# Using Supervised Classification Techniques for Producing Recent Soil Map for Northern Sinai Peninsula from Landsat TM-5 Data


**Prof. Dr. Ali Ahmed ElSagheer**
Prof. of Surveying and Geodesy

**Assoc. Prof. Maher Mohamed Amin**
Assoc.Prof. of Surveying and Geodesy

**Dr. Eng. Farag Bastawy Farag**
Lecture of Surveying

**Eng. Khaled Mahmoud Abdel Aziz**
Administrator of Surveying

Shoubra Faculty of Engineering - Benha University



**ملخص البحث**

**هذا البحث يقدم دراسة عن استخدام صور الأقمار الصناعية في أنتاج خرائط تصنيف التربة. في هذا البحث تم استخدام صورة القمر الصناعي Landsat TM-5 لجزء من منطقة شمال سيناء ذات قدرة تحليلية 28٫5 متر، وتم عمل تصنيف موجة(Supervised) للصورة لإنتاج خريطة لتصنيف التربة لمنطقة الدراسة واعتمدنا على اخذ بصمات التصنيف(Signature) من خريطة جيولوجية لنفس منطقة الدراسة بمقياس رسم 1:250000 .**

**في هذا البحث تم استخدام ثلاث طرق للتصنيف الموجة وهى (Minimum distance, Mahalanobis Distance, and Maximum Likelihood). ثم تم دراسة تأثيرعدد أصناف التربة في كل طريقة من الطرق المستخدمة (تدريجيا) فى اتجاة تنازلى حيث كانت البداية أكبر عدد من الأصناف الموجودة في الخريطة(32) ثم تقليل العدد في كل مرة بأهمال مساحة الأصناف الأقل في الخريطة الأساسية وكان التدرج من 32 صنف,28 ,24, 20, حتى16 صنف بأستخدام برنامج ENVI 4.4. وتم مقارنة الخرائط المنتجة من خلال طرق التصنيف المذكورة مع الخريطة الجيولوجية والتى تمثل خريطة الأساس (Base Map) لإختيار بصمات التربة المختلفة وكذلك للتحقيق من الخرائط المنتجة فى كل حالة من حالات الدراسة. وكانت أفضل طريقه من الطرق الثلاثة للتصنيف مع التنوع في عدد الأصناف هى Maximum Likelihood مع استخدام 16 صنف.**

## Abstract

This research presents a study using satellite image of Landsat TM-5 to obtain a soil classified map for Northern Sinai using the supervised classification techniques. These techniques depend on taking the training areas (signatures) on the image from reference such as geological map for the same area. Three methods of supervised classification were used, (Minimum Distance, Mahalanobis Distance and Maximum Likelihood) with different number of classes. Firstly, the maximum number of classes (32) has been used


---


2010, Civil *Engineering Research Magazine CERM, , Published by Faculty of Engineering,*

*Al-Azhar University, Cairo, Egypt.*

and then reducing the number of classes in each classification gradually, by neglecting small areas of classes in the reference soil map in each stage. The number of classes used in these techniques was 32, 28, 24, 20 and 16, respectively using ENVI 4.4 software. The output results of classification techniques have been compared with the available geological map. The comparison indicates that the method of Maximum Likelihood gives the best results in all used number of classes.

**Keywords:** Remote Sensing, Landsat TM-5 imagery, Soil Map, Supervised Classification, Accuracy Assessment.

## 1. Introduction

Satellite imagery has been used since 1970's as an accurate and cost- effective tool for extracting data, such as land cover mapping, environmental modeling and monitoring, updating of geographical databases and soil classification, with greater details and more reliability.

These information are fundamental requirements to support long-term, agricultural, industrial and economical development in any region. For these applications, usually a thematic map is required [16]. A thematic map displays the spatial variation of a specified phenomenon, such as land cover type. The trustworthiness and reliability of these thematic maps depend on how one classifies and analyzes remotely sensed images [9]. Classification of satellite images is one of the most commonly applied techniques used in remote sensing data processing. Classification involves performing a transformation from the numerical spectral measurements into a set of meaningful classes or labels, which can describe a landscape. Classification effects a transformation from a physical measurement into a cartographic or thematic description of the earth's surface, for examples, forest, built-up area, water bodies, etc. As such, classification can be viewed as a signal inversion process [15].

Methods of multispectral satellite image classification, founded on supervised and unsupervised classification algorithms, are based on the so-called pixel method of analysis and image classification [21].

In this paper the pixel based image analysis approach is selected normally, multispectral data are used to perform the classification and, indeed the spectral pattern present within the data for each pixel is used as the numerical basis for categorization. That is different feature types with different combination [16].

