

APPLICATION OF SIZE-FREE CANONICAL DISCRIMINANT ANALYSIS TO STUDIES OF GEOGRAPHIC DIFFERENTIATION

Sérgio F. dos Reis¹, Leila M. Pessôa² and Richard E. Strauss³

ABSTRACT

Canonical discriminant analysis (CDA) is a multivariate procedure employed in the study of geographic variation, species differentiation, and macroevolution. However, the application of CDA to study organisms where character size-frequency distribution varies within samples due to sampling bias, may result in artifactual discrimination due to shifts in mean character values. In such cases it would be desirable to discriminate among samples that have been corrected for within-group size differences. In this note we illustrate the application of size-free canonical discriminant analysis. In this procedure the effect of size variation within groups is removed by regressing each character separately on the first pooled within-group principal component, which is an estimate of general size. The application of this procedure is illustrated by a study of geographic differentiation in the echimyid rodent *Proechimys dimidiatus*. A command file of SAS-PC procedures necessary for the implementation of size-free CDA is also provided.

INTRODUCTION

Canonical discriminant analysis (CDA) is a multivariate statistical procedure employed in evolutionary biology and systematics (Campbell and Atchley, 1981; Neff and Marcus, 1980). CDA is used in studies of geographic differentiation and in the analysis of species differentiation and macroevolution (Thorpe, 1983; Lessa and Patton, 1989; Patton and Smith, 1989). The usefulness of CDA resides in the mathe-

¹ Departamento de Parasitologia, Instituto de Biologia, Universidade Estadual de Campinas, Caixa Postal 6109, 13081 Campinas, SP, Brasil. Send correspondence to S.F.R.

² Departamento de Zoologia, Centro de Ciências da Saúde, Universidade Federal do Rio de Janeiro, 21941 Rio de Janeiro, RJ, Brasil.

³ Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA.

mathematical deduction of the method itself which permits comparison of the degree of variability existing among populations to that present within populations (Morrison, 1976; Chatfield and Collins, 1980).

In general, populations or species under study are considered to be groups defined *a priori* and linear measurements of morphological characters are obtained from the specimens within each sample. However, when applying this procedure it is important that within-sample sources of variation be controlled in such a manner that the possible variation among samples will not be masked or result from sampling errors. Thus, factors such as sex dimorphism in size, different developmental stages and indeterminate growth should be considered before different samples are submitted to a discriminant study for the analysis of geographic differentiation (Thorpe, 1983, 1987).

The application of discriminant methods to organisms in which variation in size cannot be easily controlled owing to any of the causes mentioned above may lead to spurious results, since discrimination among populations may represent a mere sampling artifact concerning the groups under study. For example, in organisms such as fish and mammals which show indeterminate growth, the samples utilized often incorporate a sampling error reflecting the presence of different age classes in the group under study. Thus, size frequency distribution of individuals in the different populations will be a function of the ontogenetic development of individuals present in the different samples.

A way of correcting the problem of the existence of variation in size due to ontogeny or to other causes would be to remove statistically the effect of size present within the samples of each group and then apply CDA to the samples already corrected for within-group differences in size.

In this communication we present the application of size-free canonical discriminant analysis to a study of geographic differentiation of quantitative cranial characters in the rodent *Proechimys dimidiatus* (Guenther) (Echimyidae). This taxon represents good material for demonstrating the usefulness of the above procedure since it is considered to be a monomorphic species (Moojen, 1948); in addition, echimyid rodents present wide individual variation in size due to the elevated age variation component existing among adult individuals (Patton and Rogers, 1983). We also provide a command file of SAS-PC procedures that permit rapid and simple implementation of size-free CDA.

MATERIAL AND METHODS

A total of 215 *P. dimidiatus* specimens were examined. The specimens were divided into age classes as defined by Patton and Rogers (1983) for *P. brevicauda* on the basis of tooth eruption and wear. These age classes were used because the pattern of tooth eruption and wear of *P. dimidiatus* is very similar to that of *P. brevicauda*.

