QUESTIONABLE MULTIVARIATE STATISTICAL INFERENCE IN WILDLIFE HABITAT AND COMMUNITY STUDIES

ERIC A. REXSTAD, Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, CO 80523
DIRK D. MILLER,[1] Department of Fishery and Wildlife Biology, Colorado State University, Fort Collins, CO 80523
CURTIS H. FLATHER, Department of Fishery and Wildlife Biology, Colorado State University, Fort Collins, CO 80523
ERIC M. ANDERSON,[2] Department of Fishery and Wildlife Biology, Colorado State University, Fort Collins, CO 80523
JERRY W. HUPP,[3] Department of Fishery and Wildlife Biology, Colorado State University, Fort Collins, CO 80523
DAVID R. ANDERSON, Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, CO 80523

Abstract: We analyzed a data set constructed from functionally unrelated, easily collected observations (e.g., meat, stock, and liquor prices) around Fort Collins, Colorado, using principal components analysis (PCA), canonical correlation analysis (CC), and discriminant function analysis (DFA). Each produced seemingly significant results and suggested strong relationships between the variables measured. We suggest that multivariate techniques can provide invalid inferences when used with data containing no relationships. We question the use of these techniques in studies of wildlife habitat.

J. WILDL. MANAGE. 52(4):794–798

Studies of wildlife habitat often attempt to identify biotic and abiotic variables that can be used to evaluate niche separation among species or assess habitat suitability for a single species. A common approach is to compare characteristics of sites used by a species to random sites or sites where the species is believed not to occur. Because biologists often have limited knowledge of a species's niche, a priori hypotheses regarding animal relationships to specific habitat variables are unstated. Instead, variables are measured in hope of identifying variables that are meaningful in describing a species's habitat or niche.

Because of the multidimensional nature of habitat and community studies, researchers have increasingly used multivariate statistical procedures to identify variables that may be useful in habitat assessment. Three multivariate procedures are generally used in analysis of habitat data. Principal components analysis is used to identify variables that account for maximum variation in data and to produce a smaller number of uncorrelated variables that are linear combinations of original variables (Bhattacharyya 1981, Seber 1984:176–203). Canonical correlation evaluates correlation between ≥ 2 groups of variables by means of linear combinations of those variables (Gittins 1985:4–9). Discriminant function analysis develops linear models of variables that maximize the separation of groups (e.g., use vs. random sites) and classify additional observations (Williams 1981).

Two problems may arise when multivariate procedures are used in a posteriori searches for linear combinations of variables in the hope of learning about the structure or dynamics of a system. First, obvious relationships may not be detected. Armstrong (1967) illustrated this problem with data from a simple, deterministic geometry model. He studied 63 metal objects of varying sizes in a system that was completely described by 5 variables (length, width, ht, density, and cost). He used PCA and factor analysis to select factors that explained the system. A large proportion of the variance was explained; however, the methods failed to identify the 5 variables that truly described the system. It seems surprising that these popular methods of analysis failed to identify the obvious variables and relationships. A second problem, which is the converse of the first, is that meaningless variables yield apparently meaningful linear equations. Spurious relationships among the data may cause a researcher to incorrectly conclude that certain variables are useful or important.

Our objective is to illustrate the latter of the 2 problems. To demonstrate that apparently significant results can be derived from meaningless data, we examined a collection of 15 unrelated variables using analysis techniques employed in

[1] Present address: Wyoming Game and Fish Department, Laramie, WY 82070.
[2] Present address: Center for Environmental Studies, Unity College, Unity, ME 04988.
[3] Present address: U.S. Fish and Wildlife Service, Alaska Fish and Wildlife Research Center, 1011 E. Tudor Road, Anchorage, AK 99503.

wildlife research. We evaluated the extent to which PCA would explain the variance of the data set and reduce its dimensionality, and tested the null hypotheses that CC would not show high correlation between 2 sets of variables, and classification accuracy using DFA would be no greater than that expected by chance. We attempted to illustrate some common errors in the use and interpretation of multivariate statistical procedures, and raise general questions about the use of these methods as valid inference procedures for habitat and community studies.

We thank W. J. deVergie, E. J. Hernandez, C. W. Jeske, O. B. Myers, K. E. Ragotzkie, S. S. Rosenstock, J. J. Rotella, D. G. Saltz, J. A. Schmutz, M. A. Schroeder, G. Schwab, and T. M. Shenk for their assistance with data collection and their discussion of multivariate statistical procedures. We also thank D. E. Capen, D. H. Johnson, and T. E. Martin for reviewing an earlier draft of this manuscript.

