Ecology, 66(6), 1985, pp 1693–1698

A COMPARISON OF PRINCIPAL COMPONENTS FROM REAL AND RANDOM DATA[1]

DEAN F. STAUFFER
Department of Fisheries and Wildlife Sciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061 USA

EDWARD O. GARTON
Department of Fisheries and Wildlife, University of Idaho, Moscow, Idaho 83843 USA

AND

R. KIRK STEINHORST
Department of Mathematics and Applied Statistics, University of Idaho, Moscow, Idaho 83843 USA

Abstract. We compared principal components derived from sets of real data with dimensions of 120 × 7, 120 × 4, 150 × 11, 150 × 8, 150 × 5, 454 × 12, 454 × 8, and 454 × 5, to those from sets of randomly generated data of corresponding size. Principal components from subsets of 25, 50, 75, and 100 observations from the 120- and 150-observation data sets and those from subsets of 25, 50, 75, 100, 150, 200, 300, and 400 observations from the 454-observation data sets were compared. Percent variance associated with components from real data was relatively constant over all sample sizes; percent variance decreased with larger samples of random data. A bootstrap method was used to develop standard error estimates on percent variance and percent of remaining variance associated with components from real data. Percent of remaining variance associated with the first four components from real data was significantly higher than analogous components from random data.

Key words: bootstrap; confidence intervals; principal components analysis; random data; significance tests; standard error estimates.

INTRODUCTION

Multivariate statistical techniques have been used more commonly in recent years with the advent of easily applied computer packages. Indeed, a symposium has been devoted to the use of multivariate techniques in wildlife studies (Capen 1981). Principal components analysis (PCA) is one of the multivariate methods commonly used to investigate relationships in ecological studies (e.g., Fuju 1969, James 1971, Conner and Adkisson 1977, Smith 1977, Whitmore 1977, McCrimmon 1978, Rotenberry and Wiens 1980, Collins et al. 1982, Nudds 1983).

PCA usually is used to reduce a data set with a relatively large number (p) of correlated variables to a data set with fewer ($m < p$), uncorrelated variables (components) that retain most of the information content of the original data. These components are linear combinations of the original variables, usually derived by eigen analysis of a correlation matrix of the variables. The correlation matrix is used throughout this paper. The first component derived represents a certain portion of the generalized variance present in the original data set; successive components account for decreasing proportions of the variance while remaining uncorrelated with previous components (Rummel 1970). The criteria for deciding the "best" number of components to consider in an analysis are varied (Rummel 1970, Krzanowski 1983), but usually components associated with $<(1/p) \times 100\%$ of the total variance are not considered because they represent less information than expressed in a single variable.

Despite the attractiveness of PCA for reducing multidimensional data, some concerns regarding its use have been raised. Karr and Martin (1981) presented information indicating that the percent variance attributed to principal components derived from real data may not be substantially greater than percent variance for principal components from random data sets of the same dimensions as the real data. In particular, percent variance of the second and third components derived from random data often was higher than that for the corresponding real data. They also found that the percent variance was related negatively to sample size; as data set size increased, percent variance of each component decreased.

Our objectives for this paper are to investigate further the relationship between principal components derived from real and random data sets. Specifically, we evaluate PCA results in relationship to sample size; we apply the bootstrap method (Efron 1979) to develop estimates of standard error and confidence limits for variance associated with principal components; and we demonstrate an approach to testing the significance of

[1] Manuscript received 3 January 1984; revised 3 December 1984; accepted 16 January 1985.

TABLE 1. Percent variance (% var.) and cumulative percent variance (Cum. %) accounted for by principal components extracted from real and random data sets of various dimensions.

