A comparison study of dimension estimation algorithms

Ariel Schlamm,[a] Ronald G. Resmini,[b] David Messinger,[a], and William Basener[c]

[a]Rochester Institute of Technology, Center for Imaging Science, Digital Imaging and Remote Sensing Laboratory, 54 Lomb Memorial Drive, Rochester, NY, 14623, USA

[b]George Mason University, Geography and Geoinformation Science Department, 4400 University Drive, Fairfax, VA, 22030, USA

[c]Rochester Institute of Technology, School of Mathematical Sciences, 85 Lomb Memorial Drive, Rochester, NY, 14623, USA

ABSTRACT

The inherent dimension of hyperspectral data is commonly estimated for the purpose of dimension reduction. However, the dimension estimate itself may be a useful measure for extracting information about hyperspectral data, including scene content, complexity, and clutter. There are many ways to estimate the inherent dimension of data, each measuring the data in a different way. This paper compares a group of dimension estimation metrics on a variety of data, both full scene and individual material regions, to determine the relationship between the different estimates and what features each method is measuring when applied to complex data.

Keywords: hyperspectral, inherent dimension, fractal dimension, correlation dimension, box counting

1. INTRODUCTION

Hyperspectral imagery (HSI) typically has hundreds of bands of spectral data; thus, each HSI data set may have hundreds of dimensions. Much of the information contained in these bands is not unique and as a result, the spectral distribution of the data is contained in a lower dimensional space. Algorithm processing time and results may be improved by taking advantage of the knowledge that the data reside in a lower dimensional space. Principal Components Analysis (PCA) is the most common method for data dimension estimation and reduction.[1] This lossy technique eliminates redundancy in the data but also discards information that may be useful for analysis. Dimension reduction algorithms are limited because they are generally based simply on reducing the number of spectral bands. An alternative to dimension reduction is dimension estimation.[2–4] The inherent dimension of a dataset may be a useful metric for understanding the content and complexity of a scene.

Many of the common techniques used in HSI data analysis, like PCA, typically require that the data meet certain statistical or geometric assumptions. Though often these algorithms do perform well, they do not perform consistently across sensors, diverse scene content, and spatial resolution. This is because the assumptions do not accurately represent the true nature and complexity of hyperspectral data. Other algorithms exist which strive to limit the assumptions about the data, including fractal dimension estimation. Currently, the results from these algorithms applied to hyperspectral imagery are less well understood compared to standard techniques like PCA. For this reason, two different fractal dimension estimation algorithms are compared with PCA dimension estimation and are applied to a large collection of spectral imagery. The remainder of this report is organized as follows. Section 2 describes the dimension estimation algorithms. Section 3 describes the methods used for analysis and comparisons of the dimension estimates. Section 4 describes the data utilized for this research followed by the results given in Section 5.

Further author information:
Ariel Schlamm: aas1510@cis.rit.edu
Ronald Resmini, rresmini@gmu.edu, and also with the MITRE Corporation, rresmini@mitre.org

2. APPROACH

There are many different mathematical dimensions that can be estimated for real data. These include the spanning or encapsulating, intrinsic, topological, and fractal dimensions.[5] For example, the number of endmembers used in a linear spectral mixture analysis is an estimate of the spanning or encapsulating dimension. Alternatively, PCA gives an estimate of the intrinsic dimension. The topological dimension is the dimension of the manifold that the data lie on.[6] The fractal dimension describes how well a fractal data set fills its available space.[7] There are three independent types (and thus estimates) of fractal dimension: capacity, correlation, and information. The three data dimension estimates obtained and analyzed in this study are intrinsic, correlation, and capacity.

PCA performs eigenvalue decomposition on the spectral covariance matrix in order to orthogonalize and decorrelate the hyperspectral data. The eigenvalues are related to the variance of the data in each corresponding eigenvector's direction. Typically, to reduce the dimensionality of the data, a threshold is set on the total percentage of variability, sometimes termed "information," in the eigenvalues.[1] The data corresponding to the largest eigenvalues are retained and the less varying data are discarded from further processing. This results in a smaller data set to pass on to a classification or target detection algorithm which will run faster and in some cases provide better results. Each band in the new, reduced data set is a linear combination of all the originally measured spectral bands. For dimension estimation, the number of eigenvalues that are needed to reach the total percentage of variability threshold is used as the dimension estimate of the data.

