PRINCIPAL COMPONENTS AS A DATA REDUCTION AND NOISE REDUCTION TECHNIQUE

MARC L. IMHOFF

WILLIAM J. CAMPBELL

NASA/Goddard Space Flight Center
Greenbelt, Maryland 20771

ABSTRACT

The objectives of this study were to: (1) Assess the potential of principal components as a pipeline data reduction technique for thematic mapper data, and (2) Examine principal components analysis and its transformation as a noise reduction technique.

Two primary factors were considered:

1. How might data reduction and noise reduction using the principal components transformation affect the extraction of accurate spectral classifications, and

2. What are the real savings in terms of computer processing and storage costs of using reduced data over the full 7-band TM complement?

An area in central Pennsylvania was chosen for a study area. The image data for the project were collected using the Earth Resources Laboratory's Thematic Mapper Simulator (TMS) instrument. The TMS records data in seven band widths (.46-.52, .53-.60, .63-.69, .77-.90, 1.53-1.72, 2.04-2.24, and 10.43-12.33 μm) with a ground instantaneous field of view (GIFOV) of 30 meters. A set of surface feature verification sites corresponding to desired land cover/land use classes were geodetically measured and photographed using field teams and low altitude color infrared aerial photography. The photographs with the surface verification site boundaries were digitized, registered, and merged with the TMS data. A percentage of the surface feature verification sites was used for spectral signature training while the remaining sites were utilized for accuracy assessment.

A principal components analysis and its associated transformation was applied to six of the seven spectral bands. The thermal band 7 was not included in the initial transformation. Cost and classification accuracy comparisons were made using a supervised classification procedure applied to selected subsets of the transformed data and compared with results obtained by applying the same procedure to the full 6 band compliment.

Classifications were made on a subset consisting of three principal components axes and the full 6 band contingent for comparison. Overall classification accuracy for the transformed and reduced data was down 4 percent from that achieved using the full 6 bands. Processing costs for the transformed and reduced data were less than 53 percent of the costs required to process the 6 band data.

INTRODUCTION

The Thematic Mapper (TM) instrument on Landsat-4 is collecting nearly 15 times more data per unit area than the Multispectral Scanner (MSS). As a result of this increased data volume, data reduction techniques may be desired, or even required, by some users to reduce cost impacts in computer processing and personnel time.

Three basic techniques that are commonly used to affect the reduction of digital satellite data are: (1) Band selection, (2) Data transformations based on preliminary spectral classification, i.e., canonical analysis, and (3) data transformations based on overall data statistics, i.e., principal components.

Recent studies using band selection techniques applied to Thematic Mapper Simulator (TMS) data have yielded mixed results. Dottavio and Williams (1982) found that a band subset of three carefully selected channels slightly improved classification accuracies over those attained using the full compliment for forest cover types in North Carolina. Gervin et al. (1982), however, found that using all bands yielded higher classification accuracies than band subsets for mapping cover types in Michigan. Dottavio and Williams used the NS-001/MS Thematic Mapper Simulator and Gervin used the 18ML Scanner (also used in this study). The difference in their results may be due to the differences in the instruments themselves.

Data reduction techniques using statistical transformations have also been explored. Canonical analysis is known to improve class separability for most of the classes input to the transformation but the separability of some classes are sacrificed for the increased separability of other classes. At this writing it is not well documented as to exactly how the transform affects overall classification accuracy in different situations and objectives.

Canonical analysis also requires a preliminary spectral classification in order to develop the transformation matrix, a factor that makes this technique actually more expensive in terms of processing costs than simply classifying the full band contingent (Imhoff and Petersen, 1980).

Principal components can be used to exert a mathematical transformation requiring little preliminary spectral analysis prior to its application. Care must be exercised to include all important target features in the initial development of the transformation matrix to achieve an optimal result, but this requires only a fraction of the effort required for creating a preliminary spectral classification.