Data set dimensions and type	Component							
	1		2		3		4	
	% var.	Cum. %	% var.	Cum. %	% var.	Cum. %	% var.	Cum. %
120 × 7								
Real	62.9	62.9	14.6	77.5	9.8	87.3	7.9	95.2
Random	18.2	18.2	17.5	35.7	15.6	51.3	13.6	64.9
120 × 4								
Real	71.7	71.7	21.5	93.2	6.4	99.6	0.4	100.0
Random	29.0	29.0	26.7	55.7	23.6	79.3	20.7	100.0
150 × 11								
Real	34.9	34.9	25.8	60.7	15.1	75.8	8.3	84.1
Random	13.6	13.6	12.0	25.6	10.9	36.5	10.5	47.0
150 × 8								
Real	40.9	40.9	25.4	66.3	13.2	79.5	8.9	88.4
Random	16.8	16.8	15.8	32.6	14.2	46.8	12.4	59.2
150 × 5								
Real	41.5	41.5	26.0	67.5	15.0	82.5	10.6	93.1
Random	23.4	23.4	22.5	45.9	21.2	67.1	17.5	86.4
454 × 12								
Real	38.2	38.2	19.9	58.1	8.7	66.8	7.8	74.6
Random	10.3	10.3	10.0	20.3	9.8	30.1	9.3	39.4
454 × 8								
Real	50.3	50.3	15.4	65.7	12.5	78.2	7.6	85.8
Random	15.2	15.2	14.4	29.6	13.6	43.2	13.4	56.6
454 × 5								
Real	54.6	54.6	24.6	79.2	11.9	91.1	6.8	97.9
Random	23.7	23.7	21.7	45.4	19.7	65.1	18.1	83.2

principal components from real data in relation to random data.

METHODS

We used real data from three data sets. The first consisted of 120 observations of seven habitat variables (percent of surrounding area in deciduous cover or open, percent tree and ground cover, canopy height, and number of trees [>7 cm dbh] and small stems [<7 cm dbh]) recorded at 0.01-ha circular plots randomly located in mixed-shrub habitats in southeastern Idaho (Stauffer 1983). We also analyzed a subset of this data set containing four variables (percent open, deciduous, and ground cover, and canopy height). The second data set constituted 150 observations of 11 variables (percent of surrounding area in coniferous or deciduous cover or open, percent tree and ground cover, canopy height, and number of small stems, all trees, coniferous trees, small aspen [*Populus tremuloides*] [7–23 cm dbh], and large aspen [>23 cm dbh]) recorded at 150 circular plots located randomly in aspen stands in southeastern Idaho (Stauffer 1983). Subsets of these data with eight variables (percent of area in deciduous cover or open, tree canopy cover, canopy height, and number of small stems, all trees, deciduous trees, and coniferous trees) and five variables (tree canopy cover, canopy height, and number of small stems, deciduous trees, and coniferous trees) were also analyzed. The third data set was composed of 454 observations with 12 variables from sampling points along transects in southeastern Idaho (Stauffer and Peterson 1985). Variables were tree, shrub, and sapling density, mean tree dbh, mean shrub height and crown diameter, and the coefficient of variation (CV) of these six variables. We also analyzed subsets of eight (tree, sapling, and shrub density, mean tree dbh and shrub height, and CV of tree, sapling and shrub density) and five (tree, sapling, and shrub density, and mean tree dbh and shrub height) variables. We thus analyzed eight sets of real data with dimensions of 454 × 12, 454 × 8, 454 × 5, 150 × 11, 150 × 8, 150 × 5, 120 × 7, and 120 × 4.

Prior to analysis, percentage data were subjected to an arcsine–square root transformation, and count data to a square root transformation (Zar 1974).

We generated eight matrices of random data with dimensions corresponding to those of the real data for comparison. PCA was conducted on the correlation matrices of the complete real and random data sets. In addition, five random subsets of real and random data at sample sizes of 25, 50, 75, and 100 for the data sets with 120 and 150 observations and at sample sizes of 25, 50, 75, 100, 150, 200, 300, and 400 for the data sets with 454 observations were selected. PCA was

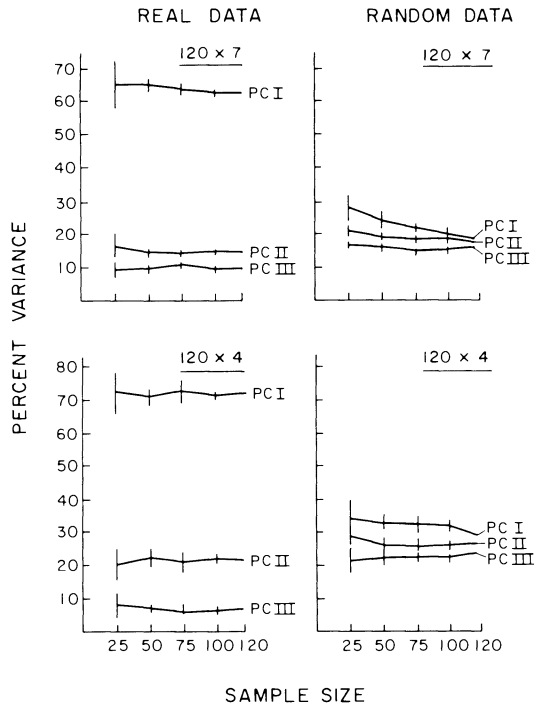

FIG. 1. Percent variance associated with the first three principal components from real and random data sets (N = 120, p = 7, 4) in relation to sample size. Vertical bars represent a 95% confidence interval on the mean value based upon five analyses at each sample size.

conducted on correlation matrices for each data subset. The mean and 95% confidence interval associated with each component at each sample size was calculated.

