ESC-TR-2004-085

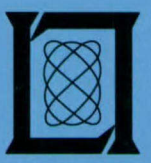
Project Report
HTAP-22

Effects of Lossy Compression of Hyperspectral Imagery

J.K. Su
M.K. Griffin
S.M. Hsu
S. Orloff
C.A. Upham

30 December 2004

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS



Prepared for the Department of the Under Secretary of Defense, S&T,
under Air Force Contract F19628-00-C-0002.

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Department of the Under Secretary of Defense, S&T, under Air Force Contract F19628-00-C-0002.

The ESC Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Gary Tutungian
Administrative Contracting Officer
Plans and Programs Directorate
Contracted Support Management

Massachusetts Institute of Technology
Lincoln Laboratory

Effects of Lossy Compression of Hyperspectral Imagery

J.K. Su
Group 103 (formerly Group 97)

M.K. Griffin
S.M. Hsu
S. Orloff
C.A. Upham
Group 97

Project Report HTAP-22

30 December 2004

Lexington Massachusetts

EXECUTIVE SUMMARY

Hyperspectral imaging (HSI) sensors provide imagery with hundreds of spectral bands, typically covering VNIR and/or SWIR wavelengths. This high spectral resolution offers promise for many applications, but it also produces enormous volumes of data, which may be problematic for storage and transmission. Lossy compression may therefore be necessary, but application performance degradation that results from compression is of concern. This report documents results for a spectral-spatial lossy compression scheme and a variety of applications: normalized difference vegetation index (NDVI), integrated column water vapor (CWV), and background classification.

The compression scheme first performs principal-components analysis spectrally, then discards many of the lower-importance principal-component (PC) images, and then applies JPEG2000 spatial compression to each of the individual retained PC images. Two different rate-allocation methods, which select the spatial compression ratios, are considered.

The assessment of compression effects considers general-purpose distortion measures, such as root-mean-square difference. It also examines changes in NDVI and CWV data products and proposes statistical tests for deciding whether compression causes significant degradations in classification results.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to CAPT Frank Garcia and Lt Col Karl Dahlhauser, DUSD (S&T), Program Managers, Hyperspectral Technology Assessment Program (HTAP), for their enthusiastic support.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

1. INTRODUCTION

A *hyperspectral imaging* (HSI) sensor gathers radiance imagery over a wide range of the electromagnetic spectrum with high spectral resolution. The sensor produces an $N_x \times N_y$ image in which each pixel has $N_c \approx$ 100–200 spectral *bands*. Such an image is commonly referred to as an $N_x \times N_y \times N_c$ *cube*. For example, the HYDICE (Hyperspectral Digital Imagery Collection Experiment [22]) airborne HSI sensor produces $320 \times 320 \times 210$ major frames of 16-bit values, with spectral coverage from 400–2500 nm and spectral resolution of 5–16 nm per band. Several major frames can be concatenated to form a larger cube; i.e., four major frames yield a sample HYDICE cube of dimensions $320 \times 1280 \times 210$. HSI thus provides both spatial coverage and rich spectral information, which facilitates applications such as background classification, material identification, and target detection.

On the other hand, the sheer volume of data produced by an HSI sensor creates problems of storage and transmission. The example $320 \times 1280 \times 210$ HYDICE cube requires 153.8 MB of storage. Lossless compression can usually reduce this requirement by a factor of 2–3, which still puts the storage requirement in the tens of megabytes and may be insufficient for timely transmission, dissemination, or archiving. *Lossy compression* then becomes necessary. Because lossy compression discards potentially useful information, researchers are interested in evaluating its effects on application performance [20, 24, 26]. This report employs a simple but effective lossy compression scheme and studies its effects on a number of applications, namely normalized difference vegetation index, integrated column water vapor, and background classification.

1.1 SELECTION OF COMPRESSION STRATEGY

This report focuses on what can best be termed "general-purpose lossy compression." To place the report in context, this section discusses the matter of selecting a compression strategy. The selection of a compression strategy involves factors such as whether or not lossy compression can be tolerated, the required amount of lossless or lossy compression, and the intended application area. Other factors include implementation complexity, onboard or ground-based processing, latency, error resilience, cost, and standardization.

Figure 1 presents a decision flowchart that considers the most general aspects of compression: namely, lossless or lossy compression, the end application for the decompressed data, and the distortion measure should lossy compression be required. The other factors mentioned above certainly also play a significant role, but the flowchart addresses the basic question of whether lossy compression is even necessary, and if so, whether existing or novel compression techniques are needed.

The top of the flowchart posits the most basic question: Is lossless compression required, and is it achievable within the problem constraints? If so, lossy compression becomes unnecessary. However, problem constraints can make lossless compression impractical or impossible. A space-borne HSI sensor with limited downlink capability is a prime example. Even though the sensor can gather large amounts of data, there is insufficient bandwidth to transmit all of the data to the ground. In such situations, lossy compression becomes necessary.

Within the realm of lossy compression, other decisions follow. The second question asks whether

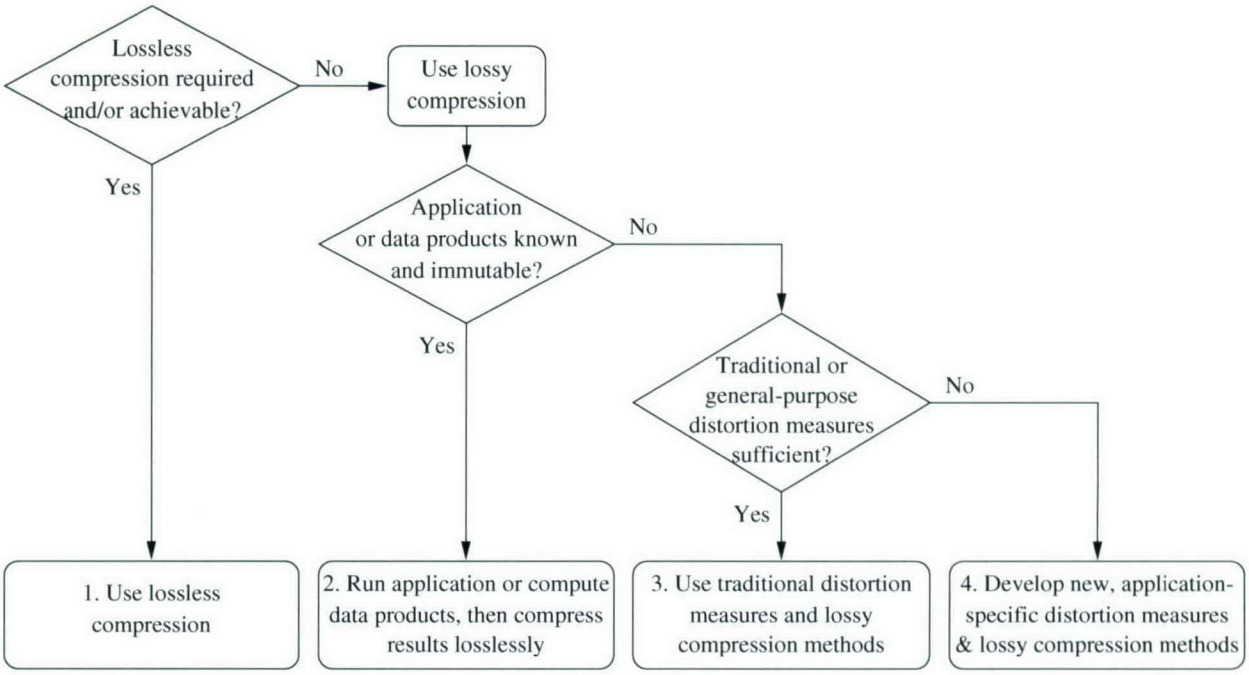


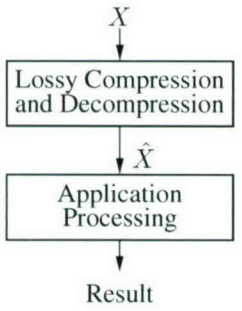
Figure 1. Flowchart for selection of compression strategy.

the intended application and/or data products are already known and will never change. If so, then a trivial solution follows: Simply perform the application processing or compute the desired data products, and then use lossless compression of the results. (This solution assumes that the products have far lower storage/transmission requirements than the unprocessed data.) An example is target detection. One could apply the detection algorithm to generate a binary-valued image, where cleared (0) pixels indicate "no target" and set (1) pixels indicate "target." This binary image requires far less storage or transmission bandwidth than the HSI data from which it was derived.

In most cases, the trivial solution is not a realistic choice. Then the third question comes into play: Are the distortion measures used in traditional lossy compression methods sufficient? If so, then the wealth of knowledge, experience, and implementations of traditional lossy compression methods can be employed. For example, an enormous amount of research and development has been devoted to the theory and practice of lossy compression under the mean square difference (also called mean square error) distortion measure. On the other hand, if traditional distortion measures do not apply, then new ones must be developed, along with suitable lossy compression methods.

1.2 APPROACHES TO LOSSY COMPRESSION AND APPLICATION PROCESSING

Lossy compression followed by application processing is a two-stage process, as depicted in Figure 2. The first stage performs lossy compression on the original, noncompressed HSI cube, denoted by X. The compressed version of the cube is a bitstream (not shown) that can be stored or transmitted more efficiently than X; decompression of the bitstream yields a decompressed, approximate cube, denoted by $\hat{X}$. In the



Conventional	Backward	Forward	Joint
Designed to min $D(X, \hat{X})$	Redesign for processing X	Designed to min $D(X, \hat{X})$	Redesign for processing $\hat{X}$
Designed for X	Designed for X	Redesign for $\hat{X}$	Redesign for $\hat{X}$

Figure 2. Lossy compression and application processing.

second stage, application processing is performed on $\hat{X}$. As a result of lossy compression, $\hat{X}$ is not an exact replica of X; the inaccuracy between X and $\hat{X}$ is given by the distortion $D(X, \hat{X}) \geq 0$. Such distortion means that the results of application processing on $\hat{X}$ will generally differ from the results that would have been obtained if processing had been applied directly to X. Moreover, good application performance — not storage or bandwidth reduction — is usually the ultimate goal.

We identify four possible approaches to this problem. In the *conventional approach*, compression and processing are designed without consideration of one another. Compression attempts to minimize $D(X, \hat{X})$ subject to a constraint on the size of the bitstream for the compressed cube. Processing is designed to operate on the original, noncompressed data X. The conventional approach is based on the idea that, if $\hat{X}$ is a sufficiently faithful approximation of X, then processing $\hat{X}$ should produce results similar to those from processing X.

In the *backward* and *forward approaches*, one of the two stages is redesigned to accommodate the other, which is held fixed. Knowledge about the fixed stage is carried backward or forward (hence the name) into the other stage. In the backward approach, knowledge about an existing processing stage — intended for use on X — is carried backward into the compression stage, which is altered in an attempt to make $\hat{X}$ "fit" as closely as possible with the processing method. Conversely, in the forward approach, the compression stage remains fixed, and knowledge about its behavior is carried forward into the processing stage, which is modified to process $\hat{X}$ instead of X.

Finally, the *joint approach* redesigns both stages together to obtain the best overall performance. Some realistic constraints must be introduced to prevent this approach from degenerating into a trivial solution (cf. Section 1.1). Consider target detection: With no constraints, a trivial joint solution simply performs detection on X, produces a binary detection-indicator image $\hat{X}$, (e.g., a pixel in $\hat{X}$ equals 0 to indicate "target present" or 1 to indicate "target absent"), and then compresses $\hat{X}$ losslessly. We adopt the conventional approach for several reasons. First, the backward, forward, and joint approaches tend to be highly specialized for a single application, while there is a wide variety of HSI applications (e.g., environmental monitoring, background classification, material identification, target detection, and anomaly detection). The assumptions and algorithms employed by different applications also vary significantly. The conventional approach means that the compression stage can be implemented separately and its effects then evaluated on a number of applications. Second, the nonconventional approaches often introduce additional assumptions about the characteristics, such as probability distributions, of the data involved. However, characterization of the statistical properties of HSI data remains an ongoing challenge [19, 29]. Finally, lossy compression

and application processing are each sophisticated fields in their own right, and the associated algorithms are very complicated. The nonconventional approaches can become intractable, and the results are limited to a particular application. For these reasons, most current work on examining the effects of HSI compression on applications has taken the conventional approach, and we do so in this report.

1.3 UNDERCOMPRESSION AND OVERCOMPRESSION

We now introduce the intuitive notions of *undercompression* and *overcompression*. Increasing the amount of compression discards more information, which may degrade application performance. As more information is discarded during compression, the more likely application performance degradation becomes. If too much information is discarded due to excessive compression, then performance becomes unacceptable. Ideally, we would like to compress as much as possible while still maintaining acceptable performance. We use the terms *undercompression* and *overcompression* to describe the failure to reach this optimal balance.

In undercompression, performance remains acceptable, but we could have compressed even more. The penalty for undercompression is excessive storage or transmission bandwidth. While such wastefulness could be expensive, if we realize that we have undercompressed the data, we might be able to compress it further or increase the compression ratio in subsequent collections.

In contrast, overcompression results in both unacceptable performance and **irretrievable loss of information**. Even if we discover that our application is underperforming and requires data with greater accuracy, there is no way to recover the discarded information. The penalty for such a mistake could be serious. We therefore assume that undercompression is preferable to overcompression. We choose to err on the side of caution, and our assessments will reflect this assumption.

1.4 NOTATION

We expand the notation here. As introduced above, we denote the original, noncompressed cube by X and a decompressed cube by $\hat{X}$. Let N_x and N_y denote the number of spatial columns and rows, respectively, where a spatial coordinate is given by (x, y), with $x \in \{1, 2, \ldots, N_x\}$, $y \in \{1, 2, \ldots, N_y\}$. Similarly, we denote the number of spectral bands by N_c, and we indicate the index of a band by $k \in \{1, 2, \ldots, N_c\}$. A single data value in X is indicated by $X[x, y, k]$. Letting $\vec{x} = (1, 2, \ldots, N_x)$ and $\vec{y} = (1, 2, \ldots, N_y)$, we use $X[\vec{x}, \vec{y}, k]$ to indicate the entire $N_x \times N_y$ *image* of data values in the kth band of X.

With $\vec{k}$ similarly defined, the notation $X[x, y, \vec{k}]$ denotes the spectrum associated with spatial location (x, y). $X[x, y, \vec{k}]$ is sometimes treated as a column vector in N_c-dimensional Euclidean space. The *spectral angle* between two spectra $\vec{u}$ and $\vec{v}$ is

$$\theta(\vec{u}, \vec{v}) = \arccos\left(\frac{\langle \vec{u}, \vec{v} \rangle}{\|\vec{u}\| \|\vec{v}\|}\right), \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product, and $\|\vec{u}\| = \sqrt{\langle \vec{u}, \vec{u} \rangle}$. Spectral angle is sensitive to differences in the spectral shapes of $\vec{u}$ and $\vec{v}$ but insensitive to illumination changes: Scalar multiplication of $\vec{u}$ or $\vec{v}$ does not change $\theta(\vec{u}, \vec{v})$.

1.5 RATE AND COMPRESSION RATIO

For any given data representation (e.g., X or $\hat{X}$), the *rate* of the representation is defined as rate = (number of bits required by the representation)/(number of data values represented), which has units of bits per data value. For hyperspectral data, one often employs units of *bits per pixel per band* (bpppb). The raw data values in X are typically stored using B-bit precision (e.g., 8-, 10-, or 16-bit integers), so the rate of X is B bpppb. Although the values in $\hat{X}$ also have a bitdepth of B, the associated bitstream is essentially an identical description of $\hat{X}$. Consequently, the rate of $\hat{X}$ is

$$R = \frac{\text{bitstream length for } \hat{X} \text{ [bits]}}{\text{number of data values represented}} \text{ bpppb.} \tag{2}$$

Similarly, define the *compression ratio* r by

$$r = \frac{\text{raw size of } X \text{ [bits]}}{\text{bitstream length for } \hat{X} \text{ [bits]}}. \tag{3}$$

Hence, the compression ratio is a dimensionless number. Finally, R and r are simply related by

$$R = \frac{B}{r}. \tag{4}$$

1.6 DISTORTION MEASURES

The preceding section discussed the rate R of the (de)compressed data $\hat{X}$. However, $\hat{X}$ only approximates X. As mentioned in the introduction, the *distortion* between X and $\hat{X}$ quantifies the inaccuracy of the approximation. Most lossy compression systems are designed with the goal of minimizing distortion for a given rate.

In lossy compression, by far the most common distortion measure is the *mean square difference* (MSD),

$$\text{MSD}\left(X, \hat{X}\right) = \frac{1}{N_x N_y N_c} \sum_x \sum_y \sum_k \left(X[x,y,k] - \hat{X}[x,y,k]\right)^2. \tag{5}$$

MSD has a natural interpretation as the power of the error $X - \hat{X}$. Additionally, its differentiability facilitates analysis, and it is easy to compute. We also make use of the *root MSD* (RMSD),

$$\text{RMSD}(X, \hat{X}) = \sqrt{\text{MSD}(X, \hat{X})}. \tag{6}$$

Although our lossy compression scheme employs MSD, we can compute additional distortion measures to see how they relate to application performance. We also use *mean absolute difference* (MAD) and *mean spectral angle difference* (MSAD), defined as

$$\text{MAD}\left(X, \hat{X}\right) = \frac{1}{N_x N_y N_c} \sum_x \sum_y \sum_k \left|X[x,y,k] - \hat{X}[x,y,k]\right|, \tag{7}$$

$$\text{MSAD}\left(X, \hat{X}\right) = \frac{1}{N_x N_y} \sum_x \sum_y \theta\left(X[x,y,\vec{k}], \hat{X}[x,y,\vec{k}]\right). \tag{8}$$

2. COMPRESSION SCHEME

The compression scheme that we are evaluating consists of the cascade of spectral compression and spatial compression. Spectral compression is accomplished by applying *principal-components analysis* (PCA) in the spectral dimension. The resulting *principal-component images* (PC images) are then compressed spatially and independently using the JPEG2000 standard. Other researchers [23, 13, 24, 20] have implemented similar schemes. In this section, we elaborate on the main parts of the scheme, which is diagrammed in Figure 3.

2.1 SPECTRAL COMPRESSION WITH PCA

In the signal-processing literature, PCA is commonly referred to as the *Karhunen-Lœve Transform* (KLT). The details of PCA/KLT are well documented in the literature [11], so we only briefly review it. Initially, one computes the spectral mean vector $\bar{X}$ and spectral covariance matrix K_{XX} of X. Next, K_{XX} is factored into the form $K_{XX} = \Phi\Lambda\Phi^{\mathsf{T}}$, where Λ is a diagonal matrix whose diagonal entries are the eigenvalues λ_k of K_{XX} arranged in descending order ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N_c}$), and Φ is a unitary matrix whose columns are the corresponding eigenvectors of K_{XX}. The energy of X is then

$$\text{energy in } X = \sum_{k=1}^{N_c} \lambda_k. \tag{9}$$

Finally, the *transform matrix* Φ^{T} is applied to each spectrum in X to create *a PCA/KLT cube* Y via

$$Y[x, y, \vec{k}] = \Phi^{\mathsf{T}} \left(X[x, y, \vec{k}] - \bar{X} \right). \tag{10}$$

The kth band $Y[\vec{x}, \vec{y}, k]$ is called the kth PC image.

To compress, some of the PC images in Y are first discarded. The retained PC images may then be compressed spatially — the subject of Section 2.2 — and stored or transmitted as a bitstream. $\bar{X}$ and Φ are also stored or transmitted as side information for decompression. We assume these values are noncompressed, B_{side}-bit quantities, so they require $S_{\text{side}} = (N_c + N_c^2)B_{\text{side}}$ bits.

