

The effect of insignificant characters on the multivariate analysis of simple patterns of geographic variation

R. S. THORPE

*Department of Zoology, University of Aberdeen,
Aberdeen AB9 2TN*

Accepted for publication April 1985

The multivariate analysis of a set of significant characters portrays a simple geographic cline which is evident even when very few characters are used. The effect of adding insignificant characters to the set of significant characters is studied, as is the effect of replacing significant characters by insignificant characters. The former (addition) causes only a slight decline in congruence between patterns of geographic variation but the latter (replacement) causes a substantial decline in congruence. The congruence between patterns of geographic variation obtained by multivariate analysis of independent character sets is plotted against character number and gives an S-shaped relationship when insignificant or significant plus insignificant characters are used. This is distinct from the convex asymptotic curve obtained when only significant characters are used. In spite of the reduction in congruence caused by the use of insignificant characters, multivariate analysis of insignificant characters consistently revealed the 'same' geographic pattern (i.e. 'cline') as the set of significant characters. However, it required 10 times as many insignificant characters to achieve this.

KEY WORDS:—Multivariate morphometrics – principal component – cline – geographic variation – congruence – race.

CONTENTS

Introduction	215
Methods	216
Results	217
Conclusions and discussion	222
References.	222

INTRODUCTION

Two previous studies on the relationship between character number and the reliability of multivariate analysis of simple pattern of geographic variation have shown that these patterns may be highly reliable with very few characters (Thorpe, 1985a, b). However, both of these previous studies used only characters that showed significant variation between geographic groups, e.g. $P < 0.05$ with a one-way analysis of variance (Thorpe, 1976).

Many multivariate studies of geographic variation do not include a test to see

if a character shows significant variation between geographic groups, and these studies presumably include both significant and insignificant characters. The advisability of this procedure can be investigated by taking a simple pattern of geographic variation based on a set of significant characters and testing the effect of both *replacing* them by insignificant characters and of *adding* insignificant characters.

A complex pattern of geographic variation may show, for example, a pair of distinct latitudinal categories with longitudinal clinal variation within these categories. A character may vary between the categories such that it exhibits significant geographic variation when all groups are considered but not when just the groups within one category are considered. In this particular sense, sets of significant characters used to study complex geographic variation can include characters that are 'significant' for some facets of the pattern but 'insignificant' for others. The effect of this can be assessed by taking a simple pattern of geographic variation based on a set of significant characters and then *adding* insignificant characters.

The simple pattern of geographic variation under investigation is the cline between five geographic groups of the grass snake (*Natrix natrix*) (Thorpe, 1984) as used by Thorpe (1985b). The broadly clinal nature of this pattern (see below), based on a principal component/coordinate analysis of normalized group means of the 18 significant characters, has been shown to be stable (Thorpe, 1985b). The 18 significant characters were selected from a total set of 89 characters, drawn from six distinct character systems (Thorpe, 1985a), which leaves 71 insignificant characters. This allows one to test the effect of insignificant characters on the multivariate analysis of this simple clinal pattern of geographic variation.

METHODS

The method for establishing a significant cline using Kendall's τ , the method for assessing congruence between patterns of geographic variation based on different character sets and the procedures (A and B) for investigating the relationship between congruence and character number are as described by Thorpe (1985b).

The procedure for assessing the relationship of character number to congruence between completely independent character sets (A) and between the total set and a subset (B) were carried out on the set of 71 insignificant characters and separately on the set of 89 characters not selected for significance (i.e. 18 significant, plus 71 insignificant characters). The former gives a ratio of 0 : 4 significant to insignificant characters whilst the latter gives a ratio of about 1 : 4 significant to insignificant characters. The ratio of 1 : 0 significant to insignificant characters is represented by the previously published study on 18 significant characters (Thorpe, 1985b) and the ratios of 1 : 1, 1 : 2 and 1 : 3 significant to insignificant characters were obtained by randomly selecting the appropriate number of characters of appropriate type. That is the 1 : 1 ratio is based on 18 significant plus 18 insignificant characters, the 1 : 2 ratio is based on 18 significant plus 36 insignificant characters, and the 1 : 3 ratio is based on 18 significant plus 54 insignificant characters. Since the maximum number of characters in independent sets is the integer value of half the total this gives a

maximum of 9, 18, 27, 36, 44 and 35 respectively for the ratios 1:0, 1:1, 1:2, 1:3, 1:4 and 0:4.

