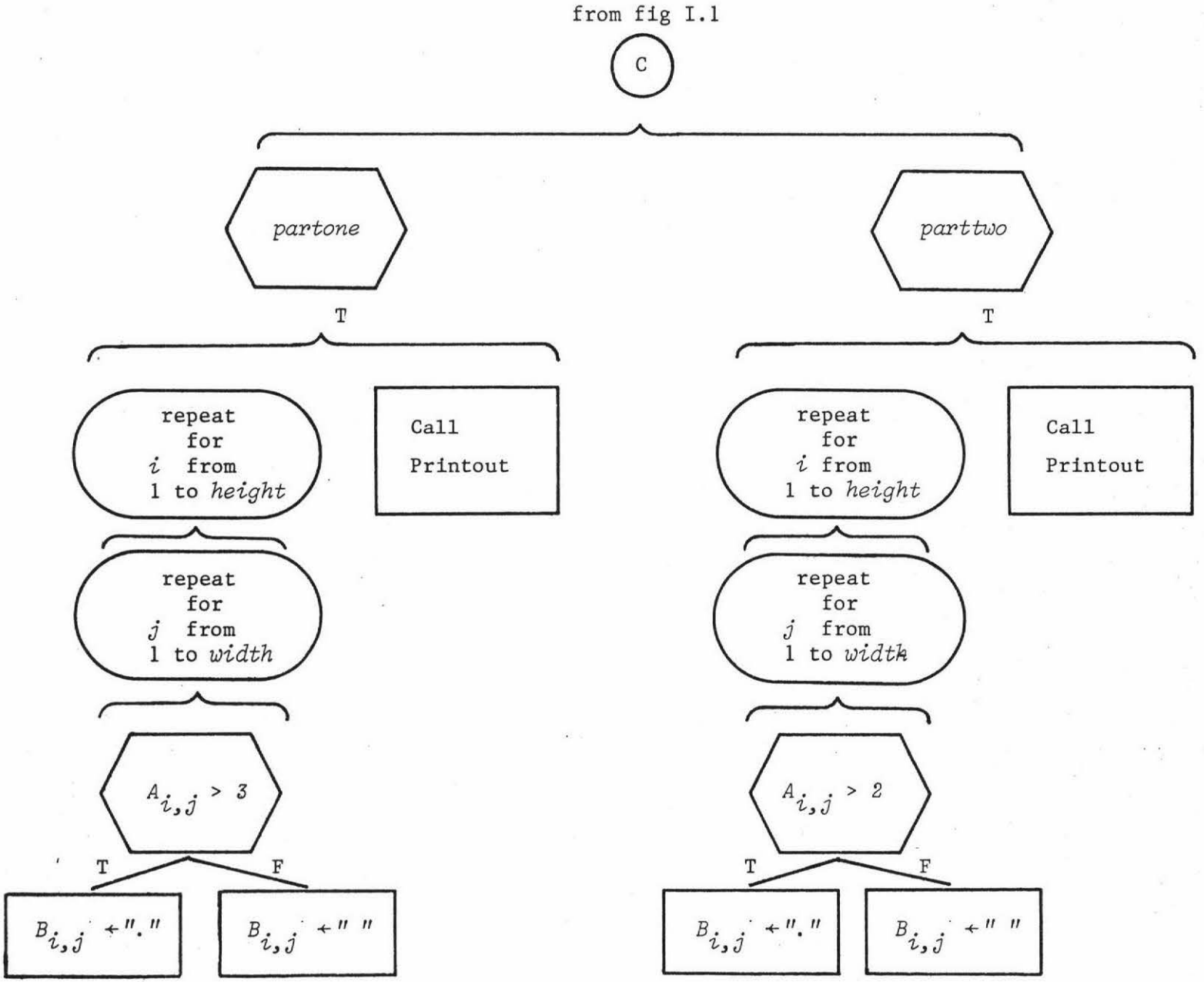


Fig I.2. Structure diagram of BOUNDARYMERGER (ii).