METHODS

The median sample size of 28 multivariate data sets published in *The Journal of Wildlife Management* during 1985–87 was 99 observations and 12 variables. We collected 75 observations on 5 "biotic" variables: students' grades, meat prices, package weights of hamburger, sagebrush (*Artemisia* spp.) cover, and produce prices, and 10 "abiotic" variables: stock prices, used car prices, book pages, street addresses, greeting card prices, telephone numbers (last 5 digits), annual mean temperatures, pitching earned run averages (ERA), diamond ring prices, and liquor prices. Each variable was collected by a different individual; therefore, each variable is independent of the others. Because multivariate statistics often assume data are continuous and normally distributed, our variables were selected to represent essentially continuous data, even though published studies often include discrete, binary, or rank data. Although our data were not distributed normally, we believed that continuous data would have less effect on the analytical techniques and interpretation of results than discrete or binary data. The data were analyzed with PCA and CC using the Statistical Analysis System (SAS) (Joyner 1985), and DFA using the Statistical Package for the Social Sciences (SPSS) (Norusis 1986).

Discriminant function analysis was performed by randomly assigning 38 observations to membership in 1 group and 37 to another. Important discriminating variables were selected by stepwise procedures (Norusis 1986:B17–22). No group assignments were associated with the rows of the data matrix because of the method of data collection. Fifty replicates of randomly assigned group membership were run to assess the mean number of variables selected, mean classification rate, number of times homogeneity of covariance (Box's M) was rejected, and number of times group differences (Wilks's λ) were significant.

RESULTS

We computed Pearson product-moment correlation coefficients for all possible pairs of variables. Only 8 of the 105 (7.6%) correlation coefficients were significant ($P < 0.05$). This is consistent with the meaningless nature of the data; i.e., no more relationships were found or suggested than would be expected from chance alone.

Principal components analysis was performed using the correlation matrix and the variance–covariance matrix. Analysis of the correlation matrix generally is used when the variables are of very different units of measure; however, the distributional properties of this analysis are poorly understood (Seber 1984:188). The PCA of the correlation matrix (Table 1) identified 7 components with eigenvalues >1.0, which explained 69% of the variance present in the data matrix. Factor loadings on the first principal component (PC) were highest for street addresses, pitching ERA, students' grades, liquor prices, meat prices, and used car prices. Principal components analysis of the variance–covariance matrix required 2 components to explain 99.5% of the variance in the data set. The first component had an eigenvalue of 241,602,064 and was loaded nearly exclusively on telephone numbers. The second component had an eigenvalue of 33,561,419 and was loaded highly on used car prices (Table 1).

Canonical correlation analysis of the 6 "biotic" versus the 10 "abiotic" variables (Table 2) had a canonical correlation coefficient of 0.601. This correlation coefficient was significant at $\alpha = 0.10$ (Wilks's $\lambda = 0.3777$; 50,277 df; $P = 0.087$). The CC loadings resulted in a linear combination for the "biotic" variables that was most highly weighted by students' grades and sagebrush canopy cover, and a linear combination of "abiotic" variables that was most highly weighted by street addresses.

Table 1. Performance of principle components (PC) analysis of a Fort Collins, Colorado, data set. Two analyses were conducted: 1 on the correlation matrix and 1 on the variance–covariance matrix.

	Correlation matrix							Variance–Covariance matrix	
Performance	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 1	PC 2
Eigenvalue	2.27	1.62	1.52	1.33	1.29	1.22	1.05	241,602,064	33,561,419
Proportion of variance	0.15	0.11	0.10	0.09	0.09	0.08	0.07	0.87	0.12
Cumulative proportion of variance	0.15	0.26	0.36	0.45	0.53	0.62	0.69	0.87	1.00

Thirty-four of the 50 (68%) stepwise DFA models were significant (Wilks's $\lambda < 0.05$). The mean percent correct classification of these models was 68.2% (range = 57.3 for a 3-variable model to 77.3% for a 10-variable model) (Fig. 1). A median of 6 variables entered the DFA models. The κ-statistic (Titus et al. 1984), was significantly different from zero in 49 of the 50 replicates. This statistic measures the improvement of the DFA model over chance group assignment and ranges from zero, which indicates no improvement, to 1, which indicates perfect assignment. Box's M-test for equality of variance–covariance matrices was rejected for 35 of the 50 (70%) random group assignments. The pooled Box's M-statistic showed that the variance–covariance matrices were not equal ($\chi^2 = 513.2$, 100 df, $P < 0.0005$).

