ROBUST CANONICAL CORRELATION OF SAGE GROUSE HABITAT[1]

Mark S. Boyce[2]

Abstract.--The distribution patterns of sage grouse (Centrocercus urophasianus Bonaparte) were studied on Atlantic Richfield Company's Coal Creek coal surface mine site in northeastern Wyoming. Fecal droppings were periodically removed from randomly located belt transects within a 36 km^2 study area. Using multiple regression and canonical correlation the seasonal distribution of droppings was related to several habitat variables which were hypothesized to be important to sage grouse. By canonical correlation, approximately 90% of the variance in a fecal count variate can be attributed to a habitat variate. The correlations between each variate and the original variables were found to be very stable when habitat variables were rotated. Similarly, the robustness of the loadings was confirmed by systematically eliminating cases (transects) from the analysis. Although canonical correlation provides little insight here over that obtained from multiple regression analysis, it does provide a succinct summary of a large and complex set of interrelationships among variables.

Key words: Canonical correlation; habitat; mine reclamation; multiple regression; patchiness; sage grouse; Wyoming.

INTRODUCTION

Applications of canonical correlation in ecology are relatively uncommon (Smith 1980). This may be due in part to unstable results or difficult interpretation (Gauch and Wentworth 1976, Johnston[3]). These problems are often attributable to assumption violations in the data, e.g., variables which are not normally distributed (Harris 1975), multicollinearity among variables (Cohen et al. 1979), or nonlinearity (Gauch and Wentworth 1976). However, if these assumptions can be approximated, the potential is great for application of canonical correlation in habitat studies. Canonical correlation searches for the maximum correlation between linear combinations of two sets of variables; for example, one set of variables may consist of measures of organism distribution and/or abundance, and another set may consist of variables characterizing habitat.

In this paper I describe an application of canonical correlation to an analysis of sage grouse habitat. One set of variables consists of the number of sage grouse fecal droppings deposited in different seasons, whereas the second set of variables consists of selected habitat variables which were hypothesized to be important factors influencing sage grouse distribution. Correlations between the original variables and the derived distribution and habitat variates were found to be robust (contra Cohen et al. 1979), even though the sample size was relatively small

[1]Paper presented at The use of multivariate statistics in studies of wildlife habitat: a workshop, April 23-25, 1980, Burlington, Vt.

[2]Assistant Professor, Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071.

[3]Personal communication with R.F. Johnston, Professor, University of Kansas.

(n = 20). I attribute the robustness of this application of canonical correlation to thorough screening for nonlinearity and normality followed by transformation or elimination of variables before conducting the analysis.

METHODS AND MATERIALS

The study area was a 36 km² coal lease (T46N, R70W) in Campbell County, Wyoming which is scheduled for development as a surface mine. An objective of this research was to characterize the habitat components which are most important to sage grouse and to develop guidelines for reclamation of mined lands for sage grouse.

Sage grouse fecal droppings are easily identified and may last for up to 3 years before deteriorating. A sampling scheme was designed to assess seasonal distribution of sage grouse by periodically checking permanent belt transects for fresh droppings, and removing all droppings deposited since transects were last searched. Twenty 2 m x 1000 m belt transects were randomly located on the study area, and checked at irregular intervals.

An independent consulting team was employed to map and characterize all vegetation types on the study area (Keammerer and Keammerer 1975). In addition, detailed data were collected on the composition of sagebrush density, height and cover along each transect by employing systematic point-quarter procedures (Seber 1973).

The total number of fecal pellets found during a particular season was summed over the three years of our study, yielding three seasonal distribution variables, i.e., number of droppings found in 1) winter, 2) spring, and 3) summer through early fall. Habitat use patterns are known to be very similar in summer and early fall in non-migratory sage grouse populations (Patterson 1952). Habitat and distribution variables employed in this analysis are listed in table 1.