from fig I.1

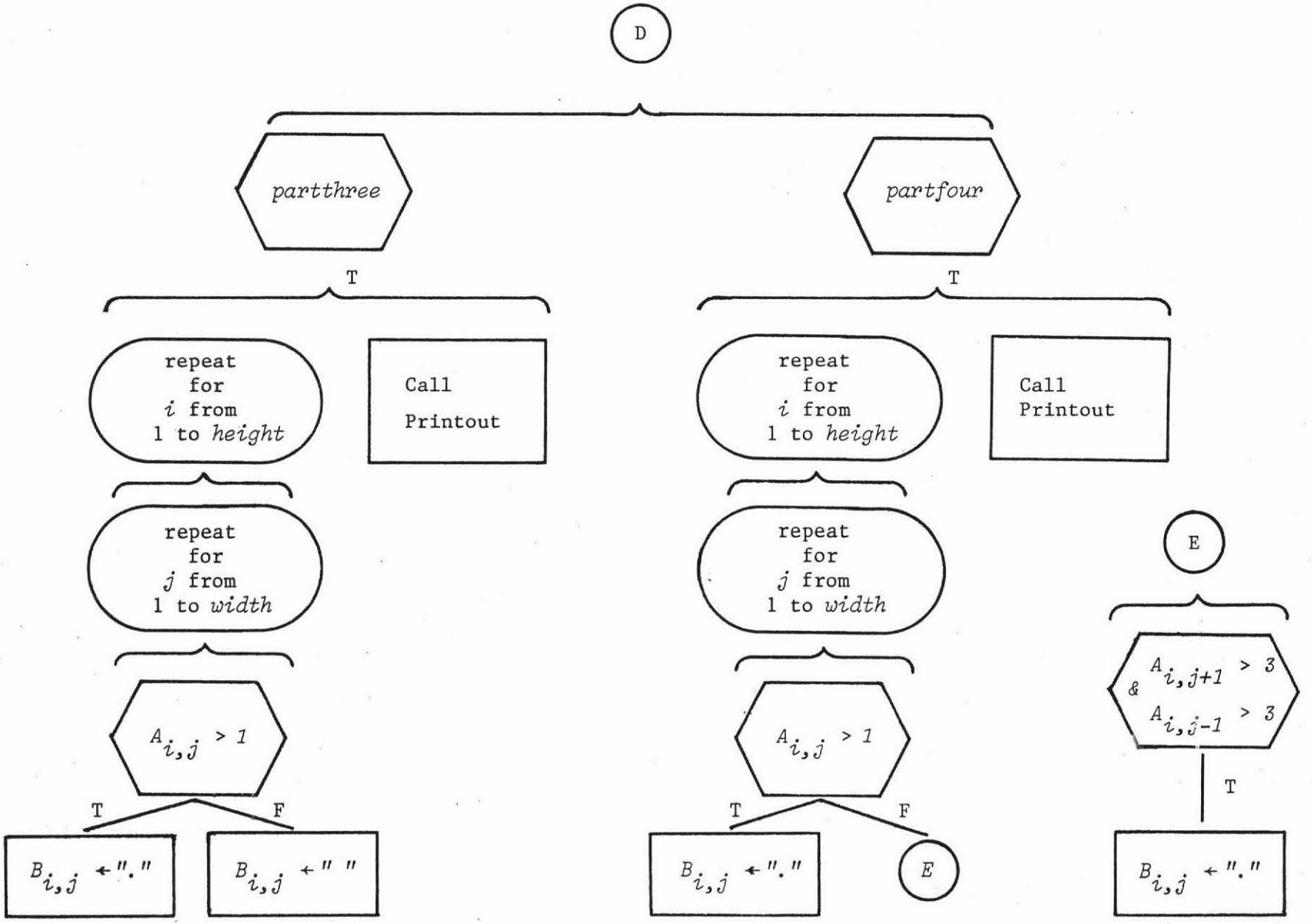


Fig I.3. Structure diagram of BOUNDARYMERGER (iii).

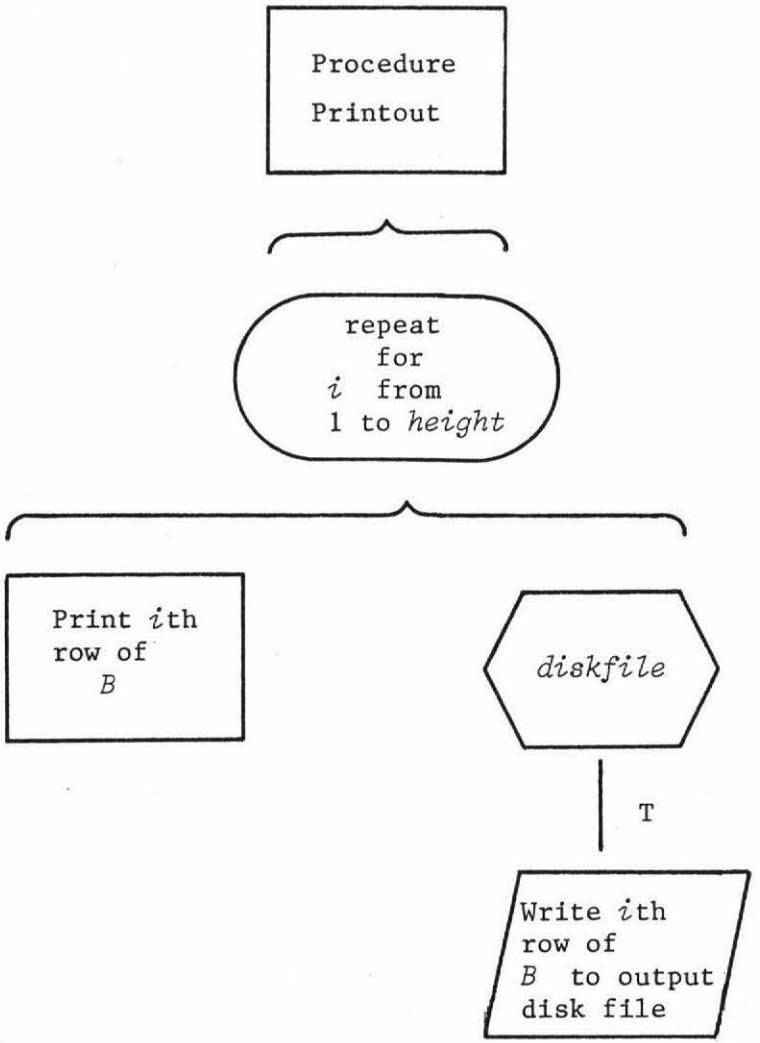


Fig I.4. Structure diagram of BOUNDARYMERGER (iv).