Box-counting algorithms estimate the capacity dimension of a dataset.[7,8] These algorithms determine the dimension of the space by counting how many boxes (in the limit obtained by successively decreasing box size) are required to cover the entire space. The number of boxes, N, needed to cover a line segment, or a one-dimensional distribution, is proportional to the length, L, of the line segment divided by the size, ϵ, of the boxes used to cover it. N for a two-dimensional distribution, such as a square, is $\left(\frac{L}{\epsilon}\right)^2$; for a three dimensional distribution, such as a cube, it is $\left(\frac{L}{\epsilon}\right)^3$. This relationship is shown in Figure 1.

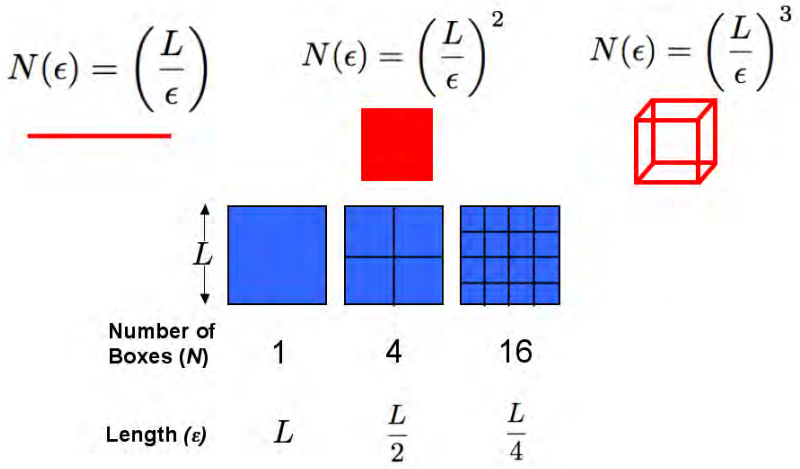

Figure 1. Graphical description of box counting algorithm to determine the capacity dimension

This is extended to a d-dimensional space as,

$$N(\epsilon) = \left(\frac{L}{\epsilon}\right)^{D_{box}}. \tag{1}$$

Approximating the solution for the capacity dimension, D_{box}, results in the expression,

$$D_{box} = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log \frac{1}{\epsilon}}. \tag{2}$$

Notice that the limit in Equation 2 is taken as ϵ approaches zero. In implementation this means reducing the box size (and increasing the number of boxes) from a single box that covers the entire dataset to a sufficiently small box size such that each box contains no more than a single point. To practically solve for D_{box}, the log number of boxes required is obtained by regression against the log reciprocal length of the box for a range of box sizes.

The point density algorithm is a different approach to estimating the correlation dimension.[2,3,9] As the name implies, the point density implementation considers the density of points within small elements, in this case hyperspheres. These algorithms determine the dimension by relating it to the volume, V, of the space. The "volume" of a line or of points on a line, shown in Figure 2(a), is proportional to the length or radius r of that line. The "volume" of a filled circle is the area of the circle, or $\pi r^2 \propto r^2$. The volume of a filled sphere is $\frac{4}{3}\pi r^3 \propto r^3$. It can be shown that the volume of a d-dimensional sphere is proportional to r^d for any dimension, d.[8] A similar relationship, between the volume of a hypercube and the dimension of the hypercube is discussed in Landgrebe.[10] In a uniformly distributed space, the number of points, N_V, in the dataset is proportional to the volume. It follows from this assumption that

$$V \propto N_V \propto r^d \leftrightarrow \log N_V \propto d \log r. \tag{3}$$

The correlation dimension, D_{cor}, can then be approximated by

$$D_{cor} \approx \frac{\log N_v}{\log r} \tag{4}$$

and practically solved for by estimating the slope of the line, shown in the point density plot (PDP) in Figure 2(b), between the log of the number of points within the volumes and the log of the radii of the volumes.

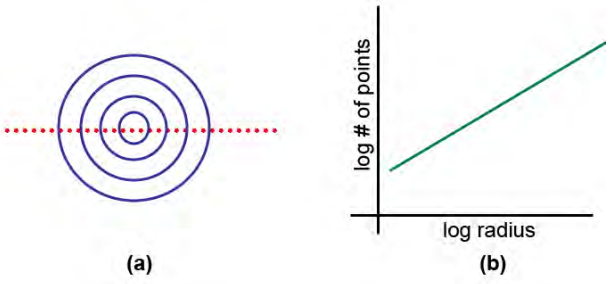

Figure 2. Illustration of counting the number of points within a 2D sphere and the resulting point density graph.