RESULTS

An average of 360 droppings was found on each transect. All variables were screened for skewness and kurtosis, and transformations conducted where appropriate. Seven variables were either eliminated or combined with other variables because their distribution contained an excessive proportion of zeros. Next, because of the sensitivity of canonical correlation to non-linearity (Gauch and Wentworth 1976), bivariate plots were constructed between the distribution variables and each habitat variable. When nonlinear patterns appeared, the respective habitat variable was eliminated from the analysis. For example, the total droppings vs. the area within each belt transect in the big sagebrush habitat type appears nonlinear (fig. 1). The availability of at least some big sagebrush is

Table 1. Reduced set of variables and appropriate transformations for variables used in this analysis.

Distribution variables	
TOTAL	= $\log_{10}$ (total droppings found on each transect over 3 years, plus 1.0).
SUMMER-FALL	= $\log_{10}$ (sum over 3 years of all droppings which accumulated between June and October, plus 1.0).
WINTER	= $\log_{10}$ (droppings accumulated during winter months, plus 1.0).
SPRING	= $\log_{10}$ (sum over 2 years of droppings accumulated during spring months through mid-May, plus 1.0).
DIST	= fecal droppings canonical variate.
Habitat variables	
LEK	= distance between the nearest sage grouse strutting ground (lek) and the closest point on each belt transect in km.
FORBS	= summed area within each belt transect within forb-producing habitats, including riparian areas, in m².
DIV	= habitat diversity = $-\Sigma p_i \log p_i$ where p_i is the proportion of the ith habitat type within each belt transect.
PATCH	= patchiness index equal to the number of times the belt transect crosses from one habitat type to another, plus 1.0.
SPRAY	= $\log_{10}$ (area within belt transect of big sagebrush habitat which had been sprayed with 2,4-D, plus 1.0).
COVER	= average sagebrush cover along transect estimated with 21 point-quarter measurements (Seber 1973).
VACOV	= standard deviation of COVER.
DENS	= average density of sagebrush plants per m² estimated from 21 point-quarter measurements (Seber 1973).
VADEN	= standard deviation of DENS.
SAGE	= area within belt transect of big sagebrush habitat type in m².
HABITAT	= canonical variate comprised of habitat variables.

essential, and transects with no big sagebrush have no sage grouse utilization. However, large homogeneous stands of sagebrush tend to have low forb availability and are not preferred habitat. Therefore, SAGE was eliminated from further analysis.

A multiple regression analysis was conducted using each distribution variable as a dependent variable. This permitted explicit hypothesis testing regarding the importance of various habitat variables. Several interesting patterns emerge; for example, habitat patchiness and the proportion of forb-producing habitats were both positively correlated with the number of droppings found on each belt transect (fig. 2). Different habitat variables are important at different times of the year due to the seasonal changes in food habits and behavior of the grouse. A few of the more interesting multiple regression models are summarized in table 2.

In an attempt to summarize the complex interrelationships between the seasonal distribution patterns and habitat, I conducted a canonical correlation analysis of the "distribution" set of variables and a selected set of habitat variables. Due to the small sample size (n = 20), I began with a small number of variables: three distribution variables, and three of the most important habitat variables. As presented in figure 3A, a habitat variate accounts for almost 90% of the variance in a distribution variate and the relationship is highly significant ($P << 0.001$). Furthermore the correlations between the first pair of canonical variates and the original variables are all intuitively satisfying and parallel the results of multiple