Of the ten age classes considered by Patton and Rogers (1983), only classes 8 to 10 represent adults, and only adults of these age classes were used in the present study.

Thirteen quantitative cranial characters defined by Patton and Rogers (1983) were measured in the present study in class 8 to 10 adults from three populations of *P. dimidiatus* obtained at the following localities in the State of Rio de Janeiro: Teresópolis (N = 18), Tijuca (N = 24) and Ilha Grande (N = 10). Measurements were made using a digital MAXCAL caliper with 0.01 mm precision. The specimens analyzed were deposited in the Museu Nacional, Rio de Janeiro. The statistical analyses presented in this study were performed using various procedures (PROC's) of the statistical program SAS-PC (SAS Institute Inc., 1988).

The localities of *P. dimidiatus* collection were considered as a random variable (Winer, 1971) and the cranial variation of the specimens was partitioned into intra-locality (residual) and inter-locality (geographic) variance components (Straney, 1978). Variance components may be estimated by different methods including the least squares procedure (Sokal and Rohlf, 1981); in the present study, however, we used maximum likelihood estimators because of their statistical properties (Searle, 1971; Lindgren, 1976; Van Vleck and Searle, 1979). Variance components were calculated by the VARCOMP procedure (SAS/STAT:967).

Size-free CDA basically consists of removing the effect of within-group size variation by regressing each character separately on the first pooled within-group principal component (a multivariate size estimate) and then applying canonical discriminant analysis to the residues obtained from the regressions (Strauss, 1985).

The first step in the analysis consists of centering the values for each character by the group mean (sample) using the PROC STANDARD (SAS/STAT:751). This first eigenvector can then be used as a multivariate size estimate if 1) all coefficients of the first eigenvector are positive, and 2) if these coefficients are positively and significantly correlated with the values of cranial characters (Strauss, 1985).

The effect of size on each character is removed by regressing the values of each cranial character on the first principal component (PC-1), thus obtaining the residuals that express variation after the removal of the within-group size effect. This step in the analysis is carried out using a combination of PROC's REG (SAS/STAT:773) and SCORE (SAS/STAT:85) instead of PROC GLM (SAS/STAT:549), which is commonly used in linear regression computation, because the above procedures, contrary to GLM, do not standardize the regression residuals by the mean. The use of PROC GLM would prevent the demonstration of inter-group differences since each group would have a mean equal to zero for each character. Canonical discriminant analysis is then performed with the residuals obtained by regressing each character on PC-1 using PROC CANDISC (SAS/STAT:173). Size-free CDA computation is then normally continued by multiplying the inter-group covariance matrix by the inverse of the within-group covariance matrix and extracting the eigenvalues and eigenvectors from the resulting matrix (Morrison, 1976).

These eigenvectors, which express the axes of greatest variation, are then linearly combined with the values of the cranial characters to compose the canonical variables that produce the individual scores. The individual scores for each population are plotted in the space of the canonical variables that determine the patterns of discrimination among populations. A command file for this procedure is provided in the Appendix.

RESULTS AND DISCUSSION

Analysis of variance components is routinely employed in quantitative genetics (Bulmer, 1980; Falconer, 1989) although its use for partitioning morphological variability components in natural populations has been limited to only a few investigations (Straney, 1978; Baker, 1980; Chesser, 1983; Patton and Rogers, 1983; Schmidly *et al.*, 1988). The method of variance components analysis was used in the present study for partitioning the cranial variability of *P. dimidiatus* into intra-locality and inter-locality components.

Estimates of variance components of the cranial characters of *P. dimidiatus* by the maximum likelihood method showed that most of the cranial variability can be attributed to the effect of locality (geographic variation) (Table I). Thus, characters such as palatal length, rostral width, diastema length, rostral height, cranial length and mandible length present a variance component of about 90% among localities. In contrast, zygomatic breadth (70.70%) and rostral length (91.19) showed a high degree of intrapopulation (residual) variation. The remaining cranial characters presented an inter-locality variation component ranging from 70.00% to 80.00%.