ANNEX J

Routine: SHAREDNN

This routine performs the Shared Near Neighbour clustering algorithm on LANDSAT MSS data. Two data points are classified as belonging to the same class if more than a threshold number, k_t, of their k nearest neighbours are common to both points. Boundary points are skipped over during all stages of the processung.

Input: All four disk files of MSS data - one per waveband.
The disk file of boundary points.
Two control cards:

1. Containing k, the number of nearest neighbours to be calculated for each point, and *size*, the number of pixels along an edge of the square image to be processed; in 2I3 Format.
2. Containing k_t, the threshold parameter, in I2 Foramt.

Output: A printed image of the clustering results with a digit representing the cluster (or class) number corresponding to each data point. Boundary points are left blank.

Running Instructions:

```
RUN SHAREDNN;                          cont:  DATA
FILE FIRST    (TITLE=BAND/4);                 020020
FILE SECOND   (TITLE=BAND/5);                 14
FILE THIRD    (TITLE=BAND/6);
FILE FOURTH   (TITLE=BAND/7);
FILE BOUNDARY(TITLE=DARFIELD/BOUNDARY);
```

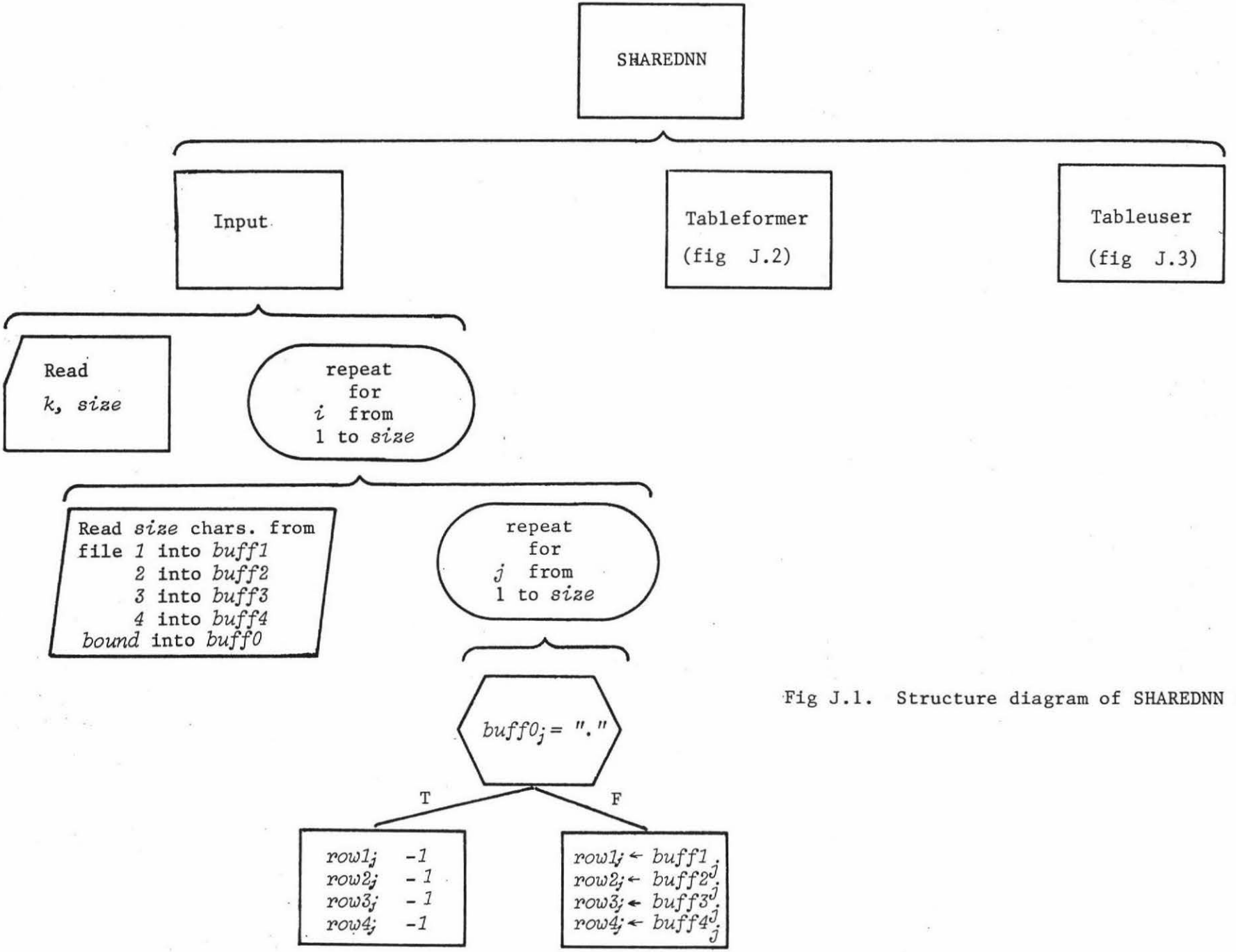


Fig J.1. Structure diagram of SHAREDNN (i).

