Hypothesis-free Land Use Classification from LANDSAT-5 TM Image Data

by Bernd-Siegfried Schulz

- Institut für Angewandte Geodäsie Frankfurt am Main -

ABSTRACT: Conceptual and methodical postulations (e.g. definition of training fields, assumptions about distributions etc.) entering the classification of multispectral data, contain hypotheses which are critically reviewed here. In some respects they proved inappropriate, which gave rise to the development of a classification method that is introduced in the following, put to the test in the course of processing a TM data set, and used for the production of a thematic map.

ZUSAMMENFASSUNG: Begriffliche wie methodische Vorgaben (z.B. Definition von Trainingsgebieten, Annahmen über Verteilungen, etc.) bei der Klassifizierung multispektraler Daten enthalten Hypothesen, die kritisch in Frage gestellt werden. Sie erweisen sich z.T. als ungeeignet. Dies gab Veranlassung zur Neuentwicklung eines Klassifizierungsverfahrens, das im folgenden vorgestellt, im Zuge der Bearbeitung von TM-Daten für ausgewählte Objekte erprobt und zur Herstellung einer thematischen Karte angewandt wurde.

RESUME: Les données d'entrée conceptuelles aussi bien que méthodiques (p. ex. définitions d'aires d'exercice, suppositions de distributions, etc.) dans la classification des données multispectrales contiennent des hypothèses qui sont soumises à un examen critique. Elles se révèlent en partie inappropriées, ce qui a donné lieu au développement d'une nouvelle méthode de classification qui sera présentée ci-après. Ce nouveau procédé a fait ses preuves au cours du traitement des données TM pour des objets sélectionnés et a été utilisé pour la production d'une carte thématique.

Introduction

In the classification of land use types from multispectral digital remote sensing data it is common practice on the one hand to apply object categories drawn from the legends of topographic maps (e.g. sparsely built-up areas, mixed forest, etc.), and on the other to summarily describe with mathematical models (random distributions) the manifold manifestations of the land use types resolved into pixels.

Such practice rests without necessity on two hypotheses which deserved to be questioned and gave rise to the development of a hypothesis-free approach

1 Review of the 1st Hypothesis: Topographic Objects of Heterogeneous Land Use Fulfill the Conditions of One-to-one Mapping.

The basis for any classification is a one-to-one multispectral mapping of objects. The spectral properties and topographic significance of the objects to which this condition applies are initially unknown. It may be satisfied for only a few of their component elements but not for the superordinate object category (e.g. sparsely built-up area) which comprises a variety of such elementary objects (e.g. concrete, vegetation, etc.). These would have to be individually defined through classification. If this notion is correct, certain objects of a priori inhomogeneous land use ought not be assigned for classification over training sites. By the same token, these objects would only become identifiable after complete classification and a synoptic representation of the typical distribution of all their elementary objects.
Such considerations touch on fundamental questions and arise independently of the considerable progress being made in the spatial resolution of remote sensing systems.

A procedure in keeping with this would resolve the inhomogeneity of objects designated in topographic terms and thus satisfy the above condition. Each topographic object would be described by a set of elements. Objects consisting of only one element must be an exception (e.g. water bodies). The current practice of using training fields for supervised classification would have to be modified and restricted to elementary objects of one-to-one mapping characteristics only.

2 Review of the 2nd Hypothesis: Gray Level Variations Within a Training Field Have Normal Distribution.

Proceeding on the undisputed finding that multispectral scanner data are not normally distributed (1), classification schemes that discriminate topographic object classes by the criterion of highest probability (e.g. maximum likelihood classifier) must be ruled out. Accordingly, the figure (cluster) delineating the variation of the topographic object - when projected into the two-dimensional feature space - is not an ellipse, as it was thought to be, but a basically irregular shape, which is readily obvious from the feature planes after classification has been completed (2). Hence the shape of the cluster has to be determined in the course of classification. This shall be conducted hypothesis-free by taking into account only actual land use types of which segments are known.

