Kontyû, Tokyo, **51** (3): 351–357. September 25, 1983

Summarization of Taxonomic Information in the Japanese Andrenid Bees by Principal Component Analysis*

Osamu TADAUCHI

Entomological Laboratory, Faculty of Agriculture,
Kyushu University, Fukuoka, 812 Japan

Synopsis Taxonomic information in morphological characters commonly used for the taxonomy of the genus *Andrena* (Hymenoptera, Andrenidae) was summarized based on correlations of characters. Twelve subsets consisting of various combinations of 130 characters obtained from 85 taxa or OTUs of the genus of Japan were analyzed by principal component analysis. Taxonomic information of each subset was summarized as principal components (P.C.s) and a summarization value (*S*). The first 14 P.C.s accounted for 71.10% of the total variance among the 130 characters. Thoracic characters were found to have much redundancy of taxonomic information, whereas head characters less redundancy.

Introduction

Taxonomic characters generally include much redundancy of information. Some characters may be largely independent, while others may form sets of correlated characters. JARDINE and SIBSON (1971) described five ways in which correlation of characters were interpreted as follows: redundant, logical, functional, statistical, and taxonomic correlations. MAYR (1969) discriminated two different kinds of correlations, redundancy, which includes the first three ways of JARDINE and SIBSON, and phyletic correlation.

Recently some attempts have been done to produce a database of taxonomic information in several organisms (e.g., Australian grass, WATSON & DALLWITZ, 1980). The author also attempts to construct a database of taxonomic information in the genus *Andrena* (Hymenoptera, Andrenidae), which includes as less redundancy of taxonomic information as possible. The purpose of the present study is to summarize taxonomic information in morphological characters commonly used for the taxonomy in the genus.

Material and Methods

Material and characters

The original data used in the present study were obtained as a by-product of numerical taxonomy in the genus *Andrena* of Japan (TADAUCHI, 1982 a). One hundred and thirty characters used in the present study are commonly employed

* Contribution from the Entomological Laboratory, Faculty of Agriculture, Kyushu University, Fukuoka (Ser. 3, No. 128).

for the taxonomy in the genus *Andrena*. Eighty-five taxa or OTUs and the 130 characters used were listed in the previous paper (TADAUCHI, 1982 a, Tables 1–2). The twelve subsets are as follows:

A. *Total characters*
   1. 130 original subset: original characters (TADAUCHI, 1982 a, Table 2, Code Nos. 1–130).

B. *Head-thoracic characters*
   2. 40 head subset: 40 characters (33 structural and 7 pubescence) derived from head region only (TADAUCHI, 1982 a, Table 2, Code Nos. 2–34 & 83–89).
   3. 58 thoracic subset: 58 characters (35 structural and 23 pubescence) derived from the thoracic region only (TADAUCHI, 1982 a, Table 2, Code Nos. 35–69 & 90–111).

C. *Total structural characters*
   4. 40 hair subset: 40 characters related to pubescence on the body (TADAUCHI, 1982 a, Table 2, Code Nos. 83–122).
   5. 50 sculptural subset: 50 characters related to sculptures of the body (TADAUCHI, 1982 a, Table 2, Code Nos. 5, 8–10, 13, 16–18, 22, 25, 26, 28, 29, 35–48, 50–57, 64, 66 & 70–82).
   6. 82 structural subset: 82 characters (one body sized and 81 structural) related to structures of the body (TADAUCHI, 1982 a, Table 2, Code Nos. 1–82).

D. *Randomly selected characters*
   7. 40 random subset: 40 characters randomly selected from the orignial data using random numbers.
   8. 70 random subset: 70 random characters selected as above.
   9. 100 random subset: 100 random characters selected as above.

E. *Total characters based on different number of OTUs*
   10. 26 OTUs subset: 130 characters based on 26 OTUs selected at least one OTU from one subgenus.
   11. 41 OTUs subset: 130 characters based on 41 OTUs selected at least two OTUs from one subgenus wherever available.
   12. 50 OTUs subset: 130 characters based on 50 OTUs selected at least three OTUs from one subgenus wherever available.