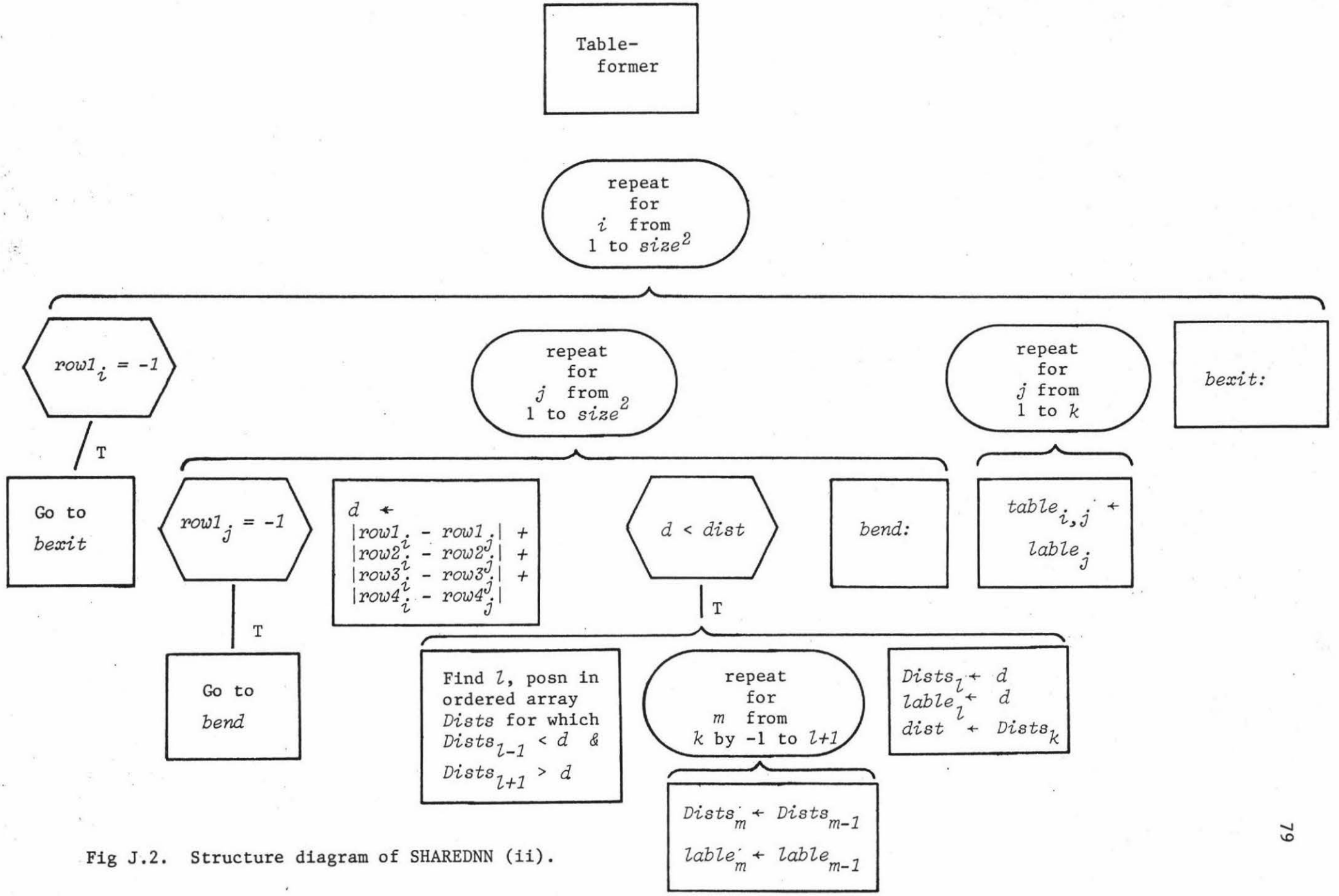


Fig J.2. Structure diagram of SHAREDNN (ii).