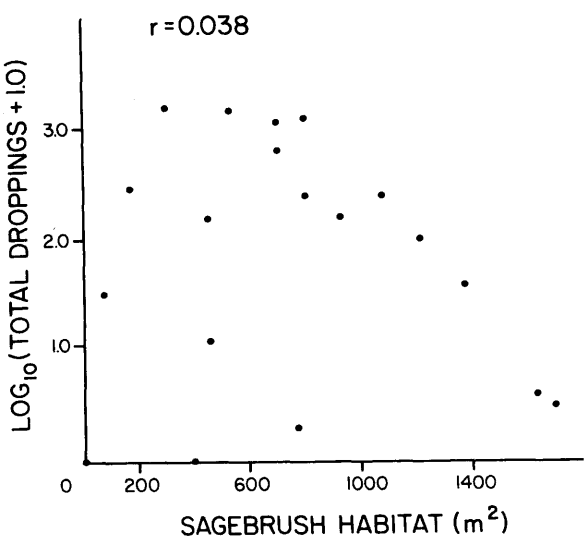

Figure 1. Total droppings over three years plotted as a function of the area within each belt transect in the big sagebrush habitat type.

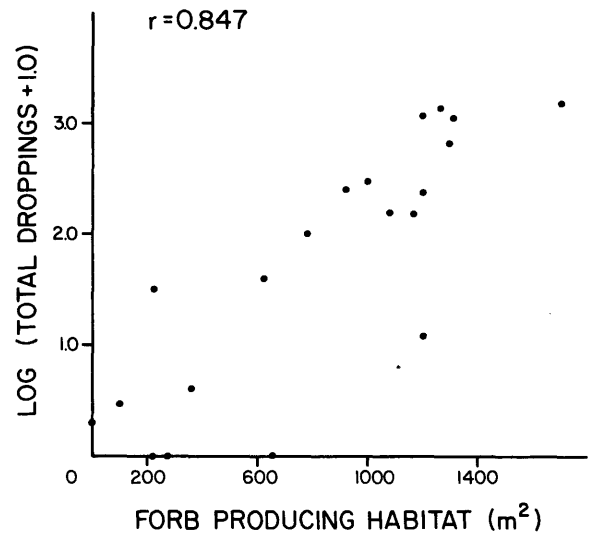

Figure 2. Total droppings over there years as a function of the area within each belt transect in forb-producing habitat types.

regression. The second canonical correlation was not statistically significant ($P < 0.05$), and loadings of the second pair of variates were not readily interpretable.

Since the 3 x 3 analysis worked so well, the number of habitat variables in the model was increased to assess whether a better synopsis of the interrelationships between habitat and distribution could be achieved. The analysis was attempted with five habitat variables (fig. 3B) and then with nine habitat variables (fig. 3C). The high stability of the loadings was impressive, i.e., the correlations between the first pair of canonical variates and the original variables remained relatively constant as additional habitat variables were added to the analysis. Even more exciting, however, the loadings were intuitively reasonable, and provided an exceptionally good representation of the interrelationships between variables which I had discovered through a tedious and detailed multiple regression analysis. All distribution variables loaded positively into a distribution variate, thus clearly the distribution variate reflected intensity of use by sage grouse. Correlations between original variables and the habitat variate all reflected quality of sage grouse habitat, i.e., variables which were positively correlated tend to be preferred habitat components and variables which were inversely correlated with the habitat variate were usually inversely correlated with the number of droppings found on transects at various times of year.

To test the robustness of the canonical correlations and variable loadings, the 3 x 9 analysis was rerun 20 times, each time leaving out one case (transect) from the analysis (similar to

Table 2. A sample of multiple regression models with droppings counts as dependent variables and habitat attributes as independent variables. All models listed account for a significant proportion of the variance in dependent variables ($P < 0.05$).

