On Statistical Band Selection for Image Visualization

M. Beauchemin and Ko B. Fung

Abstract

We show that two often-cited statistical band selection methods for image visualization provide significantly different results when applied to the same data set. The cause of the difference is first identified. Then, an alternative method based on the minimization of redundant information between bands is presented. The new measure is robust against the existence of multicollinearity. A procedure to help determine an appropriate band subset size is also proposed.

Introduction

There are only a few remote sensing publications on statistical band selection with the objective to maximize the overall information content for visual interpretation. Textbooks on remote sensing, such as Lillesand and Kiefer (1994) and Harrison and Jupp (1990), refer to a method developed by Chavez *et al.* (1982). A section devoted to that topic in a recent review by Pohl and Van Genderen (1998) reports on only a few others. Weaknesses associated with most of them can be found in Sheffield (1985) and Mausel *et al.* (1990). Sheffield (1985) proposed a method to discourage the selection of pairs of bands with high correlation based on principal component analysis. All these methods are inclusively based on statistics derived from the scene.

The image band selection problem can be stated as follow: identify the image band subset(s) of size p that contain the most information among N available bands. Statistical band selection relies on the fact that not all image bands carry the same amount of information and that adjacent bands in the spectral domain are highly correlated (Tu *et al.*, 1998). The hypothesis is that it is possible to select a subset of bands that retain most of the information of the entire data set with a negligible loss of information, or to select band combinations that contain the highest possible amount of information given a predefined subset size.

We first focus our attention on the methods of Chavez *et al.* (1982) and Sheffield (1985). It is shown that, although the goals are the same, they produce significantly different results when applied to the same data set. The origin of this departure can be understood by establishing a link in their formalisms and by reviewing the procedure used in both studies to validate their results. We then present an alternative index that is, in fact, a normalized version of the Sheffield index. A procedure to help in determining an appropriate band subset size is finally proposed. In the following discussion, the elements of the covariance matrix, **M**, are denoted as m_{ij}. The elements of the correlation matrix, **R**, are denoted as ρ_{ij}, where $\rho_{ij} = m_{ij}\,[m_{ii}\,m_{jj}]^{-1/2}$.

Canada Centre for Remote Sensing, 588 Booth Street, Ottawa K1A 0Y7, Canada (mabeauch@ccrs.nrcan.gc.ca; ko.fung@ccrs.nrcan.gc.ca).

Methods Overview

The methods developed by Chavez *et al.* (1982) and Sheffield (1985) rely on an index devised to rank band subsets according to their information content.

Chavez *et al.*

The Optimum Index Factor (OIF) was introduced by Chavez *et al.* to select a three-band combination that displays the greatest details among a maximum of 20 bands. The index is given by

$$\text{OIF} = \sum_{i=1}^{3} SD_i \,/\, \sum_{j=1}^{3} \text{ABS}(CC_j) \tag{1}$$

where SD_i is the standard deviation of band i and $\text{ABS}(CC_j)$ is the absolute value of the correlation coefficient between any two of the possible three pairs. According to Chavez *et al.*, the highest values of OIF should be the three bands having the most information content. This measure favors the selection of those bands having high variances and low pair-wise correlation. The measure can obviously be extended to any subset of arbitrary size p; then, IOF is defined as

$$\text{OIF} = \sum_{i=1}^{p} m_{ii}^{1/2} \left[\sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \text{ABS}(\rho_{ij}) \right]^{-1}. \tag{2}$$

Sheffield

Sheffield (1985) proposed a method based on the size of the hyperspace spanned by the data bands. The square root of the product of the eigenvalues of the three principal components defines the significant volume spanned by the image bands in the hyperspace (ellipsoid volume for $p = 3$). Sheffield suggests that those bands with the biggest hypervolumes be selected. According to Sheffield, the above approach would discourage the selection of those pairs having high correlation coefficients, the rationale being that highly correlated image band pairs will have the eigenvalue of one of the two image bands close to zero. Therefore, if a highly correlated pair is chosen, the resultant (hyper)volume, which is the product of the eigenvalues, will be small. Because the product of the eigenvalues (principal axis system) is equal to the determinant of the original covariance matrix, it is sufficient to rank in decreasing order the value of the determinant of each p by p sub-matrix generated from the original covariance matrix. The Sheffield index (SI) is given by

$$\text{SI} = |\mathbf{M}_{p\times p}| \tag{3}$$

Photogrammetric Engineering & Remote Sensing
Vol. 67, No. 5, May 2001, pp. 571–574.