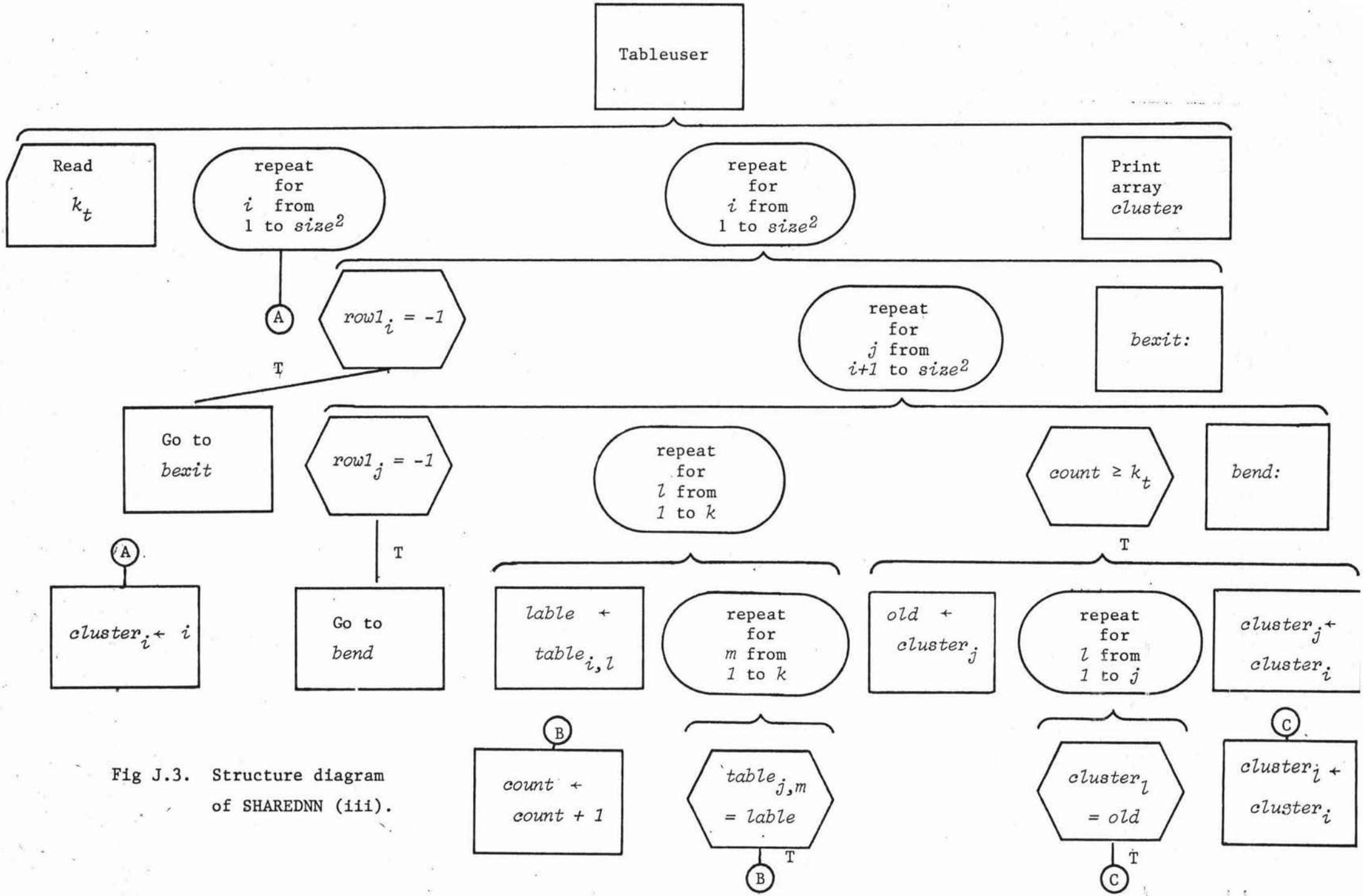


Fig J.3. Structure diagram of SHAREDNN (iii).

ANNEX K

Routine: MAXFINDER

This routine divides the pattern space into equivolume hypercubes, or cells. The population (number of contained data points) for each cell is calculated and compared with that of its immediate neighbours. Those cells with a greater population than their neighbours are considered as probable cluster centres and their coordinates are printed out.

Input: All four disk files of data - one per waveband.

The disk file of boundary points.

One control card containing the required cell edge length, *size*, in I1 Format.

Output: A list of the coordinates of (one vertex of) the cells found to contain a locally maximum number of points.

Running Instructions:

```
RUN MAXFINDER;
FILE IN0(TITLE=DARFIELD/BOUNDARY);
FILE IN1(TITLE=BAND/4);
FILE IN2(TITLE=BAND/5);
FILE IN3(TITLE=BAND/6);
FILE IN4(TITLE=BAND/7);
DATA
3
```