The relationship between congruence and character number often follows an **S**-shaped curve. It is not possible to formulate the precise nature of this **S**-curve because of the variation ('noise') in the data. Nevertheless a reasonable fit is given by the following simple formula (which is related to the logistic equation):

$$|r| = \frac{a + be^{d(c-n)} - (a+b)e^{-dn}}{1 + e^{d(c-n)}}, \quad (1)$$

where $|r|$ is the congruence, n is the number of characters, a is the upper asymptote, b is the lower asymptote, c is the inflexion point (in units of n) and d is the slope.

RESULTS

Independent sets of insignificant characters—Model A

The congruence between patterns of geographic variation portrayed by analyses based on completely independent sets of characters drawn from the pool of 71 insignificant characters is plotted against character number (Fig. 1). An **S**-curve is fitted to the scatter of mean congruence by formula (1). The curve has a lower asymptote of $0.40r$, an upper asymptote of $0.74r$, an inflexion point of 25.8 characters and a slope of 0.24. It appears then that the highest mean (Model A) congruence obtainable from insignificant characters is $0.74r$ and that it needs in excess of 35 characters to achieve this.

The minimum congruence appears to represent only the first part of an **S**-curve, as the upper asymptote is not approached. The lower asymptote is about

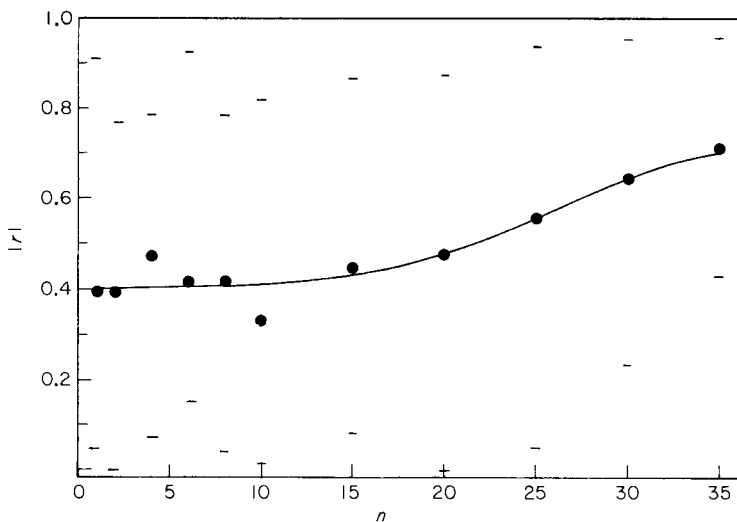


Figure 1. Congruence, $|r|$, of analyses based on independent character sets (Model A) drawn from the 71 insignificant characters, plotted against character number, n . The circles indicate the mean congruence of 10 runs for a given number of characters and the bars indicate the maximum and minimum congruence.

$0.05r$ and the congruence does not begin to increase until after 25 characters are used.

Subsets of insignificant characters—Model B

The congruence between the pattern of geographic variation portrayed by the analysis based on the total 71 insignificant characters and patterns portrayed by analyses based on subsets of these characters is plotted against character number (Fig. 2). The mean congruence rises gradually in a slightly convex curve from about $0.4/0.5r$ when few characters are used towards 1.0 when the number of characters approaches the total. This gradual increase is in sharp contrast to the distinct hyperbolic curve encountered in previous studies based on significant characters (Thorpe, 1985a, b).

The minimum congruence has an approximately concave relationship with the number of characters. Minimum congruence remains low, generally under $0.1r$ until 25 characters are used and then increases towards $1.0r$ as the total number of characters is approached.

The multivariate analysis of the 71 insignificant characters does show clinal variation (i.e. $\tau > 0.8$) but it is evident from Fig. 3 that a very large number of characters, about 55, is needed to achieve a better than 90% frequency of studies showing clines.

