Aquilo Ser. Zool. 20: 33–40. 1979

Quantitative methods for population analysis applied to a Drosophila (Diptera, Drosophilidae) collection

GERHARD BÄCHLI

Zoologisches Museum der Universität Zürich, Zürich, Schweiz

Received 20. IX. 1978

Bächli, G. 1979: Quantitative methods for population analysis applied to a *Drosophila* (Diptera, Drosophilidae) collection. — Aquilo, Ser. Zool. 20: 33–40.

Diversity analysis, niche analysis, cluster analysis, principal component analysis and discriminant analysis of a *Drosophila* collection are presented and the obtained results are discussed.

I. Introduction

Faunistic collections as well as collections made for the study of the properties of natural populations normally yield a lot of detailed data about species composition and abundance in different collection samples. These data can be used to answer such questions as: Which species are abundant, dominant or rare? Which collection sites are rich in species and specimens? Is it obvious that the species are unequally distributed over the collection sites? And, comparing the data with similar data of other collection areas, which main faunistic conformity or difference is detectable?

As a result of the papers of MacArthur (1955), Levins (1968) and others, the quantitative analysis of the population structure has gained in importance. Therefore it seems worth an additional effort to extract as much as possible of the quantitative information about population properties held in faunistic data.

In this paper I present some numerical methods suitable for a more detailed evaluation of a *Drosophila* collection and discuss some of the results obtained.

II. Material and methods

Between August 22nd and September 2nd 1977 collections of Drosophilidae were made in the central Alps near Leuk VS, Switzerland. The flies were caught by sweeping nets over fermenting banana bait exposed on open plates. The collection area was a mountain forest on a continuous slope. Collection sites 1 to 5 were arranged in a horizontal line at 1900 m above sea level, sites 6 to 10 in a line at 1400 m, sites 11 to 15 in a line at 900 m, and sites 16 to 20 in a line at 600 m, at the bottom of the valley. The results of the collections are shown in Table 1.

Table 1. Number of flies collected at Leuk in 1977. The letters in the first column correspond to the following altitudes: O = 1900 m, M = 1400 m, U = 900 m, T = 600 m. The index of diversity is corrected for sample size by the method of Stahel. The most abundant species from *D. subobscura* to *D. kuntzei,* are called "common species". Abbreviations: D = *Drosophila,* S = *Scaptomyza,* A = *Amiota,* L = *Leucophenga.*

Altitude	Collection sites	D. subobscura	D. obscura	D. testacea	D. transversa	D. alpina	D. ambigua	D. phalerata	D. nigrospar sa	D. immigrans	S. pallida	D. kuntzei	D. cameraria	D. melanogaster	S. graminum	D. funebris	D. hydei	D. bifasciata	D. confusa	D. histrio	D. subsilvestris	A. variegata	L. maculata	Total number of flies	Index of diversity
O	1	35	1	1	4	8	47		46		9		1											152	0.69
	2	7	2		3	8	16		54		1		1		3									95	0.64
	3	9	1		1	6	16		15		1													49	0.70
	4	6			1	2	5		36				3		2									55	0.55
	5	7	2		1	10	7	1	4						1									33	0.82
M	6	186	28		3	39	21		2		2		2											283	0.49
	7	168	17		13	28	23		6															255	0.51
	8	159	19		2	29	29		1				1											240	0.47
	9	183	21	1	1	38	39						1											284	0.47
	10	242	44		2	52	38		1		1							1	1				1	383	0.50
U	11	1092	68	30	4	44	10	20		5		1	6	1										1281	0.29
	12	435	26	10	9	17	4	16		1	2			1										521	0.33
	13	613	28	9	5	19	5	9																688	0.23
	14	590	26		1	21	6	1								1								646	0.17
	15	624	27	2	8	15	2	5			1					1								685	0.19
T	16	495	46	72	6	5	12	30		12		7				1								686	0.47
	17	498	50	36	2	8	4	5		9		1					1							614	0.34
	18	784	33	87	76	1	11	25		6	4	3		2		1				1		1		1035	0.42
	19	529	22	35	112	2	3	21			8	2		2			1							737	0.44
	20	701	47	189	105	4	7	58		13	1	12		1							1			1139	0.55
Total		7363	508	472	359	356	305	191	165	46	30	26	15	7	6	4	2	1	1	1	1	1	1	9861	0.48

