Paper 198-26

Partitioned Mahalanobis D^2 to Improve GIS Classification

Lynette Duncan, University of Arkansas, Fayetteville, AR
James E. Dunn, University of Arkansas, Fayetteville, AR

ABSTRACT

Recent work has pointed out the utility of using a k-component partition, say $D^2(k)$, of Mahalanobis D^2 in order to sharpen the quality of GIS-based habitat maps. This paper reviews the mathematical basis of the partition, and uses two examples from ecology to illustrate how the approach is implemented using SAS®. Dimension reduction, i.e., variable selection, is identified as a problem in confirmatory factor analysis using procedure CALIS. Robust p-values, to classify pixels as being species specific or not are obtained from the smoothed empirical cumulative distribution function of $D^2(k)$ based on procedure KDE. Finally, SAS/GIS® is used to draw prospective habitat utilization maps for the particular species. Some innovative SAS code is shown, particularly that for setting up the PROC CALIS application. Statistical ecologists and demographers involved in multivariate problems directed toward characterizing habitat selection and utilization will be interested in this paper, as well as others who are curious to see applications of these relatively new SAS procedures.

INTRODUCTION

The availability of multivariate data on habitats has increased as GIS databases have expanded. The multivariate data obtained from a GIS need to be analyzed with multivariate methods. Some such methods include logistic regression, discriminant function analysis, and Mahalanobis D^2. The first two methods require that there be observations where species are known to be present or absent. However, the nature of observational studies lends itself to false negatives, i.e., a species may be present but not observed or it may not be present while the observer is present but is present later. By using D^2, only presence data are needed, thus avoiding the use of potential false negatives.

Rotenberry, Knick, and Dunn (1999) proposed the use of partitions of Mahalanobis D^2 instead of the full D^2 to define the specific habitat requirements of a given species. They made analogy to Pearson's planes of closest fit to explain species requirements.

We have found that the SAS System is well equipped to calculate $D^2(k)$ and to manage the large amounts of data required to create maps. SAS/GIS provided a convenient tool for drawing maps.

The first three sections of this paper review the development of $D^2(k)$, the use of Pearson's planes of closest fit, and a method of variable selection from Dunn and Duncan (2000). The next two sections present two examples from ecology. These sections provide SAS code for all of the methods used including the code for SAS/GIS.

PARTITIONING MAHALANOBIS D^2

Suppose that q random variables, $y_1,...,y_q$, describe habitat. Define $\mathbf{y} = [y_1,...,y_q]'$, and let *Y* be the set of all habitats which are suitable for the species. Let $E[\mathbf{y}] = \boldsymbol{\mu}$, $\text{var}[\mathbf{y}] = \Sigma$ for $\mathbf{y} \in Y$. Mahalanobis D^2, as commonly used in GIS mapping, is defined by

$$D^2 = (\mathbf{y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu}), \qquad (1)$$

interpreted as the squared, standardized distance separating $\mathbf{y}$ and $\boldsymbol{\mu}$. Using the spectral decomposition of Σ^{-1} (Johnson, 1998), D^2 can be represented by

$$D^2 = \sum_{j=1}^{q} (\mathbf{y} - \boldsymbol{\mu})' \alpha_j \alpha_j' (\mathbf{y} - \boldsymbol{\mu}) / \lambda_j = \sum_{j=1}^{q} \left(d_j / \sqrt{\lambda_j} \right)^2, \qquad (2)$$

where $\lambda_1 \geq \ldots \geq \lambda_q$ are the eigenvalues of Σ with associated, length one eigenvectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q$ and where $d_j = (\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\alpha}_j$ (Dunn and Duncan, 2000).

In applications, $\boldsymbol{\mu}$ usually will be replaced by the centroid, $\bar{\mathbf{y}} = n^{-1} \sum_{i=1}^{n} \mathbf{y}_i$ and Σ by its unbiased estimator, $S = (n-1)^{-1} \sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$, where $\mathbf{y}_1,...,\mathbf{y}_n$ are q-dimensional vectors characterizing n habitats known to be occupied by the species.

In his 1901 paper, K. Pearson introduced the idea of a "plane of closest fit", and this plane corresponds to

$$(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\alpha}_q = 0 \qquad (3)$$

since for $\mathbf{y} \in Y$, deviations $d_q = (\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\alpha}_q$ from this plane have the smallest possible variance, namely λ_q. For any $\mathbf{y}$, $d_q / \sqrt{\lambda_q}$ in equation (2) represents its deviation from the plane defined by equation (3), in standard measure with metric defined by the smallest eigenvalue of Σ. A second-best, q - 1 dimensional hyperplane, which satisfies $\text{corr}[d_q, d_{q-1}] = 0$, is defined by

$$(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\alpha}_{q-1} = 0. \qquad (4)$$

with deviations of $\mathbf{y} - \boldsymbol{\mu}$ from this hyperplane reflected by $d_{q-1} = (\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\alpha}_{q-1}$, or $d_{q-1} / \sqrt{\lambda_{q-1}}$ in standard measure, and so forth. The net result is that D^2 represents a sum of squares of deviations, in standard measure, of a particular point with coordinates given by $\mathbf{y} - \boldsymbol{\mu}$ from each of q, q - 1 dimensional hyperplanes, all of which pass through the point $\mathbf{y} = \boldsymbol{\mu}$ in the original q-dimensional sample space.

