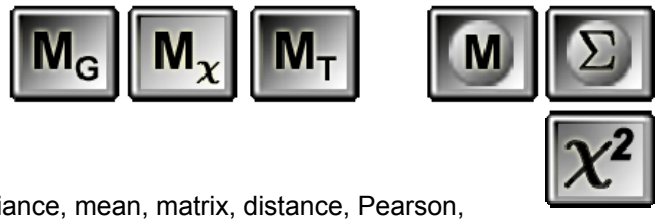
NAME: Mahalanobis Distances



Aka: mahalanobis.avx

Last modified: December 9, 2003

TOPICS: Mahalanobis, statistic, correlation, covariance, mean, matrix, distance, Pearson, Spearman, rho

AUTHOR: Jeff Jenness
Wildlife Biologist, GIS Analyst
Jenness Enterprises
3020 N. Schevene Blvd.
Flagstaff, AZ 86004 USA

jeffj@jennessent.com

DESCRIPTION: This extension produces several possible Mahalanobis-related outputs:

The View button  generates a Mahalanobis surface grid based on independent variable grids.

The View button  recodes a Mahalanobis surface grid into a *p*-value grid based on a Chi-Square distribution with *n-1* degrees of freedom, where *n* = # independent variable grids used to generate the Mahalanobis grid.

The View button  generates a Results table of Mahalanobis distances for each feature in a point, line or polygon theme.

The View tool  allows you to click on a Mahalanobis surface grid cell and calculate the Chi-square *p*-value for that cell. The *p*-values are reported in a text window.

The Table button  generates a new field in the current table containing Mahalanobis distances for each record in the table.

The Table button  generates tables containing mean vectors, covariance matrices, inverse covariance matrices, and correlation matrices for numeric fields in the current table.

REQUIRES: ArcView 3.x and Spatial Analyst. The extension will not load if Spatial Analyst is not present.

This extension also requires that the file "avdlog.dll" be present in the ArcView/BIN32 directory (or $AVBIN/avdlog.dll) and that the Dialog Designer extension be located in your ArcView/ext32 directory, which they usually are if you're running AV 3.1 or better. The Dialog Designer doesn't have to be loaded; it just has to be available. If you are running AV 3.0a, you can download the appropriate files for free from ESRI at:

http://www.esri.com/software/arcview/extensions/dialog/index.html

Recommended Citation Format: For those who wish to cite this extension, the author recommends something similar to:

Jenness, J. 2003. Mahalanobis distances (mahalanobis.avx) extension for ArcView 3.x, Jenness Enterprises. Available at: http://www.jennessent.com/arcview/mahalanobis.htm.

Please let me know if you cite this extension in a publication (jeffj@jennessent.com). I will update the citation list to include any publications that I am told about.

Discussion of Mahalanobis Distances:

General Concepts:

Mahalanobis distances provide a powerful method of measuring how similar some set of conditions is 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 among variables. 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.

Mahalanobis distances are calculated as:

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

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

For example, 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 single value as:

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

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

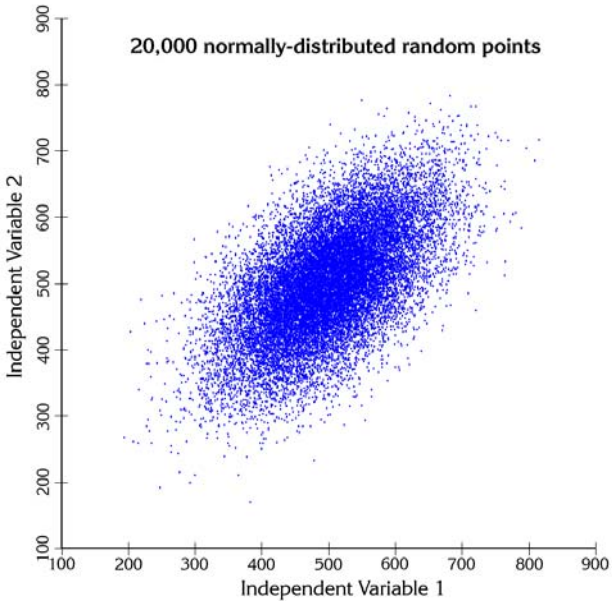
$$\mathbf{C}^{-1} = \begin{pmatrix} 6291.55737 & 3754.32851 \\ 3754.32851 & 6280.77066 \end{pmatrix}^{-1} = \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix}$$

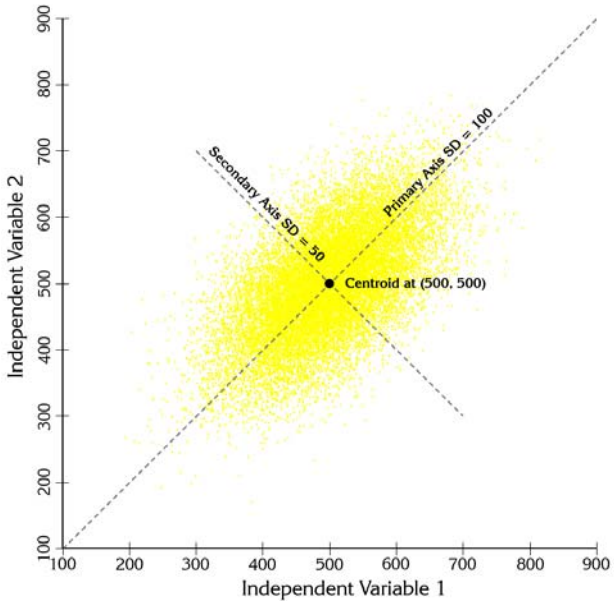
$$\text{Therefore } D^2 = \begin{pmatrix} -90 & -100 \end{pmatrix} \times \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix} \times \begin{pmatrix} -90 \\ -100 \end{pmatrix}$$

$$= 1.825$$

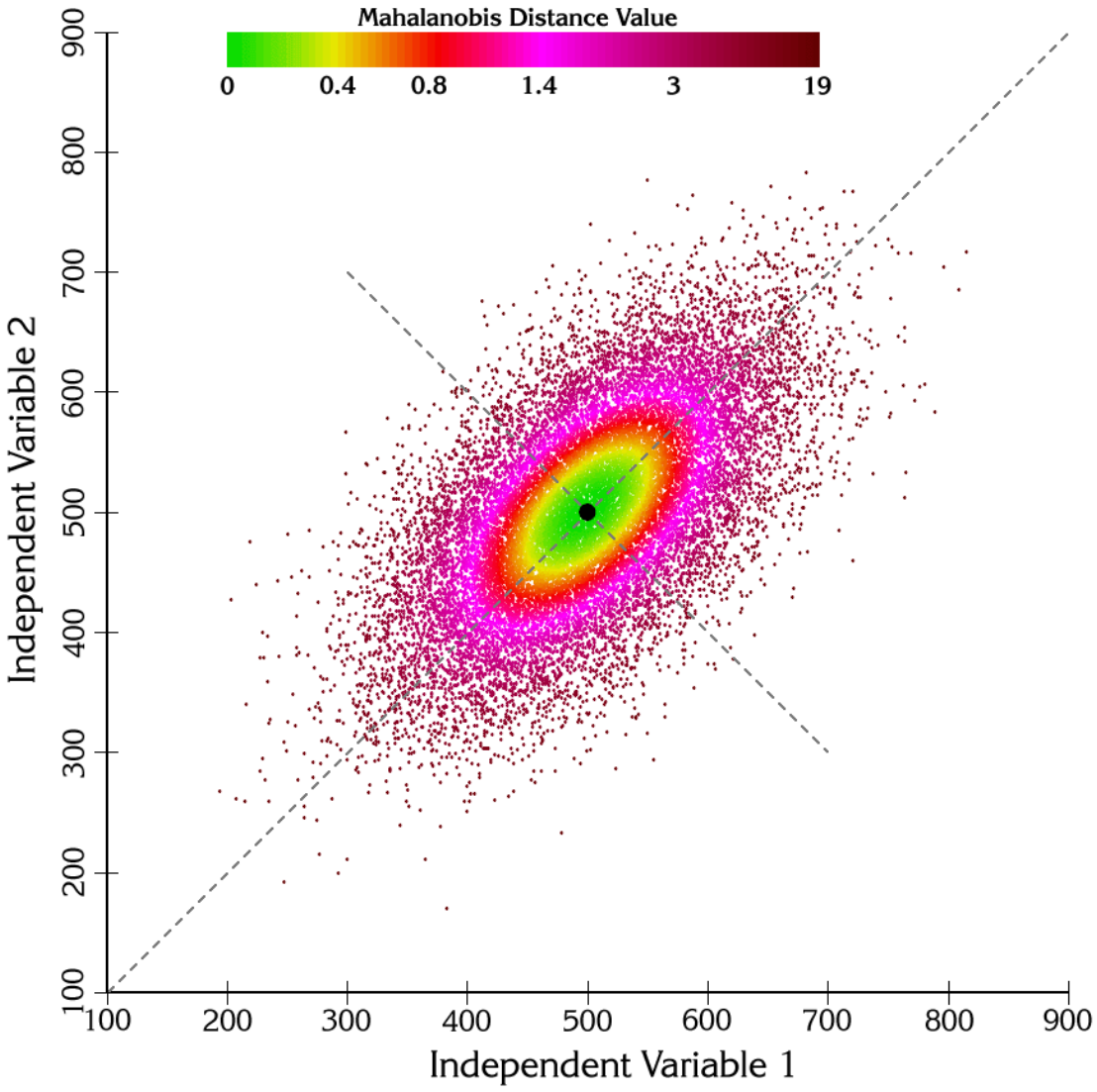
Therefore, our single observation would have a distance of 1.825 standardized units from the mean (mean is at X = 500, Y = 500).

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 below are randomly generated from the bivariate population described above:

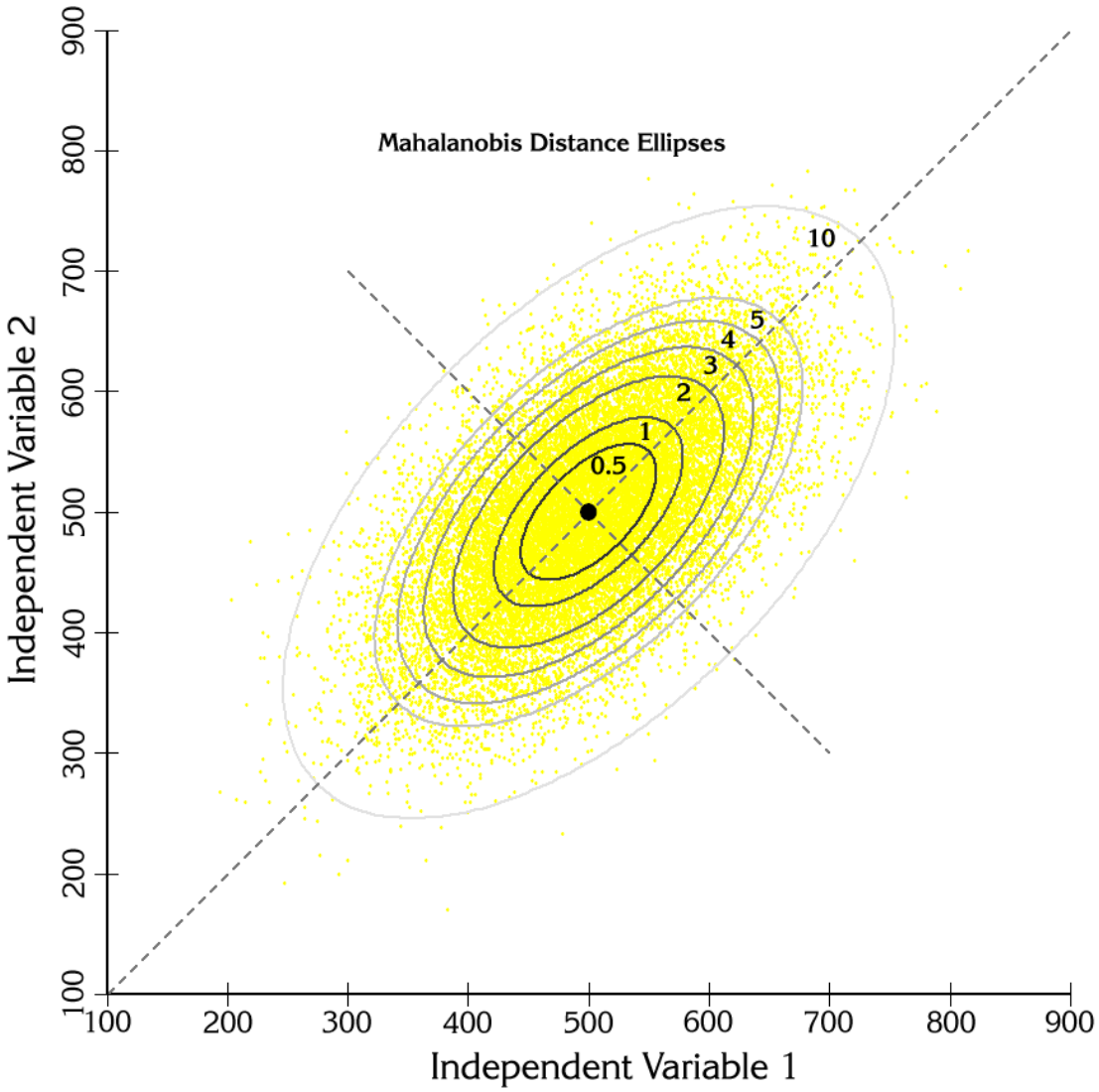




If we calculate Mahalanobis distances for each of these points and shade them according to their distance value, we see clear elliptical patterns emerge:



We can also draw actual ellipses at regions of constant Mahalanobis values:



One interesting feature to note from this figure is that a Mahalanobis distance of 1 unit corresponds to 1 standard deviation along both primary axes of variance.

Chi-Square P-values:

Mahalanobis distances are occasionally converted to Chi-square *p*-values for analysis (see Clark et al. 1993). When the predictor variables are normally distributed, the Mahalanobis distances do follow the χ^2 distribution with *n - 1* degrees of freedom (where *n* = # of habitat variables; 2 in the example above). However, Farber and Kadmon (2003) warn that wildlife habitat variables often fail to meet the assumption of normality. In cases where the predictor variables are not normally distributed, the conversion to Chi-square *p*-values serves to recode the Mahalanobis distances to a 0-1 scale. Mahalanobis distances themselves have no upper limit, so this rescaling may be convenient for some analyses.

In general, the *p*-value reflects the probability of seeing a Mahalanobis value as large or larger than the actual Mahalanobis value, assuming the vector of predictor values that produced that Mahalanobis value was sampled from a population with an ideal mean (i.e. equal to the vector of

mean predictor variable values used to generate the Mahalanobis value). *P*-values close to 0 reflect high Mahalanobis distance values and are therefore very dissimilar to the ideal combination of predictor variables. *P*-values close to 1 reflect low Mahalanobis distances and are therefore very similar to the ideal combination of predictor variables. The closer the *p-value* is to 1, the more similar that combination of predictor values is to the ideal combination.

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, however, we are limited to 8 input grids for this analysis. Spatial Analyst v. 9 is supposed to fix this limitation.

For example, suppose we have a grid of elevation values and a grid of slope values, and we are interested in identifying those regions on the landscape that have similar slopes and elevations to a mean slope and elevation preferred by some species of interest. Furthermore, we want to analyze the slope and elevations in combination so that if our species likes steep slopes at low elevations but shallow slopes at high elevations, then we won't inadvertently select steep slopes at high elevations or shallow slopes at low elevations.

Assume that the niche of our species of interest can be described in terms of Elevation and Slope with the following parameters:

$$\text{Vector of Mean Values} = \begin{pmatrix} \text{Elevation} = 2121 \\ \text{Slope} = 18 \end{pmatrix} \qquad \text{Covariance Matrix} = \begin{pmatrix} 1931 & -54 \\ -54 & 87 \end{pmatrix}$$

We can then enter the Elevation and Slope grids directly into the Mahalanobis equation to produce a Mahalanobis grid:

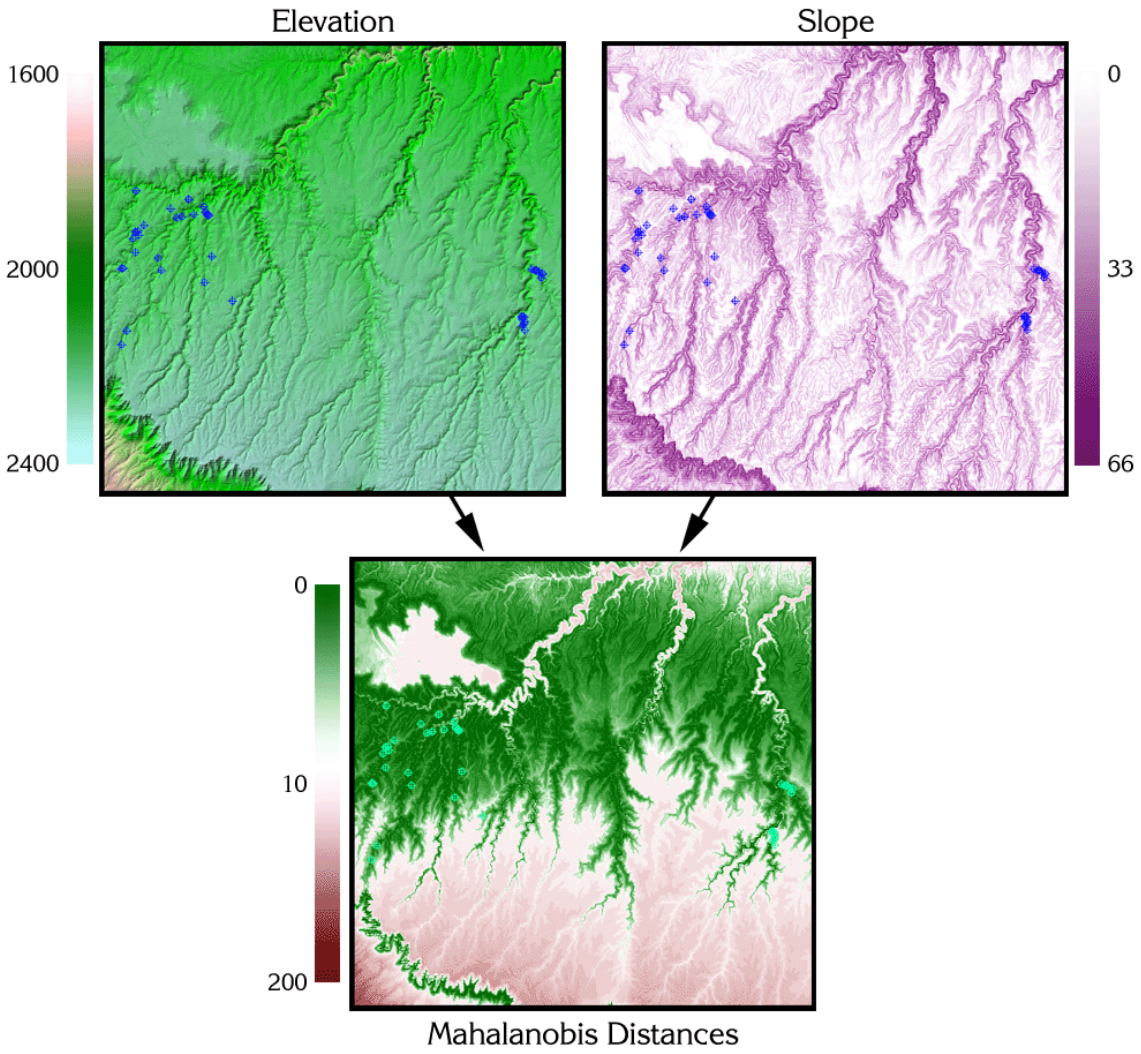
Given that $D^2 = (\mathbf{x}-\mathbf{m})^{\mathbf{T}}\,\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})$

$$(\mathbf{x}-\mathbf{m}) = \begin{pmatrix} [\text{Elevation Grid}] - 2121.41667 \\ [\text{Slope Grid}] - 18.18997 \end{pmatrix}$$

$$\mathbf{C}^{-1} = \begin{pmatrix} 1931 & -54 \\ -54 & 87 \end{pmatrix}^{-1} = \begin{pmatrix} 0.00074 & 0.00046 \\ 0.00046 & 0.01173 \end{pmatrix}$$

$$\text{Therefore } D^2 = \begin{pmatrix} [\text{Elevation Grid}] - 2121.41667 \\ [\text{Slope Grid}] - 18.18997 \end{pmatrix}^{\mathbf{T}} \times \begin{pmatrix} 0.00074 & 0.00046 \\ 0.00046 & 0.01173 \end{pmatrix} \times \begin{pmatrix} [\text{Elevation Grid}] - 2121.41667 \\ [\text{Slope Grid}] - 18.18997 \end{pmatrix}$$

$$= [\text{Mahalanobis Distance Grid}]$$



Additional Reading:

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).

Using the Mahalanobis Distances Extension:

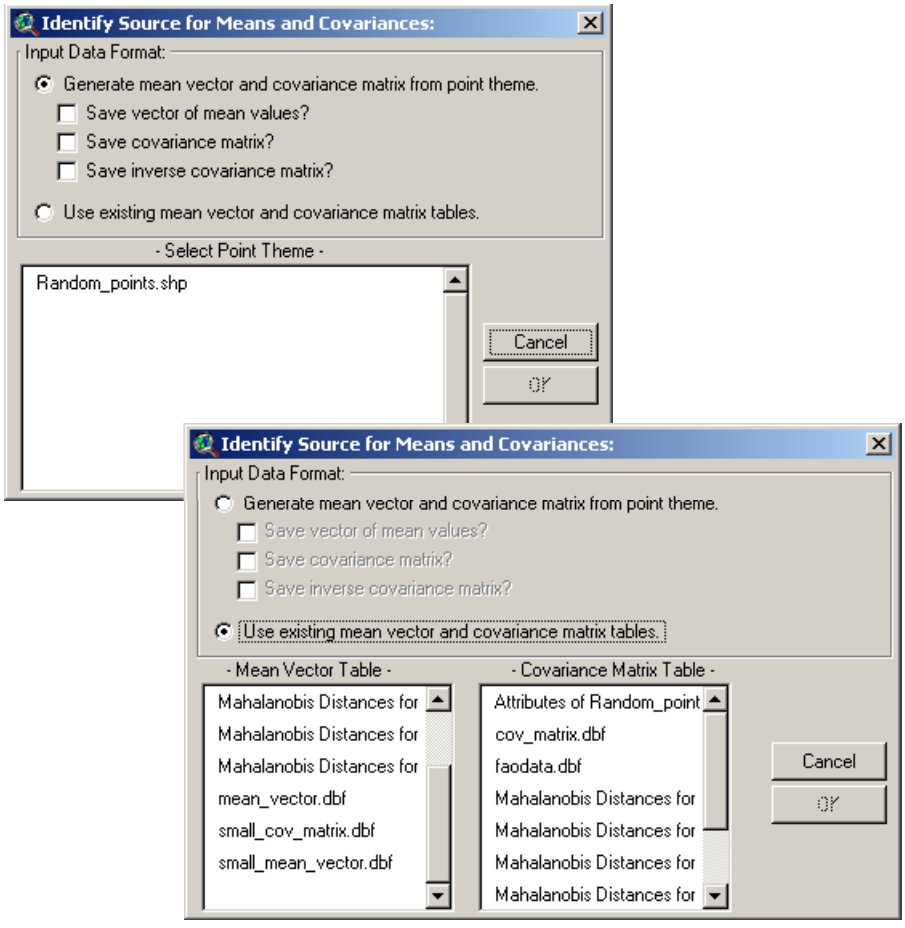
1) Begin by placing the " mahalanobis.avx" file into the ArcView extensions directory (../../Av_gis30/Arcview/ext32/).