SPRING = 2.674 - 0.452(LEK) - 0.056(SPRAY)	R = 0.780
SPRING = 1.603 - 0.359(LEK) + 0.002(FORBS)	R = 0.841
SPRING = 0.002(FORBS) + 10.6(COVER) - 0.413	R = 0.818
SPRING = 3.367(DIV) + 13.0(COVER) - 0.686	R = 0.745
SPRING = 0.936 - 0.251(LEK) + 0.002(FORBS) + 5.8(COVER)	R = 0.864
SPRING = 2.706 - 0.481(LEK)	R = 0.779
SPRING = 0.003(FORBS) - 0.163	R = 0.680
SPRING = 0.313 + 14.56(COVER)	R = 0.671
SUMMER-FALL = 0.826 + 0.002(FORBS) - 0.414(SPRAY)	R = 0.896
SUMMER-FALL = 0.467 + 0.116(PATCH) + 0.002(FORBS) - 0.416(SPRAY)	R = 0.912
SUMMER-FALL = 0.664 + 2.3(DIV) - 0.265(LEK) + 0.002(FORBS)	R = 0.879
SUMMER-FALL = 0.812 + 2.28(DIV) - 0.133(LEK) + 0.001(FORBS) - 0.325(SPRAY)	R = 0.923
SUMMER-FALL = 0.994 + 3.54(DIV) - 0.165(LEK) - 0.371(SPRAY)	R = 0.910
SUMMER-FALL = 0.623 + 3.56(DIV) - 0.528(SPRAY)	R = 0.887
SUMMER-FALL = 0.995 - 0.241(LEK) + 0.002(FORBS)	R = 0.861
SUMMER-FALL = 1.88 - 0.624(SPRAY)	R = 0.810
SUMMER-FALL = 0.003(FORBS) - 0.191	R = 0.782
WINTER = 1.747 - 2.5(DIV) + 0.002(FORBS) - 0.423(SPRAY) - 0.238(VADEN)	R = 0.680
WINTER = 1.234 + 0.001(FORBS) - 0.42(SPRAY) - 0.216(VADEN)	R = 0.649
WINTER = 1.935 - 0.555(SPRAY) - 0.207(VADEN)	R = 0.583
WINTER = 1.184 - 0.314(SPRAY)	R = 0.437
TOTAL = 1.42 + 4.798(DIV) - 0.374(LEK)	R = 0.823
TOTAL = 0.718 + 0.003(FORBS) - 0.26(SPRAY)	R = 0.881
TOTAL = 2.415 - 0.598(SPRAY)	R = 0.691

a jackknife procedure). Results for the first pair of canonical variates were impressively stable, with one exception. When transect number 16 was eliminated from the analysis, the loadings for the first pair of canonical variates were not interpretable (table 3). However, upon inspection I discovered that the second pair of canonical variates were loaded precisely in the same pattern as the first pair in all of the other rotations. Thus, the same patterns were present in all trials, but for some reason the second orthogonal pair of variates switched with the first when transect number 16 was deleted from the analysis. The second canonical correlation was statistically significant ($P < 0.05$) in this case whereas it was not in any of the other analyses.

The results of the robustness assessment are summarized in table 4 where the means and standard deviations of the canonical correlations and variable loadings are presented for 20 rotations. For the analysis where transect 16 was eliminated, the second pair of canonical variates was used, but for all other cases the results were averaged over the first pair of canonical variates.

