Multivariate Morphometrics and Cytotaxonomy of the West African *Simulium damnosum* Complex (Diptera: Simuliidae).

Thesis submitted in accordance with the
requirements of the University of Liverpool
for the degree of Doctor in Philosophy.

By

David Peter
Surtees.

August 1988

D.P. Surtees

Multivariate Morphometrics and Cytotaxonomy of the West African *Simulium damnosum* Complex (Diptera: Simuliidae).

ABSTRACT

The purpose of this project was to examine variation within the West African *Simulium damnosum* species complex using classical larval polytene chromosome analysis, multivariate statistical analysis of larval polytene chromosome variation, and multivariate statistical analysis of adult female morphological variation.

Classical polytene chromosome analysis, undertaken to provide correlated identification for adult females reared from pupae in the same samples, revealed a chromosomally distinct form within *S. sanctipauli* from Togo. This form was distinguished from typical *S. sanctipauli* by the strong Y-chromosome linkage of an inversion, 1S-21, found autosomally at low frequency in typical *S. sanctipauli*.

Multivariate statistical analysis was applied to available data of larval polytene chromosome inversion frequencies from the *S. sanctipauli* subcomplex, the first application of these methods to polytene chromosome variation. Intra- and Inter-specific variation was found to be more complex than had been found using classical methods of analysis. The most distinctive and internally homogeneous taxon was *S. soubrense* 'B', while *S. sanctipauli* was found to be heterogeneous, with OP-insecticide resistant flies represented as a distinct cluster. *Simulium soubrense* was found to be an heterogeneous assemblage of taxa. The taxa *S. soubrense* 'Chutes Milo' and *S. soubrense* 'Beffa' showed affinities for *S. sanctipauli* , while *S. soubrense* 'Menankaya/Konkoure' showed complex variation which may have been a combination of clinal and local differentiation.

Multivariate morphometric methods were applied to 28 characters of adult females of the West African *S. damnosum* complex. Some of these characters were chosen for their known taxonomic importance, and some as additional characters representing the general morphology.

Statistical methods were undertaken to ensure the integrity of the basic data base, including univariate and multivariate outlier detection procedures.

Multivariate morphometric intraspecific variation was analysed in the six main species of the *S. damnosum* complex, and shape differences correlated with chromosomal differences, seasonal size variation, and both size and shape variation with no clear cause were found in the different species. The predominant mode of variation was found to be size variation, with very strong seasonal size variation for *S. squamosum*. *Simulium soubrense* showed the most extensive morphological variation, although this did not exactly parallel the chromosomal differentiation in the group.

Multivariate interspecific variation was analysed from the perspective of allocatory discriminant analysis, and optimal subsets of characters derived for overall and species-pair analyses from regional and 'global' material.

A method of adjusting prior probabilities of species membership was derived to exploit the taxonomic potential of two non-normal characters, and two kinds of allocation, forced and typicality probability were used to identify flies. The typicality probability method was chosen because it gave approximate confidence intervals to a fly's probability of species membership without reference to the other species. This was the first application of this method to insect morphometrics.

Significant interspecific morphometric variation was found, with most successful identification being of the epidemiologically important species *S. damnosum s.s.*, and *S. sirbanum*, and for *S. squamosum*. *Simulium soubrense*, *S. sanctipauli* and *S. yahense* were similar morphologically, although a colour character could identify *S. yahense* with 96% accuracy.

ACKNOWLEDGEMENTS

I would like to thank Professor W.W. Macdonald for providing facilities within the Department of Medical Entomology for this research, and Drs. R.J. Post and J.B. Davies for their careful supervision and interest.

I thank everyone who contributed material and advice, in particular Dr. R.A. Cheke of the Tropical Diseases Research Institute for providing valuable larval and adult West African material, Dr R. Baker of the World Health Organisation's Onchocerciasis Control Programme for providing much important adult material from Guinea, and Dr. R.P. Lane of the London School of Hygiene and Tropical Medicine for advise on morphometrics. I would like to thank Dr J. McMahon, the former Director, and staff of the MRC laboratory, Bo, Sierra Leone for providing facilities and accommodation during my field study in 1986.

Special thanks to my family and Helen Thomas for their support and encouragement throughout this project.

This research was funded by a Medical Research Council quota studentship.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION

1.1 BACKGROUND

The Simuliidae is a widespread family of Nematoceran Diptera, commonly known as blackflies containing about 2000 species-level taxa (Crosskey 1987a). The larvae are aquatic, most requiring running water for filter feeding. In the tropics, larval development can be completed in a few days or weeks, after which the larva enters the pharate or pre-pupal stage, followed by the pupal stage which is spent in a sheltered cocoon spun by the pharate pupa. In the tropics eclosion of the adult often follows within a week. Male blackflies do not bite, but females of most species require a blood meal for egg development. Host species are warm blooded vertebrates, predominantly birds.

A variety of pathogens are transmitted by the females, but the most important human health problem is due to the filarial parasite *Onchocerca volvulus* Leukart, which is transmitted solely by blackflies, and which causes the debilitating disease onchocerciasis. Nearly 18 million people are infected with *O. volvulus* (WHO Technical Report 752, 1987) throughout Tropical Africa, Yemen and parts of Central and South America although this is certainly an underestimate of the true number. By far the major proportion of these people live in sub-saharan West Africa, where members of the *Simulium damnosum* complex are the sole known vectors. *Simulium bovis* De Meillon has been shown to be anthropophilic in northern Nigeria and to support filariae similar to *Onchocerca volvulus*. Its role as a vector of human onchocerciasis is uncertain, but it is unlikely to be of importance (Crosskey 1957).

In response to this enormous disease problem, the World Health Organisation (WHO) began the Onchocerciasis Control Programme (OCP) in 1974 (Walsh *et al.* 1987), which by 1989 will cover an area of 1 million km^2 in West Africa. The original aim of OCP was that in the controlled areas annual biting rates (ABR=Σ(Number of flies caught×Number of days in the month) ÷Number of catching days in the month, where summation is over the twelve monthly estimates of the monthly biting rate MBR) should be less than 1000, and annual transmission potential (ATP=Σ(MBR×Total no. of L_3 larvae of *Onchocerca volvulus* found in the head)÷No. of flies dissected, where summation is over the twelve monthly MTPs) less than 100 for two consecutive years. Up to the present, control of the disease has been effected by larviciding the rivers in which larvae of the *S. damnosum* complex live, with the insecticides temephos, chlorphoxim, permethrin, carbosulfan and *Bacillus thuringiensis* H-14. Future control will include the use of Ivermectin, a microfilaricidal drug, given on a large scale to infected human populations.

The entomological control programme has been successful in the central, well controlled area of the OCP (85% of the original OCP area), in that ABR is usually zero and rarely exceeds 100, whilst ATP is zero throughout the area. However in areas prone to reinvasion of flies by long range migration from uncontrolled areas, in Togo, Benin and Mali, ATPs of up to 1000 have been recorded.

The epidemiological impact of control reflects the success of the entomological control. The community microfilarial load (the geometric mean microfilarial load per skin snip for a cohort of adults) has decreased linearly by 70% in the central area, but has settled at a higher level than this in areas subject to reinvasion. Ocular onchocerciasis has decreased dramatically in the central area (as measured by the community microfilarial load in the anterior chamber of the eye), though less dramatically in areas subject to reinvasion (WHO technical report 752, 1987).

1.2 THE SIMULIUM DAMNOSUM COMPLEX

Simulium damnosum Theobald (1903) was first implicated as the vector of human onchocerciasis by Blacklock (1926) in Sierra Leone. For the next forty years the species was regarded as a single species. However, over a period of time it became clear that *S. damnosum s.l.* was not uniform throughout its range. Gibbins (1933) noted that adult male thoracic markings varied geographically in East Africa. Grenier and Ovazza (1951) found variation in the fronto-clypeus and in the tubercles of larval *S. damnosum* and Crosskey (1960) also found morphological variation in larval *S. damnosum*. Crisp (1956) measured variation in the size of adult female *S. damnosum* in Northern Ghana and Grenier *et al.* (1960) also found size variation in *S. damnosum*. Lewis (1960) described variation in wing length of adult female *S. damnosum* from Liberia and Cameroon, and also reported variation in proportion of nulliparous flies, biting cycle, midday lull in activity, flight range, egg production, number of flies with retained eggs and degree of man-fly contact in savanna and forest zones. He considered these differences were important enough to influence onchocerciasis transmission. Marr and Lewis (1963, 1964) found variation in the colour of the antennae of adult females from Ghana.

MacCrae (1965, 1967) was the first to consider variation from a taxonomic viewpoint. He examined anthropophilic and non-anthropophilic populations of *S. damnosum s.l.* in Uganda and considered that wing length might be correlated with anthropophilic behaviour. Lewis and Duke (1966) examined morphological and behavioural parameters of *S. damnosum s.l.*. They considered that colour variation was clinal in West Africa, with darker flies inhabiting the forest and lighter flies inhabiting the savanna. They also examined variation in the tuft of hairs on the fore tarsus, the shape of the fore basitarsus, wing length, the time of biting and the body region of biting. Overall, they thought that the differences between forest and savanna flies was partly due to factors affecting individual variation and partly due to clinal variation.

This variation led Dunbar (1966) to initiate cytotaxonomic studies of East African *S. damnosum s.l.*, using larval silk gland polytene chromosomes, following extensive research on Canadian Simuliidae (Rothfels 1956,1987, Dunbar 1959). Polytene chromosomes are a special form of polyploid nucleus in which the individual chromosomes are arranged in a highly ordered way with respect to each other (see Ashburner 1979). Because of this ordered arrangement, a consistent pattern of dark bands and light interbands can be read, which is basically constant within species. Structural rearrangements of the chromosome, most commonly paracentric inversions, can be fixed between species. In East Africa Dunbar (1966) recognised four sibling species within the *S. damnosum* complex. A further five were added by Dunbar (1969), with the complex being divided into two subgroups, 'Nile' and 'Sanje'. Dunbar and Vajime (1971, 1972) further divided the complex into 17 species from East and West Africa. In West Africa, Vajime and Dunbar (1975) described eight cytospecies within the *S. damnosum* complex, *S. damnosum s.s.*, *S. sirbanum*, *S. sudanense*, *S. diegeurense*, *S. sanctipauli*, *S. soubrense*, *S. squamosum*, *S. yahense*. The validity of certain of the West African taxa has been questioned (Quillévéré 1975, Quillévéré and Pendriez 1975), but the most recent comprehensive summary (Dunbar and Vajime 1981) reported 26 siblings within the *S. damnosum* complex throughout East and West Africa. Since 1981 the *S. sanctipauli* subcomplex has been revised (Post 1986) with the addition of a new cytospecies, *S. soubrense* 'B' from Sierra Leone and Guinea and new chromosomal forms continue to be added to the *S. damnosum* complex (e.g Meredith *et al.* 1983, Surtees *et al.* 1988).

The following is a list of the currently recognised taxa within the West African *S. damnosum* complex (Partly based on Crosskey 1987b,

S. damnosum s.s	savanna, vector
S. sirbanum	savanna, vector
S. sudanense	savanna, vector, taxonomic status uncertain (Vajime 1984)
S. dieguerense	savanna, formerly thought rare, but now considered more widespread (Boakye and Mosha 1988), vector ?
S. sanctipauli	forest, vector
S. sanctipauli 'Djodji'	forest, but may also extend range into savanna
S. soubrense	forest, vector
S. soubrense 'Chutes Milo'	forest, vector
S. soubrense 'Konkoure'	forest, non-vector?
S. soubrense 'Beffa'	forest/savanna, vector
S. soubrense 'Menankaya'	forest, vector
S. soubrense 'B'	forest, vector
S. squamosum	forest/savanna, vector
S. yahense	forest, vector

This list is probably an underestimate of the true number of West African sibling species within *S. damnosum s.l.*, as variation has been noted in some of these taxa, but not fully investigated (Post pers. comm.).

These taxa undoubtedly differ in their importance as vectors of onchocerciasis (Quillévéré 1979) although their exact relative importance has not been fully established because of the problem of identifying adult females of the *S. damnosum* complex, and because of the similar problem of distinguishing between *O. volvulus* and other species of animal *Onchocerca*. However, the basic OCP operational distinction regarding vectorial importance is between savanna dwelling and forest dwelling species (WHO technical report 597, 1976). The savanna dwelling flies (most commonly *S. damnosum s.s.* and *S. sirbanum*) are the most dangerous vectors of onchocerciasis, and control measures have been aimed specifically at controlling these two species, and extension of control has occurred in response to reinvasion of these two species (adult females of these species can migrate distances greater than 500 km, Garms and Walsh 1987) from outside the control areas (Garms *et al.* 1979, Walsh *et al.* 1987).

While the control of onchocerciasis in West Africa is based on the epidemiological and pathological differences between savanna and forest forms of the disease, the identification of adult female savanna flies is not the only important distinction. More complex discrimination between vector species arises in the context of, for example, the identification of insecticide resistant flies (Post and Kurtak 1987), the identification of reinvading flies (e.g. Cheke and Garms 1983), and more detailed local studies of disease transmission (e.g. Garms 1983, in Liberia).

1.3 ADULT IDENTIFICATION METHODS

The major practical motivation for recognising cytotaxa within the *S. damnosum* complex is that they can differ in their capacity to transmit human onchocerciasis (WHO technical report 597, 1977). Cytotaxonomy has mostly been based on larval studies (because only larvae have suitable polytene chromosomes) and a fundamental difficulty in onchocerciasis research remains the inability to distinguish accurately the adult females of most cytospecies within the complex (Phillipon 1987).

Six methods have so far been attempted to identify cytospecies of the *S. damnosum* complex as adult females.

1. Adult Polytene Chromosomes

Procunier (Procunier and Post 1986), building on a technique developed by Bedo (1976) successfully identified adult females of *S. sanctipauli* and *S. soubrense* 'B' caught biting on man. The method uses the same cytotaxonomic criteria (polytene chromosome banding patterns) as was originally used to describe the sibling species within *S. damnosum s.l.*, and is therefore potentially as accurate. The adult polytene chromosomes were taken from the Malpighian tubules. Unfortunately only a low rate of identifiable chromosomes (≃8%) could be obtained, and the females needed to be blood-fed (with the consequent ethical problem of feeding potentially infective flies on human volunteers). Wirtz and Raybould (1986) show that artificial blood feeding systems may be practicable for *S. damnosum s.l.* removing one obstacle to the use of the technique, however, unless technical advances can improve the rate of identifiable chromosomes obtained it is unlikely that the method will be used for routine identification of adult females.

2. Laboratory Reared Larval Progeny.

Raybould *et al.* (1979) reared larvae from wild caught blood fed females, which were then induced to lay eggs in the laboratory. The larvae were then reared in artificial rearing apparatus and identified chromosomally using the chromosome standards of Vajime and Dunbar (1975). The method is therefore as accurate as the chromosomal criteria for identifying flies, but is laborious since it involves the separate rearing of single egg batches. Raybould *et al.* (1979) is significant for showing that all six of the major West African cytospecies were capable of being naturally infective with L_3s indistinguishable from *O. volvulus*.

3. Enzyme Electrophoresis.

Meredith and Townson (1981) performed an electrophoretic survey of 44 enzyme systems from six species within the West African *S. damnosum* complex from 25 sites in three countries, Mali, Côte d'Ivoire and Ghana. They found that two enzyme systems, phosphoglucomutase (PGM) and trehalase had allozymes which were diagnostic for two species, *S. squamosum* and *S. yahense*. The two species can be distinguished from other members of the *S. damnosum* complex by trehalase A with 98.8% accuracy, and *S. yahense* can be distinguished from *S. squamosum* using PGM B_1 with 99.8% accuracy. These enzymes were used to identify flies caught at human bait as *S. yahense* and *S. squamosum* (Meredith 1982). Garms and Zillman (1984) used the enzyme systems in the field in Liberia to identify *S. yahense* and *S. sanctipauli*, and found that of the the results which were unequivocal, 99.3% were identified as either *S. sanctipauli* or *S. yahense*, with 0.7% being designated as 'hybrids' (i.e. heterozygotes). They compared these results with a new morphological character found in *S. yahense* (the colour of the setae on the ninth abdominal tergite, see below) to evaluate the taxonomic use of this character. Thomson *et al.* (1988) also compared the results of enzyme electrophoresis with those using morphological characters, and found that the enzyme systems successfully identified *S. yahense* and *S. squamosum*.

Townson *et al.* (1987) recorded geographic variation in allozyme frequencies between *S. squamosum* from Côte d'Ivoire and from Togo. In East Africa Mebrahtu *et al.* (1986) found significant allozymic variation in *S. damnosum s.l.* populations, but the chromosomal identity of these populations was not established.

4. DNA Probes.

The potential importance of using DNA sequences as taxonomic characters is that the genome is being sampled directly, so avoiding any possibility of environmentally mediated variation. Post (1985 and Townson *et al.* 1987) screened about 2000 random sequences from genomic libraries constructed from *S. soubrense* 'B' and *S. squamosum*, and found three DNA sequences which could separate the *S. damnosum* complex into three parts: *S. squamosum*/*S. yahense*, *S. soubrense* 'B', and *S. sirbanum* according to relative amounts of hybridisation to the three probes using the dot-blot technique. This is an improvement over the results from enzyme electrophoresis because the savanna vectors can be distinguished. However the method is still relatively new and so does not have the backup in practical and theoretical understanding from previous experience with other groups of organisms that enzyme electrophoresis enjoys. Also, the samples used for construction of the genomic libraries were collected from a limited geographic area (Sierra Leone), and it appears that there is intraspecific variation in the the probes which prevent their use east of Côte d'Ivoire (Post pers. comm.). Finally, the current use of radioactively labelled probes works against the method as a practical field technique.

5. Analysis of Cuticular Hydrocarbons.

Philips *et al.* (1985 also Townson *et al.* 1987) building on previous work on the *Anopheles gambiae* complex (Carlson and Service 1979,1980) and on the *S. damnosum* complex (Carlson and Walsh 1981) used gas liquid chromatography and gas chromatography/ mass spectrometry to analyse cuticular hydrocarbons of four species within the *S. damnosum* complex, *S. damnosum s.s.*, *S. sirbanum* (two samples), *S. sanctipauli*, and *S. yahense* from four countries in West Africa. Using multivariate analysis of the hydrocarbon profiles from the four species examined, they found that 94.6% of females were reallocated into their correct sample. This included two samples of *S. sirbanum*, so there was significant intraspecific variation in hydrocarbon profiles. It is not clear how many hydrocarbon peaks were used in their discriminant analysis, but if all 24 peaks numbered in figure one of Philips *et al.* (1985) were used then the total sample size of 131 females is too small. Lachenbruch and Goldstein (1979) suggest that the sample size in each group in a discriminant analysis should exceed three times the number of characters i.e. >72 if all 24 peaks were used, otherwise serious bias will be introduced into the model, making the results misleadingly optimistic. Two of the samples in Philips *et al.* (1985) were very small (*S. yahense*, 9, *S. damnosum s.s.*, 11). These samples are distant from the other three samples in their figure three, which may be due to sampling error.

The statistical problems with the results presented to date may be remedied once larger samples have been obtained, however the method uses expensive equipment requiring technical support, so that it is unlikely that it will be a practical field technique for some time to come.

6. Morphology.

Section 1.2 has described some of the studies which recorded morphological variation within the *S. damnosum* complex before its sibling status was known, from both East and West Africa, and in adult males, females and larvae. This section will review morphological studies which explicitly aim to distinguish adults of the West African *S. damnosum* complex, rather than those which recorded morphological variation before the recognition of the six main cytospecies (Vajime and Dunbar 1975).

Morphological methods remain the most practicable of all adult identification methods, because they are easy to use and are portable. However, morphological variation includes confounding factors such as environmentally mediated seasonal and/or geographic variation which can give misleading results unless care is taken to sample widely enough. Also, by definition, morphological differentiation between sibling species is not great, so that finding species diagnostic morphological characters is difficult.

a). Soponis and Peterson (1976, also Anon 1976) examined 55 morphological characters on 97 female flies collected from nine sites in Togo. They used univariate statistical methods, examining the sample distributions expressed as histograms for bi- or multimodality, with the flies divided into the three colour categories used by Lewis and Duke (1966). They found that nine characters were unimodal, six characters were bimodal, and 16 characters were multimodal. They found that fore and mid basitarsus length, fore femoral length, wing length, length of the fourth maxillary palp segment and the number of macrotrichia on the radial vein of the wing identified what they believed, based on correlated larval cytotaxonomic identification, to be *S. soubrense* and *S. sirbanum*. Peterson and Dang (1981) subsequently claimed these characters were not practicable for species identification.

b). Quillévéré *et al.* (1977) produced a key for the identification of females of the *S. damnosum* complex based on the length of the antenna, the relative compaction of antennal segments 4-8, and the number of maxillary teeth. This key was for the six main species of the *S. damnosum* complex, *S. damnosum s.s.*, *S. sirbanum*, *S. soubrense*, *S. sanctipauli*, *S. squamosum*, and *S. yahense*, although the latter pair could not be distinguished. They did not present formal tables of error rate, so it is not possible to evaluate the power of their key when applied to their own data. Quillévéré and Sechan (1978) examined 2468 wings from the same six West African species from five countries, and considered that the number of hairs on the radial vein of the wing could be used to distinguish *S. squamosum* from *S. yahense*, with a certain amount of overlap. This character was included as an extra section in their key.

Garms (1978) evaluated the morphological characters used in this study and confirmed that the length and shape of the antenna was taxonomically useful, but that the number of maxillary teeth and the number of hairs on the radial vein were not. Townson and Meredith (1979) also examined these characters and found that the two doubtful characters were not taxonomically useful because they were significantly correlated with overall size (*contra* Quillévéré *et al.* 1977, Quillévéré and Sechan 1978), which shows extensive and overlapping variation.

c). Garms (1978) examined seven morphological characters, wing tuft colour, length, shape and colour of the antennae, the number of maxillary teeth, the number of hairs on the radial vein of the wing and the length of the thorax in adult females of six species of the West African *S. damnosum* complex.

He found that the ratio of length of thorax to the length of antenna was a useful taxonomic character as well as wing tuft colour. These characters have been used extensively in subsequent work on the epidemiological significance of different members of the *S. damnosum* complex (e.g. Garms *et al.* 1982, Garms 1983, Cheke and Garms 1983, Garms and Cheke 1985, Cheke and Garms 1986, Cheke *et al.* 1987, Cheke and Denke 1988). The general findings of these papers has been that in the absence of *S. squamosum*, then the species pair *S. sanctipauli*/*S. soubrense* and *S. sirbanum*/*S. damnosum s.s.* could be distinguished using the thorax/antennal ratio and wing tuft colour, but *S. squamosum*/*S. yahense* overlaps with both species pairs when either is present. Garms and Zillman (1984) found a new morphological character (the colour of the setae on the ninth abdominal tergite) which was over 99% diagnostic for *S. yahense* when compared with *S.*

sanctipauli using gel electrophoresis. This character was also used by Thomson *et al.* (1987) who found 91.3% of *S. yahense* had dark abdominal setae.

To conclude, four characters emerged as being taxonomically useful, wing tuft colour, thorax length, antennal length (and shape and colour), and abdominal setal colour. However, the morphometric methods used were of the 'index' kind, rather than using multivariate statistical methods to combine these characters in an optimal way. If multivariate methods had been used, the rate of correct identification would undoubtedly have improved.

d). Dang and Peterson (1980) produced a pictorial key to six species of the West African *S. damnosum* complex, *S. damnosum s.s.*, *S. sirbanum*, *S. sanctipauli*, *S. soubrense*, *S. squamosum*, *S. yahense* from an unspecified number of countries. They used 12 characters to identify the adult females and eight to identify adult males.

The characters for identifying adult females were wing tuft colouration, length, shape and colour of the antennae, colour of the scales on the hind leg, colour of the setae on the hind trochanter, colour of the scales of the scutum and the scutellum, colour of the setae of the abdomen, colour of the setae on the vertex, colour of the setae of the postcranium and the colour of the setae on the fore coxa.

The characters for identifying the males were wing tuft colouration, colour of the haltere, colour of the scales of the lateral margin of the scutum, the colour of the setae on on the clypeus, the colour of the setae on the postcranium, the extent of the dark spot on the seventh abdominal tergite, and the scutal pattern.

No formal estimate of error rate was presented so it is not possible to evaluate how successful their key was in identifying their own data. Peterson and Dang (1981) extended this work, and showed the distribution of characters in the same six species. They described 18 characters which they considered to be taxonomically important for identifying females, and 13 characters to identify males. However, no estimate of error rate using these characters sets was given making it impossible to evaluate objectively their character sets.

Of the characters described by Dang and Peterson (1980), the male scutal patterns have been used (Meredith *et al.* 1983, Cheke *et al.* 1987), as has the colour of the postcranial hairs (e.g. Walsh *et al.* 1981) and the colour of the scutellar hairs (e.g. Garms 1983).

e). Meredith *et al.* (1983) examined variation in male scutal pattern in the *S. sanctipauli* subcomplex and found considerable variation in this character. Males from Togo and Benin (*S. soubrense* 'Beffa') were predominantly type four, while types one and two (see their figure five) were dominant in the west. They also examined the wing tuft colour character in both sexes, using the categories of Kurtak *et al.* (1981) and found that all five categories were found within *S. soubrense* 'Beffa'.

f). Cheke *et al.* (1987) examined males of *S. sirbanum* reared from pupae collected at 14 sites in four West African countries, evaluating the male scutal pattern described by Dang and Peterson (1980) as being taxonomically important in distinguishing *S. sirbanum* males from *S. damnosum s.s.* males. They considered two hypotheses to explain variation in the scutal patterns that they found between more northerly samples and more southerly samples, either that *S. sirbanum* is polymorphic or that the that the variation represents two different cytospecies. In support of the second possibility they cite Philips *et al.* (1985) who found significant intraspecific variation in cuticular hydrocarbons between northern and southern *S. sirbanum*. They conclude that the cytotaxonomic status of *S. sirbanum* needs to be further investigated.

g). Recently, Beech-Garwood *et al.* (in the press) have found that the presence of golden hairs on the mėsonotum of female *S. damnosum s.l.* in Sierra Leone is a good indicator of the presence of *S. squamosum* in areas where *S. squamosum* and *S. soubrense* or *S. soubrense* 'B' may be sympatric. The character is not diagnostic however, since a minority of *S. soubrense* and *S. soubrense* 'B' may also have these hairs. The observation has been confirmed by enzyme electrophoresis by Davies *et al.* (1988). They conclude that the character is useful at the population level, but needs to be supported by other evidence for single fly identification.

1.4 MORPHOLOGICAL DATA AND MULTIVARIATE STATISTICS

1.4.1 BASIC NOTATION

Multivariate problems are defined by Gnanadesikan (1977) as those concerned with the analysis of n points in p-space, i.e. each of the n objects (in this project, the number of flies) has associated with it a p-dimensional vector of responses (in this analysis, the 28 characters measured or scored on each fly).

The basic difference between this approach and the univariate approach is that the variation in the p-dimensional vector of characters is treated simultaneously, and any information contained in the association (correlation) between characters exploited.

Some of the basic notation and concepts which are needed for an understanding of multivariate statistics can be found in standard texts such as Seber (1984), Mardia, Kent and Bibby (1979) and Gnanadesikan (1977).

If X, is the matrix of n (number of observations) rows by p (number of characters) columns, then the mean vector can be calculated as,

$$\bar{x} = 1/n\Sigma\, x_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

where x_i is the p-dimensional vector of observations on the i-th fly, summation is over i=1,...,n.

The nxp matrix of centred observations, $\check{X}$ can be calculated,

$$[x_1-\bar{x}\ ,\ldots x_n-\bar{x}\ ]$$

and from this the p×p matrix of sums of squares and cross products (SSQPR) is given by,

$$Q = \check{X}'\tilde{X} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

The sample dispersion (variance/covariance) matrix is calculated as,

$$S = Q/(n-1) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3$$

The p×p matrix of Pearson product moment correlation coefficients is derived from the dispersion matrix by first calculating D, the diagonal matrix of variances (i.e. the principal diagonal of S). Then,

$$R = D^{0.5}SD^{0.5} \tag{4}$$

A basic statistic in multivariate analysis is Mahalanobis' distance (Mahalanobis 1936), which can be calculated as the distance between individuals in a sample, or as the distance between the individuals and the sample mean vector, or as the distance between mean vectors, e.g.,

$$D^2 = (x_i - \bar{x})'S^{-1}(x_i - \bar{x}) \tag{5}$$

which is the distance from the i-th individual to the mean vector. The importance of this distance lies in the use of S, the dispersion matrix to weight the distance, thus using the information contained in the variance of the characters and the correlation between them.

1.4.2 ASSUMPTIONS OF MULTIVARIATE ANALYSIS

The use of multivariate statistical methods includes accepting some assumptions about the data set which might not be met. It is usually assumed that the data are sampled from a multivariate normal population with mean $\bar{x}$ and dispersion matrix Σ (Seber 1984). This assumption is rarely exactly met with real data, but if the assumption is made, tested and approximately met then statistical hypotheses about the mean vector and covariance matrix can be performed. If the assumption is made and tested, then more informal data analytic multivariate methods can be used (Gnanadesikan 1977) whether the assumption is met or not. In particular, graphical and dimension reduction techniques can be applied to the data, methods which can also be used on data on which no assumptions are made.

In practise, real morphological data tend to be approximately normally (symmetrically) distributed (Campbell 1978), so the assumption of multivariate normality is often reasonable, although careful checks must be made on the characters, either jointly or separately to ensure that they do conform approximately to normality. Of course, characters which violate the assumption can be used, but not in a formal statistical hypothesis testing regime. The repertoire of nonparametric multivariate statistical methods is limited, so the use of the multivariate normal model is made partly because of the tendency for real data to approximate to it and partly because this is the best developed aspect of multivariate analysis.

In the structured data situation, it is usually assumed that the dispersion matrices of each group (samples, or species) are the same. This is assumed even though differences in covariance structure are one way in which organisms can differ in shape (Reyment 1962, Gould 1984). However organisms can differ in shape but share the same covariance structure (Campbell 1978), so unequal dispersion is not a necessary condition for shape differences to occur. The assumption can be tested (Chapter seven), although these tests are generally inadequate (Seber 1984).

1.4.3 AIMS AND METHODS OF MULTIVARIATE ANALYSIS

The analysis of a multivariate data set can be very complex (Gower and Digby 1981), and many methods have been developed to help in the interpretation of multivariate data (see for example, Gnanadesikan 1977, Gordon 1981, Seber 1984).

The particular multivariate statistical method(s) to be applied to a data set clearly depends on the original objectives of the specific project (Atchley and Bryant 1975). However, the strategy to be followed can be classified into broad regions which will to some extent dictate the methods of analysis which can be used.

First, the aim of a study might be to test a very specific hypothesis or set of hypotheses (such as testing the null hypothesis of no genetic component to size and shape in the rat, Atchley *et al.* 1982) in which case the appropriate methods include multivariate analysis of variance (MANOVA) and canonical correlation analysis. Or the aim might be to explore the data informally using data-analytic methods, perhaps revealing previously hidden patterns within the data, in which case principal components analysis (PCA), cluster analysis (Gordon 1981), non-metric multidimensional scaling (Kruskal 1977) and graphical methods (Gower and Digby 1981) are more appropriate.

Secondly, the data might be structured and the aim of the study is to explore any differences or similarities within and between data defined by these structures. This could take the form of formal hypothesis testing, using canonical variates analysis (CVA) or MANOVA, or it could be informal exploration, in which case PCA, cluster analysis etc. could be used. Or the data might be unstructured, in which case the aim of the study might be to explain and summarise the observed variation within the data, again either in a formal hypothesis testing regime (using for example canonical correlation analysis) or not (using PCA, cluster analysis, nonmetric MDS etc.).

Clearly these areas are very broad and overlepping, and a particular study is unlikely to adhere strictly to one or other of these strategies, for example, a study which aims to test a specific hypothesis will need to use informal data-analytic methods both to gain a better understanding of the data and to ensure that the assumptions of the model are not violated.

1.4.4 MULTIVARIATE ANALYSIS APPLIED TO MORPHOLOGICAL DATA

Multivariate statistics and biology have developed together (e.g. Weldon 1893, Pearson 1926, Fisher 1936, Rao 1948, Mahalanobis *et al.* 1949, Blackith and Reyment 1971, Sneath and Sokal 1973) because biological data are as a rule inherently multivariate. Multivariate methods do not have to be applied solely to morphological data (e.g. Nei 1987) but in practise this is probably the most commonly analysed form of biological data, at least using methods exploiting character correlations. If more general data are used (e.g. morphology, allozyme frequencies, immunological distance etc.) to infer evolutionary relationships between taxa then the discipline is generally known as numerical taxonomy (Sneath and Sokal 1973). Many of the methods used in numerical taxonomy are cluster analysis techniques applied to proximity matrices calculated between taxa, often not incorporating information about the correlation between the characters. The more restricted discipline, multivariate morphometrics (Blackith and Reyment 1971) in general uses statistics exploiting character correlations, but is applied only to quantifications of morphological characters. Multivariate morphometrics is not necessarily interested in evolutionary relationships in the sense of taxonomies. In practise, the distinction between the two disciplines is not great, (Blackith and Reyment 1971 suggest the term 'quantitative taxonomy' as a more general term, but much multivariate morphometrics is not taxonomic in the strict sense) and studies using both approaches are common (e.g. Lindensfelser 1984).

Insects have often been used in morphometric studies (Daley 1985), their abundance, evolutionary and ecological importance, and the large number of characters which can be measured make them ideal subjects.

Morphometric studies of the Diptera have been common, the following are examples from the literature on morphometrics of the Diptera, organised by family.

In the Culicidae, Rohlf (1963) examined the congruence between larval and adult *Aedes* classifications, using 48 species, 71 larval and 72 adult coded characters. He applied hierarchical cluster analyses to the proximity matrices calculated using the correlation similarity coefficient and a distance metric and compared the resultant dendrograms. He concluded that while there was congruence between the classifications using the two life stages, it was not good enough, and therefore a realistic taxonomy should take account of all life stages simultaneously. Ribiero (1980) examined 34 characters in the *Anopheles gambiae* complex using cluster analysis and ordination methods. Significant morphometric differentiation was found. Dahl *et al.* (1984) used pattern recognition applied to claw shape in mosquitoes, five species of *Aedes*, and one species each of *Anopheles* and *Culex*. Rohlf and Archie (1984) used fourier methods to characterise wing shape in 127 species of North American mosquitoes, which they considered a useful method for quantifying shape variation.

In the Chironomidae, Atchley and Martin (1971) examined sexual dimorphism in 17 larval head capsule characters from five species of *Chironomus*. They used discriminant analysis and canonical variates analysis to compare and contrast patterns of sexual dimorphism within the five species. They found a correlation between the degree of sexual dimorphism in a species and the amount of chromosomal polymorphism (using polytene chromosome analysis), and also that the nature of the dimorphism differed between species, as revealed by CVA ordination (for a similar but more sophisticated analysis of sexual dimorphism applied to primates see Oxnard 1984). Atchley (1971a) examined sexual dimorphism in five species of *Chironomus* using factor analysis. He found that in most species the dimorphism was along the size axis but in one species dimorphism was along a shape factor, which he explained by hypothesising ecological differences. Titmus and Badcock (1981) examined parasitic feminisation in a Chironomid, *Einfeldia dissidens* resulting from mermithid infection, using CVA, and found two axes of variation, one corresponding to sexual dimorphism and one corresponding to parasitisation. Thus parasitised males and females were closer to each other than were unparasitised and parasitised females, although all these were much closer to each other than any were to unparasitised males.

In the Ceratopogonidae, Atchley (1971b) examined geographic variation in 14 characters in the pupae of three *Culicoides* species, and found that the relative amounts and nature of the variation differed between the three species. Atchley (1971c) extended this work on the three same siblings of *Culicoides* using factor analysis and multiple regression. He found that the proportion of variation which could be accounted for by regression of climatic variables onto morphological characters varied according to species. He explained these differences in terms of Levins' (1965) adaptive models with some species responding to selection (poorly buffered) more than others (well buffered). Atchley (1973) used CVA and stepwise discriminant analysis to separate three species of the *Culicoides (Selfia)* group . Starting with 43 characters measured on 298 pupal and adult flies, he derived a subset of 7 characters for overall discrimination, and subsets of 8, 9 and 4 characters for the species-pair analyses. He found significant morphometric differentiation, but considered that the allocation rate using just adult characters was not good enough, and pupal characters were also needed, limiting the practical use of the method. Atchley (1974) examined 13 characters from 113 adult females of two species of Ceratopogonidae, *Leptoconops torrens* and *L. carteri* using CVA and nonmetric multidimensional scaling. Reducing the initial set of characters from 13 to 6 using stepwise discriminant analysis, he found complete separation between the two species. When further specimens were allocated using this 6 character set, 95% were allocated correctly. Hensleigh and Atchley (1977) examined variation in *Culicoides variipennis* (vector of blue-tongue virus in North America) in laboratory controlled conditions. Their aim was to investigate intraspecific variation which had resulted in the naming of 5 subspecies, using CVA, stepwise discriminant analysis, factor analysis, analysis of variance and multiple regression. By artificially rearing flies at different temperatures they were able to show that most of the variation found in natural populations was due to temperature variation, bringing into question the naming of subspecies. Lane (1981) examined wing spot patterns in the *Culicoides pulicaris* group, using two methods of coding the characters, one without taking account of morphogenetic parameters, the other taking these into account. Principal co-ordinates analysis and principal components analysis were used as ordination methods applied to the two methods. The method which took account of morphogenesis was considered superior in explaining observed variation within the group.

In the **Phlebotominae** , Lane and Ready (1985) examined six characters in *Lutzomyia wellcomei* and *Lu. complexus* and found significant morphometric differentiation, although there was considerable overlap between the two species.

In the Muscidae and Drososphilidae, Rohlf and Sokal (1972) examined 14 morphological characters in *Musca domestica* and *Drosophila melanogaster* using factor analysis of the correlation matrices. The five factors they extracted from the two species were considered by them to be homologous. Bryant and Turner (1978) examined variation in *Musca domestica* and *M. autumnalis* using the same characters as Rohlf and Sokal (1972). They used principal components analysis and a factor congruence method involving least squares comparison of principal components to examine geographic variation, and found that the patterns of variation were similar in the two species, but that the genetic component of variation in *M. autumnalis* was less than in the house-fly, a finding which they attributed to the recent introduction of the face-fly, resulting in a population bottleneck.

Brown (1979, also Brown and Shipp 1977, 1978) examined wing variation in the Calliphoridae and the Sarcophagidae using CVA and cluster analysis to compare and contrast the taxonomies derived using traditional taxonomic methods with those derived using numerical methods.

Apart from the Diptera, insect morphometrics has included a wide range of families, most notably in the Orthoptera (e.g. Roy and Mukherjee 1964, Blackith and Blackith 1969, Atchley and Hensleigh 1974, Campbell and Dearn 1980), in the Hemiptera and Homoptera (e.g. Sokal and Thomas 1967, Jeffers 1967, Davies and Boryatinski 1979, Bird *et al.* 1981, Simon 1983), in the Coleoptera (Lubischev 1962) and in the Hymenoptera (e.g. DuPraw 1965, Plowright and Stephen 1973), although this is by no means a comprehensive list (see Daley 1985).