The capacity and correlation dimensions are all related to each other and to the true dimension of the data through the inequalities:

$$D_{cor} \leq D_{box} \leq D, \tag{5}$$

where D is the true dimension of the data.[11] However, this theoretical relationship is not always true empirically, especially when the governing assumptions are violated. Both the capacity and correlation dimension estimation methods depend on the assumption that the data are a regular or uniform finite sample of an infinite space. Grassberger and Procaccia (1983) state that the convergence of the box counting algorithm needs a minimum of 200,000 points before the capacity dimension estimate is accurate. When applied to the same dataset, correlation dimension estimation converges to an accurate result with only a few thousand data points.[9] Point density analysis is more sensitive to deviations from the assumption than box counting. Further analysis can be done on the data from either the PDP or box-counting methodologies. In both cases, particular regions in the plots are linked back to the corresponding pixels in the original image. This feature of these algorithms may be useful in anomaly detection or spectral clustering routines.

3. METHODOLOGY

For this research, a total of 37 hyperspectral images and over 250 subset cubes derived from these images are used (see Section 4). For each image and subset, three dimension values are estimated: correlation, capacity, and principal component dimension. The PC dimension, denoted hereafter as PCA is defined as the number of principal components needed to contain 99.5% of the variability in the scene. The capacity and correlation dimensions, denoted BOX and PDP, are estimated through the box counting and point density analysis described in Section 2.

4. DATA

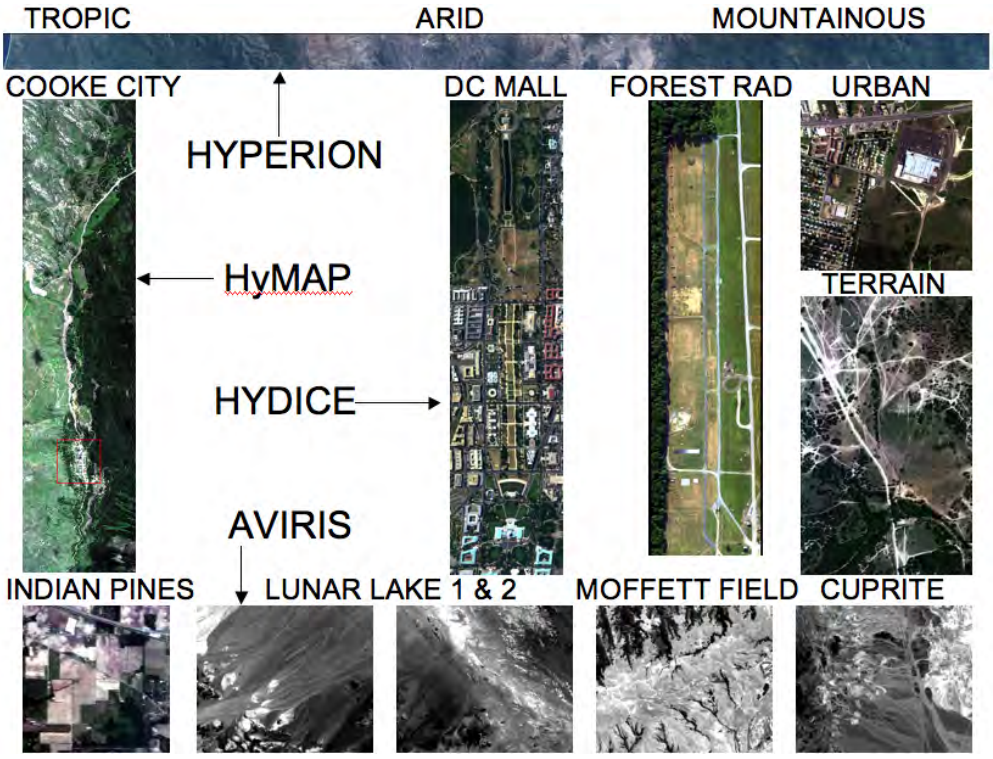

Figure 3. Image chips from the hyperspectral imagery analyzed for dimension estimation.