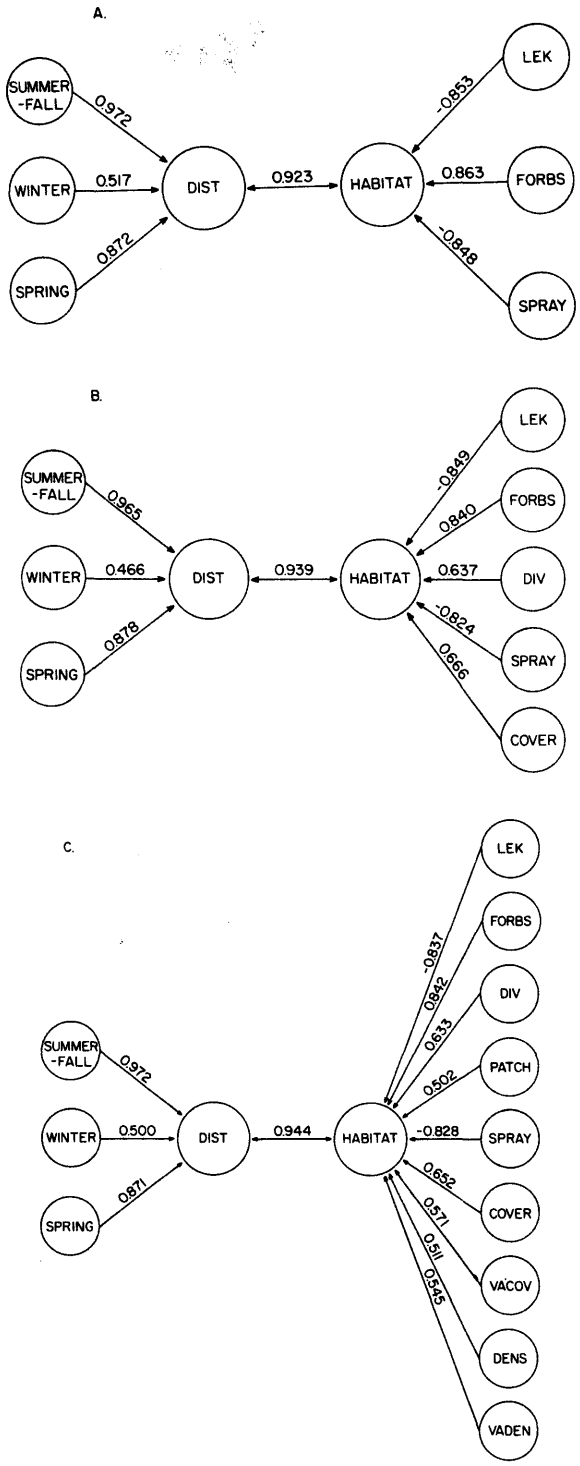

Figure 3. Path diagram of the first pair of canonical variates with three distribution variables and three (A), five (B), and nine (C) habitat variables. The significance levels based upon Bartlett's test are P = 0.00003, 0.00006 and 0.004 for the 3 x 3, 3 x 5, and 3 x 9 analyses respectively. Definitions of variables are listed in table 1.

To further assess the robustness of the 3 x 9 canonical correlation analysis, I systematically eliminated one of nine habitat variables from the analysis. The canonical correlations and variable loadings proved very stable for the first pair of canonical variates. In this experiment as well as during the rotation of cases, loadings into the second pair of canonical variates fluctuated dramatically. As can be seen in table 4, however, variation in canonical correlations and variable loadings for the first pair of canonical variates is impressively small.

The importance of distance to a lek is not clear from these results. It seems plausible that leks are located in good habitats but this may obscure results since grouse social behavior may override habitat quality in determining their distribution. My first attempt to better understand the importance of leks entailed using regression analysis to remove variation attributable to LEK from each of three distribution variables. Then canonical correlation analysis was conducted between the residuals of the distribution variables and the eight original habitat variables with LEK (sensu Boyce 1978). However, the canonical correlation was not statistically significant, and variable loadings were not interpretable. Further multiple regression analysis of the three distribution variables was conducted by studying residuals after removing variation attributable to the area in sprayed sagebrush habitat (SPRAY). The partial correlation between the fecal dropping variables and LEK was discovered to be significant only in the springtime ($r = -0.639$, $P < 0.01$). Since the leks are only actively used during March, April, and May, it seems plausible that only in spring should there be any unique contribution to distribution determined by the proximity to a lek.

DISCUSSION

Sample size guidelines are lacking for applications of many multivariate procedures. Although Bartlett's test for canonical correlation allows one to assess the overall significance of the relationship between two sets of variables, there is presently no means by which significance of individual variable loadings can be evaluated. Nevertheless, the most useful applications of procedures such as canonical correlation may be heuristic ones, thus inference procedures may not be necessary or even desirable. However, assessing the goodness of such descriptive tools does seem important, and robust procedures such as jackknifing provide such an assessment. If canonical correlations and variable loadings prove to be robust, as they were in this study, the extremely large sample sizes recommended by Thorndike (1978) should not be necessary. The impracticality of collecting extremely large data sets in wildife habitat studies should not persuade the investigator that applications of multivariate procedures, such as canonical correlation, are necessarily inappropriate.