Other invertebrates which have been analysed using multivariate statistical methods include bivalves (Ferson *et al.* 1985, Davis 1983), Foraminifers (Reyment 1982), horseshoe crabs (Riska 1981), Sea Urchins (Lessios 1981), land snails (Gould *et al.* 1975, Gould 1984) and prawns (Lindenfelser 1984).

Vertebrates have been extensively examined using multivariate morphometric methods, including Amphibia, (Reyment 1961), Reptilia (Jolicoeur and Mosimann 1960, Thorpe 1980), Birds (e.g. Schnell 1970, Johnston and Selander 1971, Rising 1970). Within the mammals, bats (e.g. Baker *et al.* 1972, Campbell and Kitchener 1980), rodents (Corbet *et al.* 1970, Thorpe and Leamy 1983, Atchley *et al* 1982), carnivores (Jolicoeur 1959) and primates (e.g. Mahalanobis *et al.* 1949, Ashton *et al.* 1965, Van Vark and Howells 1984).

1.4.5 COMPUTER PROGRAMS FOR MULTIVARIATE ANALYSIS

While it is possible to calculate some multivariate statistics without a computer, it is impossible to use the full range of statistical methods without the help of a powerful computer and well written software. Fortunately, high speed computers are widely available, and statistical packages have been written to run on these which provide most of the statistical procedures needed in a typical project.

In this project, the following statistical packages were used,

1. SAS (Statistical Analysis System, SAS Institute 1984, 1986, release 5.16) is a comprehensive system for data analysis, offering a very wide range of data management facilities, univariate and multivariate statistical procedures. Graphical procedures are provided by SAS/GRAPH (SAS Institute 1985, version 5), and matrix algebra is provided by SAS PROC MATRIX, providing a facility for developing new procedures or customising other procedures. A powerful macro

facility is provided, and the system can be run interactively using the display manager system.

2. SPSSX (Statistical Package for the Social Sciences, SPSS inc. 1985, ver. 2.2) allows the analysis of data using a wide range of univariate and multivariate statistical methods. It is generally not as flexible as SAS, but is simple to use.

3. GENSTAT (General Statistical Package, Lawes Agricultural Trust 1984, release 4.04B) provides a wide range of univariate and multivariate statistical methods, and is particularly good for the analysis of designed experiments. A macro library is provided, and the ability to write macros using matrix algebraic expressions makes it flexible. However, it is difficult to use and is poorly documented.

4. CLUSTAN (Cluster Analysis Package, Wishart 1978, release 2.1) is a specialised package offering a very wide range of cluster analysis methods, and some graphical procedures. The package is widely used across many disciplines.

5. NTSYS (Numerical Taxonomic System of Multivariate Statistical Programs, Rohlf 1985) is a specialised package allowing cluster analysis, and ordination of numerical taxonomic data. Limited matrix algebraic manipulation is allowed.

1.5 OBJECTIVES OF THE PROJECT

It is clear from the review of the literature in 1.3 that there is no simple set of morphological characters that can be used in traditional taxonomic keys to separate females of all the sibling species of the *S. damnosum* complex in West Africa. Morphometric methods based on multivariate statistical analyses have been used successfully in other groups of insects, so the basic techniques are widely available and well understood.

Therefore, the main objective of this project is to use multivariate statistical techniques to find combinations of morphological characters which can best identify adult females of the *S. damnosum* complex in West Africa.

The characters measured or scored on each adult female fly are described in Chapter four, and statistical methods used to screen the basic data set are described in Chapter five.

Chapter six introduces some of the multivariate statistical methods used for description of morphological variation, and applies these methods to intraspecific variation within cytospecies of the *S. damnosum* complex.

Chapter seven presents the statistics necessary for regional allocation of unknown adult female flies from two regions, Togo and Benin, and the area west of the Volta Lake, Ghana.

Chapter eight describes the statistics necessary for the allocation of unknown females without prior knowledge of the geographic origin of the fly, while the final chapter draws general conclusions about the method of identification, gives worked examples of the mathematics involved and suggests a protocol for the field identification of adult females based on the statistics presented in the previous two chapters. Details of the data set are presented as an appendix (Appendix one).

Prior to the multivariate morphometric analysis of adult female *S. damnosum s.l.*, it was necessary to obtain as much correlated larval cytotaxonomic identifications as possible. As a result of this work, a new cytotype within *S. sanctipauli* was found from Togo, which is presented as Chapter two. Chapter Three applies multivariate statistical methods to available data within the *S. sanctipauli* subcomplex to examine between and within species variation, the first such analysis of polytene chromosome variation.

CHAPTER TWO: THE CYTOTAXONOMY OF SIMULIUM SANCTIPAULI DJODJI FORM

2.1 INTRODUCTION

The importance of describing genetically distinct forms or geographic races within previously recognised cytospecies of the *S. damnosum* complex comes from the possible correlation of the different cytoforms with factors of significance in disease transmission, such as anthropophily (Cheke and Denke 1988), together with the use of new forms in tracing migration patterns or the distribution of insecticide resistance (Post and Kurtak 1987).

The purpose of this chapter is to describe a new cytotaxonomic form within *S. sanctipauli* from Ghana and Togo, which was discovered during routine cytotaxonomic identifications to provide correlated chromosomal identities for adults reared from pupae. The adults were used in the morphometric analyses described in Chapters six, seven, and eight.

2.2 MATERIALS AND METHODS

Breeding sites where the Djodji form of *S. sanctipauli* was collected are listed in Table 2.1. Larvae were fixed in 3:1 ethanol:acetic acid and stored in a refrigerator. For preparation of polytene chromosomes, the larvae were split open ventrally and hydrolysed in 5M Hydrochloric acid for one hour. The salivary glands were then separated from the larval body and stained in a drop of lacto-propionic orcein (Macgregor and Varley 1983) and mounted in 60% acetic acid. Photographs were taken of each chromosome arm, and the preparation made permanent by prising off the cover slip (after cooling in liquid nitrogen), immersing the slide in absolute ethanol for one minute then adding a drop of euparal onto the preparation and lowering a new cover slip onto the slide. The slide was then dried on a warm plate for some months, and stored. The larval body was washed in distilled water and put in a glass vial with Feulgen (Macgregor and Varley 1983) until the body had stained. The larval sex was determined using the shape of the developing gonads (Puri 1925). Inversions were scored by comparison with the standard maps of Post (1986), with *S. squamosum* as the reference sequence.

2.3 CHROMOSOMAL CHARACTERISTICS AND CYTOTAXONOMIC KEY

All fixed and polymorphic inversions within Djodji form are indicated on the idiogram (Figure 2.1), and frequencies of polymorphic inversions are listed as Table 2.3. The new form is fixed for the inversions 1L-P&Q, 2L-4&6&A and 3L-2, but there are no new fixed inversions unique to Djodji form, and only one new rare polymorphic inversion (1S-P, see Figure 2.2). The presence of inversion 2L-A places Djodji form in *S. sanctipauli* (Post 1986). However, 1S-21 (Figure 2.3) is strongly linked to the. Y-chromosome in Djodji form (Table 2.2), and this unique feature is the most important cytotaxonomic criterion for both description and routine identification. Since 1S-21 is Y-linked in Djodji form there is no single inversion which is diagnostic of all individuals. However, samples in which there is strong Y-linkage of the inversion can be unequivocally identified as *S. sanctipauli* 'Djodji', and mixed samples (should they exist) of the form with typical *S. sanctipauli*, will be recognised as such using population genetic analysis.

The following cytotaxonomic key can be used for the identification of *S. soubrense* 'Beffa', *S. sanctipauli* and the Djodji form of *S. sanctipauli*. The key should not be used west of Côte d'Ivoire, where typical *S. soubrense* and *S. soubrense* 'B' might also be encountered.

1) Larva homozygous for inversions 1L-P&Q, 2L-4&6 and 3L-2

....................*S. sanctipauli* subcomplex 2)

These inversions absent from larva

....................Other species of
S. damnosum complex

2) Larva homozygous for inversion 2L-A

....................*S. sanctipauli* 3)

Inversion 2L-A absent from larva

....................*S. soubrense* 4)

3) Inversion 1S-21 Y-linked in population

....................*S. sanctipauli* 'Djodji'

Inversion 1S-21 not Y-linked in population

....................*S. sanctipauli* typical

4) Inversion 2S-6b present in larva

....................*S. soubrense* 'Beffa'

Inversion 2S-6b absent from larva

....................*S. soubrense* typical

2.4 DISCUSSION

The new cytotype seems to be largely limited to the Asukawkaw and Dayi river systems in the mountainous forest on the Ghana/Togo border (see Table 2.1). Within Togo and Benin, to the north and east, *S. soubrense* 'Beffa' appears to be the sole representative of the *S. sanctipauli* subcomplex except for a few samples of the Djodji form identified further north in the savanna from the rivers Kpaza and Niankpe in October 1987. To the west of the Volta lake *S. sanctipauli* typical form and *S. soubrense* are found (Meredith *et al.* 1983, Post 1986, Fiasorgbor, Weber, Post, Surtees unpublished data, Chapter three).

In view of the absence of any sympatric samples or unique fixed inversions, there is no evidence for Djodji form being a species distinct from *S. sanctipauli* elsewhere. However, the sex-linkage of 1S-21 shows that Djodji populations are by definition genetically differentiated from other *S. sanctipauli* populations. There is also evidence for multivariate morphometric differentiation between typical *S. sanctipauli* and *S. sanctipauli* 'Djodji' (Chapter six). Therefore it seems that Djodji form should be considered to be a geographic race of *S. sanctipauli*.

The taxonomic significance of sex-linked inversions in the Simuliidae has been discussed by Rothfels (1979), Rothfels and Nambiar (1981) and Post (1982). Most blackfly species do not have distinguishable sex chromosomes, but sex chromosome differentiation can occur, often by linkage of inversions to the primary sex-determining region. Often species differ only in their sex chromosomes, which has led to the hypothesis that sex chromosome evolution may play a functional role in speciation within blackflies (Rothfels 1979).

Within the *S. damnosum* complex sex-linked inversions have been considered important in the cytotaxonomic description of several forms and species, such as *S. soubrense* 'Beffa' (Meredith *et al.* 1983), *S. yahense* (Vajime and Dunbar 1975), and Turiani form (Dunbar and Vajime 1981).

The importance of Djodji form lies in its possible importance in onchocerciasis transmission, which is discussed by Garms and Cheke (1985) and Cheke and Denke (1988). Cheke and Denke (1988) show that *S. sanctipauli* 'Djodji' is potentially a better vector than *S. squamosum* in Togo, and better than *S. sanctipauli* from Côte d'Ivoire, where *S. sanctipauli* is believed to be more zoophilic.

List of larval samples from which the Djodji form of *S. sanctipauli* has been identified.

Sample	River	Coordinates (N/E)	Date	Collectors[1]	Cytospecies[2] composition				
					squ	yah	san	dam	sir
1	Dayi	7°09' 0°29'	13.01.87	AKA	22	5	3	1	0
2	Dayi	7°07' 0°27'	22.01.87	RAC EAW YY	20	74	8	1	0
3	Dayi	7°06' 0°26'	17.03.86	AKA	31	12	4	0	0
4	Dayi	7°06' 0°26'	29.04.86	SSH	20	5	4	0	0
5	Dayi	6°57' 0°21'	12.02.86	AKA	17	15	21	55	2
6	Dayi	6°57' 0°21'	21.03.86	AKA	9	0	8	47	0
7	Dayi	6°53' 0°21'	06.05.86	SSH	2	0	4	8	3
8	Dayi	6°53' 0°21'	16.05.86	SSH	5	2	17	16	8
9	Dayi	6°53' 0°21'	13.01.87	RAC EAW YY	2	0	4	8	3
10	Dayi	6°52' 0°19'	29.05.86	AKA	2	0	8	19	2
11	Dayi	6°52' 0°19'	22.01.87	RAC EAW YY	2	0	1	7	0
12	Asukawkaw	7°54' 0°37'	05.02.87	JFW JEEH	54	0	7	0	0
13	Asukawkaw	7°54' 0°37'	27.05.86	YY	41	0	13	0	0
14	Asukawkaw	7°54' 0°37'	23.01.87	RAC EAW YY	91	0	27	0	0
15	Asukawkaw	7°52' 0°36'	18.03.86	MD	26	0	4	0	0
16	Asukawkaw	7°52' 0°29'	23.01.87	RAC EAW YY	33	0	28	0	0
17	Asukawkaw	7°41' 0°26'	28.05.86	YY	29	0	52	0	0
18	Asukawkaw	7°41' 0°25'	06.02.86		5	0	90	0	0
19	Asukawkaw	7°41' 0°25'	23.01.87	RAC EAW YY	2	0	74	0	0
20	Menou	7°37' 0°39'	28.05.86	YY	28	0	14	0	0
21	Gban-Houa	7°42' 0°38'	28.05.86	YY	3	0	25	0	0
22	Gban-Houa	7°42' 0°38'	23.01.87	RAC EAW YY	12	0	29	0	0
23	Gban-Houa	7°41' 0°37'	06.02.86	YY	12	0	18	0	0
24	Gban-Houa	7°42' 0°36'	23.01.87	RAC EAW YY	11	0	22	1	0
25	Gban-Houa	7°42' 0°35'	15.10.84	RJP CKP	4	0	49	0	0
26	Gban-Houa	7°42' 0°35'	15.03.85	RAC AMD	17	0	23	0	0
27	Gban-Houa	7°42' 0°35'	21.03.85	RAC AMD	11	0	23	0	0
28	Gban-Houa	7°42' 0°35'	26.03.85	RAC AMD	8	0	32	0	0
29	Gban-Houa	7°42' 0°35'	29.03.85	RAC AMD	10	0	29	0	0
30	Gban-Houa	7°42' 0°35'	15.10.85	RAC AMD	13	0	21	0	0
31	Gban-Houa	7°42' 0°35'	15.03.86	YY	45	0	22	0	0
32	Gban-Houa	7°42' 0°35'	27.01.87	RAC HSA	28	0	54	1	0
33	Wawa	7°43' 0°33'	20.03.86	YY	13	0	12	0	0
34	Wawa	7°43' 0°33'	23.01.87	RAC EAW YY	13	0	32	0	0
35	Wawa	7°41' 0°30'	27.03.86	AKA	17	0	6	0	0
36	Kpaza	8°33' 0°41'	15.10.87	JFW SS			-		
37	Kpaza	8°33' 0°41'	22.10.87	JFW YY AKO			-		
38	Kpaza	8°33' 0°37'	22.10.87	JFW YY AKO			-		
39	Niankpe	9°05' 0°42'	22.10.87	JFW YY AKO			-		

[1] AKA=A.K.Adzah, AKO=A.K. Opoku, AMD=A.M. Denke, CKP=C.K. Post EAW=E.A. Weber, HSA=H.S.K. Avissey, JEEH=J.E.E. Henerickx, JFW=J.F. Walsh, MA= M. Ampah, MD=M. David, RAC=R.A. Cheke, RJP=R.J. Post, SS=S. Sowah, SSH=OCP subsector Hohoe, YY=Y. Yamagata. Samples 25-30 were determined by D.P. Surtees.

[2] sq= *S. squamosum*, yah=*S. yahense*, san=*S. sanctipauli* 'Djodji', dam=*S. damnosum s.s.*, and sir=*S. sirbanum*. Other cytospecies were not found in these samples, except for 1 *S. soubrense* in sample 37. Samples 36-39 were not random samples, so numbers of flies identified for each species are not given.

Table 2.2

1S-21 karyotype frequencies

River	Samples[1]	Number of males			Number of females		
		st/st	st/21	21/21	st/st	st/21	21/21
Dayi	1,2,4,7,8 9,11	2	14	0	22	1	0
Asukawkaw	14,16,19	0	58	0	31	2	0
Gban-Houa	22,24,26,27 28,29,30,32	3	108	0	113	1	0
Wawa	34	1	12	0	8	0	0
Kpaza	36,37	0	10	0	11	0	0

Table 2.3

Autosomal polymorphic inversion frequencies

River	Samples[1]	Polymorphic inversions[2]					
		1S-A	2L-7	3L-B	3L-4.17	3L-24	IS-P
Dayi	1,2,4,7,8 9,11	0.56	1.00	0.0	1.00	0.0	0.0
Asukawkaw	14,16,19	0.42	1.00	0.0	0.97	0.0	0.0
Gban-Houa	22,24,26,27 28,29,30,32	0.45	1.00	0.0	0.996	0.002	0.002
Wawa	34	0.50	1.00	0.0	1.0	0.0	0.0
Kpaza	36,37	0.31	1.00	0.0	1.00	0.0	0.0

[1] Samples are as listed in Table 2.1. However, it was not always possible to score all inversions in every specimen, and hence sample sizes may be slightly smaller than those listed in Table 2.1.

[2] Inversions 2L-7 and 3L-B were noted heterozygously in a very few specimens from other samples which were not scored systematically for autosomal inversions.

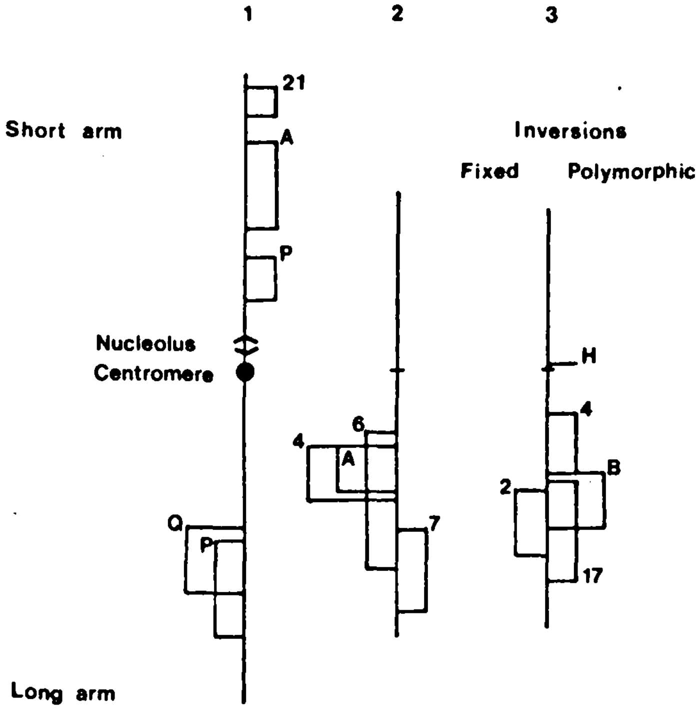


Figure 2.1

Idiogram showing the relative positions of all the inversions currently known from the Djodji form of *S. sanctipauli*. Inversions plotted on the right of each chromosome are intraspecific polymorphisms, while those to the left are fixed inversions relative to the standard sequence in *S. squamosum*. The polymorphic inversion 3L-B is based on the 3L-4.17.2 sequence. Most of these inversions are illustrated in Post (1986), although 1S-21, 1S-A and a new inversion, 1S-P are also shown on Figures 2.2 and 2.3.

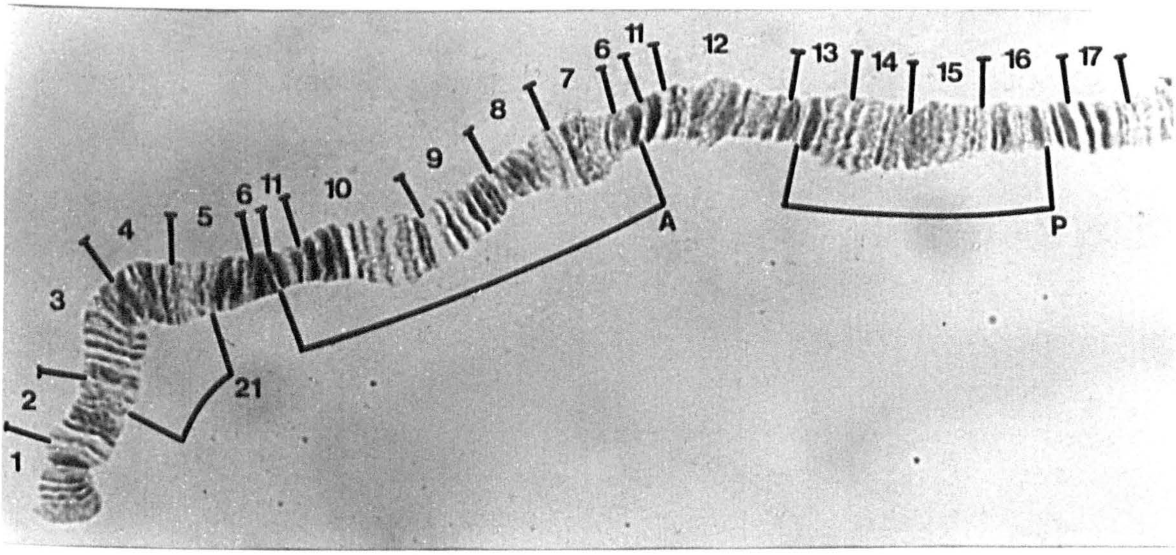


Figure 2.2

S. sanctipauli chromosome arm 1S, from the River Sassandra at Soubre, Cote d'Ivoire, 26.10.84 (see Table 3.1) showing the karyotype 1S-A/A, with the breakpoints of 1S-P and 1S-21 indicated.

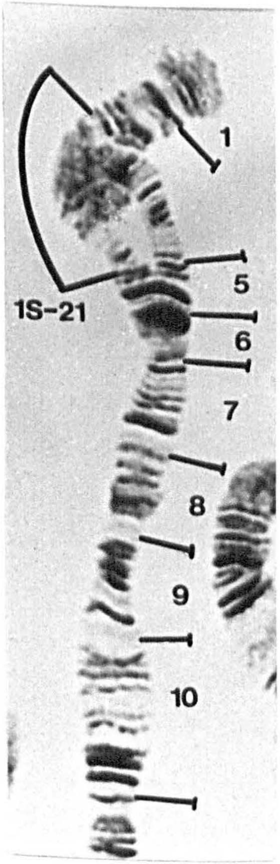


Figure 2.3

S. sanctipauli 'Djodji' chromosome arm 1S, from the river Gban-Houa, Djodji, Togo, showing the 1S-St/21 karyotype.

CHAPTER THREE: MULTIVARIATE ANALYSIS OF POLYTENE CHROMOSOME INVERSION FREQUENCIES WITHIN THE SIMULIUM SANCTIPAULI SUBCOMPLEX

3.1 INTRODUCTION

The delimitation of taxa within sibling species complexes of medical significance is an important activity before investigation of behavioural, ecological or physiological aspects relevant to disease transmission (WHO Technical Report 597, 1976). Within the *S. damnosum* complex in West Africa this delimitation of taxa was established by Vajime and Dunbar (1975), with later revision of the *S. sanctipauli* subcomplex by Post (1986).

Chapter two has emphasised the essential role that 'classical' cytotaxonomic methods play in the further development of the understanding of a species complex. The purpose of the present chapter is to use multivariate statistical techniques to analyse, objectively, polytene chromosome variation within and between members of the *S. sanctipauli* subcomplex, which is the best understood part of the *S. damnosum* complex (Post 1986), and to contrast the results thus obtained with those obtained using a 'classical' approach.

The cytotaxonomy of the *S. sanctipauli* subcomplex has been split into a number of cytoforms within the three cytospecies, *S. sanctipauli*, *S. soubrense* and *S. soubrense* 'B'. These are:

S. soubrense 'Beffa' (Meredith *et al.* 1983)

S. soubrense 'Menankaya' (Boakye personal communication)

S. soubrense 'Chutes-Milo' (Boakye personal communication)

S. soubrense 'Konkoure' (Quillévéré *et al.* 1982)

S. sanctipauli 'Djodji' (Surtees *et al.* 1988, Chapter two)

The usual approach to cytotaxonomy uses a sliding scale of chromosomal evidence for defining taxa (Rothfels 1956, Bedo 1977). The strongest evidence for specific status comes from differences between individuals maintained in sympatry, without apparent introgression (Mayr 1970). As an example from the *S. damnosum* complex, *S. squamosum*, and *S. sanctipauli*, in the River Gban-Houa at Djodji, Togo, have the fixed inversions 1L P&Q, 2L4&6 and 3L-2 maintained between them in sympatry. There are many such examples within the *S. damnosum* complex (Vajime and Dunbar 1975, Dunbar and Vajime 1981, Post 1986). The next strongest evidence for specific status comes from sex-linked inversion differences maintained in sympatry. X chromosome linked inversion differences are as diagnostic as fixed differences, while Y chromosome differences render females of the putative species homosequential. The next strongest evidence comes from polymorphic autosomal inversions found at different frequencies in sympatry. Analysis of sample inversion frequencies, testing for departures from Hardy-Weinberg equilibrium and for linkage disequilibrium can reveal the existence of non-introgressing species in sympatry.

In the absence of sympatric samples, then by analogy, the same three degrees of chromosomal evidence have been used to define taxa in allopatry (for example *S. sanctipauli*, and *S. sanctipauli* 'Djodji', Chapter two), although such evidence for specific status is always weaker than the equivalent evidence in sympatry (Mayr 1970).

When the taxonomy of a group reaches the state of knowledge achieved in the *S. sanctipauli* subcomplex, there is often subjectivity in the criteria used for defining a distinctive taxonomic or cytotaxonomic category and in the taxonomic rank to which a new cytoform should be raised. Clearly it is advantageous for more rigorous objective methods to be used to describe inter- and intraspecific variation.

Two approaches could be taken to the objective analysis of polytene chromosome variation. One method is to use standard genetic distance models to assess the evolutionary relationships between divergent taxa in a formal hypothesis-testing regime (Nei 1987). This approach has often been used for the analysis of isoenzyme variation, and could be used with cytotaxonomic data. However, within the *S. sanctipauli* subcomplex, the nature of the sampling regime and the sample sizes obtained preclude the use of this approach. Therefore it was decided to use the relatively informal methods of multivariate exploratory data analysis (Gnanadesikan 1977, Gordon 1981), using sample inversion frequencies as continuous pseudo-phenetic characters bounded by zero and one, to analyse inter- and intra- specific variation. Because the taxonomy of this subcomplex is the best known within the *S. damnosum* complex, a well established *a priori* taxonomy based on classical polytene chromosome analysis was available for comparison with the results of the multivariate analysis.

The objectives of this chapter can be clearly stated:

1. To analyse polytene chromosome inversion frequencies in the whole *S. sanctipauli* subcomplex;

2. To compare and contrast the *a priori* taxonomy of the *S. sanctipauli* subcomplex with that derived using multivariate statistical methods;

3. To analyse variation within selected *a priori* defined species.

3.2 MATERIALS AND METHODS

Samples of larvae were available from collections made between 1971 and 1986 from sites between Guinea and Nigeria in West Africa. Table 3.1 gives the details of each sample, the river, site and country of collection, the latitude and longitude of the site, the name of the cytotaxonomist who scored the inversion frequencies and the *a priori* taxonomic category of the sample.· Table 3.2 shows the sample inversion frequencies for the 47 inversions variant within the *S. sanctipauli* subcomplex, grouped by chromosome. Sample sizes listed are the maximum. It was not always possible to score all inversions for all specimens so sample sizes for each inversion are occasionally smaller than the maximum.

The chromosome preparations of the samples listed in Table 3.1 were made using the methods described in Chapter two. The larvae were identified chromosomally using the chromosome standards of Post (1986). Most of the identifications were performed by Dr R.J. Post, with others performed by D.P. Surtees.

Two sets of multivariate statistical techniques were used in the exploratory data analysis of the sample inversion frequencies listed in Table 3.2, ordination methods and cluster analysis methods (Gordon 1981). These two sets of techniques were used in a complementary way, with the results of one set helping in the interpretation of the results of the other set (Kruskal 1977). Information in a two dimensional ordination is often easier to interpret if the results from a cluster analysis of the same data are used in conjunction, and similarly an ordination can be used to assess a partition of the data resulting from a particular cluster method.

Within these two sets of techniques different methods were used, rather than using only one ordination method and one cluster analysis method to help to avoid artefacts. Cluster analysis methods in particular can sometimes produce interpretable results even in the absence of real structure in a data set, and can impose a structure on a data set which is different from that actually contained in the data (Gordon 1981). Using a plurality of methods satisfying different criteria helps to avoid this problem.

For general reference to the methods used in this analysis, see Gordon (1981), Seber (1984), Everitt (1978), Gnanadesikan (1977), Sneath and Sokal (1973).

3.2.1 ORDINATION METHODS

The raw data matrix (Table 3.2) is difficult to interpret as it stands, as it is 66 rows (samples) by 47 columns (inversions) in size. Ordination methods attempt to derive a lower dimensional graphical representation of high dimensional data, which retains as much of the information contained in the full data set as possible while producing a great simplification of the data. There are many ordination methods (Gordon 1981, Seber 1984), including principal components analysis and non-metric multidimensional scaling, which were used in this analysis. The former is an R-mode method, i.e the analysis is performed on the relationships between the characters (inversions) while the latter is a Q-mode method, i.e the analysis is performed on the relationship between individuals (samples of larvae), (Sneath and Sokal 1973).

3.2.1.1 Principal Components Analysis

PCA is a very well established technique (Pearson 1901, Hotelling 1933) which involves the extraction of eigenvalues and eigenvectors from the sample dispersion (variance/covariance) matrix, or from the correlation matrix derived from the dispersion matrix (Seber 1984). The first principal component is the linear combination of the original variables which accounts for the largest proportion of the total variance, so the first principal plane resulting from the scatter of points in the plane of the first two principal components accounts for the largest proportion of variance of all orthogonal axes. There are many uses to which PCA can be put, including interpreting patterns of covariation between characters, identifying redundant dimensions and identifying outliers (Seber 1984), but in this analysis the method was used as a low dimensional representation of the high-dimensional data set. By maximising variance it is assumed that information content is also maximised in the first few dimensions. In theory and practice this is not necessarily so (Cheng 1983), and so the results of a principal components analysis have always to be interpreted with caution.

Principal components can be extracted from either the dispersion matrix or the correlation matrix derived from it (Seber 1984, Gnanadesikan 1977). In this analysis the dispersion matrix was used, because the first two principal components accounted for a larger proportion of total variance than did the first two principal components of the correlation matrix (see section 3.3). The program used was SAS PROC PRINCOMP.

As an aid to understanding the relationship between the original inversion frequencies and the derived variables (principal components) a graphical technique called h-plotting was used (Corsten and Gabriel 1976, Seber 1984). With this method, each coefficient in the first two principal components (Table 3.3) was multiplied by the square root of the corresponding eigenvalue. The resultant co-ordinates show the strength of the relationship between the inversion and the principal components, and when superimposed on an ordination can help to explain the patterns uncovered in the data. The length of each vector is proportional to the standard deviation of the inversion and the cosine of the angle between any two vectors approximates to the correlation coefficient between the inversions. These h-plots were superimposed on the ordinations, but were not scaled to conform to the principal axes, and were not located at the origin as this would have obscured other details of the ordination.

3.2.1.2 Non-Metric Multidimensional Scaling

PCA is an R-mode method because the principal components are extracted from a matrix describing relationships between characters and individuals (samples) are then examined in the space defined by the new linear combinations of the original variables. Non-metric MDS is a Q-mode technique, in that a lower dimensional ordination of the data is derived from a matrix describing the relationship between individuals (Gordon 1981). This matrix could be defined by many proximity measures (Wishart 1978), but in this analysis the squared euclidean distance between individual samples was calculated, because it is the proximity measure also used in the cluster analysis methods described in section 3.2.2. Non-metric MDS works by finding a p-dimensional solution to the transformation of the proximity matrix (in this analysis the 66 by 66 matrix of euclidean distances between samples listed in Table 3.1) to a p-dimensional scatter of points in which the rank order of inter-point distances in the derived space matches as closely as possible the rank order of distances in the original proximity matrix (Gordon 1981). P (the dimensionality of the final solution) is a parameter defined by the user, and the adequacy of a particular p-dimensional solution is assessed using a parameter called stress (Gordon 1981) which is minimised. The algorithm used in this analysis was an iterative least-squares method, within a general multidimensional scaling package ALSCAL (Young *et al.* 1980).

A problem found with non-metric MDS is the choice of p, the dimensionality of solution. The usual method of choosing p is to plot the change in stress against increasing values of p, and to choose the lowest value of p for which the stress is tolerable. In this analysis, however, if a two-dimensional ordination did not have an acceptable stress value, higher dimensional solutions were not attempted, as this would have defeated the principle behind using the method. In practise, none of the ordinations resulting from the application of this method were used in the final interpretation of the data, because the ordinations resulting from PCA were usually clearer, and the principal axes are more easily interpretable in terms of the original inversions via the eigenvectors and the h-plot.

3.2.2 CLUSTER ANALYSIS METHODS

Ordination methods do not impose a structure onto the data matrix, instead they are a parsimonious representation of high dimensional data. Cluster analysis methods, however, impose a structure onto the data matrix (Gordon 1981). This structure can be of four types:

1. Non-overlapping partitions of the data
2. Hierarchically nested partitions of the data
3. Overlapping hierarchical partitions of the data
4. Overlapping non-hierarchical partitions of the data.

Of these four structures, methods resulting in the first three structures were used in this analysis.

3.2.2.1 Non-overlapping Partitioning Methods

In these methods the data set of n objects is divided into g-groups, where g can be set automatically or by the user depending on the algorithm used (Gordon 1981). Initially, for a g-partition of the data, the objects may be assigned randomly to the g-groups or as a result of a previous cluster analysis method. Objects are then iteratively relocated from one group to another, if doing so helps improve the parameter being optimised. The parameters optimised and the relocation procedure used are all algorithm dependent. In this analysis SAS PROC FASTCLUS was used, which iteratively minimises the sum of squared distances from the g-cluster means.

3.2.2.2 Hierarchical Cluster Methods

Hierarchical methods begin with n clusters (in this analysis n is the number of independent samples i.e. 66) which are successively fused until all belong to one cluster. Clusters formed at a lower level are completely incorporated in higher level clusters (Gordon 1987). In this analysis, four hierarchical cluster methods were used: Single linkage (nearest neighbour), Complete linkage (furthest neighbour), Group Average (UPGMA) and Ward's error sums-of-squares method (Ward 1963). These four were chosen as they are the commonest methods in use, and hence are the best understood (Gordon 1981). In addition, the four methods have quite different statistical properties which accords with the general principle of this analysis that greater confidence is obtained from similar results derived from methods satisfying different criteria.

The different methods work by fusing at each stage the two clusters which are most similar, but they differ in the way each method defines cluster similarity. Single linkage calculates the distance between two clusters as the distance between the nearest neighbours. Complete linkage calculates it as the distance between the remotest members of the two clusters. Group Average is intermediate between these methods in calculating the distance between two clusters as the average of all pairwise distances between members of the two clusters. Ward's method defines the distance between clusters as the increase in within-groups sums-of-squares which would result from the fusion of two clusters. In all cases, the two clusters for which the various definitions of distance is a minimum are fused (Gordon 1987). Hierarchical methods are usually expressed as two-dimensional branching trees called dendrograms. Within this analysis, the cluster analysis package CLUSTAN (Wishart 1978) and SAS PROC CLUSTER were used.

3.2.2.3 Overlapping Methods

The two previous methods share the restriction that an object can belong to only one cluster. Jardine (1971) argues that certain types of natural variation, including intraspecific geographic variation, is often not appropriately expressed either as an hierarchy or as a non-overlapping partition, but instead may take the form of a continuum (e.g. clinal variation, Endler 1977) or as recognised 'types' with intermediates between them.

To analyse within-taxon variation of the *a priori* taxa *S. sanctipauli*, *S. soubrense* 'Menankaya/Konkoure' and *S. soubrense* 'B' a set of overlapping cluster methods was used. These were the B_k methods of Jardine and Sibson (1968). These methods can best be understood in terms of two parameters, h and k. H is the distance between points and k is the amount of overlap allowed by the method, k-1 points being allowed to belong to the overlap of two clusters at a particular value of h. The method starts with h=0 (i.e. identical objects), h is then increased to a particular level (chosen in retrospect by the user because of perceived discontinuities in the parameter, or for other data-analytic reasons). At this value of h, all maximally complete subgraphs are drawn in (i.e. subsets of points in which all the points are connected), and any pair of subgraphs which coincide in at least k points are further fused to form a cluster. For example, if k=2, and two clusters share only one point in common, then both clusters remain distinct, but the point in common lies on the overlap of the two clusters. These methods become extremely complex to interpret for moderate numbers of individuals and values of k greater than 4, so higher values of k were not attempted. The CLUSTAN procedure KDEND (Wishart 1978) and the NTSYS procedure BKGRAPH (Rohlf 1984) were used for the analysis.

3.2.2.4 Other Methods

One problem associated with lower dimensional ordinations of high dimensional data is that certain of the interpoint distances become distorted (Seber 1984). To assess this distortion a minimum spanning tree (MST) was calculated and superimposed on the ordination. A MST is the tree connecting the n vertices (in this analysis n=66, the number of samples in Table 3.1) forming a connected graph containing no loops for which the sum of the edge lengths is a minimum (Gower and Ross 1969).

Hierarchical techniques can be seen as a transformation of the proximity matrix between individuals into a new proximity matrix satisfying the ultrametric inequality (Jardine and Sibson 1971, Gordon 1981). Generally this transformation results in some distortion, the extent of which can invalidate a hierarchical representation of the proximity matrix. In this analysis, four distortion measures were used to assess the extent of the distortion: the cophenetic correlation coefficient, r_{cop} (Sokal and Rohlf 1962) and three of the Jardine and Sibson (1968) Δ_i distortion measures.

A common problem to clustering methods is the objective estimate of the true number of clusters within a data set. For hierarchical methods this can be thought of as estimating the level at which a line should be drawn across the dendrogram, and for non-overlapping partition methods this amounts to choosing the value of g at which to stop the algorithm. For this analysis the cubic clustering criterion (CCC) within SAS PROC CLUSTER was used. This criterion compares a particular partition of the data with that expected if the data were sampled from a uniform distribution. The value for which this criterion is largest is taken as the 'true' number of clusters. This method was evaluated by Milligan and Cooper (1985) who found that it compared favourably with most other criteria in the literature, and was considered the best of the widely available methods.

Once a particular partition was obtained by a cluster method it was then compared with results obtained from other cluster methods. A combination of visual inspection of the hierarchical dendrograms to identify cohesive, isolated clusters and a cluster intersection method described in Gordon (1981) to find the maximum number of points in common to two methods was used to compare the results of cluster analyses of the data set. This procedure resulted in a consensus partition of the data defining clusters consistently uncovered using the different methods, but from which certain points were excluded because of their inconsistent classification using different clustering methods.

3.2.2.5 Summary of Methods

Ordination methods and cluster analysis methods were used to analyse the total data set presented in Table 3.2. Principal components analysis and non-metric multidimensional scaling were both used, but only the the results of the favoured method (PCA) are presented. The full data set was clustered using four hierarchical cluster methods and one non-overlapping cluster method. The optimal partition of the data as determined by the CCC was derived for each method and these partitions compared using a cluster intersection method and visual inspection, and a consensus partition of the total data set derived.

In addition to this whole data set analysis, three sub-analyses were performed on the three *a priori* groups *S. soubrense* 'B', *S. sanctipauli* and *S. soubrense* 'Menankaya/Konkoure'. A separate analysis was not performed for *S. soubrense* 'Beffa' or *S. soubrense* 'Chutes Milo' because of the small number of samples of each. An overlapping cluster analysis method was used for these intraspecific analyses as this had been shown in previous studies (Jardine 1971) to be more sensitive to intraspecific variation.