The standard error (SE) for each component's (from the 150 × 11 data sets) variance was calculated by means of a bootstrapping procedure (Efron 1979). Bootstrapping is a relatively simple, computer-intensive technique that can be used to assign accuracy to some quantity of interest. The procedure followed to derive variance estimates for variance associated with principal components was (from Efron 1979):

1) Let F be the empirical distribution of the n = 150 observations in the original data set.
2) Use a random number generator to draw n new observations independently and with replacement from F, so that each new observation is an independent random selection of one of the original n observations. Each of the original observations could appear 0 to n times in the bootstrap sample.
3) Compute the item of interest for the bootstrap sample, in this case, percent variance associated with the m components.
4) Repeat steps (2) and (3) a large number of times (N), each time using an independent set of new random numbers to generate the new bootstrap sample. This process yields N values of the m percent variances. (V_m^{*i} = i^{th} bootstrap estimate of variance for the m^{th} component [V_m]. i = 1, . . . , N; m = 1, . . . , number of components.)
5) Order the values of percent variance for each component from low to high (e.g., for component 1:

$$V_1^{*1}, V_1^{*2}, \ldots, V_1^{*i}, \ldots, V_1^{*N}).$$

6) Let [a*,b*] be the central 68% interval for the V_m^{*i} values, i.e.,

$$\{\text{number of } V_m^{*i} < a^*\}/N = 0.16,$$

$$\{\text{number of } V_m^{*i} < b^*\}/N = 0.84.$$

The bootstrap estimate for the standard error, say $\hat{\sigma}_m$ for the variance associated with component m, is:

$$\hat{\sigma}_m = (b^* - a^*)/2.$$

Once an estimate for $\hat{\sigma}_m$ is derived, a confidence interval on any component's variance can be found as:

$$V_m \pm Z_{\alpha/2}(\hat{\sigma}_m).$$

RESULTS AND DISCUSSION

Components of real and random data

Percent variance associated with the first two principal components was substantially higher for real data than random data in most cases (Table 1). For the data set with 120 observations, only the first component

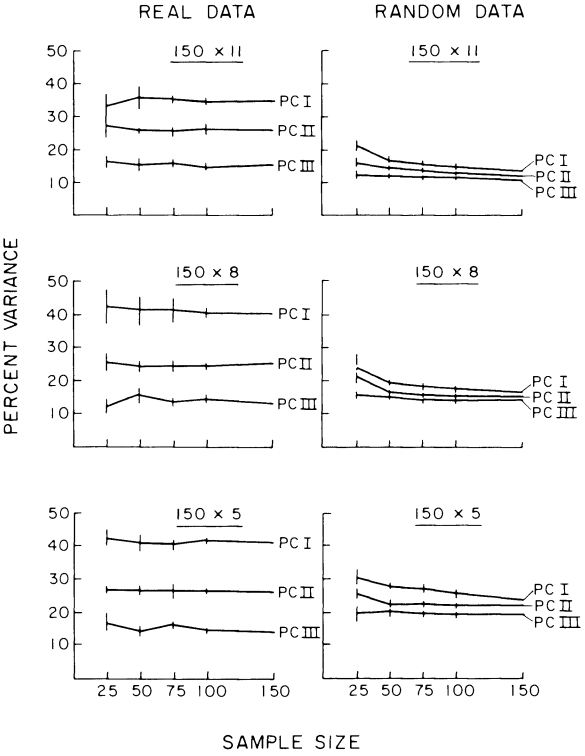

FIG. 2. As in Fig. 1, but with N = 150, p = 11, 8, 5.