To decompress, the bitstream is first decompressed to obtain approximations of the retained PC images, and any discarded PC images are filled with zeroes to make an approximate PCA/KLT cube $\hat{Y}$. Then, with the aid of the side information $\bar{X}$ and Φ, the decompressed HSI cube $\hat{X}$ is computed via

$$\hat{X}[x, y, \vec{k}] = \Phi\hat{Y}[x, y, \vec{k}] + \bar{X}. \tag{11}$$

PCA/KLT has a number of appealing properties for compression and signal approximation. First, since Φ is unitary, energy is preserved. This property is useful when the MSD distortion measure is used because $\text{MSD}(X, \hat{X}) = \text{MSD}(Y, \hat{Y})$. Second, PCA/KLT decorrelates X spectrally: the spectral covariance matrix $K_{YY} = \Lambda$, so the PC images are uncorrelated, which is convenient if they are to be processed independently. Also, the variance of the kth PC image is $\sigma_k^2 = \lambda_k$. Third, among all unitary transforms,

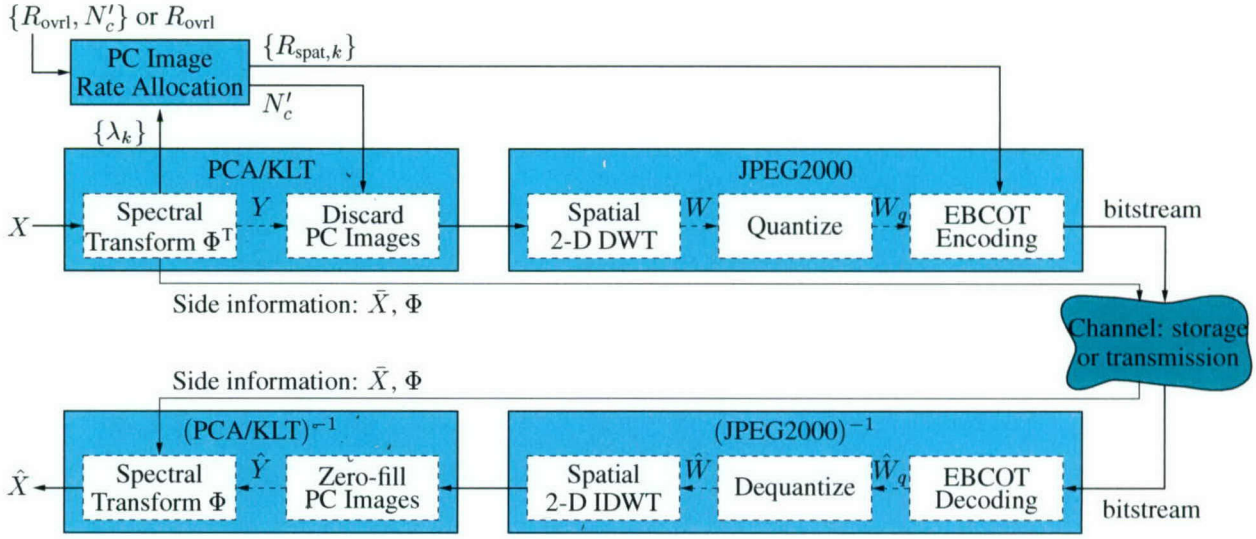


Figure 3. Block diagram of compression and decompression.

PCA/KLT "packs" the maximum amount of the energy of X into the first PC images. The percentage of total energy contained by the kth PC image $Y[\vec{x}, \vec{y}, k]$ is

$$\text{percent of total energy from } k\text{th PC image} = \frac{\lambda_k}{\sum_{\ell=1}^{N_c} \lambda_\ell} \times 100\%. \tag{12}$$

The first k PC images thus account for the following percentage of total energy:

$$\text{percent of total energy from first } k\text{th PC images} = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{\ell=1}^{N_c} \lambda_\ell} \times 100\%. \tag{13}$$

For many kinds of multidimensional data, the vast majority (e.g., $> 99\%$) of the energy is contained in the first few PC images. Figure 6 in Section 4.1 shows the eigenvalues and the individual and cumulative percentages of the total energy for an example HSI cube. Fourth, for an N_c-dimensional Gaussian source and MSD distortion, PCA/KLT is the theoretically optimum transform in a rate-distortion sense.

Despite the above properties of PCA/KLT, we would be remiss not to mention some of its potential drawbacks [30]. First, PCA/KLT is data-dependent, so Φ will change from one cube to another, and $\bar{X}$ and Φ must be stored or transmitted as side information. Second, Φ^{T} does not have any special structure, so the computational complexity of PCA/KLT is $O(N_c^2)$. In contrast, other data-independent, fast signal transformations (e.g., discrete cosine or wavelet transform) require no side information and have complexities of $O(N_c \log N_c)$. Third, most of the optimality properties of PCA/KLT are associated with Gaussian data and MSD distortion. However, HSI data do not usually have a Gaussian distribution [19, 29], and MSD may be inappropriate for HSI applications such as classification and identification.

Nevertheless, PCA/KLT remains an extremely popular and effective tool in the remote-sensing and lossy-compression communities. MSD does offer an analytically tractable distortion measure, and it is reasonable to expect acceptable application performance if $\hat{X}$ remains faithful to X. Also, the energy-preserving, energy-packing, and decorrelation properties always hold, and PCA/KLT often provides good

compression performance for non-Gaussian data. PCA/KLT forms a crucial part of many HSI compression schemes [13, 20, 23, 24, 26], including ours.

2.2 SPATIAL COMPRESSION WITH JPEG2000

Because PCA/KLT is performed only in the spectral direction, the PC images are spectrally uncorrelated, but each individual PC image $Y[\vec{x}, \vec{y}, k]$ remains spatially correlated. Spatial compression is therefore applied to each individual PC image [1]; our compression scheme employs JPEG2000, a state-of-the-art, wavelet-based image-compression standard [27]. JPEG2000 possesses a number of important properties. First and foremost, it has significantly better rate-distortion performance than many other, older image-compression methods, including JPEG. Additional features include progressive transmission, random-access and region-of-interest coding, and error resilience. The details of the standard go far beyond the scope of this paper, so we only describe the basic operations here.

Given the kth input PC image $Y[\vec{x}, \vec{y}, k]$, the JPEG2000 compression scheme computes a two-dimensional, separable discrete wavelet transform (DWT) $W[\vec{x}, \vec{y}, k]$ of $Y[\vec{x}, \vec{y}, k]$; our implementation uses the irreversible 9-7 wavelet. Next, the wavelet coefficients in $W[\vec{x}, \vec{y}, k]$ are quantized to produce the quantized wavelet coefficients $W_q[\vec{x}, \vec{y}, k]$. Each subband of $W_q[\vec{x}, \vec{y}, k]$ is then partitioned into rectangular code-blocks, and each code-block is coded independently using arithmetic encoding in conjunction with an operational rate-distortion optimization algorithm known as EBCOT (embedded block coding with optimized truncation) [27].

In decompression, arithmetic decoding and dequantization of the code-blocks give the decompressed, quantized wavelet coefficients $\hat{W}_q[\vec{x}, \vec{y}, k]$. Then the inverse DWT (IDWT) is applied to produce the approximated PC image $\hat{Y}[\vec{x}, \vec{y}, k]$.

2.3 RATE ALLOCATION AMONG PC IMAGES

To this point, we have described the two main compression steps: spectral compression via PCA/KLT and spatial compression of the individual PC images with JPEG2000. Each PC image $Y[\vec{x}, \vec{y}, k]$ can be assigned a different spatial rate $R_{\text{spat},k}$, and the EBCOT algorithm ensures that it will be spatially compressed to minimize $\text{MSD}(Y[\vec{x}, \vec{y}, k], \hat{Y}[\vec{x}, \vec{y}, k])$ for the given value of $R_{\text{spat},k}$. Thus, if the *overall rate* R_{ovrl} is given,[2] we must still choose the spatial rates $\{R_{\text{spat},k}\}$; this is the problem of *rate allocation*.

Equal-Rate Allocation The first method uses *equal-rate allocation*. Given R_{ovrl}, one simply chooses N_c' ($N_c' \leq N_c$), the number of PC images to retain, and sets $R_{\text{spat},k}$ to be the same for all of them. Because of the energy-packing property of PCA/KLT, the first PC images contain most of the energy in X, so we keep PC images 1, 2, ..., N_c' and discard the remaining ones. The spectral compression ratio is simply

$$r_{\text{spec}} = \frac{N_c}{N_c'}. \tag{14}$$

[1] The PC images are spectrally uncorrelated but not necessarily spectrally *independent*; however, as in many separable compression schemes, any remaining dependence is typically assumed to be weak and is ignored.

[2] Recall that the equivalent overall compression ratio is $r_{\text{ovrl}} = B/R_{\text{ovrl}}$.

It follows that the retained PC images all have the same spatial rate,

$$R_{\text{spat},k} = R_{\text{ovrl}} \frac{N_c}{N_c'}, \quad k = 1, 2, \ldots, N_c'. \tag{15}$$

There is nothing optimal about this method, but it is a simple approach that we use mainly as a basis for comparison with the allocation method below.

MSD-Optimal Allocation The equal-rate allocation uses the same rate $R_{\text{spat},k}$ for all retained PC images. However, the variances of the PC images can differ widely. For example, the first PC image might contain 80% of the energy, and the fifth only 1–2%; less distortion should result by making $R_{\text{spat},1}$ greater than $R_{\text{spat},5}$. This section reviews a common procedure for selecting $\{R_{\text{spat},k}\}$ for *MSD-optimal allocation.*

The distortion-rate function $D(R)$ of information theory [5] gives the theoretical lower bound D on the resulting distortion when data produced by a source are compressed at a rate R. For N_c independent sources with respective distortion-rate functions $D_k(R_k)$, the problem of rate allocation [27, 30] is to select $\{R_k\}$ to minimize the overall distortion $(1/N_c)\sum_k D_k(R_k)$ subject to the rate constraint $(1/N_c)\sum_k R_k = R_{\text{ovrl}}$. Applying the Lagrange multiplier method to solve this problem yields the equal-slope or Pareto condition $\partial D_k / \partial R_k = -\eta^*, \forall k$, where the optimum Lagrange multiplier value $\eta^* > 0$ is chosen to satisfy the rate constraint R_{ovrl}.

For the purpose of rate allocation in this work, we use MSD distortion and assume each spectrum $X[x, y, \vec{k}]$ is the output of a vector Gaussian source with mean vector $\bar{X}$ and covariance matrix K_{XX}. Since PCA/KLT is a unitary transform, the MSD in the PCA/KLT domain equals the MSD in the usual HSI domain, and the PC images $Y[\vec{x}, \vec{y}, k]$ represent outputs from N_c independent Gaussian sources with respective variances $\sigma_k^2 = \lambda_k$. With this assumption, rate allocation can be accomplished in a straightforward manner using a bisection search. A review of the rate allocation problem is given in Appendix A.

The assumption of Gaussian data is, of course, a crude approximation. Better allocation results could be achieved by modeling the distributions of the PC images more accurately. Alternatively, one could implement an operational allocation method, which compresses and decompresses each PC image over a range of rates, tabulates the resulting PCA/KLT-domain MSDs, and then performs discrete optimization. Such methods require a substantial increase in complexity and were therefore not used in this study.

3. DERIVED DATA PRODUCTS

Remotely sensed HSI data can be used to characterize the environment, both on the ground and in the atmosphere. In this section, we review two data products derived from HSI data.

3.1 NORMALIZED DIFFERENCE VEGETATION INDEX

The first data product is the *normalized difference vegetation index* (NDVI). NDVI measures spectral differences around the red edge (red and NIR wavelengths) and is commonly used to represent the health and amount of vegetation [28]. The NDVI is defined as

$$\text{NDVI} = \frac{\rho(\text{NIR}) - \rho(\text{red})}{\rho(\text{NIR}) + \rho(\text{red})}, \tag{16}$$

where $\rho(\text{NIR})$ and $\rho(\text{red})$ represent near-infrared and red band reflectances, respectively.

The rationale behind (16) can be understood as follows. The chlorophyll present in healthy vegetation absorbs visible light in the red wavelengths and reflects infrared radiation. Consequently, the surface spectra of healthy and/or dense vegetation generally exhibits a decrease in the radiance at red wavelengths and a large increase at near infrared (NIR) wavelengths. The spectra of other land-cover types may have high values in the NIR or low values in the red, but vegetation tends to be unique in having this combination. The numerator in (16) follows. The denominator in the equation normalizes for factors such as slope and changes in illumination [17].

NDVI takes on values between -1 and $+1$. Larger positive values represent increasingly healthy or dense vegetation. Values near zero indicate rock or bare soil, and negative values are associated with water, snow and ice, or barren terrain. NDVI is employed to distinguish vegetation from other land cover types and to assess the density, health, or stress of vegetation. It also finds use in the prediction of droughts and identification of areas where the risk of fire is high.

In this work, the red band is taken as the average intensity from 626.19 nm to 692.88 nm, and the near-infrared band from 779.14 nm to 807.91 nm. These bands correspond to the spectral resolution of LANDSAT bands 3 and 4, respectively. Additionally, in this report, NDVI was computed using radiance rather than reflectance values. Since the atmospheric path radiance is low in both the red and NIR bands, atmospheric correction was deemed unnecessary [25, Ch. 5].

3.2 COLUMN WATER VAPOR

The second data product considered is integrated *column water vapor* (CWV). CWV measures the water vapor content in a vertical path from the Earth's surface to the sensor. CWV is expressed in units of grams per centimeters or more often just centimeters of water vapor [7]. CWV is commonly used in the atmospheric compensation process to define the atmospheric transmission due to water vapor [9].

Following [7], the at-sensor radiance for a down-looking airborne sensor can be expressed as

$$L_{sensor}(\lambda) = L_{sun}(\lambda)\tau(\lambda)\rho(\lambda) + L_{path}(\lambda), \tag{17}$$

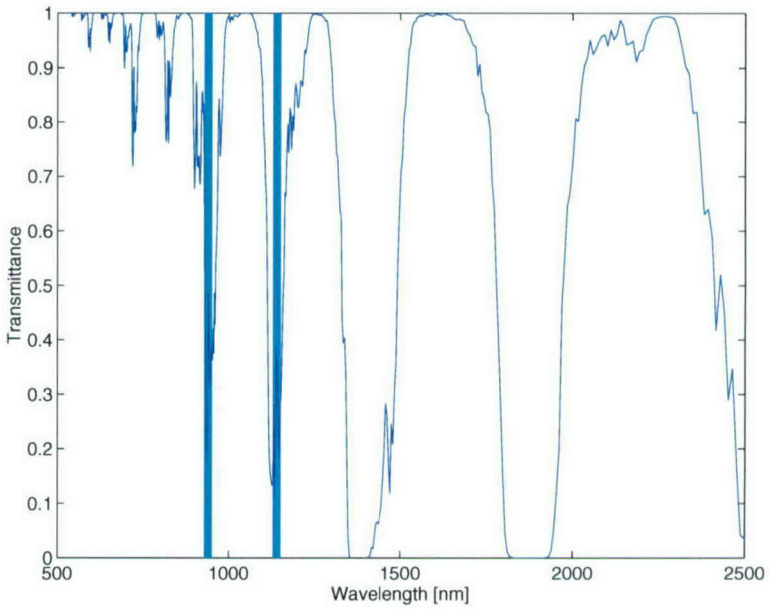


Figure 4. Atmospheric transmittance of water vapor for nominal atmospheric conditions.

where $L_{sun}(\lambda)$ is the solar radiance, $\tau(\lambda)$ ($0 \leq \tau(\lambda) \leq 1$) is the total atmospheric transmittance from the sun to the surface to the sensor, $\rho(\lambda)$ is the surface reflectance, and $L_{path}(\lambda)$ is the path-scattered radiance.

Figure 4 displays the atmospheric transmittance due to water vapor across the VNIR/SWIR for nominal atmospheric conditions. Water vapor in the atmosphere absorbs solar and surface-reflected radiance at wavelengths near 720, 820, 910, 940, 1140, 1380, 1880, and 2180 nm [7]. Around these wavelengths, the atmospheric transmittance $\tau(\lambda)$ displays sharp troughs. The wavelengths associated with these troughs are called *water vapor absorption bands*. The figure shows two absorption bands — at 940 and 1140 nm — in light blue. Typically, $\tau(\lambda)$ is greater than 0.90 or even 0.99 outside of the absorption bands, while it drops sharply to near 0.50, 0.20, or near zero in the absorption bands. The spectrally flat, high-transmittance bands adjacent to an absorption band are known as *atmospheric windows*.

The surface reflectance $\rho(\lambda)$ of natural materials generally varies smoothly in the VNIR region. In particular, $\rho(\lambda)$ is approximately linear across the water vapor absorption bands. If the transmittance were not attenuated in an absorption band, then the at-sensor radiance $L_{sensor}(\lambda)$ would be approximately linear between the atmospheric windows associated with the absorption band.

CWV is estimated based upon the above observations. A number of methods for estimating CWV exist [7, 8, 9, 14, 15], so we describe the general ideas only. First, radiance values in the atmospheric windows and those in the absorption bands are used to estimate a pseudo-transmittance $\hat{\tau}(\lambda)$ in the absorption bands. This step exploits the linearity of $\rho(\lambda)$ described above. Second, $\hat{\tau}(\lambda)$ is compared with transmittances derived from radiative transfer models with differing CWV amounts. The estimated CWV is then extracted from the model whose transmittances agree most closely with $\hat{\tau}(\lambda)$.

Some methods [7, 14] employ the absorption bands at 940 and 1140 nm, while others [9] use an additional absorption band at 910 nm. The other absorption features are usually not exploited, either because

they do not attenuate $\tau(\lambda)$ sufficiently or because their attenuation of $\tau(\lambda)$ does not vary much with CWV. This work used the method described by Griffin and Burke [9].

4. EFFECTS ON DERIVED DATA PRODUCTS

Experiments were performed on an HSI cube from AVIRIS (Airborne Visible Infrared Imaging Spectrometer). AVIRIS is a whiskbroom HSI sensor with 614 pixels in the cross-track direction and spectral coverage of 400–2500 nm in 224 bands with a nominal spectral resolution of 10 nm/band. The data was collected from an airborne platform at 20 km altitude over Moffett Field, located near the San Francisco Bay in California. The ground sample distance for this data is approximately 20 m. The original radiance cube has dimensions of $614 \times 512 \times 224$, with each radiance value stored as a 16-bit integer. The original cube, designated by X, thus occupies 134 MB. Figure 5 shows an RGB image of the original scene.

X was compressed using the method in Section 2 and a number of different compression settings. The PCA/KLT side information ($\bar{X}$ and Φ) required $B_{\text{side}} = 32$ bits per value, so $S_{\text{side}} = 1612800$ bytes. For the equal-rate allocation method, N_c' was set to 5, 10, or 20, and overall compression ratios r_{ovrl} of 50, 100, or 200 were chosen. The MSD-optimal allocation method also used $r_{\text{ovrl}} = 50$, 100, and 200.

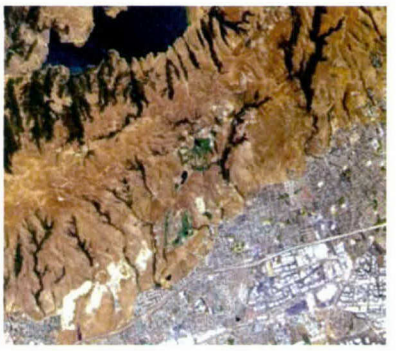


Figure 5. RGB image of Moffett Field AVIRIS cube.

4.1 COMPARISON OF RATE ALLOCATION METHODS

This section discusses results that illustrate the differences between the two rate-allocation methods. First, the upper graph in Figure 6 shows the eigenvalues λ_k associated with the PC images $Y[\vec{x}, \vec{y}, k]$ after PCA is performed on X. As is characteristic of PCA, the eigenvalues appear in descending order, which shows the well-known energy-packing property of PCA. The lower graph in the figure shows the percentage of the total variance of X formed by the PC images. The percentages for only the first 30 PC images are shown. The blue curve shows the individual percentage for each PC image (cf. (12)), and the red curve shows the cumulative percentage (cf. (13)). The first PC image $Y[\vec{x}, \vec{y}, 1]$ accounts for about 60 percent of the total variance in X, the second PC image $Y[\vec{x}, \vec{y}, 2]$ for about 30 percent, and so forth. From the red curve, the first five PC images account for over 99% of the total variance in X.