3.3 RESULTS

The first principal plane of the dispersion matrix accounted for 65% of total variance and was used in preference to the first principal plane of the correlation matrix which only accounted for 24% of variance. This was so because many inversions were not correlated, but some inversions had a relatively large variance, the effect of which is damped if the correlation matrix is used in a PCA. A two dimensional solution to a non-metric MDS of the squared euclidean distance matrix between samples resulted in a stress of 12.6%, which is only a 'fair to poor' fit (Kruskal 1964). Therefore this ordination was not used, but the scatter of points in the first principal plane of the dispersion matrix used instead (Figure 3.1). This figure is annotated with the sample numbers corresponding with those in Table 3.1. The points are connected by the minimum spanning tree derived from the squared euclidean distance matrix between samples. Also shown is the h-plot of the dispersion matrix. Table 3.3 gives the first two principal components of the dispersion matrix, demonstrating which sets of inversions have the most influence on the first two principal axes.

Figure 3.2 is the dendrogram resulting from application of single linkage cluster analysis to the squared euclidean distance matrix. The numbers at the tips of the dendrogram correspond to the sample numbers in Table 3.1. Figures 3.3 to 3.5 are the dendrograms resulting from application of complete linkage, Group Average, and Ward's method of cluster analysis.

Table 3.4 gives the measures of distortion (cophenetic correlation coefficient and the three Jardine-Sibson distortion measures) resulting from the hierarchical cluster methods. Also shown is the partition for each hierarchical method and the non-overlapping partition method for which the CCC was a maximum.

Table 3.5 shows the five partitions resulting from the cluster methods at the level suggested by the CCC for each method, while Table 3.6 is the 'consensus' partition obtained by the method described in section 3.2.2.4.

Table 3.7 gives the reduction in the Jardine-Sibson distortion measure $\Delta_{0.5}$ as the B_k methods were applied to the three data sets analysing the pre-defined groups each of which represents a single species defined by classical cytotaxonomy, *S. soubrense* 'Menankaya/Konkoure', *S. soubrense* 'B' and *S. sanctipauli*. K was increased from one (where it is equivalent to single linkage) to four (beyond which the results were extremely complex).

Figure 3.6 is the first principal plane of the dispersion matrix for *S. soubrense* 'B', with the h-plot superimposed and the cluster boundaries defined by applying the $B_{k=2}$ method at the level h=0.001. Figure 3.7 is the scatter of points in the first principal plane of the dispersion matrix of the *S. sanctipauli* samples, with the h-plot superimposed and the cluster boundaries defined by applying the $B_{k=2}$ method at h=0.001. Figure 3.8 is the first principal plane of the dispersion matrix for *S. soubrense* 'Menankaya/Konkoure' with the h-plot superimposed and the cluster boundaries defined by applying the $B_{k=2}$ method at level h=0.001.

3.4 DISCUSSION

3.4.1 SIMULIUM SOUBRENSE 'B'

Simulium soubrense B (samples 32-45) is the most clearly defined and internally most homogeneous taxon, both on the ordination (Figure 3.1) and on the dendrograms resulting from the application of the four hierarchical cluster methods (Figures 3.2-3.5). The cluster {32,...,45} was found by the CCC in the full data set for all the cluster methods used (Table 3.5) and remained in the consensus classification (Table 3.6).

This distinctiveness relative to the other taxa within the *S. sanctipauli* subcomplex can be explained by reference to the h-plot of the dispersion matrix (Figure 3.1), the original data set (Table 3.2) and the principal components (Table 3.3). Pointing directly at *S. soubrense* 'B' on the h-plot are the inversions 1L-A and 2S-7. The former is a fixed, unique derived character for this species (Post 1986) and the latter is fixed in this species, although it is shared by other members of the subcomplex. Inversions which also have a strong though less direct effect on the distinctiveness of *S. soubrense* 'B' are the inversions 2L-D and 1S-A. 2L-D is fixed in this species but shared by *S. soubrense* except for Chutes Milo form. The vector for this inversion on the h-plot is almost coincidental with the first principal axis, so samples along this axis increase in frequency for the inversion. 1S-A, however is orthogonal to the first principal axis, and hence uncorrelated with inversion 2L-D. This inversion is also fixed in *S. soubrense* 'B' but is shared with other members of the subcomplex.

When the B_k methods of cluster analysis were applied to analyse variation within *S. soubrense* 'B', the distortion measure $\Delta_{0.5}$ fell only slightly as k was increased from one to four (Table 3.7); such a high level of distortion suggests that there is little structure within the data set, either hierarchical in nature or overlapping (Jardine and Sibson 1968). The two-dimensional ordination resulting from a principal components analysis of the dispersion matrix is shown as Figure 3.6, and superimposed on this ordination are the clusters resulting from application of the $B_{k=2}$ method at h=0.001 (chosen because of a distinct moat at this level). This three cluster solution shows that the bulk of the data belong to a homogeneous cluster, with two outlying samples. Three inversions are important in this variation, 1S-C, 3L-4, 3L-17. The small angle between the last two on the h-plot reflects their strong linkage. 1S-C was only found at a low frequency in *S. soubrense* 'B' (for which species it is unique), with the maximum frequency for sample 40. Thus the variation along the second principal axis is dominated by this relatively unimportant inversion, demonstrating that there is in fact very little intraspecific variation. 3L-4 and 3L-17 are partially X-linked in *S. soubrense* 'B' (Post 1986), although this information was not included in the analysis. As expected from the h-plot, the frequency of these inversions is lowest in sample 44 and largest in sample 41. This is because sample 41 was a sample of six males and four females, while sample 44 was a sample of all females.

To conclude *S. soubrense* 'B' represents a chromosomally very homogeneous taxon which is very distinct from other members of the subcomplex, a result which supports the conclusions of classical cytotaxonomy (Post 1986). The only intraspecific variation found is likely to be because of random variation and sex ratio differences rather than being due to any systematic geographic or temporal variation. The considerable number of fixed inversions in this taxon (including inversions shared with other members of the subcomplex, Table 3.2) and the species' restricted geographic distribution suggests that the founding population for this species may have been small (i.e. the origin of the species involved a population bottleneck, Mayr 1970), or that the species is very ancient.

3.4.2 <u>SIMULIUM SOUBRENSE</u> 'CHUTES MILO/BEFFA'

By contrast *S. soubrense* 'Chutes Milo/Beffa' is an ill-defined taxon, both on the ordination (Figure 3.1) and on the dendrograms (Figures 3.2-3.5). Samples 1-4 belong to *S. soubrense* 'Beffa', but this taxon was not uncovered by any of the cluster methods and was not defined in the consensus classification (Table 3.6). This may in part be because of the information contained in the partially Y-linked inversion 2S-6b not being included in the analysis. Samples 5-8 correspond to *S. soubrense* 'Chutes Milo/Typical' and is much better defined, although only the three member cluster {6,7,8} remained in the consensus classification. On the ordination these samples lie close to *S. sanctipauli*, the only major inversion difference between these taxa being the inversion 2L-A which is fixed in *S. sanctipauli* and absent from *S. soubrense* (Post 1986).

Because of the small number of samples, *S. soubrense* 'Beffa' and *S. soubrense* 'Chutes Milo' were not analysed separately from the other samples.

To conclude *S. soubrense* 'Beffa' and *S. soubrense* 'Chutes Milo' are relatively heterogeneous taxa, although the former is more so than the latter. The very small number of samples does not allow for interpretation either in terms of geography or time. The close proximity of these samples to *S. sanctipauli* is of considerable interest, especially when compared with the large taxonomic distance between either of these taxa and *S. soubrense* 'B'.

3.4.3 SIMULIUM SANCTIPAULI

Simulium sanctipauli (samples 46-66) appears as a heterogeneous taxon both on the ordination (Figure 3.1) and on the dendrograms (Figures 3.2-3.5). The principal inversion distinguishing this species from the rest of the subcomplex is 2L-A (Post 1986), which is a unique fixed derived character for *S. sanctipauli*. On the ordination, much of the variation within *S. sanctipauli* is because of variation in the inversion 1S-A, as is clear from the h-plot and the original data matrix (Table 3.2).

The consensus partition of the whole data set (Table 3.6) resulted in three consistent clusters relative to the total data set {46,...,50}, {54,55,56,63,64} and {57, 59,...,62}. Some sample points were not included in the consensus partition because they were inconsistently classified using the different cluster methods. The first cluster corresponds to Djodji form (Chapter two), the second are samples from the Comoe, Maraoue and Baoule rivers, while the last corresponds to samples from the Bandama and Sassandra rivers.

Application of the B_k methods of Jardine and Sibson (1968) to the *S. sanctipauli* data in isolation resulted in quite a marked reduction in the distortion imposed by the resultant classification (Table 3.7). This result implies that the data can more realistically be represented by allowing samples to overlap between clusters, although the high distortion remaining at $B_{k=4}$ showed that allowing for overlap has not entirely removed distortion from the classification.

The two-dimensional ordination resulting from a principal components analysis of the dispersion matrix is shown as Figure 3.7, with the clusters defined at h=0.001 for the method $B_{k=2}$. This level was chosen because of a distinct moat between this level and the next fusion. Seven clusters were identified in the data set, but four of these overlap. The most distinctive cluster corresponds to the samples from the Rivers Maraoue, Comoe and Baoule. These samples differ from other *S. sanctipauli* principally in the inversion 2L-7 which is absent from this cluster but at a high frequency in other clusters (as can be seen from the h-plot and Table 3.2). Also the inversion 3L-B which is absent from most other *S. sanctipauli* is present at a high frequency in this cluster. Despite the diverse geographic origin of the members of this cluster, it is internally homogeneous. The most important practical aspect of this cluster is that four of the samples (55,56,63,64) are the only samples within the *S. sanctipauli* subcomplex resistant to organo-phosphate insecticide (Kurtak and Post 1987).

The samples 46-50 form another relatively tight cluster, although one sample (49) is shared in common with another cluster. These samples correspond to the Djodji form of *S. sanctipauli* defined on classical criteria by Surtees *et al.* (1988). The information contained in the strongly Y-linked inversion 1S-21 was not incorporated in the analysis. If it had been then the cluster would have been more distinct.

The cluster {57,59,60,61,62} has two samples which overlap with other clusters. One sample overlaps with the cluster {57,58} both from the Bandama river, Cote d'Ivoire, but separated by 11 years (Table 3.1), the other cluster is the Djodji cluster. The samples within this cluster are all from the rivers Bandama and Sassandra. and are distinguished by a high frequency for inversion 1S-A.

The cluster {53,58} overlaps with the Bandama cluster and includes the Ghanaian sample from the River Ejisu. The cluster {51,52} is from the River Pra, Ghana, while the geographically isolated cluster {65,66} is from the River Moa in Sierra Leone, the most westerly of the *S. sanctipauli* samples. This last cluster is particularly unusual, lacking inversion 1S-A and sharing three inversions not shared with other *S. sanctipauli*: 1S-F, 1L-B and 1L-C.

To conclude, there is considerable variation within *S. sanctipauli*, at least some of which is likely to involve restriction of gene flow between clusters. The correlation of OP insecticide resistance with one cluster is strong evidence for restricted gene flow between it and the other clusters within *S. sanctipauli*. This feature will be very useful for tracing OP resistance movement. It is unlikely, therefore, that *S. sanctipauli* will remain as a unitary taxon once more information becomes available, and the Djodji form of *S. sanctipauli* defined on classical criteria (Chapter two) will not be the only cytoform within *S. sanctipauli*.

3.4.4 SIMULIUM SOUBRENSE 'MENANKAYA/KONKOURE'

Simulium soubrense 'Menankaya/Konkoure' lies on a broad band in the lower right quadrant of the ordination (Figure 3.1) and shows a considerable degree of chromosomal heterogeneity. The superimposed MST on figure 3.1 reveals that samples 9 and 16, which are close on the ordination are in fact quite distinct, showing that the ordination has introduced some distortion into the data. The h-plot of the dispersion matrix shows that several inversions define the distinction between this taxon and the other members of the subcomplex. Inversion 2L-D, which is coincident with the first principal axis is present in all the samples of *S. soubrense* 'Menankaya/Konkoure' (Table 3.2), but it varies in frequency from fixation to 0.25. However, this inversion mainly serves to define the right half of the ordination, including *S. soubrense* 'B', but excluding *S. soubrense* 'Chutes Milo/Beffa' and *S. sanctipauli.* The inversion 1L-A defines *S. soubrense* 'B' in the upper right quadrant, while inversion 1S-A which is present throughout the subcomplex is found at a low but variable frequency in *S. soubrense* 'Menankaya/Konkoure'. The inversion 2S-7 which has a strong influence at 30° to the first principal axis varies within *S. soubrense* 'Menankaya/Konkoure', as do the inversions 2L-X and 3L-X.

The different cluster methods all found the same clusters within *S. soubrense* 'Menankaya/Konkoure' (Table 3.5) and these remained in the consensus classification (Table 3.6). There are six clusters, although two of these are singletons (samples 9 and 14).

Applying the B_k methods to *S. soubrense* 'Menankaya/Konkoure' resulted in some reduction in the distortion resulting from the transformation from distance matrix to ultrametric matrix (Table 3.7) but this reduction was not as great as would be expected if the data were sampled from genuinely overlapping clusters (Jardine and Sibson, 1968, found that the measure of distortion fell from 0.528 to 0.146 as k was increased from one to four for overlapping clusters of the annual pearlwort *Sagina apetala*).

Figure 3.8 shows the ordination resulting from a principal components analysis of the dispersion matrix, and reveals the complexity of variation within this taxon. The four cluster and two singletons are marked. The four main clusters are connected by the MST in a sequence forming a horseshoe on the ordination from *S. soubrense* 'Menankaya' (samples 10-13) from Sierra Leone/Guinea, to *S. soubrense*

'Menankaya/Konkoure' (samples 15-23) from the Rivers Tene and Bafing in the Fouta Djalon, Guinea, to *S. soubrense* 'Konkoure' (samples 30,31) from the River Koumba , Guinea, to *S. soubrense* 'Konkoure' (samples 24-29) from the Konkoure and Kakrima rivers in Guinea. A sequence of clusters connected in this way suggests clinal variation with inadequate sampling between clusters (Jardine 1971). The h-plot reveals that several inversions influence the observed pattern of variation. Inversion 2S-7 is found in the Konkoure/Koumba samples, and in the Sierra Leone samples, although its frequency is much lower in the latter. 3L-X also defines the Konkoure/Koumba samples. 3L-2 has an opposite influence to this inversion, being highest in samples 15-23 and lowest in the Konkoure form. The main inversions defining *S. soubrense* 'Menankaya' are 3L-4, 3L-17 (which are linked) and 3L-5. This last inversion is a unique polymorphic derived character defining *S. soubrense* 'Menankaya'.

To conclude *S. soubrense* 'Menankaya/Konkoure' shows a very complex pattern of intraspecific variation involving a considerable number of inversions. Two extreme hypotheses can be established to explain these results. The first is that the data have been sampled inadequately from continuous clinal variation. The sequential pattern shown on the ordination supports this hypothesis. The second hypothesis is that one or all of the clusters uncovered represents distinct forms which in sympatry would not introgress. The distinctiveness and internal homogeneity of the derived clusters supports this hypothesis. Based on the available data it is not possible to choose between these alternatives. The most likely explanation is a combination of the two i.e. that there is clinal variation within *S. soubrense* 'Konkoure', but that *S. soubrense* 'Menankaya' is a form distinct from this.

3.5 CONCLUSIONS

The multivariate statistical analyses presented in this chapter usually support the findings of classical cytotaxonomy, but also identify features of the data which classical methods have overlooked such as the distinctiveness of the OP resistant *S. sanctipauli* flies, and the complex variation within *S. soubrense* 'Menankaya/Konkoure'. These results could now be used predictively to identify and trace OP resistance.

However, the methods could undoubtedly be improved on to make them more sensitive and powerful in the future. One important feature of blackfly cytotaxonomy which has not been incorporated in this analysis is the information contained in the sex-linkage of inversions. The importance of sex-linkage in blackflies has been stated before (e.g. Post 1982). To use this information, the sexes could be treated separately within each sample, although this would require larger sample sizes. To exploit the interpretative power of the multivariate methods, systematic sampling would be an improvement over the *ad hoc* sampling of the data in this analysis. Finally, these same methods could be used on individuals within samples, to identify sympatric taxa.

Table 3.1

List of larval samples of the *S. sanctipauli* subcomplex.

Sample No. and origin[1]	River	Site	Coordinates[2] (Lat./Long)	Date	Determ--ined by[3]	*a priori* taxon
1 N	Oshun	Ede	07°44' 05°34'E	14.07.82	A	*Simulium soubrense* Beffa
2 N	Ogun	Eruwa	07°25' 04°29'E	15.07.82	A	
3 T	Mono	Tetetou	07°02' 01°32'E	20.07.80	A	
4 T	Mono	Avegode	06°48' 01°36'E	14.11.85	A	
5 C	Leraba	Leraba Bridge	10°10' 05°04'W	10.06.71	A	*Simulium soubrense* Chutes Milo
6 C	Cestos	Darlu	-	06.05.71	A	
7 C	Niandan	Rapide Pampan	10°01' 09°41'W	21.12.85	A	
8 C	Niandan	Boria	09°28' 09°57'W	14.02.86	A	
9 G	Milo	Tiekoradu	09°00' 08°57'W	-	A	*Simulium soubrense* Menankaya
10 S	Sewa	Njaime-Sewafe	08°52' 11°13'W	30.11.80	A	
11 S	Sewa	Babawahun	07°59' 11°20'W	15.12.81	A	
12 S	Seli	Badala	09°19' 11°32'W	18.12.81	A	
13 S	Rokel	Bumbuna	09°03' 11°44'W	29.11.81	A	
14 S	Mongo	Musaia	09°46' 11°28'W	17.12.81	A	
15 G	Bafing	Sokotoro	10°38' 11°45'W	13.11.86	A	*Simulium soubrense* Konkoure/ Menankaya
16 G	Bafing	Lago	10°51' 11°36'W	22.12.85	A	
17 G	Bafing	Nduria	10°45' 11°45'W	13.02.86	A	
18 G	Bafing	Yagui	11°34' 10°52'W	22.11.86	A	
19 G	Tene	below bridge	11°01' 11°49'W	13.02.86	A	
20 G	Tene	Dankolo	11°01' 11°58'W	22.11.86	A	
21 G	Tene	above Chutes	11°01' 11°49'W	22.11.86	A	
22 G	Tene	above Bafing	11°07' 11°35'W	22.11.86	A	
23 G	Bafing	Koukotamba	11°13' 11°19'W	22.11.86	A	
24 G	Konkoure	Ganiya	10°29' 12°59'W	13.02.86	A	
25 G	Konkoure	Soukia	10°25' 13°10'W	12.11.86	A	
26 G	Konkoure	Bakere	10°31' 13°10'W	12.11.86	A	
27 G	Konkoure	Kanhan	10°28' 12°47'W	12.11.86	A	
28 G	Kakrima	Bougoula	10°34' 12°58'W	12.11.86	A	
29 G	Kakrima	Kaffima	10°50' 12°57'W	12.11.86	A	
30 G	Koumba	Sidipo	11°43' 12°56'W	11.11.86	A	
31 G	Koumba	Kokou	11°42' 12°54'W	11.11.86	A	
32 S	Moa	Tiwai Island	07°32' 11°22'W	16.09.83	A	*Simulium soubrense* 'B'
33 S	Waanje	Bandajuma	07°34' 11°39'W	21.06.83	A	
34 S	Sewa	Wubunge	07°48' 11°48'W	12.06.83	A	
35 S	Sewa	Mofwe	07°40' 11°58'W	08.12.81	A	
36 S	Tabe	Gbaiima	08°06' 11°51'W	08.12.80	A	
37 S	Teye	Mongeri	08°19' 11°44'W	11.08.83	A	
38 S	Taia	Mogbamu	08°01' 12°07'W	09.12.80	A	
39 S	Taia	Mogbamu	08°01' 12°07'W	10.06.83	A	
40 S	Gbangaia	Mokasi	07°59' 12°25'W	07.12.81	A	
41 S	Rokel	Katik	08°39' 12°30'W	10.12.80	A	
42 S	Bankasoka	Port Loko	08°46' 12°47'W	12.12.80	A	
43 S	Gt Scarcies	Kanka	09°43' 12°27'W	02.12.81	A	

44	G	Kolente	Malea	10°41' 12°37'W	12.02.86	A	
45	G	Kolente	Kolente	10°04' 12°38'W	13.11.86	A	
46	T	Gban-houa	Djodji	07°40' 00°35'E	15.10.84	B	*Simulium sanctipauli* Djodji
47	T	Gban-houa	Djodji	07°40' 00°35'E	15.03.85	B	
48	T	Gban-houa	Djodji	07°40' 00°35'E	21.03.85	B	
49	T	Gban-houa	Djodji	07°40' 00°35'E	26.03.85	B	
50	T	Gban-houa	Djodji	07°40' 00°35'E	29.03.85	B	
51	Gh	Pra	Hemang	05°11' 01°32'W	08.07.80	A	*Simulium sanctipauli*
52	Gh	Pra	Hemang	05°11' 01°32'W	17.07.82	A	
53	Gh	Ofin	Ejisu	05°57' 01°42'W	25.01.86	A	
54	C	Comoe	Mbaso	06°20' 03°30'W	27.07.80	A	
55	C	Comoe	Amouakro	-	14.02.85	A	
56	C	Maraoue	Danangoro	07°10' 05°56'W	10.08.82	A	
57	C	Bandama	Tiassale	05°53' 04°49'W	08.07.82	A	
58	C	Bandama	Ahoauti	06°07' 04°57'W	23.06.71	A	
59	C	Sassandra	Soubre	05°47' 06°37'W	10.07.82	A	
60	C	Sassandra	Soubre	05°47' 06°37'W	26.10.84	B	
61	C	Sassandra	Koperagui	-	22.01.85	A	
62	C	Sassandra	Chutes Nawa	05°47' 06°37'W	06.09.84	A	
63	M	Bouale	Madina Diasso	10°40' 07°40'W	26.01.86	A	
64	M	Bouale	Konigbougeula	10°45' 07°46'W	27.01.86	A	
65	S	Moa	Maloma	08°00' 10°50'W	06.12.80	A	
66	S	Moa	Tiwai Island	07°33' 11°22'W	16.09.83	A	

[1] N=Nigeria, C=Côte d'Ivoire, T=Togo, Gh=Ghana, G=Guinea, S=Sierra Leone, M=Mali
[2] All Latitudes are North.
[3] A=Determined by Dr. R.J. Post, B=Determined by D.P. Surtees

Table 3.2

Polymorphic Inversions within the *S. sanctipauli* subcomplex

Chromosome one inversions

Sample Number	Sample Size	1L-B	1L-C	1L-D	1L-U	1L-G	1L-N	1L-X	1L-A	1L-R	1L-T
1	5	0.083	0	0	0	0	0	0	0	0	0
2	14	0	0	0	0	0	0	0	0	0	0
3	12	0	0	0	0	0	0	0	0	0	0
4	10	0	0	0	0	0	0	0	0	0	0
5	26	0.115	0.115	0	0	0	0	0	0	0	0
6	11	0	0	0	0	0	0	0	0	0	0
7	5	0	0	0	0	0	0	0	0	0	0
8	8	0.333	0.333	0	0	0	0	0	0	0	0
9	8	0.688	0.75	0	0	0	0	0	0	0	0
10	13	0.038	0.038	0	0	0	0	0	0	0	0
11	8	0	0	0.063	0	0	0	0	0	0	0
12	7	0.214	0.214	0	0	0	0	0	0	0	0
13	11	0.045	0.045	0	0	0	0	0	0	0	0
14	25	0.54	0.54	0	0	0.02	0	0	0	0	0
15	21	0	0	0	0	0	0	0	0	0	0
16	12	0	0	0	0	0	0	0	0	0	0
17	11	0.05	0.05	0	0	0	0	0	0	0	0
18	10	0	0	0	0	0	0	0	0	0	0
19	8	0	0	0	0	0	0	0	0	0	0
20	23	0	0	0	0	0	0	0.023	0	0	0
21	24	0	0	0	0	0	0	0	0	0	0
22	11	0	0	0	0	0	0	0	0	0	0
23	19	0	0	0	0	0	0	0	0	0	0
24	17	0	0	0	0	0	0	0	0	0	0
25	29	0	0	0	0	0	0	0	0	0	0
26	17	0	0	0	0	0	0	0	0	0	0
27	16	0	0	0	0	0	0	0	0	0	0
28	20	0	0	0	0	0	0	0	0	0	0
29	31	0	0	0	0	0	0	0	0	0	0
30	23	0	0	0	0	0	0	0	0	0	0
31	32	0	0	0	0	0	0	0	0	0	0
32	28	0	0	0	0	0	0	0	1	0	0
33	24	0	0	0	0	0	0	0	1	0	0
34	39	0	0	0	0	0	0	0	1	0	0
35	15	0	0	0	0	0	0	0	1	0	0
36	12	0	0	0	0	0	0	0	1	0	0
37	22	0	0	0	0	0	0	0	1	0	0
38	22	0	0	0	0	0	0	0	1	0	0
39	40	0	0	0	0	0	0	0	1	0	0.013
40	8	0	0	0	0	0	0	0	1	0	0
41	10	0	0	0	0	0	0	0	1	0	0
42	25	0	0	0	0	0	0	0	1	0	0
43	42	0	0	0	0	0	0	0	1	0	0
44	12	0	0	0	0	0	0	0	1	0	0
45	14	0	0	0	0	0	0	0	1	0	0
46	29	0	0	0	0	0	0	0	0	0	0

47	23	0	0	0	0	0	0	0	0	0	0
48	23	0	0	0	0	0	0	0	0	0	0
49	32	0	0	0	0	0	0	0	0	0	0
50	21	0	0	0	0	0	0	0	0	0	0
51	16	0	0	0	0	0	0	0	0	0	0
52	30	0	0	0	0	0	0	0	0	0	0
53	7	0	0	0	0	0	0	0	0	0	0
54	45	0	0	0	0	0	0	0	0	0	0
55	7	0	0	0	0	0	0	0	0	0	0
56	14	0	0	0	0	0	0	0	0	0	0
57	43	0	0	0	0	0	0	0	0	0.012	0
58	21	0	0	0	0	0	0	0	0	0	0
59	18	0	0	0	0	0	0.028	0	0	0	0
60	24	0	0	0	0.021	0	0	0	0	0	0
61	5	0	0	0	0	0	0	0	0	0	0
62	6	0	0	0	0	0	0	0	0	0	0
63	12	0	0	0	0	0	0	0	0	0	0
64	10	0	0	0	0	0	0	0	0	0	0
65	8	0.5	0.5	0	0	0	0	0	0	0	0
66	21	0.262	0.262	0	0	0	0	0	0.024	0	0

Table 3.2 (continued)

Chromosome one inversions

Sample Number	1S-A	1S-G.	1S-J	1S-F	1S-B	1S-X	1S-N	1S-C	1S-M	1S-P	1S-21
1	0.5	0	0	0	0	0	0	0	0	0	0
2	0.821	0.179	0.036	0	0	0	0	0	0	0	0
3	1	0.818	0.136	0	0	0	0	0	0	0	0
4	1	0.05	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	0	0
9	0.063	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0
11	0.188	0	0	0	0	0	0	0	0	0	0
12	0.357	0	0	0	0	0	0	0	0	0	0
13	0.182	0	0	0	0	0	0	0	0	0	0
14	1	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0.105	0	0	0	0	0	0
16	0.417	0	0	0	0	0	0	0	0	0	0
17	0.1	0	0.05	0	0	0	0	0	0	0	0
18	0.05	0	0	0	0.05	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0.095	0	0	0	0	0	0
21	0.026	0	0.026	0	0.132	0	0	0	0	0	0
22	0	0	0	0	0.056	0	0	0	0	0	0
23	0	0	0	0	0.033	0	0	0	0	0	0
24	0.036	0	0.071	0	0	0	0	0	0	0	0
25	0.019	0	0	0	0	0	0	0	0	0	0
26	0.033	0	0	0	0.033	0	0	0	0	0	0
27	0.125	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0.118	0.059	0	0	0	0	0
29	0.054	0	0	0	0.125	0.036	0	0	0	0	0
30	0.306	0	0	0	0.306	0	0.028	0	0	0	0
31	0.466	0	0	0	0.293	0	0	0	0	0	0
32	1	0	0	0	0	0	0	0.071	0	0	0
33	1	0	0	0	0	0	0	0.083	0	0	0
34	1	0	0	0	0	0	0	0.205	0	0	0
35	1	0	0	0	0	0	0	0.2	0	0	0
36	1	0	0	0	0	0	0	0.083	0	0	0
37	1	0	0	0	0	0	0	0.159	0	0	0
38	1	0	0	0	0	0	0	0.136	0	0	0
39	1	0	0	0	0	0	0	0.167	0	0	0
40	1	0	0	0	0	0	0	0.188	0	0	0
41	1	0	0	0	0	0	0	0.2	0	0	0
42	1	0	0	0	0	0	0	0.04	0	0	0
43	1	0	0	0	0	0	0.012	0.036	0.012	0	0
44	0.875	0	0.125	0	0	0	0	0	0	0	0
45	1	0	0	0	0	0	0	0	0	0	0
46	0.446	0	0	0	0	0	0	0	0	0	0.214
47	0.543	0	0	0	0	0	0	0	0	0	0.261
48	0.321	0	0	0	0	0	0	0	0	0	0.184

49	0.613	0	0	0	0	0	0	0	0	0.016	0.145
50	0.523	0	0	0	0	0	0	0	0	0	0.176
51	0.344	0	0	0	0.063	0	0	0	0	0	0
52	0.133	0	0	0	0	0	0	0	0	0	0.017
53	1	0	0	0	0	0	0	0	0	0	0.071
54	1	0	0	0	0	0	0	0	0	0	0
55	1	0	0	0	0	0	0	0	0	0	0
56	1	0	0	0	0	0	0	0	0	0	0
57	1	0	0	0	0	0	0	0	0	0	0
58	1	0	0	0	0	0	0	0	0	0	0
59	1	0	0	0	0	0	0	0	0	0	0
60	0.979	0	0	0	0	0	0	0	0	0	0
61	1	0	0	0	0	0	0	0	0	0	0
62	1	0	0	0	0	0	0	0	0	0	0
63	1	0	0	0	0	0	0	0	0	0	0
64	1	0	0	0	0	0	0	0	0	0	0
65	0	0	0	0.563	0	0	0	0	0	0	0
66	0	0	0	0.31	0	0	0	0	0	0	0

Table 3.2 (continued)

Chromosome two inversions

Sample Number	2L-A	2L-7	2L-D	2L-X	2L-S	2L-W	2L-B	2S-7	2S-6b
1	0	0.9	0	0	0	0	0	0	0.167
2	0	0.714	0	0	0	0	0	0	0.25
3	0	0.091	0.364	0	0	0	0	0	0.125
4	0	0.05	0.45	0	0	0	0	0	0.05
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0.938	0.938	0	0	0	0	0.625	0
10	0	1	1	0	0	0	0	0.154	0
11	0	1	1	0	0	0	0	0.25	0
12	0	0.857	1	0	0	0	0	0.071	0
13	0	0.864	1	0	0	0	0	0.227	0
14	0	0.22	1	0	0	0	0	0.02	0
15	0	0.762	0.762	0.238	0	0	0	0	0
16	0	0.833	0.833	0.167	0	0	0	0	0
17	0	0.636	0.636	0.364	0	0	0	0	0
18	0	0.85	0.85	0.15	0	0	0	0	0
19	0	0.75	0.75	0.25	0	0	0	0	0
20	0	0.826	0.826	0.174	0	0	0	0	0
21	0	0.833	0.833	0.167	0	0	0	0	0
22	0	0.909	0.909	0.091	0	0	0	0	0
23	0	0.895	0.895	0.105	0	0	0	0	0
24	0	0.794	0.794	0.206	0	0	0	0.912	0
25	0	0.821	0.821	0.179	0	0	0	0.931	0
26	0	0.875	0.875	0.125	0	0	0	1	0
27	0	0.7	0.7	0.3	0	0	0	0.875	0
28	0	0.65	0.65	0.35	0	0	0	1	0
29	0	0.683	0.683	0.317	0	0	0	0.774	0
30	0	0.25	0.25	0.75	0	0	0	0	0
31	0	0.328	0.328	0.672	0	0	0	0	0
32	0	1	1	0	0	0.018	0	1	0
33	0	1	1	0	0	0	0.022	1	0
34	0	1	1	0	0	0	0	1	0
35	0	1	1	0	0	0	0	1	0
36	0	1	1	0	0	0	0.042	1	0
37	0	1	1	0	0	0	0	1	0
38	0	1	1	0	0	0	0.023	1	0
39	0	1	1	0	0	0	0	1	0
40	0	1	1	0	0	0	0	1	0
41	0	1	1	0	0	0	0.05	1	0
42	0	1	1	0	0	0	0	1	0
43	0	1	1	0	0	0	0	1	0
44	0	1	1	0	0	0	0	1	0
45	0	1	1	0	0	0	0	1	0
46	1	1	0	0	0	0	0	0	0
47	1	1	0	0	0	0	0	0	0
48	1	1	0	0	0	0	0	0	0

49	1	1	0	0	0	0	0	0	0
50	1	1	0	0	0	0	0	0	0
51	1	1	0	0	0	0	0	0	0
52	1	0.983	0	0	0	0	0	0	0
53	1	0.714	0	0	0	0	0	0	0
54	1	0	0	0	0	0	0	0	0
55	1	0	0	0	0	0	0	0	0
56	1	0	0	0	0	0	0	0	0
57	1	1	0	0	0	0	0	0	0
58	1	1	0	0	0	0	0	0	0
59	1	0.833	0	0	0	0	0	0	0
60	1	0.854	0	0	0	0	0	0	0
61	1	1	0	0	0	0	0	0	0
62	1	0.917	0	0	0	0	0	0	0
63	1	0	0	0	0	0	0	0	0
64	1	0	0	0	0	0	0	0	0
65	1	1	0	0	0	0	0	0	0
66	1	0.976	0	0	0.024	0	0	0	0

Table 3.2 (continued)

Chromosome three inversions

Sample Number	3L-A	3L-24	3L-B	3L-25	3L-26	3L-5	3L-E	3L-I	3L-G	3L-Y
1	0	0.25	0	0	0	0	0	0	0	0
2	0	0.357	0	0	0	0	0	0	0	0
3	0	0.75	0	0	0	0	0	0	0	0
4	0	0.75	0	0	0	0	0	0	0	0
5	0	0.583	0	0	0	0	0	0	0	0
6	0	0	0	0.2	0.15	0	0	0	0	0
7	0	0	0.167	0	0	0	0	0	0	0
8	0	0	0.125	0	0	0	0	0	0	0
9	0	0	0	0	0	0.313	0	0	0	0
10	0	0	0	0	0	0.731	0.038	0.038	0.077	0
11	0	0	0	0	0	0.625	0.063	0	0	0
12	0	0	0	0	0	0.714	0	0	0	0
13	0	0	0	0	0	0.72	0	0	0	0
14	0	0	0	0	0	0.72	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0.111
18	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0.069	0	0
26	0	0	0	0	0	0	0	0.029	0	0
27	0	0	0	0	0	0	0	0.033	0	0
28	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0.	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0.02	0	0
43	0	0	0	0	0	0	0	0.024	0	0
44	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0

49	0	0.016	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0
51	0	0	0.563	0	0	0	0	0	0	0
52	0	0	0.7	0	0	0	0	0	0	0
53	0	0	0.714	0	0	0	0	0	0	0
54	0.022	0	0.911	0	0	0	0	0	0	0
55	0	0	0.857	0	0	0	0	0	0	0
56	0	0	0.893	0	0	0	0	0	0	0
57	0	0	0.267	0	0	0	0	0	0	0
58	0	0	0.643	0	0	0	0	0	0	0
59	0	0	0	0	0	0	0	0.056	0	0
60	0.021	0	0	0	0	0	0	0	0	0
61	0	0	0	0	0	0	0	0	0	0
62	0	0	0	0	0	0	0	0	0	0
63	0	0	0.917	0	0	0	0	0.042	0	0
64	0	0	1	0	0	0	0	0	0	0
65	0	0	0	0	0	0	0	0	0	0
66	0	0	0	0	0	0	0	0	0	0

Table 3.2 (continued)

Chromosome three inversions (continued)

Sample Number	3L-X	3L-2	3L-4	3L-17	3L-D	3L-K	3L-M
1	0	1	1	1	0	0	0
2	0	1	1	1	0	0	0
3	0	1	0.958	0.958	0	0	0
4	0	1	1	1	0	0	0
5	0	1	1	1	0	0	0
6	0	1	1	1	0	0	0
7	0	1	1	1	0	0	0
8	0	1	1	1	0	0	0
9	0	1	0.75	0.75	0	0	0
10	0	1	0.962	0.962	0	0	0
11	0	1	0.937	0.937	0	0	0
12	0	1	1	1	0	0	0
13	0	1	0.94	0.94	0	0	0
14	0	1	0.94	0.94	0	0	0
15	0	1	0.286	0.286	0	0	0
16	0	1	0.333	0.333	0	0	0
17	0	1	0.389	0.389	0	0	0
18	0.05	0.95	0.1	0.1	0	0	0
19	0.313	1	0.188	0.188	0	0	0
20	0	1	0.275	0.275	0	0	0
21	0	1	0.31	0.31	0	0	0
22	0	1	0.45	0.45	0	0	0
23	0	1	0.313	0.313	0	0	0
24	0.813	0.125	0.125	0.125	0	0	0
25	0.741	0.258	0.258	0.258	0	0	0
26	0.765	0.118	0.118	0.118	0	0	0
27	0.733	0.267	0.267	0.267	0	0	0
28	0.7	0.3	0.3	0.3	0	0	0
29	0.839	0.161	0.143	0.143	0	0	0
30	0.636	0.364	0.227	0.227	0	0	0
31	0.672	0.034	0.293	0.293	0	0	0
32	0	1	0.42	0.38	0.054	0	0
33	0	1	0.286	0.238	0.022	0	0
34	0	1	0.357	0.3	0.013	0	0
35	0	1	0.357	0.321	0.033	0	0
36	0	1	0.278	0.278	0.083	0	0
37	0	1	0.333	0.357	0.045	0	0
38	0	1	0.406	0.437	0.068	0	0
39	0	1	0.25	0.187	0.013	0	0
40	0	1	0.4	0.4	0.188	0	0
41	0	1	0.55	0.55	0.1	0	0
42	0	1	0.4	0.38	0	0	0
43	0	1	0.342	0.329	0	0.012	0
44	0	1	0.125	0.125	0	0	0
45	0	1	0.423	0.423	0	0	0
46	0	1	0.982	0.982	0	0	0
47	0	1	0.935	0.935	0	0	0
48	0	1	1	1	0	0	0

49	0	1	1	1	0	0	0
50	0	1	1	1	0	0	0
51	0	1	0.937	0.937	0	0	0
52	0	1	1	1	0	0	0.017
53	0	1	1	1	0	0	0
54	0	1	0.911	0.911	0	0	0
55	0	1	0.857	0.857	0	0	0
56	0	1	0.893	0.893	0	0	0
57	0	1	1	1	0	0	0
58	0	1	1	1	0	0	0
59	0	1	0.944	0.944	0	0	0
60	0	1	0.958	0.958	0	0	0
61	0	1	1	1	0	0	0
62	0	1	1	1	0	0	0
63	0	1	0.917	0.917	0	0	0
64	0	1	1	1	0	0	0
65	0	1	1	1	0	0	0
66	0	1	0.976	0.976	0	0	0

Table 3.3

First Two Principal Components of the Covariance Matrix

Inversion	PC I	PC II
1L-B	-0.016575	-0.026580
1L-C	-0.015252	-0.027132
1L-D	0.000195	-0.000967
1L-U	-0.000391	0.000290
1L-G	-0.000063	0.000068
1L-N	-0.000520	0.000399
1L-X	0.000189	-0.000742
1L-A	0.313813	0.502275
1L-R	-0.000236	0.000197
1L-T	0.000314	0.000453
1S-A	-0.080477	0.662220
1S-G.H	-0.010888	0.006406
1S-J	0.003284	-0.000580
1S-F	-0.015212	-0.014282
1S-B	0.010259	-0.052903
1S-X	0.001652	-0.003870
1S-N	0.000387	-0.000829
1S-C	0.034813	0.057303
1S-M	0.000270	0.000430
1S-P	-0.000291	0.000056
1S-21	-0.019286	0.000575
2L-A	-0.420747	0.146415
2L-7	0.170603	0.046396
2L-D	0.464544	-0.001129
2L-X	0.046113	-0.174524
2L-S	-0.000409	-0.000378
2L-W	0.000390	0.000661
2L-B	0.002974	0.005057
2S-7	0.425564	0.260556
2S-6b	-0.006319	0.000757
3L-A	-0.000921	0.000646
3L-24	-0.030320	0.013454
3L-B	-0.173934	0.098074
3L-25	-0.002764	0.001285
3L-26	-0.002073	0.000964
3L-5	0.005339	-0.049871
3L-E	0.000280	-0.001812
3L-I	0.001531	-0.002783
3L-G	0.000172	-0.001713
3L-Y	0.000472	-0.003300
3L-X	0.096833	-0.255164
3L-2	-0.099292	0.263403
3L-4	-0.341389	0.130582
3L-17	-0.346537	0.122688
3L-D	0.013501	0.022850
3L-K	0.000270	0.000430
3L-M	-0.000337	-0.000135

Table 3.4

Distortion measures for each of the hierarchical cluster methods, and the partition level suggested by the cubic clustering criterion for all methods.