2) After starting ArcView, load the extension by clicking on **File --> Extensions… ,** scrolling down through the list of available extensions, and then clicking the checkbox next to "Mahalanobis Distances."

Generating Mahalanobis Distance Surface Grids:

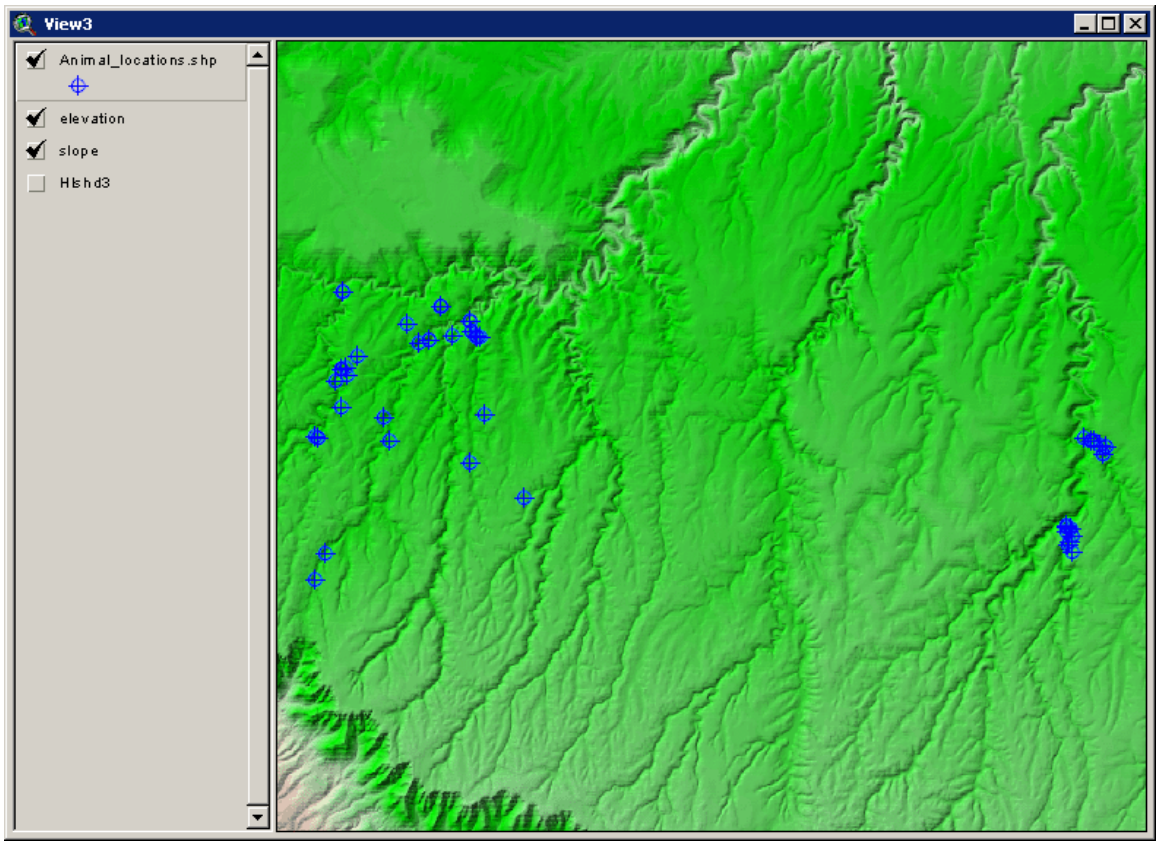
Mahalanobis surface grids require a set of independent variable data grids containing continuous numeric values, a vector of mean values for each independent variable, and a variance/covariance matrix for the set of independent variables. Users can use existing mean vector and covariance matrix tables if they have them available or they can generate them on-the-fly based on point locations distributed over the independent variable grids. IMPORTANT: Due to a limitation in ArcView Spatial Analyst, users are limited to a maximum of 8 input grids in this analysis. This limitation is expected to be fixed in Spatial Analyst v. 9.

Begin the process by clicking the "Mahalanobis Distance Surface Grid" button  in the View button bar. ArcView will prompt you to identify the source of your Mean Vector and Covariance Matrix. The "Identify Source for Means and Covariances" window is resizable by dragging on a corner.

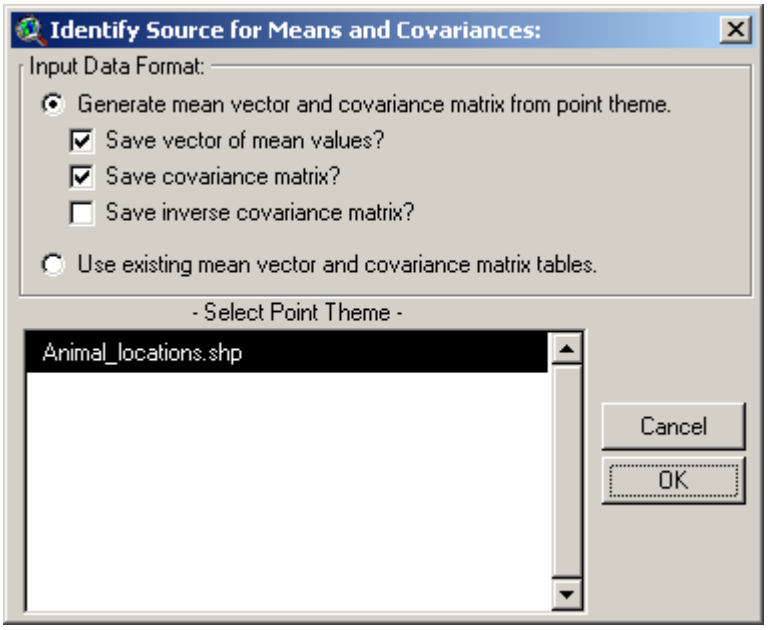


Generating Means and Covariances from point theme:

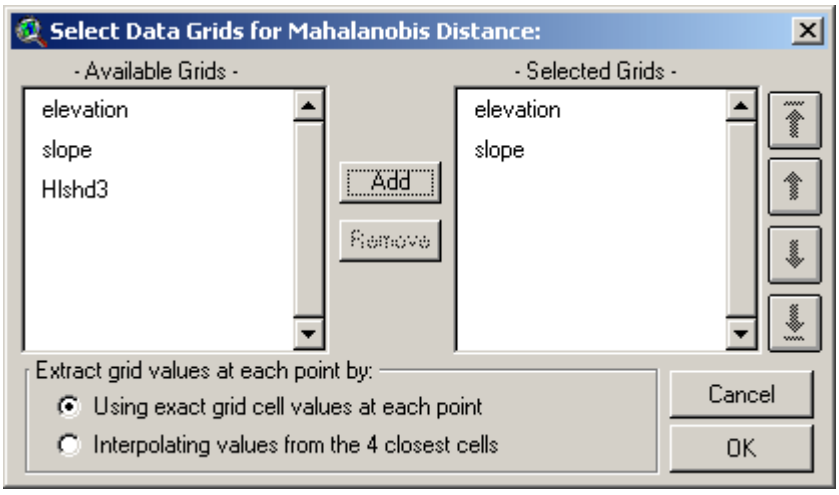
This option provides a direct way to generate a landscape surface that describes how similar any point on the landscape is to a set of sample points distributed across the landscape. For a simple example, suppose that we have a set of animal locations plus a grid of elevation and slope values, and we want to identify regions on the landscape that are similar to the animal locations. This type of analysis may be useful for identifying potential habitat for an animal species.



We can use the points directly to generate a vector of mean slope and elevation values for these animal locations, plus a covariance matrix for both slope and elevation values. Simply choose the first option in the "Identify Source for Means and Covariances" window and pick your point theme from the list at the bottom. You may choose to save tables of your mean vector, covariance matrix and inverse covariance matrix if you wish.



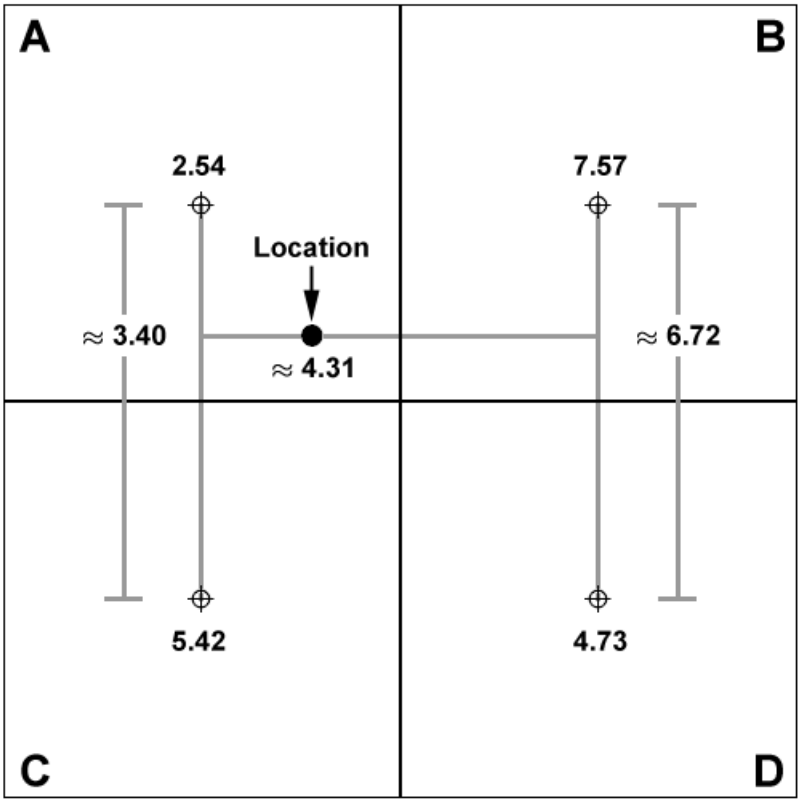
You will then be asked to identify the independent variable data grids to use with these points, and whether you wish to use exact or interpolated cell values at each point:



The list on the left shows all the grids available in your view and the list on the right shows all the grids that will be used in the analysis. Select one or more grids from the left and click the "Add" button to add them to the "Selected" list. If you need to reorder the selected grids (if, for example, you want to generate a mean vector or covariance matrix in a particular order), click on one of them and use the arrow buttons on the left to shift it up or down.