0099-1112/01/6705–571$3.00/0

TABLE 1. CHAVEZ *ET AL.* DATASET. COMPARISON OF THE RANK VALUES ASSIGNED BY DIFFERENT METHODS: OIF, SI, AND CI.

Band (ratio) Combination Number*	Rank		
	IOF	SI	CI
1,4,6	1	10	1
1,4,5	2	8	2
1,5,6	3	3	5
3,4,6	4	6	11
1,3,4	5	12	7
2,4,6	6	11	3
3,4,5	7	2	8
2,5,6	8	4	6
1,3,5	9	15	19
4,5,6	10	13	17
3,5,6	11	1	12
1,3,6	12	7	14
2,3,6	13	19	20
2,3,5	14	5	15
1,2,5	15	17	13
2,4,5	16	9	4
1,2,6	17	16	9
2,3,4	18	14	10
1,2,3	19	18	16
1,2,4	20	20	18

*The correspondence between these numbers and TM band ratios are the following: 1 = 4/5, 2 = 4/6, 3 = 4/7, 4 = 5/6, 5 = 5/7, and 6 = 6/7.

TABLE 2. SHEFFIELD (1985) DATA SET. COMPARISON OF THE RANK VALUES ASSIGNED BY DIFFERENT METHODS: OIF, SI, AND CI. ONLY THE FIRST 16 HIGHEST RANKS OBTAINED FROM SI ARE SHOWN.

Band Combinations	Rank		
	SI	IOF	CI
1,4,5	1	11	14
1,5,6	2	8	22
1,3,5	3	12	23
1,4,6	4	24	13
3,4,5	5	22	24
1,5,7	6	1	1
3,5,6	7	18	28
2,4,5	8	27	17
4,5,6	9	19	29
1,3,6	10	25	21
2,5,6	11	20	25
1,2,5	12	16	27
3,4,6	13	29	20
3,5,7	14	5	6
2,4,6	15	32	16
1,6,7	16	2	2

where $|\mathbf{M}_{p\times p}|$ is the determinant of the covariance matrix of subset size p. Assuming that the data are described by an N-dimensional normal distribution, Sheffield (1985) demonstrates that maximizing the determinant is equivalent to selecting the band subset of maximum entropy.

Comparison of the Two Methods

Let us compare these two methods using data published in both papers. Table 1 shows the results based on variances and correlation coefficients published in Chavez *et al.* (1982; their Tables I and II). The first three best combination obtained from OIF ranks 10, 8, and 3 by SI and the first three best combinations obtained from SI ranks 11, 7, and 3 by OIF. There are five combinations with ranks ≤ 10 selected from SI that are within the first ten selected from OIF. Table 2 shows ranking based on Table 2A of Sheffield (1985), where we assumed that $m_{32} = m_{23} = 125.4$. The first three best combinations derived from SI ranks 11, 8, and 12 with IOF and the first three best ones obtained from OIF ranks 6, 16, and 17 with SI. There are only two combinations with ranks ≤ 10 selected from SI that are within the first ten best combinations selected from OIF.

It can be concluded from the results listed in Tables 1 and 2 that OIF and SI applied to the same data sets generate different results. The origin for the difference will be discussed in the next section. Let us first examine the methodology adopted by these authors to verify their ranking approaches. In both studies, the ranking assessment was subjectively checked by visual inspection of RGB color composites of the subsets. All combinations were systematically inspected in the study of Chavez *et al.* and only the best triplets were selected in the case of Sheffield. However, Sheffield refers to the work of Colvocoresses (1983) in which many band and color combinations were considered.