Independent sets of significant plus insignificant characters—Model A

The congruence between patterns of geographic variation portrayed by analyses based on completely independent sets of characters, drawn from the pool of 18 significant plus 71 insignificant (i.e. 89) characters, is plotted against character number (Fig. 4). The mean congruence is clearly S-shaped and a curve is fitted by formula (1) which indicates a lower asymptote of $0.44r$, an upper asymptote of $0.86r$, an inflexion point at 20.9 characters and a slope of

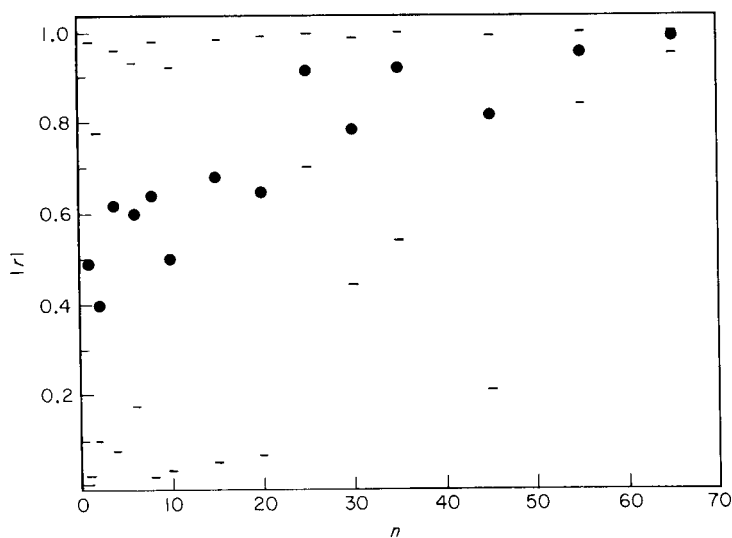


Figure 2. Congruence $|r|$, between the analysis based on the total 71 independent characters and analyses based on smaller subsets (Model B) plotted against character number, n . Symbols as for Fig. 1.

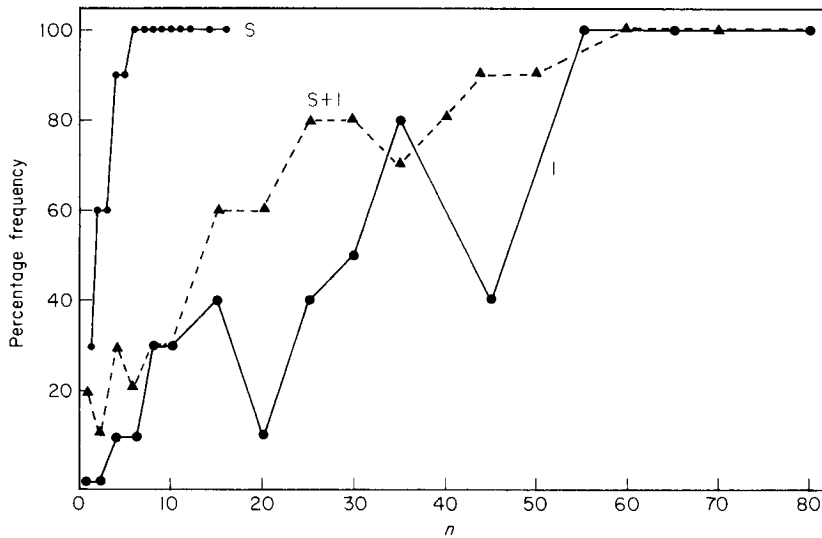


Figure 3. Percentage frequency of analyses showing 'clinal' variation (i.e. $\tau \geq 0.8$ between geographic position and component score) plotted against character number, n . S, S+I and I indicate analyses based on characters drawn from the set of significant (18), significant plus insignificant (18+71) and insignificant (71) characters, respectively.

0.15. It is apparent from Fig. 4 that a substantial number of characters, i.e. 35 or more, is needed to achieve a congruence of over 0.8 and the addition of further characters will not increase the congruence beyond 0.86.