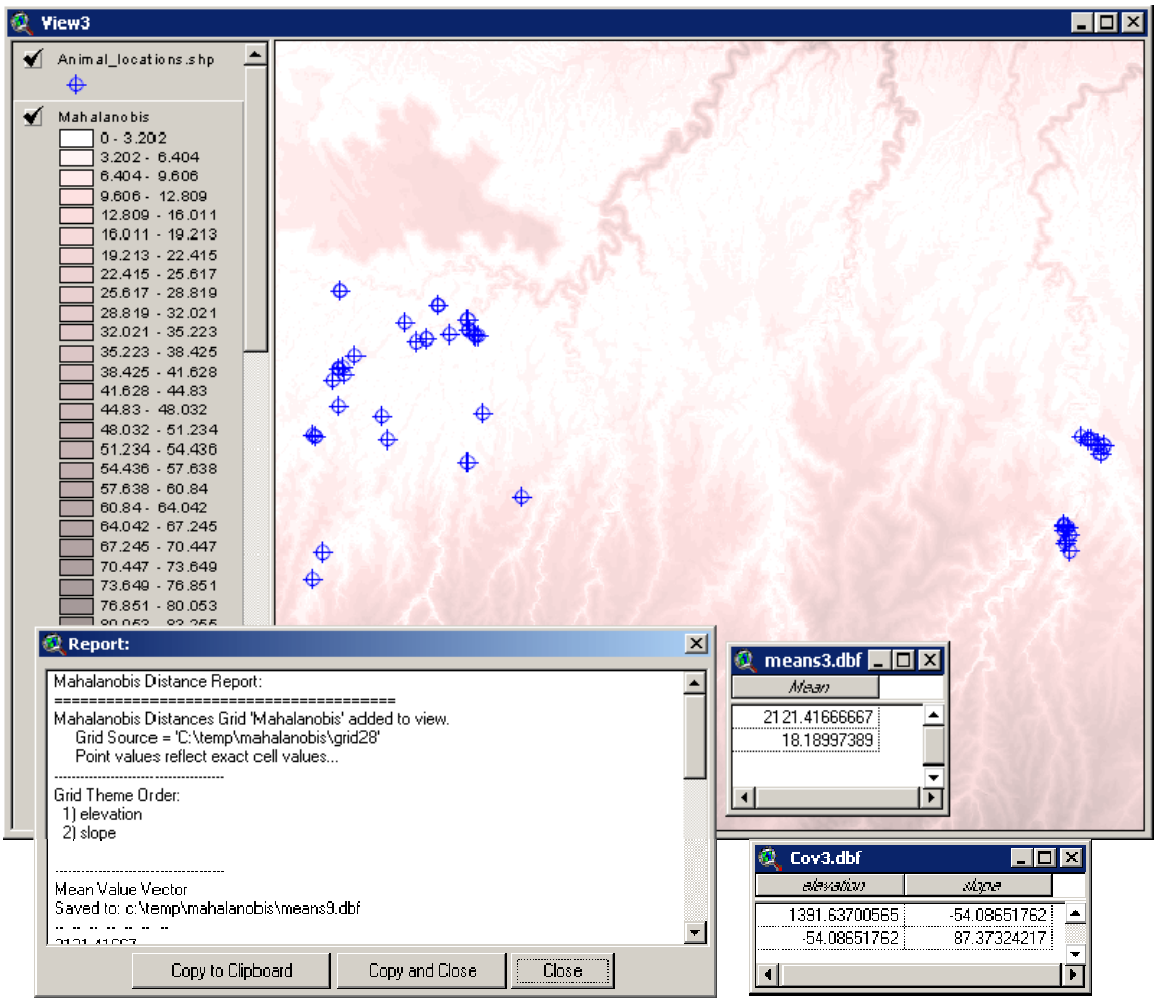
Exact Values vs. Interpolated Values:

You have the option to use the exact cell value for each of your point locations, or interpolated values based on the 4 closest cells to that point. For interpolated values, ArcView uses a 2-step method whereby values are interpolated first vertically and then horizontally. For example, given 4 cells around a particular location:



Lines are first generated between the cell centers of cells A and C, and between cells B and D, and values are interpolated along these lines at the Y-coordinate of the point location. Then a final value is interpolated along the X-axis between these two interpolated values. In this case, the interpolated value of the point is approximately 4.31, while the exact cell value of the point is 2.54.

Once you have selected your grids and point value method, click 'OK' to generate the Mahalanobis distance grid. When the computations are complete, the grid will be added to the view and you may then use it for any further classification or analyses.



In this example, we also elected to generate tables of our Mean Value vector and Covariance matrix so both of these tables will open along with the report. The values in both tables are in the order that the original grids were entered, so here the first value in the mean vector is the Elevation mean and the second value is the Slope mean. The rows in the covariance table reflect the variables in the same order as the fields, so again Elevation is in the first row and column, and Slope is in the second row and column.

The Report Window:

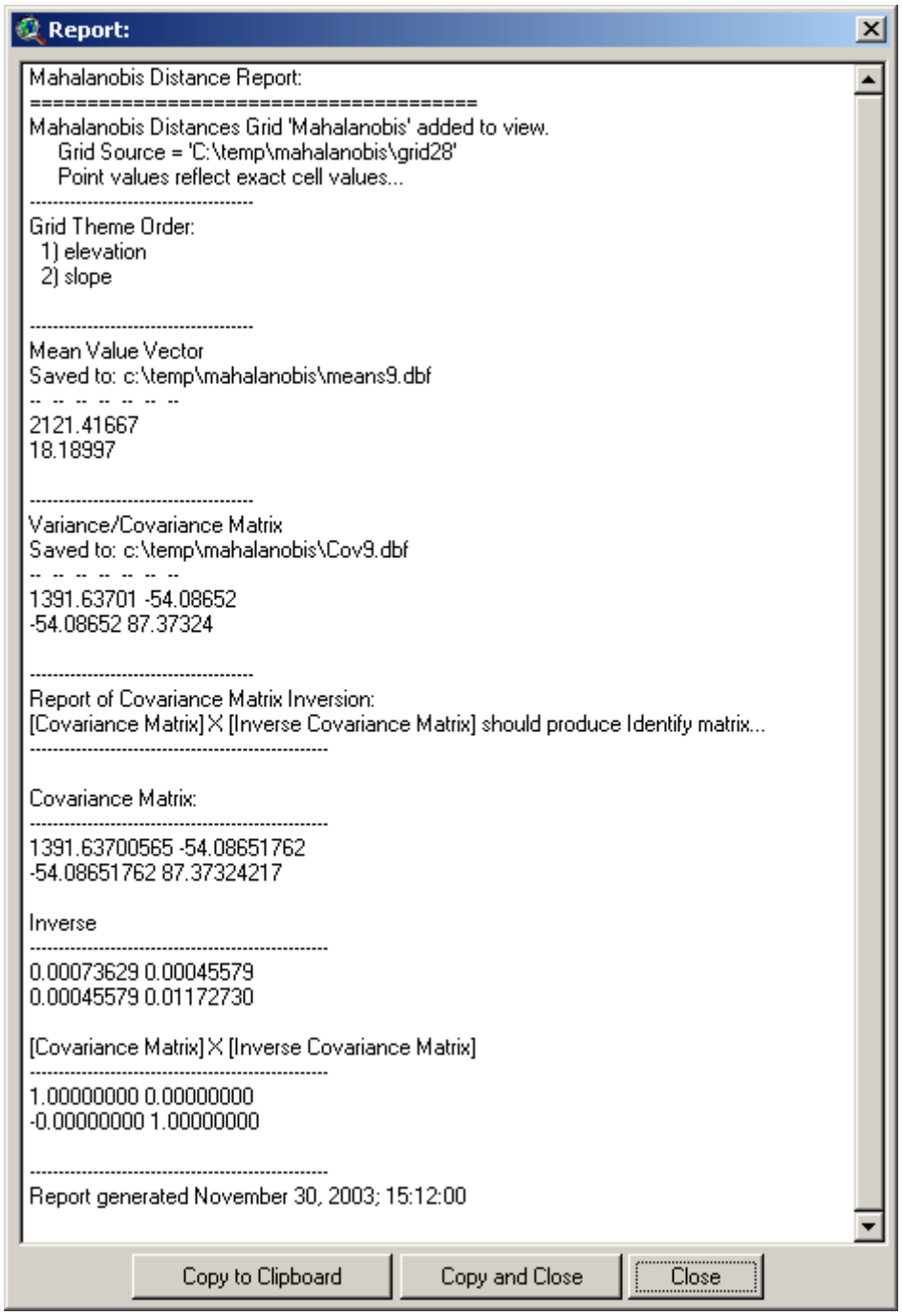
You will also see a report detailing several things that may be of interest. It begins with information on the name and hard drive location of your Mahalanobis grid and the order of the independent data grids as they were included. If any output matrices were saved, the report will also include them and show where on the hard drive they were saved. Finally, the report will allow you to check if the matrix calculations worked correctly.

Recall that the Mahalanobis equation does not use the Covariance matrix directly, but rather the inverse of that matrix:

$$D^2 = (\mathbf{x} - \mathbf{m})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m})$$

Therefore this extension must generate the inverse of the matrix before calculating the Mahalanobis distances. This extension uses the Lower/Upper Decomposition method of matrix inversion as described by Press et al. (2002).

Matrix inversion can be computationally complicated and many sources recommend checking the accuracy of the process before relying on it. The output report helps you to check that accuracy by multiplying the Covariance matrix by the Inverse covariance matrix, which should produce the Identity matrix (all 0's except for 1's down the diagonal):



The multiplied matrix appears near the bottom of the report. Do not worry about negative 0 values; these are due to rounding issues in the computer which are an inherent problem with 32-bit operating systems. These "0" values typically have non-zero values at the 10th or greater decimal place and sometimes these values are very slightly lower than 0, forcing a "-0" value

instead of a “0” value. Such matrices are still sufficiently close to perfect Identity matrices to demonstrate that the matrix inversion was successful.

Using Categorical Grids:

Categorical data do not lend themselves directly to Mahalanobis analysis. Mahalanobis values reflect how similar some set of values is to some ideal vector of values, and this ideal vector is generally assumed to be composed of the means of the variables involved. It is difficult to find the “mean” of a set of categories, and therefore they are not appropriate for Mahalanobis analysis.

However, there are aspects of categorical datasets that can be used to generate Mahalanobis distances. Clark et al. (1993) derived a numeric diversity grid from their categorical grid, where each cell value reflected the number of categories within a particular neighborhood around that cell. These data don't exactly follow a continuous distribution, but they are still reasonable as a Mahalanobis independent variable. You can generate this kind of grid using the Neighborhood Statistics function in Spatial Analyst. Generate the statistic named "Variety", which will only be available if you have an integer input grid (which is true of categorical grids).

The author has also written a tool to calculate neighborhood statistics which offers a few more options than the standard Spatial Analyst one (see Grid Tools at http://www.jennessent.com/arcview/grid_tools.htm).

Another option, also using neighborhood statistics, is to determine the proportion of the neighborhood that is composed of a particular category. You may need to generate several of these proportion grids if you want to use several categories in the Mahalanobis analysis. You can generate these category proportion grids as follows:

1) Click your "Analysis" menu, then the "Map Query..." menu item.
2) When the Map Query dialog opens, generate a query string querying your categorical grid for one particular category. For example, if you had a forest cover type category, you might enter:

   [Cover_grid] = 4

   where "4" would reflect one of the cover type categories.
3) Now you will have a "Map Query #" grid in your view, with "1" values reflecting the area represented by that category, and "0" values representing all other areas.
4) Open your Neighborhood Statistics tool and generate neighborhood statistics on your Map Query grid. You want to calculate the Sum, which will tell you the number of cells of that category within the specified neighborhood around each cell.
5) To convert this to proportions, you'll also need to find the total number of cells in that neighborhood.
6) Set your Analysis Environment to match your Map Query grid. Click the “Analysis” menu, then “Properties…”
7) In the drop-down box next to “Analysis Extent”, select your Neighborhood Sum grid.
8) In the drop-down box next to “Analysis Cell Size” , select your Neighborhood Sum grid.
9) Generate a grid of "1" values by opening the Map Calculator dialog again and entering the following calculation string:

   1.AsGrid

10) Open your Neighborhood Statistics tool again, use the exact same neighborhood, and calculate the sum of your grid of "1" values. This will tell you the total number of cells

within the specified neighborhood. Naturally, this value should always be ≥ the number of cells of each category within that neighborhood.