Figure 7 contains three graphs, each showing the resulting spatial compression ratios for a different overall compression ratio r_{ovrl}. For the equal-rate method, the graphs show essentially horizontal lines since each retained PC image is compressed at the same spatial ratio (cf. (15)). For a given overall compression

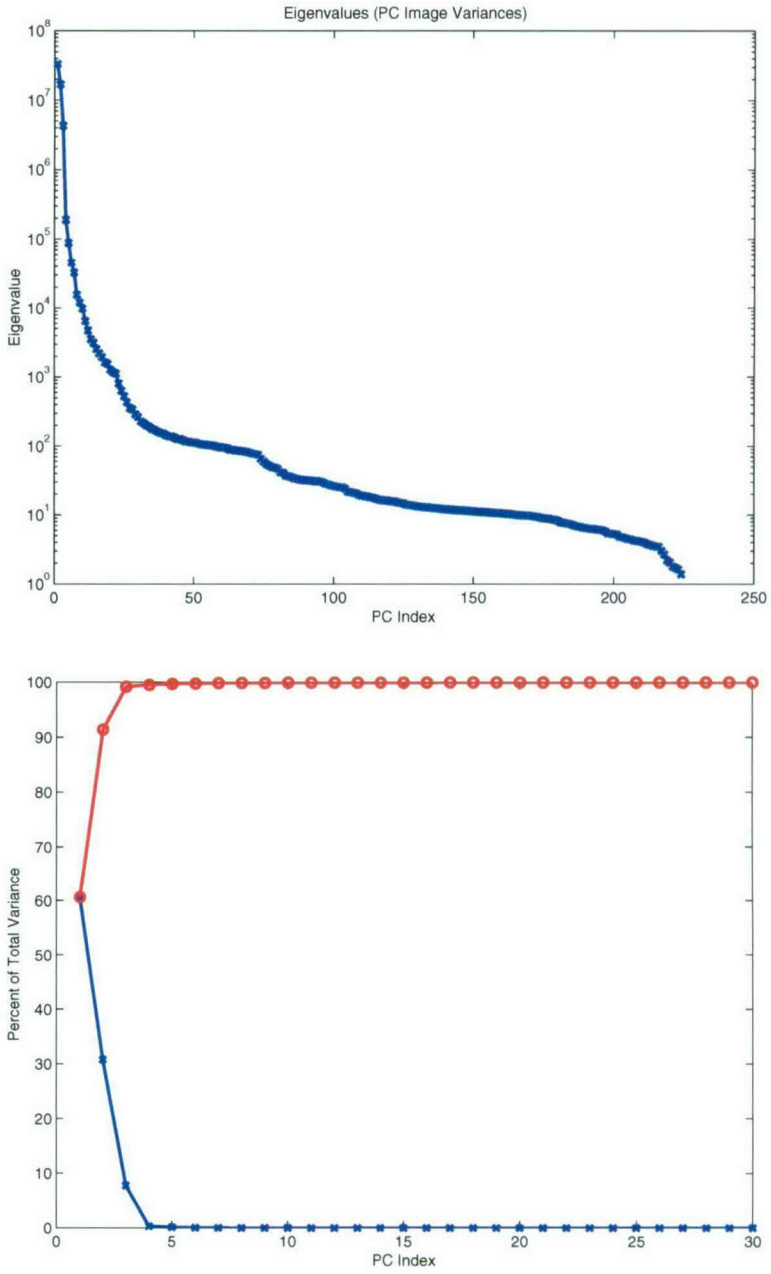


Figure 6. Eigenvalues and variance percentage for Moffett Field.

ratio, retaining more PC images means that the retained PC images must be compressed more.

However, for the MSD-optimal rate allocation method, each graph shows a monotonically increasing curve. In general, the MSD-optimal allocation method compresses the first 2–3 PC images less than the equal-rate method. The first PC image is spatially compressed the least, the second PC image is compressed slightly more, and so forth. The trend of increasing spatial compression ratios agrees with the fact that higher-index PC images contain a smaller portion of the total variance of X (cf. Figure 6). Compared to the lower-index PC images, they can be compressed at higher spatial compression ratios without affecting the overall MSD as much.

Next, Figure 8 shows two different distortion measures, RMSD$(X,\hat{X})$ and MAD$(X,\hat{X})$, plotted against the overall compression ratio. The connected curves show the trend for a given rate allocation method. The three unconnected points show the distortion and effective overall compression ratio when only PCA is used, and the retained PC images are not compressed spatially. Under both RMSD and MAD, the MSD-optimal rate allocation method performs best. Compared to using PCA only (no spatial compression), the MSD-optimal method yields overall compression ratios about five times higher. Compared to the equal-rate allocation method, the MSD-optimal method produces a much lower distortion at the same overall compression ratio.

4.2 SPECTRAL DISTORTION

Figure 9 shows the average power in X versus wavelength. There are four distinct water vapor absorption bands around 940, 1100, 1370, and 1850 nm. The green bars indicate the wavelengths used in the computation of NDVI (Section 3.1), and the light blue bars indicate the wavelengths used in the computation of CWV (Section 3.2).

Finally, Figure 10 shows plots of the ratio of the average power to the average MSD versus wavelength for different overall compression ratios and allocation methods. Not surprisingly, as the compression ratio increases from 50 to 200, the power-to-MSD ratios decrease across the entire spectrum. Consistent with Figure 8, the MSD-optimal method has the highest power-to-MSD ratio in almost every band and in all three graphs. At the NDVI wavelengths, the power-to-MSD ratio for the MSD-optimal method is several decibels higher than that for the equal-rate allocation methods. At the CWV wavelengths, the MSD-optimal method has the highest power-to-MSD ratio when $r_{\text{ovrl}} = 50$. However, when $r_{\text{ovrl}} = 100$, this ratio becomes comparable to that for the equal-rate method with $N_c' = 10$, and with r_{ovrl} reaches 200, this ratio falls several decibels below the ratio for the equal-rate method with $N_c' = 10$. In addition, in the remaining absorption bands (near 1370 and 1850 nm), all of the allocation methods produced comparable power-to-MSD ratios.

These observations can be understood by considering the graph in Figure 9 and the allocation methods. As a result of PCA, the errors due to lossy compression are distributed across all bands. The absorption bands — particularly those around 1370 and 1850 nm — already have low power, so they show correspondingly lower power-to-MSD ratios than other wavelengths. With regard to the allocation methods, the MSD-optimal method attempts to minimize the MSD; it does not consider the power-to-MSD ratio in any particular band. The equal-rate method does not perform any optimization at all. If the MSD is roughly evenly distributed among all bands, the low-power bands will suffer more than high-power bands in terms of their power-to-MSD ratios.

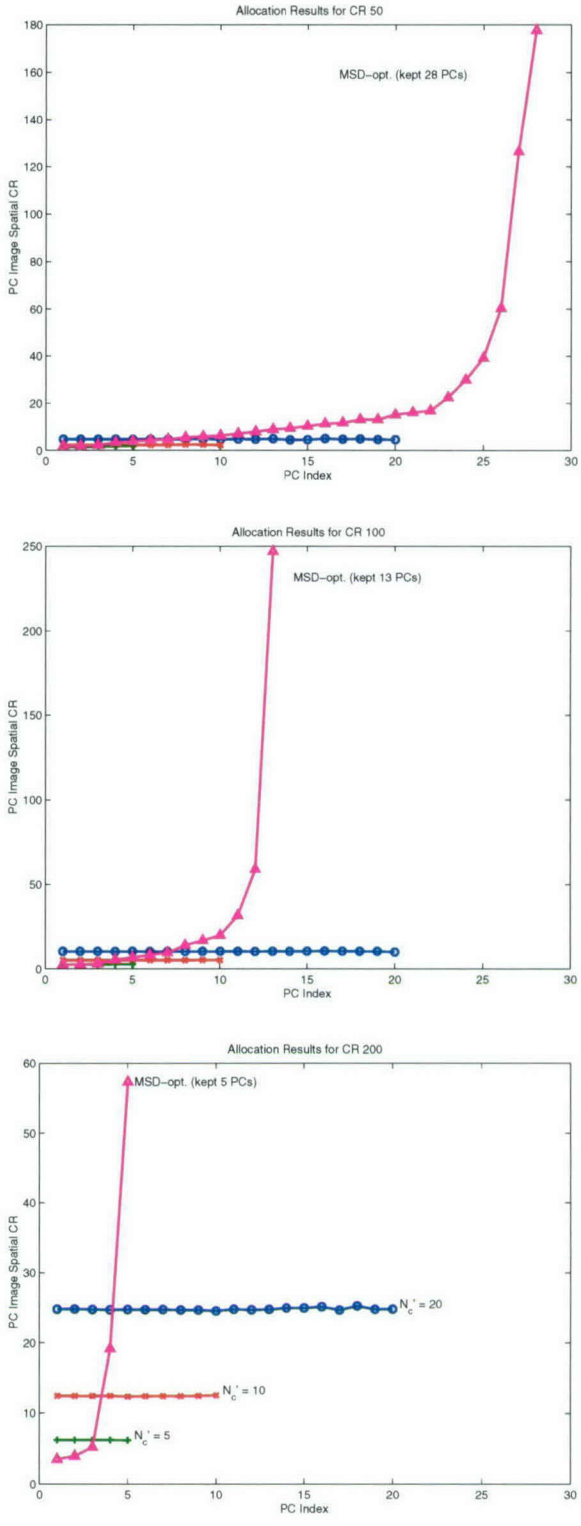


Figure 7. Allocation results for Moffett Field.

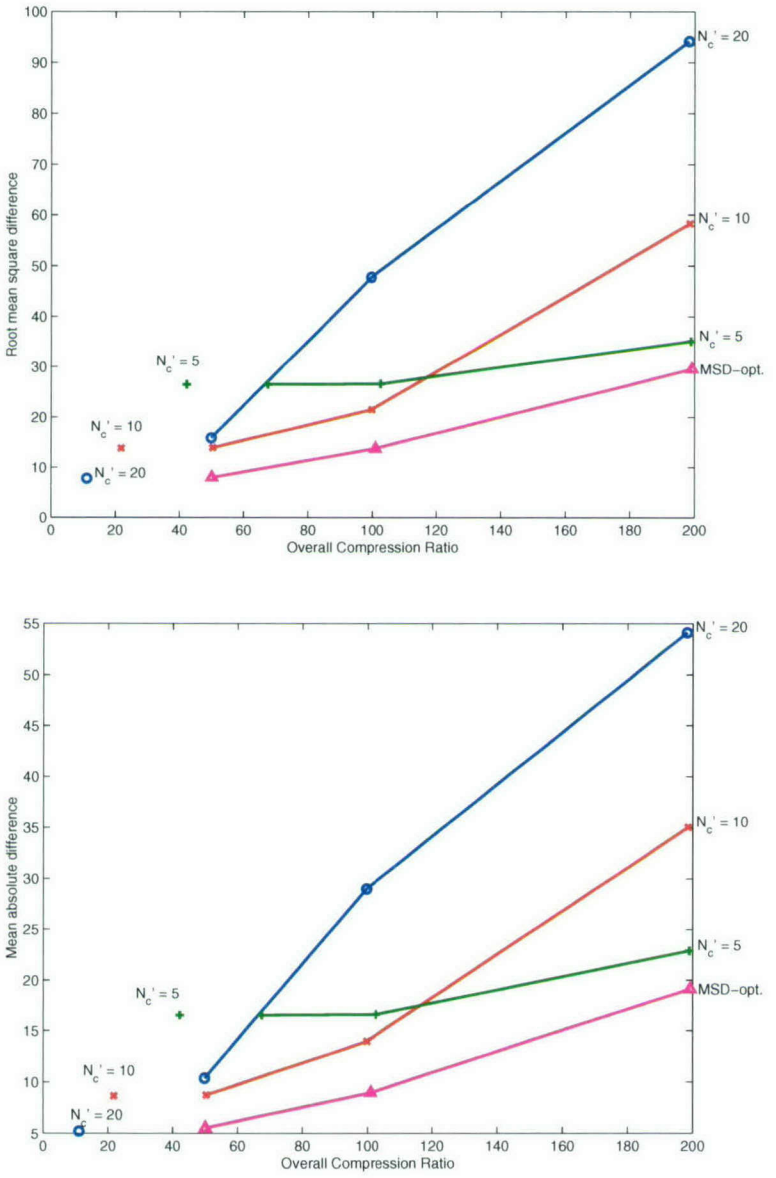


Figure 8. Distortion vs. compression ratio for Moffett Field.

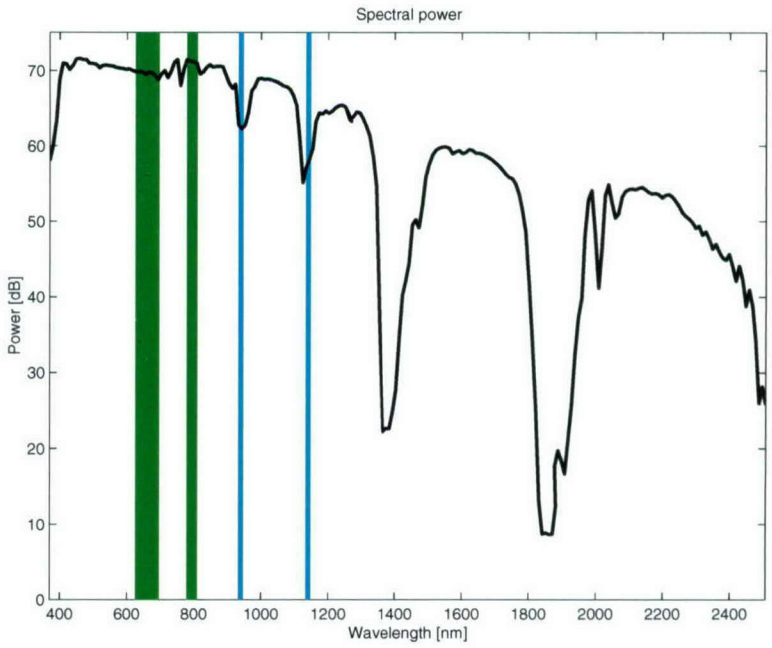


Figure 9. Power vs. wavelength for Moffett Field.

4.3 EFFECTS ON NORMALIZED DIFFERENCE VEGETATION INDEX

Figure 11 shows the NDVI image resulting from the original, noncompressed HSI data cube. The white-to-green color scale shows more positive values of NDVI in deeper shades of green and smaller or negative values in white.

NDVI difference images with and without compression are shown in Figure 12. Each column compares the difference images for different rate allocation methods. From top to bottom, the methods are: equal-rate with $N_c' = 5$, equal-rate with $N_c' = 10$, equal-rate with $N_c' = 20$, and MSD-optimal. Across each row, the desired overall compression ratio increases from 50 to 100 and then to 200. The color bar shows that differences closer to zero are white, with increasing positive and negative differences shown as deeper shades of red and blue, respectively.

The figures show that the equal-rate method with $N_c' = 5$ has the greatest NDVI differences across all compression ratios, while the MSD-optimal method has the smallest differences at each compression ratio. In addition, the MSD-optimal method with a compression ratio of 100 has differences comparable to the equal-rate method with $N_c' = 10$ and a compression ratio of 50.

Figure 13 shows a graph with the MAD between the NDVI image from the noncompressed HSI data and the NDVI image after compression for a variety of compression settings. In addition, the unconnected points show performance for PCA only and no spatial compression. The MSD-optimal rate allocation method outperforms the other allocation methods. In agreement with the previous figure, the MSD-optimal method with $r_{\text{ovrl}} = 100$ has MAD only slightly higher than the equal-rate method with $r_{\text{ovrl}} = 50$. Referring back to Figure 10, we see that the MSD-optimal method maintains a higher power-to-MSD ratio in the bands

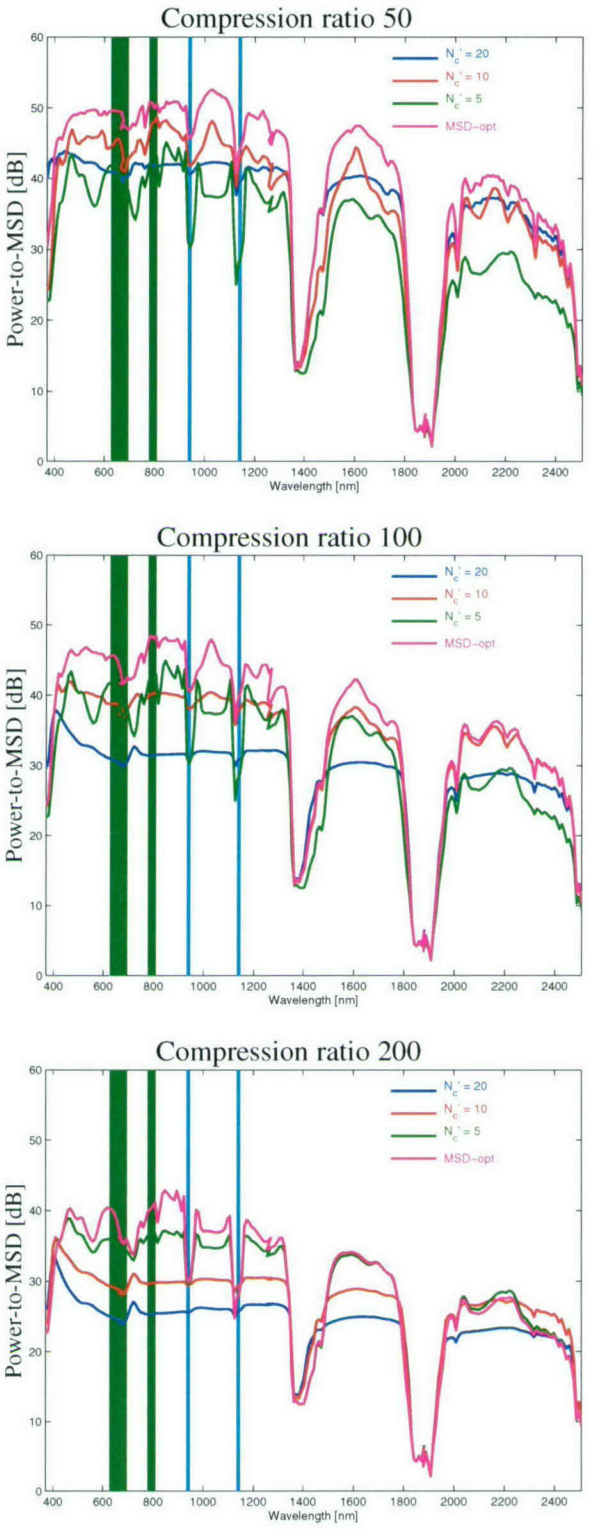


Figure 10. Power-to-MSD ratio vs. wavelength for Moffett Field.

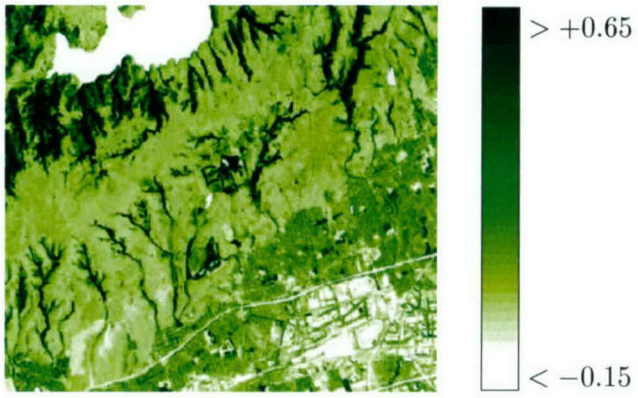


Figure 11. NDVI image for Moffett Field.

used for computing NDVI.