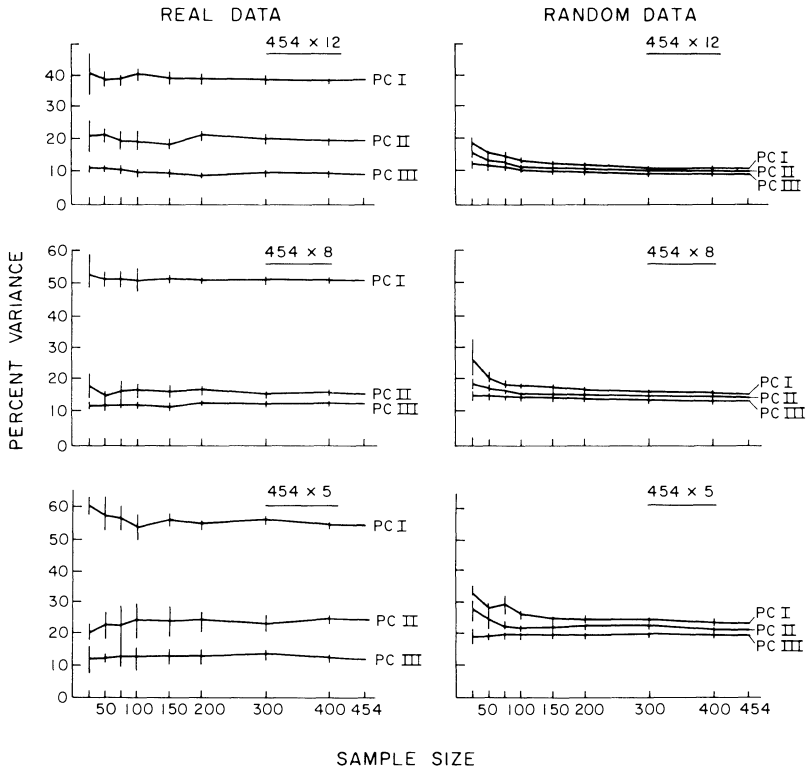

FIG. 3. As in Fig. 1, but with $N = 454$, $p = 12, 8, 5$.

had a percent variance greater than that for random data. Percent variance of the third component was apparently comparable between real and random data for the data sets with 150 and 454 observations. In all cases, percent variance of the fourth component was less for real data. Since the theoretical eigenvalues of the correlation matrix from the random data are all the same, one expects the observed percent variance to be relatively uniform as seen (Table 1). Because the percent variances decline more markedly from the first to the fourth component for real data than for the random data, a real structure among the variables is implied.

The percent variance associated with the first three components from real data did not vary substantially over the sample sizes evaluated (Figs. 1, 2, 3). In contrast, as the sample size increased, variance associated with the first three components for random data declined (Fig. 2). This decline was most pronounced for the first two components and was least evident in the smallest (120 × 4) data set. That percent variance for components from real data changed little with sample size variation indicates a structure within the real data set that can be represented by subsets of the total data. For the random data, however, as the sample size was increased, the percent variance decreased toward the theoretically expected values of $(1/p) \times 100$. In this study, the percent variances for random data with a sample size of 25 were misleadingly high.

Variance associated with the lesser components was less for real data than for random data (Table 1). This trend might be taken to mean that these lesser components contribute little, if any, information and should not be considered in analysis results. However, the percent variance associated with these lesser components is derived from the variance remaining after some portion has been captured by previous components. For example, although the second component from real data (120 × 7) accounted for 14.6% of the total variance (Table 1), 62.9% of the total variance had already been accounted for by the first component; the second component actually represented 39.4% of the variance remaining after the first component. Percent of remaining variance for the second and third components is substantially higher for real than random data (Table 2) (see also Karr and Martin 1981).

That percent of remaining variance was higher for real than random data can be interpreted to mean that the minor components contain more information than random data. For many studies, once major, obvious patterns are removed by the first few components, subtle, less obvious but ecologically meaningful patterns that are important to the organisms under study may be found (Johnson 1981). Based upon percent of remaining variance, these minor components should be considered as potentially more important than components of random data. However, Gauch (1982) has suggested that ordination axes associated with inter-

TABLE 2. Percent of remaining variance accounted for by principal components extracted from real (V_{Rt}^{A}) and random (V_{Rt}^{B}) data sets of various dimensions.