Rotenberry, Knick, and Dunn (1999) argued the premise that not all q components of D^2, as partitioned in equation (2), are likely to define limiting combinations of habitat variables for the species. Some q - k of these are included in D^2 simply because q habitat variables were measured or available in the GIS database. For example, Dettmers, *et al.* (1999) initiated their avian habitat analysis with 24 variables. Certainly, the hyperplane $(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\alpha}_1 = 0$, corresponding to the first principal component, logically cannot be considered a limitation since the variance of deviations from this hyperplane is λ_1, the maximum possible. Yet, this deviation makes its contribution to D^2 as commonly defined.

As a result, it was proposed that habitat suitability for a q-dimensional $\mathbf{y}$ be measured by

$$D^2(k) = \sum_{j=q-k+1}^{q} d_j^2 / \lambda_j \qquad (5)$$

for some $1 \leq k < q$, where the eigenvalues of Σ (or its sample analog) are ordered $\lambda_1 \geq \ldots \geq \lambda_q$. Thus, suitability of a particular habitat location $\mathbf{y}$ for a species would be measured in terms of deviations from k basic requirements for that species, to the extent that we are able to know k. In that $D^2(k) \sim \chi^2_{(k)}$ under multinormal assumptions, $p = P[\chi^2_{(k)} > D^2(k)]$ is analogous to a posterior probability resulting from use of either a Bayes discriminant function or logistic regression.

PLANE OF CLOSEST FIT

Insight may be gained by visualizing the habitat constraints implied by each of the components of $D^2(k)$. It is informative to relate variation in each of the variables to the rate of departure from the target planes. For the j^{th} component, this information is contained in the gradient vector

$$\frac{\partial d_j}{\partial \mathbf{y}} = \left[\frac{\partial d_j}{\partial y_1}, \cdots, \frac{\partial d_j}{\partial y_q}\right]' = \boldsymbol{\alpha}_j = \{\alpha_{hj}\}. \qquad (6)$$

The rank order of $|\partial d_j/\partial y_h| = |\alpha_{hj}|$ from large to small for h = 1,...,q suggests the relative importance of the habitat variables to the species requirement defined by $d_j = 0$. Variables with small $|\alpha_{hj}|$ can vary considerably while allowing the habitat to remain close to the requirement. This description is only complete, however, when one realizes that even if both α_{rj} and α_{sj} are large in absolute value, their effects will tend to cancel if they have opposite signs and the habitat still will satisfy the requirement $d_j \cong 0$. This is a perceived advantage of the model since it allows the possibility of detecting that a species can make a trade-off, balancing different habitat variables, while still maintaining habitat utility.

For the above interpretation to be effective, $y_1,...,y_q$ should be in identical units. An obvious approach is to standardize all variables preliminary to the habitat analysis, i.e., replacing **y** with $\mathbf{y}_s = D_\sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})$, where diagonal D_σ displays the standard deviations of the elements of **y**. The net effect is to replace Σ by the correlation matrix, R (Dunn and Duncan, 2000).

VARIABLE SELECTION

There usually are numerous habitat variables available in a GIS database. However, it is quite likely that the species of interest does not choose a particular habitat based on all of these variables. This often causes the researcher to want to reduce the number of measured variables to the set that the species considers most important.

Suppose a tentative k is chosen on the basis of an initial principal components solution, and we wish to test that k species requirements depend only on r habitat variables. Dunn and Duncan (2000) show that r habitat variables can be selected by testing the significance of the component loadings on the last k principal components. The testing is identifiable as one of confirmatory factor analysis and can be performed in PROC CALIS.

EXAMPLE 1: HABITAT PREFERENCES OF THE ACADIAN FLYCATCHER

This illustration was taken from Gullett (2001). Vegetation cover variables were collected for 173 sites in the Ouachita Region of Arkansas using a GIS. There were a total of twelve vegetation cover variables whose values were the percent of the site with a particular vegetation type. Through observational studies, the Acadian Flycatcher was found in 97 of these sites.

CENTERED LOGRATIOS

The compositional nature of the vegetation data led to the use of logratios to describe the habitat. Aitchison (1986) notes that in this type of analysis, centered logratios, as opposed to baseline logratios, should be used.