Table I also shows that, on average, 26.60% of cranial variability in *P. dimidiatus* populations is associated with intra-locality (residual) variation, indicating that, even though only adult animals were used in the analysis, variability in cranial character size still exists within populations. This residual variance component may be attributed to random, environmental or ontogenetic factors (Staney, 1978; Chesser, 1983; Schmidly *et al.*, 1988) or, in the specific case of *P. dimidiatus*, to an increase in the cranial dimensions of adult individuals, i.e., indeterminate growth. Any of the above causes may confuse the study of discrimination among populations since these factors generate a variability component within populations. Thus, it is necessary to apply a procedure that will remove the effect of character size variation within populations to permit the study of differentiation among populations.

The method employed in the present study was size-free CDA, whose first step is extraction of the first principal component of within-group character covariance matrix. The first principal component accounts for 49.55% of the total variation existing in the covariance matrix among characters within the three *P. dimidiatus* populations. This component may be interpreted as a general size variable since all characters present positive and statistically significant coefficients (Strauss,

1985) (Table II). This first eigenvector can then be utilized in size-free CDA as a generalized size estimate. It is important to note that the size estimate should be a non-measurable variable (Bookstein, 1982; Bookstein *et al.*, 1985), a linear combination like the first principal component and not a separate cranial character, since, by definition, size is a multidimensional quantity (Wright, 1954; Humphries *et al.*, 1981; Bookstein, 1982; Rohlf and Bookstein, 1987).

Table I - Estimates of intra-locality and inter-locality variance components for 13 quantitative cranial characters in *P. dimidiatus*.

Character	Variance component	
	Intra-location	Inter-location
Palatal length	4.03	95.97
Zygomatic breadth	70.70	20.30
Nasal length	28.32	71.68
Interorbital constriction	28.03	71.07
Rostral width	10.52	89.48
Diastema length	2.23	97.77
Rostral weight	4.34	95.66
Cranial length	5.40	94.60
Basilar length	24.30	75.70
Rostral length	91.19	8.81
Maxillary width	74.57	25.43
Post-palatal length	19.51	80.49
Mandibular length	8.45	91.95
Mean variance component	28.60	71.40

We then regressed the values for each cranial character on the first principal component scores. Table III shows that all cranial characters presented significant regressions on the first principal component scores, indicating the presence of significant variation in size within each group even though only adult individuals were employed in the analysis. Size-free CDA was then applied to the three *P. dimidiatus* populations on the basis of the residuals of each character. The plot of the canonical variables (CV) 1 and 2 shows that the individual scores for the Teresópolis and Tijuca populations did not overlap along CV-1, whereas the Ilha Grande population was discriminated from the other two along CV-2 (Figure 1).

Table II - Coefficients of principal components (PC) analysis and of size-free canonical discriminant (CV) analysis for three samples of *Proechimys dimidiatus*.^a

Character	PC-1	r	CV-1	CV-2
Palatal length	0.374	0.874	-0.058 ^{ns}	0.622*
Zygomatic breadth	0.097	0.634	0.587*	0.248 ^{ns}
Nasal length	0.278	0.766	0.264 ^{ns}	-0.227 ^{ns}
Interorbital constriction	0.263	0.609	0.051 ^{ns}	-0.159 ^{ns}
Rostral width	0.388	0.541	-0.228 ^{ns}	-0.376*
Diastema length	0.456	0.869	-0.381*	0.537*
Rostral height	0.233	0.611	-0.412*	-0.235 ^{ns}
Cranial length	0.246	0.879	-0.122 ^{ns}	-0.221 ^{ns}
Basilar length	0.236	0.840	0.374*	0.000 ^{ns}
Rostral length	0.255	0.838	0.721*	-0.036 ^{ns}
Maxillary width	0.139	0.394	0.145 ^{ns}	0.160 ^{ns}
Post-palatal length	0.141	0.547	0.023 ^{ns}	-0.449&
Mandibular length	0.265	0.630	-0.211 ^{ns}	0.181 ^{ns}

^a The canonical discriminant analysis coefficients are expressed as Pearson correlation coefficients between cranial character values and scores obtained by canonical analysis. r is the correlation coefficient between the first eigenvector (PC-1) and the cranial characters.