Cluster Method ion	Distortion Measure				CCC Parti-
	r(cop)/	Δ(0)	Δ(1)	Δ(2)	
Single Linkage	0.7761	0.807	0.501	0.5412	12
Complete Linkage	0.7898	0.8811	0.8384	0.8355	13
Group Average	0.8428	0.4415	0.2588	0.2119	14
Ward's Method	0.7968	-	-	-	14
Iterative Relocation	n.a	n.a	n.a	n.a	13

Table 3.5

Classification results for the five cluster methods on the full data set; partition level set by the cubic clustering criterion; Numbers refer to the samples listed in Table 3.1.

a). Single linkage 12 cluster solution:

Cluster Number	Sample Number	*a priori* taxon
Cluster 1	1 2	Beffa
Cluster 2	3	Beffa
Cluster 3	4,...,8	Beffa/Chutes Milo
Cluster 4	9	Menankaya
Cluster 5	10,...,13	Menankaya
Cluster 6	14	Menankaya
Cluster 7	15,...,23	Konkoure/Menankaya
Cluster 8	24,...,29	Konkoure/Menankaya
Cluster 9	30,31	Konkoure/Menankaya
Cluster 10	32,...,45	*S. soubrense* 'B'
Cluster 11	46,...,53	*S. sanctipauli*
	57,...,62,65,66	*S. sanctipauli*
Cluster 12	54,55,56,63,64	*S. sanctipauli*

b). Furthest neighbour 13 cluster solution:

Cluster Number	Sample Number	*a priori* taxon
Cluster 1	1 2	Beffa
Cluster 2	3,4,5	Beffa
Cluster 3	6,7,8	Chutes Milo
Cluster 4	9	Menankaya
Cluster 5	10,...,13	Menankaya
Cluster 6	14	Menankaya
Cluster 7	15,...,23	Konkoure/Menankaya
Cluster 8	24,...,29	Konkoure/Menankaya
Cluster 9	30,31	Konkoure/Menankaya
Cluster 10	32,...,45	*S. soubrense* 'B'
Cluster 11	46,...,53	*S. sanctipauli*
	57,...,62	
Cluster 12	54,55,56,63,64	*S. sanctipauli*
Cluster 13	65,66	*S. sanctipauli*

c). Group Average 14 cluster solution:

Cluster Number	Sample Number	*a priori* taxon
Cluster 1	1 2	Beffa
Cluster 2	3	Beffa
Cluster 3	4,...,8	Beffa/Chutes Milo
Cluster 4	9	Menankaya
Cluster 5	10,...,13	Menankaya
Cluster 6	14	Menankaya
Cluster 7	15,...,23	Konkoure/Menankaya
Cluster 8	24,...,29	Konkoure/Menankaya
Cluster 9	30,31	Konkoure/Menankaya
Cluster 10	32,...,45	*S. soubrense* 'B'
Cluster 11	46,...,50,53, 57,...,62	*S. sanctipauli*
Cluster 12	51,52	*S. sanctipauli*
Cluster 13	54,55,56,63,64	*S. sanctipauli*
Cluster 14	65,66	*S. sanctipauli*

d). Ward's Method 14 cluster solution:

Cluster Number	Sample Number	*a priori* taxon
Cluster 1	1 2	Beffa
Cluster 2	3,4,5	Beffa/Chutes Milo
Cluster 3	6,7,8	Chutes Milo
Cluster 4	9	Menankaya
Cluster 5	10,...,13	Menankaya
Cluster 6	14	Menankaya
Cluster 7	15,...,23	Konkoure/Menankaya
Cluster 8	24,...,29	Konkoure/Menankaya
Cluster 9	30,31	Konkoure/Menankaya
Cluster 10	32,...,45	*S. soubrense* 'B'
Cluster 11	46,...,52	*S. sanctipauli*
Cluster 12	53,57,...,62	*S. sanctipauli*
Cluster 13	54,55,56,63,64	*S. sanctipauli*
Cluster 14	65,66	*S. sanctipauli*

e). Iterative Relocation 13 cluster solution:

Cluster Number	Sample Number	*a priori* taxon
Cluster 1	1 2	Beffa
Cluster 2	3,4,5	Beffa/Chutes Milo
Cluster 3	6,7,8	Chutes Milo
Cluster 4	9	Menankaya
Cluster 5	10,...,13	Menankaya
Cluster 6	14	Menankaya
Cluster 7	15,...,23	Konkoure/Menankaya
Cluster 8	24,...,29	Konkoure/Menankaya
Cluster 9	30,31	Konkoure/Menankaya
Cluster 10	32,...,45	*S. soubrense* 'B'
Cluster 11	46,...,52 57,...,62	*S. sanctipauli*
Cluster 12	53,...,56,63,64	*S. sanctipauli*
Cluster 13	65,66	*S. sanctipauli*

Table 3.6

Consensus classification[1]

Cluster Number	Sample Number	*a priori* taxon
Cluster 1	1,2	Beffa
Cluster 2	3	Beffa
Cluster 3	6,7,8	Chutes Milo
Cluster 4	9	Menankaya
Cluster 5	10,...,13	Menankaya
Cluster 6	14	Menankaya
Cluster 7	15,...,23	Konkoure/Menankaya
Cluster 8	24,...,29	Konkoure/Menankaya
Cluster 9	30,31	Konkoure/Menankaya
Cluster 10	32,...,45	*S. soubrense* 'B'
Cluster 11	46,...,50	*S. sanctipauli* 'Djodji'
Cluster 12	54,...,56,63,64	*S. sanctipauli*
Cluster 13	57,59,...,62	*S. sanctipauli*

[1] Numbers refer to the sample numbers given in Table 3.1, if a specific name is not given for the *a priori* taxon, then the species is *S. soubrense*.

Samples not included in the consensus classification:
{4,5,51,52,53,58,65,66}

Table 3.7

Effect of the B_k methods on the distortion measure $\Delta_{0.5}$ for each *a priori* defined taxon.

a priori taxon	B(k=)			
	1	2	3	4
S. soubrense 'B'	0.6983	0.6588	0.6019	0.5153
S. sanctipauli	0.7408	0.5233	0.4227	0.4008
S. soubrense Menankaya/Konkoure	0.5696	0.529	0.4563	0.4332

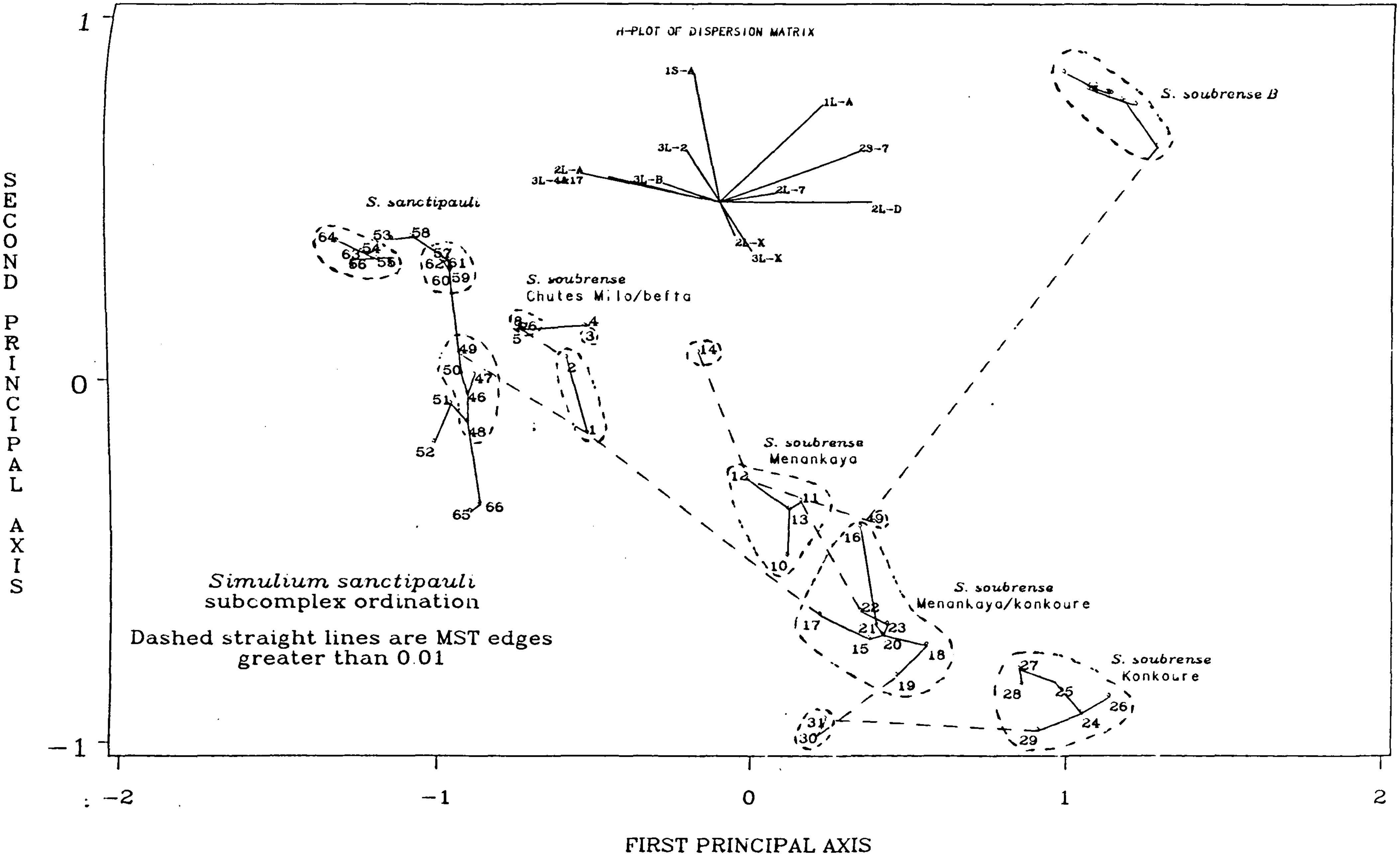

H-PLOT OF DISPERSION MATRIX
1S-A
1L-A
3L-2
2S-7
2L-A
3L-4&17
3L-B
2L-7
2L-D
2L-X
3L-X
S. soubrense B
S. sanctipauli
64
53
58
63
54
56
55
57
62
61
60
59
S. soubrense
Chutes Milo/befta
8
6
4
5
3
49
14
2
50
47
0
51
46
1
48
52
S. soubrense
Menankaya
12
11
13
65
66
49
16
10
Simulium sanctipauli
subcomplex ordination
Dashed straight lines are MST edges
greater than 0.01
S. soubrense
Menankaya/konkoure
22
17
21
23
15
20
18
19
S. soubrense
Konkoure
27
28
25
26
31
24
30
29
1
−1
−2
−1
0
1
2
SECOND PRINCIPAL AXIS
FIRST PRINCIPAL AXIS

Figure 3.1

Figure 3.2
Hierarchical Cluster Analysis using Single Linkage

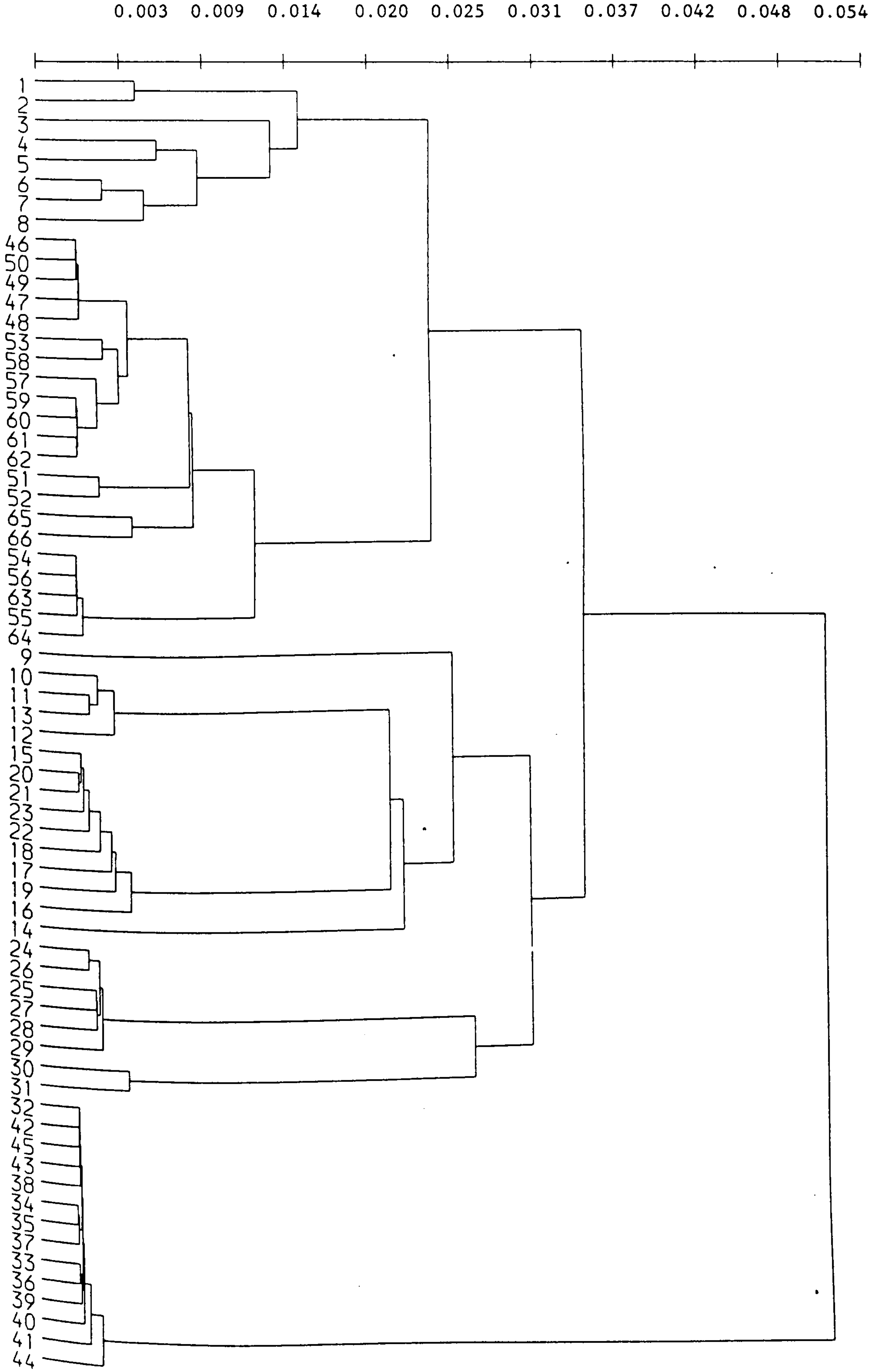


Figure 3.3
Hierarchical Cluster Analysis using Complete Linkage

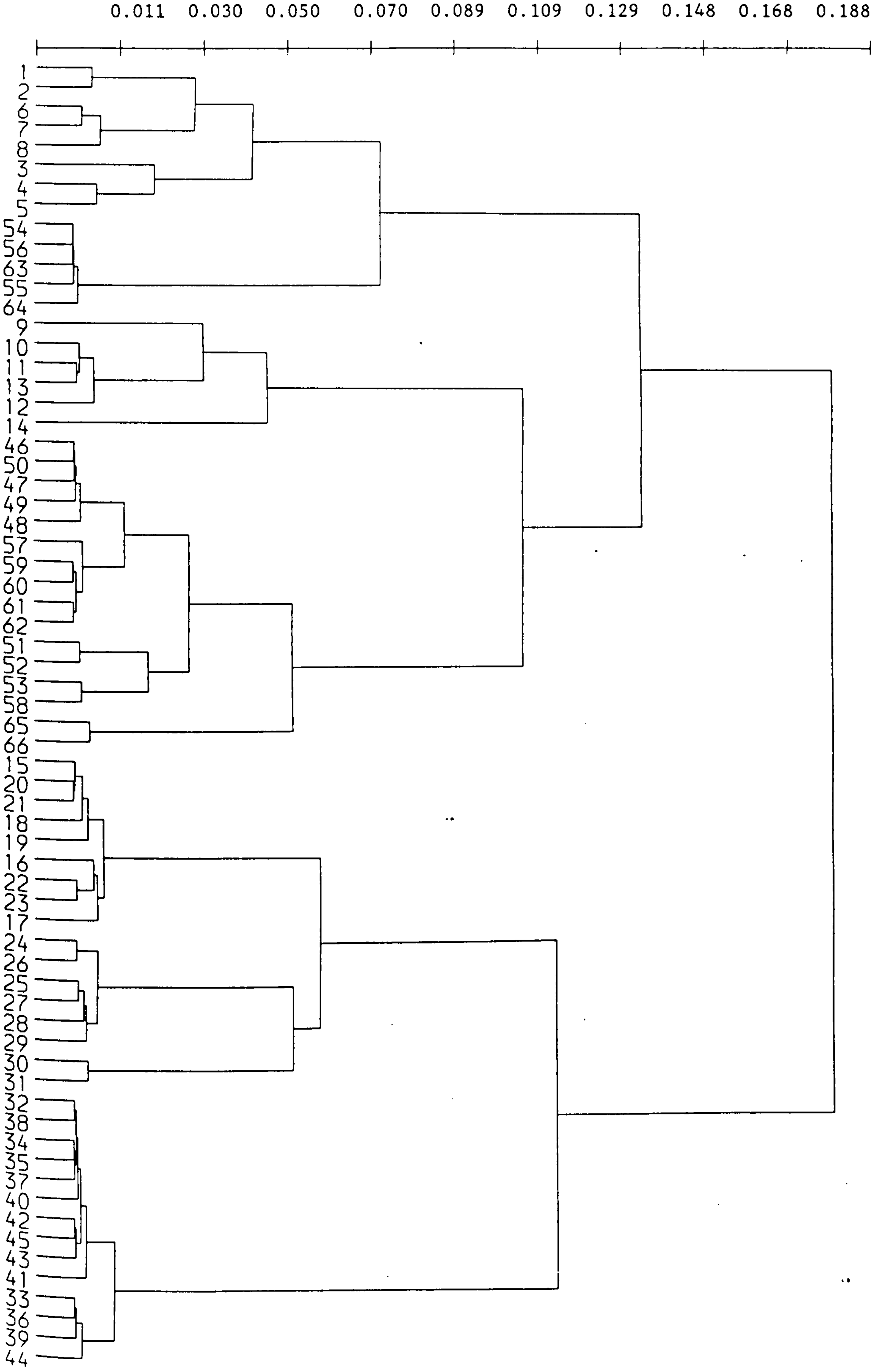


Figure 3.4
Hierarchical Cluster Analysis using Average Linkage

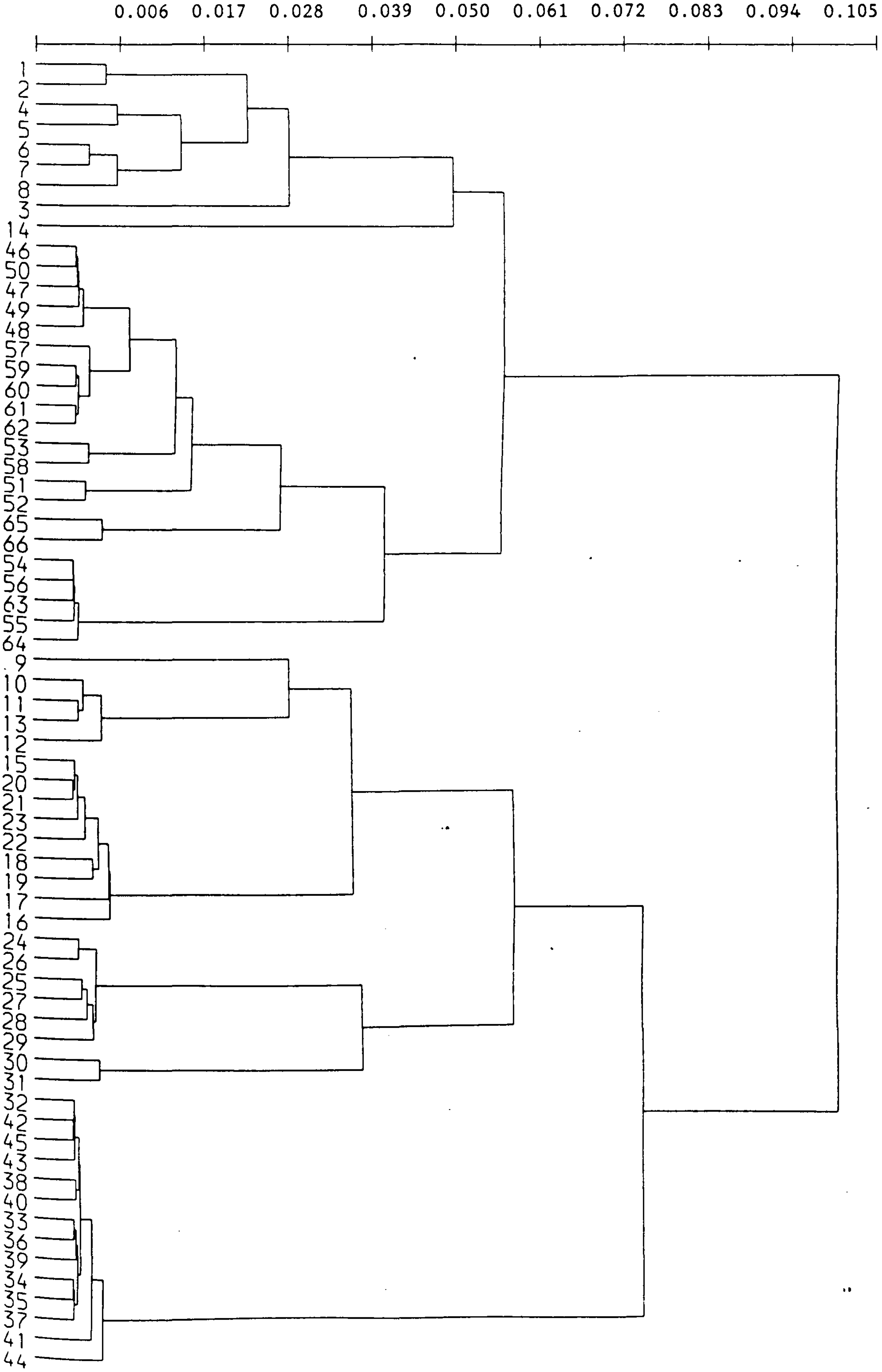
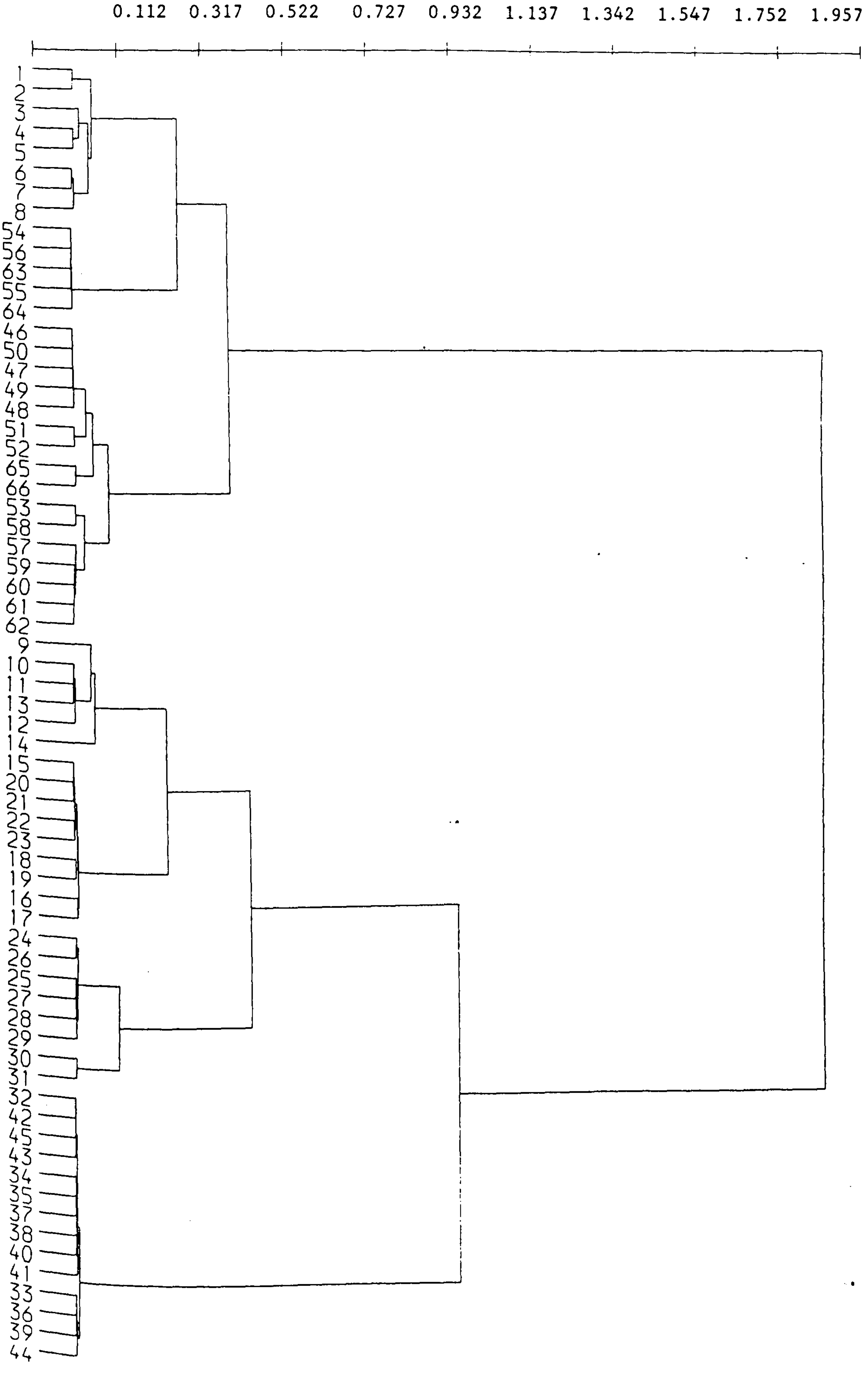


Figure 3.5
Hierarchical Cluster Analysis using Ward's Method



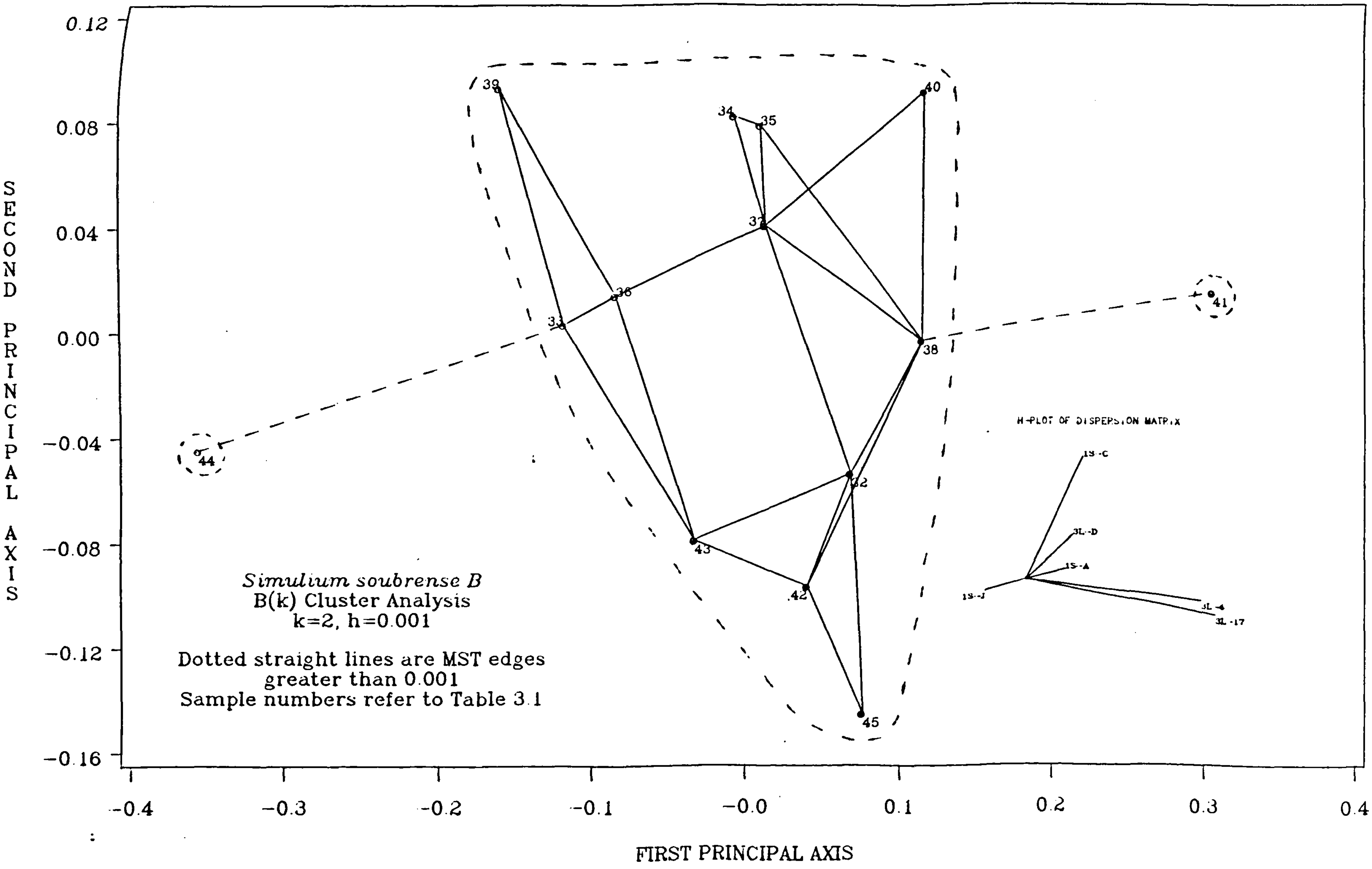

0.12
0.08
0.04
0.00
−0.04
−0.08
−0.12
−0.16
SECOND PRINCIPAL AXIS
-0.4
-0.3
-0.2
-0.1
-0.0
0.1
0.2
0.3
0.4
FIRST PRINCIPAL AXIS
39
34
35
40
37
36
33
41
38
44
32
43
42
45
H-PLOT OF DISPERSION MATRIX
1S--C
3L--D
1S--A
1S--J
3L-4
3L-17
Simulium soubrense B
B(k) Cluster Analysis
k=2, h=0.001
Dotted straight lines are MST edges
greater than 0.001
Sample numbers refer to Table 3.1

Figure 3.6

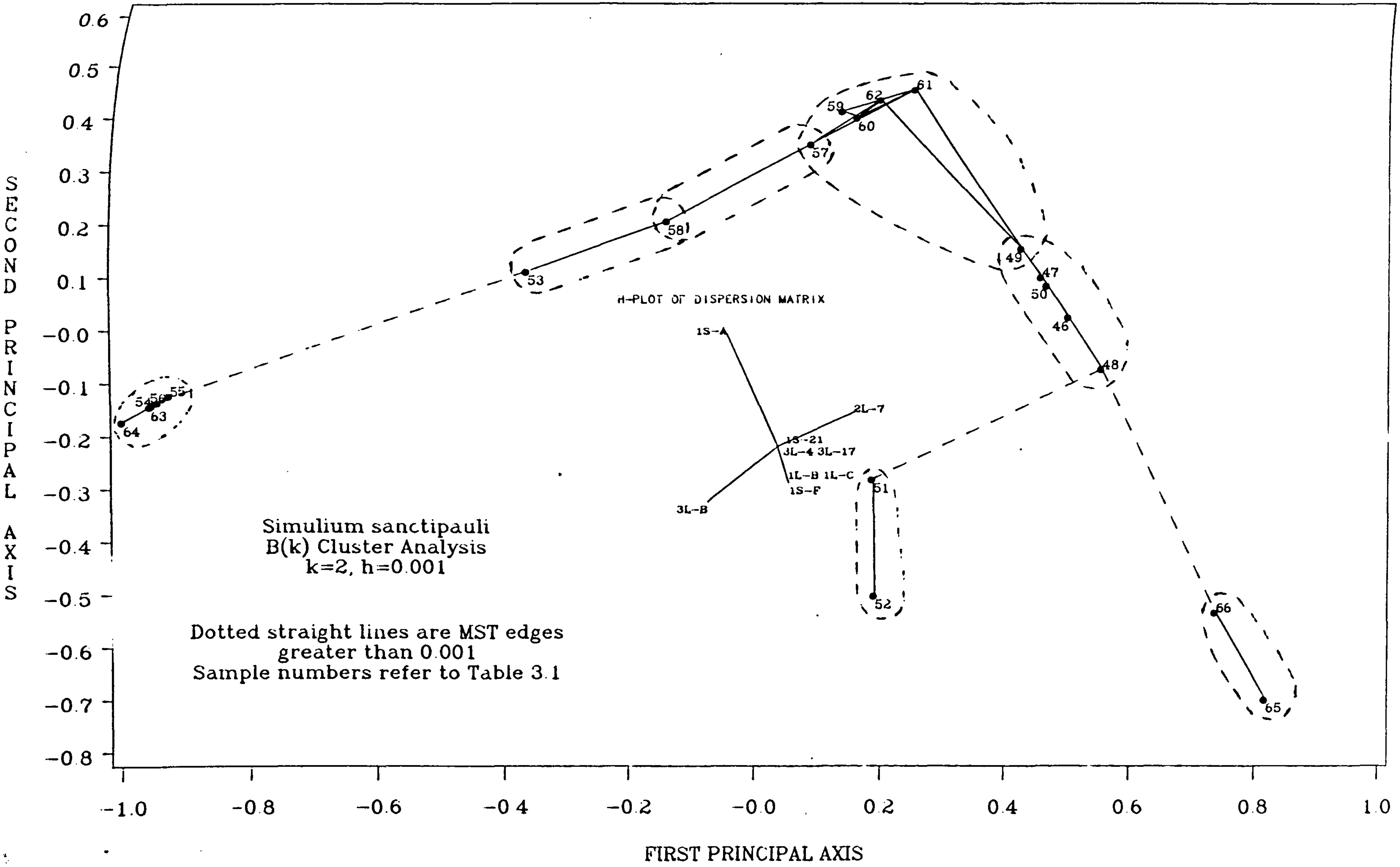

Simulium sanctipauli
B(k) Cluster Analysis
k=2, h=0.001
Dotted straight lines are MST edges
greater than 0.001
Sample numbers refer to Table 3.1
H-PLOT OF DISPERSION MATRIX
1S-A
2L-7
1S-21
3L-4 3L-17
1L-B 1L-C
1S-F
3L-B
SECOND PRINCIPAL AXIS
FIRST PRINCIPAL AXIS
0.6
0.5
0.4
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-1.0
-0.8
-0.6
-0.4
-0.2
-0.0
0.2
0.4
0.6
0.8
1.0
64
54
56
63
55
53
58
57
59
60
62
61
49
47
50
46
48
51
52
66
65

Figure 3.7

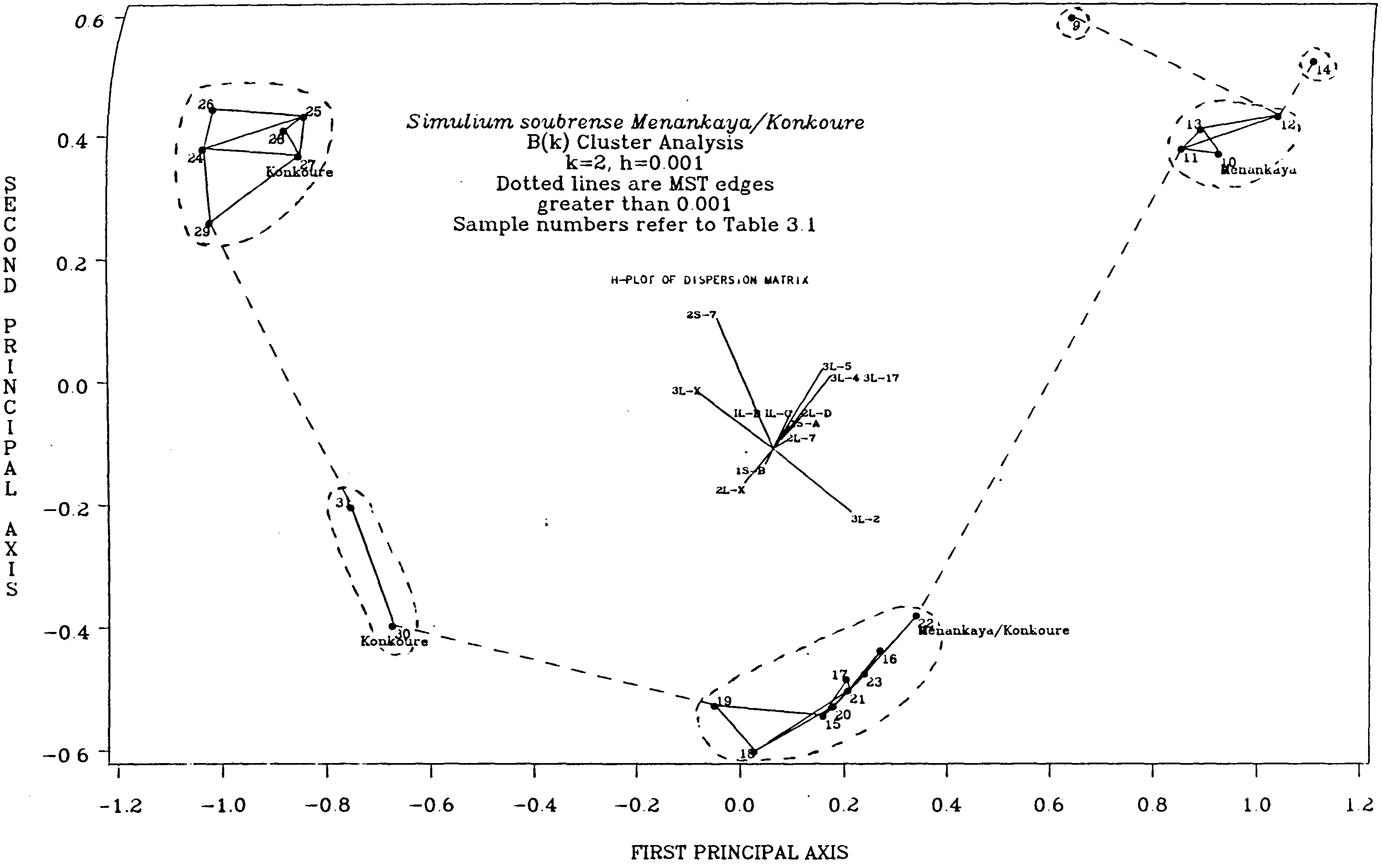

Simulium soubrense Menankaya/Konkoure
B(k) Cluster Analysis
k=2, h=0.001
Dotted lines are MST edges
greater than 0.001
Sample numbers refer to Table 3.1
H-PLOT OF DISPERSION MATRIX
2S-7
3L-5
3L-4 3L-17
3L-X
IL-B IL-C 2L-D
S-A
2L-7
1S-B
2L-X
3L-2
26
25
28
24
27
Konkoure
29
31
30
Konkoure
9
14
13
12
11
10
Menankaya
22
Menankaya/Konkoure
16
17
23
21
19
20
15
18
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
SECOND PRINCIPAL AXIS
−1.2
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
FIRST PRINCIPAL AXIS

Figure 3.8

CHAPTER FOUR: ADULT FEMALE MORPHOMETRIC CHARACTERS

4.1 INTRODUCTION

The overall aim of the multivariate morphometric analysis of the *S. damnosum* complex is to provide a scheme for allocating an adult female *S. damnosum s.l.* to its correct cytospecies (Chapter one). The statistics used in this allocation scheme are developed in Chapters seven and eight, whilst Chapter nine describes the protocol to be followed for the actual allocation.