3 Objectives of this Approach

Apart from hypothesis-free clustering for so-called elementary objects, further objectives are in:

3.1 Minimizing the impact of sensor system errors (4),
3.2 Improving the performance of the classification results in transfer and extrapolation from smaller study sites to other, resp. larger, areas of application,
3.3 Reducing the number of original bands by compacting information
3.4 Minimizing the effect of terrain relief on the classification results.

Through appropriate data preprocessing, such as the procedure described in the following, points 3.1 and 3.3 can be attained, and examination of the results indicates that 3.2 and 3.4 are feasible as well. This will be illustrated with examples.

4 Data Preprocessing

4.1 Standardization

The optimal rotation of an n-dimensional feature space, subsequent to standardization, for the purpose of minimizing the sum of the squared vertical distances of the sample points to the axes of the rotated coordinate system (principal component transformation), presupposes an orthogonal Cartesian coordinate system with axes of equal scale. If, for example, the assumption of equal-scale axes is not satisfied, the transforms are biassed with a model error showing in the sequence of the transformed images as an increased tendency of image information to persist (fig. 1, left column). Hence, optimal standardization is achieved when the tendency to information retention is lowest, resp. with maximum information compression (fig.1, right column). In the last transform only stochastic components remain. If we consider standardization as a procedure of displacing and stretching, resp. compressing, the histograms derived from the individual bands to the same means and variances, the principal component transformation packs 5-band LANDSAT-5 TM data (bands 1-5) and also, as recent studies showed, 3-band SPOT HRV data, into 2 bands of significant image content.

4.2 Principal Component Transformation

The standardized n-spectral sample data are subjected to principal component transformation (as per 4.1), producing n new data sets. (This data processing sequence has been recommended earlier in a different context (1).) The computation of the eigenvector $\underline{\lambda}$ and the eigenvector matrix $\underline{V}$ follows from the condition (cf. (3)):

$$\underline{\lambda} \cdot \underline{V}^T \cdot \underline{V} = \underline{V}^T \cdot \underline{S} \cdot \underline{V} \qquad (4.2.1)$$

The variance-covariance matrix results from:

$$S_{jk} = \frac{1}{N} \sum_{i=1}^{N} (g_j - \bar{g}_j) \cdot (g_k - \bar{g}_k) \qquad (4.2.2)$$

where j,k are the bands under consideration,

N the number of pixels in the data set,

$\bar{g}_j, \bar{g}_k$ the means of the gray levels within bands j,k,

and g_j, g_k the pixel values of bands j,k.

The transformated pixel values $\underline{g}_N$ are obtained from the standardized values $\underline{g}_A$ by:

$$\underline{g}_N^T = \underline{g}_A^T \cdot \underline{V} \qquad (4.2.3)$$

This transformation leads to an increase in the range of gray values in dependence on the number of spectral bands involved, so that values greater than 255 can occur. This extension of values can be undone by dividing by $\sqrt{n}$, where n is the number of spectral bands. With this, the last equation goes over into:

$$\underline{g}_N^T = \frac{1}{\sqrt{n}} (\underline{g}_A^T \cdot \underline{V}) \qquad (4.2.4)$$

The resulting transformed data enter into further evaluation. The transformation of the overall scene, which is about 6.7 times larger than the study site, is performed according to Eq.(4.2.4) utilizing the previously computed eigenvector matrix.

4.3 Effect of the Principal Component Transformation

The results of the principal component transformation can be summarized as follows:

- Information compaction from 5 to 2 significant bands for LANDSAT-5 TM (from 3 to 2 for SPOT HRV) in conjunction with substantial contrast enhancement (cf. plates 2-3),
- Decorrelation of data in the resultant bands after transformation as per Eq. (4.2.4),
- Considerable reduction of striping, resp. data errors, caused by faulty sensors.

To verify the latter effect, a LANDSAT-2 scene over the North Sea was selected (including Helgoland). The two plates 4 and 5, made before, resp. after, transformation as per Eq. (4.2.4), illustrate this effect. (The turbidity of the water in plate 4 around Helgoland does not disappear completely as plate 5 would suggest, its insufficient rendition has reprographic reasons.) The first two transforms of the LANDSAT-5 TM images (plates 2 and 3) are free of any discernible distortion, the images (and consequently the

data) are of excellent quality and rich in detail. They constitute the basis for classification. At the same time the five-dimensional problem has been reduced to a two-dimensional one. All other transforms contain only sensor system related errors and show no reference to the objects contained in the frame.