4.4 EFFECTS ON INTEGRATED COLUMN WATER VAPOR

The next set of results examines the effects of compression on the CWV product. The CWV image for the noncompressed HSI data appears in Figure 14, where the blue color scale shows higher CWV values as brighter shades of blue. Since the CWV calculation is not reliable over bodies of water, they have been blocked out in the image.

Figure 15 shows CWV difference images. The top row, which shows results for the equal-rate allocation method and $N_c' = 5$, indicates that retaining only 5 PC images is likely insufficient regardless of how much or how little they are compressed spatially. Results with the equal-rate method are better when 10 or 20 PC images (second and third rows) are retained. The MSD-optimal method performs best when $r_{\text{ovrl}} = 50$; at this compression ratio, it retained 25 PC images. At $r_{\text{ovrl}} = 100$, the MSD-optimal method performs comparably to the equal-rate method with $N_c' = 10$ and slightly better than the equal-rate method with $N_c' = 20$. Finally, at $r_{\text{ovrl}} = 200$, the equal-rate method with $N_c' = 10$ has the best performance, and the MSD-optimal method retained only 5 PC images and so its CWV difference image looks very similar to those in the top row.

The MAD between the CWV image from the noncompressed HSI data and the CWV images after compression is graphed in Figure 16. The individual points for PCA only with no spatial compression show small MAD values when either 10 or 20 PC images are retained but a sharp increase in MAD when only 5 PC images are retained. The connected curves show results when spatial compression and the different rate allocation methods are applied. With an overall compression ratio of fifty, the MSD-optimal method and equal-rate method with $N_c' = 10$ or $N_c' = 20$ perform comparably. As the overall compression ratio increases, the MSD-optimal method begins to perform worse than the equal-rate methods.

Recall from Figure 7 that for overall compression ratios of 50, 100, and 200, the MSD-optimal method retained 28, 13, and 5 PC images, respectively. Hence, at a compression ratio of 200, the MSD-optimal method cannot perform any better than only using PCA and retaining the first 5 PC images. In addition, Figure 10 shows that in the bands used to compute CWV, the power-to-MSD ratio for the MSD-optimal

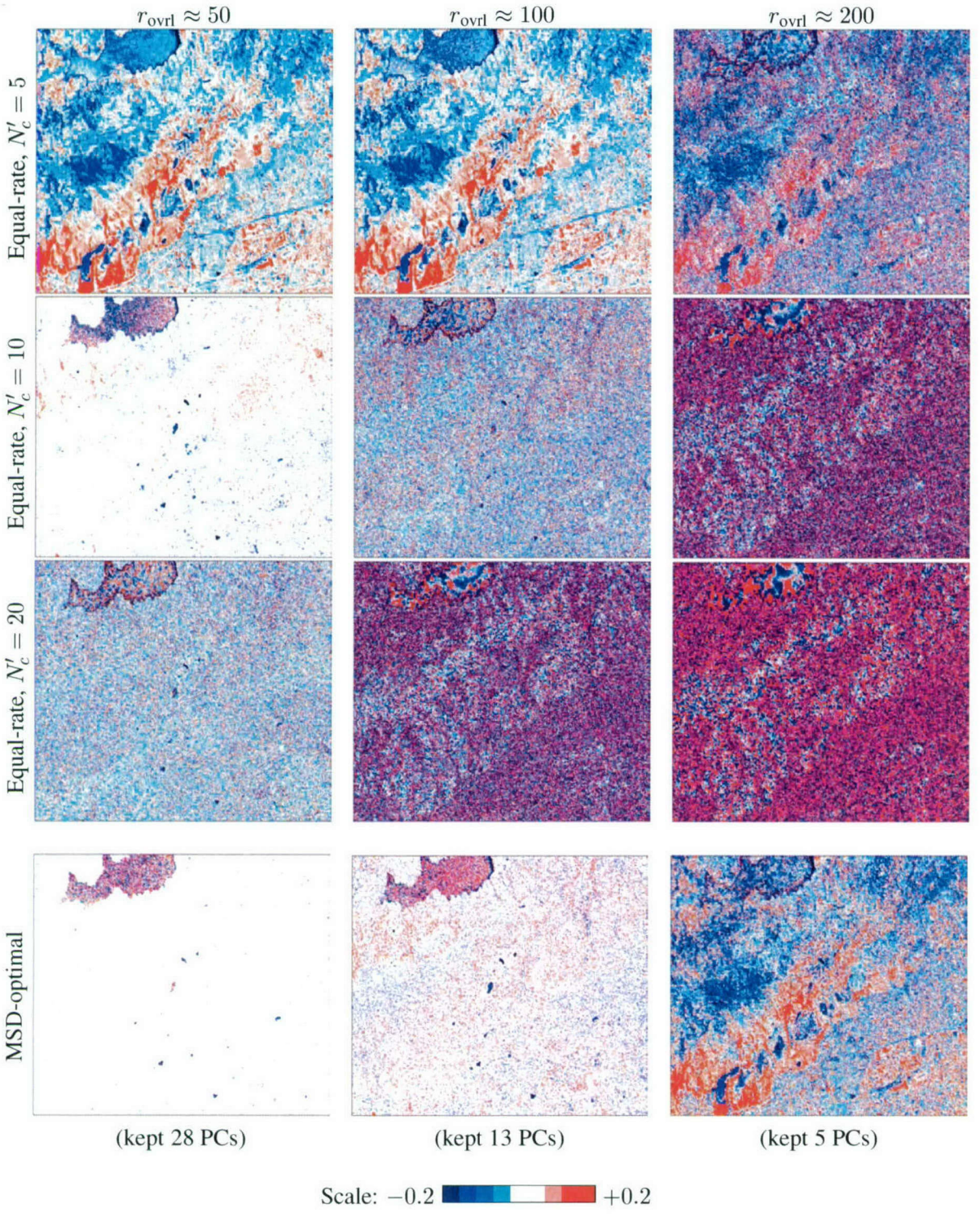


Figure 12. NDVI difference images for Moffett Field.

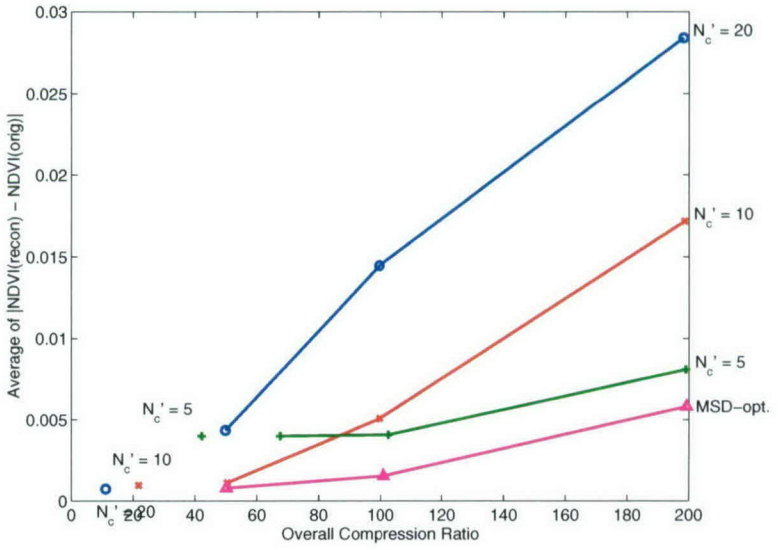


Figure 13. NDVI difference vs. compression ratio for Moffett Field.

method drops below that for the equal-rate method with $N_c' = 10$ and $N_c' = 20$.

4.5 DISCUSSION

These results indicate that the example data products (NDVI and CWV) can tolerate a certain amount of lossy compression. They also show that, compared to using PCA alone without any spatial compression, PCA and spatial compression of the retained PC images can offer significant gains in compression without degrading results.

In addition, the importance of proper rate allocation among the PC images is evident. In general, the MSD-optimal rate allocation method produced better results than the equal-rate allocation method. With

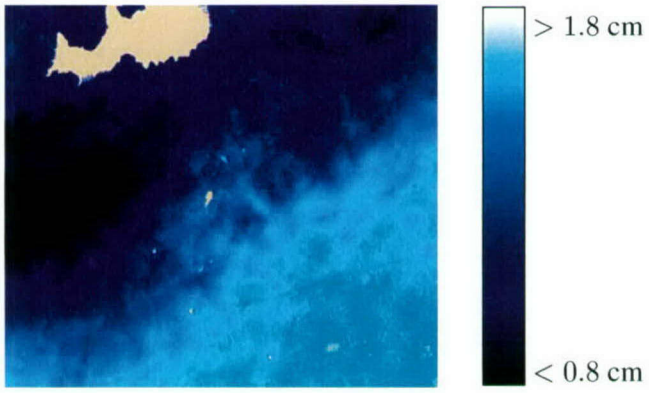


Figure 14. CWV image for Moffett Field.

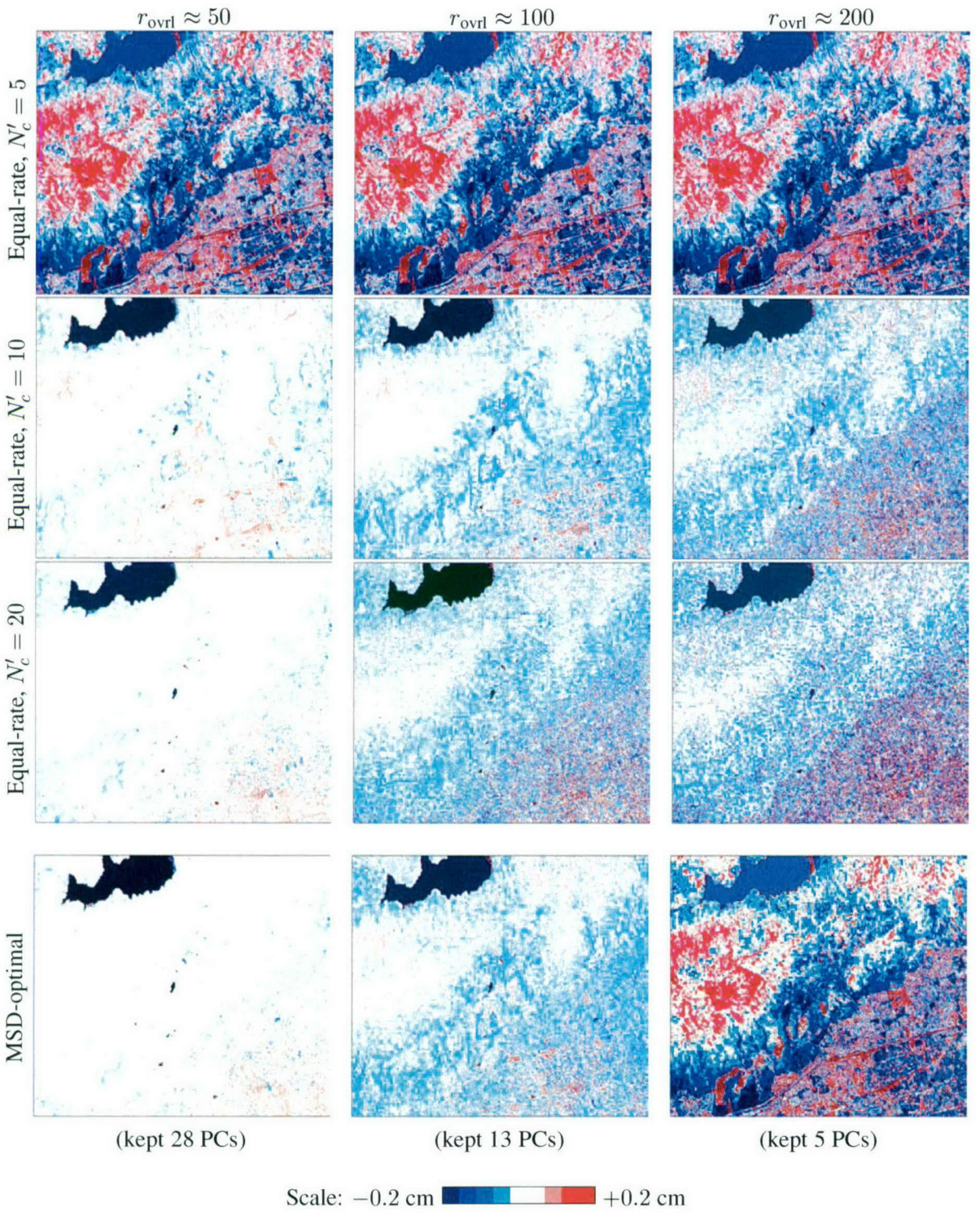


Figure 15. CWV difference images for Moffett Field.

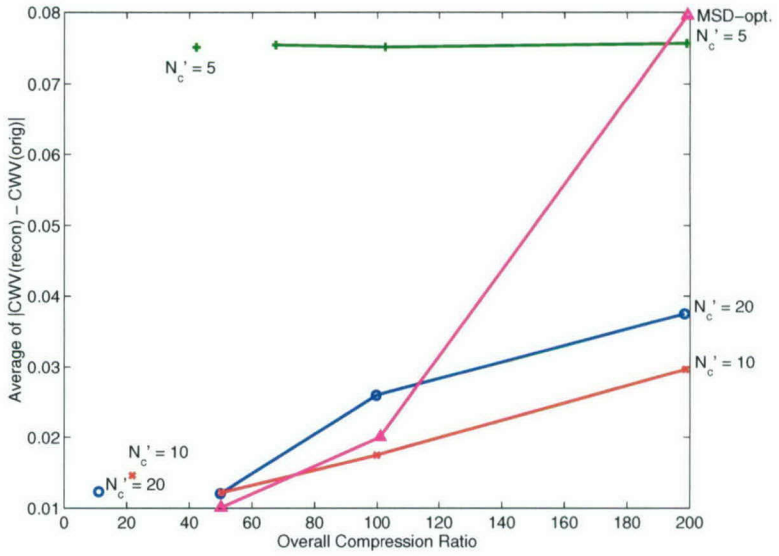


Figure 16. CWV difference vs. compression ratio for Moffett Field.

regard to distortion measures like RMSD and MAD, it yielded consistently lower distortions than the equal-rate method. With NDVI, the MSD-optimal method also had the best performance. However, with CWV, it did not always perform best.

This behavior is consistent with the fact that MSD distortion is an average measure and therefore does not consider the distortion in a particular spectral band. As the desired compression ratio increases, the MSD-optimal allocation method retains fewer PC images and spatially compresses them more, and PCA/KLT distributes the distortion across all bands. NDVI uses bands that have high power, and therefore can tolerate some distortion before the power-to-MSD ratio becomes too low and the NDVI data product becomes degraded. In contrast, CWV employs water-vapor absorption bands, which have somewhat lower power.

The equal-rate methods perform no optimization in their selection of N_c' or the spatial compression ratio for each retained PC image. Provided that r_{ovrl} and N_c' are large enough, then the distortion in the absorption bands remains small enough so that the CWV data product does not become unacceptably degraded. Note, however, that there is no way to predict what r_{ovrl} and N_c' should be.

5. CLASSIFICATION SCHEMES

The previous section considered a number of data products derived from HSI data. Another important application of HSI data is classification. The data products considered above yield real-valued images of NDVI or CWV, so it is straightforward to examine the error between the images with and without lossy compression. However, classification assigns a discrete label to each pixel, so some care is required when evaluating the effects of compression. The next few sections of this report review some basic classification schemes, discuss the evaluation of classifier performance, and then study how compression affects classification results.

Given an $N_x \times N_y \times N_c$ HSI cube Z, *classification* is the process of assigning each pixel $Z[x, y, \vec{k}]$ to one of L possible *classes*. The L classes must be mutually exclusive and totally exhaustive to ensure that each pixel is assigned to one and only one class [4]. Each class is a collection of pixels, which need not be spatially connected but are deemed to be alike in some way. A class might be labelled with an associated high-level meaning such as "road," "grass," or "forest." Given the variability of natural materials and imaging conditions, such high-level designations may require the expertise of an image analyst. With machine classification, a class usually consists of pixel spectra that are close together in some quantifiable sense. In any case, classification reduces the multidimensional, variable details of a pixel spectrum to a single, discrete designation. The result of classification is an $N_x \times N_y$ integer-valued *class image* C, where $C[x, y] = \ell \in \{1, 2, \ldots, L\}$ indicates that the pixel $Z[x, y, \vec{k}]$ has been assigned to class ℓ.

5.1 CLASSIFICATION ALGORITHMS

In this work, we consider two simple but popular supervised machine classification algorithms: the *Euclidean minimum distance* (EMD) classifier, and the *spectral angle mapper* (SAM). The algorithms are supervised in the sense that a user must provide them with a *spectral library*, a set $\mathcal{L} = \{\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_L\}$ of L distinct spectra representing the different classes.

Central to both classifiers is a distance measure $d(\cdot, \cdot)$ that quantifies how close a spectrum $Z[x, y, \vec{k}]$ is to a library spectrum $\vec{s}_\ell \in \mathcal{L}$. EMD uses Euclidean distance,

$$d(\vec{u}, \vec{v}) = \|\vec{u} - \vec{v}\|.$$

SAM employs spectral angle,

$$d(\vec{u}, \vec{v}) = \theta(\vec{u}, \vec{v}).$$

Given the cube Z and spectral library $\mathcal{L}$, the classifier operates as follows: For each spectrum $Z[x, y, \vec{k}]$, the classifier finds

$$\ell^* = \arg\min_{\ell} d(Z[x, y, \vec{k}], \vec{s}_\ell).$$

That is, ℓ^* is the index of the closest library spectrum. Then the classifier sets $C[x, y] = \ell^*$ to make the class assignment. Note that both algorithms operate on individual pixel spectra only; they do not consider spatial relationships.

In some versions of these classification algorithms, the distance $d(Z[x, y, \vec{k}], \vec{s}_{\ell^*})$ must also be below some threshold; otherwise, the pixel is labelled as unclassified ($C[x, y] = 0$). This option was not exercised for the work in this report.

These classifiers are very simple but nonetheless worth investigating. Euclidean distance corresponds directly to MSD distortion, so we might expect a simple relationship between MSD and compression-induced changes in EMD classification. In contrast, the spectral angle between two spectra can be very small even if the Euclidean distance between them is large. For example, consider $\vec{v} = a\vec{u}$ with and $a \gg 1$; then $\|\vec{u} - \vec{v}\| = |a - 1|\|\vec{u}\|$, but $\theta(\vec{u}, \vec{v}) = 0$. Consequently, SAM provides a case where the distortion measure used for compression does not relate to the distance measure used for classification.

5.2 CONFUSION MATRIX

Classification is a hard-decision process; each pixel must be assigned to exactly one class. Pixels will occasionally be assigned to the wrong class for a number of reasons. Natural materials possess inherent variability. Also, the spatial area being imaged by a single HSI pixel might contain multiple materials. Finally, energy reflected or emitted from nearby materials might bleed into the pixel as a result of scattering. The *confusion matrix* is a standard tool for summarizing the correct and incorrect assignments generated by a classification method [4].

Let $C_{\text{ref}}[x, y]$ be a truth or *reference* classification image. Also, let $C[x, y]$ be the class image returned by a classifier. The confusion matrix M is an $L \times L$ matrix of nonnegative integers. It compares the actual or assigned class image $C[x, y]$ with the reference class image $C_{\text{ref}}[x, y]$. Matrix element M_{ij} gives the number of times that $C[x, y] = i$ and $C_{\text{ref}}[x, y] = j$, i.e., the number of times the classifier assigned a pixel to class i when the pixel belongs to class j in the reference class image.

Portions of the subsequent discussion will treat C and C_{ref} as realizations of two ergodic, multinomial random processes, $\mathbf{C}$ and $\mathbf{C}_{\text{ref}}$. In practice, such a distinction is often ignored. However, it is necessary in the development of statistical tests for evaluating classification performance. We use the boldface symbols to distinguish between realizations (C, C_{ref}) and stochastic processes ($\mathbf{C}$, $\mathbf{C}_{\text{ref}}$).