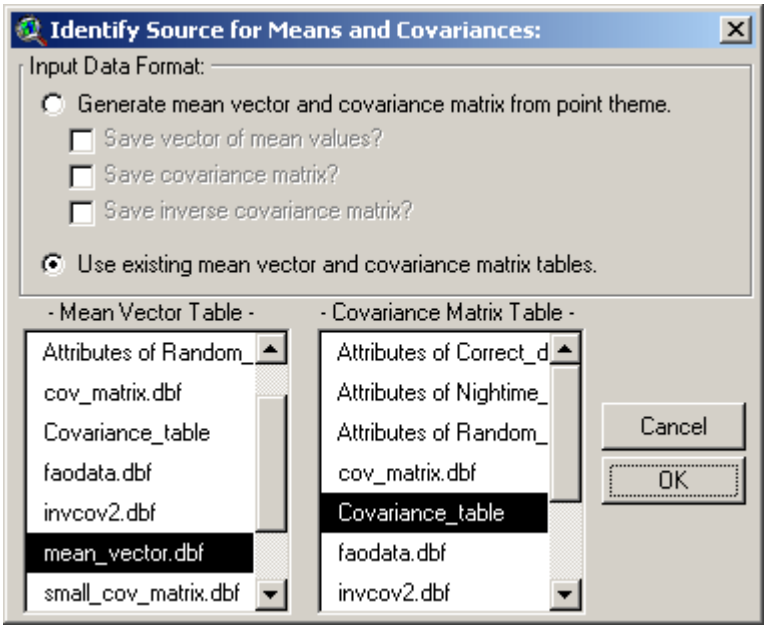
11) Now you have two Neighborhood Sum grids; one representing the number of cells of that category in your neighborhood, and the other representing the total number of cells in that neighborhood. Divide the Category Sum grid by the Total Sum grid and you'll get the proportion grid.

12) Use that Proportion grid as one of the independent grids in the Mahalanobis tool.

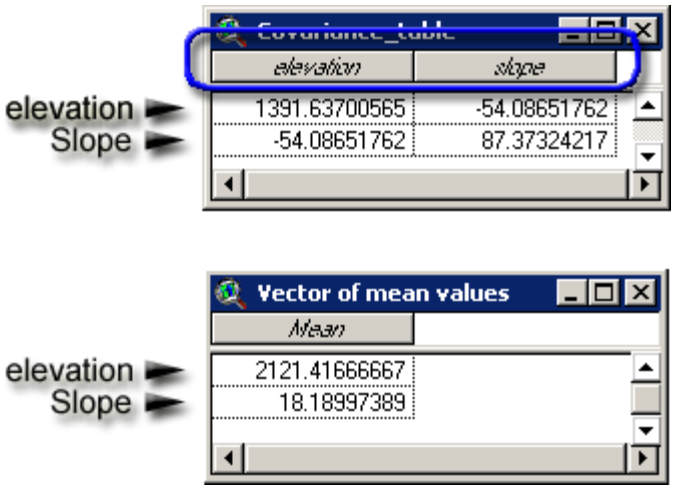
Using Existing Mean and Covariance Data:

This option allows you to use an existing mean vector and covariance matrix in your analysis rather than generating them on-the-fly from point locations. This option is useful if you have already derived your means and covariances using this extension or some other software, or if you would like to generate comparative Mahalanobis surface grids using slightly different mean vectors. Knick and Dyer (1997) describe a method of substituting a weighted mean and covariance matrix when certain input variables are better measured than others.

If you choose this option, you will need to identify the tables containing your Mean Value vector and your Covariance matrix before clicking 'OK':

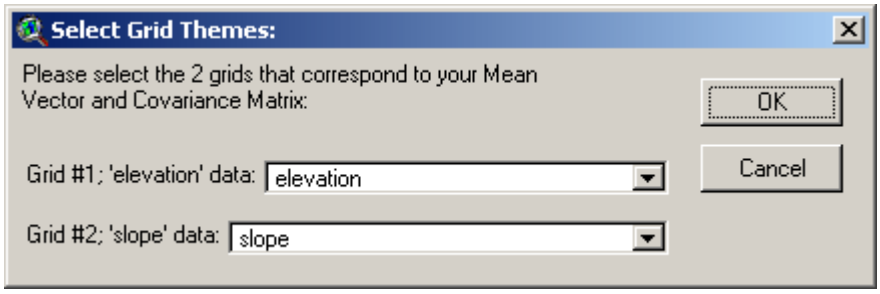


IMPORTANT: These tables must be in the correct format for the analysis to work! Both the mean vector table and the covariance matrix table must contain only numeric values and they must be ordered correctly. The field order of the Covariance matrix table should apply to the row order of both the Covariance matrix table and the Mean Vector table:



The tool will check the tables to see if they appear to contain valid matrices before letting you continue.

Next, you will be prompted to identify your independent variable grids. These grids must be selected in the order of the matrices above, and the query window is designed to facilitate this:

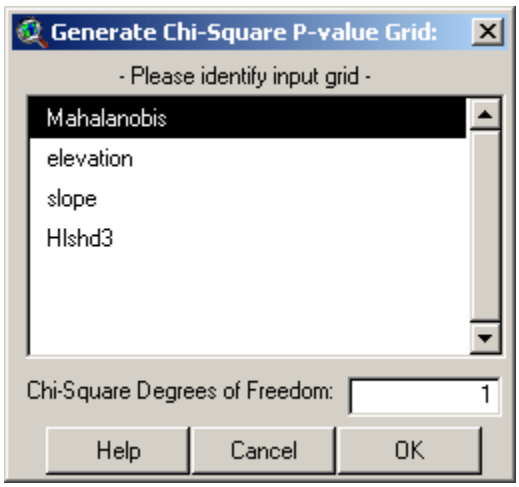


Click 'OK' to generate your grid. Your Mahalanobis Distance grid will be added to your view and you will see a report describing the analysis. See the discussion above on the Report window for an explanation of the report.

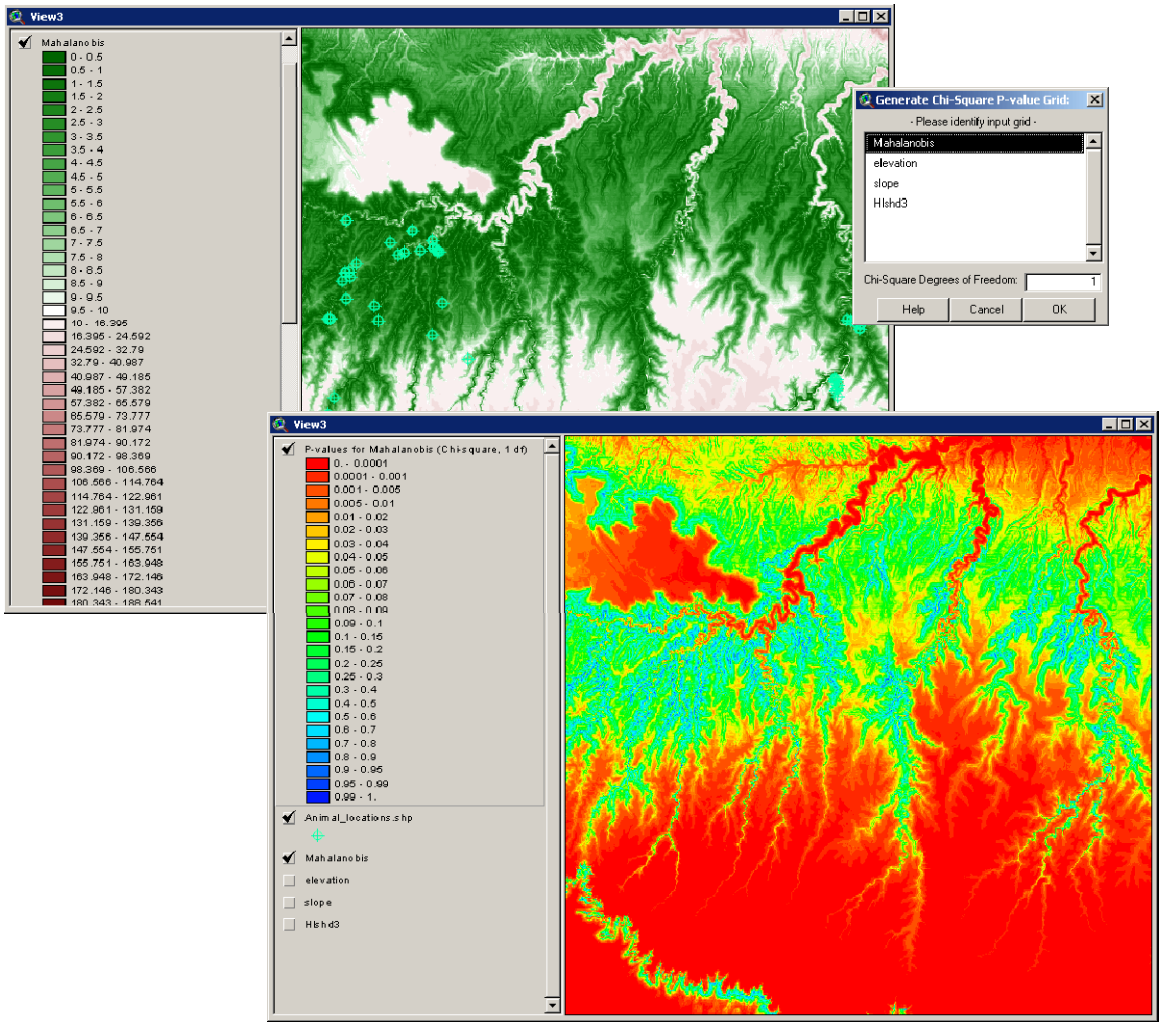
Generating P-value grid from Mahalanobis Distance Grid:

When the predictor variables used to generate the mean vector and covariance matrix are normally distributed, then Mahalanobis distances are distributed approximately according to a Chi-square distribution with *n-1* degrees of freedom. In such cases it may be useful to convert the Mahalanobis distance grid into a grid of *p-value*s. If the predictor variables are not normally distributed, it may still be useful to make this conversion because it rescales the unbounded Mahalanobis values such that they are all between 0 and 1. For more details on the uses of converting to *p*-values, see the discussion on Chi-Square values on page 5 or refer to Clark et al. 1993.

The  button provides an easy way to convert Mahalanobis grids into *p*-value grids. Click the button and you will be prompted to identify your Mahalanobis grid and the appropriate degrees of freedom:



The degrees of freedom should be equal to *n-1*, where *n* = # predictor variables used to generate the original mean vector and covariance matrix. The Mahalanobis grid in this example was generated from 2 predictor variables (Slope and Elevation), so there would only be 1 degree of freedom. Click the OK button to generate the *p*-value grid:

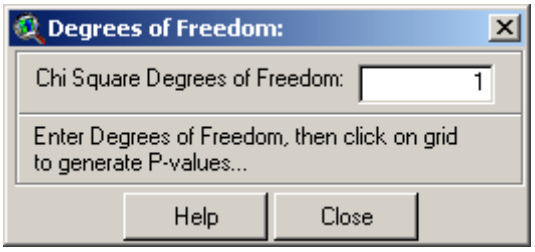


Due to a limitation in Spatial Analyst, this function does not generate exact *p*-values for each cell but rather classifies the grid into 26 *p*-value ranges. The *p*-value for each cell reflects the probability of seeing a Mahalanobis value as large or larger than the actual Mahalanobis value for that cell, assuming the vector of predictor values that produced that Mahalanobis value was sampled from a population with an ideal mean (i.e. equal to the vector of mean predictor variable values used to generate the original Mahalanobis grid). *P*-values close to 0 reflect high Mahalanobis distance values and are therefore very dissimilar to the ideal combination of predictor variables. *P*-values close to 1 reflect low Mahalanobis distances and are therefore very similar to the ideal combination of predictor variables.