III. Diversity analysis

A common measure of the diversity of biological collections is the index of Shannon and Wiener, H'. Hurlbert (1971) criticizes the use of the index as meaningless and ill-applied. He suggests that the number of species (species richness) estimated or found in a sample is a clearer and sufficiently exact indicator of biotic diversity. However, the number of species contained in the 20 samples of Table 1 is more or less constant and does not discriminate between samples and altitudes. The index of diversity on the contrary is much more distinguishing: The highest diversity index was found in the samples 1 to 5, the lowest in 11 to 15, while the samples 6 to 10 and 16 to 20 yielded an index of medium size. In relation to the different altitudes we see a clinical reduction of the diversity index from the top samples group, 1 to 5, to the samples group 11 to 15, whereas the somewhat higher diversity index in the bottom samples group, 16 to 20, could be explained by a different type of woodland at the bottom of the valley.

Lloyd and Ghelardi (1964) showed that diversity can be split into the components "number of species" and "equitability". In our data the equitability index is highly correlated with the diversity index ($r = 0.96$). Because the number of species is reasonably constant the altitudinal differences of the diversity index are caused by different values of the equitability component.

It can be expected that the *Drosophila* population at a higher altitude is less stable than one at a lower altitude due to the rigors of environment. Our data indicate in general

that diversity and equitability of the studied populations were contrary to the postulated trend of stability.

The index of Shannon and Wiener systematically underestimates the community diversity, particularly in small sample sizes (Pielou 1966, Alatalo and Alatalo 1977). If diversity indices based on different sample sizes are to be compared, this bias should be eliminated, and we can use the methods of Basharin (1960), Pielou (1966) or Stahel (personal note) for correction. In Table 1 the index of diversity is corrected by the method of Stahel (personal note).

Another proposal for standardization in order to equalize sample sizes is the rarefaction methodology of Sanders (1968), which allows one to interpolate the expected number of species in subsamples of 50, 100, 200, 300 etc. specimens and to estimate the diversity index that would be expected for such subsamples. For simplification the collections obtained at the same altitude were pooled, and estimates of the diversity index for different sample sizes were calculated. The results for the four altitudes are shown in Fig. 1. The shape of the four curves indicates that in subsamples of about 50 specimens the diversity index is only a little lower than that of the whole sample, and the indices for the altitudes 1900 m (O), 1400 m (M), and 900 m (U) are quite different. If we were interested only in this diversity index, the samples could be very small, and a larger, more expensive collection would hardly improve these estimates.

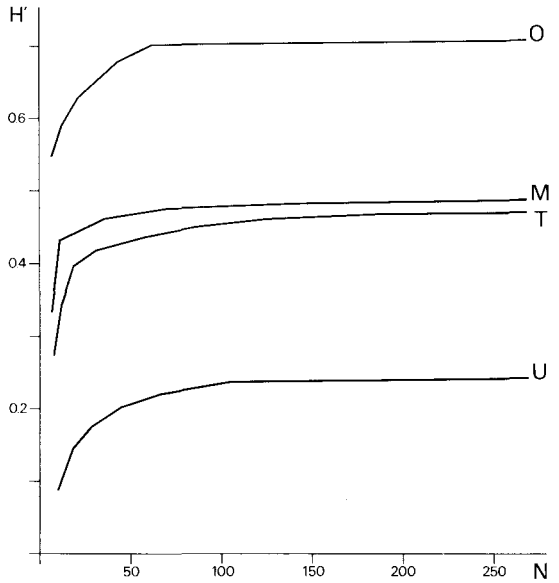

Fig. 1. Relationship of the species diversity H' to the rarefaction data. Collections of equal altitude are pooled. Letters correspond to the altitudes (Table 1). N = rarefied sample size.

IV. Niche analysis

Starting from the hypervolume niche model of Hutchinson (1958) several parameters have been proposed for the analysis of niche relationships: niche breadth, niche overlap, niche dimension, e.g. by Levins (1968) and Pielou (1972). Some of these parameters have also been estimated for *Drosophila* collections by Martinez et al. (1965), Levins (1968), Shorrocks (1974, 1975), Heed et al. (1976), Atkinson and Shorrocks (1977), Bächli (1977) and others.