If f_{ij}, i=1,…,n, j=1,…,q, is the fraction of site i covered with vegetation type j, then the centered logratio for site i with vegetation type j is,

$$y_{ij} = \ln\left(\frac{f_{ij}}{\tilde{f}_i}\right),$$

where $\tilde{f}_i$ = geometric mean of $f_{i1}, f_{i2}, \ldots f_{iq}$. Use of centered logratios automatically yields at least one zero eigenvalue; thus, $D^2(k)$ must be computed as,

$$D^2(k) = \sum_{j=m-k+1}^{m} d_j^2 / \lambda_j ,$$

where m is the number of non-zero eigenvalues of R.

COMPUTATION OF $D^2(K)$ IN SAS

The computation of $D^2(k)$ depends only on d_j and λ_j for a given **y**. Both d_j and λ_j are available in procedure PRINCOMP. The calculation of $D^2(k)$ for all k=1, …, m, where m is the number of non-zero eigenvalues, is performed in a data step. The calculation of the associated chi-square p-value also is performed in the data step.

```
* Call PRINCOMP to compute the PC solution
* for habitats where species is present;
PROC PRINCOMP DATA=PRESENT OUTSTAT=EIGEN;
  VAR habitat variables;
RUN;

* Call SCORE to compute PC scores for all
* data including sites where the species
* was not observed.;
PROC SCORE DATA=ALL
  OUT=SCORES(DROP=habitat variables)
  SCORE=EIGEN;
  VAR habitat variables;
  ID id variables;
RUN;

%LET P=12; *Set P to number of variables;

* Compute D^2(k) for k=1,...,m where m
* is the number of non-zero eigenvalues.;
DATA SCORE;
  IF _N_=1 THEN SET
    EIGEN(WHERE=(_TYPE_='EIGENVAL'));
  SET SCORES;
  ARRAY P[&P] PRIN1-PRIN&P;
  ARRAY LAMBDA[&P] habitat variables;
  ARRAY D[&P] D2_1-D2_&P;
  ARRAY DC[&P] D2C_1-D2C_&P;
  ARRAY PV[&P] PVALUE1-PVALUE&P;
  ARRAY DF[&P] DF1-DF&P;
  K=0;
  DO I=&P TO 1 BY -1;
    IF LAMBDA[I] NE 0 THEN DO;
      K=K+1;
      IF K NE 1 THEN DO;
        DC[I]=(P[I]**2)/LAMBDA[I];
        D[I]=DC[I]+D[I+1];
      END;
      ELSE IF K=1 THEN DO;
        DC[I]=(P[I]**2)/LAMBDA[I];
        D[I]=DC[I];
      END;
      PV[I]=1-PROBCHI(D[I],K);
      DF[I]=K;
    END;
  END;
  KEEP X Y PRIN1-PRIN&P D2C_1-D2C_&P
       D2_1-D2_&P PVALUE1-PVALUE&P DF1-DF&P;
RUN;
```

The eigenvalues and eigenvectors output from PROC PRINCOMP were used to help determine k. The p-values from the data step were used to create maps based on several values of k. Based on these maps, k=3 was chosen.

VARIABLE SELECTION

The initial principal components solution produced eigenvalues λ_1 = 5.315, ...,λ_8 = 0.160, λ_9 = 0.115, λ_{10} = 0.105, λ_{11} = 0.028, λ_{12}=0, so that k = 3 was chosen on the basis of subjective

judgment. The eigenvectors 9, 10, and 11 are reproduced in table 1.

Table 1 Eigenvectors 9, 10, and 11 from Acadian Flycatcher data.

	(α_9)	(α_{10})	(α_{11})
LR1	**0.644440 (6)**	0.046337 (2)	0.031315 (2)
LR2	-0.029892 (2)	0.032734 (2)	**-0.764507 (9)**
LR4	**-0.683056 (8)**	-0.072230 (2)	-0.030420 (2)
LR5	-0.101627 (2)	0.258358 (2)	**0.583588 (4)**
LR6	0.019247 (1)	0.047659 (1)	-0.010451 (1)
LR10	-0.065695 (2)	-0.325416 (3)	0.252936 (2)
LR11	0.258443 (1)	-0.073353 (1)	-0.010231 (1)
LR12	0.084056 (2)	**-0.625734 (5)**	0.075962 (2)
LR17	0.058122 (1)	0.027835 (1)	0.000670 (1)
LR18	-0.025355 (2)	**0.644464 (7)**	-0.003296 (2)
LR19	0.101714 (1)	0.010231 (1)	0.044381 (1)
LR20	0.118314 (1)	0.048385 (1)	0.033271 (1)

In order to select variables, confirmatory factor analysis needed to be performed using PROC CALIS. When there is no singularity problem, the confirmatory factor analysis is conducted by (1) inputting the factor loadings of the first q-k factors into the _F_ matrix in PROC CALIS, (2) zeroing suspected non-significant factor loadings in the last k factors of the _F_ matrix, and (3) allowing PROC CALIS to estimate the factor loadings that are considered significant. If the X^2 statistic is non-significant, then it is safe to exclude the variables for which the factor loadings were zero on all of the last k factors.