In order to obtain a broad understanding of what each dimension estimation algorithm is measuring, a large collection of data was assembled. A total of 37 hyperspectral images from four visible/near infrared through shortwave infrared (VNIR/SWIR, $0.4\mu m$ to $2.5\mu m$) sensors of 11 unique scenes (shown in Figure 3) and 7 ground sample distances (GSD) were used for this research. All processing was performed on radiance data, but could easily be done in reflectance or digital counts. Hyperion is the 220 band space-based sensor mounted on the NASA EO-1 satellite with 30 m GSD.[12] AVIRIS is a NASA airborne hyperspectral sensor with 20 m GSD and 224 spectral bands.[13] HyMap is a commerical hyperspectral sensor operated by HyVista with 126 spectral bands and approximately 3 m GSD.[14] HYDICE (Hyperspectral Digital Imagery Collection Experiment) is an airborne sensor (now defunct) which collected a large library of 210 band data in the mid 1990s.[15–17] Table 1 contains information regarding the VNIR/SWIR HSI data analyzed in this study, including the scene content, sensor, number of images from each sensor and for different scene content, GSD, and number of bands. Within a particular scene content (i.e., predominant physical/geographic background), the number of images is the number of flightlines in the collection that cover approximately the same location and share the same scene content. In the case of HYDICE and HyMap data, the multiple flightlines were imaged within hours or days of each other. The four Hyperion images were collected over a period of 10 months and are divided into three independent climate regions: tropical, arid, and mountainous.

Table 1. Hyperspectral Image Data Details

Scene Content	Sensor	Number	GSD (m)	Bands
Forest	HYDICE	15 (5 per GSD)	0.8,1.5,3	170
Desert	HYDICE	3 (1 per GSD)	0.8,1.5,3	170
Washington DC Mall	HYDICE	1	1.5	191
Urban	HYDICE	1	1.5	170
Terrain	HYDICE	1	1.5	170
Cooke City	HyMap	7	3	126
Cuprite, NV	AVIRIS	1	20	220
Lunar Lake, NV	AVIRIS	2	20	220
Moffett Field, CA	AVIRIS	1	20	220
Indian Pines	AVIRIS	1	20	220
Oaxaca Valley, Mexico	Hyperion	4	30	158

5. RESULTS

Multiple comparisons of the dimension estimation results were conducted in this analysis, including within and across sensor platforms, full scenes, material subsets (i.e., subsets of a single data/background class), and GSD. Selected results that highlight the differences and similarities between the different dimension estimates are presented next.

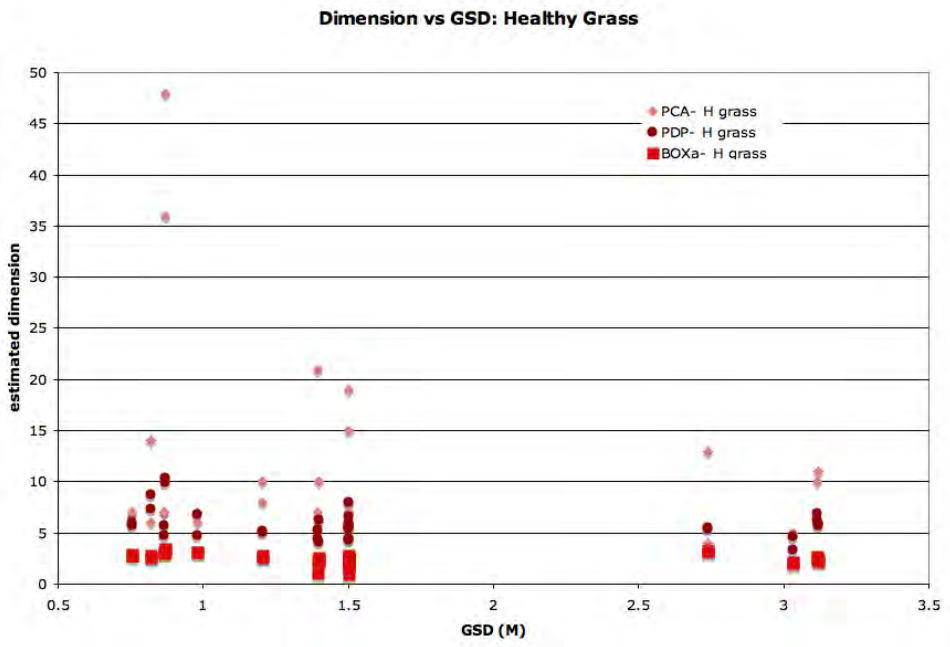

Figure 4. Estimated dimension as a function of GSD for healthy grass subsets of HYDICE data.