We quickly review some quantities commonly associated with the confusion matrix. First, let $P_{\text{prod}}(j)$ denote the *producer's accuracy probability for class* j,

$$P_{\text{prod}}(j) = \Pr(\mathbf{C}[x, y] = j | \mathbf{C}_{\text{ref}}[x, y] = j).$$

In practice, $P_{\text{prod}}(j)$ is estimated from the confusion matrix M. Denote the sum down the jth column of M by $M_{\cdot j} = \sum_{i=1}^{L} M_{ij}$. $M_{\cdot j}$ equals the true number of pixels from class j that are present. Then $\hat{P}_{\text{prod}}(j)$, the estimated producer's accuracy probability for class j, is

$$\hat{P}_{\text{prod}}(j) = \widehat{\Pr}(\mathbf{C}[x, y] = j | \mathbf{C}_{\text{ref}}[x, y] = j) = \frac{M_{jj}}{M_{\cdot j}} = \frac{M_{jj}}{\sum_{i=1}^{L} M_{ij}}. \tag{18}$$

$\hat{P}_{\text{prod}}(j)$ is frequently referred to as the *producer's accuracy for class* j. It tells us, of all the reference class-j pixels, how many were correctly placed into class j by the classifier.

Similarly, the *user's accuracy probability for class* i is denoted by $P_{\text{user}}(i)$,

$$P_{\text{user}}(i) = \Pr(\mathbf{C}_{\text{ref}}[x, y] = i | \mathbf{C}[x, y] = i).$$

This probability is also estimated from the confusion matrix. The sum across the ith row of M is given by $M_{i\cdot} = \sum_{j=1}^{L} M_{ij}$, the total number of pixels put into class i by the classifier. The estimate of $P_{\text{user}}(i)$ is then

$$\hat{P}_{\text{user}}(i) = \widehat{\Pr}(\mathbf{C}_{\text{ref}}[x,y] = i | \mathbf{C}[x,y] = i) = \frac{M_{ii}}{M_{i\cdot}} = \frac{M_{ii}}{\sum_{j=1}^{L} M_{ij}}. \tag{19}$$

$\hat{P}_{\text{user}}(i)$ is commonly called the *user's accuracy for class i*, and it tells us, of all the pixels that the classifier placed — both correctly and incorrectly — into class i, how many are truly in class i.

Finally, the *overall accuracy probability* P_{ovrl} is defined as

$$P_{\text{ovrl}} = \Pr(\mathbf{C}[x,y] = \mathbf{C}_{\text{ref}}[x,y]) = \sum_{j=1}^{L} \Pr(\mathbf{C}[x,y] = j, \mathbf{C}_{\text{ref}}[x,y] = j). \tag{20}$$

P_{ovrl} is also often referred to as the probability of correct classification (P_{cc}). P_{ovrl} can be estimated from M as follows. Let $M_{\cdot\cdot}$ denote the sum of all elements of M (so $M_{\cdot\cdot} = N_x N_y$). Then the estimate of P_{ovrl} is

$$\begin{aligned}
\hat{P}_{\text{ovrl}} = \widehat{\Pr}(\mathbf{C}[x,y] = \mathbf{C}_{\text{ref}}[x,y]) &= \sum_{j=1}^{L} \widehat{\Pr}(\mathbf{C}[x,y] = j, \mathbf{C}_{\text{ref}}[x,y] = j) \\
&= \sum_{j=1}^{L} \widehat{\Pr}(\mathbf{C}[x,y] = j | \mathbf{C}_{\text{ref}}[x,y] = j)\, \widehat{\Pr}(\mathbf{C}_{\text{ref}}[x,y] = j) \\
&= \sum_{j=1}^{L} \frac{M_{jj}}{M_{\cdot j}} \frac{M_{\cdot j}}{M_{\cdot\cdot}} = \frac{1}{M_{\cdot\cdot}} \sum_{j=1}^{L} M_{jj},
\end{aligned} \tag{21}$$

where $\Pr(\mathbf{C}_{\text{ref}}[x,y] = j)$ is estimated via

$$\widehat{\Pr}(\mathbf{C}_{\text{ref}}[x,y] = j) = \frac{M_{\cdot j}}{M_{\cdot\cdot}} = \frac{1}{M_{\cdot\cdot}} \sum_{i=1}^{L} M_{ij}, \quad j = 1, 2, \ldots, L. \tag{22}$$

The estimate $\hat{P}_{\text{ovrl}}$ is known as the *overall accuracy*.

5.3 KAPPA COEFFICIENT

The producer's, user's, and overall accuracy probabilities measure different aspects of classification performance. Taken individually, none is sufficient for summarizing the performance of a classifier [4]. For a particular class, the producer's accuracy probability might be high, but the user's accuracy probability low, and vice versa. Also, the overall accuracy $\hat{P}_{\text{ovrl}}$ captures only the diagonal of the confusion matrix. Note that the probability of error P_E is simply $P_E = 1 - P_{\text{ovrl}}$, so its estimate $\hat{P}_E = 1 - \hat{P}_{\text{ovrl}}$ only includes the diagonal of the confusion matrix as well.

The *kappa coefficient* K provides an alternative scalar measure of classification performance [3, 4] that compares the overall accuracy probability against the probability of chance agreement. K is defined as

$$K = \frac{P_{\text{ovrl}} - P_{\text{chance}}}{1 - P_{\text{chance}}}, \tag{23}$$

where

$$P_{\text{chance}} = \sum_{i=1}^{L} \Pr(\mathbf{C}[x,y] = i)\ \Pr(\mathbf{C}_{\text{ref}}[x,y] = i) \tag{24}$$

is the probability that a pixel is correctly classified by chance.

$\hat{K}$, an estimate of K, can be computed from the confusion matrix. First, the estimate follows the form of (23):

$$\hat{K} = \frac{\hat{P}_{\text{ovrl}} - \hat{P}_{\text{chance}}}{1 - \hat{P}_{\text{chance}}}. \tag{25}$$

Second, P_{ovrl} is estimated from M as in (21). Third, P_{chance} is estimated by combining $\widehat{\Pr}(\mathbf{C}_{\text{ref}}[x,y] = j)$ in (22) with

$$\widehat{\Pr}(\mathbf{C}[x,y] = i) = \frac{M_{i\cdot}}{M_{\cdot\cdot}} = \frac{1}{M_{\cdot\cdot}} \sum_{j=1}^{L} M_{ij}, \quad i = 1, 2, \ldots, L, \tag{26}$$

so that

$$\hat{P}_{\text{chance}} = \sum_{i=1}^{L} \widehat{\Pr}(\mathbf{C}[x,y] = i)\widehat{\Pr}(\mathbf{C}_{\text{ref}}[x,y] = i) = \frac{1}{M_{\cdot\cdot}^2} \sum_{i=1}^{L} M_{i\cdot} M_{\cdot i}. \tag{27}$$

Finally, the resulting maximum likelihood estimate of K is obtained by substituting the equations for$\hat{P}_{\text{ovrl}}$ and $\hat{P}_{\text{chance}}$ into (25):

$$\hat{K} = \frac{M_{\cdot\cdot} \sum_{i=1}^{L} M_{ii} - \sum_{i=1}^{L} M_{i\cdot} M_{\cdot i}}{M_{\cdot\cdot}^2 - \sum_{i=1}^{L} M_{i\cdot} M_{\cdot i}}. \tag{28}$$

Equation (28) is the expression for kappa given in many documents (e.g., [21, Ch. 6]) but without much explanation of the associated stochastic interpretation.

Unlike $\hat{P}_{\text{prod}}(j)$, $\hat{P}_{\text{user}}(i)$, and $\hat{P}_{\text{ovrl}}$ (or $\hat{P}_E$), $\hat{K}$ incorporates both the diagonal and marginals (column and row sums) of the confusion matrix. It does not, however, take into account the producer's and user's accuracies.

Some intuition into the form (23) of K can be gained by writing P_{ovrl} as

$$P_{\text{ovrl}} = \sum_{i=1}^{L} \Pr(\mathbf{C}[x,y] = i | \mathbf{C}_{\text{ref}}[x,y] = i)\ \Pr(\mathbf{C}_{\text{ref}}[x,y] = i). \tag{29}$$

In this equation, the conditional probability $\Pr(\mathbf{C}[x,y] = i|\mathbf{C}_{\text{ref}}[x,y] = i)$ reflects the dependence of the classifier output $\mathbf{C}[x,y]$ upon the actual reference class value $\mathbf{C}_{\text{ref}}[x,y]$. If the classifier accurately labels a pixel as class i when the pixel truly belongs to class i, then $\Pr(\mathbf{C}[x,y] = i|\mathbf{C}_{\text{ref}}[x,y] = i)$ should be close to unity. If the classifier performed perfectly, then $\Pr(\mathbf{C}[x,y] = i|\mathbf{C}_{\text{ref}}[x,y] = i)$ would equal 1 for $i = 1$, $2, \ldots, L$, and P_{ovrl} would equal 1.

Next, the chance agreement probability P_{chance} can be considered the probability of correct classification if the classifier output $\mathbf{C}[x,y]$ were conditionally independent of the reference pixel value $\mathbf{C}_{\text{ref}}[x,y]$, i.e., if

$$\Pr(\mathbf{C}[x,y] = i|\mathbf{C}_{\text{ref}}[x,y] = i) = \Pr(\mathbf{C}[x,y] = i).$$

Then (24) can be written as

$$P_{\text{chance}} = \sum_{i=1}^{L} \Pr(\mathbf{C}[x,y] = i | \mathbf{C}_{\text{ref}}[x,y] = i)\ \Pr(\mathbf{C}_{\text{ref}}[x,y] = i)\Big|_{\mathbf{C} \text{ and } \mathbf{C}_{\text{ref}} \text{ conditionally independent}}$$
$$= \sum_{i=1}^{L} \Pr(\mathbf{C}[x,y] = i)\ \Pr(\mathbf{C}_{\text{ref}}[x,y] = i).$$

Conceptually, P_{chance} is the probability of correct classification that would result if the classifier output $\mathbf{C}[x,y]$ bore no relation to the actual class value $\mathbf{C}_{\text{ref}}[x,y]$. It can also be thought of as the expected probability of correct classification if the class labels in $C[x,y]$ were assigned randomly using the distribution $\Pr(\mathbf{C}[x,y] = i)$.[3]

From this discussion, even if the classifier output $\mathbf{C}[x,y]$ bears no relation to the actual class labels $\mathbf{C}_{\text{ref}}[x,y]$, we expect to have correct classification with probability P_{chance}. Also, the best possible probability of correct classification is unity. In (23), the kappa coefficient K first removes the bias due to chance agreement and then normalizes the result $P_{\text{ovrl}} - P_{\text{chance}}$ by the best possible result, $1 - P_{\text{chance}}$.

Values of K close to unity indicate that classification is better than chance, while a value of zero means the classification is no better than chance. (K can even take on negative values, although this occurs very rarely and indicates that the classifier is performing worse than chance. In such a case, one would be better off simply assigning class labels in $C[x,y]$ at random, rather than using the classifier.) Some authors [16] have also suggested the following scale for evaluating the agreement between C and C_{ref}:

$0.8 < K \leq 1$	strong agreement;
$0.4 < K \leq 0.8$	moderate agreement;
$K \leq 0.4$	poor agreement.

5.4 KAPPA ANALYSIS

Kappa analysis is a method for determining whether or not classification differences are statistically significant [4]. In the stochastic interpretation of kappa analysis, K is itself a random variable, and$\hat{K}$ in (28) is an estimate of K. Kappa analysis also provides an estimate $\hat{\sigma}^2_{\hat{K}}$ of the variance of $\hat{K}$. The estimates of $\hat{K}$ and $\hat{\sigma}^2_{\hat{K}}$ can be used to provide confidence intervals on K. We repeat the expression from Congalton and Green [4, p. 50]:

$$\hat{\sigma}^2_{\hat{K}} = \frac{1}{M_{\cdot\cdot}} \left[\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2 - \theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4 - 4\theta_2^2)}{(1-\theta_2)^4} \right], \tag{30}$$

where

$$\theta_1 = \frac{1}{M_{\cdot\cdot}} \sum_{i=1}^{L} M_{ii}, \qquad \theta_2 = \frac{1}{M_{\cdot\cdot}^2} \sum_{i=1}^{L} M_{i\cdot} M_{\cdot i}, \qquad \theta_3 = \frac{1}{M_{\cdot\cdot}^2} \sum_{i=1}^{L} M_{ii} \left(M_{i\cdot} + M_{\cdot i} \right), \tag{31}$$

[3] This is roughly analogous to rolling a pair of fair dice many times: one-sixth of the time, "doubles" will occur, even though the dice do not influence one another.

and

$$\theta_4 = \frac{1}{M_{\cdot\cdot}^3} \sum_{i=1}^{L} \sum_{j=1}^{L} M_{ij} \left(M_{j\cdot} + M_{\cdot i} \right)^2 . \tag{32}$$

Statistical significance tests can be applied because, under the multinomial sampling model of kappa analysis, $\hat{K}$ is asymptotically normally distributed. For example, if $K = 0$, then the classifier does not perform any better than random classification. A statistical test to determine whether or not $K = 0$ consists of the competing hypotheses [4]:

$$H_0 : K = 0;$$
$$H_1 : K \neq 0.$$

With the estimates $\hat{K}$ and $\hat{\sigma}^2_{\hat{K}}$, the appropriate test statistic is

$$Z = \frac{\hat{K}}{\sqrt{\hat{\sigma}^2_{\hat{K}}}},$$

which has a standard normal distribution. H_0 is rejected if $Z \geq z_{\alpha/2}$ at significance level α. Otherwise, H_0 is accepted.

5.5 OTHER STATISTICAL MEASURES OF CLASSIFICATION PERFORMANCE

A number of measures of classification performance similar to K have been proposed. Although they are not used in this study, they nonetheless deserve mention. Like K in (23), these other measures have the form

$$\frac{P_{\text{ovrl}} - P_{\text{bias}}}{1 - P_{\text{bias}}},$$

where P_{ovrl} is the overall accuracy probability, and P_{bias} is a *bias probability* that is subtracted to eliminate the influence of correct classifications that occur purely by chance. P_{ovrl} is given by (20) and estimated using (21). The classification measures differ in their choice of the bias probability P_{bias}.

For the kappa coefficient, the bias probability is

$$P_{\text{bias}} = P_{\text{chance}} = \sum_{i=1}^{L} \Pr(\mathbf{C}[x, y] = i) \ \Pr(\mathbf{C}_{\text{ref}}[x, y] = i).$$

In practice, P_{chance} is estimated from the empirical estimates $\widehat{\Pr}(\mathbf{C}[x, y] = i)$ and $\widehat{\Pr}(\mathbf{C}_{\text{ref}}[x, y] = i)$ of (26) and (22).

Brennan and Prediger [1], Foody [6], and Ma and Redmond [18] suggest using $P_{\text{bias}} = 1/L$. Ma and Redmond argue that, when C_{ref} is generated by unsupervised classification methods, there is no prior knowledge of how many pixels will likely be assigned to each class, so the *a priori* probabilities should be the same for all classes; i.e., $\Pr(\mathbf{C}_{\text{ref}}[x, y] = i) = 1/L$, $i = 1, 2, \ldots, L$. Consequently,

$$P_{\text{bias}} = \sum_{i=1}^{L} \Pr(\mathbf{C}[x, y] = i) \, \frac{1}{L} = \frac{1}{L}, \tag{33}$$

which results in the performance measure

$$T_e = \frac{P_{\text{ovrl}} - 1/L}{1 - 1/L}. \tag{34}$$

Ma and Redmond call T_e, "Tau for use with classifications based on equal probabilities of [class] membership" [18]. The same measure results in the case of supervised classification when the operator supervising classification assumes $\Pr(\mathbf{C}_{\text{ref}}[x, y] = i) = 1/L$. The estimated value of T_e is $\hat{T}_e = (\hat{P}_{\text{ovrl}} - 1/L)/(1 - 1/L)$.

Ma and Redmond also suggest a generalization in which $\Pr(\mathbf{C}_{\text{ref}}[x, y] = i)$ is estimated from other sources, such as "a previous survey of group population, an estimation of group population from aerial photos, or an arbitrary decision for the importance of groups" [18]. In this case, they denote the measure by $T_p = (P_{\text{ovrl}} - P_{\text{bias}})/(1 - P_{\text{bias}})$, with the corresponding estimate

$$\hat{T}_p = \frac{\hat{P}_{\text{ovrl}} - \hat{P}_{\text{bias}}}{1 - \hat{P}_{\text{bias}}}, \tag{35}$$

where $\hat{P}_{\text{ovrl}}$ is given by (21), and

$$\hat{P}_{\text{bias}} = \sum_{i=1}^{L} \widehat{\Pr}(\mathbf{C}[x, y] = i)\, \widehat{\Pr}(\mathbf{C}_{\text{ref}}[x, y] = i) = \frac{1}{M_{\cdot\cdot}} \sum_{i=1}^{L} M_{i\cdot}\, \widehat{\Pr}(\mathbf{C}_{\text{ref}}[x, y] = i), \tag{36}$$

Equations for the variances of $\hat{T}_e$ and $\hat{T}_p$ and test statistics for evaluating classifier performance appear in [18].

Additional measures of classifier performance that are related to kappa are "conditional kappa" and "weighted kappa" [4]. Conditional kappa is useful when one is interested in the classifier performance for a particular class. Weighted kappa is appropriate when certain classification errors are more important than others and can be weighted accordingly.

5.6 ONE-KAPPA TEST FOR ASSESSMENT OF COMPRESSION EFFECTS

Kappa analysis as described in Section 5.4 is typically used to determine whether or not a classifier performs significantly better than chance assignment, or whether or not two confusion matrices are significantly different. In a similar way, we apply kappa analysis on one or two confusion matrices to evaluate the (dis)similarity between classifications before and after lossy compression. These tests allow one to determine whether or not lossy compression has had a significant effect on the classification results.

The first method, or *one-kappa test*, involves a single confusion matrix or kappa coefficient. The test assesses the agreement between classification results from the same classifier before and after compression. A classifier is first applied to X, the entire original, noncompressed cube, and the resulting class image provides C_{orig}. Then the same classifier is applied to $\hat{X}$, a decompressed version of the cube, to produce a *post-compression class image* C_{comp}. The question is, *"For a given classifier, does compression significantly change its results?"*

To answer this question, we obtain the confusion matrix and kappa coefficient K from C_{orig} and C_{comp} and then perform kappa analysis. A diagram of the approach is shown in Figure 17. We assume that some

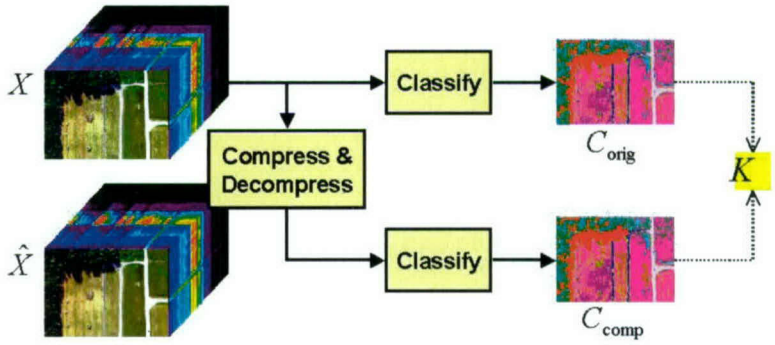


Figure 17. Diagram of one-kappa test.

minimum acceptable kappa coefficient $K_{\min} \in [0, 1]$ is given; $K_{\min}$ should be close to unity. If $K > K_{\min}$, then we declare that C_{comp} is acceptable; otherwise, we declare that C_{comp} differs too greatly from C_{orig}. This idea yields the following test with maximum significance level α:

$$\begin{aligned} H_0 &: K = K_{\min}, \quad \text{(overcompression)}; \\ H_1 &: K > K_{\min}, \quad \text{(acceptable)}. \end{aligned} \tag{37}$$

The test statistic is

$$Z_1 = \frac{\hat{K} - K_{\min}}{\hat{\sigma}_{\hat{K}}}, \tag{38}$$

where Z_1 has a standard normal distribution [4]. We accept H_1 if $Z_1 > z_\alpha$ and reject H_1 if $Z_1 \leq z_\alpha$.