The purpose of the present chapter is to describe the morphometric characters which were used in the analyses following this chapter, and to describe how the characters were measured. The simple statistics used to describe each character were derived from the total data matrix following the procedures described in Chapter five.

4.2 MATERIALS AND METHODS

4.2.1 MATERIALS

Appendix one gives the full details of each sample used in the multivariate morphometric analyses presented in Chapters six, seven and eight. The initial sample size, and the final sample size following the methods given in Chapter five are shown.

4.2.2 METHODS

Appendix one describes the different sources of the samples of adult female *S. damnosum s.l.* available in this analysis. The females were either reared from pupae (which were collected attached to substrate at breeding sites and maintained at room temperature until eclosion), with or without correlated cytotaxonomic identification (for cytotaxonomic methods see Chapter two), or caught at human bait. Each whole sample was then stored either in 95% propanol or in 70% ethanol, depending on the collector. The effect that storage in different media has on morphology was not evaluated in this analysis, but would certainly be worthy of future investigation.

Each adult female was examined in a culture dish in the same medium as it was preserved in, using a binocular microscope at 40X magnification. If the fly was missing for any of the characters described in the next section, it was discarded.

Four characters, thorax length, thorax width, clypeus width and head width, were measured at this stage, and the abdominal setal colouration recorded. In order to keep the fly in the correct plane, a plastic ring was glued to the base of the culture dish and the fly measured on this using an eyepiece graticule (Beech-Garwood *et al.* in the press). The head was removed for ease of measurement. The fly was then dissected, and the left antenna, left fore and mid legs, and the left wing mounted on a microscope slide in a small drop of Berlese mountant. After careful manipulation of the insect parts to ensure that all of the characters were clearly visible and not distorted, a cover slip was lowered at an angle onto the mounted parts. Only the left side of the insect was used to ensure that any variation due to systematic asymmetry of the flies was not detected.

The slide-mounted insect was then examined using a compound microscope at 32X, 100X, 250X and 400X magnification. The magnification for a particular character was used so that the character filled a large proportion of the eyepiece graticule. Measurements were recorded in graticule units on a standard record sheet, which contained details of the sample from which the fly came, and a code number. The slide was marked with a code number and stored. The remainder of the fly was stored individually in the same medium as was used for the original sample.

At regular intervals during the data collection, data were typed into an IBM 3083 mainframe computer at the University of Liverpool Computer Laboratory. It was important that too much data were not input at any one time because the primary source of problems in the data derived from operator error (Chapter five). The data were edited in a temporary data set and carefully checked against the original record sheets. Once errors were corrected the temporary data set was appended to the main data set. The data were input in graticule units and a SAS data step program used to convert these units into microns, according to the calibration factor for each magnification, which was calculated from a standard stage micrometer slide.

4.2.3 ADULT FEMALE CHARACTER CHOICE

The choice of characters and the number of characters to be chosen is a problematic aspect of applied multivariate morphometric analysis (Blackith and Reyment 1971, p32). For technical reasons, the characters in this analysis were restricted to colour characters, counts and continuous linear characters. More sophisticated methods of direct shape assessment such as finite element scaling (Cherverud and Archie 1984) or Fourier methods (Rohlf and Archie 1984) could not be attempted. The problem of homology of characters does not usually arise in a closely related species complex, as it can in higher level taxa (Blackith and Reyment 1971). Thus a potentially enormous number of characters describing variation within and between members of the *S. damnosum* complex were available for measurement. Ideally both a very large character set and very large sample sizes should be obtained in a multivariate morphometric analysis, so that no important character combinations are missed, and the estimation of parameters is accurate. In practise this was not possible, and a balance had to be struck between the number of characters and sample size.

Other practical considerations which were taken into account when choosing the initial characters were that the characters should:

1. be robust, i.e. not easily lost from flies,

2. be well defined, e.g. having a well defined landmark at each end of the body part;

3. be easily measured; i.e. a character was not included if it involved a large amount of time in dissection.

Besides these practical and statistical considerations, characters were also chosen which had been shown to have taxonomic value by earlier authors (Chapter one). Overall, a 28 eight character set was initially derived, which are individually described in section 4.3.

4.3 DESCRIPTION OF ADULT FEMALE CHARACTERS

An initial set of 28 characters was scored or measured for each individual fly. These characters were chosen according to the general principles described in section 4.2.3. The simple statistics reported for each character are those resulting from the total data matrix following the screening methods described in Chapter five. The relative taxonomic importance of these characters is assessed in Chapters six to nine. The characters are illustrated in Figures 4.1 to 4.7.

THORAX MEASUREMENTS

Two thorax characters were measured, length and width. It would have been preferable to have included a measure of thorax depth, so that the whole thorax volume was measured, but clearly defined landmarks were not available to make such a measurement reliable. In the past, thorax length has been used as a measure of overall size (Garms 1978), however it is unlikely that a single character can be wholly representative of size, and the thorax measurements were chosen in this analysis as characters in themselves.

4.3.1 THORAX LENGTH (V3)

The length of the thorax was measured from the posterior of the anterior thoracic spiracle to the middle of the posterior thoracic spiracle (Figure 4.1). This measure differs from the usual thorax length measurement (Garms 1978) which is the length from the anterior margin of the thorax to the posterior margin of the scutellum. This new measure was used because of the well-defined landmarks, which increases the accuracy of the measurement, although this new measure is smaller than the older measure, which in Garms (1978) ranged from 800µm to 1300µm. Thorax length measured in the old way is known to be a useful taxonomic character (Garms 1977), as a measure of size.

Mean	635.87 µm
Standard error of the mean	1.882
Standard deviation	53.243
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.207

4.3.2 THORAX WIDTH (V4)

The width of the thorax was measured across the greatest width of the scutum (Figure 4.2) This character has not been used in the morphometrics of the *S. damnosum* complex, and was chosen for its clearly defined landmarks.

Mean	882.08 µm
Standard error of the mean	2.3011
Standard deviation	65.0856
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.069

Figure 4.1

Thorax Length

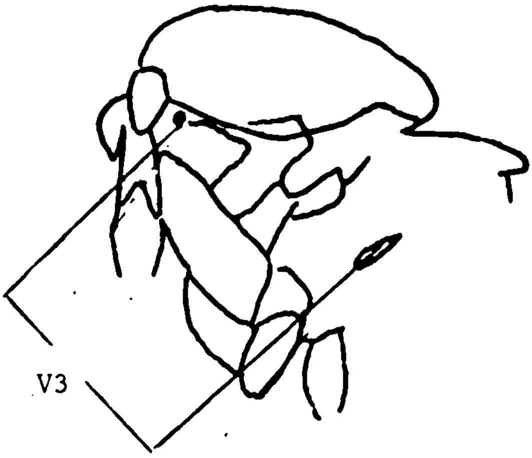
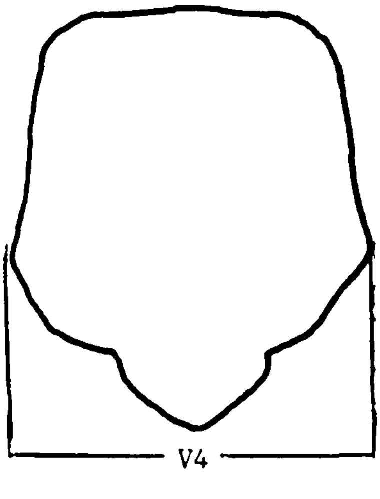


Figure 4.2

Thorax Width



HEAD MEASUREMENTS

Two head characters were measured, clypeus width and head width. Only these two head characters were measured because other characters did not have clearly defined landmarks. The number of maxillary teeth was not included in the analysis because it was an awkward character to score, and because Townson and Meredith (1979) and Garms (1978) both showed that it was not useful in distinguishing *S. yahense* form *S. squamosum*, as had been found by Quillévéré and Sechan (1978). The palpal segments were not measured because they were difficult to dissect and were not linear.

4.3.3 CLYPEUS WIDTH (V5)

The width of the clypeus was measured between the extreme lateral points, just ventral to the antennae (Figure 4.3). The large proportion of samples for which the null hypothesis of normality was rejected was due to the low resolving power at 40X magnification for this small character. The sample distribution, whilst it was usually symmetrical was always stepped, due to rounding up or down of eyepiece graticule units. This character was therefore treated as being of questionable value prior to the analyses described in Chapters six to eight.

Mean	211.798 µm
Standard error of the mean	0.5628
Standard deviation	15.919
Proportion of samples for which the null hypothesis of normality was rejected, $\alpha=0.05$	0.862

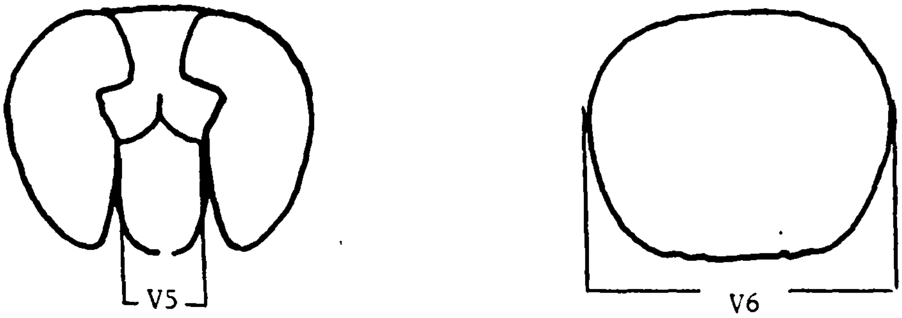
4.3.4 HEAD WIDTH (V6)

The width of the head was measured from the extreme lateral margins, viewed in posterior aspect (Figure 4.3). Anon (1976) measured head width in *S. damnosum s.l.* from Togo (Mean=874.26µm, s.e.=4.94), but found that it was not useful taxonomically.

Mean	806.46 µm
Standard error of the mean	1.779
Standard deviation	50.3101
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.0345

Figure 4.3

Clypeus and Head Width



COLOUR CHARACTERS

Dang and Peterson (1980) and Peterson and Dang (1981) presented a key to the species of *S. damnosum s.l.* based largely on colour characters. In general, colour characters were not included in this analysis because of the greater subjectivity involved with scoring such characters, and because the statistical procedures used in the analysis are more appropriate for analysing continuous linear characters. However two colour characters were included despite these problems because of their proven taxonomic importance in previous work.

4.3.5 WING TUFT COLOUR (V7)

The colour of the tuft of hairs at the base of the radial vein of the wing was recorded. The standard five state method was used for scoring the wing tuft colouration categories (Kurtak *et al.* 1981). The character states were as follows:

1. All pale hairs
2 Between one and five dark hairs (inclusive)
3. Mixed pale and dark hairs
4. Between one and five pale hairs (inclusive)
5. All dark hairs

The character was scored at high magnification using a compound microscope for greater accuracy.

Wing tuft colour has been used extensively in *S. damnosum* complex morphological studies in the past (Chapter one). In general it has been found that darker species are found in the forest while lighter species are found in the savanna. Some species (e.g. *S. squamosum*, Kurtak *et al* 1981, *S. soubrense* 'Beffa', Meredith *et al.* 1983) are known to show all five character states.

4.3.6 ABDOMINAL SETAL COLOUR (V8)

The colour of the setae on the ninth abdominal tergite was recorded (Garms and Zillman 1984) as a two state character: character state two was defined as all hairs towards the middle of the tergite being long thick and dark, character state one was defined as anything other than this unless it was not possible to score the variable.

Previous work has shown this character to be 99.7% diagnostic for *S. yahense* (Garms and Zillman 1984), although more recent work has shown that it is 91% diagnostic (Thomson *et al.* 1987).

ANTENNAL MEASUREMENTS

A number of antennal measurements were made, reflecting the importance of this character in previous studies on the taxonomy of the *S. damnosum* complex (see Chapter one). Two measurements included more than one segment, measures V9 and V10. Neither of these included the third antennal segment as this was sometimes lost in dissection. V9 was recorded at a higher magnification than V10 which was more comprehensive of the whole antenna. The usual measure of antennal length (e.g. Garms 1978) is from the base to the tip.

In addition, five antennal segments were measured at 400X magnification. These characters were chosen because previous work has shown that the relative compaction of the antennal segments is taxonomically important (Quillévéré *et al.* 1977, Garms 1978).

4.3.7 ANTENNAL LENGTH 1 (V9)

The antenna was measured from the base of the sixth to the tip of the eleventh antennal segment (Figure 4.4).

Mean	287.233 µm
Standard error of the mean	1.111
Standard deviation	34.4266
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.0

4.3.8 ANTENNAL LENGTH 2 (V10)

The antenna was measured from the base of the fourth antennal segment to the tip of the eleventh antennal segment (Figure 4.4)

Mean	428.188 µm
Standard error of the mean	1.6418
Standard deviation	46.438
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.0345

4.3.9 ANTENNAL SEGMENT 4 (V11)

The fourth antennal segment was measured at 400X magnification from the proximal to the distal margin (Figure 4.4).

Mean	39.894 µm
Standard error of the mean	0.199
Standard deviation	5.628
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.5172

4.3.10 ANTENNAL SEGMENT 5 (V12)

The fifth antennal segment was measured at 400X magnification from the proximal to the distal margin (Figure 4.4).

Mean	39.26 µm
Standard error of the mean	0.1935
Standard deviation	5.472
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.655

4.3.11 ANTENNAL SEGMENT 6 (V13)

The sixth antennal segment was measured at 400X magnification from the proximal to the distal margin (Figure 4.4).

Mean	42.055 µm
Standard error of the mean	0.2129
Standard deviation	6.0216
Proportion of samples for which the null hypothesis of normality was rejected, $\alpha=0.05$	0.5172

4.3.12 ANTENNAL SEGMENT 7 (V14)

The seventh antennal segment was measured at 400X magnification from the proximal to the distal margin (Figure 4.4).

Mean	41.387 µm
Standard error of the mean	0.1989
Standard deviation	5.625
Proportion of samples for which the null hypothesis of normality was rejected, $\alpha=0.05$	0.6897

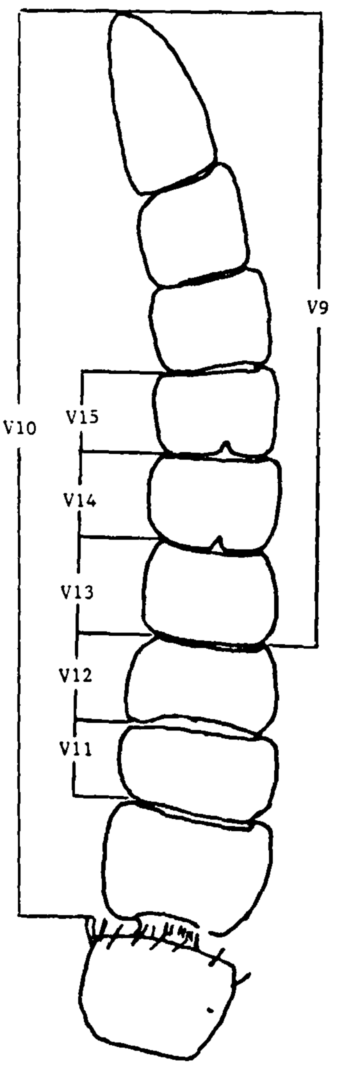
4.3.13 ANTENNAL SEGMENT 8 (V15)

The eighth antennal segment was measured at 400X magnification from the proximal to the distal margin (Figure 4.4).

Mean	40.785 µm
Standard error of the mean	0.1905
Standard deviation	5.3873
Proportion of samples for which the null hypothesis of normality was rejected, $\alpha=0.05$	0.7586

Figure 4.4

Antennal Measurements



WING CHARACTERS

In all, six wing characters were measured or scored: wing tuft colouration and the ones presented in this section, of which four were continuous measurements and one a count.

4.3.14 WING LENGTH 1 (V16)

The length of the wing was measured from the humeral cross vein to the fusion of the subcostal and costal veins (Figure 4.5). This is not the whole wing length but only the length of the cell. This character has not been used in previous studies of the *S. damnosum* complex.

Mean	737.791 µm
Standard error of the mean	2.0485
Standard deviation	57.939
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.0345

4.3.15 WING LENGTH 2 (V17)

The wing was measured from the point of fusion of the humeral cross vein and the costa to the radius media crossvein (Figure 4.5).

Mean	450.943 µm
Standard error of the mean	1.222
Standard deviation	34.567
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.069

4.3.16 WING WIDTH 1 (V18)

The width of the wing was measured from the point of fusion of the two medial veins to the meeting point of the second cubital vein and the wing margin (Figure 4.5).

Mean	1010.969 μm
Standard error of the mean	2.489
Standard deviation	70.402
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.2414

4.3.17 WING WIDTH 2 (V19)

The width of the wing was measured from the meeting point of the second cubital vein and the wing margin to the end of the radial sector (Figure 4.5).

Mean	1428.8996 μm
Standard error of the mean	3.5159
Standard deviation	99.443
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.1034

4.3.18 WING LENGTH 3 (V20)

The length of the wing was measured from the end of the radial sector to the radius media crossvein (Figure 4.5).

Mean	1487.153 μm
Standard error of the mean	3.791
Standard deviation	107.217
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.06897

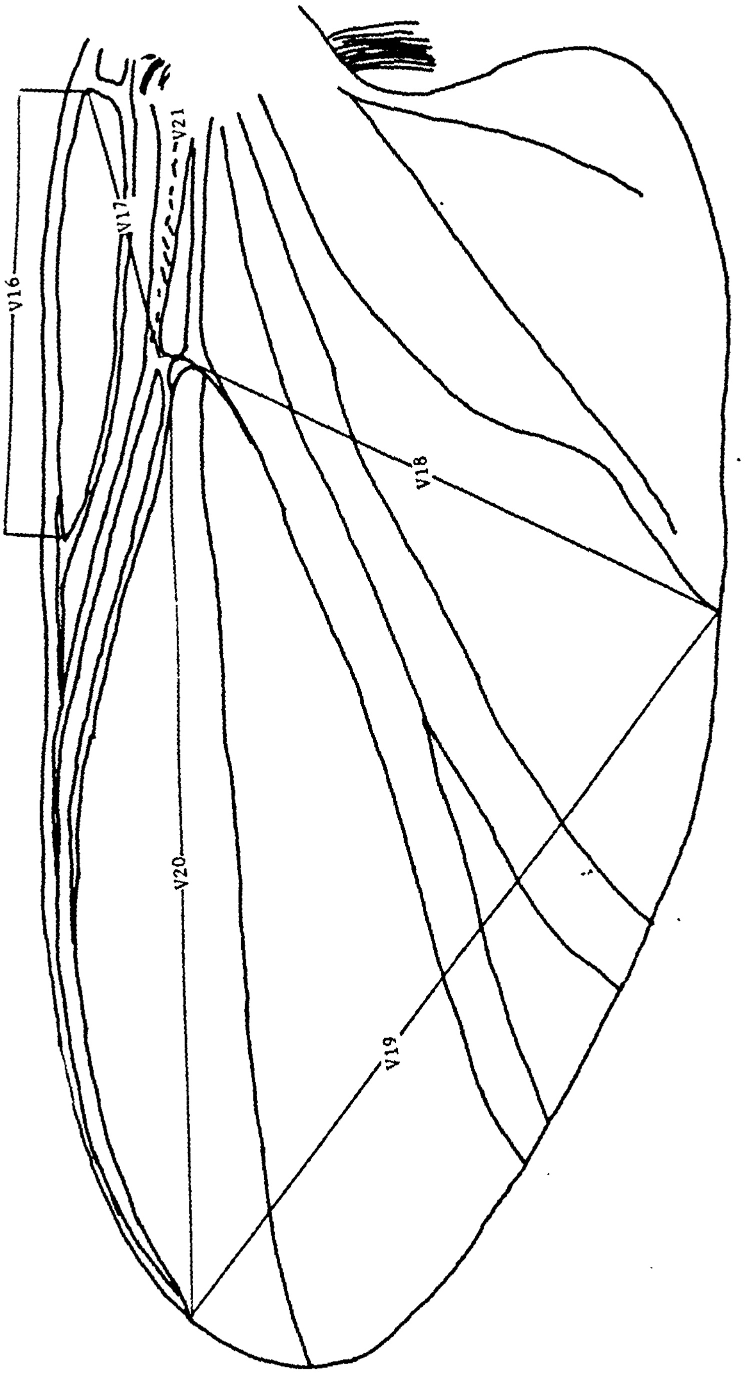
4.3.19 NUMBER OF HAIRS ON THE RADIAL VEIN (V21)

The number of hairs on the radial vein up to the radius media cross vein were counted (Figure 4.5). This character was previously used by Quillévéré and Sechan (1978) as a taxonomic character in the *S. damnosum* complex, but was subsequently shown to be of little taxonomic value by Garms (1978) and Townson and Meredith (1979).

Mean	15.274
Standard error of the mean	0.1192
Standard deviation	3.371
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.3103

gure 4.5

ing Characters



LEG CHARACTERS

Nine leg characters were measured or scored, eight continuous measurements and one a count. Leg shape has previously been shown to be variable within *S. damnosum s.l.* (Lewis and Duke 1966). Anon (1976 and Soponis and Peterson 1976) also measured a number of leg segments and found them to be taxonomically useful.

4.3.20 FEMUR LENGTH 1 (V22)

The length of the femur of the left fore leg was measured in anterior aspect from the strongly sclerotized ventral articulation to the extreme dorsal end (Figure 4.6). Anon (1976) measured a similar character (Mean=616.87µm, s.e.=3.5) and considered it to be taxonomically useful.

Mean	633.352 µm
Standard error of the mean	1.7022
Standard deviation	48.144
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.1034

4.3.21 TIBIA LENGTH 1 (V23)

The length of the tibia of the left fore leg was measured in anterior aspect from the extreme dorsal end of the proximal end to the end of the tibia (Figure 4.6). Anon (1976) measured this character (Mean=695.87µm, s.e.=3.91) but did not consider it to be taxonomically useful.

Mean	695.971 µm
Standard error of the mean	1.7783
Standard deviation	50.2925
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.1724

4.3.22 BASITARSUS LENGTH 1 (V24)

The length of the basitarsus of the left fore leg was measured in anterior aspect from the point of articulation with the tibia to the distal end (Figure 4.6). Anon (1976) measured this character (Mean=437.75µm, s.e.=1.03) and considered it to be taxonomically useful.

Mean	438.79 µm
Standard error of the mean	1.202
Standard deviation	39.999
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.1035

4.3.23 TARSAL SEGMENT 2 (V25)

The second tarsal segment of the left fore leg was measured from its proximal to its distal end (Figure 4.6). Anon (1976) measured this character (Mean=161.5µm, s.e.=1.03) but did not consider it to be taxonomically useful.

Mean	165.769 µm
Standard error of the mean	0.425
Standard deviation	12.02
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.0178

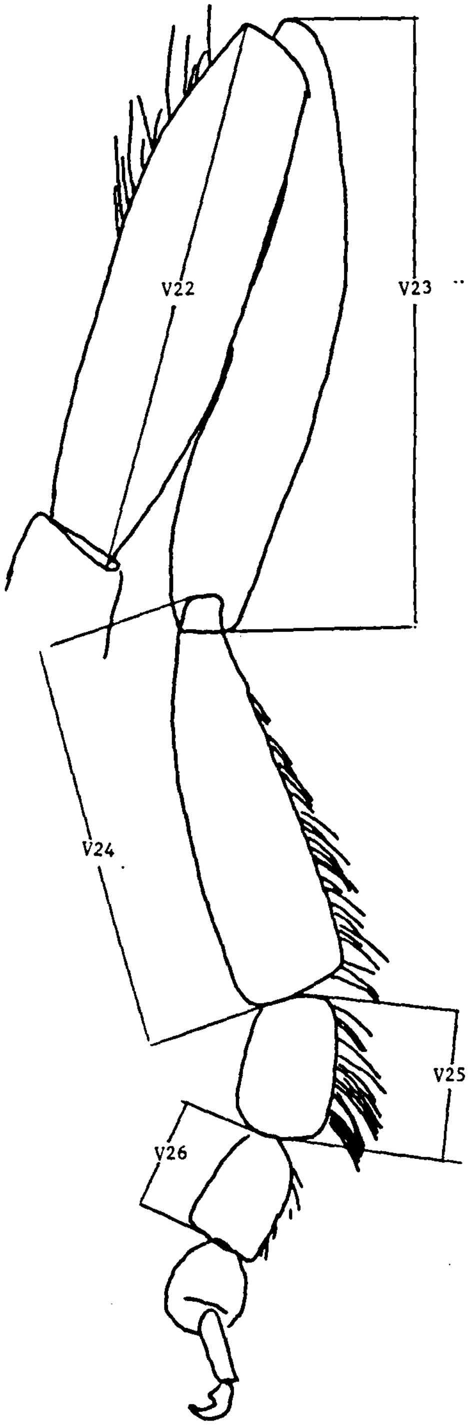
4.3.24 TARSAL SEGMENT 3 (V26)

The third tarsal segment of the left fore leg was measured in anterior aspect from its proximal end to its distal end (Figure 4.6). Anon (1976) measured this character (Mean=120.1µm, s.e.=1.03) but did not consider it to be useful.

Mean	124.605 µm
Standard error of the mean	0.3526
Standard deviation	9.9744
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.069

Figure 4.6

Fore-leg Characters



4.3.25 FEMUR LENGTH 2 (V27)

The length of the femur of the left mid-leg was measured in posterior aspect from the ventral articulation with the coxa to the dorsal distal end (Figure 4.7). Anon measured this character (Mean=628.82µm, s.e.=3.6) but did not consider it to be taxonomically useful.

Mean	663.247 µm
Standard error of the mean	1.7036
Standard deviation	48.185
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.138

4.3.26 TIBIA LENGTH 2 (V28)

The length of the tibia of the left mid-leg was measured in posterior aspect from the dorsal/proximal articulation with the femur to the distal end (Figure 4.7). Anon (1976) measured this character (Mean=620.99µm, s.e.=3.6) but did not consider it useful.

Mean	623.229 µm
Standard error of the mean	1.5927
Standard deviation	45.047
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.1035

4.3.27 BASITARSUS LENGTH 2 (V29)

The basitarsus of the left mid-leg was measured in posterior aspect from the point of articulation with the tibia to the distal end (Figure 4.7). Anon (1976) measured this character (Mean=316.72µm, s.e.=2.01) and considered it to be taxonomically useful.

Mean	328.755 µm
Standard error of the mean	0.996
Standard deviation	28.1706
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.0345

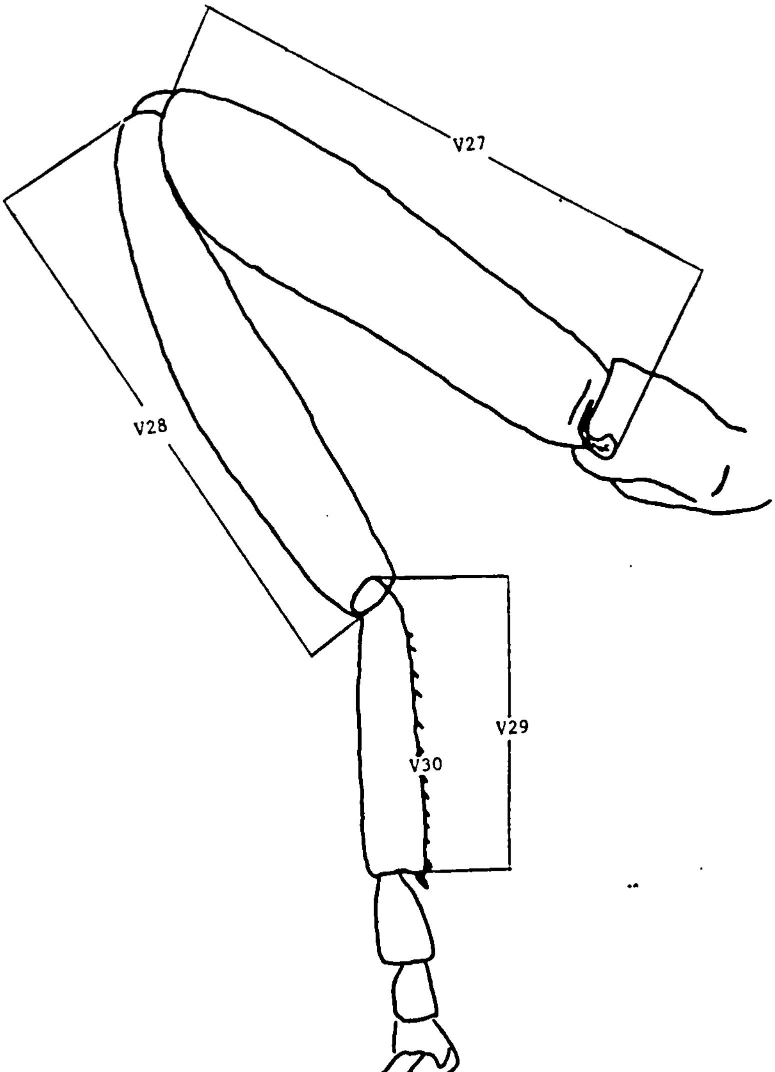
4.3.28 BASITARSAL SPINE NUMBER (V30)

The number of spines on the dorsal margin of the basitarsus of the left mid-leg were counted (Figure 4.7). The sample size for this character was smaller (N=777) than all the other characters in this section (N=800) because this character was rejected from the analysis at an early stage (Chapter 5) due to a large number of missing values and poor discriminatory power.

Mean	24.51
Standard error of the mean	0.1122
Standard deviation	3.128
Proportion of samples for which the null hypothesis of normality was rejected, α=0.05	0.1724

Figure 4.7

Mid-leg Characters



4.4 DISCUSSION

The character set chosen for this analysis comprised 28 characters, of which 15 were similar or identical to characters previously considered taxonomically useful in the *S. damnosum* complex. The number of characters measured or scored could in principle have been extended indefinitely, but this would have severely restricted the range of variation which could have been sampled.

The time taken to measure or score all 28 characters was roughly 20 minutes. The lengthiest measurements to take were those taken on the dissecting microscope because it was more difficult to ensure that the fly was positioned in the correct plane, while those characters measured on the compound microscope were slide-mounted and so in the correct plane. Dissection of the fly was also time consuming, and was made more or less difficult according to the preservative in which the fly was stored. Ethanol preserved material was less brittle than Propanol preserved material. Propanol was preferred because the brittleness of the fly made dissection quicker and easier.

CHAPTER FIVE: WITHIN SAMPLES VARIATION

5.1 INTRODUCTION

The major objective of the multivariate morphometric analysis of the *S. damnosum* complex is to develop a technique for the identification of adult female *S. damnosum s.l.* to their correct cytospecies. To develop the appropriate statistics for this method it is necessary to obtain sufficient samples of each cytospecies which together comprehensively cover the range of variation within each cytospecies.

The purpose of this chapter is to describe the statistical methods used to ensure that the variation observed within each sample was not unduly influenced either by mistakes within the data set or by rogue individuals.

There were two basic requirements which the available samples needed to meet for them to be included in the analyses presented in Chapters six to nine: the chromosomal identity of the sample was known, and the individuals within the samples were not atypical of the range of variation found within the cytospecies.

Examining Appendix one shows that the strength of evidence supporting the chromosomal identity of samples ranged from the moderately strong to the relatively weak. Thus some samples were known from correlated larval cytotaxonomy to be pure for one species (e.g sample V1=1, *S. squamosum*), or to be a mixture of species (e.g sample V1=5, *S. squamosum* and *S. sanctipauli*); other samples did not have correlated larval identifications but were believed, from other evidence to be a certain cytospecies (e.g. sample V1=15, *S. squamosum*) or cytospecies mixture (e.g. sample V1=10, *S. damnosum s.s.*, *S. squamosum*/*S. yahense* from correlated DNA probes identifications).

The pretreatment that a sample underwent depended on the *a priori* information available as to its species composition and purity. In the case of pure samples, the problem of pretreatment was restricted to the identification of outliers and occasional contaminants, whilst in the case of mixed samples the problem was one of separating a mixture of distributions followed by outlier detection in the separated parts.

5.2 MISSING VALUES

Not all individuals could be measured or scored for every character described in Chapter four, either because of damage to or loss of the character during dissection, or because of distortion during the slide-mounting stage. Some characters were missing more often than others (Appendix one). This created two data-analytic problems, firstly what to do with a character for which missing values were relatively common, and secondly what to do with individual flies with one or more missing values.

Characters which could not be scored for a relatively large number of individuals were tested in the course of the data-collection phase of the analysis (i.e. prior to the separation of mixtures or the detection of outliers) to see if they had any discriminatory power, using one-way ANOVA and discriminant analysis. If the character proved to be taxonomically unimportant relative to the number of missing values, it was discarded from subsequent analyses and individuals missing for only that character but complete for the rest were brought into the main data set. This procedure resulted in the basitarsal spine number being rejected due to the combination of a large number of missing values and poor discriminatory value. By contrast, wing tuft colouration was kept because of its good discriminatory value in spite of a relatively large number of missing values.

Any fly missing for any of the remaining 27 characters in the analysis was rejected from the main data set and not used again.

5.3 UNIVARIATE OUTLIERS

Once a sample was cleared of flies with missing values, each character was screened for univariate outliers. Barnett and Lewis (1984) define an outlier as an observation (or subset of observations) which appears to be inconsistent with the remainder of the data set. Outlying individuals can arise in a number of ways (Barnett and Lewis 1984) including:

1. operator error (i.e. typing error, recording error, measurement error);
2. as outliers in relation to a specific model (e.g. normal distribution);
3. as contaminants from other distributions.

The underlying model assumed for the characters in this analysis was the univariate normal distribution (unless the character departed from this model in a large proportion of samples, see Chapter four). Outliers were assessed using the following methods.

5.3.1 INFORMAL GRAPHICAL METHODS

There are many informal graphical methods for examining a sample distribution (Tukey 1977, Seber 1984) Amongst the simplest is the 'stem-and-leaf' plot (Mosteller and Tukey 1977), which is an informal histogram, where the 'stem' is equivalent to the midpoints of a standard histogram and the 'leaves' represent the number and value taken by individuals within the class interval defined by the midpoints. This plot was used to assess the shape of the sample distribution, in particular its symmetry and the length of the tails, aspects of the distribution which have most influence on the usual estimates of parameters (Barnett and Lewis 1984). 'Box-and-whisker' plots (Tukey 1977) were also used in this analysis to assess the same sample properties (Figure 5.1). This is a graphical representation of the data which identifies the mean, median, inter-quartile range, and outlying individuals. Finally, the ordered data for each sample were plotted against the quantiles of the standard normal distribution N(0,1), giving a quantile-quantile (Q-Q) plot (Gnanadesikan 1977). A straight line results if the data are sampled from a normal distribution. This method was a useful informal method of assessing distributional assumptions and identifying outlying individuals.

These methods were all output as part of the procedure SAS PROC UNIVARIATE. Figure 5.1 is a sample of the output from such an analysis showing the three plots used. The individuals identified as outliers in this example were from sample V1=17 (Appendix one), the character was thorax length. In practise these methods resulted in voluminous printed output, so to be practicable most analyses were performed interactively, using the Display Manager System within SAS.

5.3.2 FORMAL STATISTICAL TESTS

Whilst the emphasis at this stage of the analysis was on informal data-analytic methods, three statistical tests were performed. Graphical methods are very useful but it is quite easy to miss important features of a graphical representation of the data, especially if the analysis is largely interactive. The following tests were used:

1. Shapiro-Wilk W test (Seber 1984,p452) for univariate normality; If the null hypothesis of normality was rejected, the graphical plots described in section 5.3.1 were examined, as departure from normality can result from the following reasons:

a). genuine non-normality (i.e. the sample is drawn from another distribution, such as the log-normal distribution) b). inadequate measurement of a normally distributed character such as clypeus width (Chapter four), identified by approximate symmetry, ordinary tails but a stepped distribution

c). from the presence of outliers.

2. Single upper or lower outlier test (Barnett and Lewis 1984, p167 Table VIII)

3. Upper and lower outlier pair test (Barnett and Lewis 1984, p171, Table XIIa).

These two outlier test were chosen for their power and relative simplicity to calculate.