The minimum congruence appears to have a concave relationship to character number, i.e. the first part of the S-curve. The lower asymptote is under 0.1 and the congruence does not noticeably increase until many characters are used.

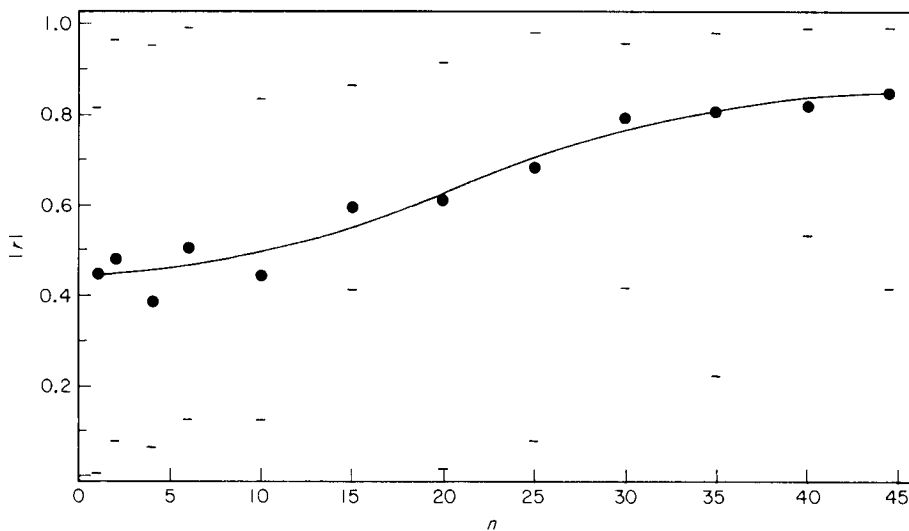


Figure 4. Congruence, $|r|$, of analyses based on independent character sets (Model A) drawn from the 18 significant plus 71 insignificant characters, plotted against character number, n . Symbols as for Fig. 1.

Subsets of significant plus insignificant characters—Model B

The congruence between the pattern of geographic variation portrayed by the analysis based on the total 89 (18 significant plus 71 insignificant) characters and patterns portrayed by subsets of these characters is plotted against character number (Fig. 5). The relationship between mean congruence and character number is not sharply hyperbolic but rises gradually from about $0.6r$ when few characters are needed towards $1.0r$ when the number of characters used approaches the total. When an S-curve is fitted by formula (1) it indicates a lower asymptote at $0.59r$ and an upper asymptote at $1.00r$.

The minimum congruence (Fig. 5) also fails to be hyperbolic in relation to character number but instead can be fitted by an S-curve (formula 1). The S-curve has a lower asymptote of $0.27r$, an upper asymptote of $0.98r$, an inflexion point at 35.5 characters and a slope of 0.18.

The analysis based on the full set of 89 characters portrays clinal geographic variation but it is evident from Fig. 3 that a very large number of characters, about 60, is needed to achieve a better than 90% frequency of analyses showing clines.

Replacement by, and addition of, insignificant characters

The previous analysis of 18 significant characters indicated a congruence of 0.93 between independent sets (Model A) when the maximum number of characters, i.e. 9, are used (N.B. the maximum number of usable characters in this procedure is the integer value of half the total). The effect of adding insignificant characters is revealed by comparing this congruence to that obtained when the maximum number of characters are drawn (i.e. 44) from the set of 18 significant plus 71 insignificant characters (Fig. 4). In this latter study, where the ratio between significant and insignificant characters is 1:4, the