We can distinguish between average measures, giving an ecological synopsis, and separate measures for each species, allowing comparison between species. "Niche dimensions" or "niche components" belong to the former. The total diversity index can be divided up into components associated with distinct dimensions (Levins 1968, Alatalo and Alatalo 1977). The samples in Table 1 can be grouped according to altitude, and the contribution of the dimension "altitude" to the total diversity can be estimated using the methods of Allan (1975), but in the antilogarithmic form (Alatalo 1978). The component "4 different altitudes" has an influence of 23.9 %, the component "20 different collection sites" has an influence of 28.6 % to the total diversity. So the hierarchical dimension "sites within altitude" has an influence of 4.7 %. In other words, faunistic differences are high between the 4 altitudes, but low between the sites of a given altitude.

The estimation of the average niche breadth for all species seems to me to be of little value. It is perhaps better to evaluate the relative niche breadth for every species separately using the formula for B' given by Levins (1968).

This measure is strongly influenced by the number of specimens of each species in question. Generally it seems better to exclude all data about rare species, because some methods react very sensitively to zeros or because the statement "not found" may be misleading. For this reason the following analyses are made for the common species only.

The relative niche breadth was found to be high (0.68 to 0.74) for the four species of the *obscura* group, and low (0.23 to 0.40) for the seven other species. This corresponds well to a first look at Table 1 and shows that the four species of the *obscura* group are faunistic generalists compared with the other species.

The niche overlap between every pair of species can be estimated by distance coefficients (Martinez et al. 1965, Shorrocks 1974) or by the percentage similarity (Renkonen 1938, Colwell and Futuyma 1971, Hurlbert 1971). This index gives values between 0 (when two species never occur in the same sample) and 1 (when two species have the same proportional distribution among the samples). The niche overlap values can be used as similarities for a cluster analysis. The dendrogram for the common species is shown in Fig. 2. The first cluster contains species found more frequently in lower altitudes *(Drosophila kuntzei, D. testacea, D. phalerata, D. immigrans),* in combination with *D. transversa* and *Scaptomyza pallida.* The abundant species *D. subobscura, D. obscura, D. alpina* and *D. ambigua,* present at nearly all collection sites, join in the second cluster. *D. nigrosparsa,* occuring only in higher altitudes, is somewhat apart.

Interpreting the clusters with regard to diet, we see a cluster of fungus feeders *(D. kuntzei, D. testacea, D. phalerata, D. transversa)* and a cluster of fruit feeders (some *obscura* group species).

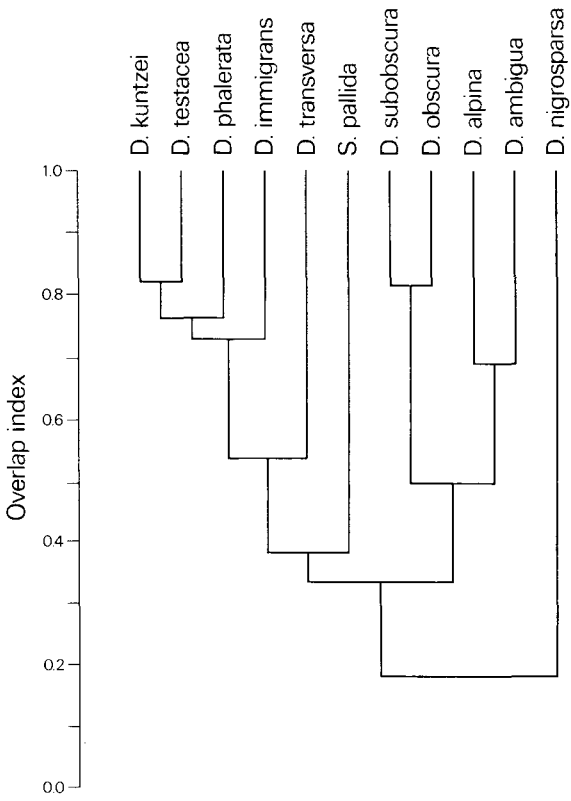

Fig. 2. Dendrogram showing the common species according to their niche overlap. Clustering by UPGMA.

V. Cluster analysis of the samples

Collection samples can be compared by evaluating the similarity of their species composition. Two samples are similar if about the same species occur with about the same frequencies in both. We proceeded as follows: After some transformations to be discussed below, we standardized the frequencies of the species for each sample to unit variance. Similarity was measured by the distance coefficient proposed by Sokal and Sneath (1963), excluding zero frequencies as "missing data".

