Reprinted from

Ninth International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Natural Resources Evaluation

June 21-23, 1983

Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

FACTORIAL ANALYSIS OF CORRESPONDENCES APPLIED TO LANDSAT DATA

B.M. RACHED

Ecole Nationale D'Ingenieurs
Tunis, Tunisia

ABSTRACT

Semetimes, Classifications techniques have a subjective aspect and their results may differ from one to another. On the other side, the abundance of physical data relevant to the studied region and need not more identification may reduce to a minimum mis-classification and solve some of the confusion between different themes. The use of ancillary data needs, however, the use of some processing techniques allowing to process large quantities of various thematic data.

"Factorial Analysis of Correspondences" conceived by Professor BENZECRI (France) enables to regroup in the same analysis informations extracted from different sources. This method may be applied in two ways:
- an analysis of a composite image (mixture of MSS and ancillary data);
- an analysis from a result of a classification on the MSS image with statistical data.

According to the method used, this data is digitized in the first case and put in the form of a numerical array of correspondance in the second case.

This ancillary data may be extracted from a geographic Data Base. This data allows an automatical access to files by means of points of known coordinates.Around these points we build "windows" the size of which depends on the theme.Within or outside each window a classification is made. The obtained classes may be correlated to the statistical data.

The main results of these methods are the obtainment of "factorial" images where the pixels are classified in a hierarchical system and where the themes are well-separated in each "factorial" band. The number of factorial bands is lower than the number of multispectral bands used The first two factorial bands explain almost all the information except in the case of a multitemporel image where the third factorial band is also usefull. Thus, with successive thresholding, the factorial bands provide very interesting thematical classes.

I. INTRODUCTION

The distribution of Landsat MSS bands on the electromgnetic spectrum, the wealth of information provided by these bands, thematic confusions, interpretation difficulties are as many as problems that we have tried to solve by applying the "Factoriel analysis of correspondences" method (FAC) to the MSS and ancillary data.

Application of FAC can be done in two ways (figure 1) :

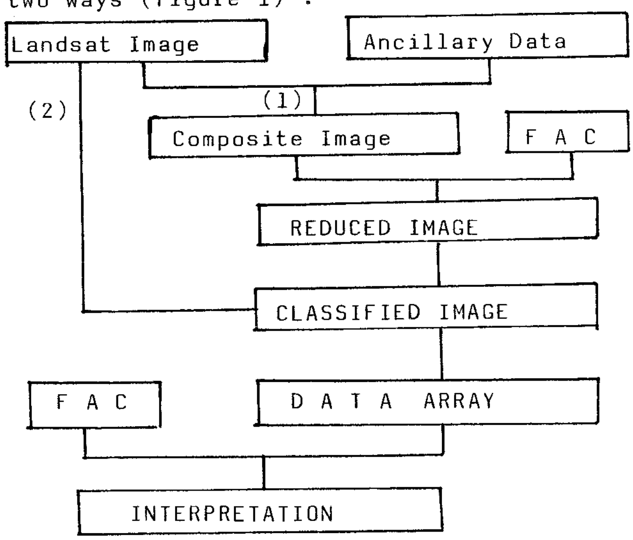

Figure 1. A brief description of the applications of FAC method to remote sensing data.

- On the Landsat image, or on an image having similar structures;
- On an array where remote sensing and ancillary data cross.

The results of these two applications have neither the same value nor the same finality. The first one enables to solve some ambiguities and confusions, and find correlations between the different types of data used, whereas the second one is rather an aid to intertpretation, since the FAC is not applied directly to the pixels.

After some definitions, we shall give, first an insight on the theoritical aspects of the FAC, and a methodology of its applications to an image and to a data array. Then, we shall define the type of statistical date to use, and a local classification technique.