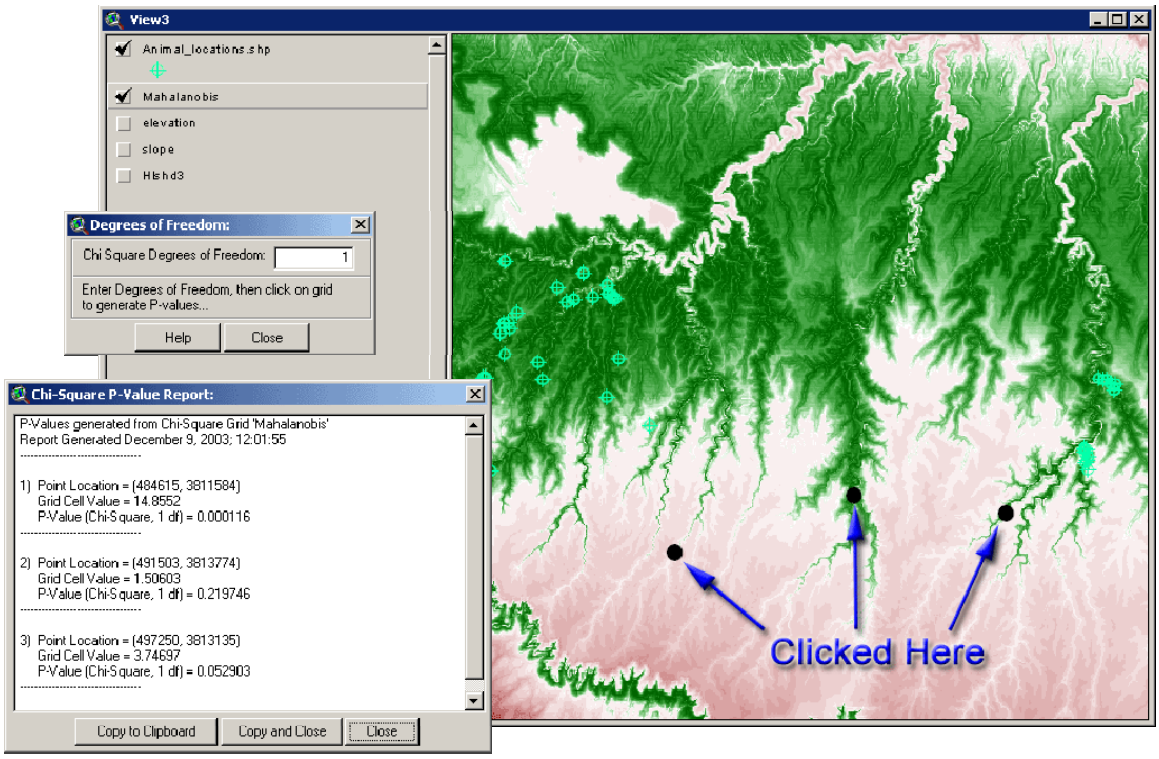
Calculating P-values for individual Mahalanobis Distance Grid cells:

The χ^2 tool allows you to calculate exact *p*-values for individual Mahalanobis surface grid cells, assuming a Chi-square distribution with *n-1* degrees of freedom where *n* = # predictor variables used to generate the mean vector and covariance matrix. See the discussion of Chi-Square values on page 5 for an explanation of this concept.

When you initially click the χ^2 tool, you will be asked to identify the degrees of freedom to use in the calculations:



After you enter a number, you can start clicking on the screen to calculate *p*-values. Your cursor should be represented with a ⊙ symbol. As you click on different places on the grid, you will generate a running list of *p*-values:

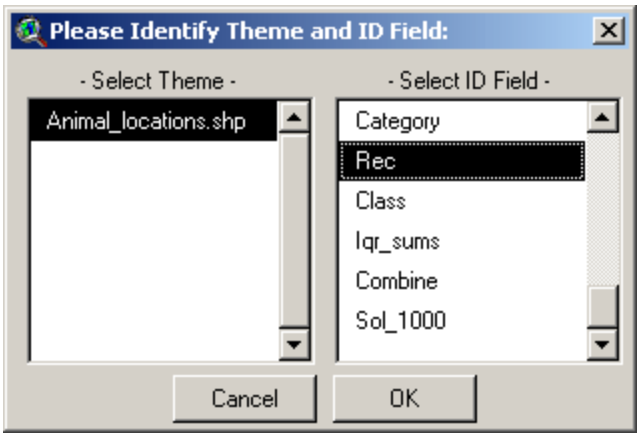


Generating Mahalanobis Distances for Feature Themes:

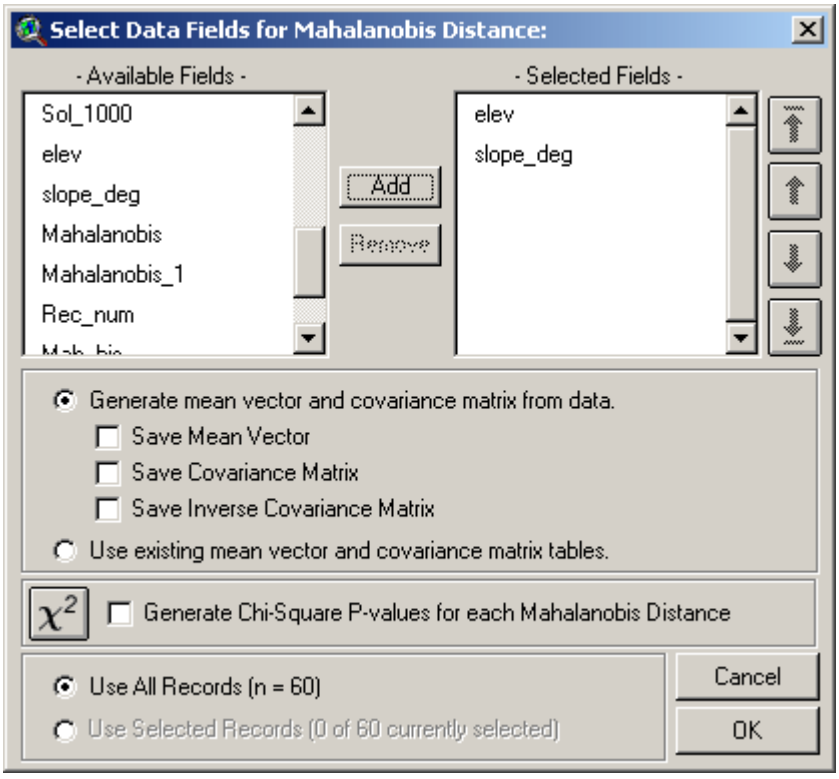
This process does not generate a Mahalanobis surface grid, nor does it use any predictor variable grids in the analysis. Rather, it generates individual Mahalanobis distances for each feature in a feature theme based on attribute data contained in the feature attribute table. The output is presented in a Results table containing ID values for each feature and the Mahalanobis distance of that feature, and the table can easily be joined to the feature attribute table for further analysis.

If you wish to add a field with Mahalanobis distances directly to the attribute table, open the table using the  button and use the "Generate Mahalanobis Distances for Tables" option (page 22). The tools for generating Mahalanobis distances for feature themes and for tables are essentially identical, except that the Table function adds a field to the table while the Theme function creates a separate Results table.

Click the "Theme Mahalanobis Distances" button  to start the process. Because this button works directly on feature themes, the button will only be enabled if there is at least one point, line or polygon theme present in the view. You will first be prompted to identify your theme and an ID field containing unique ID values for each feature in the theme. These ID values are necessary for you to join your Results table with your theme attribute table:



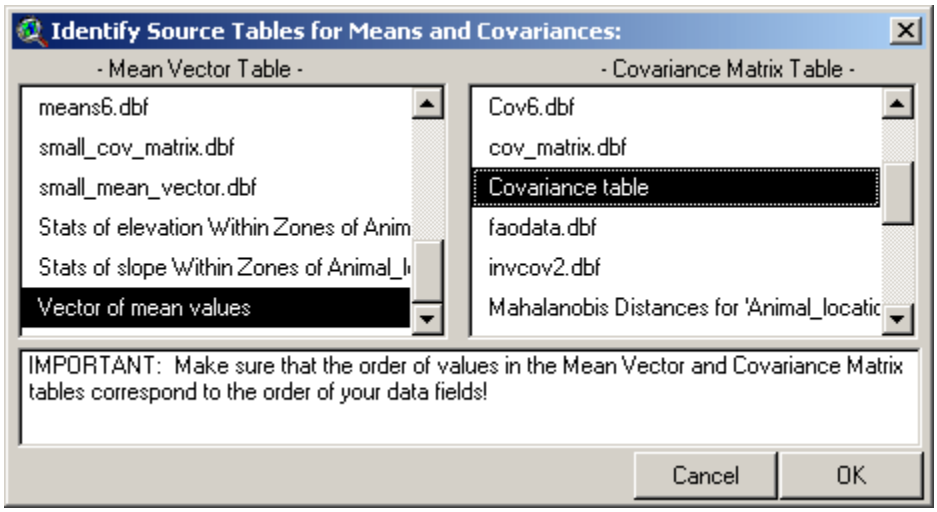
Next, identify the fields containing the independent variable values for each record, and specify whether you would like to generate the mean vector and covariance matrix directly from the data or use existing mean vector and covariance matrix tables:



The “Available Fields” list on the left contains all the numeric fields available in the theme attribute table, and the “Selected Fields” list on the right contains all the fields to be used in the analysis. Select one or more fields from the “Available” list and click the “Add” button to add them to the “Selected” list. If you need to reorder the selected fields (if, for example, you need to generate a mean vector or covariance matrix in a particular order, or if you need to reorder your fields to match an existing mean vector or covariance matrix), use the arrow buttons on the left to shuffle the fields up or down.

You have the option to generate your mean vector and covariance matrix directly from the data, in which case the Mahalanobis distances will reflect the distance of each individual feature from the internal mean vector of the group. Such values may be useful for determining within-group

variability. Alternatively you can generate distances of each record from some other mean vector, possibly generated from a control group or based on earlier research, by clicking the "Use existing mean vector and covariance matrix tables" option. See Knick and Dyer (1997) for an example of substituting a weighted mean and covariance matrix when certain input variables are better measured than others. If you choose this second option, you will next be asked to identify the tables containing your Mean vector and Covariance matrix:



This window is resizable by dragging on a corner.