Principal components therefore appears to be a well suited technique for quick, inexpensive data reduction. The question remains, however, as to how many transformed data channels can be removed without reducing classification accuracy below tolerable limits. It also remains to be determined what those tolerable accuracy limits are in relation to the limitations of data processing and analysis costs. This is most probably a variable that will change with each user and application.

OBJECTIVE

The objective of this paper was to examine the potential of principal components analysis and its transformation as a pipeline data reduction and noise reduction technique. The two criteria by which the success of this technique was assessed were: (1) classification accuracy, and (2) data processing and analysis costs.

MATERIALS AND METHODS

The Thematic Mapper Simulator (TMS) data used in this study was acquired using the 18ML Scanner and is composed of seven bands (0.45-0.52, 0.52-0.60, 0.63-0.69, 0.76-0.90, 1.55-1.75, 2.08-2.35, 10.4-12.5 μm) (ORI, 1982). The instrument has a GIFOV of 30 x 30m and the data were collected from an airborne Learjet aircraft at an altitude of 45,000 feet. The area chosen for analysis was a site surrounding the Susquehanna nuclear steam electric generating facility in Berwick County, Pennsylvania (figure 1). This project was carried out as part of the overall NASA/NRC Energy Facility Siting Program (Campbell, 1982).

The TMS data suffered from several problems:

a. Considerable image distortions due to aircraft flight path movement were apparent in all bands,

b. Calibration problems and electronic noise appeared in the imagery in the form of line striping and beat patterns, and

c. A high frequency spatial distortion possibly due to aircraft and/or instrument jitter was present in all bands.

The distortions caused by aircraft/scanner jitter were not immediately removable and left unchanged. The effects of problems a and b above were handled as described below.

Prior to collecting training site statistics some extensive preprocessing was undertaken to remove some of the radiometric and geometric distortions inherent in the aircraft-collected TMS data. Two primary steps were taken:

a. Radiometric adjustment for scan angle effects. The look angle and subsequently the atmospheric path lengths vary systematically as the TMS scans across the flight path. As a result, reflectance data recorded for a particular target feature near nadir appear different from that of the same target feature off nadir with a longer atmospheric path length. This factor causes confusion in the classification of land cover categories over the image. In order to compensate for this effect, the raw TMS data were normalized to a predicted radiometric response at nadir.

