

Reprinted from

**Symposium on
Machine Processing of
Remotely Sensed Data**

June 21 - 23, 1977

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
77CH1218-7 MPRSD

Copyright © 1977 IEEE
The Institute of Electrical and Electronics Engineers, Inc.

Copyright © 2004 IEEE. This material is provided with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the products or services of the Purdue Research Foundation/University. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

A MEASURE OF RELATIVE NORMALITY FOR LANDSAT DATA MULTIVARIATE DISTRIBUTIONS

ROBERT M. RAY III

Center for Advanced Computation, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801

Often within computer classification of LANDSAT multispectral scanner data into terrain classes, we assume multivariate normality for distributions of data samples within each class to bring to bear on our problem the extensive body of statistical methods applicable for analysis of normally distributed multivariate data. Thus, it would seem desirable to have some means of evaluating, at least relatively, the legitimacy of our assumption of the multivariate normal or Gaussian model.

Now we know from information theory that, over all n -dimensional distributions having identical variance-covariance matrices Σ , the distribution having maximum entropy is multivariate normal. Furthermore, the entropy of the n -dimensional normal distribution associated with the random variable X , $X \sim N(M, \Sigma)$ is given by $H_{\max}(X) = \log_2 \left[\frac{(2\pi e)^{n/2}}{|\Sigma|^{1/2}} \right]$. Hence, $H_{\max}(X)$ represents the limiting value of entropy for any n -dimensional random variable X as its distribution approaches the multivariate normality. To the extent that X is non-normally distributed, $H(X)$ will be less than $H_{\max}(X)$.

While the above result from information theory applies explicitly to continuous distributions, it can be employed for characterization of the relative normality of discrete multivariate distributions as well. Let X be a four-dimensional random variable representing the four spectral reflectance levels quantized by LANDSAT for some large set of N sample observations known to correspond to a particular land cover type, say corn. These N data samples may then be sorted and distributed over K intervals ($K \leq N$) within a four-dimensional lattice with frequencies $\rho_k = n_k/N$, $k = 1, \dots, K$ and $\sum_{k=1}^K \rho_k = 1$. Presumably, many four-dimensional data values will occur several times and hence $K \ll N$.

Now the mean vector M and the variance-covariance matrix Σ for the sample may be

computed directly from the N data values given. Consequently, the limiting value of entropy $H_{\max}(X)$, the entropy of a normal distribution characterized by M and Σ , may be computed as a function of Σ using the formula given above. Now if the elements of M and Σ are computed from data values expressed in units corresponding to uniformly spaced intervals of the four-dimensional data lattice, then a comparable value of actual distribution entropy may be computed as

$$H_{\text{act}}(X) = - \sum_{k=1}^K \rho_k \log_2 \rho_k.$$

To the extent that N is large (say $N > 1000$) and the distribution of X is normal,

$H_{\text{act}}(X)$ will be close to $H_{\max}(X)$, and we may define the relative normality of X as $RN(X) = H_{\text{act}}(X)/H_{\max}(X)$.

The same concepts employed in defining our measure of relative normality for multivariate distributions appear also well equipped for treatment of certain problems central within multivariate cluster analysis methodologies. Our paper discusses and illustrates the utility of this measure of relative normality in devising a heuristic for cluster splitting and cluster merging within LANDSAT data cluster analysis applications.