The important point we would like to stress is that, in almost all circumstances, original data are transformed for RGB display to highlight visual contrast in each band. A widely used technique consists of stretching linearly each variable (band) according to the relation $DN' = K\,(DN - min)/(max - min)$, where DN and DN′ are the data values before and after transformation, K is a constant to scale the result to within a given dynamic range (usually $K = 255$), and *min* and *max* are the minimum and maximum histogram bounds (Lillesand and Kiefer, 1994). The *min* and *max* values can be obtained statistically, for example, in terms of distance in units of standard deviation from the average DN. This is the procedure used by Chavez *et al.* in his study. However, under such an operation, each transformed band will have the same variance; in other words, the band variance effect is removed. Therefore, the data inspected visually by Chavez *et al.* to assign the level of information in each triplet is not necessarily the same as the ones used to calculate the OIF values. Applying Equation 1 to the transformed data (DN′), the $\sum_{i=1}^{3} SD_i$ term will be constant for all combinations and only the correlation term in Equation 1 will apply. Note that correlation coefficient values are unaffected under this linear stretching transformation. In Table II of Chavez *et al.*, the first and second ranked subsets do not change rank if only the pair correlation term is considered (third column).

In the case of the Sheffield index (Equation 3), it can be shown that the determinant of the covariance matrix in the stretched system, $\mathbf{M}_{DN'}$, is proportional to the determinant of the correlation matrix in the original domain, $\mathbf{R}_{DN}$. Consequently, the rank assigned to each triplet will be the same for both cases ($\mathbf{M}_{DN'}$ or $\mathbf{R}_{DN}$). This property is linked to the fact that the covariance matrix of standardized variables is equal to the correlation matrix. The latter is sometimes referred to as the normalized covariance matrix.

An Alternative Measure for Band Selection

We propose an alternative index that does not emphasize the individual band variance in the ranking process. The emphasis is on the amount of correlation between band pairs, which reflects the level of complementary information. Although it is straightforward to eliminate the influence of individual band variance by keeping only the correlation term in Equation 2, the way the band pair correlation coefficients are combined in OIF has no statistical basis. On the opposite end, the Sheffield approach offers a tractable theoretical framework. Let us first consider the simplest case, $p = 2$. It can be shown that the determinant of a 2 by 2 covariance matrix $\mathbf{M}$ can be expressed as a function of m_{ij} and ρ_{ij} (Anderson, 1958): i.e.,

$$|\mathbf{M}_{2\times 2}| = m_{ii}\, m_{jj}\,(1-\rho_{ij}^2). \qquad (4)$$

By substituting the off-diagonal elements m_{ij} of the covariance matrix with $\rho_{ij}(m_{ii}\, m_{jj})^{1/2}$, it can be shown after some algebraic manipulations that a similar formulation can be obtained for $p = 3$: i.e.,

$$|\boldsymbol{M}_{3\times3}| = m_{ii}\, m_{jj}\, m_{kk}\,(1 + 2\rho_{ij}\rho_{ik}\rho_{jk} - \rho_{ij}^2 - \rho_{ik}^2 - \rho_{jk}^2). \quad (5)$$

Both Equations 4 and 5 are the product of two functions: one that takes into account the individual band variance information (diagonal elements of the covariance matrix), the other that considers band pair correlation coefficients. Similarly, in Equation 1, the index can be split into two terms. The difference noted above (Tables 1 and 2) comes from the different expressions attributed to each of the two functions. Particularly, the difference in the first terms involving the weights attributed to individual band variances will produce marked differences in ranking.