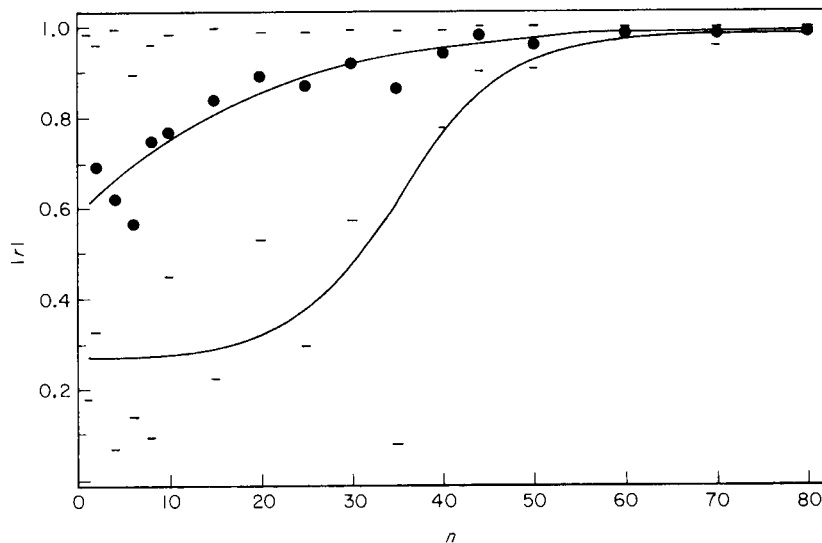


Figure 5. Congruence, $|r|$, between the analysis based on the total 89 characters (18 significant plus 71 insignificant) and analyses based on smaller subsets (Model B), plotted against character number, n . Symbols as for Fig. 1.

mean congruence between independent sets has declined from 0.93 to 0.85 even though more characters are used.

A more complete picture of the decline in mean congruence with the addition of insignificant characters is given by Fig. 6 where the significant:insignificant ratios of 1:1, 1:2, 1:3 are included. In spite of the increase in characters with the addition of insignificant characters the congruence between patterns of geographic variation based on independent sets declines. However, the decline in congruence is only slight and cannot continue indefinitely since it must flatten to an asymptote as the proportion of significant characters decreases to approach zero. This asymptote should not be lower than that obtained by entirely insignificant character, i.e. a congruence of 0.74.

The effect of replacing significant by insignificant characters can be seen by comparing studies based on a constant number of characters, i.e. nine characters, when the ratio of significant to insignificant characters changes from 1:0 to 0:4 through 1:1, 1:2, 1:3 and 1:4. The decline in congruence due to replacement of significant characters by insignificant characters is sharp (Fig. 6). The mean congruence in patterns of geographic variation based on independent sets of nine significant characters is 0.93 whereas it declines progressively to 0.47 when nine insignificant characters are used.