II. SOME DEFINITIONS

In addition to the vocabulary in use in remote sensing, we introduce the following definitions (Boussema, 1981) :
a composite image is an image formed by a number of MSS bands, and other ones coming from the digitization of any graphical data and which can be registered to the original image ;

a factorial image is an image resulting from applying the FAC method to a multispectral, multidate, or composite image ;

a factorial band is any of the factorial image bands ;

a factorial value is each of the factorial coordinates of a pixel,

III. FACTORIAL ANALYSIS OF CORRESPONDENCES

A. Introduction

The FAC is a multidimensional statistical method which is used when one have to deal with a large number of correlated variables. Thus, it makes it possible to account for almost all of the information in a much lower-dimensions space. The new dimensions, which are not correlated, are linear functions of the original variables they are called "factors" and chosen so that the total variance is maximized.

This technique was conceived by professor J.P. BENZECRI about 1964-1965, and was applied with success in a great number of statistical studies (BENZECRI 1976).

The FAC technique differs from Principal components analysis mainly by the fact that the vectors which represent individuals are "profiles", by the use of the chi square ($\chi 2$) distance, and by the symmetry that exists between the individuals and variables.
In the FAC we are rather interested in individuals profiles through its characteristics than in the raw variables. This notion is also important in remote sensing data analysis since the spectral values profile within the whole MSS bands is more interesting than each of these taken separately, and quite often used in visual interpretation.

B. MATHEMATICAL PRINCIPALS

The FAC enables to process arrays of positive numbers with the help of a computer. The basic data is therefore a rectangular array noted K_{IJ} (BENZECRI, 1976) :

$$K_{ij} = \left\{ K(i,j) / i \in I, j \in J \right\} \quad (1)$$

Where :
I and J are the sets of individuals and variables, respectively.
I and J play symmetrical rôles.
We start at first by computing K_I, K_J and K, the row and column margins and their total, respectively (BENZECRI, 1976) :

$$K_i = \left\{ K(i) / i \in I \right\} \; ; \; \forall \, i \in I : K(i) = \sum_{j \in J} K(i,j) \quad (2)$$

$$K_j = \left\{ K(j) / j \in J \right\} \; , \; \forall \, j \in J \; : \; K(j) = \sum_{i \in I} K(i,j) \quad (3)$$

$$K = \sum_{i \in I} \sum_{j \in j} K(i,j) \quad (4)$$

These margins enable to get the marginal laws noted f_I, f_J and f_{IJ} (BENZECRI, 1976)

$$f_I = \left\{ fi / i \in I \right\} \; , \; fi = K(i) / K \quad (5)$$

$$f_J = \left\{ fj / j \in J \right\} \; ; \; f_j = K(j) / K \quad (6)$$

$$f_{IJ} = \left\{ fij / i \in I, j \in J \right\} \; , \; f_{ij} = K(i,j)/K \quad (7)$$

Then the marginal laws enable to compute the row and columnprofiles noted f_J^i and f_I^j respectively :

$$f_J^i = \left\{ f_j^i \; / j \in J \right\} \; ; \; f_j^i = fij/fi \quad (8)$$

$$f_I^j = \left\{ f_i^j \; / i \in I \right\} , \; f_i^j = fij / fj \quad (9)$$

Starting from a set of points, we end up with row and column sets of profiles noted N(I) and N (J), respectively (BENZECRI, 1976)

$$N(I) = \left\{ (f_J^i \, , \, f_i \,) / \; i \in I \right\} \quad (I0)$$

$$N(J) = \left\{ (f_I^j \, , \, fj) \; j \in J \right\} \quad (II)$$

Thus, the profile array is an array composed of numbers all of them greater or equal to zero f^I_J such that :

$$f^I_J = \left[f^i_j \,/ i \in I, j \in J \right] \text{ with } f^i_j = fij/fi \text{ and } \forall\, i \in I : \sum_{j \in J} f^i_j = 1 \qquad (12)$$

The N(I)and N(J) sets have principal stretch dimensions termed factorial axis and determined as principal axis of inertia. Thus, the problem consist in looking for these principal axis of inertia by choosing chi square (X2) métric.