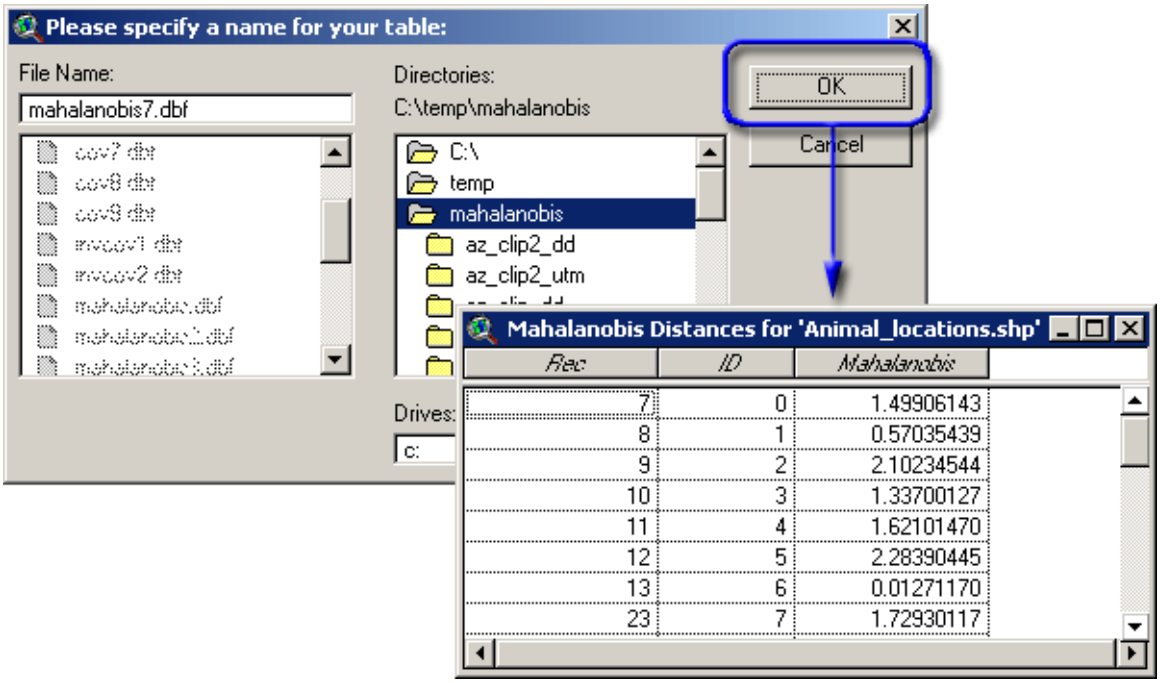
Recall that using existing tables requires that your input fields be correctly ordered. In our previous example, we don't want our Elevation values to be evaluated based on our Slope mean and variance! Use the arrow buttons on the previous window to adjust your field order.

Additional Options:

If any of your features are currently selected, you will have the option to use either all features in the analysis or only the selected features.

You also have the option to generate *p*-values for each Mahalanobis value, based on a Chi-square distribution with *n-1* degrees of freedom. See the discussion of Chi-square *p*-values on page 5 for a description of the relationship between Chi-square *p*-values and Mahalanobis values. The χ^2 button opens up a help window briefly discussing Chi-square *p*-values.

Click 'OK' to generate the Results table of Mahalanobis distances. You will be asked where you want to save your table, and the table will then be generated and opened. If you elected to generate additional tables, these tables will open also.



If you wish, you can join this new table with your original attribute table with the following steps:

1) Click the ID field of your new Mahalanobis Distance table ("Rec" in our example).

2) It not already open, open your theme attribute table by clicking the Open Table button  on the View button bar.

3) Click the ID field of your theme attribute table ("Rec" in our example).

4) Click the Join button on the Table button bar.

5) Your tables are now joined, and the Mahalanobis distances will appear as a field in the attribute table as long as the tables stay joined. The dbf files themselves have not been altered, but any analyses done on the theme can now include the Mahalanobis distances.

You will also see a report describing several aspects of the analysis which may be of interest, including the hard drive location of any new tables and a check on the matrix inversion calculations (see explanation of the Report window in "Generating Mahalanobis Distance Surface Grids" [page 12] for more details).

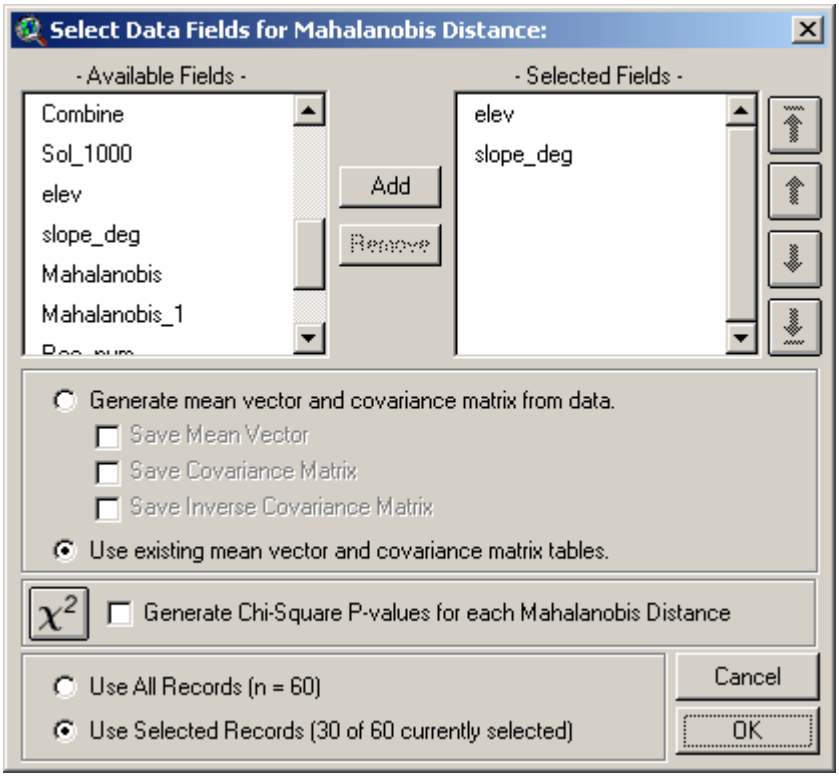
Generating Mahalanobis Distances for Tables:

This function generates individual Mahalanobis distances for each record in a table based on independent variable fields contained in the table. The function adds a field to the table labeled "Mahalanobis" containing the Mahalanobis distances.

If you cannot or do not wish to modify your table by adding a field with Mahalanobis distances, use the "Generate Mahalanobis Distances for Feature Themes" function (page 19) to create a separate Results table. This separate table can then be joined with the current table for further analysis. The tools for generating Mahalanobis distances for feature themes and for tables are essentially identical, except that the Table function adds a field to the table while the Theme function creates a separate Results table. If you have no theme associated with this table, you

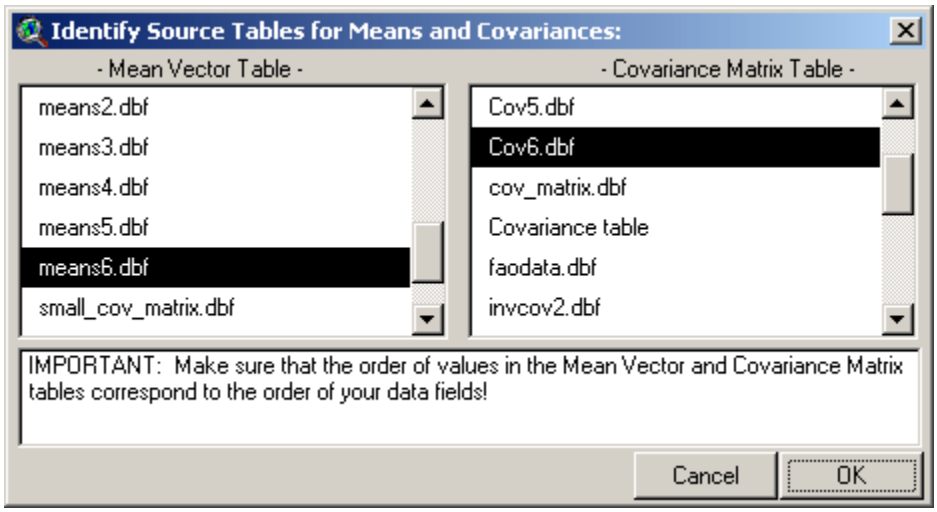
can create a temporary one using the "Add Event Theme" menu item in the View menu (see ArcView Help files).

Begin by clicking the "Calculate Mahalanobis Distances" button  in the Table button bar. You will be prompted to identify the fields in the open table containing the independent variable values for each record, and specify whether you would like to generate the mean vector and covariance matrix directly from the data or use existing mean vector and covariance matrix tables:



The "Available Fields" list on the left contains all the numeric fields available in the current table, and the "Selected Fields" list on the right contains all the fields to be used in the analysis. Select one or more fields from the "Available" list and click the "Add" button to add them to the "Selected" list. If you need to reorder the selected fields (if, for example, you need to generate a mean vector or covariance matrix in a particular order, or if you need to reorder your fields to match an existing mean vector or covariance matrix), click on any of the selected fields and use the arrow buttons on the left to shuffle it up or down.

You have the option to generate your mean vector and covariance matrix directly from the data, in which case the Mahalanobis distances will reflect the distance of each individual feature from the internal mean vector of the group. Such values may be useful for determining within-group variability. Alternatively you can generate distances of each record from a separate mean vector, possibly generated from a control group or based on earlier research, by clicking the "Use existing mean vector and covariance matrix tables" option. See Knick and Dyer (1997) for an example of substituting a weighted mean and covariance matrix when certain input variables are better measured than others. If you choose this second option, you will next be asked to identify the tables containing your Mean vector and Covariance matrix:



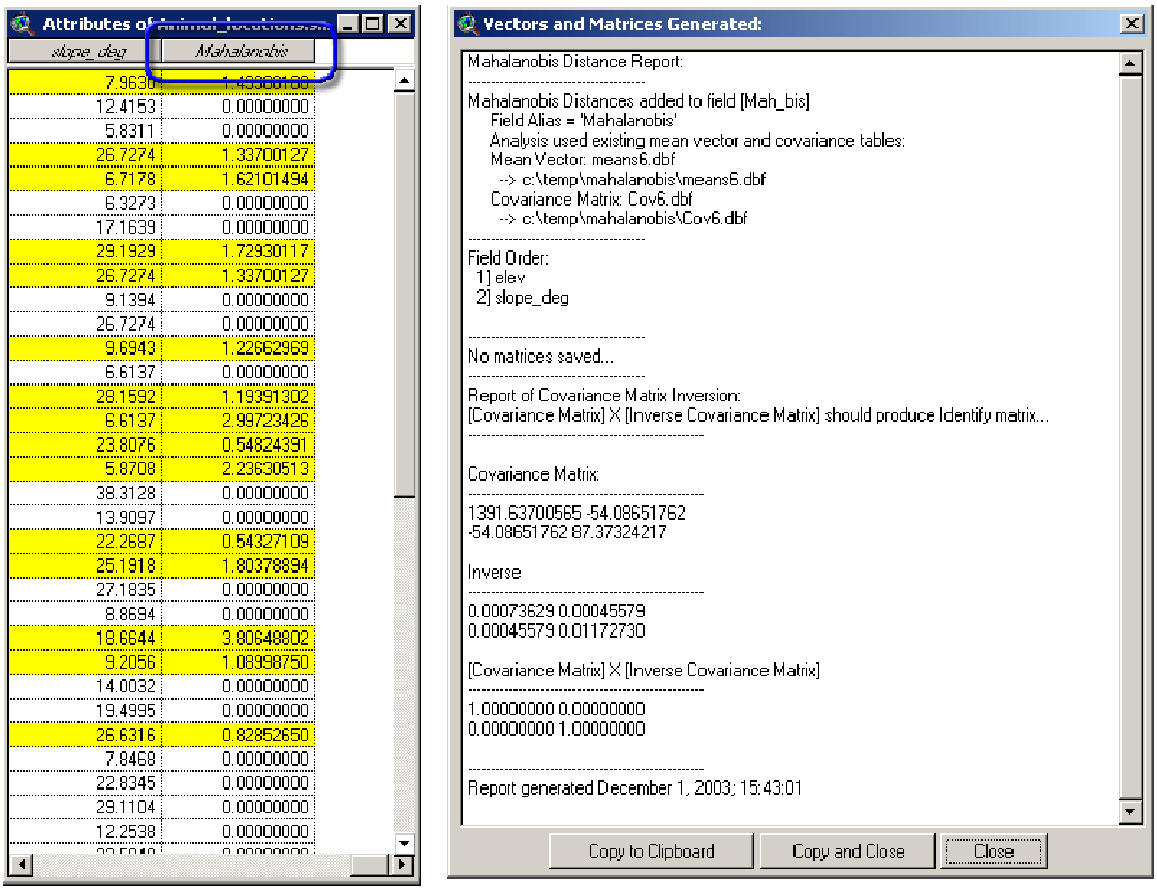
Recall that using existing tables requires that your input fields be correctly ordered. Using tables in which the independent variable fields follow a different order will cause the data to be analyzed based on incorrect means and covariances, and will produce invalid Mahalanobis distances.