Figures 4 and 5 show similar relationships between the three estimated dimensions as a function of GSD for image subset regions of healthy and unhealthy grass from HYDICE data. Each point on the plot represents a single dimension estimate from an image subset from one of 15 images. For each individual image subset, there are 3 points on the plot. In both cases, the PDP and BOX dimension estimates are constant with little variability in the 0.5-3.5m GSD range when compared to PCA results. The PCA dimension estimates are larger and have more variability at small GSDs. Both the magnitude and the variability of this estimate decrease as the GSD increases. For estimating the dimension of single material image subsets, PDP and BOX dimensions will provide more uniform results within and across different GSDs.

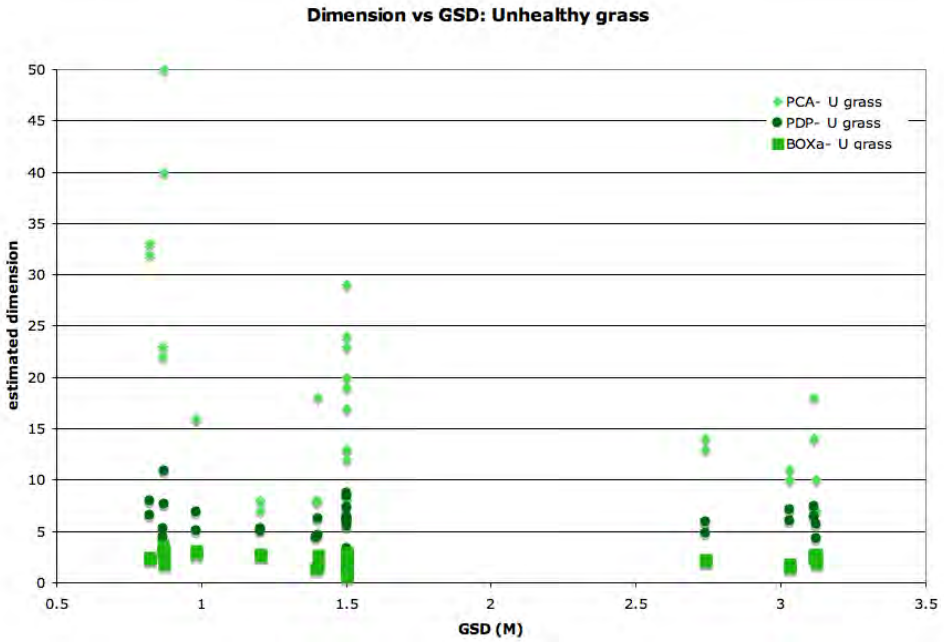

Figure 5. Estimated dimension as a function of GSD for unhealthy grass subsets of HYDICE data.

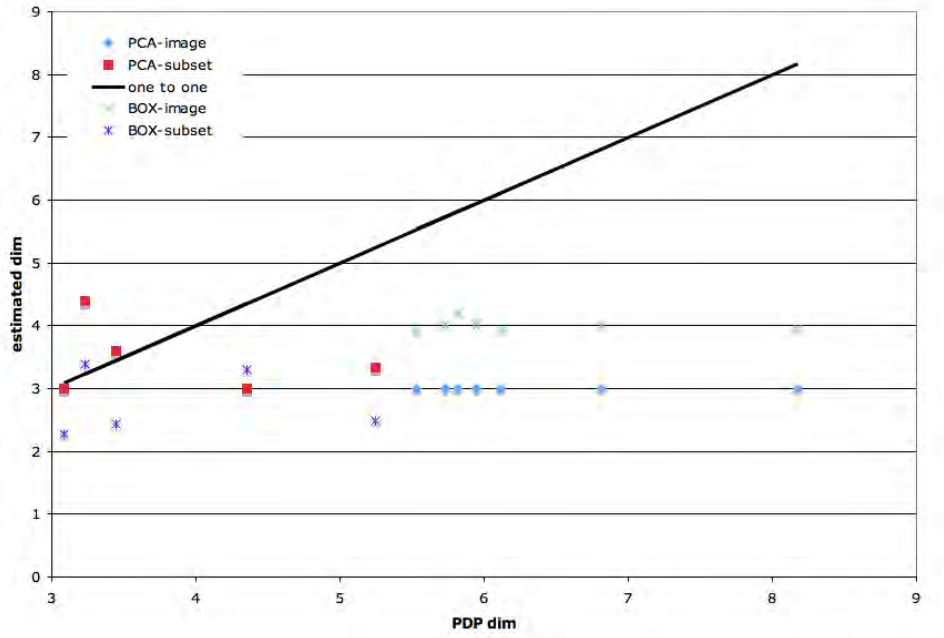

Figure 6. Estimated dimension as a function of PDP dimension for HyMap data.