With this data, the singularity problem caused by using centered logratios prevents the use of PROC CALIS directly because there are only eleven vectors of factor loadings and PROC CALIS expects twelve vectors. The use of an adjusted correlation matrix allowed PROC CALIS to proceed.

We use the fact that **j**, a column vector of 1's, is an eigenvector associated with the zero eigenvalue of R in order to restore R to full rank. Algebraically, from the identity

$$R + c_1\mathbf{j}\mathbf{j}' = [\Lambda \vdots c_2\mathbf{j}][\Lambda \vdots c_2\mathbf{j}]' = \Lambda\Lambda' + c_2^2\mathbf{j}\mathbf{j}',$$

if we add any scaler $c_1>0$ to all elements of R, then the first q-1 columns of component loadings are unchanged and the last column of loadings is the constant $c_2 = \sqrt{c_1}$. The non-zero eigenvalues of rank-adjusted R are unchanged, while the zero eigenvalue is replaced by $\lambda=c_1q$, which is implied by

$$[(R+c_1\mathbf{j}\mathbf{j}')-\lambda I]\mathbf{j} = \mathbf{0}$$

since $R\mathbf{j}=\mathbf{0}$. In order to identify this artificial component as the q^{th} component, we choose c_1q to be slightly less than the smallest non-zero eigenvalue of the original R matrix.

We have no means of adjusting R if PROC CALIS is provided with the raw data. However, PROC CALIS accepts TYPE=CORR and TYPE=COV data as input. When we adjust the correlation matrix, the diagonal is no longer only ones, and PROC CALIS will not accept this data set as TYPE=CORR. However, after the R matrix is adjusted, PROC CALIS will accept it as input if it is labeled as TYPE=COV. The following steps will allow PROC CALIS to accept the adjusted R matrix:

1. Use procedure CORR to generate a TYPE=CORR data set that contains the R matrix.
2. Use a data step to make a TYPE=COV data set which is created by setting the TYPE=CORR data set and changing the _TYPE_ variable from CORR to COV.
3. Use the TYPE=COV data set as input to PROC CALIS.

The SAS program shown below performs these steps. It also uses a macro to write the factor scores to PROC CALIS.

```
%LET P=12; *** P IS NUMBER OF VARIABLES;

***  Step 1;
***  Creating a TYPE=CORR data set;
PROC CORR DATA=LR OUTP=CORR(TYPE=CORR)
     NOPRINT;
  VAR LR1 LR2 LR4 LR5 LR6 LR10 LR11 LR12
      LR17 LR18 LR19 LR20;
RUN;

***  Step 2;
***  Adjusting the correlation matrix,
***  calling it a covariance matrix, and
***  and changing the data set to TYPE=COV;
DATA COV(TYPE=COV);
  SET CORR(TYPE=CORR);
  ARRAY L[12] LR1--LR20;
  IF _TYPE_='CORR' THEN DO I=1 TO 12;
     _TYPE_='COV'; L[I]=L[I]+.0004;
  END;
RUN;

%LET K=3;  ***  Choose k;

***  Calling FACTOR to generate and
***  capture the factor loadings;
PROC FACTOR METHOD=PRIN DATA=LR OUTSTAT=PAT
            EIGENVECTORS NFACT=11 NOPRINT;
  VAR LR1 LR2 LR4 LR5 LR6 LR10 LR11 LR12
      LR17 LR18 LR19 LR20;
RUN;

***  Rearranging the loadings data for the
***  _F_ matrix;
DATA PAT;
  SET PAT(WHERE=(_TYPE_='PATTERN'));
  DROP _TYPE_;
RUN;
PROC TRANSPOSE DATA=PAT OUT=PAT1;
RUN;

***  Creating macro variables that contain
***  the factor loadings;
%MACRO SP;
  %DO I=1 %TO &P;
    CALL SYMPUT
    (TRIM(LEFT(PUT(_NAME_,4.)))||"_"
          ||"&I",FACTOR&I);
  %END;
%MEND SP;
DATA _NULL_;
  SET PAT1;
  N=_N_;
  %SP;
RUN;

***  This macro writes the factor loadings
***  into the _F_ matrix in CALIS;
%MACRO SP1(V);
  %DO I=1 %TO %EVAL(&P-&K-1);
    &&&V&I
  %END;
%MEND SP1;

***  Step 3;
***  Calling CALIS for the confirmatory
***  factor analysis.  Note that the data
***  set is the adjusted correlation matrix
***  and is TYPE=COV.  The macro %SP1( )
***  will write the factor loadings for
***  first p-k-1 factors.  The 0.02 comes
***  from adjusting the correlation matrix;
PROC CALIS DATA=COV(TYPE=COV)
PCORR MAXITER=5000 MAXFUNC=5000;
FACTOR N=12 COMPONENT;
  MATRIX _F_
       [1, ]=%SP1(LR1_) a9 0 0 .02,
```

```
        [2, ]=%SP1(LR2_) 0 0 b11 .02,
        [3, ]=%SP1(LR4_) c9 0 0 .02,
        [4, ]=%SP1(LR5_) 0 0 d11 .02,
        [5, ]=%SP1(LR6_) 0 0 0 .02,
        [6, ]=%SP1(LR10_) 0 0 0 .02,
        [7, ]=%SP1(LR11_) 0 0 0 .02,
        [8, ]=%SP1(LR12_) 0 h10 0 .02,
        [9, ]=%SP1(LR17_) 0 0 0 .02,
        [10, ]=%SP1(LR18_) 0 j10 0 .02,
        [11, ]=%SP1(LR19_) 0 0 0 .02,
        [12, ]=%SP1(LR20_) 0 0 0 .02;
  VAR LR1 LR2 LR4 LR5 LR6 LR10 LR11 LR12
      LR17 LR18 LR19 LR20;
  TITLE1 'Acadian Flycatcher, k=3';
RUN;
```