In regression, if one band is assumed independent and the other one dependent, then ρ_{ij}^2 is referred to as the coefficient of determination and represents the fraction of variance of the independent variable explained by the dependent variable. Note that the interpretation of ρ_{ij}^2 is more straightforward than the ρ_{ij} upon which Equation 1 is based. When the value of ρ_{ij}^2 decreases, i.e., when $(1 - \rho_{ij}^2)$ increases, the level of redundant information between the two variables decreases. Therefore, an index based on this quantity is a logical choice for band selection. Maximizing $(1 - \rho_{ij}^2)$ is equivalent to maximizing the ratio $|\mathbf{M}_{2\times2}| / m_{ii}\, m_{jj}$. By taking the square root of this ratio, $|\mathbf{M}_{2\times2}|^{1/2}$ is proportional to the surface spanned by the data measured in the principal axis system, while $(m_{ii}\, m_{jj})^{1/2}$ determines the extent of the space spanned by the data in the original axis system. The same observation applies for Equation 5: i.e., $|\mathbf{M}_{3\times3}|^{1/2}$ is proportional to the ellipsoidal volume spanned by the data in the principal axis system and $(m_{ii}\, m_{jj}\, m_{kk})^{1/2}$ measures the volume extent spanned by the data in the original axis system. The ratio of these two quantities can be uniquely expressed as a function of the correlation coefficients between bands. This leads to the following generalization for the alternative index:

$$\text{CI} = |\mathbf{M}_{p\times p}| \,/\, \prod_{i}^{p} m_{ii}. \quad (6)$$

The square root of CI is related to the ratio of the hypervolume defined by the data scatter measured in the principal axis system to the hypervolume defined by the data scatter in the original axis system. An intuitive geometrical explanation for Equation 6 is that CI measures how the data subset fills the space defined by the size of its projection on each axis in the original system (bands). In fact, CI is equal to the determinant of the correlation matrix $\mathbf{R}$; $\text{CI} = |\mathbf{R}_{p\times p}|$ with $p \geq 2$. The link with the previous section on band stretching for visualization is obvious. The higher the value of $|\mathbf{R}_{3\times3}|$, the higher the value of the (standardized) ellipsoid volume and the higher the data filling the RGB cube, providing a more colorful image and presumably indicative of higher information content. Essentially, the less the bands are correlated, the greater is the volume they define.

Ranking based on CI is reported in the last column of Tables 1 and 2. Ranking based on CI tends to show agreement with OIF in selecting the first few best combinations: the two highest ranked combinations are the same for both data sets. This confirms the previous observation that individual band variance has a much more limited influence on OIF than on SI.

Statistical Band Selection for Image Classification

Mausel *et al.* (1990) analyzed four separability measures among six classes to determine the best subset of four bands selected from an eight-channel video system for a parametric classification of an agricultural area. The divergence, the transformed divergence, the Bhattacharyya, and the Jeffreys-Matusita distances were considered. The number of possible four-band combinations is 70. The classification accuracy was determined using a supervised maximum-likelihood classification method. Their study shows that both Jeffreys-Matusita and transformed divergence distances predict the best combinations for classification. They have also evaluated band selection based on eigenvector analysis as well as band variance size considerations. They concluded that eigenvector analysis provided a "reasonable alternative" because it selected the sixth best four-channel combination. On the other hand, ranking based on (the sum of) individual variance gives poor results, the best combination being ranked 56th. Inspection of their (Mausel *et al.*, 1990) Tables 1 and 7 reveals that eight of the ten best combinations contain a band that accounts for only about 3 percent of the total variance. According to Mausel *et al.*, ". . . additional future research should consider additional comparisons with algorithms which have shown promise such as those suggested by Chavez *et al.* and Sheffield." Such a comparison is presented in this section. Band selection methods developed for image visualization are applied to the Mausel *et al.* data.

TABLE 3. MAUSEL *ET AL.* (1990) DATASET. COMPARISON OF THE RANK VALUES ASSIGNED BY DIFFERENT METHODS: OIF, SI, $(\Sigma\ |\mathbf{CC}_j|)^{-1}$, AND CI.