* P < 0.01; ns, nonsignificant.

CDA also indicates the characters that most contribute to discrimination among populations (Pimentel, 1979; Neff and Marcus, 1980). The criterion commonly used for character selection is based on the relative magnitude of the eigenvector coefficients (Neff and Marcus, 1980; Campbell and Atchley, 1981). However, these coefficients are difficult to interpret in biological terms since, if two characters are correlated, the canonical coefficient of the character having the lower F will be close to zero (Morrison, 1976) even though each character may be as important as the other. The importance of each character for discriminating among populations may be better evaluated by transforming the coefficients into correlation vectors, which may be calculated from the correlation between individual scores for the canonical variables and the values of the characters for each individual (Strauss, 1985).

The correlation vectors between the original cranial characters and the canonical variables 1 and 2 are presented in Table II. These coefficients show that the Teresópolis and Tijuca populations differ in measurements of length, including diastema, rostrum and cranial base, and also in zygomatic breadth and rostral height. In contrast, the Ilha Grande population differed from the others in another set of

characters including palatal length, rostral width and post-palatal length. Diastema length also differentiated the Ilha Grande population from the Teresópolis and Tijuca populations.

Table III - Parameters concerning the regression analyses of the values of the 13 quantitative cranial characters of *P. dimidiatus* on the scores of the first principal component.

Character	b	t	F
Palatal length	0.363	12.62**	159.16**
Zygomatic breadth	0.159	5.74**	33.00**
Nasal length	0.315	8.33**	69.36**
Interorbital constriction	0.269	5.37**	28.86**
Rostral width	0.319	4.50**	20.26**
Diastema length	0.404	12.30**	151.23**
Rostral height	0.183	5.40**	29.18**
Cranial length	0.239	12.88**	165.90**
Basilar length	0.266	10.85**	75.70**
Rostral length	0.376	10.74**	115.32**
Maxillary width	0.168	2.998*	8.99*
Post-palatal length	0.145	4.58**	20.94**
Mandibular length	0.234	5.67**	32.20**

b, angular coefficient of the regression line; t, null hypothesis test for $b = 0$; F, value of analysis of variance of regression.

* $P < 0.01$; ** $P < 0.0001$.

P. dimidiatus is apparently restricted to the state of Rio de Janeiro (Moojen, 1952). In a study of speciation and evolution in the genus *Proechimys*, when analyzing the variation in morphological characters observed in *P. dimidiatus*, Moojen (1948, p. 373) concluded that "samples studied of *P. dimidiatus* are notably uniform throughout the geographic range of the species. The few biotypes detected seemed unworthy of subspecific rank". However, application of size-free CDA to the three *P. dimidiatus* populations studied here showed the existence of geographic variation in quantitative cranial characters in this species, even though Moojen (1948) had considered it monomorphic. The present analysis also indicated that the nature of morphological variation in *P. dimidiatus* populations differs between the continental populations and the insular population. It is interesting to note that the geographically closer populations (Teresópolis and Tijuca) showed the greatest morphometric differentiation (Figure 1).