In table 1, numbers in parenthesis indicate the order in which confirmatory tests were performed using PROC CALIS (*SAS OnlineDoc® Version* 8) with all tests based on 66 degrees of freedom (df) except the first one which was based on 52 df. Step (1) attempted to zero entire rows LR6, LR11, LR17, LR19 and LR20, yielding $X^2 = 2.2628$ with p = 1.0. This suggests that these 5 logratios do not contribute to the 3 habitat requirements of the species. In addition to the zero elements specified in step (1), the hypothesis of step (2) also specified certain zero elements in those rows containing at least one sizable element, e.g., LR1. Again, this hypothesis was accepted since $X^2 = 8.8117$, p = 1.0. Step (3) gave a $X^2 = 17.5215$, p = 1.0. Steps (4) - (9) included the hypotheses of steps (1) (2) and (3) as well as zeroing one additional element. X^2 = 147.1, 609.5, 816.2, 409.0, 669.0, and 167.7 respectively, in steps (4) - (9) with $p < 0.0001$ in every case, thus leading to retention of the 6 boldface entries shown above in table 1.

The geometric mean and the centered logratios were recomputed for the six retained variables, LR1, LR2, LR4, LR5, LR12, and LR18. A principal components analysis of these variables yielded eigenvalues $\lambda_3 = 0.236$, $\lambda_4 = 0.181$, and $\lambda_5 = 0.0270$ with associated eigenvectors shown in table 2.

Table 2 Eigenvectors 3, 4, and 5 from reduced set of variables for the Acadian Flycatcher data.

	(α_3)	(α_4)	(α_5)
LR1	**0.678593**	-0.162090	0.076478
LR2	0.069868	0.027896	**-0.665625**
LR4	**-0.663151**	0.174703	-0.026249
LR5	-0.045755	0.049612	**0.739637**
LR12	0.303344	**0.683515**	0.040161
LR18	-0.027383	**-0.687585**	0.041589

Aside from an immaterial reflection of signs in one case, the eigenvalues and eigenvectors for the reduced set of 6 variables closely mimic those of the original set of 12 variables. These three eigenvectors ultimately were the basis for $D^2(3)$ used to map potential habitat of the Acadian flycatcher. The boldface elements suggest that Acadian flycatcher habitat requirements will be met if (a) all 6 logratios are close to their respective means, or if (b) LR1 and LR4 simultaneously vary in the same direction from their means, or if (c) LR12 and LR18 simultaneously vary in the same direction from their means, or if (d) LR2 and LR5 simultaneously vary in the same direction from their means, or if (e) any combination of (b), (c), or (d) holds.

USING THE EMPIRICAL CDF FROM PROC KDE

Figure 1 shows an overlay of the empirical c.d.f. of $D^2(3)$ for the 97 known Acadian Flycatcher sites, its kernel density smooth by means of PROC KDE, and the c.d.f. of the approximating $\chi^2_{(3)}$. We used 51 bins and a bandwidth multiplier of 1.3 for the kernel density estimation shown here. It is typical that the empirical c.d.f. tends to accumulate faster than the normal-based chisquared distribution, suggesting that one or both tails of the values defining $D^2(3)$ have been trimmed in the calibration data set. This corresponds to preferred habitat being in the middle of the ranges of the habitat variables, a situation where both Bayes linear discriminant functions and logistic regression tend to perform poorly.