When an individual was identified as an outlier using the formal and informal methods, the data were checked for typing errors. This was a common source of outliers in the data, and was also the simplest to correct. If, however, the record sheet and the data on the computer agreed, then the original measurement was checked for recording error, and the individual re-measured for that character. If a fly appeared to be genuinely discordant then it was removed from the data set and the whole procedure repeated until all such individuals were identified. Appendix one gives the number of flies rejected from each sample as a result of either univariate or multivariate outliers.

5.4 MULTIVARIATE OUTLIERS

The detection of univariate outliers described in the previous section is a relatively straightforward task, involving a simple ordering of the data. However outlier detection in multivariate data is an extremely complex procedure for the following reasons:

1. there is no unique ordering of multivariate data;
2. graphical representation of high dimensional data is not easy;
3. outliers in multivariate data can distort the association between characters (correlation) as well as the estimates of location and spread affected by univariate outliers. (see Seber 1984, Barnett and Lewis 1984, Gnanadesikan 1977, Gnanadesikan and Kettenring 1972 for details). In addition bias is introduced due to sampling error in the estimate of the parameters of the joint multivariate distribution for a small sample size and moderately large dimensionality (Seber 1984). Because of these problems a number of different methods were used to identify multivariate outliers in an informal, interactive approach.

5.4.1 PRINCIPAL COMPONENTS ANALYSIS

One way of imposing order on multivariate data is to find linear combinations of the original characters (Seber 1984) and to examine the ordering of individuals along these new variables. A straightforward and well tried method for finding linear combinations is principal components analysis (Seber 1984, Gnanadesikan 1977). The first principal component is that linear combination of the original variables along which the variance is a maximum, subsequent principal components account for progressively less variance until r principal components have been extracted (where r is the rank of the dispersion (variance/covariance) matrix, equal to the number of characters, p, if the dispersion matrix is not singular). For data sampled from a multivariate normal population, the scatter of points along each component is expected to be univariate normally distributed (Mardia, Kent and Bibby 1977, p230), and so can be examined using the graphical and statistical methods described in section 5.3. Furthermore, the component on which an individual is an outlier gives information as to the nature of the distinctiveness of the fly (Gnanadesikan 1977). For example, if an outlier was found on the first principal component, usually a size vector in multivariate morphometric analysis (Blackith and Reyment 1971), then its distinctiveness was due to systematic differences in all characters simultaneously (i.e. size), whereas an outlier on one of the shape vectors could be due to the association between one or more characters (i.e. shape).

Principal components are extracted either from the sample dispersion (variance/covariance) matrix or from the correlation derived from it. Because the estimates of dispersion are very sensitive to outliers it has been suggested that a robust estimate of dispersion matrix (i.e an estimate which is insensitive to outlying individuals, Huber 1981) should always be calculated and the principal components extracted from this (Gnanadesikan 1977). This was attempted using a method for calculating robust principal components (Campbell 1980, using a GENSTAT macro written by Matthews 1984). This method has the effect of making discordant observations more obvious, because observations away from the bulk of the data are given reduced weight, and so influence the estimates of mean and dispersion to a smaller extent than usual. The method was very expensive computationally and could not be used in the interactive way that the usual principal components analysis computed by SAS PROC PRINCOMP could.

To summarise, each sample was subjected to a principal components analysis, the scatter of points along the principal components were examined for normality and the presence of outliers. Bivariate scatters of selected principal components were also examined, in particular the first two (accounting for most variance) and the last two (accounting for least variance). Characters which were known *a priori* not to be univariate normally distributed were not usually included in the principal components analysis. As a result of this method a list of multivariate outliers for a sample was obtained.

5.4.2 MAHALANOBIS' DISTANCE

The Mahalanobis' squared distance of each individual to the sample mean can be used to order the individuals in a sample. The distance: $(x_i-\bar{x})'S^{-1}(x_i-\bar{x})$ where x_i is the vector of observations for the i-th individual, $\bar{x}$ is the mean vector of the sample, and S is the sample dispersion matrix, is distributed as a χ^2 variable with p (number of variables) degrees of freedom (Seber 1984). For p up to and including five, a discordancy test for a single multivariate normal outlier was used (Barnett and Lewis 1984, Table XXX). Higher dimensionalities than this could not be tested for discordant individuals in this way. Instead, an informal graphical method was used, which involved plotting the ordered Mahalanobis' squared distances against the quantiles of a Beta distribution with parameters $a=\frac{1}{2}p$ and $b=\frac{1}{2}(n-p-1)$, where n is the sample size and p is the number of characters. If the data were multivariate normal, the result was a straight line (Seber 1984, p153). Figure 5.2 is an example of such a plot, with the outlier clearly visible on the top Q-Q plot. Computationally this was performed using a routine written in SAS PROC MATRIX. A simpler Q-Q plot was also drawn by plotting the cube roots of the ordered distances against the quantiles of a standard normal distribution (Campbell 1984). These plots were used interactively to assess the distributional properties of the sample and to identify discordant individuals within the sample, however the high dimensionality and small sample size often rendered the results of this method equivocal, and so it was usually used only on relatively small numbers of characters at a time.

As well as these plotting methods, Mardia's (1970) estimates of multivariate skewness and kurtosis were calculated in the same SAS PROC MATRIX routine, as these could be derived in earlier steps of the program to calculate the Mahalanobis' squared distances. These statistics have poor small sample properties (Seber 1984), but were used as general indicators of the adequacy of the data set. The multivariate kurtosis statistic was especially useful in the identification of multivariate outliers, to which it is particularly sensitive.

Using these techniques, a list of potential multivariate outliers was obtained for each sample. At each stage, the individual which was the most 'extreme' outlier was discarded, until all outliers were identified and excluded from the data set.

5.5 MIXED SAMPLES

A number of samples were known *a priori* to be mixtures of cytospecies (Appendix one), from correlated larval cytotaxonomy. Therefore, before the set of pure-sample outlier detection techniques described in previous sections could be applied it was necessary to separate mixed samples into their constituent species. Two approaches to this problem were taken: internal analysis and external analysis.

5.5.1 INTERNAL ANALYSIS

With internal analysis, only the sample data were used to separate the mixture into its constituent parts. A number of dimension reduction methods were used to view the data comprehensively, and so identify groupings within the data (Gordon 1981), including principal components analysis and cluster analysis methods (Gordon 1981, Chapter three). Using these methods, the sample was split into its component parts and the parts identified as a particular species with reference to the expected proportions from correlated cytotaxonomy, or from external analysis (section 5.5.2).

5.5.2 EXTERNAL ANALYSIS

With external analysis, the data used for separating the mixture came from outside of the sample data. For a small number of samples, independent DNA probes identifications were available (Post pers. comm.), whilst other samples were separated using linear discriminant functions calculated from other samples belonging to the species known to be contained in the mixed sample.

Once a mixture was separated using these techniques, each part was then examined as if it were a pure sample, for univariate and multivariate outliers using the methods described in sections 5.3 and 5.4.

5.6 SUMMARY OF METHODS

For samples which were known *a priori* to be pure, the initial sample was examined for missing values in the 27 characters for which missing values were to be rejected. These flies were discarded from the main data set. Univariate outlier tests were applied to each character in turn, typing and recording errors were recognised and remedied, any flies which were outlying were rejected. Multivariate outlier tests were then applied, either to the whole character set or to subsets of characters, and any multivariate outliers which were identified were rejected.

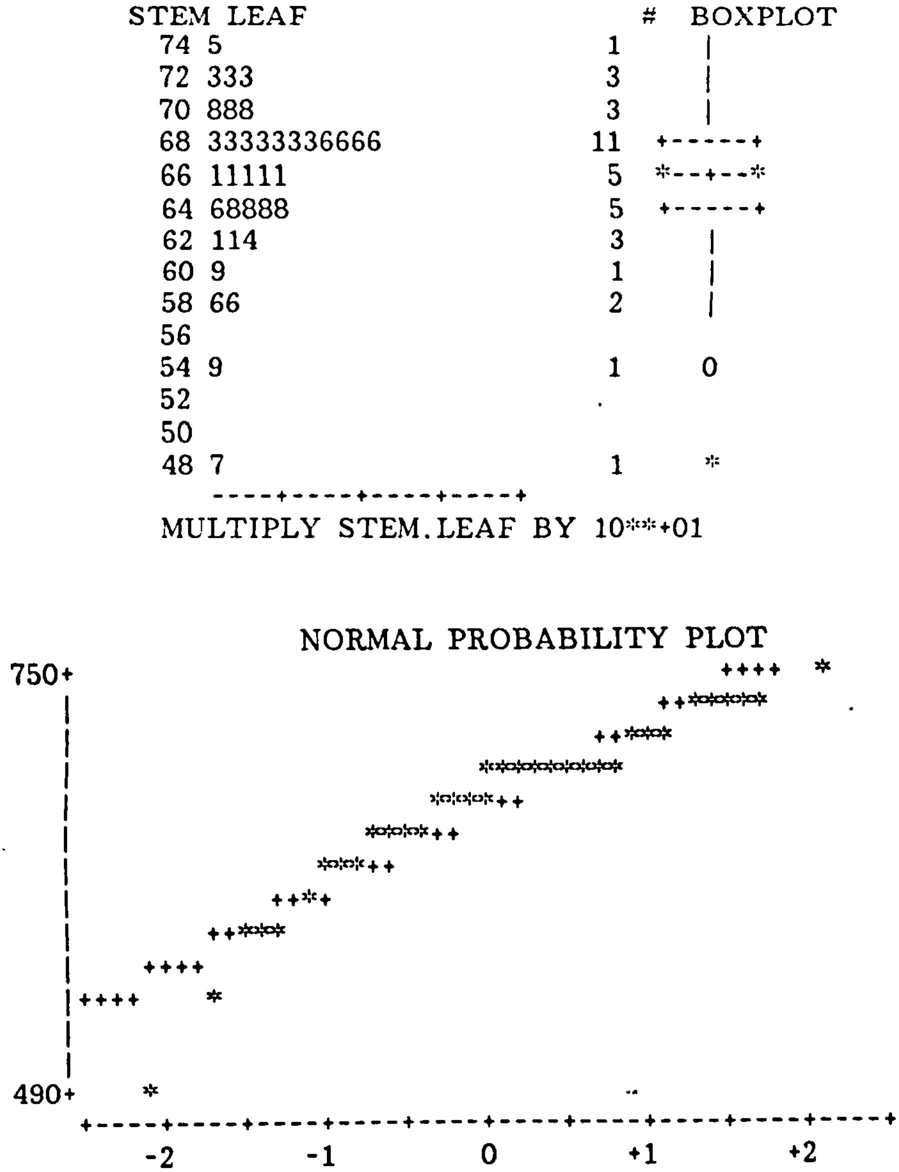
For samples which were known *a priori* to be mixed, the exact procedure used depended upon the proportion of each species in the mixture. Thus if one species was present only as a small proportion of the total sample, the methods applied to the sample were basically those applied to a pure sample with outliers and contaminants. Initially, missing values were identified and rejected, and typing or recording errors corrected. Then the internal (section 5.5.1) and external analyses (secticn 5.5.2) were applied as appropriate, any flies which did not fall clearly into any of the clusters revealed by these method were rejected, and the initial sample divided into its component parts. The 'new' samples produced in this way were then treated in exactly the same way as if they were pure samples, with univariate and multivariate outliers being detected and rejected.

5.7 DISCUSSION

The methods described in this chapter aimed to identify problems within each sample of *S. damnosum s.l.* and to solve them either by correcting the data, or by removing rogue individuals. To some extent the choice of flies to be rejected involved some subjectivity, because some methods identified flies as outliers which were not recognised by other methods. However, this problem was of less importance than the fact that obviously influential flies were identified by all methods without equivocation, so that the minimum effect of applying the methods described in this chapter was to identify flies whose presence in the data set would have seriously affected the subsequent analyses.

Figure 5.1

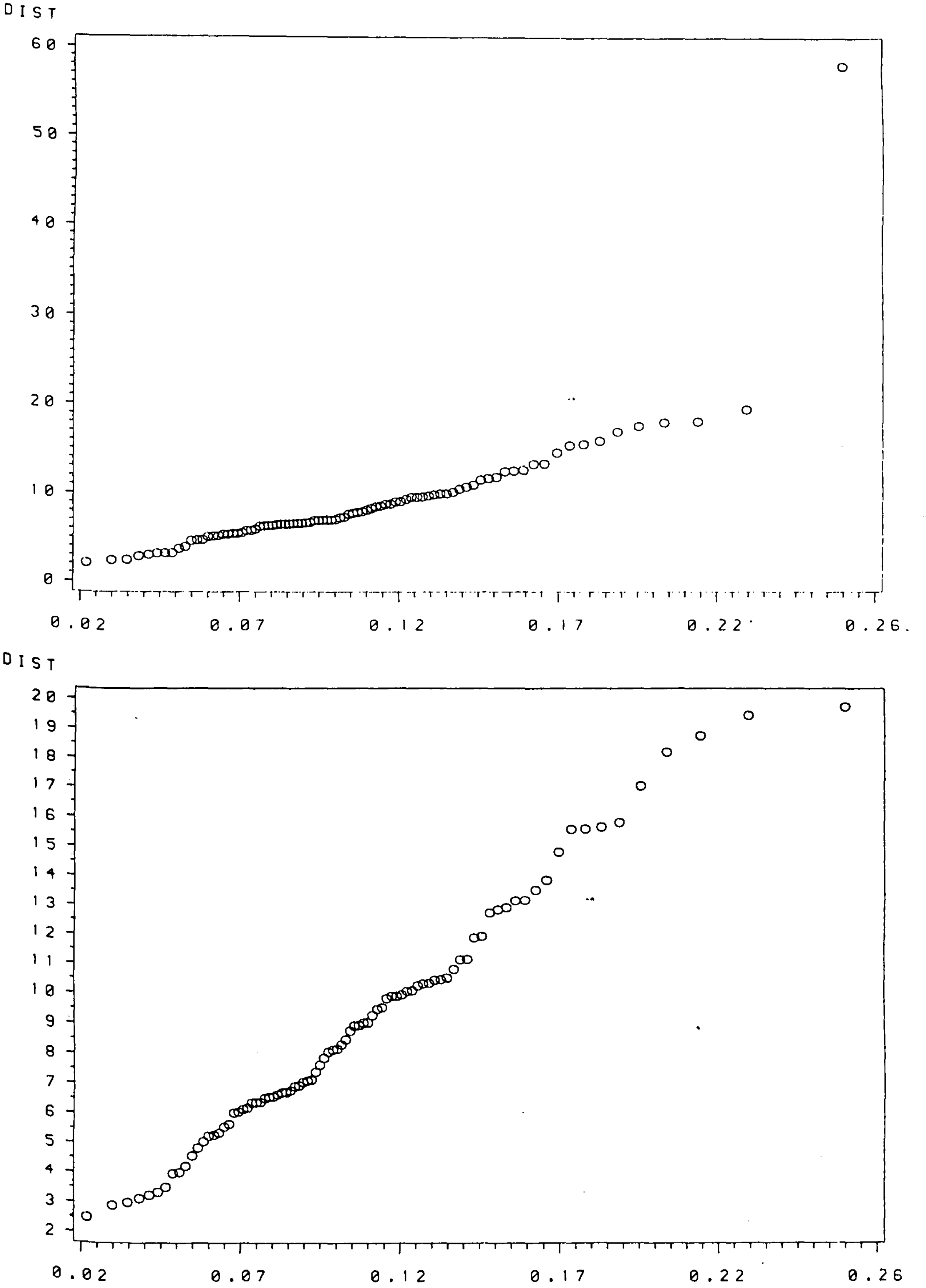
Example printout from SAS PROC UNIVARIATE showing stem-and-leaf plot, boxplot and Q-Q plot, to identify an outlier.



```
STEM LEAF                  #   BOXPLOT
 74 5                      1      |
 72 333                    3      |
 70 888                    3      |
 68 33333336666           11   +-----+
 66 11111                  5   *--+--*
 64 68888                  5   +-----+
 62 114                    3      |
 60 9                      1      |
 58 66                     2      |
 56
 54 9                      1      0
 52
 50
 48 7                      1      *
    ----+----+----+----+
 MULTIPLY STEM.LEAF BY 10**+01
```

Figure 5.2

Q-Q plots of Mahalanobis' squared distances against quantiles of Beta distribution, before and after identification of outlier



CHAPTER SIX: WITHIN SPECIES VARIATION

6.1 INTRODUCTION

The analysis of variation within species using multivariate morphometric techniques is taxonomically important but inferentially very complex (Thorpe 1976, Blackith and Reyment 1971). Its importance lies in the deeper understanding of the biology of a species, which may have implications for its vectorial importance, but the complexity arises from difficulties in identifying the source and nature of the variation which has been uncovered.

The observed variation within a taxonomic unit is likely to be a combination of environmentally mediated variation and variation with a genetic basis, and can be expressed either as size differences, shape differences, differences in the covariation between characters, or any combination. This expressed variation could take the form of geographic and/or temporal variation between samples, and the patterns of variation could conform to any of a number of theoretical possibilities (Endler 1977).

In order to investigate fully intraspecific morphometric variation careful experimental planning and sampling regimes need to be adhered to. However, in the absence of carefully controlled experiment, valuable information can still be obtained from data acquired in an *ad hoc* manner. Thus the purpose of this chapter is to analyse, for the first time, multivariate morphometric variation within the *S. damnosum* complex, using data obtained primarily for purposes of discrimination between these species (Chapters seven and eight), to describe this variation and to contrast it with known chromosomal or other variation, where appropriate.

6.2 MATERIALS AND METHODS

6.2.1 MATERIALS

The samples used in these analyses are listed as Appendix one, and were previously screened for outliers and/or contaminants (Chapter five). Of the 28 characters originally measured or scored, the number of spines on the basitarsus of the second leg was not analysed because it showed little interspecific variation and had many missing values (Chapter four). Two others, wing tuft colouration and abdominal setal colouration were analysed separately rather than jointly with the other characters because they both showed consistent departure from univariate normality (Chapter four).

6.2.2 DIMENSION REDUCTION

Even though three characters were not included in the analyses, 25 characters were measured or scored for each fly. This is clearly very large relative to most of the sample sizes obtained (Appendix one). Therefore, for the results not to be dominated by sampling error the number of characters jointly analysed needed to be reduced (Van Ness and Simpson 1976).

In the context of discriminant analysis, this is usually achieved by stepwise discriminant analysis (McKay and Campbell 1982a,b, Seber 1984). The pitfalls associated with this method are discussed in Chapter seven, where the inadequacy of the method in deriving a subset of characters with a low error rate is discussed. The analyses in the present chapter are not involved with the derivation of statistics for the future allocation of unknown flies in the field, but with describing intraspecific morphometric variation parsimoniously. In this context (i.e. descriptive discrimination rather than allocatory discrimination, Geisser 1977), stepwise discriminant analysis is more appropriate for character subset generation, as the method optimises a measure of overall separation (Wilks' lambda) at each stage of the analysis (Seber 1984). The program used was SAS PROC STEPDISC, the algorithm that this program uses is described in section 7.2.1.

The initial subset of characters derived using stepwise discriminant analysis was examined for departures from univariate and multivariate normality, using SAS PROC UNIVARIATE for the marginal distributions and Mardia's multivariate skewness and kurtosis statistics for the joint distributions (see Chapter five). The null hypothesis of equal dispersion was also tested at this stage using the likelihood ratio test (see Chapter seven).

Characters which contributed to the departure from the null assumptions of equal dispersion and normality were removed interactively, until a character set was formed which was acceptable.

6.2.3 MULTIVARIATE ANALYSES

The Mahalanobis' squared distance between the sample mean vectors was calculated as part of the procedure SAS PROC DISCRIM and tested for significance. A cancnical variates analysis was performed, involving extracting the eigenvalues and eigenvectors from the matrix $W^{-1}B$, where W is the pooled within-samples SSQPR matrix and B is the between-samples SSQPR matrix. The standardised eigenvectors (Canonical variates) show the association between each character and each discriminant vector, while the proportion of variation accounted for by each canonical (discriminant) vector is given by the associated eigenvalue (canonical root). The number of canonical variates extracted is min(g-1,p), where g is the number of groups and p is the number of characters. Thus for two groups there is one canonical variate, which is directly related to the linear discriminant function.

The first canonical variate is that linear combination of the original characters along which the ratio of between groups to within groups variation is a maximum (Campbell and Atchley 1981), with subsequent vectors accounting for progressively smaller proportions of variance. Thus, the plane defined by the first two canonical variates is the best two dimensional representation of discriminant space amongst all such planes (Gnanadesikan 1977), and was used in these analyses as a parsimonious representation of this space. The procedure used was SAS PROC CANDISC.

A multivariate analysis of variance (MANOVA) was performed (using SAS PROC GLM) to assess the degree of overall separation *via* Wilks' lambda, which is the ratio of the determinant of the between-samples SSQPR to the determinant of the total SSQPR (Seber 1984).

The influence of size variation on between samples variation was assessed by calculating the eigenvectors of the pooled within-samples correlation matrix. The first eigenvector (principal component) is usually a size vector in morphometric studies (Rao 1964), and the scores along this vector were introduced as a covariable in a multivariate analysis of covariance (MANCOVA). The effect on the individual canonical roots was used as an informal measure of the influence of size on discrimination. The pooled within-samples correlation matrix was calculated by a routine written in SAS PROC MATRIX, while the MANCOVA was calculated using SAS PROC GLM.

6.3 RESULTS AND DISCUSSION

6.3.1 SIMULIUM DAMNOSUM S.S

Three samples of this species were available for analysis (Appendix one), one from Sierra Leone (V1=10, N=28), one from Benin (V1=28, N=16) and one from Togo (V1=28, N=27).

A stepwise discriminant analysis on the 25 character set resulted in an initial subset of eleven characters :

[V5,V9,V12,V15,V16,V17,V21,V22,V23,V25,V27]

i.e one head, three antennal, three wing and four leg characters. Mardia's multivariate skewness and kurtosis statistics and examination of the marginal distributions revealed some departure from the assumption of multivariate normality. The characters responsible for this distortion where removed interactively until the five character subset:

[V9,V16,V22,V25,V27]

resulted which conformed to the assumption of multivariate normality. In addition, the likelihood ratio test for equality of dispersion was not rejected at $p<0.0001$, so two of the basic assumptions of the multivariate linear model were met.

Table 6.1 shows the overall mean, pooled coefficient of variation, and proportion of variance among localities for the five character subset. The CVs are all very similar, but the proportion of variance among samples is lower for antennal length 1 than for the other characters.

Using the five character subset, the matrix of Mahalanobis' squared distances shown in Table 6.2 was obtained. All of these distances were significant at $p<0.001$, but clearly the two eastern samples (V1=10, V1=28) are closer to each other than either is to the western sample (V1=10).

Examination of the standardised canonical variates shown in Table 6.3 reveals that the first canonical variate, with a canonical root of only 0.9004 (accounting for 80% of total variance) loads strongest on two characters, wing length 1 and femur length 2. The group means along this vector were: [-1.15,0.83,0.7] for samples V1=10,11,28 respectively, so it discriminates principally the western sample from the eastern samples, as can also be seen from Figure 6.1. The second canonical variate, with a canonical root of only 0.224 is therefore probably not biologically significant, even though it is statistically significant (Campbell 1982). This vector discriminates between the two eastern samples (Figure 6.1).

The first principal component of the pooled within-groups correlation matrix was a size vector accounting for 74% of pooled within-groups variance, with coefficients,

[0.33,0.47,0.49,0.46,0.48]

When the scores along this vector were introduced into the model as a covariable the canonical roots fell from 0.9004,0.2243 to 0.7169,0.122. With size controlled in this way, there was still significant morphometric differentiation between the samples, but this was reduced.

To conclude, there is significant morphometric differentiation between the three samples of *S. damnosum s.s.*, but this is not great when compared with the results from some other groups (e.g. see section 6.3.4), and includes a significant proportion of size variation.

One sample (V1=10) is separated from the other two by two years and about 1500 km, and this sample is morphometrically further from the other two samples. However, *S. damnosum s.s.* is known to migrate distances greater than 500 km, (Garms and Walsh 1987), so it is possible that the eastern and western samples are from the same gene pool. Therefore the source of the multivariate variation found among samples within this species cannot be attributed to any specific causal factor such as seasonal or geographic variation.

Table 6.1

Overall mean, coefficient of variation (CV), and proportion of among-samples variance ($s^2(A)$)

Character	Mean	CV	$s^2(A)$(%)
Antennal Length 1	256.55μm	4.71	6.01
Wing Length 1	731.35μm	4.51	27.61
Femur Length 1	631.42μm	4.47	14.66
Tarsus Segment 2	165.93μm	4.56	20.28
Femur Length 2	663.23μm	4.05	25.14

Table 6.2

Mahalanobis' squared distances between samples

From sample	10	11	28
10	0.0	4.44	3.68
11	4.44	0.0	1.53
28	3.68	1.53	0.0

all $p<0.001$

Table 6.3

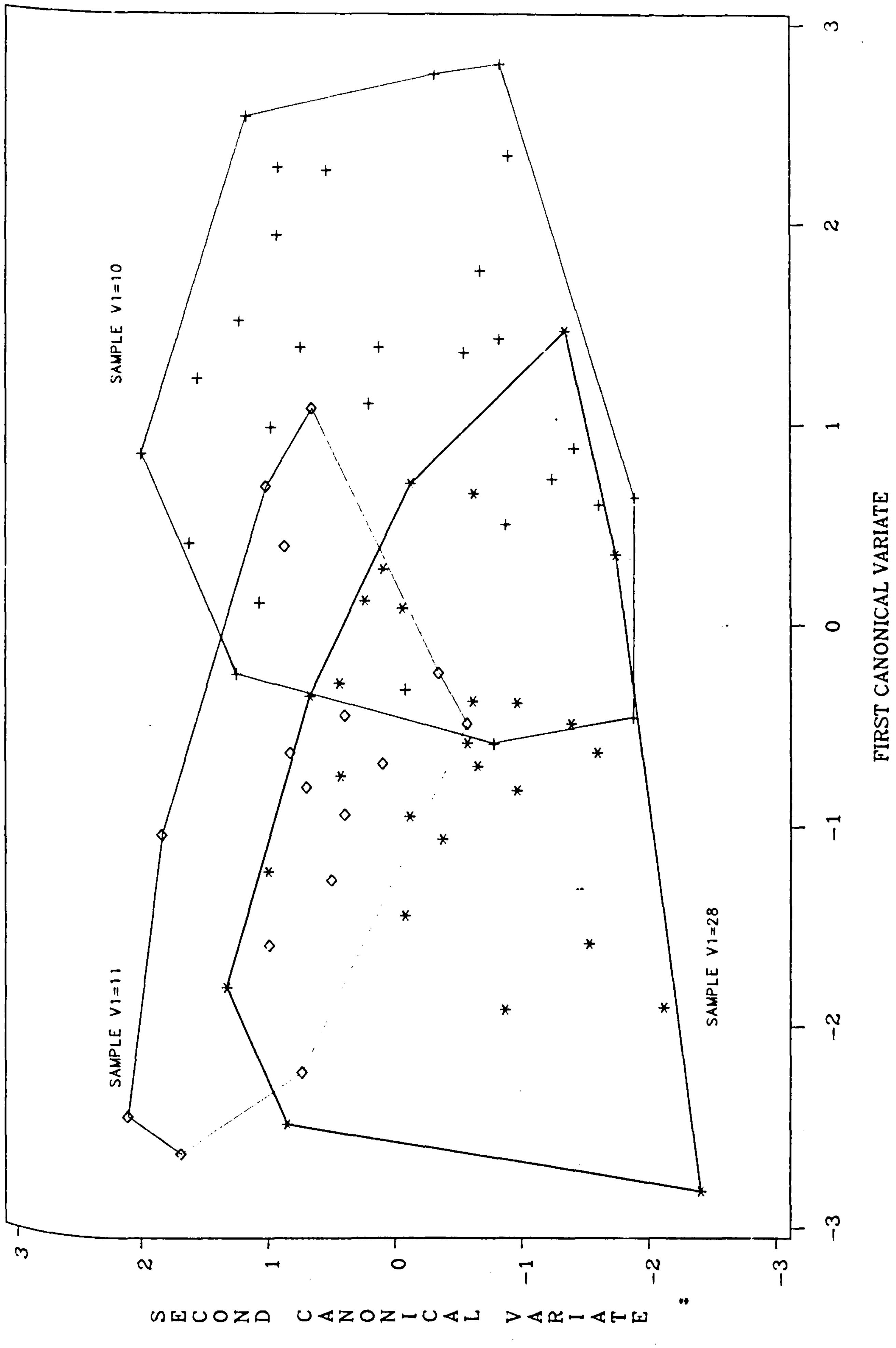
Standardised Canonical Variates for *S. damnosum s.s.*

Character	CV I	CV II
Antennal Length 1	-0.7088	-0.3105
Wing Length 1	1.2212	-0.4399
Femur Length 1	-0.8888	0.5433
Tarsus Segment 2	-0.8196	1.4641
Femur Length 2	1.4460	-0.6032
Canonical Root	0.9004[1]	0.2243[2]

[1] $p<0.001$
[2] $0.001<p<0.01$

Figure 6.1

Canonical Variates Plot for *S. damnosum s.s.*



6.3.2 SIMULIUM SIRBANUM

Five samples of *S. sirbanum* were available for analysis, four from Guinea (V1=21 N=44,V1=22 N=29,V1=23 N=43,V1=24 N=33), and one from Mali (V1=3 N=29) (Appendix one). A stepwise discriminant analysis on the 25 character set excluding wing tuft colouration, abdominal setal colouration, and basitarsal spine number resulted in an initial subset of 11 characters:

[V3,V4,V5,V6,V11,V17,V18,V20,V21,V23,V25]

i.e. two thorax, two head, one antennal, four wing and two leg characters. Examination of this character set using the methods described in Chapter five revealed some departure from multivariate normality. The characters having the greatest influence on non-normality were removed interactively, until a seven character subset which showed only minor departure from joint normality was obtained:

[V3,V4,V6,V17,V18,V20,V23]

i.e two thorax, one head, three wing and one leg character. The likelihood ratio test for equality of dispersion was not rejected at $p<0.0001$ so two of the assumptions of the multivariate linear model were not violated.

Table 6.4 gives the overall mean, coefficient of variation and proportion of variance among samples for the seven character subset. The proportion of variance among samples varies between characters, with wing width being considerably less variable among samples than the other characters. The two wing length characters show similar proportions of among samples variance as each other, implying that wing length varies among samples, while wing width varies less, i.e. wing shape differs among samples. When each sample's CVs were examined in greater detail using the variability profile method demonstrated in Bird *et al.* (1981), it became clear that the characters differed in their variability within profiles among samples. Thus head width, wing width 1 and wing length 3 were significantly less variable than the other characters, and this relationship was consistent among samples (Friedman randomised block test, $p<0.001$). Therefore wing shape is less variable than most other characters. A possible explanation for this might be that wing shape is under stronger natural selection for aerodynamic reasons.

Using this seven character set in a discriminant analysis, the matrix of Mahalanobis' squared distances given as Table 6.5 was obtained. Examination of this matrix shows that the principal morphometric differentiation within *S. sirbanum* involves the sample from Mali (V1=3) from the other four samples.

The standardised canonical variates are given as Table 6.6. The first two of the canonical variates are statistically significant, but the second canonical variate, with a canonical root of only 0.2579 is therefore unlikely to be of biological significance (Campbell 1982). Figure 6.2 shows the scatter of points in the first discriminant plane defined by the first two canonical variates. This figure clearly shows the differentiation of the Mali sample from the other four samples. The first canonical variate (Table 6.6) is principally a contrast between the wing characters wing width 1 and wing length 3, i.e. wing shape, with some influence from head width, confirming the result inferred from the univariate analyses.

The first principal component of the pooled within-species correlation matrix accounted for 85% of pooled within-species variation, and was a size vector with coefficients,

[.34,.39,.38,.37,.39,.38,.39].

When the scores along this vector were introduced into the model as a covariable the first canonical root fell from 1.1635 to 0.81101, showing that size has some importance in between-samples variation along this vector. The second canonical root fell from 0.2539 to 0.2432 showing that size is of no importance along this (biologically insignificant) vector.

To conclude, the major differentiation among the five samples of *S. sirbanum* involved one sample (V1=3) which is different from the other four in both shape and size. The main shape difference is wing shape, with the four Guinea samples having longer wings than the Mali sample, and all five having approximately the same wing width.

The Mali flies were sampled in November 1984, V1=21 and V1=22 in September 1986, V1=23 in August 1985 and V1=24 in December 1985, so that simple seasonal size variation can be discounted as the source of between samples variance. The Mali sample is geographically most isolated, but this species is known to migrate distances greater than 500 km (Garms and Walsh 1987), so all five samples are likely to be from the same gene pool. The only other difference between the Mali sample and the others is that it was reared from pupae rather than caught at human bait, so that the age distributions in the samples is likely to be different. Whether this is the source of variation, or some other unspecified source is responsible for the differentiation can not be decided on the available data.

Table 6.4

Overall mean, coefficient of variation (CV), and proportion of among-samples variance ($s^2(A)$)

Character	Mean	CV	$s^2(A)$(%)
Thorax Length	612.42µm	6.8	19.4
Thorax Width	846.99µm	6.42	19.17
Head Width	767.4µm	5.37	28.43
Wing Length 2	432.46µm	6.35	25.95
Wing Width 1	962.68µm	5.6	4.19
Wing Length 3	1394.33µm	5.04	24.15
Femur Length 2	657.29µm	5.52	12.77

Table 6.5

Mahalanobis' squared distances between samples

Sample	3	21	22	23	24
3	0.0	9.27	6.98	9.64	6.78
21	9.27[1]	0.0	0.47	1.23	2.16
22	6.98[1]	0.47[3]	0.0	1.41	1.25
23	9.64[1]	1.23[2]	1.41[1]	0.0	0.99
24	6.78[1]	2.16[1]	1.25[2]	0.99[3]	0.0

[1] $p<0.001$
[2] $0.01<p<0.05$
[3] $p>0.05$

Table 6.6

Standardised Canonical Variates for *S. sirbanum*

Character	CVI	CVII	CVIII	CVIV
V3	0.37	0.31	0.99	-0.35
V4	-0.29	-2.45	-0.27	-0.79
V6	1.09	-0.5	-0.76	1.8
V17	0.89	0.10	0.32	-0.59
V18	-1.85	0.05	-0.41	-0.37
V20	1.63	1.4	-1.2	-1.24
V23	-0.91	0.87	1.8	1.42
Canonical Root	1.164[1]	0.254[1]	0.08[2]	0.012[2]

[1] $p<0.001$
[2] $p>0.05$

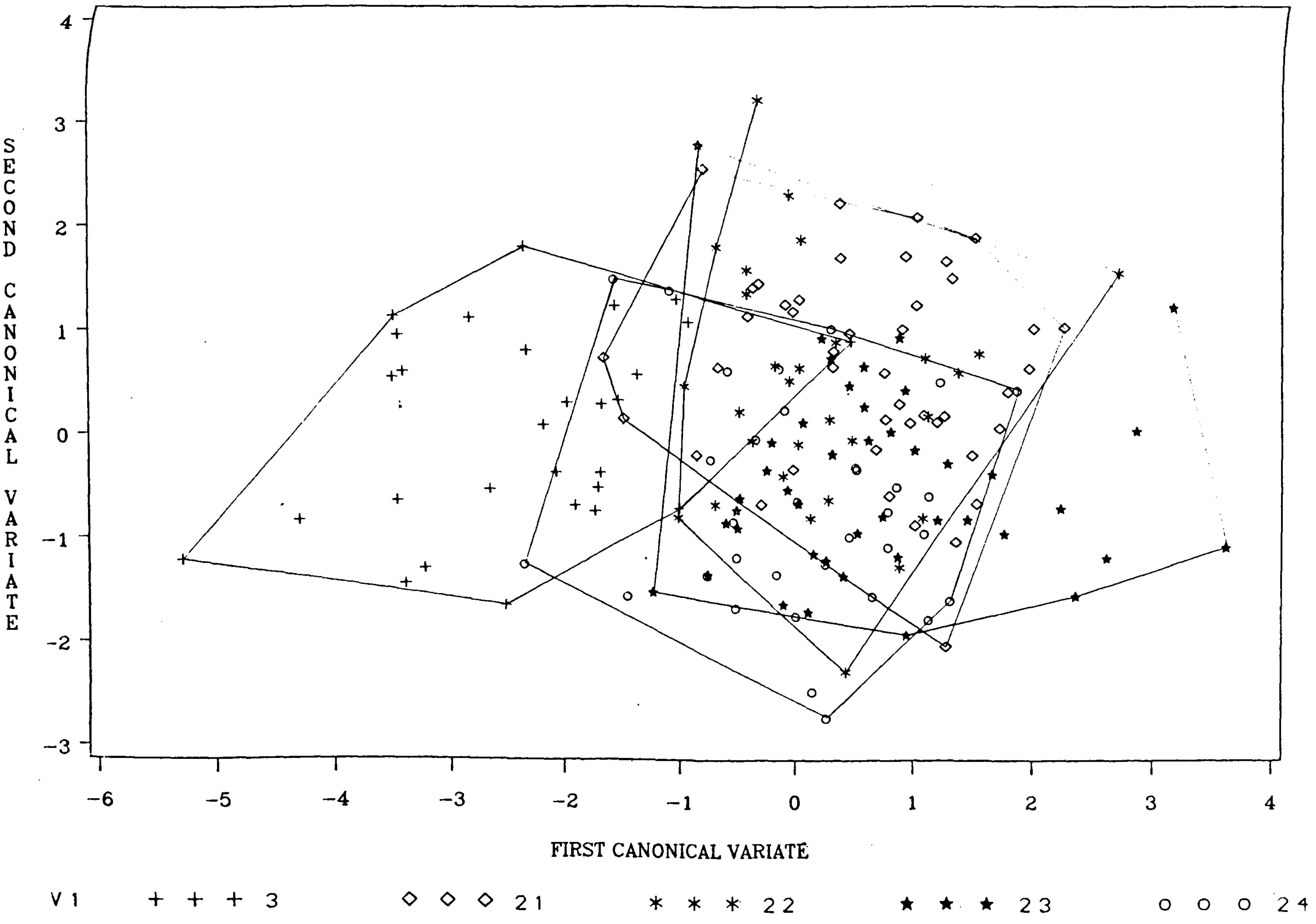


Figure 6.2
Canonical Variates Plot for *S. sirbanum*

6.3.3 SIMULIUM SANCTIPAULI

Only two samples of this species were available for analysis (Appendix one), but these are chromosomally and geographically distinct. Sample V1=4 (N=35) is typical *S. sanctipauli* from Côte d'Ivoire, while sample V1=5 (N=26) is *S. sanctipauli* 'Djodji' form from Togo (Chapter two, Chapter three, Surtees *et al.* 1988).

A stepwise discriminant analysis between the two samples using the 25 character set resulted in an initial subset of eight characters:

[V5,V6,V9,V15,V16,V17,V19,V20]

however this character set showed departure from multivariate normality, using Mardia's skewness statistic and so was reduced to the two character subset:

[V9,V20]

i.e. antennal length and wing length. Table 6.7 gives the overall means, coefficient of variation and proportion of variance among samples for the two characters. The CVs are similar, but the proportion of variance among samples is strikingly different.

