Mahalanobis Distances:

Identifying Suitable Areas based on Multiple Datasets with Complex Inter-Relationships

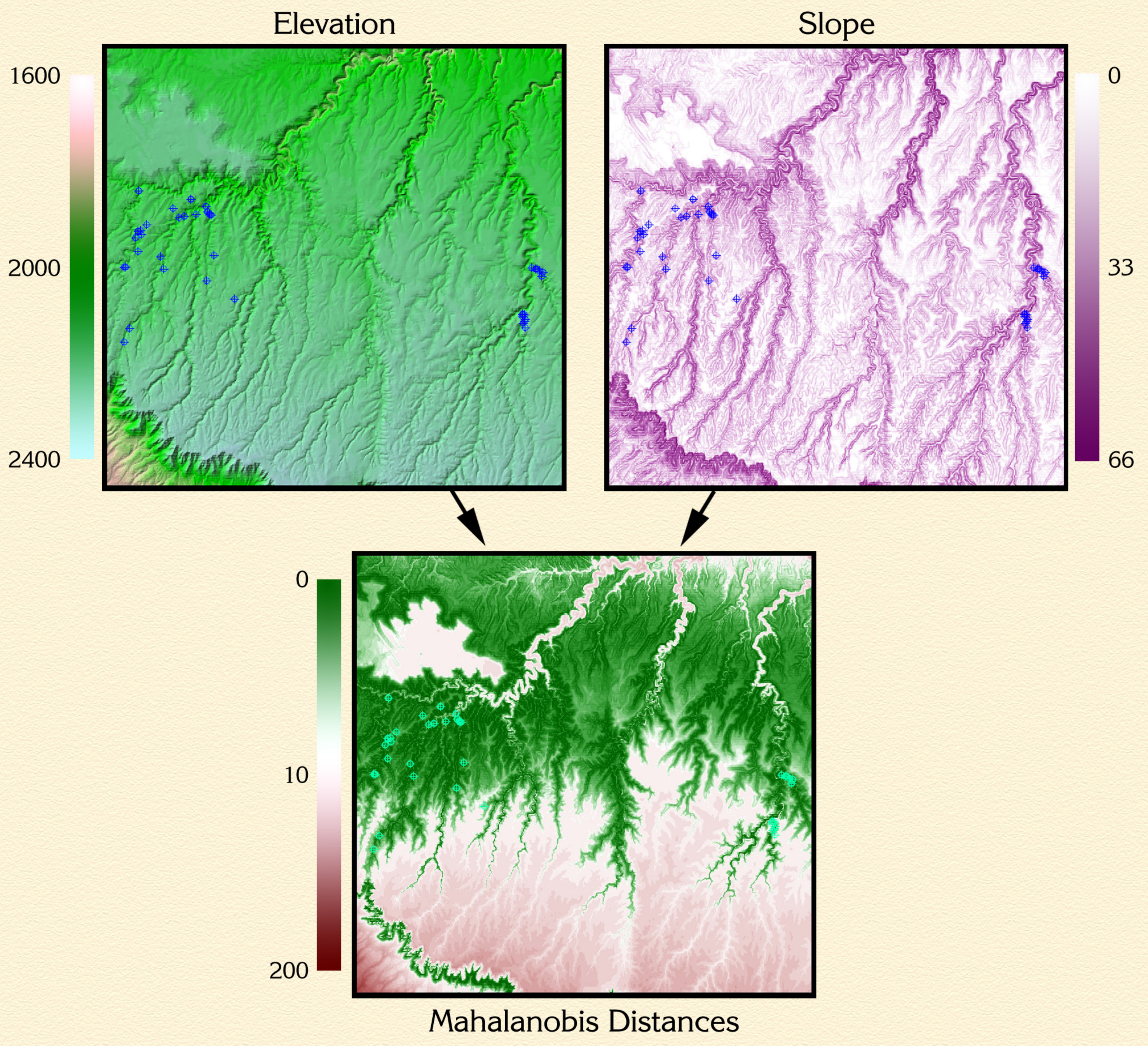
Award-Winning Software
1st Place
Best Avenue Application
2004 ESRI International Users Conference
San Diego, California

Introduction

Mahalanobis distances provide a powerful method of measuring how similar some set of conditions are to an ideal set of conditions, and can be very useful for identifying which regions in a landscape are most similar to some "ideal" landscape.

For example, in the field of wildlife biology we might define an "ideal" landscape as that which best fits the niche of some wildlife species. Through observation, we may find that a wildlife species typically occurs within a particular elevation range, on slopes of a particular steepness, and perhaps within a certain vegetation density. Using Mahalanobis distances, we can quantitatively describe the entire landscape in terms of how similar it is to the ideal elevation, slope and vegetation density of that animal.