The data were processed on FACOM M-200 computer at the Computer Center of Kyushu University.

Method

Each of the twelve subsets was analyzed by principal component analysis (PCA) in SAC (system for *Andrena* classification, TADAUCHI, 1981) in order to achieve a parsimonious summarization of data from multistate morphological characters. This procedure constructs a new set of characters as weighed linear combinations of the original set of characters. The new set has the convenient properties of

Table 1. Eigenvalue (E), percentage (P), and accumulated percentage (A.P.) of total variance among characters of 130 original subset contributed by each principal component (P.C.) with eigenvalue greater than 1.0.

P.C.	E	P	A.P.	P.C.	E	P	A.P.
1	22.25	17.11	17.11	15	2.15	1.66	72.76
2	15.77	12.13	29.24	16	1.94	1.49	74.25
3	7.94	6.11	35.35	17	1.85	1.42	75.67
4	7.11	5.47	40.82	18	1.64	1.27	76.94
5	5.80	4.46	45.28	19	1.55	1.19	78.13
6	5.46	4.20	49.48	20	1.46	1.12	79.25
7	4.94	3.80	53.28	21	1.39	1.07	80.32
8	4.42	3.41	56.69	22	1.35	1.04	81.36
9	4.08	3.14	59.83	23	1.31	1.01	82.37
10	3.39	2.61	62.44	24	1.25	0.96	83.33
11	3.11	2.39	64.83	25	1.17	0.90	84.23
12	3.00	2.30	67.13	26	1.13	0.86	85.09
13	2.67	2.06	69.19	27	1.06	0.82	85.91
14	2.48	1.91	71.10	28	1.01	0.78	86.69

Table 2. The number of principal components (P.C.s) and the summarization values (S) of the nine character subsets at the criterion level of 70.0% of the total variance and of eigenvalue greater than 1.0.

Subset	70.0%		Eigenvalue>1.0	
	No. of P.C.s	S	No. of P.C.s	S
40 random	10	4	13	3
70 random	11	6	17	4
100 random	13	8	22	5
130 original	14	9	28	5
40 hair	7	6	9	4
50 sculptural	8	6	11	5
82 structural	12	7	19	4
40 head	10	4	12	3
58 thoracic	6	10	12	5

being uncorrelated and summarizes most of the total variance in the original data as minimum principal components as possible.

Results and Discussion

A correlation matrix of each character subset was calculated and subjected to PCA. Table 1 shows eigenvalue, percentage and accumulated percentage of the total variance among 130 original subset contributed by each principal component (P.C) with eigenvalue greater than 1.0. It was observed that the first 2

P.C.s had large percentage of the total variance, 17.11% and 12.13%, respectively. The first 14 P.C.s accounted for 71.10% of the total variance among the 130 characters. Fig. 1 shows the accumulated percentage of the total variance contributed by P.C.s with eigenvalue greater than 1.0 for the nine subsets excluding the three subsets based on different number of OTUs. The 58 thoracic subset on the utmost left of the figure had only 12 P.C.s to account for 85.48% of the total variance. On the other hand, the 130 original subset on the utmost right of the figure had 28 P.C.s to account for 86.69% of the total variance.

Table 2 shows the number of P.C.s and summarization values (S) in the nine character subsets at the criterion level of 70.0% of the total variance and of the eigenvalue greater than 1.0. The S was calculated by the number of characters divided by the number of P.C.s obtained at the criterion level. The result showed that the number of P.C.s and the S generally increased according to an increase of the number of characters. This was especially observed in the result of the random subsets. The 130 original subset had 14 P.C.s and considerably high S (9). On the other hand, the 40 random subset had 10 P.C.s and the lowest S (4). The similar result was obtained at the criterion level of the eigenvalue greater than 1.0, having 13 P.C.s and S of 3 for the 40 random subset in contrast to 28 P.C.s and S of 5 for the 130 original subset. The result suggests that the more the characters are randomely employed, the more the correlated characters would increase. Accounting for the 70.0% of the total variance, the 40 random subset needed 10 P.C.s and the 130 original subset needed only 14 P.C.s.