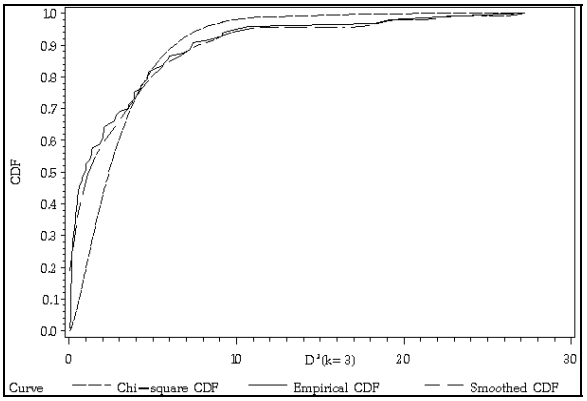

Figure 1 PROC KDE was used to smooth the empirical c.d.f. In this case the bandwidth multiplier was 1.3 using 51 bins.

EXAMPLE 2: HABITAT PREFERENCES OF THE TIMBER RATTLESNAKE

The following illustration is taken from Browning (2000). Thirty-six Timber Rattlesnakes were tracked and their dens located in Madison County, Arkansas. A GIS was used to obtain habitat data at each of the den locations where a GPS was used to find precise coordinates of the dens. The habitat data collected included elevation, slope, aspect, relief, and various soil characteristics for a total of sixteen variables. The nature of the soil characteristic data led to linear dependencies among the variables, and thus, to zero eigenvalues of the correlation matrix.

The last seven eigenvalues were zero, so that the computation of $D^2(k)$ was based only on the first nine eigenvalues. Because the largest two eigenvalues were much larger than the others (see table 3), the computation of $D^2(k)$ was initially based on principal components 3 through 9.

Table 3 Eigenvalues of the correlation matrix for the sixteen habitat variables for the Timber Rattlesnake data.

	Eigenvalue	Difference
1	8.5815	5.3295
2	3.2519	1.5118
3	1.7401	0.6691
4	1.0711	0.4912
5	0.5798	0.1573
6	0.4225	0.2245
7	0.1981	0.1076
8	0.0905	0.0260
9	0.0645	0.0645
10-16	0	0

This solution assigned p-values of greater than 0.05 to all dens except dens 14 and 18. This indicated that all dens except these two were in suitable habitats; however, dens 14 and 18 must also have been suitable since there were confirmed rattlesnake dens present at these sites.

BOOTSTRAPPING AND CROSSVALIDATION

Crossvalidation was performed on the principal component scores to find if any den had undue influence on the eigenvalues and eigenvectors.

Let $d_j(y_i)$ be the j^{th} principal component score for den i when all dens are used in the principal components analysis. Let $d_j(y_{(i)})$ be the j^{th} principal component score for den i when den i is excluded from the principal components analysis. Similarly, $\lambda_{j\,(i)}$ is the j^{th} eigenvalue from the principal components analysis when

den i is excluded. The measure of influence used was

$$\text{dffits pc } j_i = \left| \frac{d(y_i) - d(y_{(i)})}{\sqrt{\lambda_{j(i)}}} \right| \text{, where } j = 1, \ldots, 9.$$

Crossvalidation on the pc scores indicated that when either den 14 or 18 was excluded from the data, eigenvalue 9 was zero.

Bootstrapping techniques were used to obtain approximate 90% confidence intervals for the p-values associated with $D^2(k)$. The steps in the bootstrapping follow:

1. A random sample of size 36 was drawn with replacement from the 36 dens.
2. $D^2(k)$ was computed for all 36 dens based on the sample from step 1.
3. Steps 1 and 2 were repeated 1000 times.
4. The mean of the p-value associated $D^2(k)$ for each of the dens was computed, and the upper and lower 90% confidence intervals were taken as the 5 and 95 percentiles of the p-value associated with $D^2(k)$ for each of the dens.

In twelve instances in the bootstrapping, eigenvalues 7-16 were all zero; in 197 instances, eigenvalues 8-16 were all zero; and in 662 instances, eigenvalues 9-16 were all zero.

In addition to considering the bootstrapping and the crossvalidation, the contribution of each component of D^2 was generated, and the largest contribution to den 14 came from the first D^2 component (associated with pc 9), and the largest contribution to den 18 came from the third D^2 component (associated with pc 7). Based on the results from crossvalidation and bootstrapping and from the contributions of the first and third components of D^2, we decided not to use the contributions of principal components 7, 8, and 9 in the computation of D^2.

$D^2(k=4)$ was computed based on principal components 3 through 6; $D^2(k=5)$ was computed based on principal components 2 through 6; and $D^2(k=6)$ was computed based on principal components 1 through 6. The p-values associated with $D^2(k)$ were based on χ^2 tests with k degrees of freedom.