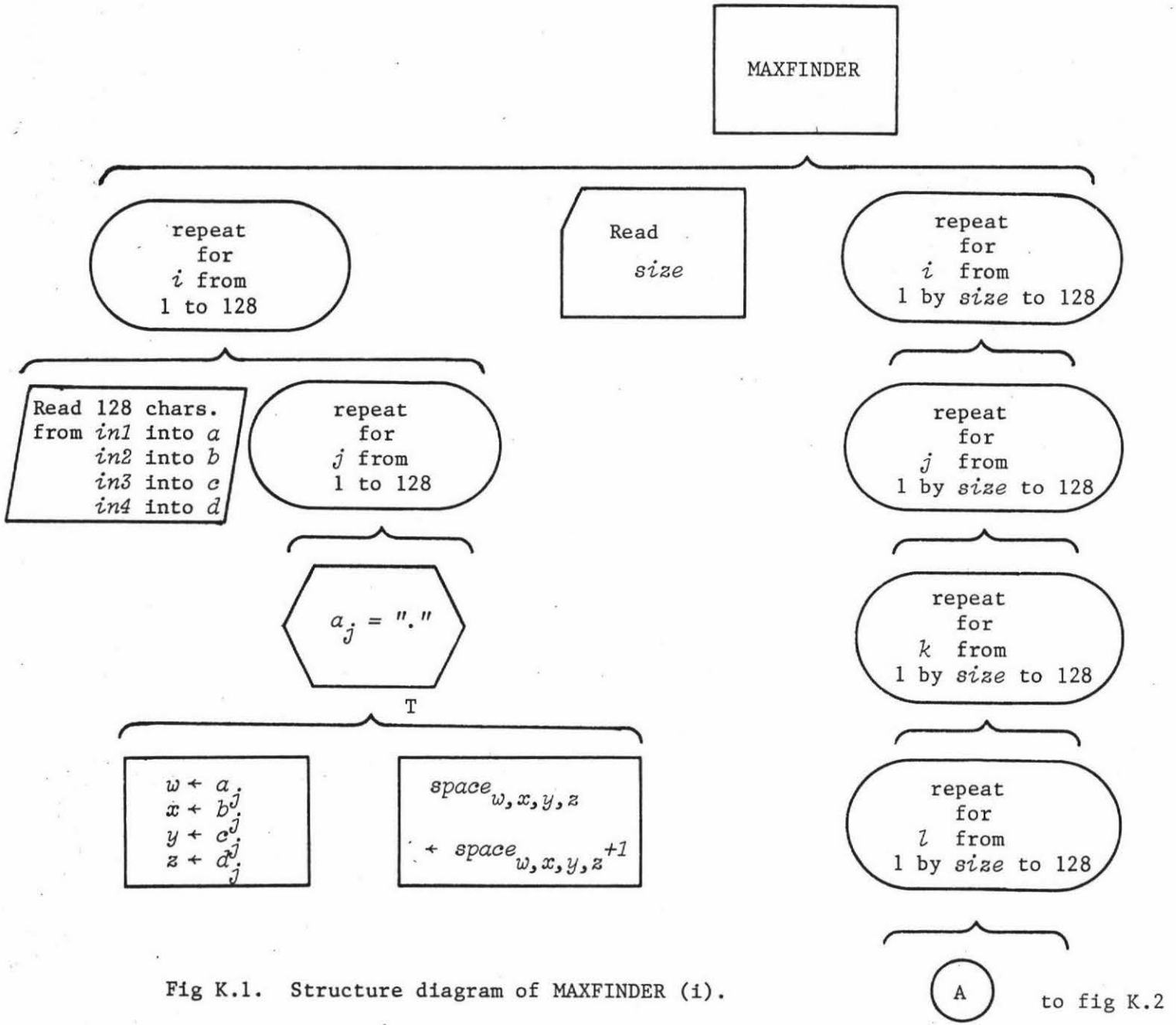


Fig K.1. Structure diagram of MAXFINDER (i).