Table 3. Correlation between original variables and the first two pairs of derived canonical variates when transect number 16 was eliminated from the analysis. Note the similarity between figure 3C and the loadings for the second pair of canonical variates. The probability levels are from Bartlett's test.

	Distribution variate		Habitat variate	
First Canonical	SUMMER-FALL	-0.542	LEK	0.090
r = 0.973	WINTER	-0.548	FORBS	-0.566
(P = 0.0002)	SPRING	-0.040	DIV	-0.229
			PATCH	-0.052
			SPRAY	0.544
			COVER	0.053
			VACOV	-0.041
			DENS	0.072
			VADEN	-0.082
Second Canonical	SUMMER-FALL	0.827	LEK	-0.858
r = 0.931	WINTER	0.424	FORBS	0.647
(P = 0.048)	SPRING	0.924	DIV	0.589
			PATCH	0.573
			SPRAY	-0.655
			COVER	0.707
			VACOV	0.567
			DENS	0.551
			VADEN	0.425

Cohen et al. (1979) claim that canonical weights (loadings) are necessarily unstable in canonical correlation because each variable accounts for a unique portion of the variability in the other set of variables. However, we have seen in this analysis that unstable loadings or "bouncing betas" need not occur and that removal or addition of variables need not influence the qualitative interpretation. It seems more likely that nonlinearity or multicollinearity are behind many applications where "bouncing betas" appear to be a serious problem (see Gauch and Wentworth 1976).

The results depicted in figure 3 confirm many of the subjective impressions which were developed based upon field obsevations of the major habitat components important to sage grouse. During winter, sage grouse feed exclusively on leaves of big sagebrush (Artemisia tridentata). However, between April and October, forbs such as dandelion (Taraxacum officianale), curly-cup gumweed (Grindelia squarrosa), sweet clover (Melilotus officianalis), false dandelion (Agoseris glauca), alfalfa (Medicago sativa), and salsify (Tragopogon dubius) constitute a major portion of the diet for both young and adult birds (Patterson 1952). Although sagebrush provides important visual cover from aerial predators, sagebrush plants compete with forbs for nutrients and water. Thus optimal habitats are often patchy where forbs and sagebrush cover occur in close proximity.

Each of the variables which we selected to measure habitat patchiness (DIV, PATCH, VACOV, VADEN) is positively correlated with the number of droppings found on the transects. Similarly, these same variables are positively correlated with the "habitat quality" variate. SPRAY is clearly an important variable which is negatively loaded into the "habitat quality" variate. Since spraying with herbicides such as 2,4-D kills both forbs and sagebrush, it is easy to appreciate the devastating impact of herbicide spraying on the use of areas by sage grouse.

It is important to caution against cavelier interpretations of canonical correlation analysis. As noted earlier, explicit hypothesis testing regarding any single variable is not possible with canonical correlation but may be accomplished with linear regression techniques. Some patterns can be masked by simple reliance upon the overall trends summarized by canonical correlation as I demonstrated above for LEK. Also, to achieve a robust canonical model, variables must often be eliminated for statistical reasons even though some of these variables may be biologically important, e.g., SAGE. I strongly recommend that careful attention be given to any variables of postulated biological significance but which must be eliminated on statistical grounds. Various statistical procedures may provide insight into the behavior of these variables, e.g., nonlinear regression, ridge regression, discriminant analysis, or nonparametric techniques.

In summary, my application of canonical correlation does not provide much insight into the

Table 4. Robust estimates of canonical correlations and variable loadings. The top portion of the table lists means and standard deviations (in parentheses) of values where one case (transect) was systematically eliminated from the analysis (n = 20). The bottom portion lists means and standard deviations when one habitat variable was eliminated (n = 9).