In case of total structural character subsets 7 P.C.s and S of 6 were obtained for the 40 hair subset, and 8 P.C.s and S of 6 were obtained for the 50 sculptural subset at the 70.0% criterion level. These scores of S are very high in comparison

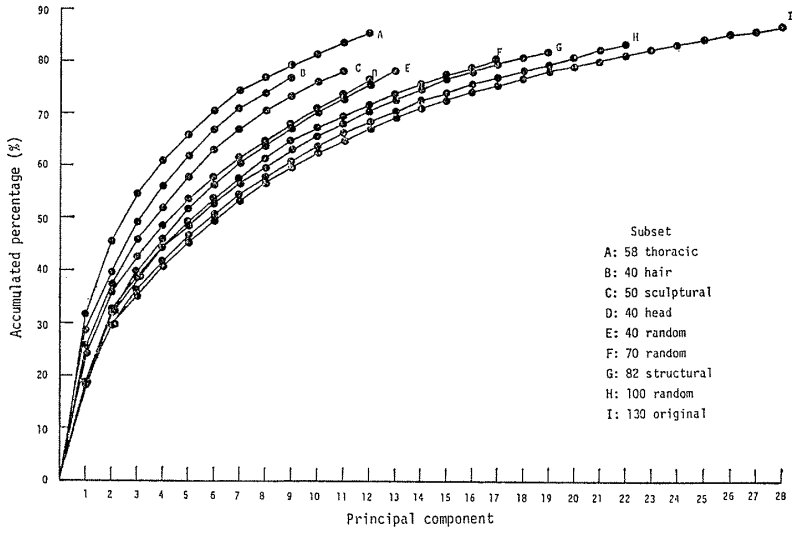

Fig. 1. Accumulated percentage of total variance among characters contributed by principal components in the nine character subsets studied.

with the other subsets, viz. both the 40 head subsets and the 40 random subset showed 10 P.C.s and *S* of 4 at the same criterion level. Similar result was obtained at the criterion level of eigenvalue greater than 1.0. The *S* of 4, 3, 3 were obtained from the 40 hair, the 40 head, and the 40 random subsets, respectively. It seems reasonable that the homogeneous characters such as those related to pubescence obtained from various parts of the body highly correlates with each other, showing high *S*. It is considered that the subsets consisting of homogeneous characters such as those related to pubescence on the body have much redundancy and lower information.

As for the subsets consisting of head-thoracic characters different results were obtained. The 40 head subset had 10 P.C.s and *S* of 4 at the 70.0% criterion level. The scores were the same as those derived from the 40 random subset. The 58 thoracic subset had, on the other hand, 6 P.C.s and very high *S* of 10 at the 70.0% criterion level.

Table 3 shows the number of P.C.s and the *S* in the four subsets which consist of 130 characters different in number of OTUs treated. The number of P.C.s increased according to an increase of the number of OTUs, whereas *S* decreased according to an increase of the number of OTUs. The more the OTUs are added, the more the information seems to increase. At the 70.0% criterion level 10 P.C.s and *S* of 13 were obtained from the 26 OTUs subset. On the other hand, 14 P.C.s and *S* of 9 were resulted from the 85 OTUs subset.