Moreover, Mahalanobis distances are based on both the mean and variance of the predictor variables, plus the covariance matrix of all the variables, and therefore take advantage of the covariance between variables. For example, suppose that our hypothetical species likes steep slopes at low elevations and shallow slopes at high elevations. This example implies a certain covariance relationship between slope and elevation, and the Mahalanobis distance for that sample will be based on that covariance relationship. The region of constant Mahalanobis distance around the mean forms an ellipse in 2D space (i.e. when only 2 variables are measured), or an ellipsoid or hyperellipsoid when more variables are used.



Methods

Mahalanobis distances are calculated as: $D^2 = (\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})$

where:

D^2 = Mahalanobis distance

$\mathbf{x}$ = Vector of data

$\mathbf{m}$ = Vector of mean values of independent variables

$\mathbf{C}^{-1}$ = Inverse Covariance matrix of independent variables

$\mathbf{T}$ = Indicates vector should be transposed

Suppose we took a single observation from a bivariate population with Variable X and Variable Y, and that our two variables had the following characteristics:

Variable X: mean = 500, SD = 79.32
Variable Y: mean = 500, SD = 79.25

Variance/Covariance Matrix		
	X	Y
X	6291.55737	3754.32851
Y	3754.32851	6280.77066

If, in our single observation, X = 410 and Y = 400, we would calculate the Mahalanobis distance for that value as:

Given that Mahalanobis Distance $D^2 = (\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})$

$$(\mathbf{x}-\mathbf{m}) = \begin{pmatrix} 410-500 \\ 400-500 \end{pmatrix} = \begin{pmatrix} -90 \\ -100 \end{pmatrix}$$

$$\mathbf{C}^{-1} = \begin{pmatrix} 6291.55737 & 3754.32851 \\ 3754.32851 & 6280.77066 \end{pmatrix}^{-1} = \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix}$$

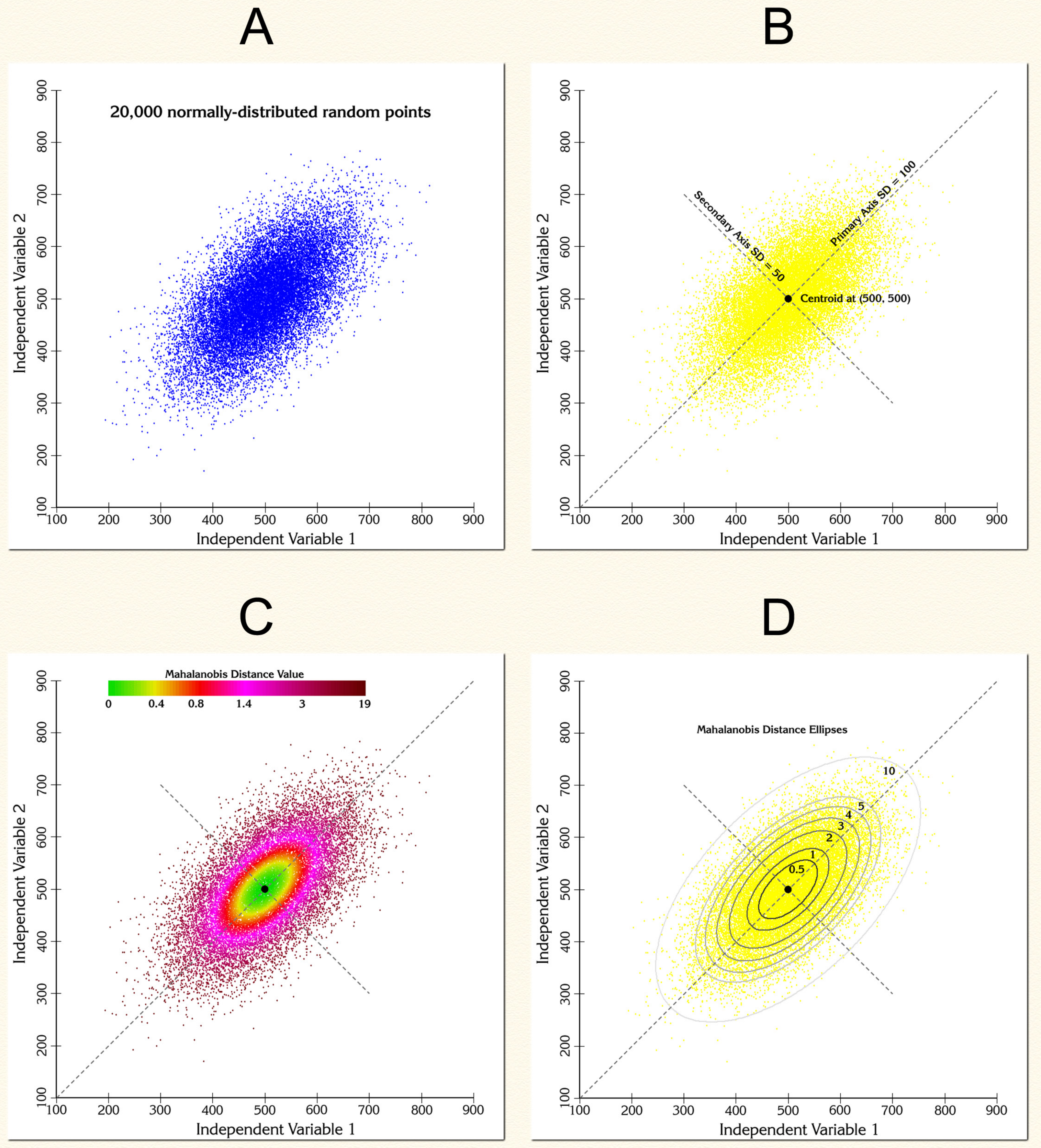
$$\text{Therefore } D^2 = \begin{pmatrix} -90 & -100 \end{pmatrix} \times \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix} \times \begin{pmatrix} -90 \\ -100 \end{pmatrix}$$