The Mahalanobis' squared distance between samples using the two characters was 3.33, which was significant at $p<0.001$.

The standardised canonical variate, given as Table 6.8 was a shape vector, representing a contrast between antennal length 1 and wing length 3. Figure 3.3 shows the bivariate scatter of points using the two characters, and the differentiation between the two samples is clear.

The first principal component of the pooled within-samples correlation matrix was a size vector, with coefficients,

[0.71,0.71]

and accounted for 69.5% of pooled within-samples variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 0.8292 to 0.8279, indicating that the influence of size variation is negligible.

To conclude, there is significant morphometric differentiation between the two samples of *S. sanctipauli.* They were collected in the same month of the same year, so that simple temporality can be discounted as the source of the variation. The two samples do not differ along the size vector, so the observed morphometric variation is more likely to have a genetic basis. This variation is entirely shape variation involving the relative size of the antenna, a feature of *S. damnosum s.l.* morphology which has frequently been used to distinguish between species (e.g. Garms 1978), and which is presumably developmentally independent from the rest of the morphology. This species is known not to migrate large distances (Garms and Walsh 1987), allowing for the evolution of localised forms. Therefore, it appears that the chromosomal differentiation of *S. sanctipauli* 'Djodji' from typical *S. sanctipauli* has been parallelled by morphological evolution, although the extent of this will have to be established by more comprehensive sampling.

Table 6.7

Overall mean, coefficient of variation (CV), and proportion of among-samples variance ($s^2(A)$)

Character	Mean	CV	$s^2(A)$(%)
Antennal Length 1	328.21µm	3.75	52.18
Wing Length 3	1522.33µm	4.13	2.73

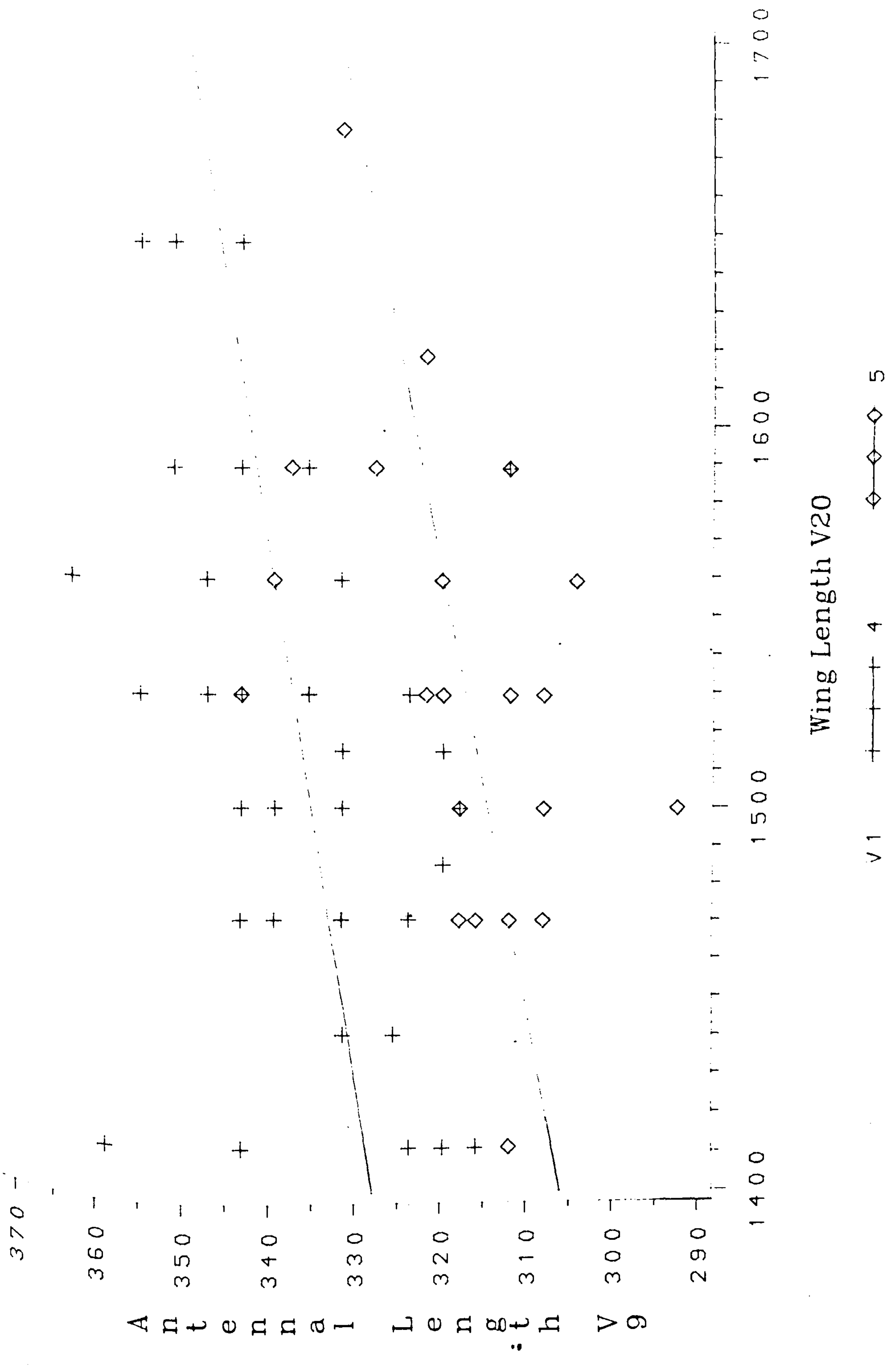
Table 6.8

Standardised Canonical Variate for *S. sanctipauli*

Character	CVI
V9	1.32
V20	-0.61

Figure 6.3

Scatter Plot of Wing Length against Antennal Length for *S. sanctipauli*



6.3.4 SIMULIUM SOUBRENSE AND S. SOUBRENSE 'B'

Eight samples were available within *S. soubrense* s.l., representing a considerable degree of chromosomal variation. One of these samples was *S. soubrense* 'B' (V1=6, N=28) which has been described as a separate species (Post 1986). However, it is chromosomally closest to *S. soubrense* (Chapter three), so to make the present analysis as general as possible, this sample was included in the analysis. In addition to this sample, there were two samples of *S. soubrense* 'Beffa'(V1=14 N=25,V1=29 N=11), one of *S. soubrense* 'Chutes Milo' (V1=16 N=30), three of *S. soubrense* 'Menankaya/Konkoure' (V1=17 N=32,V1=19 N=31,V1=25 N=27) of which V1=19 is chromosomally closest to *S. soubrense* 'Konkoure' *sensu* Quillévéré *et al.* 1982, V1=17 and V1=25 are closer to Menankaya form, see Chapter three). Finally one sample was of a form of unknown chromosomal affinities designated by Dr. R. Baker (personal communication) as *S. soubrense* 'Forest' form (V1=18 N=38).

Wing tuft colour was very variable between the samples of *S. soubrense*, with the null hypothesis of equal wing tuft colouration being rejected at p<0.001 using a Kruskal-Wallis non-parametric one-way analysis of variance. Figure 6.4 shows the histograms for this character for the five subgroups within *S. soubrense s.l.*. The darkest flies belong to *S. soubrense* 'Forest' form, which confirms the original criterion for describing this morphological form. *S. soubrense* 'Beffa', *S. soubrense* 'B' and *S. soubrense* 'Chutes Milo' have similar frequency distributions which are basically symmetrical. The lightest of all the flies was *S. soubrense* 'Menankaya/Konkoure'. This result suggests that the chromosomal evolution within these species and forms has been parallelled by the expression of wing tuft colouration. It was unfortunate that the standard scheme for scoring this character was adopted (Kurtak *et al.* 1981), as this was insensitive to the considerable taxonomic information in this character.

A stepwise discriminant analysis on the 25 character set resulted in an initial subset of 21 characters, which was too large for a parsimonious description of between-samples variation. Mardia's multivariate skewness and kurtosis statistics, together with examination of the marginal distributions and probability plotting of the joint distributions (Chapter five) resulted in a nine character subset:

[V4,V10,V17,V18,V20,V24,V27,V28,V29]

i.e. one thorax, one antennal, three wing and four leg characters. This character subset showed only moderate departures from multivariate normality, but the likelihood-ratio test for equality of dispersion was rejected at $p<0.001$. Despite this, the pooled dispersion matrix was still used, for relative computational simplicity. The rejection of the likelihood ratio test for equal dispersion could be because of the relatively large number of characters, and the slight departure from normality, to which this test is not robust (Seber 1984).

Table 6.9 gives the overall means, pooled coefficient of variation, and proportion of among-samples variation for each character. The CVs suggest that antennal length and the two larger wing measurements might be less variable than the other characters. The proportion of among-samples variation is high for all characters, implying a considerable degree of morphological heterogeneity.

The nine character subset resulted in the matrix of Mahalanobis' squared distances shown as Table 6.10. All these distances were significant at p<0.001. As it stands this matrix is difficult to interpret, so a hierarchical cluster analysis method was used to simplify the matrix. The method used was the group average method (UPGMA, see Chapter three). The cophenetic correlation coefficient was only 0.58 so that the results of the cluster analysis should be treated cautiously.

The dendrogram resulting from the cluster method is shown as Figure 6.5. The tightest cluster on this dendrogram includes the *S. soubrense* 'B' sample and *S. soubrense* 'Chutes Milo' form, with a Mahalanobis' squared distance of 3.27 (significant at p<0.001). The next tightest cluster contains the two *S. soubrense* 'Menankaya' samples, followed by a cluster consisting of the two *S. soubrense* 'Beffa' samples. Beyond these clusters, the inadequacy of the hierarchical representation allows only a limited interpretation, however, the two-cluster solution consists of one cluster containing all the western samples and the other cluster consisting of *S. soubrense* 'Beffa'. *S. soubrense* 'Forest' joins the western cluster last, and may therefore represent a morphologically distinct taxon relative to other *S. soubrense*.

The standardised canonical variates (Table 6.11) show the character combinations of importance in between-samples variation. The scatter of points in the first discriminant plane are shown as Figure 6.6. The first canonical variate has a canonical root of 2.1693, and is a contrast between two positively loading characters, wing width 1 and wing length 3, and two negatively loading characters, basitarsus length 1 and femur length 2, and therefore expresses a relationship between wing size and leg size (i.e. shape). This vector discriminates, at the positive end, the *S. soubrense* 'Menankaya/Konkoure' samples, while at the negative end there is a sample of *S. soubrense* 'Beffa'. The other samples lie intermediate between these samples. The second canonical variate has a canonical root of 1.5858, and contrasts two positively loading characters: antennal length 2 and basitarsus length 2 against two negatively loading characters: wing width 1 and tibia length 2. This vector discriminates principally between *S. soubrense* 'Forest' and *S. soubrense* 'Konkoure'.

The first principal component of the pooled within-samples correlation matrix was a size vector, with coefficients,

[0.33,0.28,0.33,0.33,0.34,0.34,0.35,0.35,0.35]

and accounted for 83.7% of pooled within-samples variation. When the scores along this vector were introduced into the model as a covariable, the canonical roots fell from,

[2.1693,1.5858,0.5681,0.3244,0.1095,0.0964,0.023] to,

[1.829,1.2055,0.3509,0.1626,0.1051,0.041,0.0229].

Therefore, the first two canonical variates are both influenced a little by size variation, but discrimination between samples is still effective when size is controlled.

To conclude, there is extensive morphometric variation within *S. soubrense*/*S. soubrense* 'B'. This differentiation reflects the chromosomal heterogeneity of this group (Post 1986, Chapter three) but does not exactly parallel it. Thus, *S. soubrense* 'B' is chromosomally very distinct but morphologically it is very similar to other *S. soubrense*. The migratory ability of *S. soubrense* is not well known (Garms and Walsh 1987), but the considerable chromosomal and morphological heterogeneity within *S. soubrense s.l.* supports the hypothesis that it does not migrate far, allowing for localised differentiation into new chromosomal and morphological forms.

Table 6.9

Overall mean, coefficient of variation (CV), and proportion of among-samples variance ($s^2(A)$)

Character	Mean	CV	$s^2(A)(\%)$
Thorax Width	872.95µm	5.48	35.96
Antennal Length 2	452.31µm	4.66	49.84
Wing Length 2	443.02µm	5.77	46.7
Wing Width 1	997.16µm	4.92	45.82
Wing Length 3	1478.65µm	4.79	55.23
Basitarsus Length	432.3µm	5.31	43.75
Femur Length 2	659.01µm	5.01	38.48
Tibia Length 2	617.97µm	5.11	37.49
Basitarsus Length	325.77µm	5.30	46.15

Table 6.10

Mahalanobis' squared distances between samples

Sample	6	14	16	17	18	19	25	29
6	0.0							
14	10.04[1]	0.0						
16	3.27[1]	11.16[1]	0.0					
17	9.85[1]	25.95[1]	9.07[1]	0.0				
18	6.82[1]	14.86[1]	7.53[1]	12.49[1]	0.0			
19	6.87[1]	17.6[1]	6.51[1]	6.98[1]	18.46[1]	0.0		
25	10.72[1]	24.21[1]	8.0[1]	4.66[1]	7.26[1]	12.7[1]	0.0	
29	8.52[2]	5.81[3]	8.23[2]	10.71[2]	10.0[2]	10.68[1]	11.22	0.0

[1] $p<0.001$
[2] $0.001<p<0.01$
[3] $p>0.05$

Table 6.11

Standardised Canonical Variates for *S. soubrense s.l.*

Character	CV I	CV II	CV III	CV IV	CV V	CV VI	CV VII
V4	-0.43	0.16	0.17	0.75	1.66	-1.23	-0.75
V10	0.01	1.51	-0.59	-0.88	-0.32	0.23	0.47
V17	0.95	-0.59	1.09	-0.99	-0.05	-0.63	1.37
V18	1.17	-1.26	-0.61	-0.01	-1.1	2.08	-0.08
V20	2.55	0.85	-0.74	-0.47	1.04	-1.87	-1.2
V24	-1.2	0.38	1.14	-0.74	-1.1	0.53	-1.1
V27	-1.7	-0.19	1.86	-2.1	2.59	0.98	-1.98
V28	-0.34	-1.22	-0.91	1.63	-0.94	-0.31	2.58
V29	0.09	1.3	-0.73	2.86	1.23	0.47	0.83
Canonical Root	2.169[1]	1.586[1]	0.568[1]	0.324[1]	0.11[1]	0.096[2]	0.023[3]

[1] $p<0.001$

[2] $0.001<p<0.01$

[3] $p>0.05$

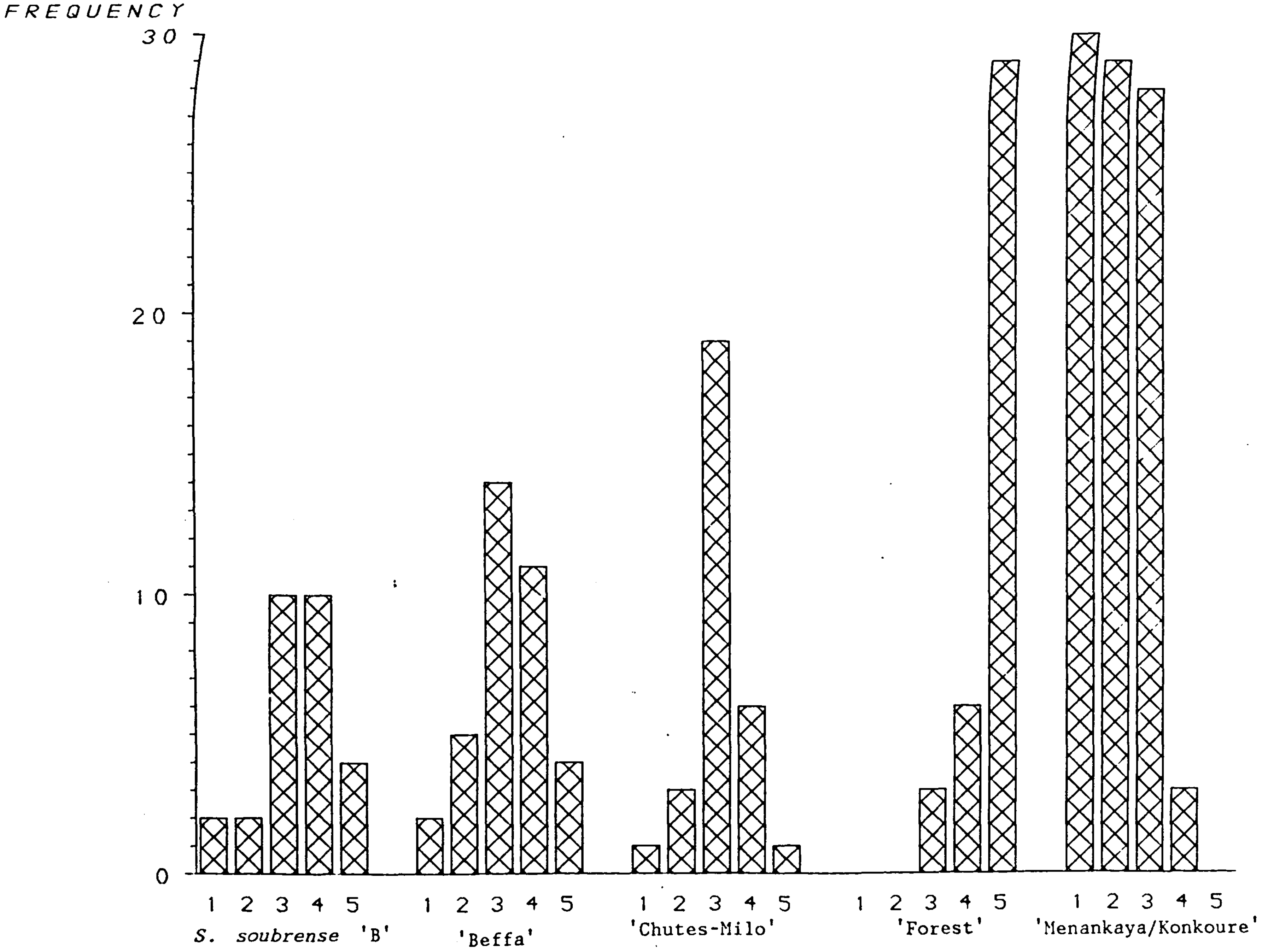
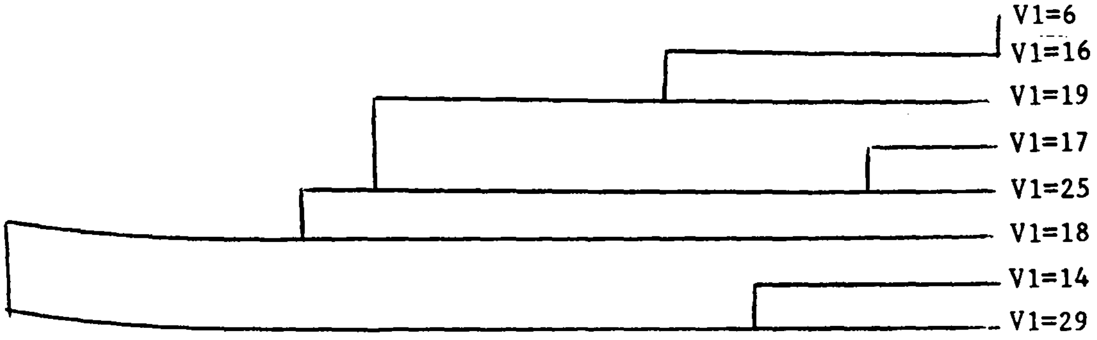


Figure 6.4
Histograms of Wing Tuft Colour for *S. soubrense s.l.*

Figure 6.5

Dendrogram resulting from hierarchical cluster analysis of Mahalanobis' distance matrix between samples of *S. soubrense*



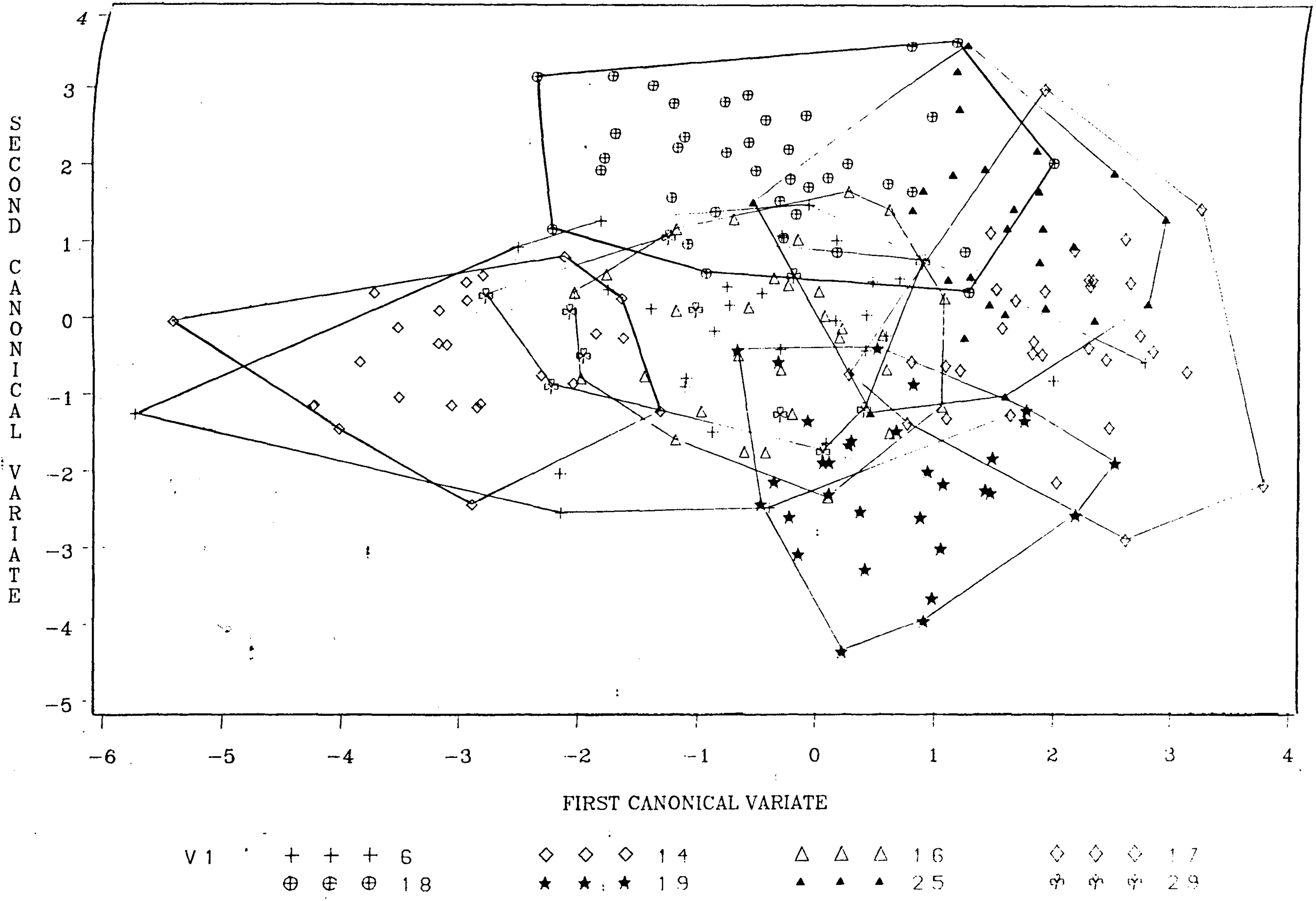


Figure 6.6
Canonical Variates Plot for *S. soubrense*

6.3.5 SIMULIUM SQUAMOSUM

Seven samples of *S. squamosum* were available (Appendix one), four from Togo (V1=1 N=34, V1=12 N=14, V1=13 N=20, V1=30 N=17) and three from Guinea (V1=15 N=40, V1=20 N=34, V1=26 N=13). Chromosomally the *S. squamosum* from the two countries differ, with the eastern *S. squamosum* showing less polymorphism than the western *S. squamosum* (Surtees and Post unpublished data), there is also evidence for allozymic differences between eastern and western *S. squamosum* (Townson *et al.* 1987) and DNA probes differences (Post personal communication).

A stepwise discriminant analysis on the 25·character set resulted in an initial subset of 13 characters:

[V4,V6,V9,V11,V12,V17,V19,V20,V21,V22,V26,V28,V29]

however, this subset showed some departure from multivariate normality and so was further reduced to the seven character subset:

[V9,V11,V19,V20,V22,V28,V29]

i.e. two antennal, two wing, and three leg characters.

The likelihood ratio test for equality of dispersion was not rejected at $p<0.0001$.

Table 6.12 gives the overall means, pooled coefficients of variation and proportion of variance among samples. Some of the among-groups proportions of variance are very large implying considerable morphological differentiation. Wing length tends to be a less variable character, while antennal segment 4 is comparatively very variable, although this may be an artifact resulting from the very small size of this character (Lande 1977).

The matrix of Mahalanobis' squared distances between samples is given as Table 6.13. The most striking feature of this matrix is the proximity of two of the Togolese samples (V1=1 and V1=13, D^2=0.59) but the distance of this pair from all other samples, including V1=12 which was sampled from the same site on other date as sample V1=1.

To examine this in further detail, the Togolese samples were analysed separately. Table 6.14 gives the matrix of Mahalanobis' squared distances between the Togolese samples using the seven character subset. Some of these distances are very large, and suggest considerable multivariate morphometric variation between samples of Togolese *S. squamosum*. However, when the scores along the first principal component of the pooled within-samples correlation matrix (which was a size vector accounting for 55.5% of pooled within-samples variation) were introduced into the model as a covariable, the first canonical root fell from 3.704 to 0.2193. Thus the observed variation within Togolese *S. squamosum* is size variation. This size variation is seasonal in origin, as the principal difference is between the pair: V1=1, V1=13 (sampled in October 1984 and October 1985 respectively) and V1=12 (sampled in March 1985). V1=30 was also sampled in October 1984 and is closest to the other October samples.

The western samples were also analysed separately for comparison, and the matrix of Mahalanobis' distances is shown as Table 6.15. The canonical roots changed from (0.935,0.272) to (0.695,0.24) when the scores along the first principal component of the pooled within-samples correlation matrix were introduced as a covariable, showing that size has some influence on variation between samples of Guinean *S. squamosum*.

The Mahalanobis' squared distance between the eastern *S. squamosum* and the western *S. squamosum* was 4.55 and the canonical root 1.1509. When size was controlled for by the introduction of the scores along the first principal component of the pooled within-samples correlation matrix, the canonical root fell to 0.638, showing that there was little east/west differentiation beyond that already found within these areas.

Table 6.16 gives the standardised canonical variates for the eastern and western samples combined. The first canonical variate is a contrast between wing length 3 and tibia length 2/basitarsus length 2. The samples V1=1 and V1=13 are the main samples to be discriminated along this vector and when size is controlled for, the first canonical root fell from 2.35 to 0.927, indicating that the first canonical variate is dominated by the influence of size variation.

To conclude, there is a large amount of seasonal size variation within Togolese *S. squamosum*. Cheke and Harris (1980) recorded seasonal size variation in *S. damnosum s.l.* from Côte d'Ivoire and Cheke and Denke (1988) also noted seasonal size variation in *S. squamosum* from Togo, so the phenomenon is not unknown in *S. damnosum s.l.*. Unfortunately all of the western samples were collected in the same month of 1986, so it is not possible to establish whether the extensive size variation is a species specific phenomenon, or whether it is a feature only of the eastern *S. squamosum*.

There was no evidence for morphometric differentiation between eastern and western *S. squamosum*, beyond that which was found within countries, once size was controlled for. Wing tuft colour showed some tendency to be differently distributed in the east than the west, with the former being lighter. Garms and Walsh (1987) suggest that *S. squamosum* in the West might be isolated from that in Togo, but morphology does not support the evidence for separate gene pools that is implied by other sources of variation, including chromosomal, biochemical and ecological variation.

Table 6.12

Overall mean, coefficient of variation (CV), and proportion of among-samples variance ($s^2(A)$)

Character	Mean	CV	$s^2(A)(\%)$
Antennal Length 1	279.51μm	5.11	31.8
Antennal Segment 4	37.7μm	6.83	44.05
Wing Width 2	1466.4μm	4.73	56.44
Wing Length 3	1540.1μm	4.28	56.21
Femur Length 1	654.56μm	4.76	65.65
Tibia Length 2	637.4μm	4.71	65.88
Basitarsus Length	343.4μm	4.77	66.2

Table 6.13

Mahalanobis' squared distances between samples

Sample	1	12	13	15	20	26	30
1	0.0						
12	13.21[1]	0.0					
13	0.59[3]	12.13[1]	0.0				
15	10.17[1]	8.01[1]	11.47	0.0			
20	14.9[1]	2.84[1]	15.33	4.76[1]	0.0		
26	8.2[1]	3.64[2]	9.48[1]	4.14[3]	3.06[3]	0.0	
30	3.59[2]	4.96[1]	3.91[1]	6.74[1]	5.17[1]	2.23[2]	0.0

[1] $p<0.001$
[2] $0.001<p<0.01$
[3] $p>0.05$

Table 6.14

Mahalanobis' squared distances between Togo samples

Sample	1	12	13	30
1	0.0			
12	26.93	0.0		
13	0.62[1]	23.01	0.0	
30	5.84	9.36	4.81	0.0

[1] $p>0.05$

Table 6.15

Mahalanobis' squared distances between Guinea samples

Sample	15	20	26
15	0.0		
20	4.12	0.0	
26	3.72	2.72	0.0

Table 6.16

First Three Standardised Canonical Variates for *S. squamosum*

Character	CV I	CV II	CV III
V9	-0.37	0.09	-0.93
V11	0.45	-0.29	-0.12
V19	0.43	-1.67	-0.49
V20	-1.16	1.93	-1.3
V22	0.39	2.13	2.64
V28	0.91	-1.32	-0.83
V29	1.09	-0.63	0.27
Canonical Root	2.35[1]	0.74[1]	0.16[1]

[1] $p<0.001$

[2] $0.001<p<0.01$

6.3.6 SIMULIUM YAHENSE

Four samples of *S. yahense* were available for analysis (Appendix one), one from Côte d'Ivoire (V1=2 N=37), two from Sierra Leone (V1=7 N=14, V1=8 N=25) and one from Guinea (V1=27 N=20). Chromosomally these are differentiated in that the Côte d'Ivoire *S. yahense* is not sex-linked for the inversion 2L-18 (it is fixed), while in Sierra Leone the inversion is X-linked (Vajime and Dunbar 1975, Surtees and Post unpublished data). This is the only chromosomal heterogeneity recorded for *S. yahense*.

A stepwise discriminant analysis resulted in an initial subset of 11 characters:

[V5,V6,V12,V15,V17,V18,V19,V20,V22,V27,V28]

This character set showed considerable departure from multivariate normality, so the data set was reduced further to give the following seven character subset:

[V6,V17,V19,V20,V22,V27,V28]

i.e. one head, three wing and three leg characters. However, this subset also showed multivariate skewness, and the likelihood ratio test for equality of dispersion was rejected at p<0.001. Thus two of the assumptions of the model were rejected, meaning that interpretation of the results obtained should be cautious.

Table 6.17 gives the overall mean, coefficients of variation and proportion of among-samples variance for the seven characters. The CVs are comparable although there is some indication that head width and two of the wing characters are less variable than the other characters. This was supported by Friedman's test for randomised blocks (p<0.01) applied to each sample's variability profile, showing that the profiles are consistent among samples (Bird *et al.* 1981).

Table 6.18 gives the matrix of Mahalanobis' squared distances between samples using the seven character subset. It is apparent that one of the Sierra Leone samples, V1=7 is differentiated relative to the other samples.

The standardised canonical variates are shown in Table 6.19. The first canonical variate is mainly a contrast between femur length 2 and tibia length 2 with some influence from wing length. This vector discriminates sample V1=2 at its positive end from V1=7 at the negative end. The second canonical variate only has a canonical root of 0.447 and is therefore not biologically significant (Campbell 1982). It contrasts two positive characters wing width 2 and wing length 3 with two negative characters: femur length 1 and femur length 2. This vector discriminates the sample V1=8 from the other samples. Figure 6.7 shows the scatter of points in the first principal plane.

The first principal component of the pooled within-samples correlation matrix was a size vector with coefficients,

$$[0.36,0.36,0.37,0.38,0.39,0.40,0.39]$$

and accounted for 86% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable, the canonical roots changed from [1.223,0.447,0.359] to [0.453,0.364,0.245] indicating that the first canonical variate was substantially influenced by size variation.

To conclude, there is significant multivariate morphometric variation between the samples of *S. yahense*, but this variation is principally size variation. There is no indication whether this variation is geographic or temporal in origin, and the chromosomally distinct Côte d'Ivoire sample is not morphologically distinct. This species is known to migrate only very short distances (Garms and Walsh

1987), but despite this there is no evidence for localised morphological differentiation.

Table 6.17

Overall mean, coefficient of variation (CV), and proportion of among-samples variance ($s^2(A)$)

Character	Mean	CV	$s^2(A)$(%)
Head Width	828.68µm	4.53	38.81
Wing Length 2	471.04µm	5.95	40.4
Wing Width 2	1495.93µm	4.76	47.44
Wing Length 3	1576.96µm	4.57	50.47
Femur Length 1	657.64µm	5.61	48.29
Femur Length 2	690.34µm	5.22	51.26
Tibia Length 2	648.42µm	5.59	40.81

Table 6.18

Mahalanobis' squared distances between samples

Sample	2	7	8	27
2	0.0			
7	10.36	0.0		
8	4.11	5.96	0.0	
27	4.32	5.93	3.63	0.0

Table 6.19

Standardised Canonical Variates for *S. yahense*

Character	CV I	CV II	CV III
V6	-0.38	-0.44	-1.17
V17	0.14	0.05	-1.86
V19	0.18	2.03	0.57
V20	1.3	1.2	-1.22
V22	-0.07	-1.83	0.06
V27	2.67	-1.54	1.64
V28	-2.51	0.62	1.85
Canonical Root	1.223[1]	0.447[1]	0.359[1]

[1] $p<0.001$

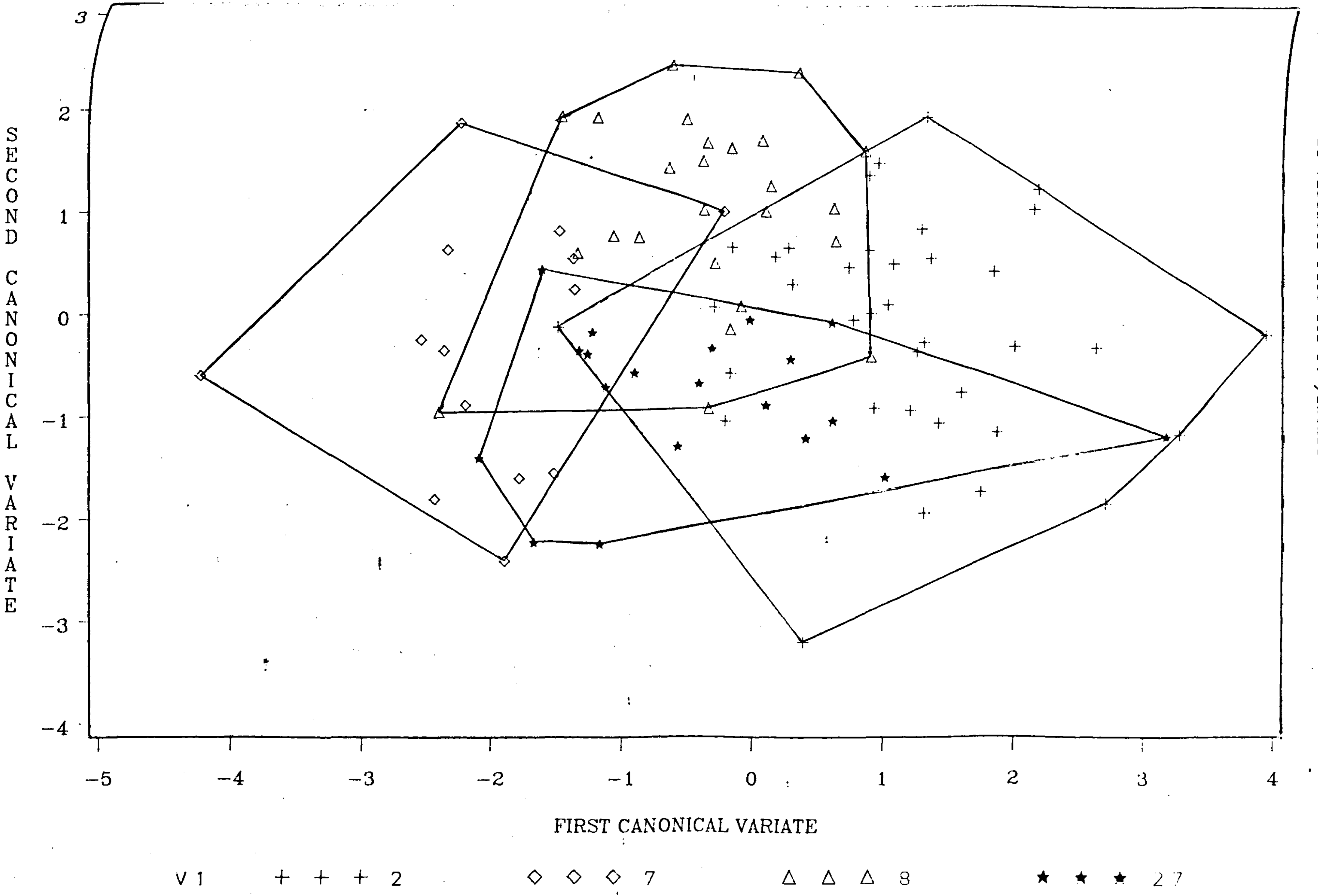


Figure 6.7
Canonical Variates Plot for *S. yahense*

6.4 DISCUSSION

Table 6.20 summaries the results presented in sections 6.3.1-6.3.6. The percentage of variation along the first principal component of the pooled within-samples correlation matrix (the size vector) differed between species, but was consistently high. This reflects the high intercorrelation between most of the characters used to describe variation within species.

The size-free canonical roots can be used as an informal indicator of the amount of shape differentiation within the six taxa. Clearly, *S. soubrense* shows the greatest morphological differentiation between samples, as even the second canonical root is larger than any other canonical root for the other species. The other species show similar degrees of morphological differentiation, which is generally not great, especially when compared with between-species analyses (see Chapters seven and eight).

The influence of size variation on among-samples variation differs greatly between species. *Simulium yahense* and *S. squamosum* are both very heavily influenced by size variation, as shown by the ratio of size-free canonical root to total canonical root, expressed as a percentage. For both of these species, among-samples size variation seems to be the predominant mode of variation. Whether this similarity between the two species is a reflection of their chromosomal relatedness can not be stated on the present data. *S. sirbanum*, *S. damnosum s.s.* and *S. soubrense* are decreasingly influenced by size variation among samples, while *S. sanctipauli* in uninfluenced by size variation. This result could be an artifact of the limited sampling of *S. sanctipauli*, and only additional data will clarify this.