Data set dimensions and type	Component			
	1	2	3	4
	Percent remaining variance			
120 × 7				
Real	62.9	39.4	43.3	62.3
Random	18.2	21.4	24.3	28.0
120 × 4				
Real	71.7	76.0	94.8	100.0
Random	29.0	37.6	53.3	100.0
150 × 11				
Real	34.9	39.6	38.4	34.3
Random	13.6	13.9	14.7	16.5
150 × 8				
Real	40.9	43.1	39.3	43.7
Random	16.8	19.0	21.2	23.2
150 × 5				
Real	41.5	44.4	46.0	60.5
Random	23.4	29.4	39.2	53.2
454 × 12				
Real	38.2	32.3	20.9	23.6
Random	10.3	11.1	12.3	13.2
454 × 8				
Real	50.3	31.0	36.5	34.9
Random	15.2	17.0	19.2	23.5
454 × 5				
Real	54.6	54.2	57.5	77.2
Random	23.7	28.5	36.1	51.5

mediate eigenvalues may represent spurious polynomial relationships. Scatter plots of data points from lower order axes against higher axes can be inspected for nonlinear relationships, which, if found, should be interpreted with caution. Polynomial relationships most likely will occur in the higher axes when the primary gradient for a data set is ≈2–3 times longer than a secondary gradient (Gauch 1982). Careful screening of variables prior to analysis should minimize potential problems with nonlinear relationships.

Variance estimates and hypothesis testing

Given that there is a difference between components from the real data with a nontrivial correlation structure, and random data, it is desirable to evaluate the significance of the observed differences. The first step towards evaluating the significance is to derive an estimate of variance on the parameter of interest, in this case, percent variance (or percent remaining variance) associated with each component.

To illustrate this procedure we calculated estimates of the standard error (SE) for percent variance and percent of remaining variance, and the associated 95% confidence intervals, from 600 bootstrap replications on the 150 × 11 real data set (Table 3). Since an estimate of the SE of a component can be calculated, hypotheses concerning the components can be tested.

We tested hypotheses concerning the principal components of real data based on the percent of remaining variance. The percent of remaining variance associated with a particular component from real data was considered a point estimate and the SE (σ_{Rm}) of that estimate was estimated by bootstrapping (Table 3). A Z test was then conducted for each of the first four components to test the null hypothesis that the percent of remaining variance from real data was not greater than the percent remaining variance from random data of a corresponding size data set. The percent of remaining variance values for the random data components were the mean of analyses of three hundred 150 × 11 data sets of random data.

In all cases, the percent of residual variance associated with each of the first four principal components from real data was significantly higher than that associated with the respective component from random data (Table 4). These results indicate that for the set of real data used here, there exist significant differences between principal components derived from real and random data. The bootstrap method used to derive the SE estimates is easily programmed on a computer, and its application to data such as these is straightforward. Morrison (1976:294) also presents a means to calculate confidence intervals on the characteristic roots of correlation matrices.

Hence, any researcher using principal components who may desire to evaluate his/her data in relation to random data of the same dimensions can do so easily. We suggest that for comparison the mean percent variance (or percent of remaining variance) be calculated from several sets of random data, because if only one set of random data is used, a significant result may be more likely to occur by chance alone. Because of the relatively low variability found in the principal components from random data sets (coefficient of variation ranged from 3.2 to 5.9% for the means of the first four components from $n = 300$ sets of 150 × 11 random data), a sample from 20 or more random data sets of the same dimensions should provide an adequate estimate of the percent variance associated with components from random data.

TABLE 3. Bootstrap estimates of SE ($\hat{\sigma}_m$) and 95% confidence intervals for percent variance and percent remaining variance accounted for by the first four principal components.*

Component	Percent variance		Percent remaining variance	
	$\hat{\sigma}_m$	95% CI	$\hat{\sigma}_{Rm}$	95% CI
1	2.06	30.9–38.9	2.06	30.9–38.9
2	1.12	23.6–28.0	1.60	36.4–42.8
3	1.13	12.9–17.3	2.02	34.4–42.4
4	0.66	7.0–9.6	2.70	29.0–39.6

* Derived from a 150 × 11 matrix of real data based on $N = 600$ bootstrap trials.