Pixel based approach is based on conventional statistical techniques, such as supervised and unsupervised classification [16]. In unsupervised classification, the computer or algorithm automatically group pixels with similar spectral characteristics (means, standard deviations, covariance matrices, correlation matrices, etc.) into unique clusters according to some statistically determined criteria. The analyst then re-labels and combines the spectral clusters into information classes [20]**.**

In a supervised classification, the identity and location of some of the land cover types (e.g., urban, agriculture, or wetland) are known a priori through a combination of fieldwork, interpretation of aerial photography, map analysis, and personal experience. The analyst attempts to locate specific sites in the remotely sensed data that represent homogeneous examples of these known land-cover types. These areas are commonly referred to as training sites because the spectral characteristics of these known areas are used to train the classification algorithm for eventual land-cover mapping of the remainder of the image. Multivariate statistical parameters (means, standard deviations, covariance matrices, correlation matrices, etc.) are calculated for each training site. Every pixel both within and outside the training sites is then evaluated and assigned to the class of which it has the highest likelihood of being a member [19].

The main objective of this paper is to obtain the soil classification map from Landsat TM-5 data by supervised approach and comparing the output map with its corresponding existing reference geological map.

## 2. Study Area, and Data Sources

The paper was carried out for apart from the Northern Sinai which located in the northern east of Egypt, Figure (1). Geographical coordinates of the investigated area lies between 30º 00' to 32º 15' N and 32º 15' to 33º 45' E.

The following data sources are available for the study area:

a- A multispectral Landsat TM-5 image of Sinai Peninsula, acquired on 1990, was used for this investigation Figure (2), which has three spectral bands with 28.5 spatial resolutions. This is a modified image obtained from internet site [22].

b- Geological map of the area of the area study (Sheet No.5), 1: 250 000, 1992. Sponsored by Academy of Scientific Research and Technology (Sinai Development Research Council).

## 3. Methodology

Soil classification map of investigation of the study area will be developed according to the following stages:

1. Scanning the available existing hard copy of the geological map (Sheet No.5), 1: 250 000.
2. Georeferencing the scanned map.
3. Image to map georeferencing.
4. Supervised classification using the following methods:
    a. Minimum distance technique.
    b. Mahalanobis distance technique.
    **c.** Maximum likelihood technique**.**

In the following subsections a brief discussion of the above mentioned stages is given.

### 3.1 Scanning the Available Existing Hard Copy of the Geological Map

Scanning is a very common procedure used for transforming hard copy maps into a digital format. The 1:250 000 hard-copy geological map has been scanned using Xerox scanner. The scanned image of geological map is in TIFF format with resolution of 200 dpi (dot per inch).

### 3.2 Georeferencing the Scanned Map

The process involved georeferencing of geological map to the Universal Transverse Mercator projection (UTM) using map corner coordinates in ENVI 4.4 software, where the projection information are, Project type is UTM, Reference Ellipsoid is WGS-72 and Zone Number is 36. The total Root Mean Square Error (RMSE) was 110.42652 meter as shown in Table (1).

## 3.3 Image to Map Georeferencing

Georeferencing was done by performing image registration and geometric correction utilities to reference pixel-based (raster) images to geological map. The geological map is a georeferenced image in UTM coordinates map projection. Map Reference Points (MRPs) are selected from the geological map as master/ base image to rectify the Landsat TM-5 image. It is necessary to spread MRPs throughout the study area on clearly identifiable points on both the master image and warp image for tie points in the image to image registration, Figure (3). After selecting enough MRPs, the images were warped using second order polynomial with Nearest Neighbour option to make a geometric correction or georeferenced image. The Root Mean Square Error (RMSE) obtained is 14.281 meter from 11 ground control points. The (RMSE) of the selected MRPs are shown in Table (2). For polynomial warping, the degree available is dependent on the number of MRPs defined where # MRPs > $(\text{degree} + 1)^2$ [14].

Therefore the required MRPs for rectification should be > 9 points.

## 3.4 Supervised Classification

The procedure of supervised classification will be more controlled by the user than unsupervised classification. Supervised classification starts with training, which results in various signatures, and followed by class assignment using a decision rule [5]. Training is the process of defining the criteria by which patterns are recognized [8]. The computer must be trained to recognize patterns in the data. Training samples for each land-use type are first selected by use of ground truth data, aerial photos, and maps. Signatures for each class are then generated from the training samples. Signatures are statistical criteria for the corresponding classes. Finally the pixels in the image are sorted into classes based on the signatures, by use of a classification decision rule. The decision rule is a mathematical algorithm that performs the actual sorting of pixels into distinct classes [5].