Band Combinations (Mausel *et al.*)	Rank				
	Mausel *et al.*	OIF	SI	$[\Sigma\text{ABS}(\mathbf{CC}_j)]^{-1}$	CI
3,4,7,8	1	2	7	2	5
2,4,7,8	2	4	12	3	6
4,6,7,8	3	9	26	4	3
4,5,7,8	4	7	10	6	7
1,4,7,8	5	1	15	1	1
3,4,5,8	6	10	1	19	15
2,4,6,8	7	14	13	13	9
3,4,6,8	8	13	9	12	14
2,4,5,8	9	12	2	20	13
1,4,6,8	10	8	14	5	2

Mausel *et al.* (1990) do not provide the covariance matrix of their data set. However, we estimated the covariance matrix of their data by performing the inverse of the principal component transformation in Tables 6 and 7 of their published paper. Because the measurements in these two tables are rounded off to two decimals, the recovered covariance matrix elements cannot exactly match the original elements that generated Tables 6 and 7. Ranking based on the recovered covariance matrix is shown in Table 3 for OIF, SI, the correlation term in IOF ($[\Sigma\text{ABS}(\mathbf{CC}_j)]^{-1}$), and CI. Combinations having high discriminating power are selected by all indices, even though they are not optimized for such a specific problem. Six of the first ten best combinations based on OIF are contained in the top six best combinations. However, the first best combination (1,4,7,8) ranks fifth for all three indices. Ranking relying uniquely on correlation (the last two columns) give results comparable to OIF in that six ($[\Sigma\text{ABS}(\mathbf{CC}_j)]^{-1}$) or seven ($|\mathbf{R}|$) combinations with assigned ranks less than or equal to 10 are among the ten best combination determined by Mausel *et al.* (1990). Ranking based on SI provides results that are slightly inferior to the other three indices. Three of the first ten best combinations based on SI are contained in the top six best combinations. It is apparent that removing the individual variance factor from $|\mathbf{M}|$ results in a much more pronounced effect on the ranking selection than in the case of IOF. Although the best results are obtained with separability measures, those indices relying mainly on inter-band correlation may constitute a reasonable alternative to predict useful subsets for classification.

TABLE 4. MAXIMUM VALUE OF THE DETERMINANT OF THE CORRELATION MATRIX, MAX($|\mathbf{R}_{p\times p}|$), AS A FUNCTION OF THE DATA SUBSET SIZE, p (MAUSEL *ET AL.* (1990) DATA SET; SEE TEXT).

p	max($\|\mathbf{R}_{p\times p}\|$)
2	0.9995
3	0.8406
4	0.3159
5	0.0985
6	0.0199
7	0.0032
8	0.0003

Band Subset Size Determination

An assumption behind band selection is that a reduced number of bands can minimize redundancy and can "closely" represent the original data set in terms of information content. No procedure is described in Sheffield (1985) and Chavez *et al.* (1982) for selecting the size of the subset that will retain an amount of information close to that in the original data. However, Sheffield (1985) has shown that maximizing the determinant of the covariance matrix is equivalent to selecting the band subset having the maximum volume, as well as the maximum entropy if N-dimensional normal distributions are assumed for the data. Because the determinant of the correlation matrix is equivalent to the determinant of the covariance matrix of the standardized data, the same conclusions apply for $|\mathbf{R}|$. An inspection of the behavior of $|\mathbf{R}|$ as a function of the subset size can therefore serve as a guide for selecting the subset size. Consider the intuitive geometric explanation furnished above. As the subset size increases, if the added variable is highly correlated with some others, the volume of the hypercube in the original domain will increase much faster than the volume of the hyper ellipsoid it "encloses," thus generating a marked decrease in $|\mathbf{R}|$. On the other hand, the less the added variable is correlated with the other variables, the more the value of $|\mathbf{R}|$ remains high. Table 4 presents max($|\mathbf{R}_{p\times p}|$) as a function of p derived from the recovered covariance matrix of Mausel *et al.* (1990). It can be seen that max($|\mathbf{R}_{p\times p}|$) drops rapidly with p. The value of max($|\mathbf{R}_{p\times p}|$) becomes rather small for a subset of more than five bands. This is not far from the choice of Mausel *et al.* (1990). They have restricted their subset size to four bands based on considerations which are not directly related to their data set.