Figure 6 shows the estimated dimension of HyMap data vs. the PDP dimension (also derived from the HyMap data). The points on the left (labeled "subset") are the average estimated dimensions of seven flightlines for five unique material subsets. The points on the right (labeled "image") are the estimated dimension of seven full flightlines covering the same area. The individual material subsets have no regular pattern across dimension estimates and are dependent on the image content of the subsets. In general, the PCA and PDP dimension estimates are higher than the BOX estimate, except for a single point. These patterns are not observed for the image-wide estimates. The PCA and BOX dimension estimates of full images are consistent across the seven

flightlines. The PDP estimates of full images are consistently larger than both the PCA and BOX dimension estimates, but they are not constant and range from approximately 5.5 to 8.2. Also, the BOX dimension estimate is higher than the PCA dimension estimate—the opposite of what is seen for smaller image subsets. For full image dimension estimation, the BOX and PCA algorithms will provide more consistent results than PDP.

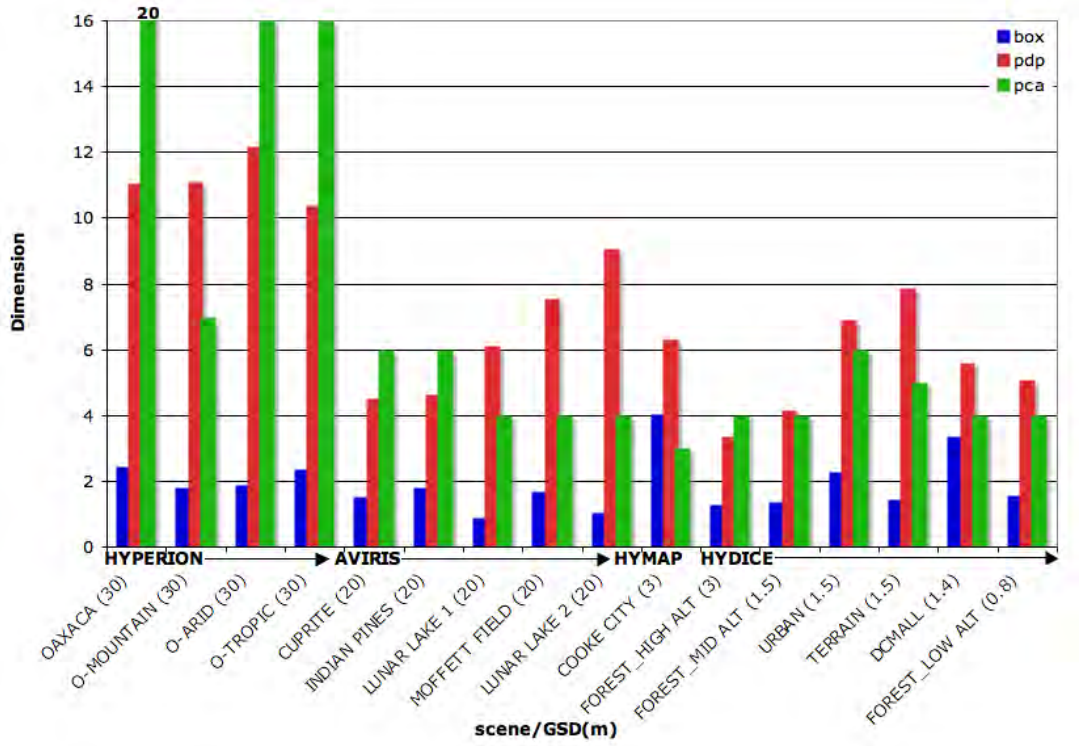

Figure 7. Estimated dimensions of full scene hyperspectral data as a function of sensor, image content, and GSD.