The χ^2 distance is a particular Euclidian distance which enables to calculate the differences between row or column profiles, and which may be written, for example, for the set N(I) having f_J as a center :

$$\forall\, i_1, i_2 \in I, \; D^2(i_1, i_2) = \sum_{j \in J} (f^{i_1}_j - f^{i_2}_j)^2 \, (1/f\, j) \qquad (13)$$

The factorial axis are determined by the axial factorial vectors (or their density) of the set of points; to each of these correspnds an eigenvalue λ .

We define :

U_J = axial vector of the set N(I)relative to λ ;

Q^J = density of U_J relative to f_J such that $Q^j = Uj / fj$.

Then, Q^J is a factor if and only if Q^J has a mean equal to zero, a variance equal to 1, and verifies the following equation, (BENZECRI,1976) :

$$Q^J \text{ o } f^J_I = \lambda Q^J \qquad (14)$$

The factors are coordinates of the set of points on the factorial axis (or factorial bands for the image). Thus the factorial value for a given pixel i depends not on the total mass, but rather on the row profile which it describes.

We define :

F^I is a factor such that F(i) be the coordinate of the profile f^i_J on the axis determined by f_J and defined by U_J.

Then we have (BENZECRI,1976) :

$$Q^J \text{ o} f^I_J = F^I \; ; \; F^I \text{ o} f^J_I = \lambda Q^J \qquad (15)$$

The principal axis of inertia of the sets N (I) and N(J) relative to a non zero eigenvalue correspond bi-univocally ; (BENZECRI,1976) :

If (Q^I, F^I, λ) is a triple relative to N(I), $(\lambda^{-\frac{1}{2}} F^I, \lambda^{\frac{1}{2}} Q^J, \lambda)$ is a triple relative to N(J), and inversely. Thus, the same function Q^J with a variance of 1, defines an axis for N(I) and a factor ($G^J = \lambda^{\frac{1}{2}} Q^J$) for N(J).

The factors F^I and G^J of two sets of points are linked by the following equations,(BENZECRI,1976) :

$$\forall i \in I : F(i) = \lambda^{-\frac{1}{2}} \sum_{j \in J} G(j)\, f^i_j \qquad (16)$$

$$\forall j \in J : G(j) = \lambda^{-\frac{1}{2}} \sum_{i \in I} F(i)\, f^j_i \qquad (17)$$

The successive factorial axis are orthogonal two by two (in the sense of the χ^2 metric with center f_J).

The FAC enables to include one or several individuals (or one or several variables) in the factors calculation and to study the behavior of these individuals or variables, set as "supplementary elements" in the system of factorial axis (or bands) set up without them. The coordinates F (s) of a given profile f^s_J are calculated by the following equation (BENZECRI,1976) :

$$F(s) = \lambda^{-\frac{1}{2}} \sum_{j \in J} G(j)\, f^s_j \qquad (18)$$

IV. APPLICATION TO AN IMAGE

If n is the number of bands, the FAC may be applied to a multispectral (n=2 to 4), multidate or composite (n=2 to p,with p may be greater than 4) image.

These bands constitute the set of variables which are previously called J.

As input, we have as individuals, the set of pixels of the original image, and as output a q-dimension factorial image (q less than or equal to n-1).

For a composite image, the ancillary data is recorded as an image. In order to integrate this data with the multispectral one, it has to be digital, as subtle as possible, and its grid similar to the remote sensing data (Landsat MSS data for example). Most of this ancillary data is graphic and often comes from the digitization of maps and charts;we used, for example, the Digital Topographic Elevation Models (DEM) in BOUSSEMA(I98I). All these data must be geometrically corrected and refer to the same geographic base. The first correction is made on the multispectral image. This corrected image is then used as a reference to make the other corrections. This comes to consider the multispectral image as a planimetric base.

The following approach is then used:
1) digitization,
2) geometrical corrections, 3) Integration (or registration), and 4) processing. It should be noted that the integration have to be preceeded by a reformatting of data in order that these be in a usable form.

At the factorial image level, the factorial values may be very small.In order to have a better visualization, these values are stretched linearly between 0 and 255. Thereby, the means have often values of about 127-128, and do not contribute, practically, to explain the factorial bands. We then have in each factorial band an opposition between themes represented by lower gray levels and those represented by higher gray levels.