	Distribution variate		Habitat variate	
Rotation of Cases				
Canonical	SUMMER-FALL	0.957(0.036)	LEK	-0.827(0.051)
$\bar{r}$ = 0.950(0.011)	WINTER	0.488(0.063)	FORBS	0.826(0.047)
($\bar{P}$ = 0.0077[0.005])	SPRING	0.864(0.062)	DIV	0.625(0.043)
			PATCH	0.498(0.059)
			SPRAY	-0.811(0.048)
			COVER	0.646(0.068)
			VACOV	0.564(0.065)
			DENS	0.539(0.08)
			VADEN	0.502(0.075)
Rotation of Habitat Variables				
Canonical	SUMMER-FALL	0.969(0.024)	LEK	-0.837(0.019)
$\bar{r}$ = 0.940(0.003)	WINTER	0.506(0.065)	FORBS	0.839(0.014)
($\bar{P}$ = 0.007[0.014])	SPRING	0.846(0.069)	DIV	0.630(0.012)
			PATCH	0.497(0.012)
			SPRAY	-0.838(0.015)
			COVER	0.634(0.061)
			VACOV	0.558(0.046)
			DENS	0.527(0.073)
			VADEN	0.495(0.060)

interactions between sage grouse distribution and habitat that I could not glean from multiple regression analysis. This is not surprising since canonical correlation is somewhat of a generalization of multiple regression (Blackith and Reyment 1971). However, canonical correlation does provide a succinct and synoptic way to summarize a bulky and unwieldly data set. This is certainly one of the most important functions of multivariate statistical analysis.

ACKNOWLEDGMENTS

This work was supported by a contract from ARCO Coal Co. under the helpful supervision of James Tate, Jr. Several people have provided useful assistance in the field including B. Colenso, D. Rothenmaier, B. Holz, and J. Tate, Jr. I thank L. McDonald, K.G. Smith, and J.R. Karr for useful discussions about canonical correlation; and D.E. Capen and J. Rice for suggesting revision in the manuscript.

LITERATURE CITED

Blackith, R.E., and R.A. Reyment. 1971. Multivariate morphometrics. 412 p. Academic Press, New York, N.Y.

Boyce, M.S. 1978. Climatic variability and body size variation in the muskrats (Ondatra zibethicus) of North America. Oecologia 36:1-19.

Cohen, P., E. Gaughran, and J. Cohen. 1979. Age patterns of childbearing: a canonical analysis. Multivariate Behavioral Research 14:75-89.

Cooley, W.W., and P.R. Lohnes. 1971. Multivariate data analysis. 364 p. John Wiley and Sons, New York, N.Y.

Gauch, H.G., Jr., and T.R. Wentworth. 1976. Canonical correlation analysis as an ordination technique. Vegetatio 33:17-22.

Harris, R.J. 1975. A primer of multivariate statistics. 332 p. Academic Press, New York, N.Y.

Keammerer, W.R., and D.H. Keammerer. 1975. Floristic studies and vegetation mapping of the Coal Creek lease area. 55 p. Stoecker-Keammerer and Assoc., Boulder, Colo.

Patterson, R.L. 1952. The sage grouse in Wyoming. 341 p. Sage Books, Denver, Colo.

Seber, G.A.F. 1973. The estimation of animal abundance and related parameters. 506 p. Griffin, London.

Smith, K.G. 1981. Canonical correlation analysis and its use in wildlife habitat studies. In Capen, D.E., editor. The use of multivariate statistics in studies of wildlife habitat: Proceedings of a workshop [Burlington, Vt., April 23-25, 1980]. USDA Forest Service General Technical Report. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colo. (In press).

Thorndike, R.M. 1978. Correlational procedures for research. 340 p. Halsted Press, New York, N.Y.