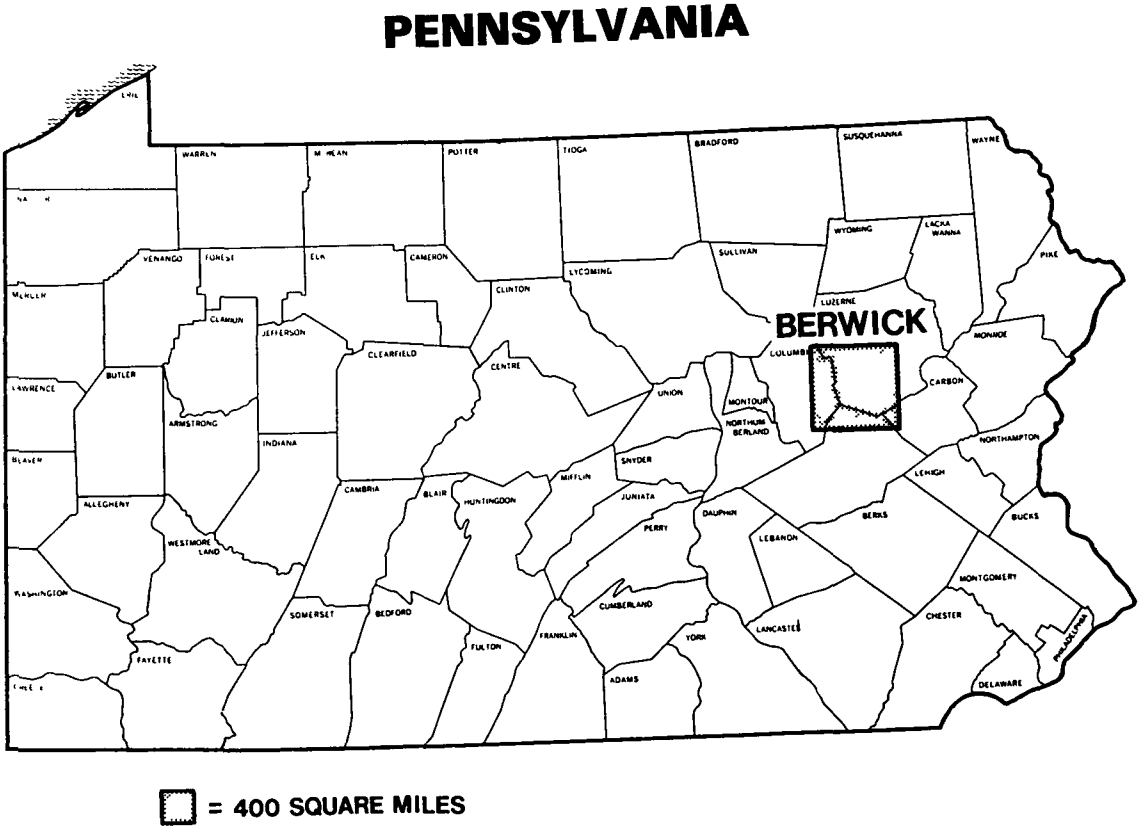


FIGURE 1. LOCATION OF STUDY AREA

b. Geometric correction. Geometric distortions perpendicular to the flight line were also inherent in the TMS data. These distortions were caused by variations in the look angle during data collection. A control grid was developed with a cell size equal to the TMS resolution cell size at nadir. A nearest neighbor resampling algorithm was then used to fit all of the TMS image data to the control grid.

GROUND TRUTH

Once the primary geometric and radiometric distortions were removed, the TMS data were geodetically precision registered to a series of digitally encoded, low altitude color IR aerial photographs which were in turn geodetically registered to US Geological Survey (USGS) 7.5 minute topographic quadrangles.

In order to exercise scientific control in comparing the classification accuracies achieved for the unaltered and transformed data sets, precise ground truth data were collected. The method that was designed for the study was to collect ground truth data coincident with the low altitude overflights. In reality, scheduling the concurrence of these events with cloudless, clear weather proved to be an impossible task. However, the time interval between the three events was 6 weeks, not optimum but adequate.

A rigorous cluster sampling procedure was designed to combine the ground truth surveys with the low altitude color IR digitized photography. Areas were randomly selected from USGS 7.5 minute quadrangle maps covering each test site. The randomly selected sites were visited and photographed in color and color IR. A professional survey team provided locational accuracy to within ± 1 foot with a laser geodimeter. The survey data were then combined with the digitized color IR data which were digitally registered with the USGS 7.5 minute quadrangle maps to produce georeferenced ground truth which were in turn used to generate pixels of known identity in the areas sampled for both training set generation and accuracy assessment. The main advantage of this procedure is that cluster sampling provides identities for more pixels per area visited than systematic sampling or simple random sampling.

Using the cluster sampling and survey technique 180 training sites were documented and registered. Approximately half of the training sites were used for signature derivation and the remainder were reserved for classification accuracy assessment.

DATA REDUCTION--PRINCIPAL COMPONENTS

In order to effect a data reduction, a general principal components analysis and its transformation was used to create a new set of data channels whereby more of the system variance might be explained by a fewer number of data channels or axes.

Principal components analysis and its transformation was selected due to its relative simplicity and general availability. Jet Propulsion Laboratory's VICAR, ESL's IDIMS, and the Pennsylvania State University's ORSER system all have principal components options. Principal components is a technique whereby a new set of axes is defined for the data such that the first principal component or axis explains as much of the total variance as can be explained by any single variable or axis. The second principal component or axis explains as much of the remaining variance as can be explained by any axis or orthogonal (uncorrelated) to the first. The third principal component continues this process and so on until the dimensionality of the data is exhausted (Merembeck and Borden, 1978). The effect is that most of the information inherent in the many spectral bands is combined or explained by one, two, or three of the principal components.