5 Evaluation of the Histogram

The further evaluation steps are not performed on the original LANDSAT-5 TM data but on their principal component transforms HKT1 and HKT2 (see 4). They contain a value range from 0 to 255 which determines a number plane of corresponding size. Calculation of the frequencies of the respective value pairs (gray value combinations) and subsequent logarithmic scaling and standardization to 255 produces, with a line interval of 10 units, the frequency distribution represented in diagram 1. Thicker lines bearing numbers indicate integer multiples of 100 units. Small crosses mark points of relative frequency maxima within a 3 by 3 matrix. If we now took each maximum, say, in decreasing order of frequency, as the center of a cluster that, with its surrounding field of variance defined by the method of unsupervised classification, describes an object, the result would certainly be useless. However, by linking these maxima with the maxima of the larger area, we obtain clues to land use types that are relevant for the topographic map. Seen from this aspect, the object group "forest" extends as a long ridge of frequencies between the coordinates (95,33) (coniferous) and (113,58) (deciduous). Near the lower left margin of the scattergram, between coordinates (100,30) and (130,15), five different types of water bodies can be found. All other maxima are irrelevant for further object definition. Here the principally problematic nature of the unsupervised classification method becomes evident, which defines individual classes on the basis of maxima automatically drawn from the histogram. Also, the shape of the histogram and the separability of objects depends on the initial data: a data set comprising a variety of objects of completely different spectral properties does not permit the fine discrimination between objects (e.g. different forest types) that would be possible if the data selection were initially limited to specific objects, i.e. object-specific data selection enhances object separation. This holds only for the procedure proposed here with all its individual steps but not for direct processing of original sample data. In addition, the standard variance to be selected in the course of standardization must not be set too low lest the subsequent histogram compression lead to loss of detail (see 4.1).

6 Hypothesis-free Cluster Development and Classification

From the evaluation of the two-dimensional histogram (see 5), the approximate centers of the clusters can be determined in a first step. These constitute the relative frequency maxima within the extensive histogram maxima. The shape of the clusters is roughly defined by lines of equal frequency and can be approximated by means of a series of empirically determinable circle areas which may vary in radius. The procedure is as follows: The first cluster center with the first circle area and all cluster elements it encompasses, allows a first partial classification that provides only a very incomplete areal representation of the land use under consideration. Yet it is the very gaps appearing here which yield information about direction and amount of the extension of the first partial cluster: if we search within this first partial result for those pixels which, being surrounded by already classified ones, are themselves not yet classified, and determine their coordinates (gray values of the principal component transforms), sort them in

descending order of frequency, check if they plausibly belong to the existing partial cluster to reject or include them for cluster expansion - we arrive at a new classification. This process is iterated until a planimetrically correct and complete representation of a land use type has been achieved. As can be seen from diagram 2, the emerging final cluster shape is, contrary to the assumption of the maximum likelihood method, not elliptic but of a principally different and hence arbitrary form.

7 Method

This method is kept hypothesis-free in that the selection of data sets (e.g. 500 by 500 pixels) is object-oriented and the transforms are examined for object-specific distributions. This allow conclusions about object discriminability. It avoids the pitfall of unqualified assignment of training fields which consists in imposing on a wide variety of land uses a catalog of expected object classes that is likely to exceed the class separation possible with remotely sensed data. The method thus reduces the risk of obtaining ambiguous object classes. Also, there are no a priori expectations as to the distribution functions. Instead, the clusters are empirically approximated by a series of elementary functions (circle areas), using as a reference, in checking the spatial distribution of the classified land use types, the actual land use extracted from IR color airphotos or other sources. Moreover, this classification is an iterative procedure with regard to the completion of a single cluster as well as the total of all clusters.