Discussion

An advantage of CI (and SI) over $[\Sigma\text{ABS}(\mathbf{CC}_i)]^{-1}$ and OIF is its robustness against the existence of multicollinearity. Consider the following correlation coefficients ($p = 3$): $\rho_{12} = 0.96$, $\rho_{13} = 0.8$, and $\rho_{23} = 0.6$. The value of $|\mathbf{R}|$ for this specific case is exactly zero (so as for $|\mathbf{M}|$; see Equation 5). In regression, this indicates the existence of multicollinearity, meaning that the variance of one of the three bands can be entirely (100 percent) explained by a linear function of the other two (Edwards, 1979). Such a case remains undetectable under OIF. Another advantage of CI is its physical interpretability in terms of standardized hypervolume size. Moreover, its use for band size determination does not find its equivalent under the other two indices.

There are, however, inherent problems related to that approach, the most important one being that such techniques cannot replace optimum procedures for specific applications such as the one developed in Mausel *et al.* (1990) or Tu *et al.* (1998; and references therein). Different combinations of correlation coefficients may produce identical index values, although they may represent different distribution shapes. Also, like the case using principal components, the rejected bands (or components) that are ranked very low may, nevertheless, have high discriminating power for specific classes. The same conclusion applies for the subset size determination procedure.

Conclusion

Two often-cited statistical band selection methods for image visualization have been compared and shown not to agree when applied to the same data set. This discrepancy is mainly caused by the weight assigned to the individual band variance term in each index. An alternative index based on the minimization of redundant information between bands has been introduced. It is in fact a normalized version of the Sheffield index. The new index is robust against the existence of multicollinearity. A procedure to help determine an appropriate band subset size has also been proposed. Its application on the covariance matrix of Mausel *et al.* (1990) provides satisfactory results. Although values of $p \leq 4$ and $N \leq 8$ were considered in this paper, the method can be extended, in principle, to higher data dimension. However, the number of possible band combinations increases drastically with the number of bands, resulting in unacceptable computational cost. For example, with $N = 64$ and $p = 16$, the number of band combinations to evaluate is on the order of 10^{14}.

Statistical band selection methods for overall information certainly are important tools for reducing subset size to carry out analysis of data sets of high dimension. However, great care must be taken in adopting techniques such as the ones described here. Although helpful, they cannot replace optimization procedures for specific applications.

References

Anderson, T.W., 1958. *An Introduction to Multivariate Statistical Analysis*, Chapter 2, John Wiley and Sons, Inc., New York, N.Y., 374 p.

Chavez, P.S., Jr., G.L. Berlin, and L.B. Sowers, 1982. Statistical method for selecting Landsat MSS ratios, *Journal of Applied Photographic Engineering*, 8(1):23–30.

Colvocoresses, A.P., 1983. Presentation at the Landsat-4 Early Results Symposium, March, Goddard Space Flight Center, Maryland.

Edwards, A.L., 1979. *Multiple Regression and the Analysis of Variance and Covariance*, Chapter 5, W.H. Freeman and Company, San Francisco, California, 212 p.

Harrison, B.A., and D.L.B. Jupp, 1990. *Introduction to Image Processing*, Chapter 13, CSIRO Publications, Melbourne, Australia, 255 p.

Lillesand, T.M., and R.W. Kiefer, 1994. *Remote Sensing and Image Interpretation*, Chapter 7, John Wiley and Sons, Inc, New York, N.Y., 750 p.

Mausel, P.W., W.J. Krambler, and J.K. Lee, 1990. Optimum band selection for supervised classification of multispectral data, *Photogrammetric Engineering & Remote Sensing*, 56(1):55–60.

Pohl, C., and J.L. Van Genderen, 1998. Multisensor image fusion in remote sensing: Concepts, methods and applications, *International Journal of Remote Sensing*, 19(5):823–854.

Sheffield, C., 1985. Selecting band combinations from multispectral data, *Photogrammetric Engineering & Remote Sensing*, 51(6):681–687.

Tu, T.-M., C.-H. Chen, J.-L. Wu, and C.-I. Chang, 1998. A fast two-stage classification method for high-dimensional remote sensing data, *IEEE Transactions on Geoscience and Remote Sensing*, 36(1):182–191.

(Received 07 September 1999; accepted 23 March 2000; revised 16 May 2000).