In this application a general principal components (PC) analysis was used to derive the transformation matrix for the TMS data. A simple polygon targeting training site selection and statistical calculation program was used to determine mean responses for each band and a variance-covariance matrix for a general cross section of the data. The transformation training site transected all of the major cover types and/or target features found in the data. A set of Eigen values was calculated from the training statistics and the transformation matrix was generated.

Once the transformation matrix was applied, a variance-covariance matrix and correlation matrix was generated to determine the effectiveness of the transform and compare the new axes with the unaltered data (figure 2). In this case the data represented by axes 1-3 accounted for 98 percent of the total system variance and raw channels 2, 3 and 4 had the highest correlations with the first three principal axes. For this application the data represented by axes 1, 2 and 3 were used for classification purposes for comparison with the full 6-band contingent. The data represented by axes 4, 5, and 6 were discarded.

For the purpose of simplicity, throughout the remainder of the text, the full 6-band unaltered data will be referred to as "raw" or "raw 6-band" data and the transformed and reduced data will be referred to as "PC" or "PC 3 axes" data.

ACCURACY ASSESSMENT

As described above, the training site boundaries were delineated on paper copies of the low altitude aerial photography. These boundaries were then transferred to the digitized version of the same photography using an interactive CRT and track ball-driven cursor.

Once in digital format the randomly selected training site boundaries for each target feature were divided in two categories:

CORRELATION MATRIX PRINCIPAL COMPONENTS AXES 1 2 & 3

VS

UNALTERED 6 CHANNEL DATA

UNALTERED CHANNELS	PRINCIPAL COMPONENTS AXES 1	2	3
1	-.72	.65	.27
2	-.67	.79	.20
3	-.72	.80	.25
4	.91	-.10	-.69
5	.23	.55	-.35
6	-.35	.65	-.10

COVARIANCE MATRIX FOR PRINCIPAL COMPONENTS TRANSFORMED DATA

6 CHANNEL

PC AXES	1	2	3	4	5	6
1	526.51					
2	-74.02	310.24				
3	15.11	18.05	48.10			
4	10.36	7.63	-1.22	9.98		
5	5.99	1.45	1.71	-.12	6.48	
6	-5.41	.28	-.83	.19	-.13	3.62

FIGURE 2. VARIANCE-COVARIANCE MATRIX AND CORELATION MATRIX FOR 6 BAND AND PRINCIPAL COMPONENTS TRANSFORMED DATA

a. A statistical (STATS) category from which spectral signatures for classification were developed, and

b. An accuracy assessment category (ACC) against which the classification was to be tested.

The two sets of training site boundaries or polygons were stored as images, each polygon retaining its geometry and spatial juxtaposition as it appeared on the georegistered digital data. The two sets of polygon images were then used for comparison with classified data. The STATS polygon boundaries were transferred to both the PC TMS data and the raw TMS data for the generation of spectral signature statistics for classification. Once the spectral signature banks had been developed for the PC and raw TMS data, the two scenes were classified using a maximum likelihood classifier. The same algorithm was used to generate the spectral signatures and classify both the PC 3 axes and raw 6-band data sets.

After classification, accuracy assessment was made by creating contingency tables of the classified PC and raw TMS data sets against the ACC image. Calculations derived from the contingency tables provided classification accuracy statistics in the form of:

a. Probability that a pixel classified as class i is class i,

b. probability that a pixel that is class i is classified as class i, and

c. overall (combined for all classes) probability of correctly classifying a pixel given this set of circumstances.

COST ASSESSMENT

Cost assessment was made by documenting the time required to generate the classifications for both the raw 6-band and PC 3-axes data sets. The items measured were:

a. Central Processing Unit (CPU) time (seconds)

b. Computer connect time (minutes)

c. Man-hours

The costs were documented in the form of time and not dollars since the dollar/time relationship changes for each user and set of circumstances.

The time required for the principal components analysis and transformation was included in the costing of PC 3-axes classifications. The time costs of preprocessing were common to both data sets and not included.