$D^2(k=6)$ gave p-values of greater than 0.01 for all 36 dens; however, it used the first principal component which accounts for the most variation in the data, i.e., the least likely to represent a species requirement. $D^2(k=5)$ gave more p-values less than 0.10 than did $D^2(k=4)$; thus, we did not consider k=5 to be useful. $D^2(k=4)$ produced five p-values between 0.03 and 0.08, and the other p-values were all greater than 0.10. The choice of a decision-making-p-value does not need to be 0.05, and since we know that dens were found on all 36 sites, we can assign our decision-making-p-value to be 0.03. To make the choice between k = 4 or 6, we produced maps of the area based on both solutions, and based on the researchers' knowledge of the region, k = 4 was the final choice.

MAP CREATION

The GIS data for this example were collected using Geographic Resource Analysis Support Software (GRASS, v. 4.1), and the data were put into an ASCII file and read into SAS using a standard datastep. The data included all of the habitat variables as well as the x,y coordinates of the dens. To create a map of the entire county, the habitat data for each pixel on the map were collected in an ASCII file along with the x,y coordinates, and these data were also read into SAS. After $D^2(k)$ was computed, the x,y coordinates and their associated p-values were exported to an ASCII file, and they were read into Idrisi32 to create the maps. Moving data from platform to platform and in and out of SAS seemed quite cumbersome. This motivated the use of SAS/GIS to create the maps for different values of k.

The principal component solution from the original 36 dens was computed in PROC PRINCOMP and was used in procedure SCORE to generate the principal component scores for every pixel on the map. Next, a data step computed $D^2(k)$ and corresponding chi-square p-values for all of the values of k in which we were interested for every pixel on the map and for the known dens. A data set was created that contained the following variables for k = 4 and 6:

1. ID — each pixel's unique id number
2. X — x-coordinate
3. Y — y-coordinate
4. LAYER — _4DF, _6DF, or DEN
5. PVALUE_4 df — p-value for each pixel based on 4 df
6. PVALUE_6 df. — p-value for each pixel based on 6 df.

The following code shows how these data were read into SAS/GIS as generic point data, and a map was subsequently generated with each of the values of LAYER forming a different layer. The p-value variables formed the theme for the _4DF and _6DF layers. The program used batch importing to create the initial map, and it used procedure GIS to create and modify the themes. Figure 2 shows the end result, which was a single map with the dens and the p-value layers. In order to view the map for the different values of k, one need only make the desired layer active and the other layers hidden.

```
***  Importing the data into GIS
***  using an automatic SAS macro;
%LET IMP_TYPE=GENPOINT;
%LET INFILE=DAWN.FULL_MAP;
%LET NIDVARS=1;
%LET IDVAR1=LAYER;
%LET MAPLIB=DAWN;
%LET MAPCAT=MAPS;
%LET MAPNAME=D2_FULL1;
%LET CATHOW=CREATE;
%LET SPALIB=DAWN;
%LET SPANAME=D2_FULL1;
%LET SPAHOW=CREATE;

DM 'AF C=SASHELP.GISIMP.BATCH.SCL';

*** Creating a theme variable based on the
*** p-value variable.;

PROC GIS CATALOG=DAWN.MAPS;
  SPATIAL DAWN.MAPS.D2_FULL1;
  LAYER UPDATE DENS/
    DEFAULT=(POINT=(SIZE=5
                    COLOR=RED
                    FONT=MARKER
                    CHARACTER='W'))
  RUN;
  LAYER UPDATE _4DF/
    THEMATIC
    DEFAULT=(
      POINT=(SIZE=1))
    THEME=(
      CREATE
      DATASET = DAWN.FULL_MAP
      THEMEVAR = PVALUE_4
      DATAVAR = ID
      COMPOSITE = ID
      LINK = FULL_MAP
      RANGE=DISCRETE
      POINT=((LEVEL=1 COLOR=VLIBG SIZE=1)
             (LEVEL=2 COLOR=VLIYG SIZE=1)
               more lines for each p-value
             (LEVEL=20 COLOR=VDAG SIZE=1)
             (LEVEL=21 COLOR=VDABG SIZE=1)
               ));
  RUN;
  LAYER UPDATE _6DF/
    THEMATIC
    DEFAULT=(
      POINT=(SIZE=1))
    THEME=(
      CREATE
      DATASET = DAWN.FULL_MAP
```

```
        THEMEVAR = PVALUE_6
        DATAVAR = ID
        COMPOSITE = ID
        LINK = FULL_MAP
        RANGE=DISCRETE
        POINT=((LEVEL=1 COLOR=VLIBG SIZE=1)
               (LEVEL=2 COLOR=VLIYG SIZE=1)
                 more lines for each p-value
               (LEVEL=19 COLOR=VDAYG SIZE=1)
               (LEVEL=20 COLOR=VDAG SIZE=1)
                 ));
  RUN;
QUIT;
```