TABLE 4. Results of the hypothesis test that the remaining percent of variance accounted for by a principal component from real data ($V_{Rm}{}^{A}$) is not more than that attributable to a corresponding component from random data ($V_{Rm}{}^{B}$).*

Component	$V_{Rm}{}^{A}$	$\hat{\sigma}_{Rm}$†	$V_{Rm}{}^{B}$‡	Z§	P
1	34.9	2.06	13.3	10.48	<.0001
2	39.6	1.60	13.9	16.06	<.0001
3	38.4	2.02	14.8	11.68	<.0001
4	34.3	2.70	16.2	6.70	<.0001

* For data sets of dimensions 150 × 11. $H_0: V_{Rm}{}^{A} \geq V_{Rm}{}^{B}$.
† $\hat{\sigma}_{Ri}$ is based upon a bootstrap sample of 600 observations derived from the set of real data.
‡ Represents the mean percent remaining variance based upon PCA analyses of 300 random sets of data of dimensions 150 × 11.
§ $Z = (V_{Rm}{}^{A} - V_{Rm}{}^{B})/\hat{\sigma}_{Rm}$.

That the percent variance associated with components of random data was highest at small sample sizes (Fig. 1, see also Karr and Martin 1981) should caution workers against conducting principal components analyses on small data sets. At times, the correlation matrix for PCA is derived from mean vectors of the variables (e.g., James 1971). In using this approach, information about variation within each group being considered is lost, and the dimensions of the original data matrix are considerably reduced. Wherever possible, all observations in the data matrix should be used to preserve information on variation of subgroups of interest within the data.

ACKNOWLEDGMENTS

The use of the bootstrap method for estimating variance was first suggested by C. Potratz. Funds for computer analysis were provided by User Services at the University of Idaho and Virginia Polytechnic Institute and State University. Reviews of an earlier draft by Louis B. Best, Robert H. Giles, Jr., and two reviewers are appreciated.

LITERATURE CITED

Capen, D. E., editor. 1981. The use of multivariate statistics in studies of wildlife habitat. United States Forest Service General Technical Report **RM-87**.

Collins, S. L., F. C. James, and P. G. Risser. 1982. Habitat relationships of wood warblers (Parulidae) in northern central Minnesota. Oikos **39**:50–58.

Conner, R. N., and C. S. Adkisson. 1977. Principal component analysis of woodpecker nesting habitat. Wilson Bulletin **89**:122–129.

Efron, B. 1979. Computers and the theory of statistics: thinking the unthinkable. Society for Industrial and Applied Mathematics (SIAM) Review **21**:460–480.

Fuju, K. 1969. Numerical taxonomy of ecological characteristics and the niche concept. Systematic Zoology **18**:151–153.

Gauch, H. G., Jr. 1982. Noise reduction by eigenvector ordinations. Ecology **63**:1643–1649.

James, F. C. 1971. Ordinations of habitat relationships among breeding birds. Wilson Bulletin **83**:215–236.

Johnson, D. H. 1981. The use and misuse of statistics in wildlife habitat studies. Pages 11–19 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. United States Forest Service General Technical Report **RM-87**.

Karr, J. R., and T. E. Martin. 1981. Random numbers and principal components: further searches for the unicorn. Pages 20–24 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. United States Forest Service General Technical Report **RM-87**.

Krzanowski, W. J. 1983. Cross-validatory choice in principal component analysis; some sampling results. Journal of Statistical Computation and Simulation **18**:299–314.

McCrimmon, D. A., Jr. 1978. Nest site characteristics among five species of herons on the North Carolina coast. Auk **95**: 267–280.

Morrison, D. F. 1976. Multivariate statistical methods. McGraw-Hill, New York, New York, USA.

Nudds, T. D. 1983. Niche dynamics and organization of waterfowl guilds in variable environments. Ecology **64**:319–330.

Rotenberry, J. T., and J. A. Wiens. 1980. Habitat structure, patchiness, and avian communities in North American steppe vegetation: a multivariate analysis. Ecology **61**:1228–1250.

Rummel, R. J. 1970. Applied factor analysis. Northwestern University Press, Evanston, Illinois, USA.

Smith, K. G. 1977. Distribution of summer birds along a forest moisture gradient in an Ozark watershed. Ecology **58**:810–819.

Stauffer, D. F. 1983. Seasonal habitat relationships of ruffed and blue grouse in southeastern Idaho. Dissertation. University of Idaho, Moscow, Idaho, USA.

Stauffer, D. F., and S. R. Peterson. 1985. Seasonal habitat use by ruffed and blue grouse in southeastern Idaho. Journal of Wildlife Management **49**:459–466.

Whitmore, R. C. 1977. Habitat partitioning in a community of passerine birds. Wilson Bulletin **89**:253–265.

Zar, J. H. 1974. Biostatistical analysis. Prentice-Hall, Englewood Cliffs, New Jersey, USA.