In (37), the probability of Type I error (accept H_1 when H_0 is true) is greatest for $K = K_{\min}$ and lower for $K < K_{\min}$; that is, $\Pr(\text{accept } H_1 | H_0 \text{ true}) = \alpha$ for $K = K_{\min}$, and $\Pr(\text{accept } H_1 | H_0 \text{ true}) < \alpha$ for $K < K_{\min}$. Thus, we protect ourselves against the worst possible probability of Type I error [12]. Such an error corresponds to an *overcompression error*: We incorrectly decide that the post-compression classification result is acceptable, when in reality lossy compression has discarded too much information. In contrast, a Type II error (accept H_0 when H_1 is true) is an *undercompression error*. We cannot determine the probability of Type II error unless we replace H_1 with an equality such as $H_1 : K = K_1$, and then this probability only holds for that particular value of K_1. Since overcompression may have more serious consequences than undercompression (cf. Section 1.3), we have arranged the test to limit the probability of overcompression error.

5.7 TWO-KAPPA TEST FOR ASSESSMENT OF COMPRESSION EFFECTS

In the one-kappa test, the reference C_{orig} does not represent the truth; rather, C_{orig} is just the class image produced by the classifier without compression. Hence, the test decides whether or not a classifier remains consistent **with itself** following compression. Thus, if the classifier performs poorly in the absence of compression, the test determines whether or not the classifier continues to perform poorly after lossy compression. The test says nothing about how closely the class images agree with truth. Even without compression, a classifier will not typically classify all pixels correctly. This fact leads us to a *two-kappa test*,

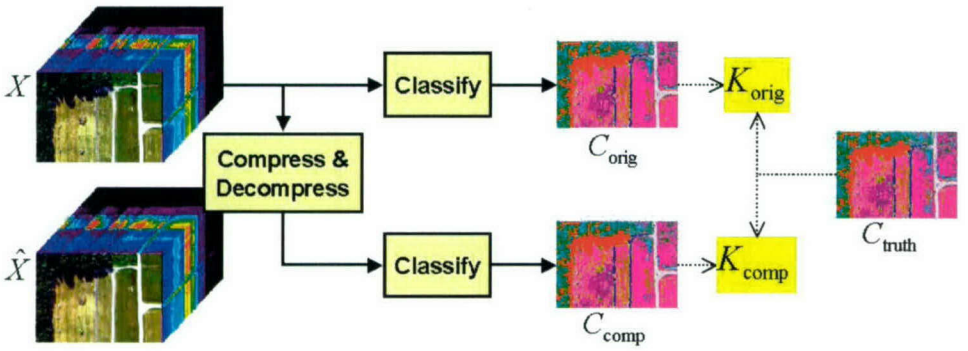


Figure 18. Diagram of two-kappa test.

which compares two kappa coefficients to answer the question, *"For a given classifier, does compression significantly reduce its agreement with truth?"*

This approach appears in Figure 18. For this test, we use C_{truth} to emphasize that the reference class image represents truth. First, we apply a classifier to X to produce the class image C_{orig} and compute the kappa coefficient K_{orig} from C_{truth} and C_{orig}. Next, we apply the same classifier to $\hat{X}$, obtain the post-compression class image C_{comp}, and determine a second kappa coefficient K_{comp} from C_{truth} and C_{comp}. We assume we can tolerate some maximum degradation $\Delta K_{\text{tol}} > 0$ after compression and construct a test with maximum significance level α:

$$\begin{aligned} H_0 &: K_{\text{orig}} - K_{\text{comp}} = \Delta K_{\text{tol}}, \quad \text{(overcompression)}; \\ H_1 &: K_{\text{orig}} - K_{\text{comp}} < \Delta K_{\text{tol}}, \quad \text{(acceptable)}. \end{aligned} \tag{39}$$

The test statistic is

$$Z_2 = \frac{\hat{K}_{\text{orig}} - \hat{K}_{\text{comp}} - \Delta K_{\text{tol}}}{\sqrt{\hat{\sigma}^2_{\hat{K}_{\text{orig}}} + \hat{\sigma}^2_{\hat{K}_{\text{comp}}}}}. \tag{40}$$

where $\hat{\sigma}_{\hat{K}_{\text{orig}}}$ and $\hat{\sigma}_{\hat{K}_{\text{comp}}}$ are the estimates of the standard deviations of K_{orig} and K_{comp}. Z_2 has a standard normal distribution [4].

If $Z_2 \leq -z_\alpha$, we accept H_1; otherwise, we reject H_1. Since $\Delta K_{\text{tol}} > 0$, the test also accepts H_1 (with maximum significance level α) in the case where $K_{\text{comp}} > K_{\text{orig}}$, i.e., where compression **improves** classifier performance; such behavior could occur in the conventional approach because lossy compression can have a denoising effect [2]. Finally, like (37), (39) is posed such that a Type I error corresponds to an overcompression error.

6. EXPERIMENTAL RESULTS

This section documents results for an HSI cube of a forest scene (Forest Radiance, run05) in the HYDICE data collection. This data was collected from an airborne platform at an altitude of 1.5 km; at this altitude, the ground sample distance is approximately 0.8 m. The HYDICE sensor returns data as 12-bit unsigned integers. However, calibration increases the dynamic range and occasionally produces negative values [13]; hence, the radiance values are stored as 16-bit signed integers.

The initial radiance cube had dimensions of $320 \times 1280 \times 210$ but was modified prior to the compression experiments. Spatially, the ten leftmost and rightmost columns were discarded because they corresponded to all-zero pixels at the edges of the sensor focal plane array. Spectrally, bands corresponding to the water-absorption wavelengths were eliminated. This preparation step yielded an original cube X of size $300 \times 1280 \times 145$, requiring 106 MB ($111.36 \times 10^6$ bytes) of storage. All rates and compression ratios are reported relative to this cube. The actual dynamic range of X was -121 to 12911.

The compression scheme described in Section 2 was then applied to X for a variety of compression settings. The PCA/KLT side information ($\bar{X}$ and Φ) required $B_{\text{side}} = 32$ bits per value, so $S_{\text{side}} = 84680$ bytes. The equal-rate allocation method used $N_c' = 145$, 50, 20, 15, 10, and 5, and overall compression ratios r_{ovrl} ranging from 14.5 up to about 340. Recall that for this method, with r_{ovrl} and N_c' given, $r_{\text{spat},k} = (N_c'/N_c)/r_{\text{ovrl}}$ for each retained PC image. The MSD-optimal allocation method used the same values of r_{ovrl}.

Figure 19 shows a graph of the eigenvalues for the modified radiance cube. Like Figure 6 for Moffett Field, the first few eigenvalues have much larger values than subsequent eigenvalues. Next, Figure 20 shows an example of the allocation results for $r_{\text{ovrl}} = 30$. The equal-rate allocation method compresses each retained PC image at almost the same spatial compression ratio, while the MSD-optimal rate allocation method compresses the first few PC images very little and increases the amount of spatial compression for the later PC images. Finally, Figure 21 shows the values of N_c' selected by the MSD-optimal method as a function of r_{ovrl}. Note also that for each choice of N_c', the retained PC images were compressed with different spatial compression ratios.

Figure 22 shows RGB images for a portion of the original cube X and several decompressed cubes. At $r_{\text{ovrl}} \approx 30$, the decompressed cubes look very similar to X. At $r_{\text{ovrl}} \approx 340$, compression artifacts are obvious for the equal-rate allocation, less so for the MSD-optimal allocation.

In order to generate an initial set of classes, we employed the ISODATA unsupervised clustering algorithm [10]. ISODATA begins by randomly choosing a number of initial mean spectra as cluster centroids; then an iterative procedure follows: First, each pixel in X is assigned to the cluster whose centroid is closest in a Euclidean-distance sense. Second, the standard deviations of the resulting clusters are then computed, and clusters with large standard deviations are split. Third, the centroids of the resulting clusters are computed, and clusters whose centroids are close together are merged. Finally, the procecedure is repeated until it converges, producing L clusters. We denote the class image associated with the final clustering by C_{iso}, and the corresponding set of class means by $\mathcal{L}_{\text{iso}}$. A portion of C_{iso} appears in Figure 23, along with the corresponding RGB image. For the given cube, ISODATA produced 15 clusters, which served as the basic classes: **In the following classification results, the number of classes is** $L = 15$.

6.1 DISTORTION RESULTS

Figure 24 shows plots of both RMSD$(X, \hat{X})$ and MSAD$(X, \hat{X})$ versus the overall compression ratio.[4] The curves associated with the equal-rate allocation scheme are marked with the corresponding value of N_c'. The curve corresponding to the MSD-optimal allocation is marked "MSD-opt."

As expected, the MSD-optimal allocation has substantially lower RMSD over the entire range of r_{ovrl}. With regard to MSAD, the MSD-optimal allocation has the smallest distortion for $r_{\text{ovrl}} < 200$, while the equal-rate allocation with $N_c' = 5$ is slightly better for larger values of r_{ovrl}. Recall that for these large values of r_{ovrl}, the MSD-optimal allocation retained only 3–5 PC images.

6.2 CLASSIFICATION RESULTS

To prepare for the one-kappa test, we adopted $\mathcal{L}_{\text{iso}}$ from ISODATA as our spectral library. We applied either the EMD or SAM classifier to X, and the resulting class image served as the reference class image C_{orig}; the ISODATA class image C_{iso} was not used with this test. For each decompressed cube $\hat{X}$, the same classifier was again applied to produce a post-compression class image C_{comp}. Lastly, $\hat{K}$ was determined from C_{orig} and C_{comp}. Figures 25 and 26 show portions of the class images produced by the EMD and SAM classifiers, respectively. Consistent with Figure 22, C_{comp} is similar to C_{orig} at $r_{\text{ovrl}} = 29$ for both allocations, but many differences appear for the equal-rate allocation at $r_{\text{ovrl}} = 340$.

The test was performed with $\alpha = 0.01$ and $K_{\min} = 0.95$. Figure 27 shows the results for EMD and SAM classification as r_{ovrl} is varied. The $100(1-\alpha)\%$ confidence intervals associated with $\hat{K}$ were extremely narrow and are not shown. The dotted line indicates $K_{\min}$: If $\hat{K}$ lies entirely above this line, then we decide that the classification differences are insignificant. For EMD classification, the MSD-optimal allocation results in considerably less performance degradation than the equal-rate allocations. It does not cause significant degradation provided that $r_{\text{ovrl}} < 200$, while the equal-rate allocations yield acceptable results for $r_{\text{ovrl}} < 105$. For SAM classification, the MSD-optimal allocation again produced better results than the equal-rate allocation, but the amount of compression that could be tolerated was lower.

For the two-kappa test, we again used $\mathcal{L}_{\text{iso}}$ as the spectral library, but we also employed C_{iso} as the truth class image; i.e., $C_{\text{truth}} = C_{\text{iso}}$. Recall that a portion of C_{iso} appears in Figure 23. Either the EMD or SAM classifier was applied to X to generate C_{orig}, and $\hat{K}_{\text{orig}}$ followed from C_{iso} and C_{orig}. For each decompressed cube $\hat{X}$, we again performed classification to obtain C_{comp} and determined $\hat{K}_{\text{comp}}$ from C_{iso} and C_{comp}.

Figure 28 shows the results with $\alpha = 0.01$ and $\Delta K_{\text{tol}} = 0.05$. Again the confidence intervals were narrow and are omitted. We decided that the classification differences are not significant whenever $\hat{K}_{\text{orig}} - \hat{K}_{\text{comp}}$ lies entirely below the dotted line for ΔK_{tol}. For EMD classification, $\hat{K}_{\text{orig}} = 0.9553$, indicating that C_{orig} agreed very closely with C_{iso} (cf. Figures 22 and 25). With compression, the MSD-optimal allocation outperformed the equal-rate allocations again; it caused no significant change in classification performance up to $r_{\text{ovrl}} = 305$. For SAM classification, $\hat{K}_{\text{orig}} = 0.5297$, which indicates moderate agreement between

[4] Peak signal to noise ratio (PSNR) is often used instead of RMSD. However, the calibrated radiance data usually span only a portion of the possible dynamic range, which makes computation of PSNR difficult [13], so we do not report it.

C_{orig} and C_{iso} (cf. Figures 22 and 26). This result is not too surprising because ISODATA employs Euclidean distance, but SAM uses spectral angle. The MSD-optimal allocation resulted in the fewest differences and did not cause significant degradation over the entire range of r_{ovrl} tested. Here, the equal-rate allocation also performed well for $N_c' = 5$.

6.3 DISTORTION AND CLASSIFICATION RESULTS

The graphs in Figures 24, 27, and 28 also show that both distortion and classification differences change monotonically as r_{ovrl} increases. Such behavior suggests that distortion might be used to anticipate classification degradation.

An example appears in Figure 29. For the MSD-optimal allocation, EMD classification degrades with increasing RMSD and becomes unacceptable for RMSD > 65; SAM classification remains acceptable until RMSD > 45. Similar behavior for the two-kappa test can be observed in Figure 30.

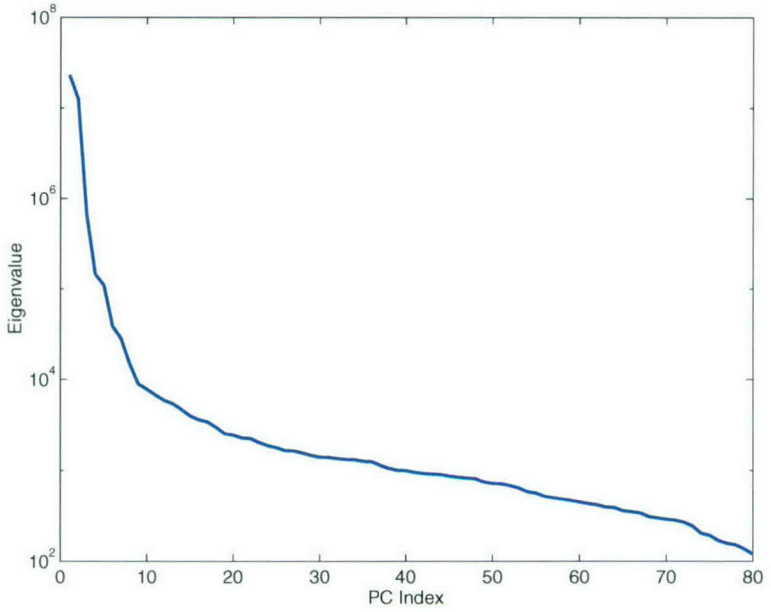


Figure 19. Eigenvalues for Forest Radiance. Only the first eighty eigenvalues are shown.

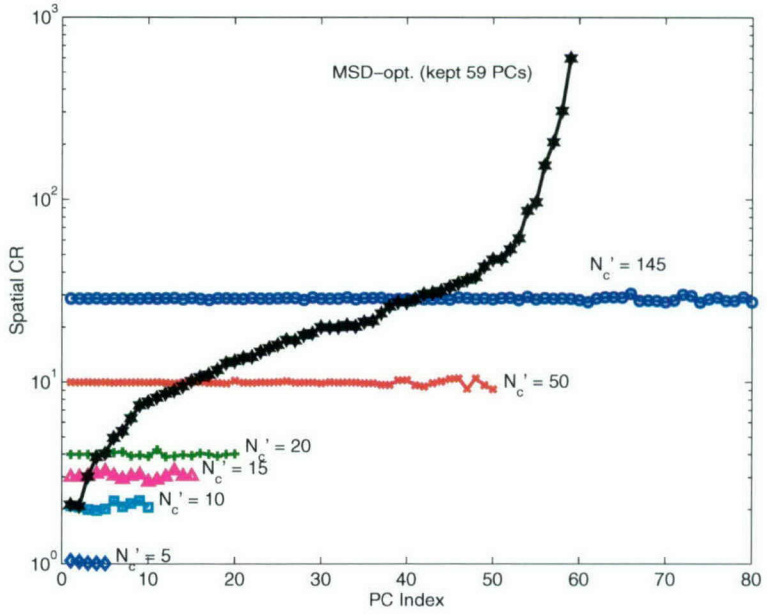


Figure 20. Example allocation results for Forest Radiance for an overall compression ratio of 30.

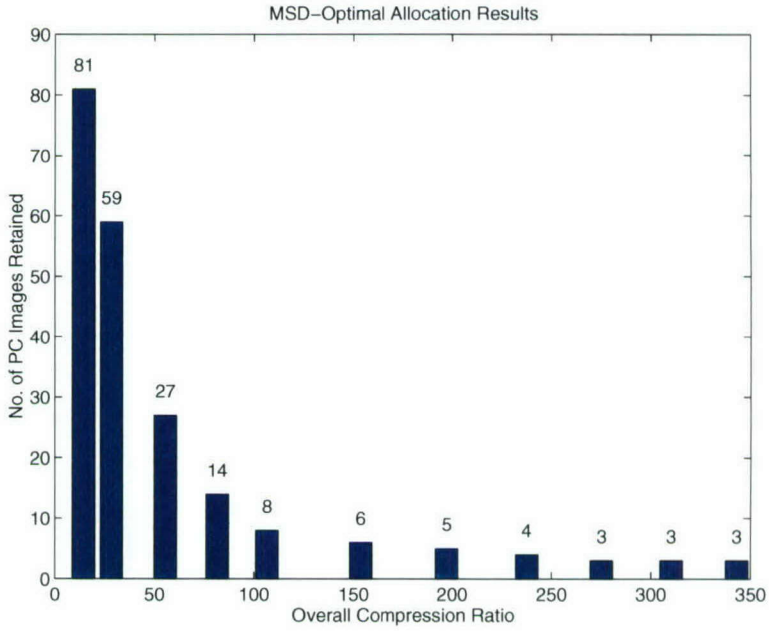


Figure 21. Number of PC images retained by MSD-optimal allocation for Forest Radiance.