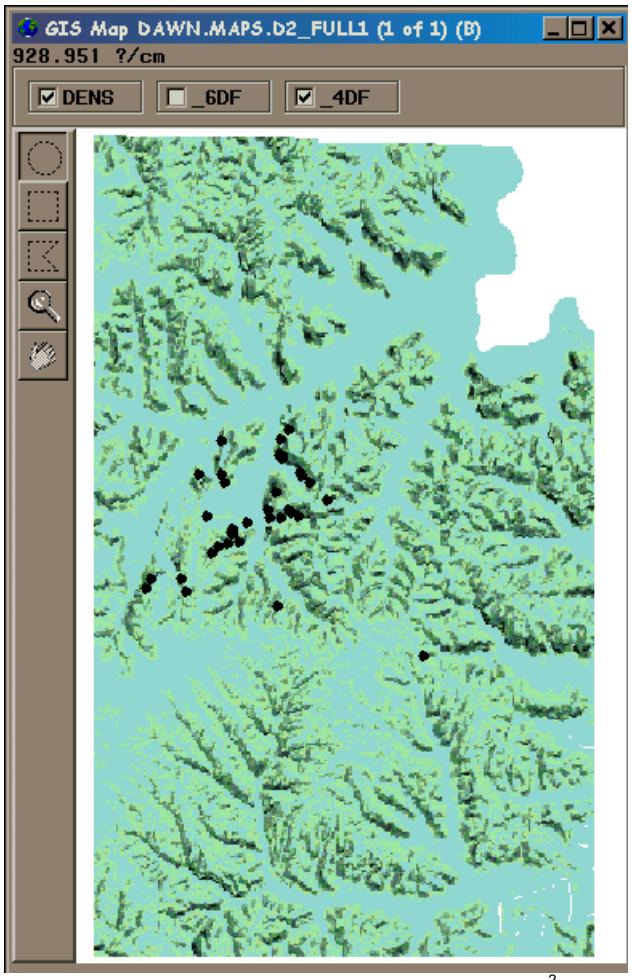

Figure 2 SAS/GIS map of the p-values associated with $D^2(4)$ for Timber Rattlesnakes in Madison County, Arkansas. Dots represent snake dens.

CONCLUSION

As the availability of data increases, the need for methods to analyze and summarize the data also increases. In the context of habitat selection, there are many variables that can be measured. However, the interest of most studies of this type is to find what is important to the species. The use of partitions of Mahalanobis D^2 allows the researcher to focus on what the species requires.

REFERENCES

Aitchison, J. (1986), *Statistical Analysis of Compositional Data*, London and New York: Chapman and Hall, Inc.

Browning, Dawn M. (2000), *A comparison of a modified form of Mahalanobis Distance Statistic and Boolean methods to devise GIS-based models of the denning habitat of timber rattlesnakes (Crotalus horridus) in northwest Arkansas*, M.S. Thesis, University of Arkansas, Fayetteville, Arkansas, 63 pp.

Dettmers, R., Buehler, D.A., and Bartlett, J.G. "A test and comparison of wildlife-habitat modeling techniques for predicting occurrence on a regional scale," *Predicting Species Occurrence: Issues of Scale and Accuracy*, conference October 18-22, 1999, Snowbird, Utah.

Dunn, J.E., and Duncan, L. (2000) "Partitioning Mahalanobis D^2 to sharpen GIS classification", *Management Information Systems 2000: GIS and Remote Sensing,* C.A. Brebbia and P. Pascolo, ed., WIT Press, pp. 195-204.

Johnson, D.E., (1998), "Distances and Angles" (Appendices A.3 & A.4*). Applied Multivariate Methods for Data Analysis*, London and New York: Duxbury Press, 528-529.

Gullett, B.W. (2001) *Combining Arkansas breeding bird atlas and GAP vegetation data to predict avian distributions within the Ouachita Mt. Physiographic Region*. Thesis submitted to the University of Arkansas, Dept. of Biological Sciences.

Pearson, K. (1901), "On lines and planes of closest fit in systems of points in space," *Philosophical Magazine*, 2, 559-572.

Rotenberry, J.T., Knick, S.T., and Dunn, J.E. (1999), "A minimalist approach to mapping species habitat: Pearson's planes of closest fit," *Predicting Species Occurrence: Issues of Scale and Accuracy*, conference, Snowbird, Utah.

SAS Institute, Inc. (2000), *SAS OnlineDoc Version 8,* Cary, NC: SAS Institute, Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Lynette Duncan
University of Arkansas
Department of Mathematical Sciences
538 HOTZ
Fayetteville, AR 72701
(501) 575-6429
duncan@uark.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.