In this research Training samples for each land-use type are selected by the use of the only available mean i.e., geological map. In addition to, the Gaussian Minimum Distance, Mahalanobis Distance, and Maximum Likelihood/ Bayesian methods classifier was utilized to classify the image using ENVI 4.4 image processing software. These decision rules are discussed in the following subsections.

## 3.4.1 Minimum Distance Classifier

The mean spectral value in each band for each information class is determined. These spectral values comprise the mean vector for each class [10]. The minimum distance decision rule calculates the spectral distance between the measurement vector for the pixel to be classified and the mean vector for each signature [7]. The unclassified pixel is assigned to class membership based on the closest mean class value, or minimum distance [4].

The equation for classifying by spectral distance is based on the equation for Euclidean distance:

$$SD_{xyc} = \sqrt{\sum_{i=1}^{n} (\mu_{ci} - X_{xyi})^2} \qquad (1)$$

Where:

***n*** : Number of bands (dimensions)

***i*** : A particular band

***c*** : A particular class

***Xxyi*** : Data file value of pixel x,y in band i

***μci*** : Mean of data file values in band i for the sample for class c

***SDxyc*** : Spectral distance from pixel x,y to the mean of class c

When spectral distance is computed for all possible values of c (all possible classes), the class of the candidate pixel is assigned to the class for which SD is the lowest [12].

### 3.4.2 Mahalanobis Distance Classifier

The Mahalanobis distance algorithm assumes that the histograms of the bands have normal distributions. In addition to the mean value of the signature class the covariance matrix (covariance and variance) is included in the calculation [1]. Mahalanobis distance is similar to minimum distance, except that the covariance matrix is used in the equation. Variance and covariance are figured in so that clusters that are highly varied lead to similarly varied classes, and vice versa. For example, when classifying urban areas (typically a class whose pixels vary widely) correctly classified pixels may be farther from the mean than those of a class for water, which is usually not a highly varied class [12].

The equation for the Mahalanobis distance classifier is as follows:

$$D = (X\text{-}M_c)^T\,(Cov_c^{-1})\,(X\text{-}M_c) \qquad (2)$$

Where:

$\boldsymbol{D}$ : Mahalanobis distance

$\boldsymbol{c}$ : A particular class

$\boldsymbol{X}$ : The measurement vector of the candidate pixel

$\boldsymbol{M_c}$ : The mean vector of the signature of class c

$\boldsymbol{Cov_c}$ : The covariance matrix of the pixels in the signature of class c

$\boldsymbol{Cov_c^{-1}}$ : Inverse of $\boldsymbol{Cov_c}$

$\boldsymbol{T}$ : Transposition function

The pixel is assigned to the class, c, for which D is the lowest [6].

### 3.4.3 Maximum Likelihood/Bayesian Classifier

This classification method uses the training data as a means of estimating means and variances of the classes, which are then used to estimate probabilities. Maximum likelihood classification considers not only the mean or average values in assigning classification, but also the variability of brightness values in each class [2]. The maximum likelihood algorithm is the most common decision rule for supervised classification. This decision rule is based on the probability that a pixel belongs to a particular class. It assumes that these probabilities are equal for all classes, and that the input bands have normal distributions [7].

If you have a priori knowledge that the probabilities are not equal for all classes, you can specify weight factors for particular classes. This variation of the maximum likelihood decision rule is known as the Bayesian decision rule. Unless you have a priori knowledge of the probabilities, it is recommended that they not be specified. In this case, these weights default to 1.0 in the equation.

The equation for the maximum likelihood/Bayesian classifier is as follows [8]:

$$D = \ln(a_c) - [0.5 \ln(|Cov_c|)] - [0.5 (X-M_c) T (Cov_c^{-1}) (X-M_c)] \qquad (3)$$

Where:

***D*** : Weighted distance (likelihood)

***c*** : A particular class

***X*** : The measurement vector of the candidate pixel

$\boldsymbol{M_c}$ : The mean vector of the sample of class c

$\boldsymbol{a_c}$ : Percent probability that any candidate pixel is a member of class c (defaults to 1.0, or is entered from a priori knowledge)

$\boldsymbol{Cov_c}$ : The covariance matrix of the pixels in the sample of class c

$\boldsymbol{|Cov_c|}$ : Determinant of Covc (matrix algebra)

$\boldsymbol{Cov_c^{-1}}$ : Inverse of Covc (matrix algebra)

***Ln*** : Natural logarithm function

***T*** : Transposition function (matrix algebra)

The pixel is assigned to the class, c, for which D is the lowest [6].