On the other hand, one or more bands may be set as supplementary elements. This is all the more useful that we process either multidate or composite images.

It could also be interesting to set pixels as supplementary elements. But this seems to be more complex to put in practice.

V. APPLICATION TO A DATA ARRAY

FAC may also be applied to an array where remote sensing and ancillary data cross (figure 2).

We will take as an example, the methodology used in a study carried out by the author within his thesis research (BOUSSEMA,I98I)

We define (figure 3) :

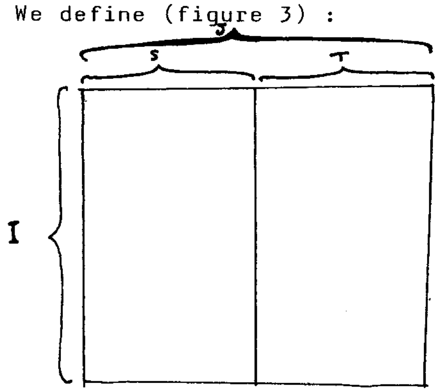

Figure 3.The crossing data array $K_{I,J}$.

I= set of town-ships; J=set of variables, S= set of statistical variables ; T= set of remote sensing variables; with $J=S \cup T$ and $S \cap T = \emptyset$

The statistical variables for each town-ship are of various types : demographic, geological,etc.

The remote sensing variables are the set of classes resulting from the different classifications made on the images corresponding to the different townships.

Array $K_{I,SuT}$ shows that the K_{IS} and K_{IJ} arrays are disjointed and may be studied separately. The first one statistically describes the townships and enables to bring out their similarities and dissimilarities. The second one enables to comprehend a classification on a landsat image.However, these two arrays risk to have different forms ; for example the first one may be a contingency array grouping counting data and that is positive or nul integers the values of which may be less or more high. On the other hand, K_{IJ} array may be logical containing ones and zeros, that is to say a class belongs or not to a town-ship. Regrou-

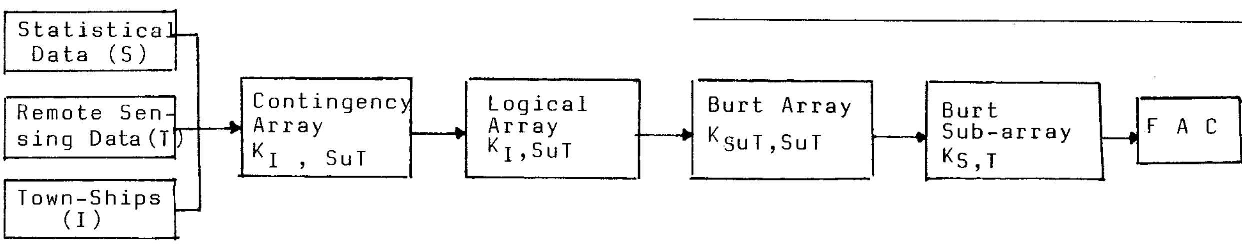

Figure 2. A description scheme of an analysis of a crossing data array.

ping these data of different natures may cause difficulties when analysing the results. For this reason it is necessary to re-codify the data.

K_{IJ} array may be best re-codified by transforming it into a logical description one. For this, the variables have to be grouped in different classes so that to each individual and for each variable one and only one class is associated. This partition may be done in various ways, for example after the variables histograms. (Figure 4)

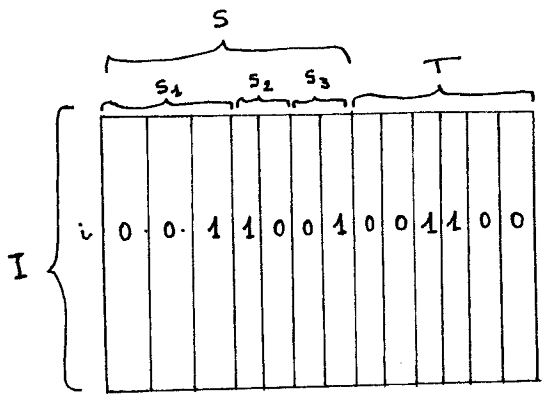

Figure 4. The logical array $K_{I,\ SuT}$

When the codification is done, we proceed with the analysis of the arrays with two objectives in mind :

1) Look for the features of the town-ships through remote sensing and statistical data. In this case we can apply directly the FAC to the recodifred array.