The standardization of the raw data as well as many other statistical procedures are designed for normally distributed variables. Numbers of individuals in a faunistic collection rarely follow a normal distribution. The following analyses were performed on raw as well as square root, arctan, and log transformed data. Tested by the calculation of skewness and kurtosis no one of these transformations was found good enough for all species. Heuristically I found that results of the analysis on square root transformed data corresponded best to my subjective impressions. This coincides with the

statistical justification of the root transformation for Poisson and negative binomial variables. In any case statistical statements made for transformed data may not be valid for raw data.

The results of the cluster analysis are shown in Fig. 3. We expect that samples taken at the same altitude are more similar than those at different altitudes. This is true for the samples 1 to 5 and 6 to 10, also to some extent for the samples 12 to 15, whereas the sample 11 is somewhat separated by the samples 17 and 16. The samples 16 to 20 are widely split, perhaps because of some vegetational differences in their collection area.

VI. Principal component analysis

In terms of multivariate statistics our 20 collection sites can be represented as points in a 11-dimensional hyperspace, whose axes are abundances of species. Principal component analysis projects the hyperspace into a plane in such a way that as much of the total variance as possible is retained.

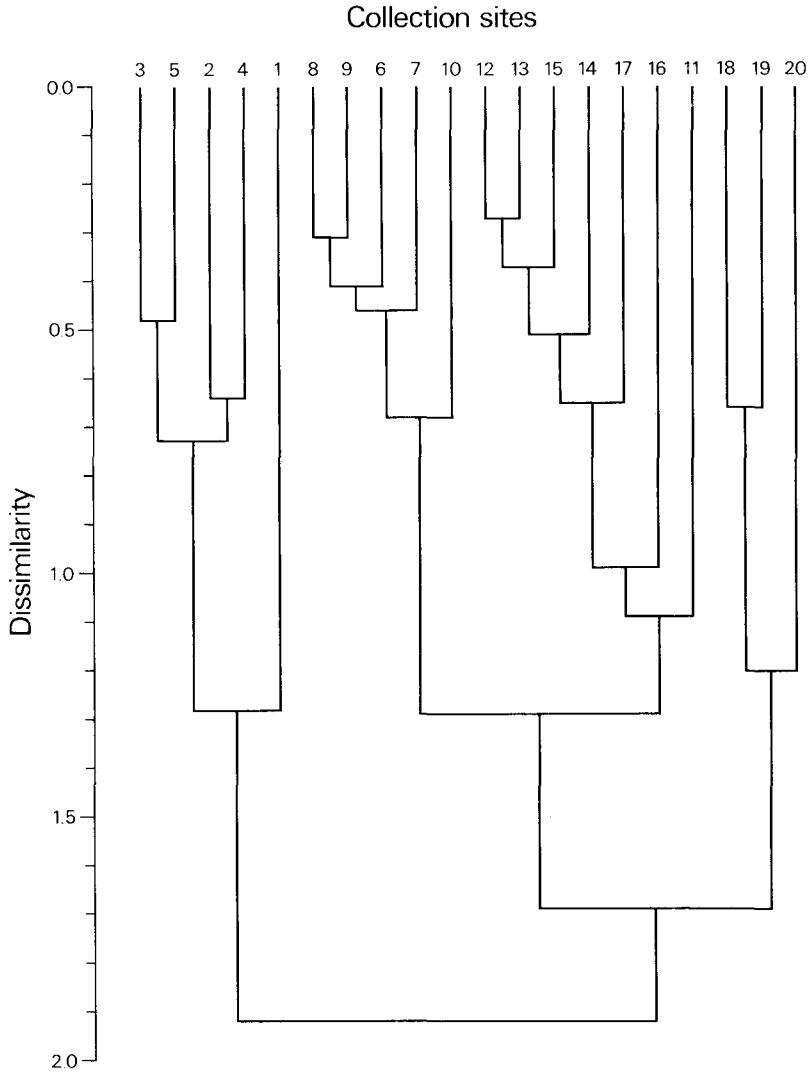

Fig. 3. Dendrogram showing the faunistic relations of the collection sites. Dissimilarity based on the common species, with square root transformation, standardisation, Sokal's distance and UPGMA clustering.