## 4. Accuracy Assessment

Accuracy assessment forms important integral part of the classification process. No classification is complete until its accuracy has been assessed [3]. Accuracy assessment is done by comparing the supervised classification maps with the available geological map, this geological map represented the reference of training areas (signatures) taking on the image [18]. The error matrix and kappa statistics are two methods commonly used to do the accuracy assessment and are given under.

### 4.1 The Error Matrix

Error Matrix is the most effective way of representing classification map accuracy in which the individual accuracies of each category are mainly described along with both error of commission and error of omission. User’s accuracy, producer’s accuracy and overall accuracy can be judged with the help of error matrix method. The brief description of the accuracy indexes are given below.

#### 4.1.1 Overall Accuracy

The proportion of the reference pixels which are classified correctly is known as the overall accuracy. It is computed by dividing the total number of correctly classified pixels by the total number reference pixels. It is a very coarse measurement and does not provide the information about the classes that are classified with good accuracy [3]. Table (3) illustrates the error matrix for the maximum likelihood soil classification map with the 16 class, and the value for the overall classification accuracy and overall Kappa Statistics for this soil classification map.

#### 4.1.2 Producers Accuracy

Producer's accuracy is the probability of a reference (training) pixel being correctly classified, i.e., a measure of omission error. It is the number of pixels correctly classified as a land cover class divided by the total number of reference pixels for that land cover class [17].

### 4.1.3 User's Accuracy

User's accuracy indicates reliability, or the probability that a pixel classified in the image is really that land cover class on the ground. It is the number of pixels correctly classified as a land cover class divided by the total number of pixels that were classified in that land cover class [17]**.** Table (4) illustrates the producer's accuracy and the user accuracy for the maximum likelihood classification map with 16 classes**.**

## 4.2 Kappa Statistics

Overall Kappa statistics and individual Kappa values for each class are calculated for the thematic maps. Kappa coefficient measures the difference between the agreement of the reference data and classification results and the chance agreement of the reference data and a random classifier [13]. It provides a better measure of the accuracy of a classifier than the overall accuracy, and it takes into account the whole confusion matrix rather than the diagonal elements alone [11].

## 5. Results and Analysis

The main aim of this investigation is to assess the relative performance of three classification approaches in remote sensing, which are, maximum likelihood, mahalanobis distance and minimum distance methods, by using different number of classes which, affect their performance in terms of the overall classification accuracy. Tables 5, 6 and 7 show the overall accuracy and kappa results obtained for the three classifier methods in the study area.

These tables indicate that the obtained classification results from the maximum likelihood classifier are the best of the three methods for all different classes. In this method the number of 16 classes gives the best overall accuracy. Finally, the produced classified map based on this technique is shown in Figure (4).

## 6. Conclusion and Recommendations

This research investigate the use of satellite image of Landsat TM-5 to obtain a soil classified map for Northern Sinai using the supervised classification technique. This technique depends on taking the training areas (signatures) on the image from a reference geological map for the same area of scale 1:250 000. Three methods of supervised classification were used, (Minimum Distance, Mahalanobis Distance and Maximum Likelihood). The following are the results of this study:

1- The maximum likelihood gives the best results in all used number of classes. When using maximum likelihood, the overlay between the classified image and the reference soil map were 40.2584%, 44.2044%, 45.1891%, 47.3693% and 49.8404% in the cases of using 32, 28, 24, 20, and 16 classes respectively.

2- Water is almost classified perfectly, because both the producer's and user's accuracy are 95.94% and 97.90% respectively, due to the higher compatibility between the reference geological map and the spectral resolution in the classified image for water class.

3- More accurate soil classified maps could be obtained, if satellite data with high spatial and spectral resolution is available, beside the reliability in the ground truth information of the study area.

## The Following Section Illustrates the Definition of Each Class in Map Classification

**Qst** : Mixture of black and white sands, and silt.

**Qsd** : Sand dunes and sheets.

**Telmn**: Fossiliferous limestone, nodular in parts, with nummulites bassuni, alveolina frumentiformis, and orbitolites

**Khll** : Alternate beds of dolomite, limestone, marl, and clay; commonly rich in fossil.

**Qfg** : Fanglomerate

**Jmj** : Limestone with marl intercalations, sandy limestone in the upper part.

**Qw** : Wadi deposits.

**Qh** : Alluvial hamadah deposits.

**Tplhj** : Conglomerate composed of limestone cobbles and pebbles in sandy matrix.

**Ksd** : White chalk intercalated with thin clay bands and marl near top.

**Jbgh** : Coralline limestone intercalated with calcareous clays and algal limestone.

**Teleg** : Chalky limestone with flint bands and nodules at base and thin successive chert bands at top.

**Jsh** : Brown, fine to coarse-grained sandstone with intecalations of carbonaceous clay.