Additional Options:

If any of your records are currently selected, you will have the option to use either all records in the analysis or only the selected records.

You also have the option to generate *p*-values for each Mahalanobis value, based on a Chi-square distribution with *n-1* degrees of freedom. See the discussion of Chi-square *p*-values on page 5 for a description of the relationship between Chi-square *p*-values and Mahalanobis values. The χ^2 button opens up a help window briefly discussing Chi-square *p*-values.

Click the 'OK' button to start the analysis. As soon as computations are done, the tool will add a new field to your table named "Mahalanobis" containing the Mahalanobis distance values and open a report window containing information about the analysis:



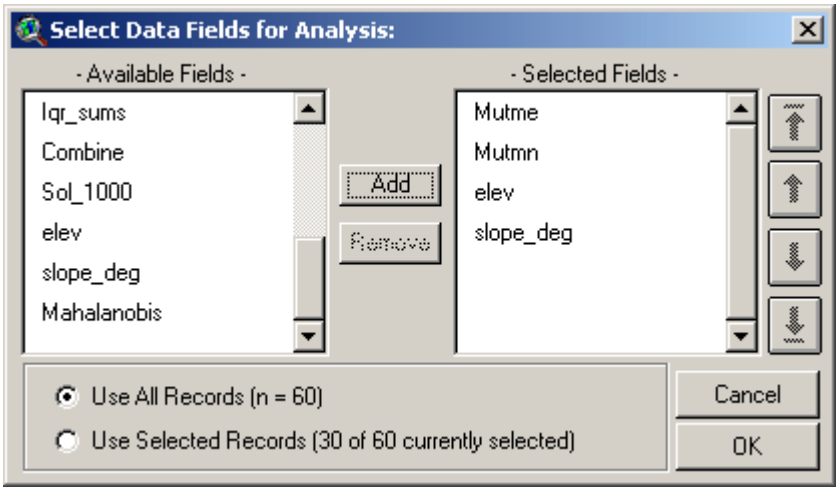
See the description of the Report window (page 12) for more details on the report components.

Generating Statistical Matrices:

This function provides a quick way to generate tables containing the mean vector, covariance matrix, inverse covariance matrix, Pearson's *r* correlation matrix, and Spearman's rho rank correlation matrix from multiple fields in the current table. The mean vector and covariance matrix tables can be used with the Mahalanobis functions described elsewhere in this manual. Options to generate a Pearson's *r* and/or Spearman's rho correlation matrix are included because the author of the extension needed them for something and decided to leave them available for others to use.

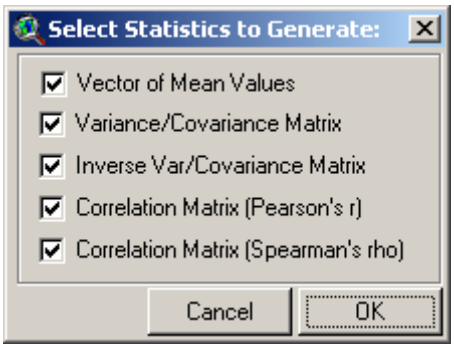
Pearson's *r* correlations measure how much one variable changes as a second variable changes, and in which direction. Values range between -1 and 1, with negative values implying a negative relationship (i.e. as one variable increases, the other decreases). Values close to 1 or -1 have high correlation while values close to 0 have low correlation. Spearman's rho correlations are identical to Pearson's *r* except that they are calculated from the relative rank of each value rather than the value itself (see Conover 1980:252). Spearman's rho correlations are generally considered more appropriate when the variables are not normally distributed or when the researcher wants to reduce the importance of outliers.

Open your table and click the "Create Statistical Matrices" button  to start the process. You will be prompted to identify the fields to include in the analysis:

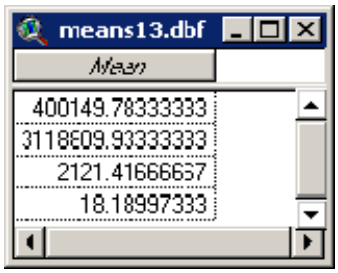


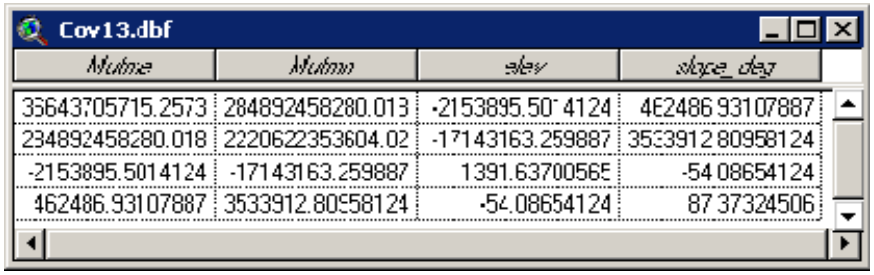
The list labeled “Available Fields” contains all the numeric fields available in your open table and the list labeled “Selected Fields” contains all the fields currently selected for analysis. Select the fields you would like to analyze and click the “Add” button to add them to the Selected list. If you need to change the order of the selected fields for any reason, click on one of the fields and use the arrow buttons on the left to shuffle it up or down. If any of your records are selected, you have the option to analyze either the full set of records or only the selected set.

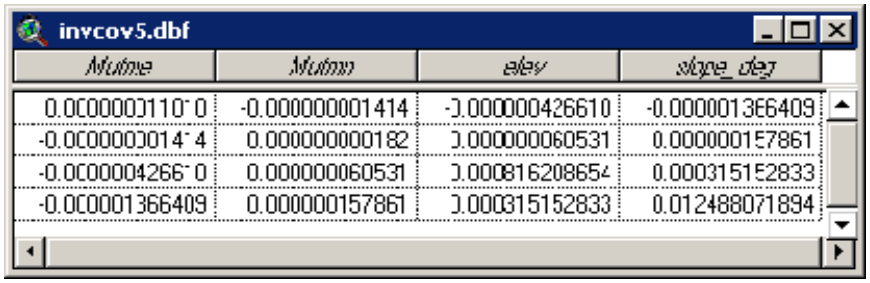
Next, choose which matrices you would like to generate:

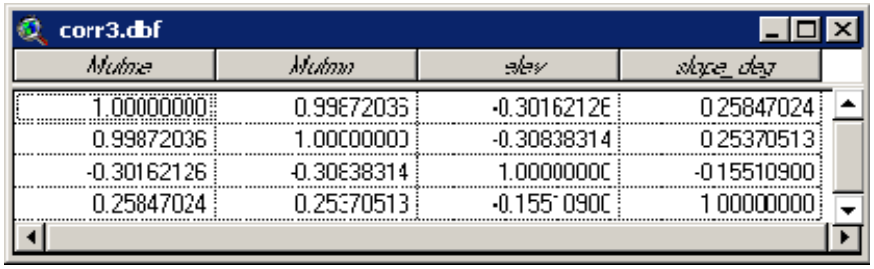


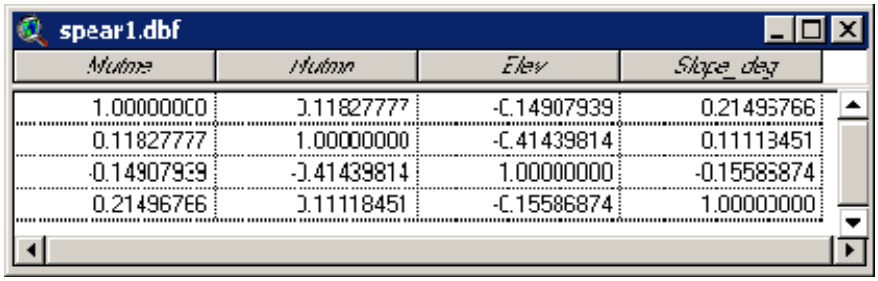
Click ‘OK’ and the tables will be generated and opened, along with a Report dialog describing the analysis. See the description of the Report window (page 12) for more details on the report components.











Statistical Matrix Methods:

- *Mean:* $\frac{\sum_{i=1}^{n} X_i}{n}$
- *Variance/Covariance Matrix:*

Variance of variable $X = \sigma_{xx} = \sigma^2$; estimated by $\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

Covariance between x and $y = \sigma_{xy}$; estimated by $\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Therefore, given p variables:

$$\text{Covariance Matrix } Cov(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$\text{estimated by} = \begin{bmatrix} \dfrac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}{n-1} & \dfrac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{n-1} & \cdots & \dfrac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{ip} - \bar{x}_p)}{n-1} \\ \dfrac{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)(x_{i1} - \bar{x}_1)}{n-1} & \dfrac{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}{n-1} & \cdots & \dfrac{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)(x_{ip} - \bar{x}_p)}{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\sum_{i=1}^{n}(x_{ip} - \bar{x}_p)(x_{i1} - \bar{x}_1)}{n-1} & \dfrac{\sum_{i=1}^{n}(x_{ip} - \bar{x}_p)(x_{i2} - \bar{x}_2)}{n-1} & \cdots & \dfrac{\sum_{i=1}^{n}(x_{ip} - \bar{x}_p)^2}{n-1} \end{bmatrix}$$

- *Inverse Covariance Matrix:* Matrix inversion is computationally complex, and the author refers interested readers to the Lower/Upper (LU) Decomposition method in chapter 2 of Press et al (2002).
- *Pearson Correlation Matrix:*

$$\text{Given a Covariance Matrix } Cov(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$\text{the Pearson Correlation Matrix} = \begin{bmatrix} \dfrac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \dfrac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \dfrac{\sigma_{21}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \dfrac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \dfrac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\sigma_{p1}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \dfrac{\sigma_{p2}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \cdots & \dfrac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix}$$

- *Spearman Correlation Matrix:* Computationally identical to the Pearson Correlation Matrix except that ranks are used in place of original values. For example, the list of values {12, 3, 56, 23, 1} would be replaced with {3, 2, 5, 4, 1}, and the replacement list would then be used to generate the correlation matrix.

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.

Enjoy! Please contact the author if you have problems or find bugs.

Jeff Jenness
Jenness Enterprises
3020 N. Schevene Blvd.
Flagstaff, AZ 86004
USA

jeffj@jennessent.com
http://www.jennessent.com
(928) 607-4638

Updates to this extension and an on-line version of this manual are available at

http://www.jennessent.com/arcview/mahalanobis.htm

Please visit *Jenness Enterprises* ArcView Extensions site for more ArcView Extensions and other software by the author. We also offer customized ArcView-based GIS consultation services to help you meet your specific data analysis and application development needs.