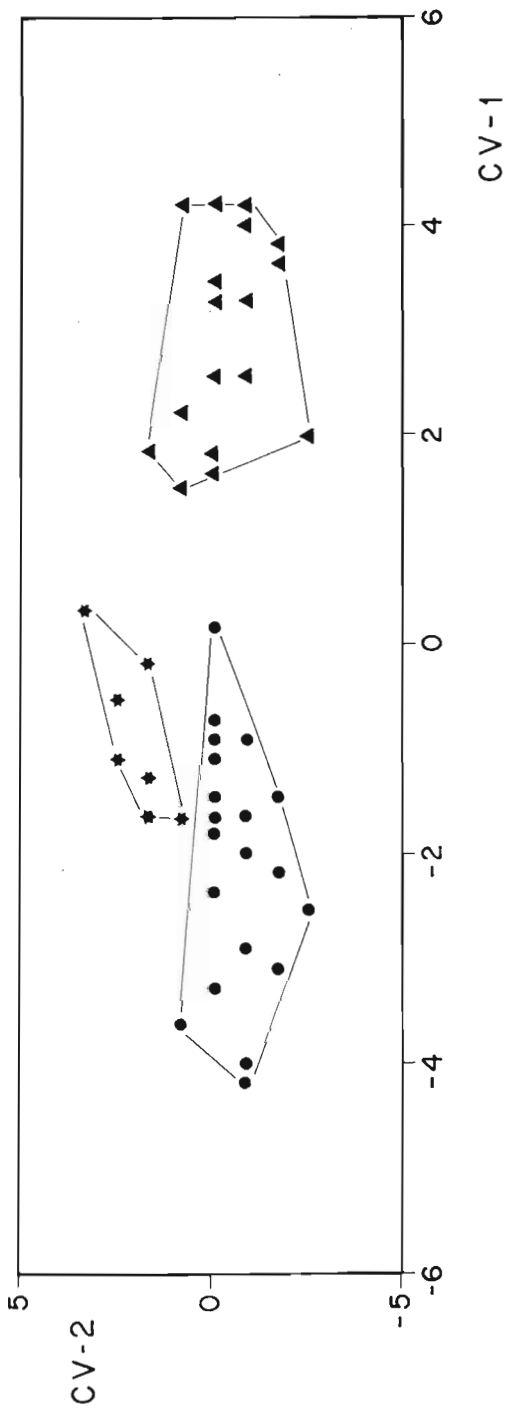


Figure 1 - Projection of individual scores for the *P. dimidiatus* populations from Teresópolis (triangles), Tijuca (circles) and Ilha Grande (asterisks) in the space of the canonical variables (CV) 1 and 2.

Full discrimination among the three *P. dimidiatus* populations showed that size-free CDA is an efficient procedure, considering that this species presents a moderate degree of intrapopulation variation in size. This intrapopulation variation component may indeed be a source of confusion in the analysis of interpopulation differentiation since, as demonstrated by Patton and Rogers (1983), an age component (intrapopulation) variation of 25% in *P. brevicauda* may reduce the discrimination in this taxon by as much as 30%. However, as demonstrated in the present study, size-free CDA did discriminate among the three *P. dimidiatus* populations despite the variation in size existing within each population.

Obviously, quantitative analysis of intraspecific variation at the morphological level is only one of the approaches to be used in studies of evolutionary biology and systematics. Other levels, including cytogenetic variation and allozyme and molecular variation, should also be investigated. The nature and magnitude of variation at these different levels may or may not be congruent (Patton, 1985; Schmidly *et al.*, 1988), although the primary objective of these studies is the integration of different sets of data in order to identify the geographic structure that represents intraspecific evolutionary units (Smith and Patton, 1988; Patton and Smith, 1989).

In the present study, application of size-free CDA was limited to an example of geographic differentiation in *P. dimidiatus*. However, this procedure could be applied to any group of organisms in which sources of intrapopulation variation may confuse discriminant analysis.

ACKNOWLEDGMENTS

We wish to thank Ivan Sazima, Augusto S. Abe, Rubens A. da Cunha and José Negreiros for a critical reading of the text which resulted in an improved manuscript. We are also indebted to Ulisses Caramaschi for permitting access to the mammal collection of the Museu Nacional. We also thank Elettra Greene for translating the original version of the manuscript.