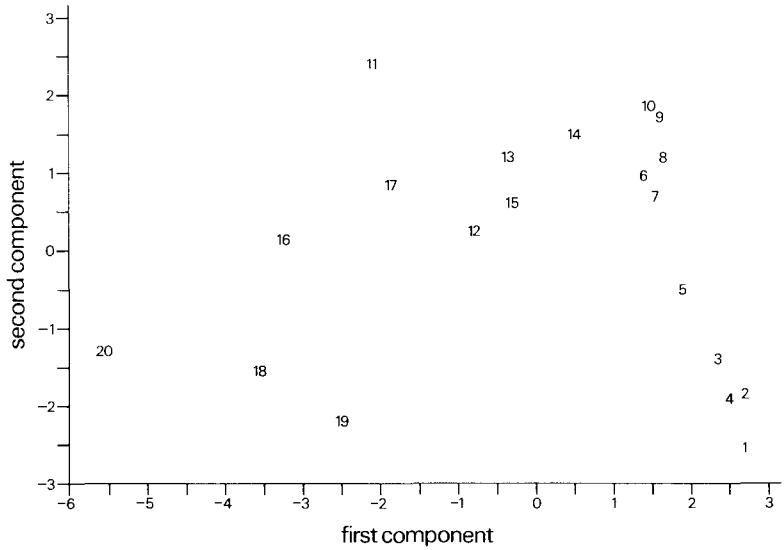

Fig. 4. Principal component analysis based on the common species with square root transformation. Plot of first and second components. Cumulative proportion of total variance = 0.75. Numbers correspond to collection sites.

Using the BMDO1M computer program (Dixon 1971) we got the coordinates for the first two components, containing three quarters of the total variance. The two-dimensional representation of the 20 collection sites is shown in Fig. 4. The samples 1 to 5 and 6 to 10 form distinct and compact clusters, while the samples 11 to 15 and 16 to 20 are somewhat spread out and confluent. Compared with the dendrogram (Fig. 3), the principal component plot represents the relationships between samples in more detail. While cluster analysis expelled samples 11, 16 and 17 from their respective groups, the principal component plot shows that they are nevertheless not too far apart. So, in this case we would prefer principal component to cluster analysis. If the percentage of explained variance in principal component analysis would be lower, cluster analysis could be superior.

VII. Discriminant analysis

If faunistic samples are taken in discrete groups, or if such groups are postulated or supposed either by cluster analysis or by principal component analysis, the degree of separation between these groups may be of interest. It can be measured by the percentage of wrongly classified samples. Further, if we want to know which species, alone or in combination with others, are best in discriminating the given groups, it is preferable to make a stepwise discriminant analysis. At each step the best discriminating variable is entered into the set of discriminating variables.

We checked the separation between the predefined altitudinal groups using the program BMDO7M (Dixon 1971). The most separating species were *D. subobscura, D. alpina, D. obscura* and *D. ambigua*. Fig. 5 gives the representation of the samples in the plane of the first two canonical variables. It shows that there is no misclassified sample. This excellent result is achieved already by the two species *D. subobscura* and *D. alpina*. We conclude that the principal component representation was not the best one for the redetection of the natural clusters in the data. Normally, however, such a predefined grouping is not available.

VIII. Conclusions

Statistical methods usually require numerous and correctly taken samples. This should be

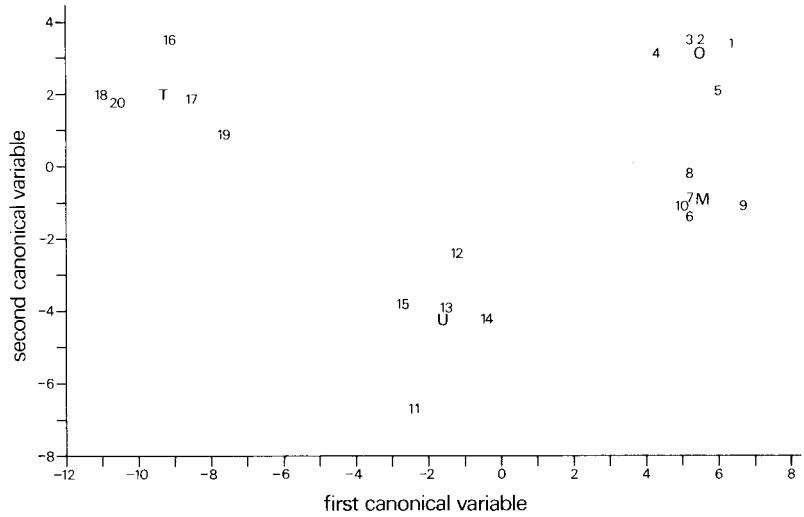

Fig. 5. Stepwise discriminant analysis based on the common species with square root transformation. Plot of first and second canonical variables. Numbers correspond to collection sites, letters to the altitudinal group means. Included species: *D. subobscura, D. alpina, D. obscura* and *D. ambigua*.