RESULTS

The contingency tables comparing the classified data sets with the ACC image or accuracy data set provided information concerning the classification accuracies for each data set. Overall statistics and statistics for each class were calculated from the contingency tables comparing classification accuracies or classification perfoemance level for the raw 6-band data and the PC 3-axes data (figure 3). In general, the accuracy statistics were fairly good.

Accuracies for some classes such as coniferous forest, orchards, mixed forest and meadow were low due to the lack of good training sites for these cover types in the Berwick area.

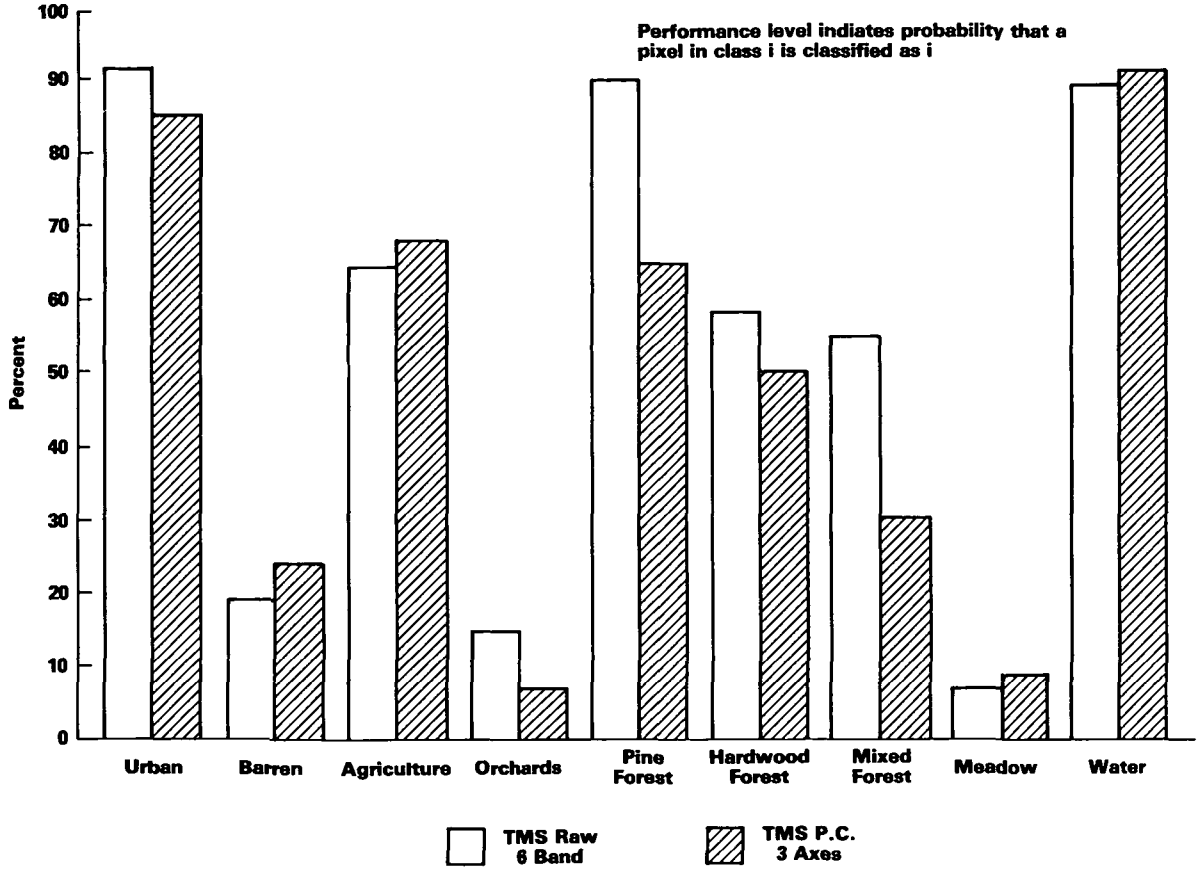
The classification performance comparison revealed that the overall probability of correctly classifying a pixel was slightly lower for the PC 3-axes data (4.3 percent) than for the 6-band raw data (figure 4). A class by class analysis reveals that for most classes the probability that a pixel classified as class i is class i and the probability that a pixel in class i is classified as such both decreased slightly using data reduction. The probabilities of correct classification for a few classes, however (barren, meadow, and water), actually increased using the reduced data.