Figure 7 shows the estimated dimension of full images as a function of scene content, sensor, and GSD. In general, large GSD data sets, such as those from Hyperion, have larger dimension estimates than higher spatial resolution data. Some scenes with small GSDs show increased dimension estimates in at least two of the algorithms. These scenes, mainly Moffett Field, Cooke City, Urban, and the DC Mall, all contain urban regions and may be assumed to be more complex spectrally than the scenes containing only natural materials. An increased dimension estimate may indicate increased complexity of—or material variability within—the scene.

6. CONCLUSION

This analysis describes a comparison of three independent dimension estimation algorithms: principal component, point density, and box counting analysis. A large collection of hyperspectral images and image subsets was assembled for use in evaluating the three algorithms. Results were presented across sensor platform, GSD and scene content for full flightlines and individual material image subsets. In general, PCA is not a reliable method for estimating the inherent dimension of a hyperspectral image. The PCA dimension estimates vary significantly within scene content within sensor, especially for individual material subsets, such as grass. The box counting algorithm is best suited for full image dimension estimation as the results are the most reliable and consistent. Point density analysis is most reliable for estimating the dimension of individual material subsets. The main assumption of PDP that the data set is regularly sampled is severely violated when applied to a full hyperspectral scene but significantly less so for uniform material regions. When the BOX and PDP dimension estimates are higher, it indicates a complex scene, possibly containing an urban area. Future research will involve using the box counting and point density analysis in order to estimate the local complexity of hyperspectral image tiles.

ACKNOWLEDGMENTS

The authors wish to thank to Dr. John Kerekes for providing the HyMAP imagery of Cooke City, MT, Dr. William D. Middleton for providing the Hyperion imagery of Oaxaca Valley, Mexico, the Spectral Information Technology Applications Center for providing the HYDICE data and Dr. Emmett Ientilucci for preparing the HYDICE forest and desert imagery.

REFERENCES

[1] Schott, J. R., [*Remote Sensing: The Imaging Chain Approach*], ch. 10.2: Issues of Dimensionality and Noise, Oxford University Press, 2nd ed. (2007).

[2] Schlamm, A., Messinger, D., and Basener, B., "Geometric estimation of the inherent dimensionality of a single material cluster in multi- and hyperspectral imagery," *Proc. SPIE* **6966** (2008).

[3] Schlamm, A., Messinger, D., and Basener, W., "Geometric estimation of the inherent dimensionality of a single and multi-material clusters hyperspectral imagery," *JARS* **3**, 033527 (Feb 2009).

[4] Resmini, R. G., "A tool for the nonparametric characterization of the geometry of spectra in hyperspace," *Proc. SPIE* **7334**, 73341S (2009).

[5] Kirby, M., [*Geometric Data Analysis: An Empiracle Approach to Dimensionality Reduction and the Study of Patterns*], 8–11, John Wiley & Sons, INC (2001).

[6] Bachmann, C. M., Ainsworth, T. L., and Fusina., R. A., "Exploiting manifold geometry in hyperspectral imagery," *IEEE Trans. Geosci. Rem. Sens.* **43**, 441–454 (2005).

[7] Baker, G. L. and Gollub, J. P., [*Chaotic Dynamics: An Introduction*], ch. 5: The characterization of chaotic attractors, Cambridge University Press, New York, first ed. (1990).

[8] Theiler, J., "Estimating fractal dimension," *J. Opt. Soc. Am.* **7**, 1055–1073 (1990).

[9] Grassberger, P. and Procaccia, I., "Characterization of strange attractors," *Phys. Rev. Lett.* **50**, 346–349 (1983).

[10] Landgrebe, D., "Hyperspectral image analysis," *IEEE Signal Process. Mag.* **19**, 17–28 (2002).

[11] Grassberger, P. and Procaccia, I., "Measuring the strangeness of strange attractors," *Phys. Nonlinear Phenom.* **9**, 189–208 (1983).

[12] http://edcsns17.cr.usgs.gov/eo1/hyperion.php.

[13] http://aviris.jpl.nasa.gov/.

[14] http://www.intspec.com/products/HyMap/overview/.

[15] Mitchell, P. A., "Hyperspectral digital imagery collection experiment: Hydice," *Proc. SPIE* **2587**, 70–95 (1995).

[16] Basedow, R. W., Carmer, D. C., and Anderson, M. E., "Hydice system: implementation and performance," *Proc. SPIE* **2480**, 258–267 (1995).

[17] Aldrich, W. S., Kappus, M. E., Resmini, R. G., and Mitchell, P. A., "Hydice postflight data processing," *Proc. SPIE* **2758**, 354–363 (1996).