8 Extrapolation

In order to classify forest and water bodies in the area covered by the topographic map sheet C 5914 "Wiesbaden" , we started with a subscene of 512 E-W by 1024 N-S pixels which is located in the SE corner of that sheet and contains various kinds of forest areas and water bodies (see 1:50,000 map). Standardization and principal component transformation was first confined to that subscene. The eigenvectors are taken over for standardization and principal component transformation of the overall scene. Through this forced application of subscene parameters, the study site is mapped identically (cf. 4.1, 4.2).

9 Verification of the Results

The classification results developed over the SE corner of sheet C 5914 "Wiesbaden" were compared with contemporary IR aerial photography (July 30, 1984) as to forest classes. The land use types thus obtained and their spatial distributions determined the cluster shapes. After the forced application of the parameters relevant in the subscene classification to the entire extension of the map sheet, results were sampled and checked in the map boundary regions of Grävenwiesbach, Kelsterbach, Eltville and Kettenbach (defined as per topographic map 1:25,000) as well as in the sheet center region of Königstein. Here we utilized IR color imagery acquired two years ago in an aerial survey conducted in 1986 by the Land Survey Office of the State of Hesse. As far as the object classes "forest" and "water" are concerned, this time difference dhould pose no problem. The samples collected in this way showed very good accuracy without any gross errors. Particular attention was paid to forest classification in the Taunus hills but exposition-dependent misclassifications were not found. Also, corresponding boundary area checks corroborated that the signatures developed over the area of the municipal forest of Frankfurt are extrapolatable to the entire map sheet.

10 Prospects

The fact that classification results can be extrapolated with sufficient accuracy to larger regions has direct bearings on the automatized production of large-scale thematic maps (covering the Federal Republic of Germany, for instance). Further studies on this aspect as well as on more detailed object discrimination (e.g. different deciduous or coniferous forest classes) will follow.

References

(1) Ekenobi, S.L., 1986. Klassifizierung multispektraler Digitalbilder mit der Methode der trennenden Hyperflächen. Bul 54:23-29

(2) Quiel, F., 1986. Landnutzungskartierung mit LANDSAT-Daten. in: Fernerkundung in Raumordnung und Städtebau, 17 a publication of the Bundesforschungsanstalt für Landeskunde und Raumordnung, Bonn

(3) Wolf, H., 1968. Ausgleichsrechnung nach der Methode der kleinsten Quadrate. Bonn: Ferd. Dümmler, p.529

(4) Schulz, B.-S., 1986. Analyse der Datenqualität multispektraler Sensorzeilenabtaster. Bul 54:241-248