MAYR (1969) stated, "One can count and evaluate in classification only those characters that are reasonably independent of each other. Just exactly 'independent' is, and how this can be determined is still controversial." One promising approach to solving this problem involves the use of PCA and related techniques. MOLINA-PARDO (1973) investigated two subsets of characters consisting of 69 measurable and 148 commonly used in conventional taxonomy obtained from 73 OTUs of the new world andrenid bees by PCA. He found that the measurable characters involved little information having only 2 P.C.s accounting for 81.05% of the total variance. On the other hand, the characters used in taxonomy of the genus *Andrena* involved more information having 18 P.C.s and *S* of 8 (calculated

Table 3. The number of principal components (P.C.s) and the summarization values (*S*) of the four subsets based on different number of OUTs at the criterion level of 70.0% of the total variance and of eigenvalue greater than 1.0.

Subset	70.0%		Eigenvalue>1.0	
	No. of P.C.s	*S*	No. of P.C.s	*S*
26 OTUs	10	13	24	5
41 OTUs	11	12	26	5
50 OTUs	12	11	27	5
85 OTUs	14	9	28	5

by the author) at the 70.0% criterion level. The present study reveals that there is some difference in taxonomic information among characters commonly used in taxonomy. The 58 thoracic subset having the highest S showed the highest correlation among characters, whereas the 40 head subset having the lowest S the lowest correlation. From the view point of independence of characters, the 58 thoracic subset involves the lowest information for the number of characters, whereas the 40 head subset the highest one. Above results showed that among the 130 characters commonly used for the taxonomy of the genus *Andrena* the thoracic characters have much redundancy of taxonomic information. Observation of the 130 original subset showed that the P.C.s with low eigenvalues increased in number according to an increase of the number of characters (Fig. 1). These lower P.C.s are considered to be related with only one or two characters. TADAUCHI (1982 b, 1982 c) investigated 40 and 50 characters from hairs and integumental sculptures respectively by factor analysis. In that study several characters which had no connection with common factors (P.C.s) extracted by factor analysis were obtained. These characters were considered to have their own variations independent of the other characters. Therefore, it is necessary not only to summarize information about correlated characters but also to give attention to the independent characters for compilation of taxonomic information. Once these characters are selected, recording for taxonomic information of a database will be easier and time saving.

Acknowledgements I am grateful to Prof. Y. HIRASHIMA of Kyushu University for his valuable advices on andrenid taxonomy. My sincere thanks are due to Prof. Ch. ASANO of Research Institute of Fundamental Information Science of Kyushu University for his kind advices on multivariate statistics. I am also indebted to Assoc. Prof. K. MORIMOTO of Kyusyu University, Prof. Emerit. T. OKADA of Tokyo Metropolitan University and Prof. S. SAKAI of Daito Bunka University for their useful advices and constant encouragements. This work was supported in part by a Grant-in-Aid Scientific Research (No. 57480043) from the Ministry of Education, Japan.

References

JARDINE, N., & R. SIBSON, 1971. Mathematical Taxonomy. 286 pp. Wiley, London.

MAYR, E., 1969. Principles of Systematic Zoology. 428 pp. McGraw-Hill, New York.

MOLINA-PARDO, A., 1973. A phenetic analysis of new world bees of five subgenera of the genus Andrena (Hymenoptera: Apoidea). 181 pp. Doctor thesis for the Ph. D. in Univ. Illinois.

TADAUCHI, O., 1981. Taxonomic working system by computer (SAC) with application to Japanese Andrenid bees. *Esakia, Fukuoka*, (17): 161–182.

——— 1982 a. A numerical taxonomic study of the genus *Andrena* (Hymenoptera, Andrenidae) of Japan. *J. Fac. Agr., Kyushu Univ.*, **26**: 169–191.

——— 1982 b. Character correlations of hairs in the Japanese andrenid bees. *Kontyû, Tokyo*, **50**: 411–424.

——— 1982 c. Factor analysis of integumental sculptures in the Japanese Andrenid bees. *Esakia, Fukuoka*, (19): 135–150.

WATSON, L., & M. J. DALLWITZ, 1980. Australian Grass Genera. Anatomy, Morphology, and Keys. 209 pp. Australian Univ. Press, Canberra.