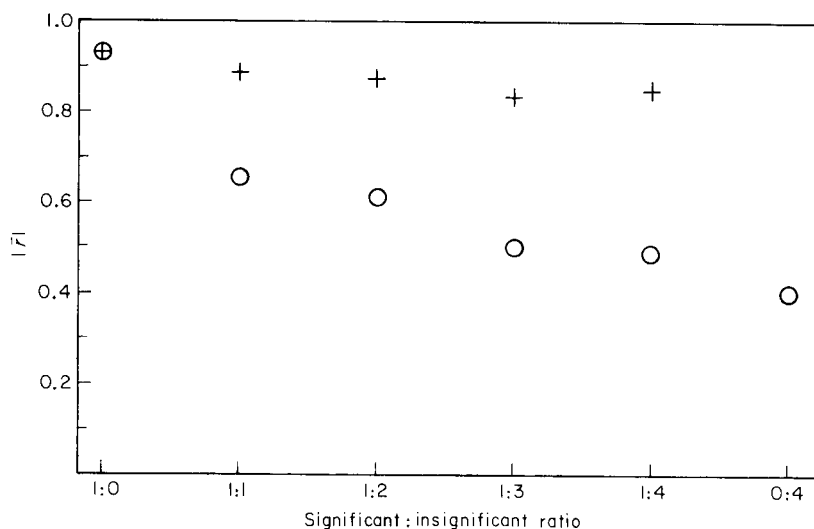


Figure 6. The decline in mean congruence, $|r|$, between analyses based on independent sets of characters (Model A) due to the progressive *addition* (+) of insignificant characters to the nine significant characters and due to the progressive *replacement* (o) of the nine significant characters by insignificant characters. The cross in the circle indicates the mean congruence between nine-character analyses drawn from the 18 significant characters (N.B. the maximum number of characters in Model A is the integer value of half the total set). The crosses indicate the congruence between analyses with the progressive addition of insignificant characters, i.e. 1:1 indicates a total set of 18 significant plus 18 insignificant from which are drawn sets of 18 characters, 1:2 indicates a total set of 18 significant plus 36 insignificant characters from which are drawn sets of 27 characters, 1:3 indicates a total set of 18 significant plus 71 insignificant characters from which are drawn sets of 44 characters (i.e. Fig. 4). The circles indicate the mean congruence when the nine significant characters are replaced by insignificant characters. In this series the number of characters used in the analyses remains the same, i.e. 9, for 1:0, 1:1, 1:2, 1:3, 1:4 and 0:4 even though the total set from which they are drawn is the same as above (N.B. 0:4 is the set of 71 insignificant characters only).

CONCLUSIONS AND DISCUSSION

Irrespective of whether the analyses are based on insignificant characters, significant characters or a mixture of both, they all consistently indicate clinal geographic variation if enough characters are used (Fig. 3). The essential point is that whilst one requires only a few significant characters to achieve this consistency (i.e. six) one needs around 10 times as many characters as this if one uses insignificant characters or a 1:4 mix of significant/insignificant characters.

The congruence between patterns of geographic variation based on independent sets of characters is only slightly depressed by the addition of even large numbers of insignificant characters. Consequently, analysis of complex patterns of geographic variation should not be perturbed by characters that are significant over all groups but not for subsets of groups exhibiting a particular facet of the geographic variation (see above). Since congruence does decrease, albeit slightly, with the addition of insignificant characters these may be best excluded from studies of geographic variation; however, the case for exclusion is only weak.

Whilst the addition of insignificant characters has little effect on congruence the replacement of significant characters by insignificant characters has a marked effect on congruence and rapidly reduces congruence to low levels. Moreover, the analyses based on, or including, insignificant characters do not show the sharp hyperbolae in Model B (Figs 2 & 5) characteristic of reliable patterns based on significant characters (Thorpe, 1985a, b), where high levels of congruence are reached with very low numbers of characters. Consequently, multivariate studies of geographic variation that do not test for significance may not be presenting reliable patterns unless they are based on very large numbers of characters. However, such studies are based on few characters (Thorpe, 1976, 1983) irrespective of whether insignificant characters are excluded.

Previous studies have shown that the relationship between congruence of simple patterns of geographic variation based on independent sets of significant characters (Model A) and character number is a simple, convex asymptotic curve. This study indicates that this relationship may perhaps be more generally thought of as an S-shaped curve. Studies based on significant characters depicting a reliable pattern may emphasize just the convex part of the curve indicating the upper asymptote (Thorpe, 1985a, b), whilst studies based on insignificant characters may emphasize just the concave part of the curve indicating the lower asymptote (Fig. 1). On the other hand, studies of simple patterns based on a mix of significant and insignificant characters may show the complete S (Fig. 4) with both lower and upper asymptotes.

The S-curve of formula (1) fits the scatter of mean congruence against character number (Model A, Figs 1 & 4) well but little meaning can be attributed to the precise formulation of the fitted curve because of the variability ('noise') in the data.

REFERENCES

- THORPE, R. S., 1976. Biometric analysis of geographic variation and racial affinities. *Biological Reviews*, 51: 407-452.
- THORPE, R. S., 1983. A review of the numerical methods for recognising and analysing racial differentiation. In J. Felsenstein (Ed.), *Numerical Taxonomy: Proceedings of a NATO Advanced Studies Institute*:

404–423. NATO Advanced Studies Series G (Ecological Sciences), No. 1. Berlin, Heidelberg and New York: Springer-Verlag.

- THORPE, R. S., 1984. Geographic variation in the western grass snake (*N. natrix helvetica*) in relation to hypothesized phylogeny and conventional subspecies. *Journal of Zoology*, 203: 345–355.
- THORPE, R. S., 1985a. Character number and the multivariate analysis of simple patterns of geographic variation: categorical or “stepped clinal” variation. *Systematic Zoology*, in press.
- THORPE, R. S., 1985b. Clines: Character number and the multivariate analysis of simple patterns of geographic variation. *Biological Journal of the Linnean Society*, 26: 201–214.