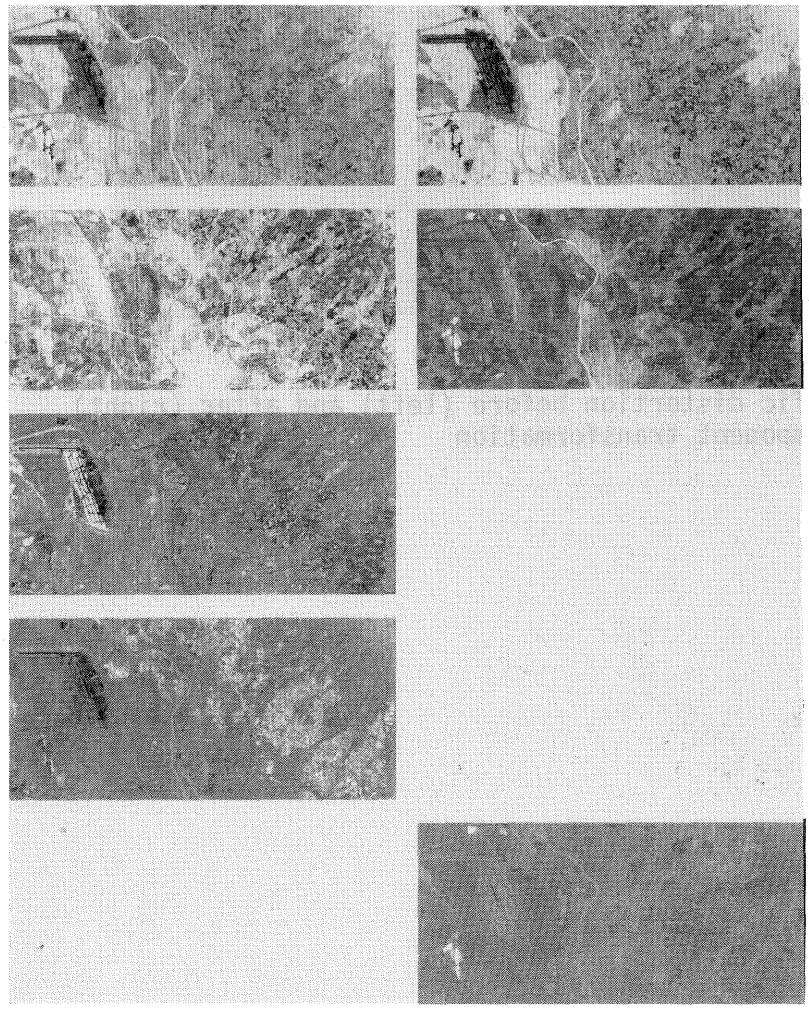
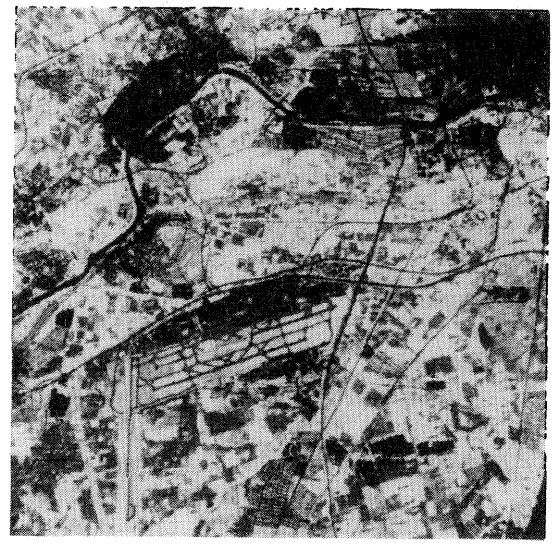
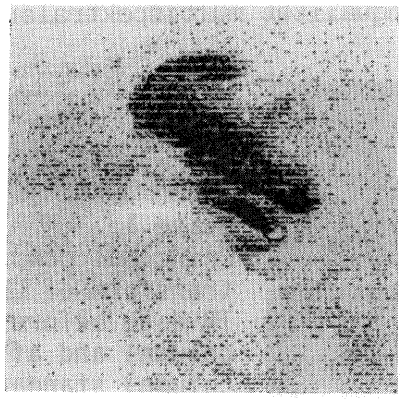
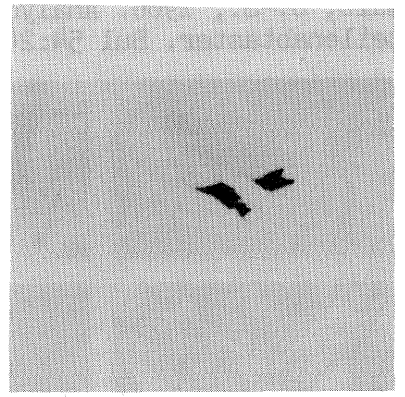


plate 1:
Principal component transform of LANDSAT-5 TM data before (left column) and after (right column) standardization. Area size: 480x1024 pixels



plates 2,3: First and second principal component transform (HKT1 and HKT2) of LANDSAT-5 TM data





plates 4,5: Sensor-specific distortion before (left) and after (right) principal component transformation