$$= 1.825$$

Therefore, our single observation would have a distance of 1.825 standardized units from the mean (mean is at X = 500, Y = 500).

If we took many such observations, graphed them and colored them according to their Mahalanobis values, we can see the elliptical Mahalanobis regions come out. For example, the cloud of data points in Figure A to the right are randomly generated from the bivariate population described above. These points are actually normally distributed along 2 major axes of variance,, with the standard deviation along the primary axis set to twice the standard deviation on the secondary axis (Figure B).

If we calculate Mahalanobis distances for each of these points and shade them according to their distance value, we see clear elliptical patterns emerge (Figure C): Figure D includes actual ellipses drawn at constant Mahalanobis distances. Notice that a Mahalobis Distance of 1 equals 1 standard deviation along both major axes of variance.



Applications to Landscape Analysis

A nice feature of ArcView Spatial Analyst is that we can use actual grids in the Mahalanobis Distance equation rather than numbers, so we can input a vector of habitat grids in place of the vector of input values. We still need the vector of mean values and the covariance matrix, but Spatial Analyst will treat each of these values as an individual landscape-scale grid of that value, and therefore the mathematical functions in Spatial Analyst will work correctly and produce a final grid of Mahalanobis values. Due to a limitation in Spatial Analyst for ArcView 3.x, however, we are limited to 8 input grids for this analysis. Spatial Analyst v. 9 is supposed to fix this limitation.

This Mahalanobis Distances ArcView extension is available for free download at:
http://www.jennessent.com/arcview/mahalanobis.htm

The author recommends Clark et al. (1993), Knick & Dyer (1997), and Farber & Kadmon (2002) for a few good papers illustrating the use of Mahalanobis distances in ecological applications. For anyone interested in the details of matrix algebra and computational/statistical algorithms, the author recommends Conover (1980), Neter et al. (1990), Golub and Van Loan (1996), Draper and Smith (1998), Meyer (2000) and Press et al. (2002).

REFERENCES:

Clark, Joseph D., Dunn, James E., and Smith, Kimberly G. 1993. A multivariate model of female black bear habitat use for a geographic information system. Journal of Wildlife Management. 57:519-526.

Conover, W. J., 1980. Practical nonparametric statistics, 2nd Ed. John Wiley and Sons. 493 p.

Draper, Norman R. and Smith, Harry. 1998. Applied Regression Analysis, 3rd Ed. Wiley Series in Probability and Statistics. 706 p.

Farber, Oren. and Kadmon, Ronen. 2002. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. Ecological Modelling. Elsevier Science. 160:115-130.

Golub, Gene H. and Van Loan, Charles F. 1996. Matrix computations, 3rd Ed. Johns Hopkins University Press. 694 p.

Knick, Steven T. and Dyer, Deanna L. 1997. Distribution of black-tailed jackrabbit habitat determined by gis in southwestern Idaho. Journal of Wildlife Management. 61:75-85.

Meyer, Carl D. 2000. Matrix analysis and applied linear algebra. Society for Industrial and Applied Mathematics. Philadelphia. 718 p.

Neter, John., Wasserman, William., and Kutner, Michael H. 1990. Applied Linear Statistical Models, 3rd Ed. 1181 p.

Press, William H.; Teukolsky, Saul A.; Vetterling, William T., and Flannery, Brian P. 2002 Numerical recipes in C: the art of scientific computing. 2nd ed. Cambridge: Cambridge University Press; 994 pages.

Jeff Jenness, GIS Analyst/Wildlife Biologist

USDA Forest Service, Rocky Mt. Station
2500 S. Pine Knoll Drive.
Flagstaff, Arizona, USA 86001
jjenness@fs.fed.us
(928) 556-2012

Jenness Enterprises
3020 N. Schevene Blvd.
Flagstaff, Arizona, USA 86004
jeffj@jennessent.com
(928) 607-4638