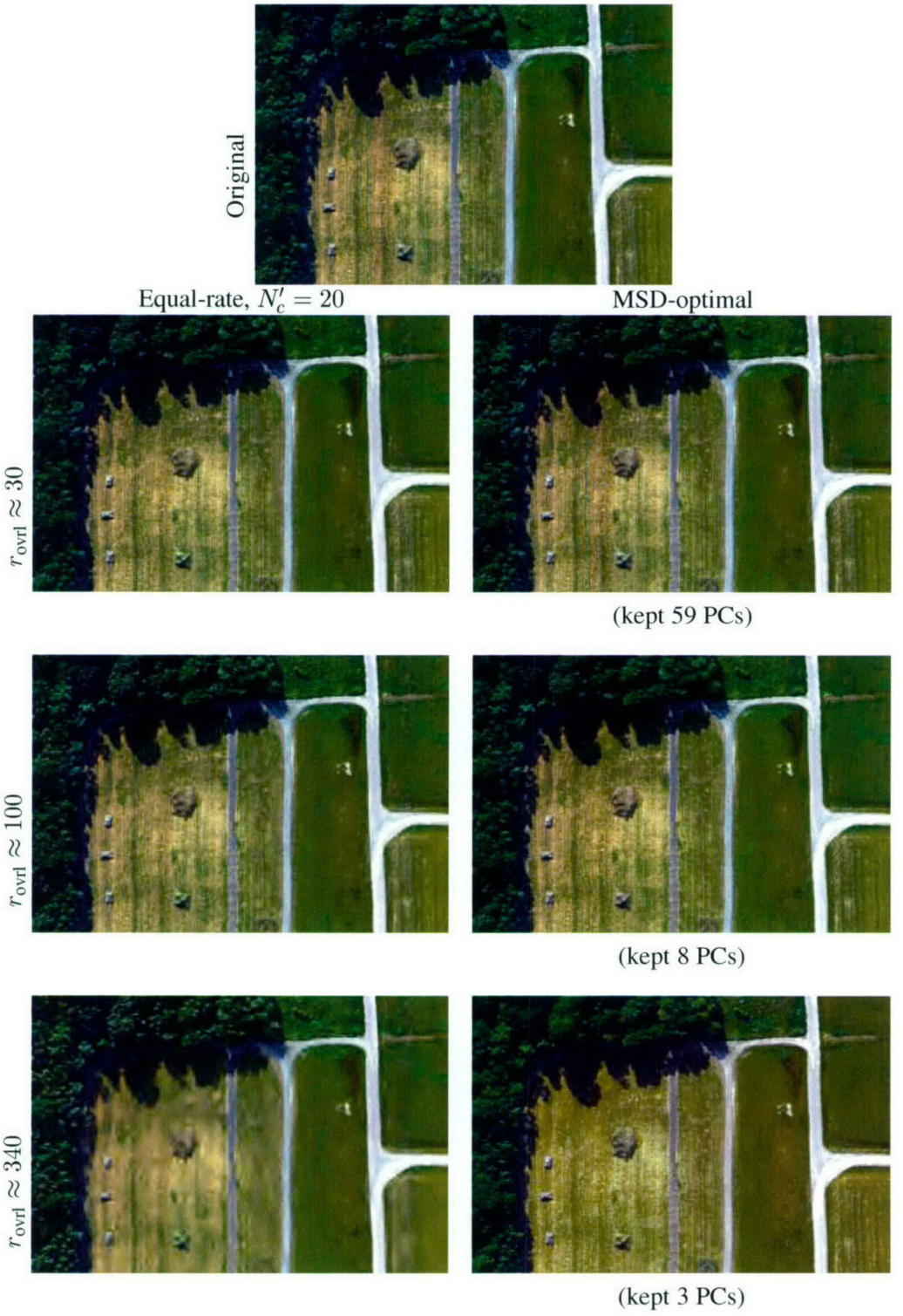
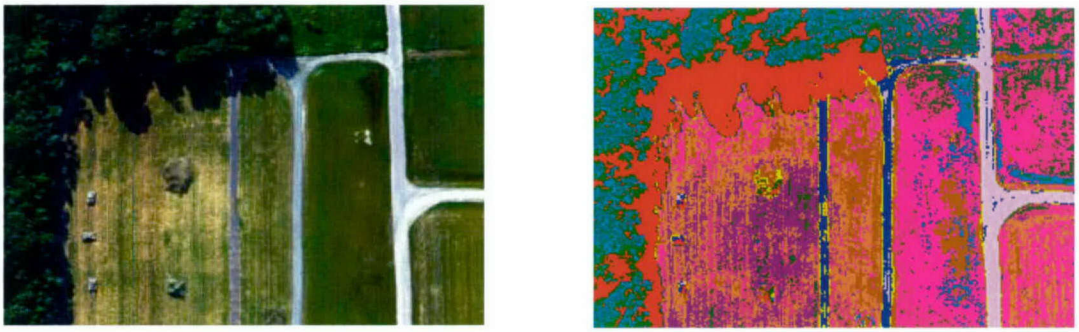


Figure 22. RGB images of a portion of the original and decompressed Forest Radiance cubes.



Original ISODATA result

Figure 23. Original RGB image and ISODATA class image C_{iso} for Forest Radiance.

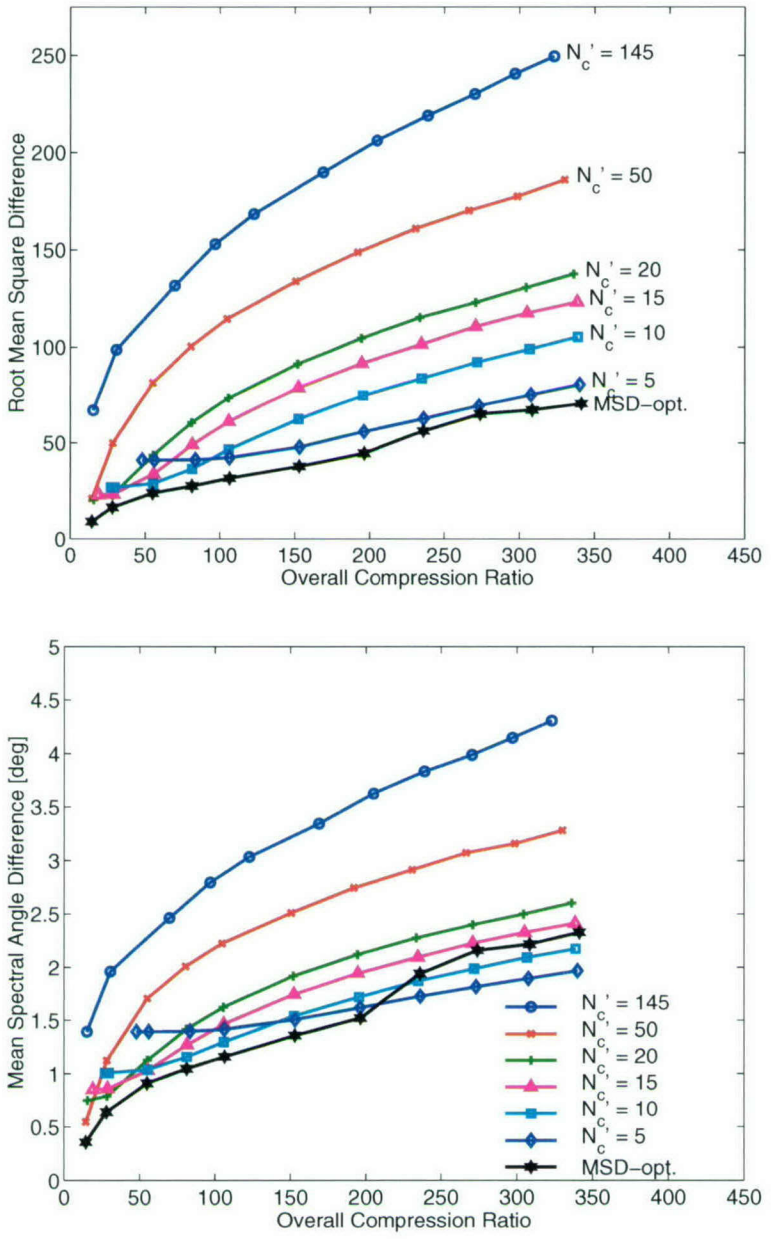


Figure 24. Distortion vs. compression ratio for Forest Radiance.

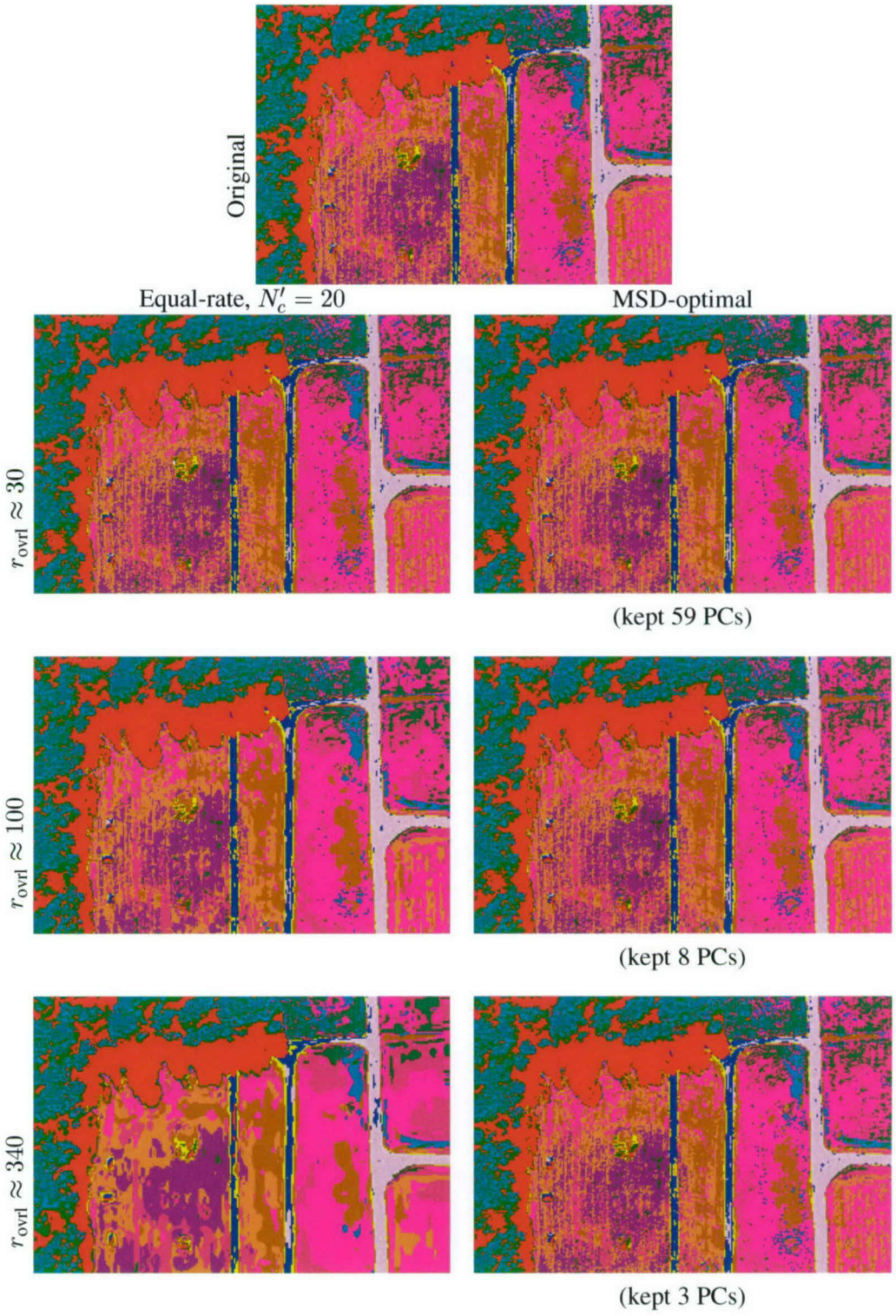


Figure 25. Class images for EMD classifier and Forest Radiance.

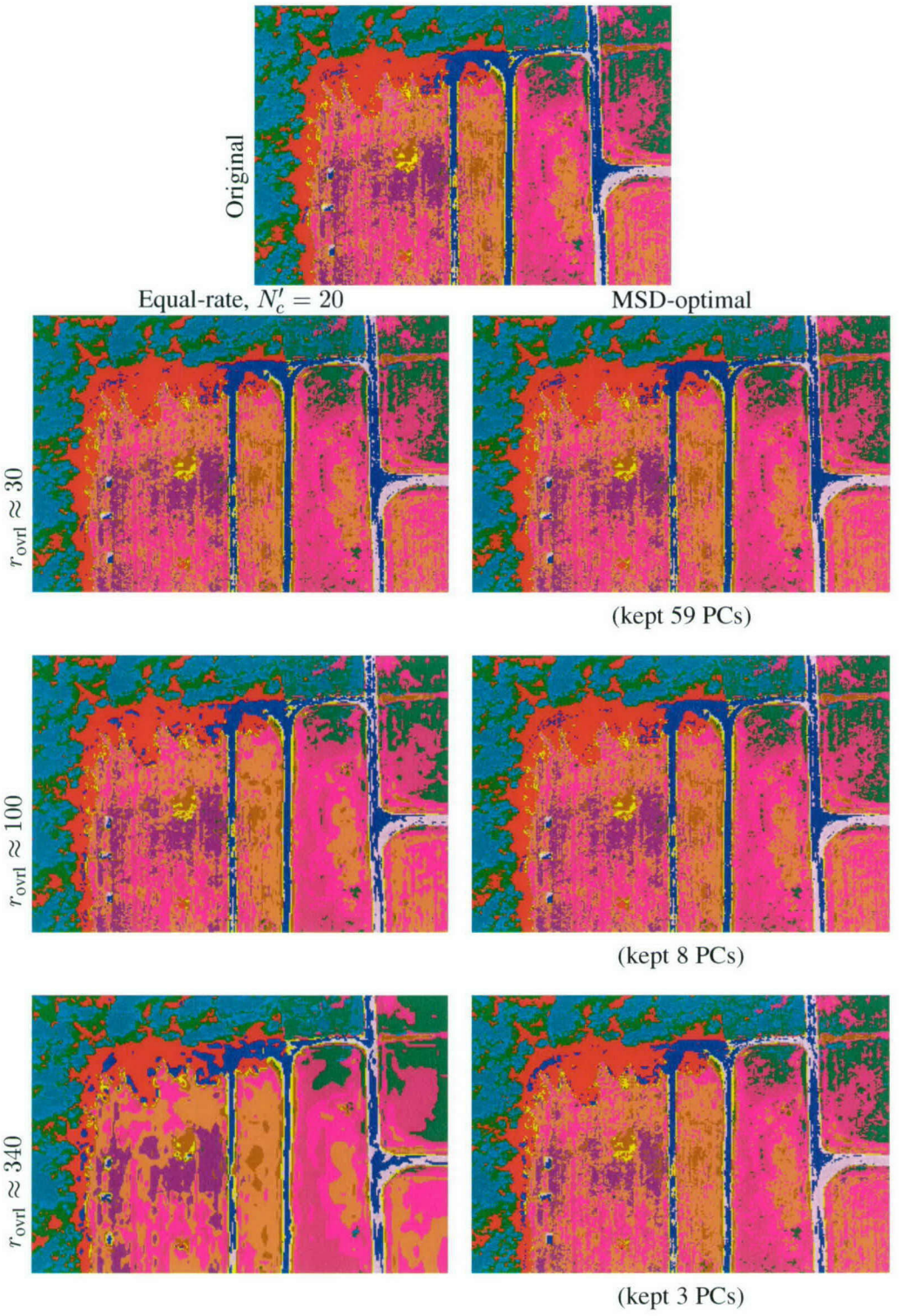


Figure 26. Class images for SAM classifier and Forest Radiance.

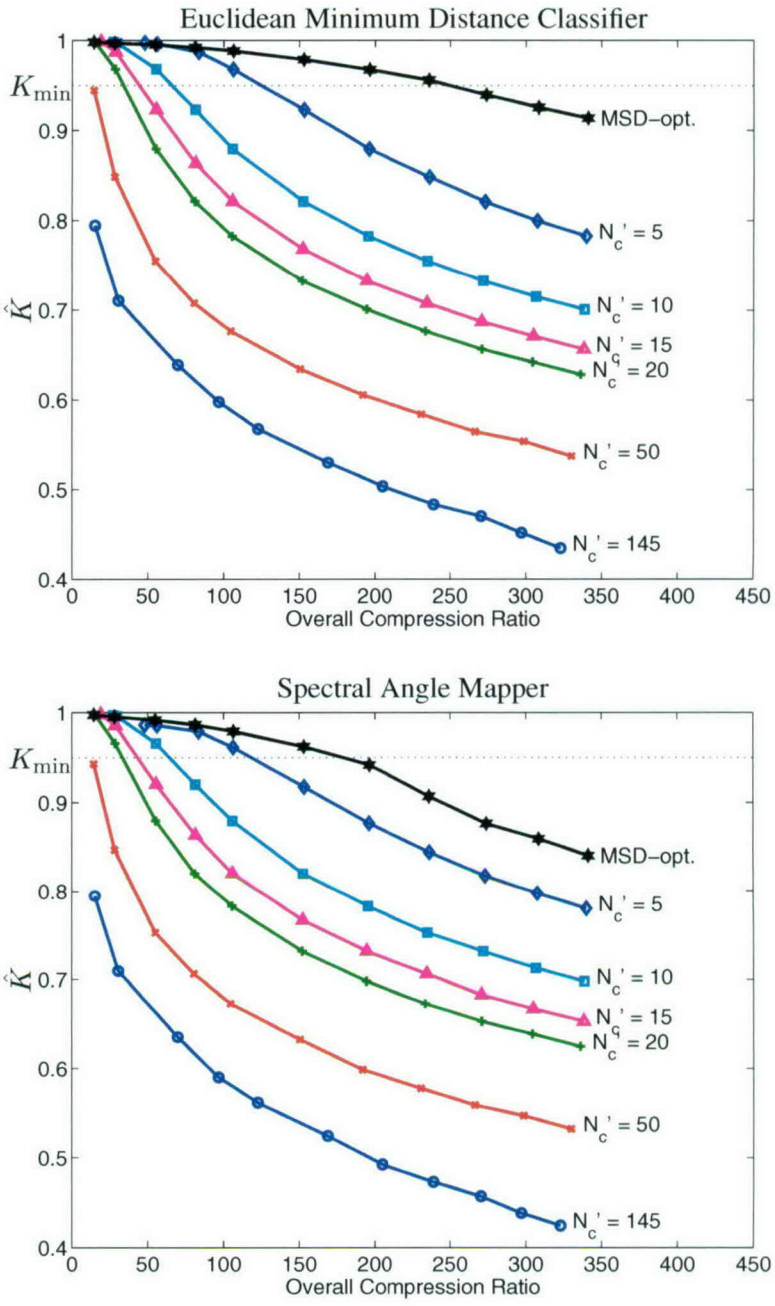


Figure 27. One-kappa test vs. compression ratio for Forest Radiance.

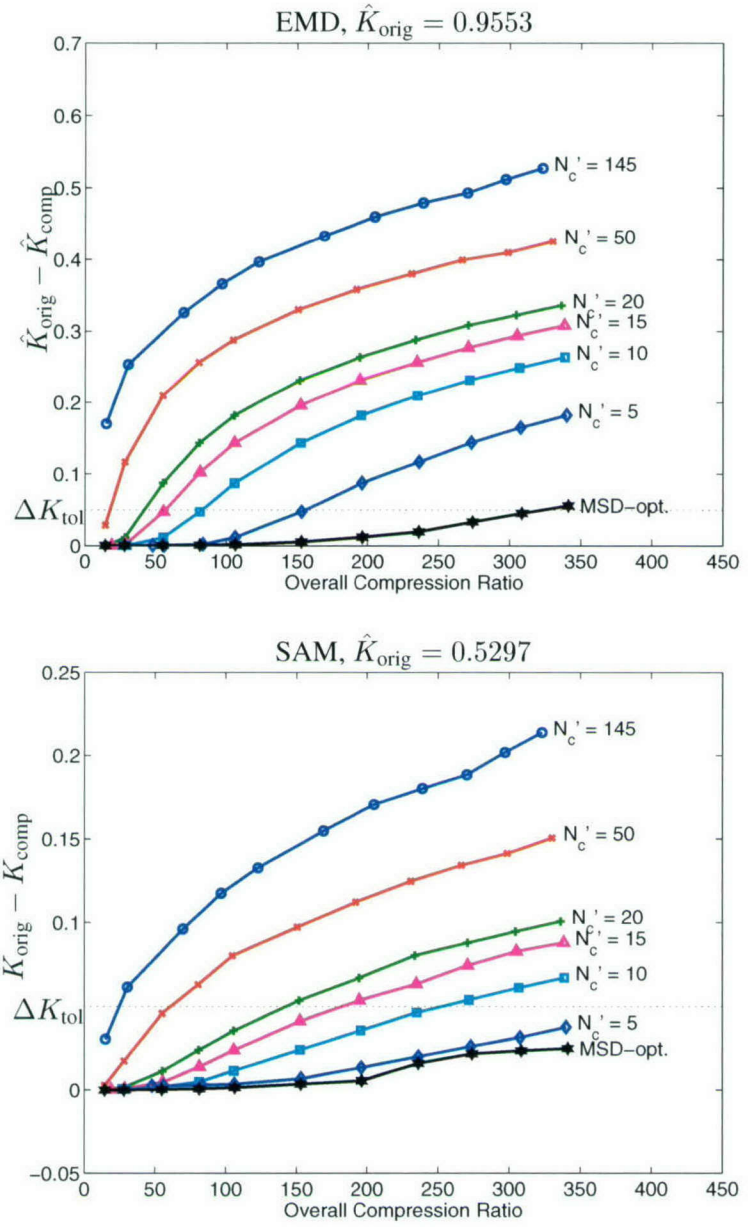


Figure 28. Two-kappa test vs. compression ratio for Forest Radiance.