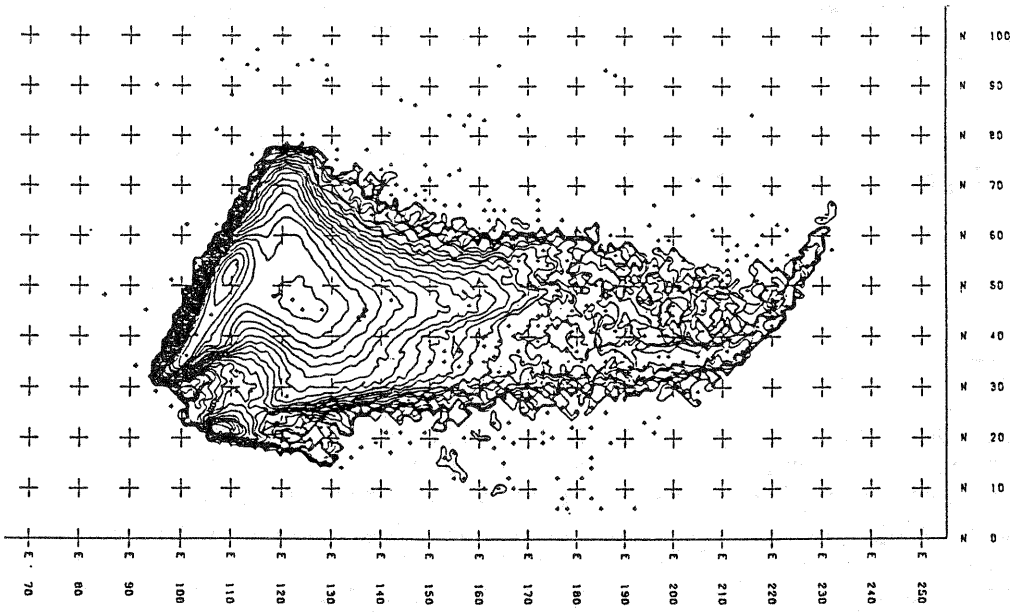


diagram.1: Two-dimensional frequency distribution (log. scale) of the first two principal component transforms HKT1 and HKT2 (HKT1 in direction X-(=E) and HKT2 in Y-(=N))

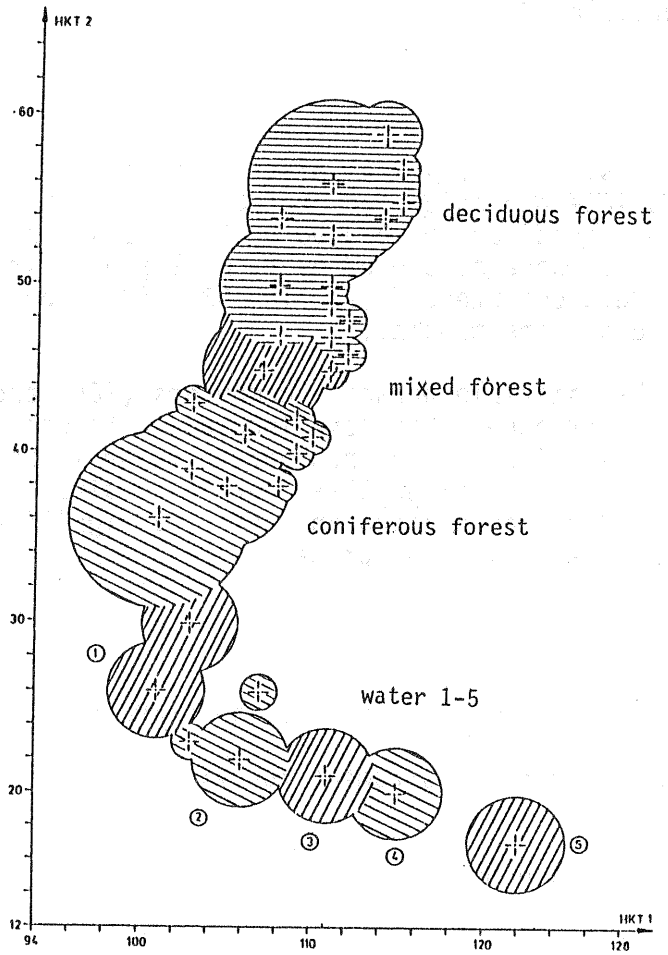


diagram.2: Hypothesis-free developed cluster for classification of forest and water bodies from TM data