The cost analysis calculated the central processing unit (CPU) time (in seconds), the man-hours, and the computer connect time required for the generation of training site statistics and the actual classification of the raw 6-band and PC 3-axes data sets. The time required for the TMS preprocessing was not included as it was the same in both cases. The cost statistics for generating the PC analysis and transformation were included in the costing of the transformed data.

The cost comparison showed that the raw 6-band data required 13000 CPU seconds, 40 man-hours, and 2580 minutes of computer connect time. The addition of two extra spectral bands (over the usual four associated with MSS data) greatly increased the time required to process the classification statistics.

The PC 3-axes data required 7000 CPU seconds, 22 man-hours and 1260 minutes of connect time (cost of data reduction technique included). This represents a 46.15 percent decrease in CPU time, a 45.00 percent decrease in man-hours and 51.16 percent decrease in computer connect time (figure 5).

Due to the technically complex and fiscally demanding nature of the project of which this study was a part, it was not possible to generate results for raw band selection. Studies performed on data collected using this same instrument, however, have indicated that raw band selection also did not improve overall classification accuracy (Gervin, et al., 1982).



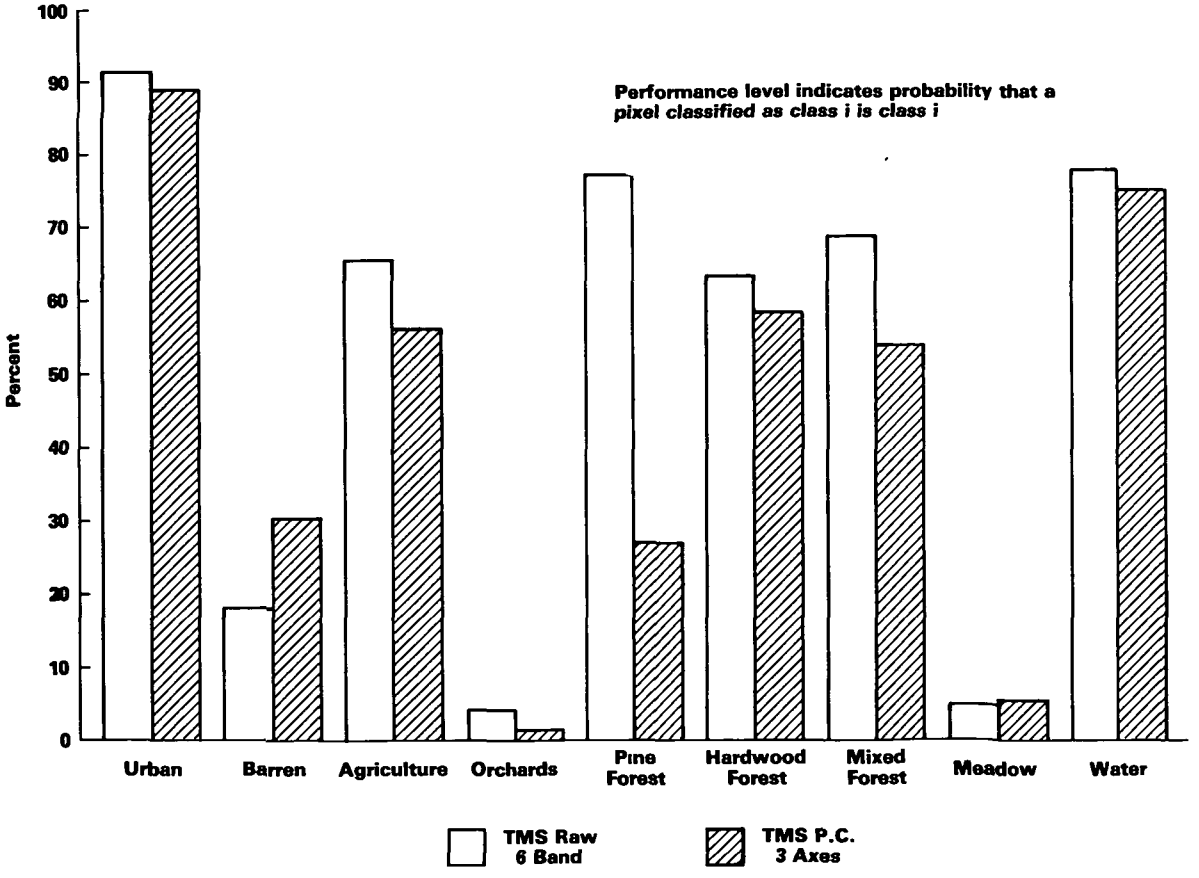


Figure 3. Comparison of TM Simulator 6 band raw and PC transformed and reduced 3 axes classification performance levels.

PROBABILITY OF CORRECTLY CLASSIFYING A PIXEL

PRINCIPAL COMPONENTS 3 AXES	UNALTERED (RAW) 6 BAND
70.30%	74.61%

Figure 4. Overall performance level

COMPARISON OF WORK REQUIRED TO PROCESS 6 BAND VS. 3 BAND (AXES) DATA

6 BAND DATA		3 BAND (PC 3 AXES) DATA	
CPU SECONDS	13,000	CPU SECONDS	7,000
MAN HOURS	40	MAN HOURS	22
CONNECT TIME (MINUTES)	2,580	CONNECT TIME (MINUTES)	1,260

SAVINGS OVER 6 CHANNEL DATA	
45.15	% LESS CPU TIME
45.00	% LESS MAN HOURS
51.16	% LESS CONNECT TIME

FIGURE 5. COST COMPARISON FIGURES FOR CLASSIFYING THE RAW 6 BAND TMS AND THE REDUCED PC TRANSFORMED DATA

CONCLUSIONS