Four of the 50 models obtained correct classification rates >77%. Of these, 1 group assignment met the assumption of equality of variance–covariance matrices. This model incorporated 8 variables, with Wilks's $\lambda = 0.596$ ($P < 0.001$) and Box's $M = 47.81$ ($P = 0.220$). The discriminant function was most highly correlated with book pages, sagebrush cover, students' grades, mean temperatures, and street addresses (Table 3).

DISCUSSION

Principal components analysis reduced the dimensionality and explained a high proportion of the variance and our hypotheses concerning CC and DFA were rejected. The first CC coefficient between "biotic" and "abiotic" variables was 0.601, and group membership was correctly classified by DFA at a rate greater (68%) than that explained by chance. These results are disconcerting given their similarity to values reported in the literature on wildlife habitat, and they raise important questions about valid inference from such analyses.

Seven PC's based on the correlation matrix explained 69% of the variance in the original data. This seemingly important result was due to chance correlation among the variables. However, the sphericity test (Hope 1969:63), which tests the equality of all eigenvalues, failed to reject the hypothesis that all PC's are of equal magnitude; i.e., all PC's explain an equal amount of variation in the data, characteristic of random data. Interpretation of factor loading on the first PC therefore was futile. Based on the variance-covariance matrix, 2 principal components explained 99.5% of the total variance. This was because of the large variance in 2 variables (telephone numbers and used car prices) relative to other variables. We observed a 53–87% reduction in the number of variables when we used totally spurious data. This raises an important question: what criteria should be used to assess the number of important factors, or the interpretation of those factors with biological data?

The CC results appeared to suggest a strong relationship ($r = 0.601$) between "biotic" and "abiotic" variables. Rejection of the null hy-

Table 2. Performance of canonical correlation (CC) analysis of a Fort Collins, Colorado, data set. The first 5 variables in the data set were "biotic," and the remaining 10 variables were "abiotic."

Performance	CC1[a]	CC2	CC3	CC4	CC5
Eigenvalue	0.567	0.285	0.182	0.059	0.050
Proportion of variance	0.496	0.249	0.159	0.052	0.044
Cumulative proportion of variance	0.496	0.745	0.904	0.956	1.000
Canonical correlation	0.601	0.471	0.392	0.237	0.219
Approximate F[b]	1.319	0.962	0.745	0.493	0.536
Significance of F[b]	0.087	0.536	0.800	0.933	0.779

[a] Canonical correlates.
[b] Likelihood ratio test. H_0: canonical correlation is zero.

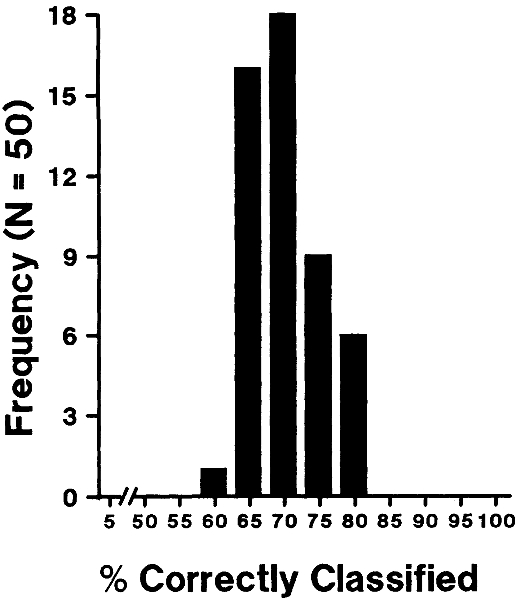

Fig. 1. Histogram of correct classification rates for stepwise discriminant function models from a Fort Collins, Colorado, data set using 50 random group assignments.

pothesis (H_0: $\rho = 0$) at $\alpha = 0.10$ may cause researchers to conclude that there was a relationship between these sets of unrelated data.

Like PCA and CC analyses, DFA showed seemingly impressive results. Four models obtained correct classification rates >77%. These results were greater than expected for a random data set with 75 observations fitted with a discriminant function. Like PCA, questions arise concerning the interpretation of discriminant functions. Interpretation is generally accomplished 2 ways: considering the relative size of standardized coefficients, or determining which variables are most closely correlated with the discriminant scores. Because the 2 techniques often do not identify the same important variables (Table 3), interpretations may be conflicting.