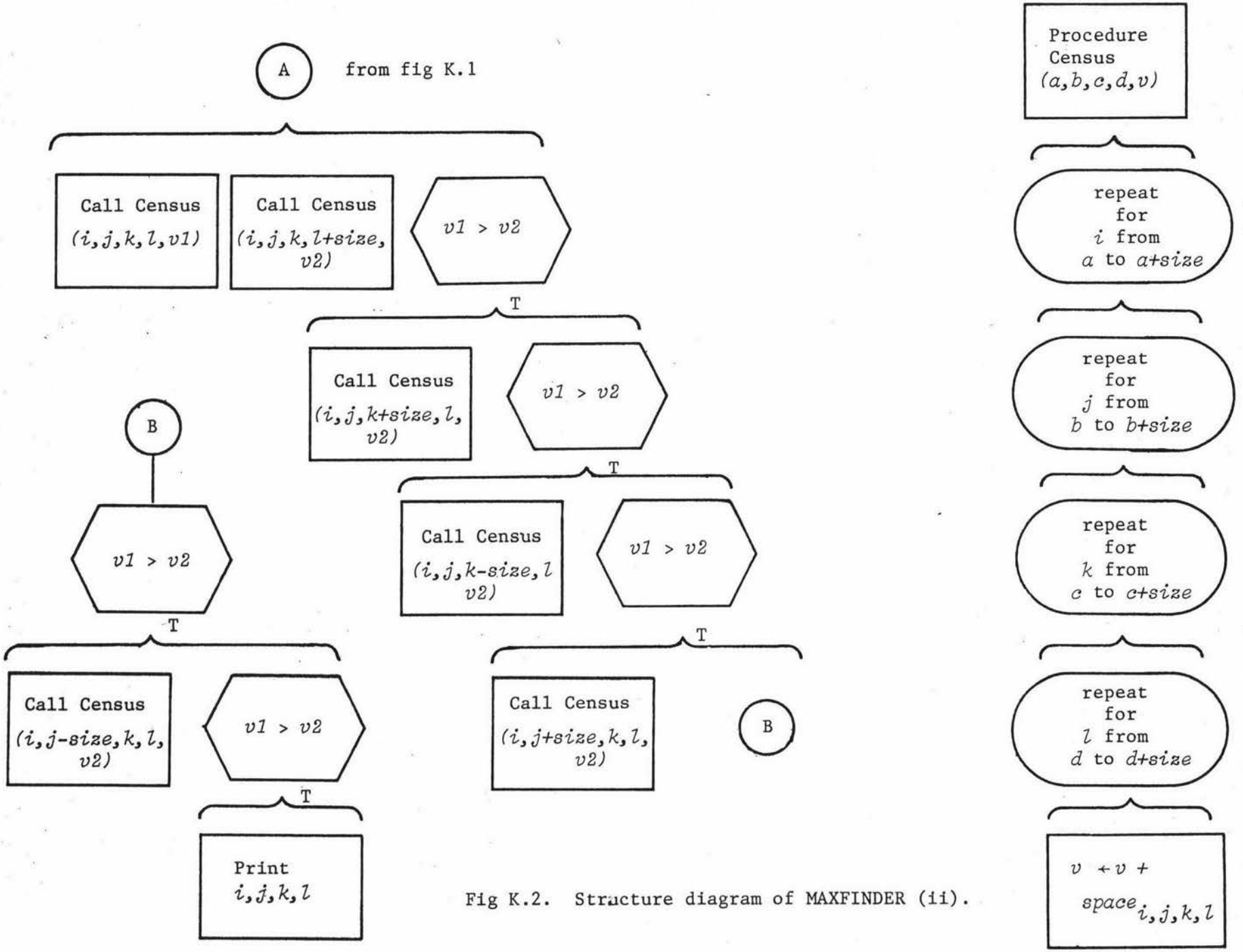


Fig K.2. Structure diagram of MAXFINDER (ii).

ANNEX L

Routine: CENTREFINDER

This routine performs the k-means clustering algorithm using LANDSAT MSS data. Initial cluster centres are progressively refined in an iterative fashion by classifying the data points according to the nearest cluster centre and then recalculating the centres by averaging the member points' coordinates. Boundary points are ignored during the first phase; then when this has converged they are classified with the nearest cluster centre, providing they lie within the value of spread calculated for that cluster. The number of clusters to be considered, k, is input as a parameter.

Input: All four disk files of MSS data - one per waveband.

The disk file of boundary points.

Control cards:

1. Containing the height and width of the image to be processed; and k, in 3I3 Format.

k cards each containing the coordinates of one initial cluster centre, in 4I3 Format.

Output: The cluster centres and populations for each iteration, then a printed image of the clustering results both before and after classifying the boundary points.

Running Instructions:

```
RUN CENTREFINDER;
FILE IN0(TITLE=DARFIELD/BOUNDARY);
FILE IN1(TITLE=BAND/4);
```

Running Instructions cont.

```
FILE IN2(TITLE=BAND/5);
FILE IN3(TITLE=BAND/6);
FILE IN4(TITLE=BAND/7);
DATA
 40 40  8
 18 16 69 38
 30 35 71 41
 33 41 59 49
 13 41 47  8
 26 29 59 49
 22 29 23 24
 17 17 59 31
 13 12 23 24
```