2) Look for a correlation between the remote sensing and statistical data. In this case, we will have to manipulate again the re-codified array.

To reach the second objective, we have to go through the two following steps:

1) The set I of town-ships being fixed once and for all, each class may be identified as a part of I. It will be the set of town-ships having this class. We then, define a symmetrical array K_{JJ} called table of Burt (BENZECRI,1976) :

K(j1,j2)= number of town-ships having classes j1 and j2 in common.

2) The study of correlation between remote sensing and statistical data consist in studying K_{ST} array (figure 5) :

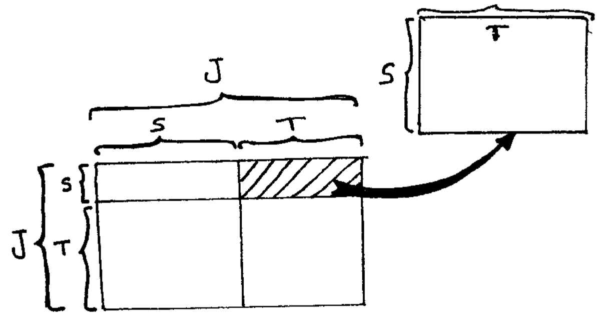

Figure 5. Array of Burt $K_{JJ}(=K_{SuT,SuT})$

K(s,t)= Number of town-ships having both "statistical" class s and "remote sensing" class t.

It is a question here of a sub-array of Burt which, in fact, is a bloc of the array of Burt. It is to this K_{ST} array that FAC will be applied.

VI. CLASSIFICATION USED

The selection of test zones and the ground truth data may be fed to the computer automatically through computer files.

With appropriate algorithms and accurate coordinates, a window is created around a point which we suppose to belong to the theme being studied. The window size may vary according to the importance of the central point .

Whithin this window, a classification is made. This classification is then used to classify the entire image.

This approach enables more precise cartographic transfer thanks to the coordinates of windows centers which are determined with an accuracy in accordance with the scale of work. Furthermore, Knowing that it is hard for the eye to detect subtle spectral differences, this approach also, makes it possible to extract more information.

The access to this kind of data bases may be done by the use of coordinates or other identification elements: region numbers, map numbers, aerial photos numbers, etc.

In this way, each zone, may be taken out from this system and processed separately. The wanted characteristics of this zone may then be examined quickly.

VII. CONCLUSION

Application of FAC technique has shown that the number of MSS spectral bands could be reduced considerably with very little loss of information and that the factorial image has a lower dimensions number. So, this approch optimizes the codification of images and improves their visibility.

Integration of ancillary data has improved the image dimensionality.

On the other hand, FAC help to grade and improve spectral data, favors a better comprehension and usage of Landsat data. A classification is made by simply threshholding the factorial bands. The legend of classes is obtained by their correlations with ancillary data.

There might also advantegeous to have multidate remotly-sensed data and reduce it to a lower dimension.

VIII. REFERENCES

BENZECRI J.P., 1976. L'Analyse des données. T.2.- L'Analyse des correspondances, DUNOD, PARIS.

BOUSSEMA M.R., 1981. Utilisation de l'Analyse Factorielle dans l'Etude de Corrélation entre données de Télédétection et données statistiques, Doctor-Ingeneer theses, ENSG, PARIS.

AUTHOR BIOGRAPHICAL DATA

Boussema M. Rached was born in Teboulba Tunisia on July 20, 1953. He received the Doctor Ingineer Dipoom in Geodetic Sciences from the Ecole Nationale des Sciences Georgrapgpqsiq.