The principal components analysis and transformation was successful in removing noise. By concentrating the noise on lower order axes, color composite images of increased quality could be produced from axes 1, 2 and 3. In this application, data reduction effected by a principal components analysis and its transformation and the removal of the lower order axes did indeed adversely affect the overall classification accuracy. The reduction in classification accuracy, however, was minimal and may be insignificant in most applications. On the other hand, the cost savings afforded by the reduced data were substantial, > 47 percent--more than enough to offset the decrease in accuracy.

More research needs to be performed to compare raw band selection against data reduction techniques such as principal components which require transformations. It is imperative that this research be done using the actual TM data itself as recent studies appear to indicate that the TM data are quite different in character and quality from the sensors designed to simulate them. It is also important to test this procedure on a variety of target areas as the effectiveness of this and other transformations as a data reduction technique may vary depending upon the character of survey area.

REFERENCES

1. Campbell, W. J., 1982. Integration of Remotely Sensed Data with Geographic Information Systems for Application in Energy Management. Proceedings of the 1982 Conference on Energy Resource Management, Baltimore, Maryland.

2. Gervin, J., 1982. Comparison of Land Cover Information from Landsat MSS and Airborne TMS for Input to Hydrological Models: Preliminary Results. Proceedings of the 1982 Conference on Energy Resource Management, Baltimore, Maryland.

3. Dottavio, L. L. and D. L. Williams, 1982. Mapping a Southern Pine Plantation with Satellite and Aircraft Scanner Data: A Comparison of Present and Future Landsat Sensors. Journal of Applied Photo graphic Engineering. Volume 8, Number 1.

4. Imhoff, M. L. and G. W. Petersen, 1980. The Role of Landsat Data Products in Soil Surveys. Institute for Research on Land and Water Resources, The Pennsylvania State University, University Park, PA.

5. Merembeck, B. and F. Y. Borden, 1978. Principal Components and Canonical Analysis for Reduction of Dimensionality of Large Data Sets. Technical Report 5-78, Office for Remote Sensing of Earth Resources, The Pennsylvania State University, University Park, PA.