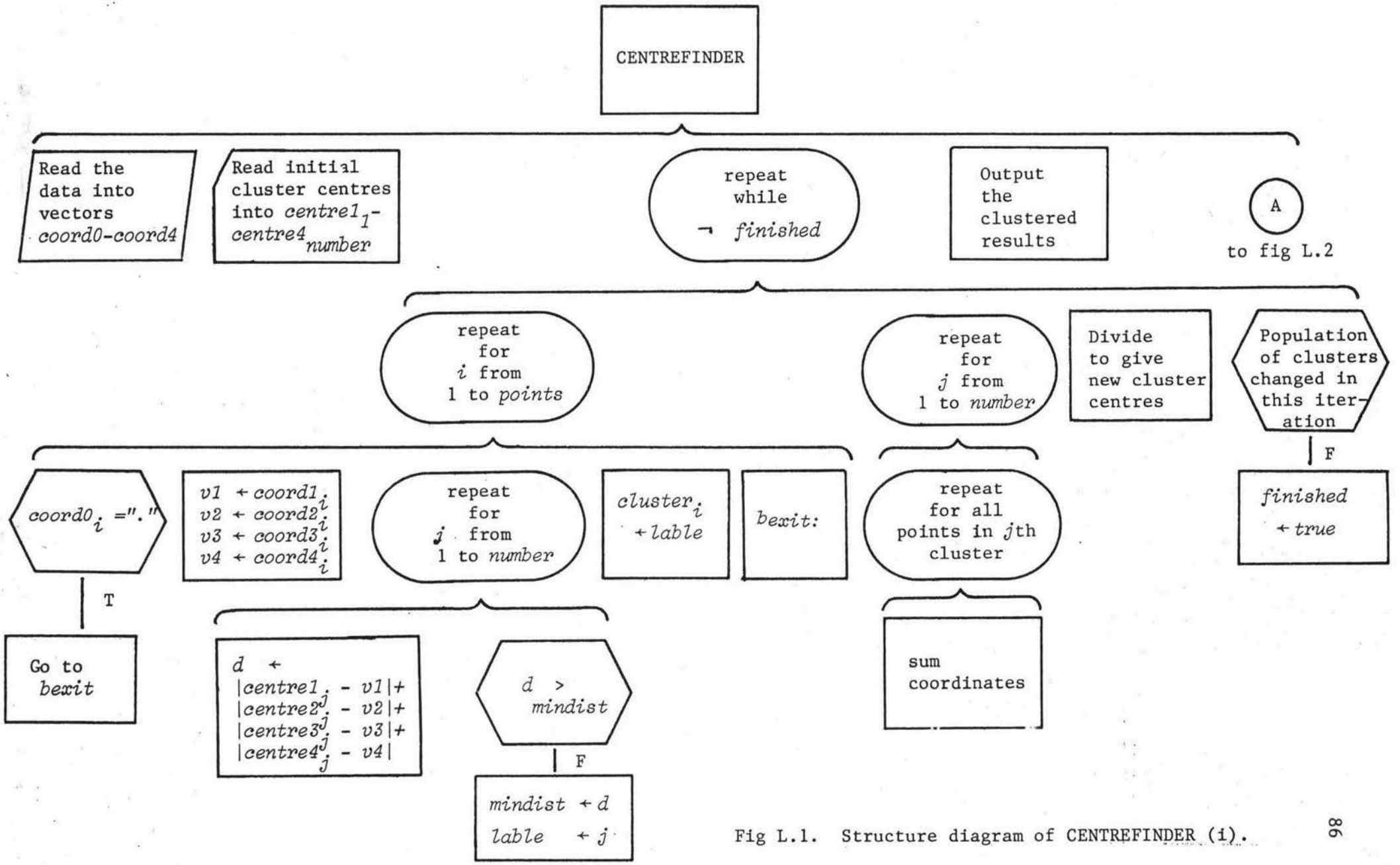


Fig L.1. Structure diagram of CENTREFINDER (i).

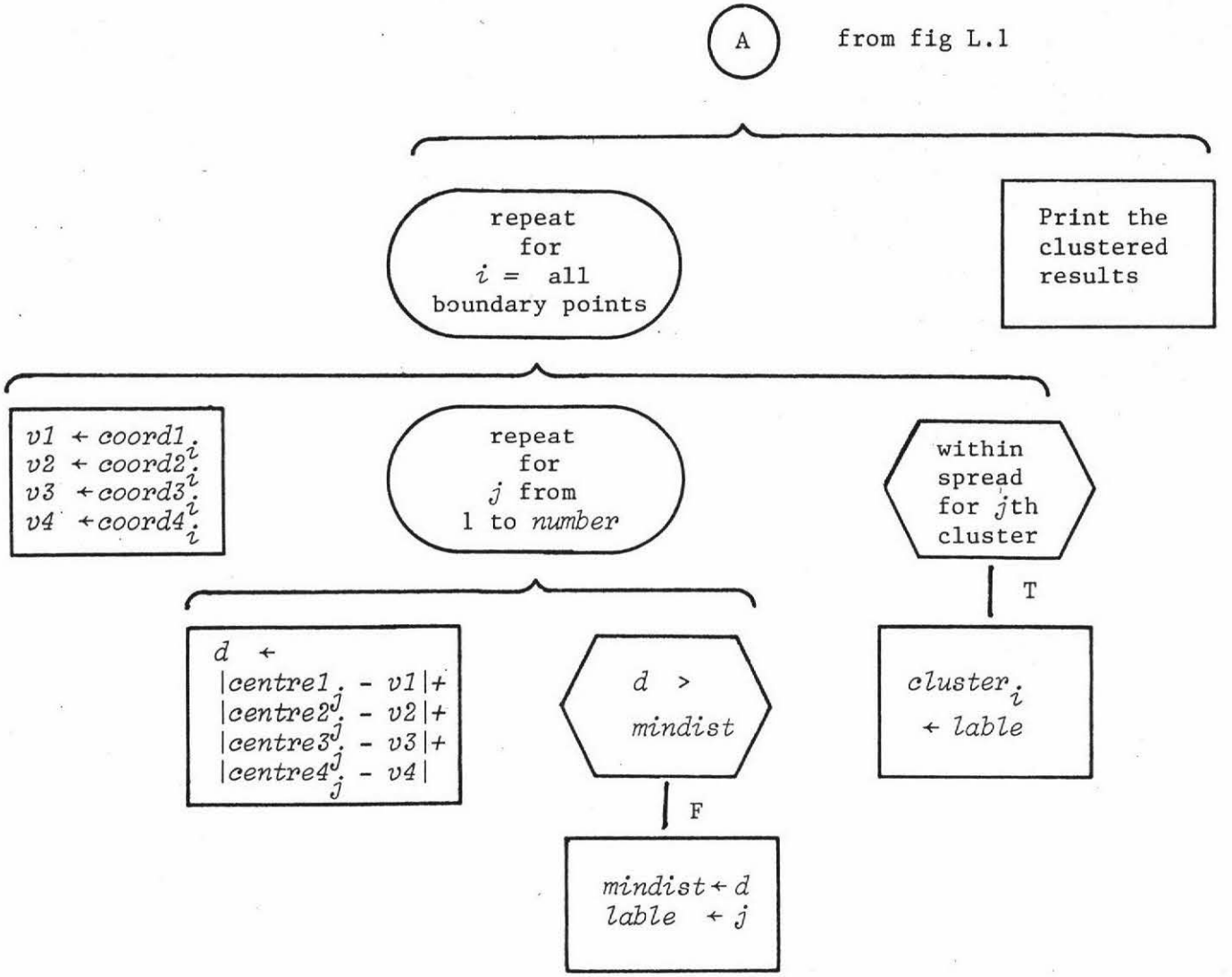


Fig L.2. Structure diagram of CENTREFINDER (ii).