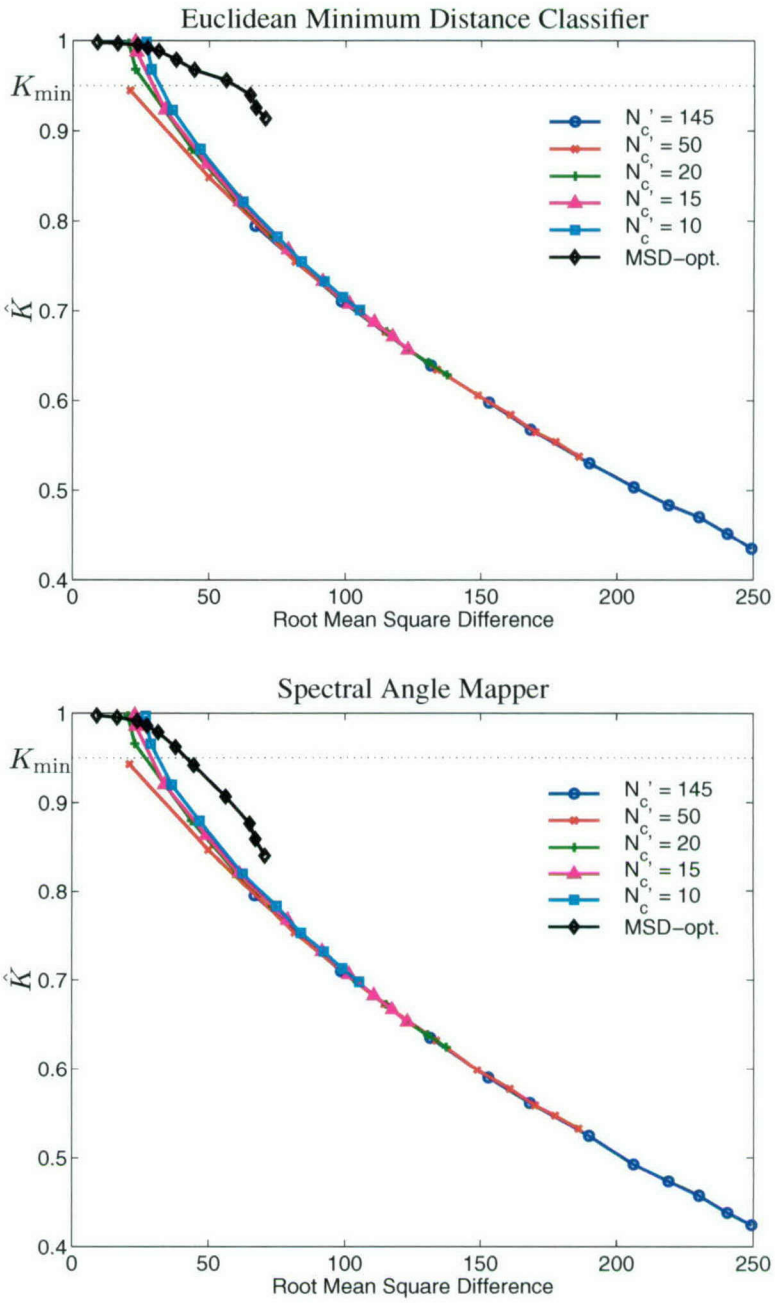


Figure 29. One-kappa test vs. RMSD for Forest Radiance.

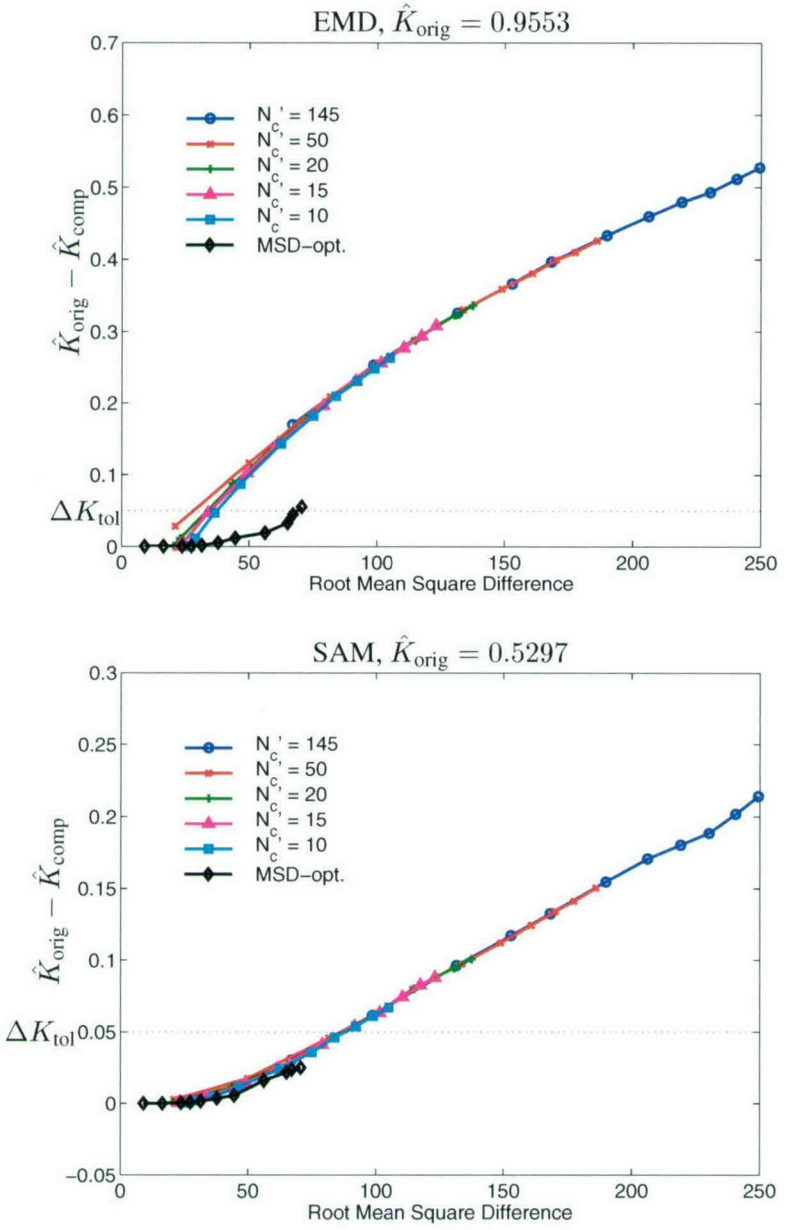


Figure 30. Two-kappa test vs. RMSD for Forest Radiance.

7. SUMMARY AND CONCLUSIONS

We have demonstrated the need for lossy compression of HSI data and discussed strategies and approaches to the problem. We have adopted the conventional approach, in which traditional compression methods and traditional application processing algorithms are employed. Other approaches, while potentially beneficial, tend to be specific to a single application and can become unwieldy given the complexities of both the compression scheme and application processing.

We have examined some effects of lossy compression of HSI data for data products and background classification. The compression scheme consists of spectral compression via PCA/KLT followed by spatial compression of the PC images via JPEG2000. Data products considered were the normalized difference vegetation index (NDVI) and integrated column water vapor (CWV). To assess compression-induced changes in classification results, we proposed two statistical tests based on kappa analysis; the tests allow us to limit the probability of overcompression error, which we deem more serious than an undercompression error.

The experimental results demonstrated the importance of spatial rate allocation among the PC images. The equal-rate allocation method does not consider distortion at all. It simply compresses a selected number of PC images at the same spatial compression ratio. In contrast, the MSD-optimal method considers the average squared distortion over all spectral bands, not the distortion in each individual spectral band. Although it attempts to produce the smallest RMSD between the noncompressed and decompressed data, it may do so at the expense of increased distortion in some bands.

The MSD-optimal allocation method produced the smallest RMSD and MAD between the noncompressed and decompressed data at all compression ratios. In addition, this method produced the smallest MSAD at lower (< 200) compression ratios. At high compression ratios (> 200), the MSD-optimal method retained only 3–5 PC images, and the equal-rate method slightly outperformed it.

With regard to NDVI, the MSD-optimal allocation method yielded smaller errors than the equal-rate allocations did. For CWV, it displayed the best results at lower compression ratios (50), but at high compression ratios (200), the equal-rate method performed better when a sufficient number of PC images were retained. This behavior is explained by recognizing that CWV is calculated from water-absorption bands, which have low signal power and therefore can incur significant spectral distortion under the MSD-optimal allocation method at high compression ratios.

For classification, the MSD-optimal method consistently produced significantly better results than equal-rate allocations. Remarkably, compression did not significantly change classification results even at overall compression ratios as high as 100–340. We note that the spectral library was derived from in-scene data and thus represented a kind of "best-case" library. Also, background classification is perhaps not as demanding as some other HSI applications. Lossy compression and background classification are somewhat alike in the sense that compression tends to preserve coarse features and discard fine ones, and classification tends to classify together spectra that are roughly similar but might differ in their fine details. However, for a similar compression scheme, Shen and Kasner [26] reported "no appreciable degradation" for anomaly detection and material identification at compression ratios up to 48, the highest they tested.

In summary, these results show that HSI data possesses a large amount of redundancy, which can be exploited by lossy compression to produce high compression ratios without significantly degrading appli-

cation performance. Rate allocation is an important topic, since it determines how the limited resource of storage or transmission bandwidth is consumed. For the applications tested, the MSD-optimal allocation method generally outperformed the simple equal-rate method. Finally, statistical tests are encouraged as a means for evaluating the effects of lossy compression on application results and protecting against an overcompression error. The example application of background classification proposed two such tests.

Some avenues for future research follow naturally from this study. Although we have examined some data products and classification, another important area of interest is the effect of lossy compression on the detection of targets or anomalies in hyperspectral data. This topic is the subject of another ongoing study. Regarding rate allocation, the methods employed here are both standard techniques in lossy compression. An area of high potential benefit is the development of other allocation methods that consider the importance of different spectral bands. Statistical tests to assess effects of compression on applications can also be developed for other applications. For detection applications, tests could use a measure such as the area under the receiver operating characteristic (ROC) curve. For quantitative data products, standard tests such as the t-test [12, Ch. 7] could be employed.

A. RATE ALLOCATION

Assume there are N independent sources with respective distortion-rate functions $D_i(R_i)$, $i = 1, 2, \ldots, N$. Let the desired rate be R_{des}. For the allocation $\mathbf{R} = (R_1, R_2, \ldots, R_N)$, the average distortion is

$$D(\mathbf{R}) = D(R_1, R_2, \ldots, R_N) = \frac{1}{N}\sum_{i=1}^{N} D_i(R_i),$$

and the rate is

$$R(\mathbf{R}) = \frac{1}{N}\sum_{i=1}^{N} R_i.$$

The rate allocation problem is then

$$\min_{\mathbf{R}} D(\mathbf{R}) = \min_{\{R_i\}} \frac{1}{N}\sum_{i=1}^{N} D_i(R_i) \text{ such that } R(\mathbf{R}) = \frac{1}{N}\sum_{i=1}^{N} R_i = R_{\text{des}}. \tag{41}$$

Equation (41) is a classic constrained optimization problem, which can be solved by applying the Lagrange multiplier method. Construct the Lagrangian cost function $J(\mathbf{R})$,

$$J(\mathbf{R}) = D(\mathbf{R}) + \eta\left[R(\mathbf{R}) - R_{\text{des}}\right] = \frac{1}{N}\sum_{i=1}^{N} D_i(R_i) + \eta\left[\frac{1}{N}\sum_{i=1}^{N} R_i - R_{\text{des}}\right].$$

Setting $\partial J/\partial R_i = 0$ for each i yields the equal-slope or Pareto condition

$$\frac{\partial D_i}{\partial R_i} = -\eta^*, \quad \forall i, \tag{42}$$

where the optimum Lagrange multiplier value $\eta^* > 0$ is chosen so that $(1/N)\sum_{i=1}^{N} R_i = R_{\text{des}}$.

For a scalar Gaussian source with variance σ^2 and MSD distortion, $D(R) = \sigma^2 2^{-2R} = \sigma^2 e^{-R\ln 4}$ [5]. For other sources, $D(R)$ is often approximated as $D(R) \approx C\sigma^2 e^{-\gamma R}$, $R \geq 0$, where C and γ are positive constants [30]. In the remaining discussion, assume that

$$D_i(R_i) = C_i\sigma_i^2 e^{-\gamma_i R_i}, \quad R_i \geq 0, i = 1, 2, \ldots, N. \tag{43}$$

The Pareto condition (42) gives

$$\frac{\partial D_i}{\partial R_i} = -C_i\sigma_i^2\gamma_i e^{-\gamma_i R_i} = -\eta^*, \quad i = 1, 2, \ldots, N.$$

Solving this equation for R_i gives the solution

$$R_i^* = \max\left\{\frac{1}{\gamma_i}\ln\frac{C_i\sigma_i^2\gamma_i}{\eta^*}, 0\right\}, \quad i = 1, 2, \ldots, N. \tag{44}$$

Convexity of the assumed distortion-rate functions means that η^* can be determined efficiently using a bisection search.

The associated distortions can be calculated by substituting (44) into (43):

$$D_i(R_i^*) = \begin{cases} \eta^*/\gamma_i, & \text{if } R_i^* > 0; \\ C_i\sigma_i^2, & \text{if } R_i^* = 0; \end{cases} \qquad i = 1, 2, \ldots, N.$$

Hence, the sources that are assigned a positive rate have distortions that are proportional to γ_i^{-1}, the reciprocal of their exponential constant γ_i. In the case where all sources have the same exponential constant $\gamma_i = \gamma, \forall\, i$, the sources with positive rates operate at the same distortion, η^*/γ.

REFERENCES

[1] R. Brennan and D. Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educ. and Psych. Meas.*, 41:687–699, 1981.

[2] S. Grace Chang, Bin Yu, and Martin Vetterli. Adaptive wavelet threshold for image denoising and compression. *IEEE Trans. Image Processing*, 9(9):1532–1546, Sep. 2000.

[3] J. Cohen. A coefficient of agreement for nominal scales. *Educ. and Psych. Meas.*, 20:37–46, 1960.

[4] Russell G. Congalton and Kass Green. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. Lewis, 1999.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Interscience, 1991.

[6] G. M. Foody. On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Eng. & Remote Sensing*, 58(10):1459–1460, Oct. 1992.

[7] B.-C. Gao and A. F. H. Goetz. Column atmospheric water vapor and vegetation liquid water retrievals from airborne imaging spectrometer data. *J. Geo. Res.*, 95:3549–3564, 1990.

[8] B.-C. Gao and Y. J. Kaufman. The MODIS near-IR water vapor algorithm, 1997. MODIS Algorithm Technical Background Doc. ID: MOD05 - Total Precipitable Water.

[9] M. K. Griffin and H. K. Burke. Compensation of hyperspectral data for atmospheric effects. *Lincoln Laboratory Journal*, 14(1):29–54, 2003.

[10] D. J. Hall and G. B. Ball. ISODATA: A novel method of data analysis and pattern classification. Technical report, Stanford Research Inst., 1965.

[11] Anil K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, 1989.

[12] Richard A. Johnson. *Miller & Freund's Probability and Statistics for Engineers*. Prentice-Hall, 6th edition, 2000.

[13] James H. Kasner, Ali Bilgin, Michael W. Marcellin, A. Lan, B. V. Brower, Sylvia S. Shen, and Timothy S. Wilkinson. JPEG-2000 compression using 3D wavelets and KLT with application to HYDICE data. In *Imaging Spectrometry VI, Proc. SPIE*, volume 4132, pages 157–166, 2000.

[14] Y. J. Kaufman and B.-C. Gao. Remote sensing of water vapor in the near IR from EOS/MODIS. *IEEE Trans. Geoscience & Remote Sensing*, 30:871–884, 1992.

[15] Y. J. Kaufman and C. Sendra. Algorithm for atmospheric corrections. *Intl. J. Remote Sensing*, 9:1357–1381, 1998.

[16] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977.

[17] Thomas M. Lillesand and Ralph W. Kiefer. *Remote Sensing and Image Interpretation*. John Wiley and Sons, 2000.

[18] Z. Ma and R. L. Redmond. Tau coefficients for accuracy assessment of classification of remote sensing data. *Photogrammetric Eng. & Remote Sensing*, 61(4):435–439, Apr. 1995.

[19] D. Manolakis, D. Marden, J. Kerekes, and G. Shaw. On the statistics of hyperspectral imaging data. In *Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imaging VII, Proc. SPIE*, volume 4381, pages 308–316, 2001.

[20] Mihaela D. Pal, Christopher M. Brislawn, and Steven P. Brumby. Feature extraction from hyperspectral images compressed using the JPEG2000 standard. In *Fifth IEEE Southwest Symp. Image Analysis and Interpretation*, 2002.

[21] Research Systems, Inc. *ENVI User's Guide*, Sep. 2001 edition, 2001.

[22] L. J. Rickard. HYDICE: An airborne system for hyperspectral imaging. In *Imaging Spectrometry of the Terrestrial Environment, Proc. SPIE*, volume 1937, pages 173–179, 1993.

[23] J. A. Saghri, A. G. Tescher, and J. T. Reagan. Practical transform coding of multispectral imagery. *IEEE Signal Processing Magazine*, 12(1):32–43, 1995.

[24] John A. Saghri and Andrew G. Tescher. Impact of lossy compression on the classification of remotely-sensed imagery data. *Journal of Imaging Science and Technology*, 46(6):575–582, Nov./Dec. 2002.

[25] Robert A. Schowengerdt. *Remote Sensing: Models and Methods for Image Processing*. Academic Press, Chestnut Hill, MA, USA, 2nd edition, 1997.

[26] Sylvia S. Shen and James H. Kasner. Effects of 3D wavelets and KLT based JPEG-2000 hyperspectral compression on exploitation. In *Imaging Spectrometry VI, Proc. SPIE*, volume 4132, pages 167–176, 2000.

[27] David S. Taubman and Michale W. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic, 2002.

[28] C. J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.*, 8:127–150, 1979.

[29] J. S. Tyo, J. Robertson, J. Wollenbecker, and R. C. Olsen. Statistics of target spectra in HSI scenes. In *Imaging Spectrometry VI, Proc. SPIE*, volume 4132, pages 306–314, 2000.

[30] Martin Vetterli and Jelena Kovacevic. *Wavelets and Subband Coding*. Prentice-Hall, 1995.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30 December 2004	Project Report	

4. TITLE AND SUBTITLE
Effects of Lossy Compression of Hyperspectral Imagery

5a. CONTRACT NUMBER
F19628-00-C-0002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
J.K. Su, M.K. Griffin, S.M. Hsu, S. Orloff, C.A. Upham

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

HTAP-22

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Department of the Under Secretary of Defense, S&T
Rosslyn Plaza North, Suite 9030
1777 N. Kent Street
Rosslyn, VA 22209

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
ESC-TR-2004-085

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Hyperspectral imaging (HSI) sensors provide imagery with hundreds of spectral bands, typically covering VNIR and/or SWIR wavelengths. This high spectral resolution offers promise for many applications, but it also produces enormous volumes of data, which may be problematic for storage and transmission. Lossy compression may therefore be necessary, but application performance degradation that results from compression is of concern. This report documents results for a spectral-spatial lossy compression scheme and a variety of applications: normalized difference vegetation index (NDVI), integrated column water vapor (CWV), and background classification.

The compression scheme first performs principal-components analysis spectrally, then discards many of the lower-importance principal-component (PC) images, and then applies JPEG2000 spatial compression to each of the individual retained PC images. Two different rate-allocation methods, which select the spatial compression ratios, are considered.

The assessment of compression effects considers general-purpose distortion measures, such as root-mean-square difference. It also examines changes in NDVI and CWV data products and proposes statistical tests for deciding whether compression causes significant degradations in classification results.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified	Same as report	69	