considered when planning the collections. On the premise that good samples are obtained, one might expect that such collection data contain a lot of information, which should be further analysed. Some of the methods presented qualify for a synopsis of the distributions of the species collected. Other analyses are suitable for the extraction of special parameters for single species or comparisons of species pairs. The methods may give hints for faunistic and ecological conclusions, and it seems worth the effort to apply them tentatively.

Acknowledgements. — The collecting was planned and executed in cooperation with Professor Dr. Hans Burla and Dr. Hans Jungen, who used part of the material for karyotypic and electrophoretic analyses, respectively. H. Vicentini and K. Rust helped with collecting the flies. I am also indebted to Prof. Burla and W. Stahel for making valuable suggestions on the manuscript.

Literature

Alatalo, R. 1978: Resource partitioning in Finnish woodpeckers. — Ornis Fenn. 55: 49—59.

Alatalo, R. & Alatalo, R. 1977: Components of diversity: Multivariate analysis with interaction. Ecology 58: 900—906.

Allan, J. D. 1975: Components of diversity. Oecologia 18: 359—367.

Atkinson, W. & Shorrocks, B. 1977: Breeding site specificity in the domestic species of Drosophila. — Oecologia 29: 223—232.

Basharin, G. P. 1960: On a statistical estimate for the entropy of a sequence of independent random variables. Theory Probab. Applic. 4: 333—336.

Bächli, G. 1977: Über Drosophiliden (Diptera) an der oberen Waldgrenze in der Schweiz. — Mitt. Schweiz. Ent. Ges. 50: 47—55.

Colwell, R. K. & Futuyma, D. J. 1971: On the measurement of niche breadth and overlap. — Ecology 52: 567—576.

Dixon, W. J. 1971: BMD. Biomedical computer programs. — 600 pp. Berkely.

Heed, W. B., Starmer, W. T., Miranda, M., Miller, M. W. & Phaff, H. J. 1976: An analysis of the yeast flora associated with cactiphilic Drosophila and their host plants in the Sonoran desert and its relation to temperate and tropical associations. — Ecology 57: 151—160.

Hurlbert, S. H. 1971: The nonconcept of species diversity: A critique and alternative parameters. — Ecology 52: 577—586.

Hutchinson, G. E. 1958: Concluding remarks. — Cold Spring Harbor Symp. Quant. Biol. 22: 415—427.

Levins, R. 1968: Evolution in changing environments. — 120 pp. Princeton.

Lloyd, M. & Ghelardi, R. J. 1964: A table for calculating the "equitability" component of species diversity. — J. Anim. Ecol. 33: 217—225.

MacArthur, R. H. 1955: Fluctuations of animal populations, and a measure of community stability. — Ecology 36: 533—536.

Martinez Pico, M., Maldonaldo, C. & Levins, R. 1965: Ecology and genetics of Puerto Rican Drosophila: I. Food preferences of sympatric species. — Carib. J. Sci. 5: 29—37.

Pielou, E. C. 1966: The measurement of diversity in different types of biological collections. — J. Theoret. Biol. 13: 131—144.

Pielou, E. C. 1972: Niche width and niche overlap: A method for measuring them. — Ecology 53: 687—692.

Renkonen, O. 1938: Statistisch-ökologische Untersuchungen über die terrestrische Käferwelt der finnischen Bruchmoore. — Ann. Zool. Soc. Zool. Bot. Fenn. Vanamo 6: 1—231.

Sanders, H. L. 1968: Marine benthic diversity: A comparative study. — Amer. Nat. 102: 243—282.

Shorrocks, B. 1974: Niche parameters in domestic species of Drosophila. — J. Nat. Hist. 8: 215—222.

Shorrocks, B. 1975: The distribution and abundance of woodland species of British Drosophila (Diptera: Drosophilidae). — J. Anim. Ecol. 44: 851—864.

Sokal, R. R. & Sneath, P. H. A. 1963: Principles of numerical taxonomy. — 359 pp. San Francisco.