Researchers should remember the definition of Type I error in statistical analyses; i.e., the probability of rejecting a true hypothesis. This risk is determined by setting the α-level. Our results demonstrate the dangers of committing this type of error in multivariate analyses with data sets of the magnitude often used in studies of wildlife habitats and communities. Given that there were no relationships in the data we collected, all 3 hypotheses we posed should have been accepted, but were not.

The application of multivariate techniques to

Table 3. Standardized coefficients and correlation coefficients of stepwise discriminant function analysis of a Fort Collins, Colorado, data set (random group assignment 14 with Wilks's $\lambda = 0.5964$ [$P < 0.001$] and Box's $M = 47.81$ [$P = 0.2203$]).

Variables	Standardized coefficients	Correlation
Students' grades	−0.31447	−0.28229
Meat prices	−0.22746	−0.20696
Meat weights	0.47877	0.15551
Sagebrush cover	0.61193	0.43347
Produce prices	0.33445	0.26423
Book pages	−0.49705	−0.44752
Street addresses	−0.24607	−0.27265
Mean temperatures	0.53073	0.27612

the Fort Collins data clearly shows that statistical and biological significance are not necessarily related. Selection of variables and choice of statistical procedures will have a profound effect on the conclusions reached by a study (Johnson 1981). Additionally, interpretation of coefficients from multivariate procedures is often arbitrary and not well founded in statistical theory. Sophisticated multivariate techniques cannot be expected to replace careful a priori thinking and design.

Multivariate techniques often are applied in habitat and community studies in an exploratory way to illuminate ecologically meaningful relationships. This study, along with the work of Armstrong (1967), represent extremes in the spectrum of problems that can arise through use of multivariate methods. Armstrong (1967) used known, deterministic relationships that multivariate techniques were unable to discern. Our data, containing no inherent relationships, had relationships manufactured by the statistical methods. It seems questionable that multivariate techniques, as currently applied, can consistently identify meaningful relationships in complex, natural systems.

LITERATURE CITED

ARMSTRONG, J. S. 1967. Derivation of theory by means of factor analysis, or Tom Swift and his electric factor analysis machine. Am. Stat. 21: 17–21.

BHATTACHARYYA, H. 1981. Theory and methods of factor analysis and principal components. Pages 72–79 *in* D. E. Capen, ed. The use of multivariate statistics in studies of wildlife habitat. U.S. For. Serv. Gen. Tech. Rep. RM-87.

GITTINS, R. 1985. Canonical analysis: a review with applications in ecology. Springer-Verlag, Berlin, Germany. 351pp.

HOPE, K. 1969. Methods of multivariate analysis. Gordon and Breach, New York, N.Y. 288pp.

JOHNSON, D. H. 1981. The use and misuse of statistics in wildlife habitat studies. Pages 11–19 *in* D. E. Capen, ed. The use of multivariate statistics in studies of wildlife habitat. U.S. For. Serv. Gen. Tech. Rep. RM-87.

JOYNER, S. P. 1985. SAS/STAT guide for personal computers. Version 6 ed. SAS Inst., Inc., Cary, N.C. 378pp.

NORUSIS, M. J., editor. 1986. SPSS/PC+ advanced statistics. SPSS, Inc., Chicago, Ill. 332pp.

SEBER, G. A. F. 1984. Multivariate observations. John Wiley & Sons, New York, N.Y. 686pp.

TITUS, K., J. A. MOSHER, AND B. K. WILLIAMS. 1984. Chance-corrected classification for use in discriminant analysis: ecological applications. Am. Midl. Nat. 111:1–7.

WILLIAMS, B. K. 1981. Discriminant analysis in wildlife research: theory and applications. Pages 59–71 *in* D. E. Capen, ed. The use of multivariate statistics in studies of wildlife habitat. U.S. For. Serv. Gen. Tech. Rep. RM-87.

Received 30 July 1987.
Accepted 27 March 1988.

CORRIGENDUM

KLETT, A. T., T. L. SCHAFFER, AND D. H. JOHNSON. 1988. Duck nest success in the prairie pothole region. J. Wildl. Manage. 52:431–440.

The second complete sentence in the right column on page 435 should be "Mallards had the greatest success in all periods in idle grassland; the habitat ranked second was planted cover in 1966–74, grassland in 1975–79, and wetland in 1980–84 (Table 4)."