This research was supported by CNPq (402265/87.4.ZO), FAPESP (88/2237-4, 89/0772-2 and 89/3405-0) and FAP (503/89). S.F.R. is the recipient of a CNPq fellowship.

Publication supported by FAPESP.

RESUMO

A análise discriminante canônica (ADC) é um procedimento empregado em estudos de variação geográfica, diferenciação interespecífica e macroevolução. Todavia, o emprego desta técnica, com organismos nos quais o tamanho dos indivíduos possa variar em função da amostragem, pode originar resultados espúrios, pois a discriminação entre as amostras será um artefato de amostragem. Apresentamos nesta nota um procedimento, a análise discriminante canônica independente do tamanho, que corrige estatisticamente o efeito da variação no tamanho dos indivíduos dentro das amostras. Neste procedimento, o efeito do tamanho é removido calculando a regressão de cada caráter sobre uma es-

timativa multidimensional de tamanho, o primeiro componente principal. A análise discriminante canônica é então efetuada sobre os resíduos resultantes da análise de regressão. A aplicação deste método é ilustrada em uma análise de diferenciação geográfica no roedor *Proechimys dimidiatus* (Echimyidae). Uma lista de comandos do programa estatístico SAS-PC, necessários para a implementação do método é também fornecida.

APPENDIX

Command file of SAS-PC procedures necessary for the execution of size-free canonical discriminant analysis. Three quantitative variables (x1-x3) and one categoric variable (taxon) are hypothetically considered in this file.

```

data dadorig;
  input taxon $ x1-x3;
  x1 = log10(x1);
  x2 = log10(x2);
  x3 = log10(x3);
  cards;
  ;
proc standard mean = 0 data = dadorig out = dadocent;
  by taxon;
  var x1-x3;
proc princomp cov n = 1 data = dadocent out = prin;
  var x1-x3;
proc corr;
  var x1-x3 prin1;
proc reg noprint data = prin outest = coefreg;
  r1: model x1 = prin1;
  r2: model x2 = prin1;
  r3: model x3 = prin1;
proc princomp cov n = 1 data = dadorig out = prin;
  var x1-x3;
proc score type = parms residual data = prin score = coefreg out = resid;
  var x1-x3 prin1;
proc candisc data = resid out = outcan;
  class taxon;
  var r1-r3;
proc corr;
  var r1-r3 can1 can2;
proc plot data = outcan;
  plot can2 * can1 = taxon / box;

```

REFERENCES

- Baker, A.J. (1980). Morphometric differentiation in New Zealand populations of the house sparrow (*Passer domesticus*). *Evolution* 34: 638-653.
- Bookstein, F.L. (1982). Foundations of morphometrics. *Ann. Rev. Ecol. Syst.* 13: 451-470.
- Bookstein, F.L., Chernoff, B., Elder, R., Humphries, J., Smith, G. and Strauss, R.E. (1985). *Morphometrics in Evolutionary Biology*. Special Publication 15. The Academy of Natural Sciences of Philadelphia, Penn.
- Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics*. Clarendon Press, London.
- Campbell, N.A. and Atchley, W.R. (1981). The geometry of canonical variate analysis. *Syst. Zool.* 30: 268-280.
- Chatfield, C. and Collins, A.J. (1980). *Introduction to Multivariate Analysis*. Chapman and Hall, London.
- Chesser, R.K. (1983). Cranial variation among populations of the black-tailed prairie dog in New Mexico. *Occas. Papers Mus., Texas Tech Univ.* 49: 1-25.
- Falconer, D.S. (1989). *Introduction to Quantitative Genetics*. 3rd. ed. Oliver and Boyd, Edinburgh.
- Humphries, J., Bookstein, F.L., Chernoff, B., Smith, G., Elder, R. and Poss, S. (1981). Multivariate discrimination by shape in relation to size. *Syst. Zool.* 30: 291-308.
- Lessa, H.P. and Patton, J.L. (1989). Structural constraints, recurrent shapes, and allometry in pocket gophers (genus *Thomomys*). *Biol. J. Linnean Soc.* 36: 349-363.
- Lindgren, B.W. (1976). *Statistical Theory*. 3rd. ed. MacMillan, N.Y.
- Moojen, J. (1948). Speciation in the Brazilian spiny rats (Genus *Proechimys*, Family Echimyidae). *Univ. Kans. Publ., Mus. Nat. Hist.* 1: 301-406.
- Moojen, J. (1952). *Roedores do Brasil*. Instituto Nacional do Livro, Rio de Janeiro.
- Morrison, D.F. (1976). *Multivariate Statistical Methods*. McGraw Hill, New York.
- Neff, N.A. and Marcus, L.F. (1980). *A Survey of Multivariate Methods for Systematics*. Privately Published, New York.
- Patton, J.L. (1985). Population structure and the genetics of speciation in pocket gophers, genus *Thomomys*. *Acta Zool. Fenn.* 170: 109-114.
- Patton, J.L. and Rogers, M.A. (1983). Systematic implications of non-geographic variation in the spiny rats genus *proechimys* (Echimyidae). *Z. Saeugetierkunde* 48: 363-370.
- Patton, J.L. and Smith, M.F. (1989). Population structure and the genetic and morphologic divergence among pocket gopher species (genus *Thomomys*). In: *Speciation and its Consequences* (Otte, D. and Endler, J.A., eds.). Sinauer, Sunderland, pp. 215-235.
- Pimentel, R.A. (1979). *Morphometrics*. Kendall/Hunt, Dubuque.
- Rohlf, F.J. and Bookstein, F.L. (1987). A comment on shearing as a method of "size correction". *Syst. Zool.* 36: 356-367.
- SAS Institute Inc. (1988). *SAS/STAT User's Guide*, Release 6.03 Edition. Cary, NC.
- Searle, S.R. (1971). *Linear Models*. Wiley, N.Y.
- Schmidly, D.J., Bradley, R.D. and Cato, P.S. (1988). Morphometric differentiation and taxonomy of three chromosomally characterized groups of *Peromyscus boylii* from East-Central Mexico. *J. Mamm.* 69: 462-480.

- Smith, M.F. and Patton, J.L. (1988). Subspecies of pocket gophers: Causal basis for geographic differentiation in *Thomomys bottae*. *Syst. Zool.* 37: 163-178.
- Sokal, R.R. e Rohlf, F.J. (1981). *Biometry*. 2nd. ed. Freeman, San Francisco.
- Straney, D.O. (1978). Variance partitioning and nongeographic variation. *J. Mamm.* 59: 1-11.
- Strauss, R.E. (1985). Static allometry and variation in body form in the South American catfish genus *Corydoras* (Callichthyidae). *Syst. Zool.* 34: 381-396.
- Thorpe, R.S. (1983). A review of the numerical methods for recognizing and analysing racial differentiation. In: *Numerical Taxonomy* (Felsenstein, J., ed.). Springer Verlag, Berlin, pp. 404-423.
- Thorpe, R.S. (1987). Geographic variation: a synthesis of cause, data, pattern and congruence in relation to subspecies, multivariate analysis and phylogenesis. *Boll. Zool.* 54: 3-11.
- Van Vleck, L.D. and Searle, S.R. (1979). *Variance Components and Animal Breeding*. Cornell University, N.Y.
- Winer, B.J. (1971). *Principles in Experimental Design*. 2nd. ed. McGraw-Hill, N.Y.
- Wright, S. (1954). The interpretation of multivariate systems. In: *Statistics and Mathematics in Biology* (Kempthorne, O., Bancroft, T.A., Gowen, J.W. and Lush, J.L., eds.). Iowa State College Press, Ames, pp. 11-33.

(Received January 10, 1990)