Thus three sources of between-samples variation have been uncovered in the six main taxa of the *S. damnosum* complex. Temporal size variation is strongest is *S. squamosum*, and the extent of this variation is quite surprising. Geographic shape variation is most marked in *S. soubrense*, although this variation does not exactly parallel the considerable chromosomal variation within this taxon. Shape variation was less in *S. sanctipauli*, but may reflect the chromosomal differences between the two samples. The remaining species show varying degrees of size and shape variation, which cannot be attributed to any one specific cause.

To conclude, there is significant morphometric differentiation within the six main taxa of the *S. damnosum* complex. The extent of this varies between species, as does the influence of size variation.

Table 6.20

Summary of intraspecific morphometric analyses of the six main *S. damnosum s.l.*taxa

Analysis[1]	% Variation along first PC[2]	Size-free canonical root	Size influence[3]
6.3.1	74%	0.717,0.122	79.6%,53.4%
6.3.2	85%	0.811,0.243	69.5%,95.8%
6.3.3	69.5%	0.828	99.84%
6.3.4	83.7%	1.829,1.206,0.35	84.3%,76%,61.7%
6.3.5	69.6%	0.927,0.348	39.5%,46.87%
6.3.6	86%	0.453,0.364	35.3%,35.3%

[1] Numbers refer to analyses in section 6.3
[2] This is the proportion of variation along the first principal component of the pooled within-samples correlation matrix, expressed as a percentage
[3] This is the ratio of the size-free canonical root to the total canonical root expressed as a percentage. Only canonical roots considered of biological significance are included.

CHAPTER SEVEN: REGIONAL DISCRIMINATION OF FLIES

7.1 INTRODUCTION

The techniques of multivariate morphometric analysis can be used to study such areas as quantitative genetics, congruence of life stages in classification and geographic or other variation within species (Chapter six). However, the ultimate aim of the multivariate morphometric analysis of the *S. damnosum* complex is to produce a scheme for the allocation of unknown flies which is as accurate and as informative as possible.

The *S. damnosum* complex is very heterogeneous chromosomally (Vajime and Dunbar 1975, Post 1986, Chapter three) with particular variants only being found in certain restricted geographic areas. For example, in Togo there are the chromosomally distinct forms *S. soubrense* 'Beffa' (Meredith *et al.* 1983) and *S. sanctipauli* 'Djodji' (Surtees *et al.* 1988, Chapter two). *Simulium squamosum* in Togo is different chromosomally (Post and Surtees unpublished data), morphologically (Chapter six), and at the molecular level (Post personal communication), from *S. squamosum* further west in Sierra Leone. Within Guinea, there are the forms *S. soubrense* 'Menankaya/Konkoure', not found in Togo (Post personal communication), and from Sierra Leone and Guinea there is the species *S. soubrense* 'B' (Post and Crosskey 1985, Post 1986).

In view of this heterogeneity it was decided to derive the statistics for species identification from samples from Togo and Benin separately from the samples of species from Mali, Guinea, Sierra Leone and Cðte d'Ivoire. The purpose of this chapter is to describe the method used to derive the best character subsets from the full set of characters described in Chapter four, in the context of allocatory discriminant analysis and to present the results obtained using these character subsets, for the two geographic areas. In Chapter eight, a 'global' approach will be taken to the allocation of unknown flies which will not assume prior knowledge of the source of the fly beyond the fact that it is West African *S. damnosum s.l.*. All the methods described in this chapter were also used in Chapter eight.

In both chapters, statistics for pairwise discrimination of species are derived in addition to statistics for the allocation of an unknown fly to any of the reference species simultaneously. The pairwise statistics can be used preferentially in areas where it is known *a priori* that only two species are likely to occur (e.g. River Gban-Houa, Djodji, Togo; Garms and Cheke 1985, Surtees *et al.* 1988, Chapter two) or subsequent to the 'all-species' allocation using typicality probabilities, to improve allocation rates.

7.2 MATERIALS AND METHODS

7.2.1 MATERIALS

The full list of samples is given in Appendix one. These were all screened for outliers or split into their component species using the methods described in Chapter five. Of the 28 characters described in Chapter four, basitarsal spine number was not included in any of the analyses to follow for the reasons given in Chapter five.

7.2.2 SELECTION OF CHARACTER SUBSETS FOR DISCRIMINATION

A major factor which determines the usefulness of an allocation scheme derived using multivariate morphometric analysis is the number of characters which need to be measured or scored for each fly. This analysis began with 28 characters (Chapter four). Such a large character set reduces the usefulness of the method for the following reasons:

1. It takes longer to identify a fly if all characters need to be measured or scored.

2. The probability that a fly will be complete for all the characters is smaller the larger the number of characters.

3. The estimates of the joint (multivariate) distribution are less accurate for a fixed sample size if the dimensionality (number of characters) is increased (Van Ness and Simpson 1976), which in practise has the effect of increasing the observed error rate (Van Vark 1984).

4. The assumptions of multivariate normality and equality of dispersion are less likely to be met for a larger character set than for a smaller one (Seber 1984).

5. More data must be stored in computer memory and more computation is involved in the field identification of a fly.

For these reasons it was necessary to find a method for finding a subset of characters which performed as well as the original full character set, in the context of allocatory discriminant analysis.

There are a number of methods available for character subset generation (Seber 1984), either within widely available computer packages, or as programs not available to this analysis (e.g. ALLOC80, Hermans *et al.* 1982), or as published algorithms (e.g. McKay and Campbell 1982a). For this analysis the only methods available were those within the statistical packages SAS and SPSSX.

The commonly available methods for deriving character subsets in the context of multiple group allocatory discriminant analysis are known to be inadequate (Seber 1984, Habbema and Hermans 1977, McKay and Campbell 1982b). The best of the widely available methods (stepwise discriminant analysis) is known to separate groups which are already well separated at the expense of groups which are close together, and may result in a character subset which performs poorly in an allocation scheme (Habbema and Hermans 1977). This is because the method optimises Wilks' lambda (a statistic describing overall discrimination) at each step in the analysis rather than optimising some statistic relevant to future allocation, such as error rate.

Because programs using the preferred methods (concentrating on error rate), were not available in this analysis, the following method was used for the generation of character subsets:

1. Stepwise discriminant analysis was used to generate an initial subset of characters (excluding wing tuft colouration and abdominal setal colouration because of their consistent non-normality, Chapter four), using the SAS PROC STEPDISC program. This method starts with a model containing no characters. At each step, characters in the model are tested, using Wilks' ratio (see Seber 1984), and any which fall below a preset constant (F-to-leave, set to 0.15, Costanza and Afifi 1979) are removed. Characters not in the model are then tested, and the one which best exceeds another preset constant (F-to-enter, also set to 0.15) is entered. The method stops when none of the characters outside of the model exceed F-to-enter and none of the characters in the model fall below F-to-leave.

The error rate resulting from this method was compared with that using the full 25 character set using the 'leave-one-out' method of error rate estimation (Lachenbruch and Mickey 1968, also known as cross-validation and jackknifing, Seber 1984) for small to medium sized data sets (this error rate was calculated using an inefficient GENSTAT macro which could work only on moderately sized data sets), or, more usually, using the resubstitution method of error rate estimation (using SAS PROC DISCRIM). The 'leave-one-out' method works by removing a fly from the data set, calculating the discriminant functions in its absence and classifying it using the resultant statistics. Resubstitution works by classifying each fly in the data set using discriminant functions derived from the whole data set, including the fly to be classified. This method is more prone to bias than the 'leave-one-out' method, especially for small data sets and large character sets (Seber 1984). If the number of characters was sufficiently low and the error rate acceptable, then this initial subset was accepted. Otherwise:

2. Characters were removed one-by-one from the model, starting with those which had entered it last in the stepwise discriminant analysis. If the removal of the character had a detrimental effect on error rate, then it was returned into the model, however, if its removal had only a slight effect on error rate, then it was rejected. The process was stopped when all of the characters remaining in the model had a detrimental effect on error rate when removed.

As it stands, this method is open to some subjectivity in the choice of a particular subset of characters, because error rate was not the only parameter influencing character choice. For example, if a character increased error rate when it was removed, but it was known to depart from normality, or it was a relatively difficult character to measure (Chapter four), then it was still rejected in spite of the detrimental effect on error rate.

Whilst it is difficult to demonstrate that the subsets of characters obtained by this method are optimal, the results which follow show that they are at least adequate.

7.2.3 PRIOR PROBABILITIES AND NON-NORMAL CHARACTERS

The previous section described the method used for the selection of subsets of characters. This method excluded two characters, wing tuft colouration and abdominal setal colouration because both have previously been shown to be non-normally distributed (Chapter four). However, these characters were included in the analysis because both have been shown to be of considerable taxonomic importance (Dang and Peterson 1980, Garms and Zillman 1984). The purpose of this section is to describe a method for including the taxonomic information these characters contain in the analyses to follow.

The simplest approach to including these characters in the analysis would be to derive the character subset of approximately normally distributed characters using the method described in the previous section and then to introduce the two colour characters into the linear discriminant functions (LDFs), keeping them in the analysis if error rate is improved. However there are two objections to this approach, one of which will be explained in this section, the other in the next section.

The first objection concerns the robustness of the LDF to departures from the assumptions of the model (the most important of which are multivariate normality and equality of dispersion). The LDF is generally robust to departures from normality (Lachenbruch 1975), however it is particularly sensitive to skewed continuous distributions (Seber 1984). The wing tuft colouration is a discrete character because of the method of scoring (Chapter four), and was often skewed, so to be safe the character was not included in the LDF. The abdominal setal colouration was a simple two-state character, and so could not be distributed normally, this character was also not included in case it affected the performance of the resultant LDFs.

An alternative approach, and the one which was adopted in this analysis, is to manipulate the prior probabilities of species membership according to the fly's score for wing tuft colouration and/or abdominal setal colouration. The method used for deriving the prior probabilities was:

1. The species in a particular analysis were tested for equality of wing tuft colouration and abdominal setal colouration using a non-parametric one-way analysis of variance (Wilcoxon two-sample rank sum test for two species, Kruskal-Wallis test for more than two, using SAS PROC NPAR1WAY). If either of the null hypotheses were rejected then:

2. For the particular analysis, LDFs were calculated using abdominal setal colouration and/or wing tuft colouration . An artificial data set was created of the ten (wing tuft colouration and abdominal setal colouration), five (wing tuft colouration), or two (abdominal setal colouration) possible outcomes of the two characters (Chapter four); this artificial data set was then classified using the LDFs. The resultant posterior probabilities of species membership were then used as prior probabilities of species membership for a fly with that particular combination of characters.

The prior probabilities affected allocation in the following way. Mahalanobis' distance is calculated to each of the reference species and substituted in the following equation:

$$\pi_i \exp(-\tfrac{1}{2} D_i^2) \div \Sigma \pi_i \exp(-\tfrac{1}{2} D_i^2)$$

where π_i is the prior probability of belonging to the i-th species (i=1...g) and D_i^2 is the distance of the fly from the i-th species. This quantity is the posterior probability of belonging to the i-th species, and the g posterior probabilities sum to one. The fly is allocated to the species for which its posterior probability of membership is the largest.

7.2.4 ALLOCATION SCHEMES The second objection to including a non-normal character in a LDF concerns the method used for allocation of unknown flies. The usual method for allocation involves calculating the fly's score on each LDF and allocating it to the species on which its score is highest (Seber 1984). This is equivalent to assigning it to the species to which it has the smallest Mahalanobis' distance (D_i^2), and also equivalent to assigning it to the species for which its posterior probability is highest. This is known as forced allocation (Campbell 1984).

An alternative approach is to calculate a fly's typicality probability of species membership, but for this approach to be used, the data need to conform approximately to multivariate normality and equal dispersion; hence the objection to including the non-normal characters in the analysis.

The typicality probability is the probability associated with the observed Mahalanobis' distance of the fly to each of the reference species. There are different ways of calculating typicality probabilities (Campbell 1984, Ambergen and Schaafsma 1984), but in this analysis it was decided to use the method of Ambergen and Schaafsma (1984) because this method allows the construction of approximate confidence intervals for the observed distance from each reference species.

The confidence intervals were calculated in the following way:

1. An unbiased estimate of Mahalanobis' distance was calculated:

$$(n-g-p-1)n^{-1} D_i^{2} - n^{-1}_i$$

2. An estimate of the variance of this unbiased distance was then calculated:

$$(n-g-p-3)^{-1}\{2D^{4} + 4(n-g-1)n^{-1}_i D^{2}+2p(n-g-1)n^{-2}_i\}$$

where n is the total sample size on which the dispersion matrix was based, g is the number of reference species, p is the number of characters and n_i is the sample size of the i-th species.

The following rules were adhered to throughout the analyses in this chapter and in Chapter eight for the typicality probability allocation of flies:

1. If all of the confidence intervals straddled α=0.01 then the fly was classified as untypical of the range of reference species, clearly 1% of flies can be expected to be unallocated in this way. Other values of α could be chosen depending on the requirements of a specific analysis.

2. If the confidence intervals for the probability of belonging to one species was greater than for all the others, and did not include 0.01, and none of the other confidence intervals included probabilities of species membership covered by this one, then the fly was classified into that species.

3. If two confidence intervals included a common range of probabilities of species membership, then the fly was allocated onto

the overlap of the two species, unless one the the confidence intervals contained 0.01, in which case the fly was classified into the species whose probability did not include 0.01.

4. If more than two of the higher probability confidence intervals included a common range of probabilities, then the fly was regarded as unclassified, overlapping.

The principal advantage that using typicality probability for allocation has over forced allocation is that a fly which is atypical of all the reference species is not forced into the species to which it closest. Also, a fly which lies on the overlap region of two or more species is identified as such and is not forced into the species to which it is closest. Because an unbiased estimate of Mahalanobis' distance is used in calculating the typicality probabilities, then some of the bias of a finite sample size is corrected for.

In summary, typicality probability allocation yields more information about the affinities of a particular fly to a particular species, without direct reference to the other species. Rather than having a simple allocate / don't allocate rule, this method of allocation allows a more informed and biologically more realistic allocation decision to be made.

7.2.5 OTHER METHODS

Apart from these methods, the other multivariate statistical techniques used in this analysis have been described in previous sections: Principal components analysis (Chapter three), Canonical variates analysis (Chapter six), and cluster analysis (Chapter three).

The importance of size in discrimination was assessed by introducing the scores along the first principal component of the pooled within-species correlation matrix into the model as a covariable in a multivariate analysis of covariance (MANCOVA). The first principal component of the pooled within-species correlation matrix is usually a size vector in morphometric studies (Blackith and Reyment 1971), meaning that a unit change in one character is accompanied by a unit change of the same sign in all (Rao 1964). The pooled within-species correlation matrix was used rather than the usual (Total) correlation matrix so that between-species variation and within-species variation were not confused. The analysis was performed using a SAS PROC MATRIX routine to calculate the pooled within-species correlation matrix and to run the principal components analysis whilst SAS PROC GLM was used for the multivariate analysis of covariance.

The assumption of equal dispersion was tested using the likelihood ratio test (Seber 1984). It is known (Layard 1974) that this test is very sensitive even to slight non-normality, and a rejection of the null hypothesis is as likely to be due to departure from multivariate normality as it is to departure from equal dispersion. Despite this criticism of the test, it was still used in this analysis as an indicator of the reliability of the sample statistics obtained. If the null hypothesis was rejected, then caution should be used in the interpretation of results using those statistics.

Strictly, if the null hypothesis is rejected, then each species' dispersion matrix should be calculated and a quadratic discriminant function (QDF) used (Lachenbruch 1975, Seber 1984), or typicality probabilities calculated using the individual species' dispersion matrices rather than the pooled within-species dispersion matrix (Ambergen and Schaafsma 1984, Campbell 1984). This option was not investigated in this analysis for the following reasons:

1. The QDF has poor small sample properties (Seber 1984);

2. The QDF is very sensitive to departures from multivariate normality;

3. More parameters need to be estimated (each species' dispersion matrix rather than the pooled dispersion matrix, as well as the mean vectors), with the result that for the same sample size the precision of estimation of these parameters is reduced (Van Ness and Simpson 1976), which has the effect of broadening the confidence intervals for the typicality probability of species membership (Ambergen and Schaafsma 1984);

4. Greater computation is needed to allocate an unknown fly and more statistics need to be stored in computer memory, which both work against the field applicability of the method.

7.3 RESULTS AND DISCUSSION, TOGO AND BENIN

7.3.1 SPECIES PAIR DISCRIMINATION

In order to deal with the situation where typicality probability allocation results in a fly being allocated to the overlap between two species, then re-allocating the fly in an allocation scheme involving just those two species should increase the chance of correct allocation. Also, in some areas it is known *a priori* that only two species will be expected (e.g. River Gban-Houa, Togo, where *S. sanctipauli* 'Djodji' and *S. squamosum* are the only two species likely to be found), so it would be advantageous to calculate the probability of species membership only in relation to these two species.

For these reasons, statistics for species-pair discrimination will be presented first, followed by those for overall discrimination.

7.3.1.1 Discrimination of Simulium soubrense 'Beffa' and S. damnosum

The squared Mahalanobis' distance between species using the 25 character set was 31.32 (unbiased D^2=20.74, e_{act}=0.0114) with a single fly misallocated using resubstitution (e_{res}=0.0127). A stepwise discriminant analysis generated a nine character subset which gave a Mahalanobis' squared distance of 26.121 (unbiased D^2=22.73, e_{act}=0.0086) and no change in the resubstituted error rate (e_{res}=0.0127).

The dimension reduction procedure described in section 7.2.1 resulted in a six character subset:

[V9,V14,V18,V19,V27,V28]

i.e. two antennal, two wing and two leg characters, with a Mahalanobis' squared distance of 20.11 (unbiased D^2=18.282, e_{act}=0.0163) and a single misallocated flies using resubstitution

(e_{res}=0.0127), which was a *S. damnosum* classified as *S. soubrense* 'Beffa'.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in one misallocated fly (1.27%), two atypical flies (2.53%), one overlapping fly (1.27%) and 75 correctly allocated flies (94.94%).

The null hypothesis of equal wing tuft colouration was rejected at p<0.001 using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were calculated using the method described in section 7.2.3 and are shown as Table 7.1. When the prior probabilities were adjusted the number of flies misallocated remained at one (e_{res}=0.0127).

The null hypothesis of equal dispersion was not rejected at p<0.001 using the likelihood ratio test, legitimising the pooling of the dispersion matrices.

The standardised canonical variate (Table 7.2) shows that the main character discriminating these species is antennal length 1, with a relationship between femur length 2 and tibia length 2 also having some effect.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.37,0.20,0.45,0.43,0.47,0.47]$$

and accounted for 66.4% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 5.1175 to 4.8085 showing that size has only a minor influence in the morphometric differentiation between these species.

The mean vectors are shown as Table 7.3, the linear discriminant functions as Table 7.4 and the pooled within-species dispersion matrix as Table 7.5.

To conclude, there is significant multivariate morphometric differentiation between *S. soubrense* 'Beffa' and *S. damnosum*. This differentiation involves mainly shape variation as in the absence of size there is still good discrimination. Garms and Cheke (1985) considered that these two species were distinguishable using thorax antennal ratios, although their histograms show some overlap. Thus, the character set derived in this analysis is an improvement over previous methods. The six character subset can be expected to classify flies to their correct species in nearly 95% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration does not improve allocation rate, so it is not recommended that this should be done, except in cases of doubt.

Table 7.1.

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. soubrense 'Beffa'	*S. damnosum s.s.*
1	0.0106	0.9894
2	0.2986	0.7014
3	0.9442	0.0558
4	0.9985	0.0015
5	1.0	0.0

Table 7.2

Standardised Canonical Variate for *S. soubrense* 'Beffa' and *S. damnosum s.s.*

Character	Canonical Variate
Antennal Length 1	1.5086
Antennal Segment 7	0.6834
Wing Width 1	-0.4247
Wing Width 2	-0.3936
Femur Length 2	0.8102
Tibia Length 2	-0.9750

Table 7.3

Mean Vectors for species *S. soubrense* 'Beffa' and *S. damnosum s.s.*

Character	*S. soubrense* 'Beffa'	*S. damnosum s.s.*
Antennal Length 1	294.17916667	259.07790698
Antennal Segment 7	43.12444444	36.62325581
Wing Width 1	961.54361111	996.17825581
Wing Width 2	1341.42277778	1419.20348837
Femur Length 2	658.46000000	653.67348837
Tibia Length 2	611.72000000	619.80558140

Table 7.4

Linear Discriminant functions for species *S. soubrense* 'Beffa' and *S. damnosum s.s.*

	S. soubrense 'Beffa'	*S. damnosum s.s.*
CONSTANT	-339.40213292	-305.70354140
V9	0.79149552	0.48709948
V14	2.88058489	2.16075641
V18	0.26957828	0.30973237
V19	-0.03963553	-0.01919285
V27	0.37263982	0.25241827
V28	-0.21198021	-0.06491194

Table 7.5

Pooled within-species dispersion matrix for species *S. soubrense* 'Beffa' and *S. damnosum s.s.*

CHARACTER	V9	V14	V18	V19	V27	V28
V9	186.83945957	12.65833021	327.11915002	507.64767399	262.81697161	264.95975717
V14	12.65833021	7.60652510	34.33590606	41.91742685	22.33336249	22.27992751
V18	327.11915002	34.33590606	1973.49030357	2759.17544091	1076.58329496	1038.32222621
V19	507.64767399	41.91742685	2759.17544091	6012.41826947	1779.96639190	1650.71783588
V27	262.81697161	22.33336249	1076.58329496	1779.96639190	919.26563866	830.83295536
V28	264.95975717	22.27992751	1038.32222621	1650.71783588	830.83295536	878.65357221

7.3.1.2 Discrimination of Simulium soubrense 'Beffa' and S. sanctipauli 'Djodji'

The squared Mahalanobis' distance between species using the 25 character set was 36.42 (unbiased D^2=20.64, e_{act}=0.0116) with no flies misallocated using resubstitution (e_{res}=0.0). A stepwise discriminant analysis produced a nine character subset which gave a Mahalanobis' squared distance of 18.261 (unbiased D^2=15.22, e_{act}=0.0256) with a single fly misallocated using resubstitution (e_{res}=0.0161).

The dimension reduction technique described in section 7.2.1 resulted in a six character subset:

[V6,V10,V17,V20,V27,V29]

i.e. one head, one antennal, two wing and two leg characters, with a Mahalanobis' squared distance of 10.91 (unbiased D^2=8.92, e_{act}=0.068) and two misallocated flies using resubstitution (e_{res}=0.0645), one into each species.

Allocation using typicality probability of species membership with atypicality defined at $\alpha=0.01$ resulted in two misallocated flies (3.23%), no atypical flies, six overlapping flies (9.68%) and 54 correctly allocated flies (87.1%).

The null hypothesis of equal wing tuft colouration was not rejected at $p<0.001$ using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were therefore not calculated.

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, legitimising the pooling of the dispersion matrices.

The standardised canonical variate (Table 7.6) shows that the primary contrast discriminating these species is the positively loading characters antennal length 2, wing length 3, and basitarsus length 2, **against** a negatively loading character head width.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.41,0.37,0.39,0.42,0.44,0.43]$$

and accounted for 80.1% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 2.5402 to 1.2334 showing that size has a significant influence in discriminating between these species.

The mean vectors are shown as Table 7.7, the linear discriminant functions as Table 7.8 and the pooled within-species dispersion matrix as Table 7.9.

To conclude, there is significant multivariate morphometric differentiation between *S. soubrense* 'Beffa' and *S. sanctipauli*

'Djodji'. Chromosomally, these two species are close (Chapter three, Post 1986, Meredith *et al.* 1984), so it is not surprising that they are close together morphometrically. Previous work using morphometrics of the *S. damnosum* complex have not distinguished between these species (e.g. Garms and Cheke 1985), but recent work on the epidemiological importance of *S. sanctipauli* 'Djodji' (Cheke and Denke 1988) shows that identification of *S. sanctipauli* 'Djodji' is important. Much of the separation between these species involves size, so that once a wider range of variation has been examined, the error rate will probably increase.

Table 7.6

Standardised Canonical Variate for *S. soubrense* 'Beffa' and *S. sanctipauli* 'Djodji'

Character	Canonical Variate
Head Width	-1.3623
Antennal Length 2	1.0467
Wing Length 2	-0.6889
Wing Length 3	1.1002
Femur Length 2	-0.5752
Basitarsus Length	1.8089

Table 7.7

Mean Vectors for species *S. soubrense* 'Beffa' and *S. sanctipauli* 'Djodji'

Character	*S. soubrense* 'Beffa	*S. sanctipauli*
Head Width	815.04746667	838.51038462
Antennal Length 2	442.52666667	476.10461538
Wing Length 2	438.70000000	467.96769231
Wing Length 3	1409.63763889	1534.97576923
Femur Length 2	658.46000000	697.88307692
Basitarsus Length 2	320.77500000	350.32500000

Table 7.8

Linear Discriminant functions for species *S. soubrense* 'Beffa' and *S. sanctipauli* 'Djodji'

Characte	*S. soubrense* 'Beffa'	*S. sanctipauli* 'Djodji'
CONSTANT	-287.63555221	-331.59567479
V6	0.19155864	0.08060720
V10	0.66202707	0.79023419
V17	-0.13911607	-0.21223907
V20	0.03484206	0.07186381
V27	0.03900777	-0.01079215
V29	0.35042500	0.61631544

Table 7.9

Pooled within-species dispersion matrix for species *S. soubrense* 'Beffa' and *S. sanctipauli*

Character	V6	V10	V17	V20	V27	V29
V6	1408.89212946	532.60455016	649.46063498	2163.13096102	989.61622415	482.10925950
V10	532.60455016	400.41592410	327.60362462	908.01100124	454.97687385	204.33990000
V17	649.46063498	327.60362462	695.39090769	1416.44027244	656.58308308	340.11345000
V20	2163.13096102	908.01100124	1416.44027244	5112.05735181	1891.56033064	968.23234375
V27	989.61622415	454.97687385	656.58308308	1891.56033064	978.09372923	442.22190000
V29	482.10925950	204.33990000	340.11345000	968.23234375	442.22190000	255.33056250

7.3.1.3 Discrimination of Simulium soubrense 'Beffa' and S. squamosum

The squared Mahalanobis' distance between species using the 25 character set was 29.21 (unbiased D^2=22.83, e_{act}=0.0073) with a single fly misallocated using resubstitution (e_{res}=0.0083). A stepwise discriminant analysis produced a 12 character subset which gave a Mahalanobis' squared distance of 27.38 (unbiased D^2=24.39, e_{act}=0.0068) with one fly misallocated using resubstitution (e_{res}=0.0083).

The dimension reduction technique described in section 7.2.1 resulted in a four character subset:

[V6,V10,V18,V20]

i.e. one head, one antennal and two wing characters, with a Mahalanobis' squared distance of 19.0 (unbiased D^2=18.2, e_{act}=0.00165) and one misallocated fly using resubstitution

(e_{res}=0.0083), which was a *S. squamosum* misallocated into *S. soubrense* 'Beffa'

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in one misallocated fly (0.83%), two atypical flies (1.65%) no overlapping flies and 118 correctly allocated flies (97.52%).

The null hypothesis of equal wing tuft colouration was rejected at p<0.001 using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were calculated using the method described in section 7.2.3 and are shown as Table 7.10. When the prior probabilities were adjusted according to wing tuft colouration, then three flies were misallocated using resubstitution (e_{res}=0.0248).

The null hypothesis of equal dispersion was rejected at p<0.001 using the likelihood ratio test, however the pooled dispersion matrix was still used for the practical and statistical reasons given in section 7.2.5.

The standardised canonical variate (Table 7.11) shows that the main contrast discriminating these species is the positively loading wing characters wing width 1 and wing length 3, and the negatively loading character antennal length 2.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.49,0.48,0.51,0.52]

and accounted for 84.4% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 4.0397 to 2.4333 showing that size has an important influence in discriminating between these species, but that shape differences are as significant.

The mean vectors are shown as Table 7.12, the linear discriminant functions as Table 7.13 and the pooled within-species dispersion matrix as Table 7.14.

To conclude, there is significant multivariate morphometric discrimination between *S. soubrense* 'Beffa' and *S. squamosum.* This differentiation involves a combination of size and shape variation. Garms and Cheke (1985) consider this species pair to be very difficult to distinguish using thorax antennal ratios, so the character subset derived in this analysis is a considerable improvement. The four character subset can be expected to classify flies to their correct species in over 97% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration does not improve allocation rate, due to *S. soubrense* 'Beffa' having wing tufts falling into all five colour categories (Chapter four). Thus, *S. soubrense* 'Beffa' with colour categories one or two are strongly penalised against 'own-group' membership. Such flies were found in this data set, and have been reported in previous work (Meredith *et al.* 1984) Therefore it is not recommended that the prior probabilities should be altered.

Table 7.10

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. soubrense 'Beffa'	*S. squamosum*
1	0.0062	0.9938
2	0.2513	0.7487
3	0.9473	0.0527
4	0.9990	0.0010
5	1.0000	0.0000

Table 7.11

Standardised Canonical Variate for *S. soubrense* 'Beffa' and *S. squamosum*

Character	Canonical Variate
Head Width	-0.4915
Antennal Length 2	-1.3310
Wing Width 1	1.3799
Wing Length 3	1.1973

Table 7.12

Mean Vectors for species *S. soubrense* 'Beffa' and *S. squamosum*

Character	*S. soubrense* 'Beffa'	*S. squamosum*
Head Width	815.04746667	839.65947529
Antennal Length 2	442.52666667	425.89835294
Wing Width 1	961.54361111	1087.65100000
Wing Length 3	1409.63763889	1574.30000000

Table 7.13

Linear Discriminant functions for species *S. soubrense* 'Beffa' and *S. squamosum*

Characte	*S. soubrense* 'Beffa'	*S. squamosum*
CONSTANT	-190.90523318	-194.88882867
V6	0.28684170	0.23993834
V10	0.43286097	0.20806258
V18	-0.02160380	0.04978747
V20	-0.01614504	0.02893111

Table 7.14

Pooled within-species dispersion matrix for species *S. soubrense* 'Beffa' and *S. squamosum*

Character	V6	V10	V18	V20
V6	1975.76775752	799.57472278	2167.22308444	3156.97831053
V10	799.57472278	613.25246529	1153.60869894	1694.35477955
V18	2167.22308444	1153.60869894	3782.23936992	4815.58203830
V20	3156.97831053	1694.35477955	4815.58203830	7760.93115189

7.3.1.4 Discrimination of Simulium damnosum and S. sanctipauli 'Djodji'

The squared Mahalanobis' distance between species using the 25 character set was 52.42 (unbiased D^2=32.08, e_{act}=0.0023) with a single fly misallocated using resubstitution (e_{res}=0.015). A stepwise discriminant analysis produced a 7 character subset which gave a Mahalanobis' squared distance of 43.15 (unbiased D^2=37.996, e_{act}=0.001) with no flies misallocated using resubstitution (e_{res}=0.0).

The dimension reduction technique described in section 7.2.1 resulted in a four character subset:

[V6,V9,V13,V29]

i.e. one head, two antennal and one leg character, with a Mahalanobis' squared distance of 36.4 (unbiased D^2=33.68, e_{act}=0.00186) and no misallocated flies using resubstitution (e_{res}=0.0).

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in no misallocated flies, one atypical fly (1.45%), no overlapping flies and 68 correctly allocated flies (98.55%).

The null hypothesis of equal wing tuft colouration was rejected at p<0.001 using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were calculated using the method described in section 7.2.3 and are shown as Table 7.15. When the prior probabilities were adjusted according to wing tuft colouration, no flies were misallocated using resubstitution (e_{res}=0.0).

The null hypothesis of equal dispersion was not rejected at p<0.001 using the likelihood ratio test, legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.16) shows that three of the characters load positively (antennal length 1, antennal segment 6, basitarsus length 2) and one negatively (head width). Antennal length relative to other characters is the most important discriminatory character.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.56,0.52,0.36,0.54]

and accounted for 58.6% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 8.803 to 3.351 showing that size has a significant influence in discriminating between these species, but that shape differences are still very important in the absence of size variation.

The mean vectors are shown as Table 7.17, the linear discriminant functions as Table 7.18 and the pooled within-species dispersion matrix as Table 7.19.

To conclude, there is significant multivariate morphometric differentiation between *S. damnosum* and *S. sanctipauli* 'Djodji'. This differentiation involves a combination of size and shape variation, but shape is very important as a discriminatory character. The four character subset can be expected to classify flies to their correct species in nearly 99% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration does not improve allocation rate, so it is recommended that this should be done only in cases of doubt following typicality probability allocation.

Table 7.15

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. damnosum s.s.	*S. sanctipauli* 'Djodji'
1	1.0000	0.0000
2	0.9738	0.0262
3	0.0138	0.9862
4	0.0000	1.0000
5	0.0000	1.0000

Table 7.16

Standardised Canonical Variate for *S. damnosum s.s.* and *S. sanctipauli* 'Djodji'

Character	Canonical Variate
Head Width	-0.7509
Antennal Length 1	1.8187
Antennal Segment 6	1.2273
Basitarsus Length	0.5989

Table 7.17

Mean Vectors for species *S. damnosum s.s.* and *S. sanctipauli* 'Djodji'

	S. damnosum s.s.	*S. sanctipauli*
Head Width	812.94496744	838.51038462
Antennal Length 1	259.07790698	317.62500000
Antennal Segment 6	37.05581395	49.64769231
Basitarsus Length 2	320.16279070	350.32500000

Table 7.18

Linear Discriminant functions for species *S. damnosum s.s.* and *S. sanctipauli* 'Djodji'

	S. damnosum s.s.	*S. sanctipauli* 'Djodji'
CONSTANT	-376.76328848	-479.42883704
V6	0.35764041	0.22851342
V9	0.65331761	1.00564940
V13	1.03505757	2.11469538
V29	0.79699938	0.97862744

Table 7.19

Pooled within-species dispersion matrix for species *S. damnosum s.s.* and *S. sanctipauli* 'Djodji'

Character	V6	V9	V13	V29
V6	1091.36893129	211.47417506	32.18249829	315.12796654
V9	211.47417506	155.38130988	14.73961229	83.65941644
V13	32.18249829	14.73961229	9.39496878	7.76922839
V29	315.12796654	83.65941644	7.76922839	181.63442858

7.3.1.5 Discrimination of Simulium damnosum and S. squamosum

The squared Mahalanobis' distance between species using the 25 character set was 16.96 (unbiased D^2=13.46, e_{act}=0.0331) with two flies misallocated using resubstitution (e_{res}=0.01563). A stepwise discriminant analysis produced an 11 character subset which gave a Mahalanobis' squared distance of 15.304 (unbiased D^2=13.85, e_{act}=0.03139) with two flies misallocated using resubstitution (e_{res}=0.01563).

The dimension reduction technique described in section 7.2.2 resulted in a nine character subset:

[V3,V4,V9,V12,V17,V19,V23,V28,V29]

i.e. two thorax, two antennal, two wing and three leg characters, with a Mahalanobis' squared distance of 13.32 (unbiased D^2=12.26, e_{act}=0.04) and three misallocated flies using resubstitution (e_{res}=0.0234), two misallocated into *S. squamosum*, the other into *S. damnosum*.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in two misallocated flies (1.56%), three atypical flies (2.34%) six overlapping flies (4.69%) and 117 correctly allocated flies (91.41%).

The null hypothesis of equal wing tuft colouration was not rejected at $p<0.001$ using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were therefore not calculated.

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.20) shows that the characters of most importance in discrimination are the positively loading characters: tibia length 1, basitarsus length 2, and the negatively loading characters: thorax width and tibia length 2.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.33,0.36,0.28,0.20,0.34,0.34,0.37,0.37,0.36]$$

and accounted for 73.5% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 3.0186 to 2.0499 showing that size has a significant influence in discriminating between these species, but that shape differences are more important.

The mean vectors are shown as Table 7.21, the linear discriminant functions as Table 7.22 and the pooled within-species dispersion matrix as Table 7.23.

To conclude, there is significant multivariate morphometric differentiation between *S. damnosum* and *S. squamosum*. This differentiation involves a combination of size and shape variation, but shape is very important as a discriminatory character. In previous work (Garms and Cheke 1985) these species were considered to be very difficult to distinguish using thorax antennal ratios, so that the character subset derived in this analysis is an improvement. However, nine characters is quite large, so it might be preferable to rely on the overall discrimination statistics described in section 7.3.2 for identification of this species pair.

Table 7.20

Standardised Canonical Variate for *S. damnosum s.s.* and *S. squamosum*

Character	Canonical Variate
Thorax Length	0.4729
Thorax Width	-1.6522
Antennal Length 1	0.2601
Antennal Segment 5	0.4221
Wing Length 2	-0.3824
Wing Width 2	0.4267
Tibia Length 1	1.1911
Tibia Length 2	-1.3827
Basitarsus Length	1.9785

Table 7.21

Mean Vectors for species *S. damnosum s.s.* and *S. squamosum*

	S. damnosum s.s.	*S. squamosum*
Thorax Length	631.52014884	680.26941882
Thorax Width	907.70187907	927.77268706
Antennal Length 1	259.07790698	282.91058824
Antennal Segment 5	34.74883721	39.35905882
Wing Length 2	454.47069767	476.02447059
Wing Width 2	1419.20348837	1522.57300000
Tibia Length 1	689.37209302	749.57647059
Tibia Length 2	619.80558140	662.40564706
Basitarsus Length 2	320.16279070	357.69882353

Table 7.22

Linear Discriminant functions for species *S. damnosum s.s.* and *S. squamosum*

	S. damnosum s.s.	*S. squamosum*
CONSTANT	-238.68582302	-270.07005463
V3	-0.15676510	-0.12640568
V4	0.17864883	0.07054880
V9	0.90987717	0.95944660
V12	0.41001607	0.82246741
V17	0.03658488	-0.00681295
V19	0.26517257	0.28239191
V23	-0.21673943	-0.13174467
V28	0.02150117	-0.09219862
V29	-0.28934526	-0.02802901

Table 7.23

Pooled within-species dispersion matrix for species *S. damnosum s.s.* and *S. squamosum*

Character	V3	V4	V9	V12	V17	V19	V23	V28	V29
V3	2719.42384454	2343.02705442	463.99135654	58.31889137	1161.84786378	2760.74649043	1741.21253327	1664.75317336	828.83677365
V4	2343.02705442	3045.02475869	497.58785243	74.78838804	1311.54845189	3398.17483861	2092.51304810	1917.48843647	999.41286012
V9	463.99135654	497.58785243	240.87287287	19.86120827	259.72038999	567.86963583	427.40714813	405.64447794	209.29048183
V12	58.31889137	74.78838804	19.86120827	9.24125212	26.34928804	99.18742789	57.79532002	51.22733913	24.78239719
V17	1161.84786378	1311.54845189	259.72038999	26.34928804	937.10625857	1815.13373298	1065.76385538	994.32845783	521.21037749
V19	2760.74649043	3398.17483861	567.86963583	99.18742789	1815.13373298	5824.41303244	2769.46224275	2448.53392002	1298.53001096
V23	1741.21253327	2092.51304810	427.40714813	57.79532002	1065.76385538	2769.46224275	1815.06318296	1606.66280946	824.62797616
V28	1664.75317336	1917.48843647	405.64447794	51.22733913	994.32845783	2448.53392002	1606.66280946	1574.31333928	771.22664678
V29	828.83677365	999.41286012	209.29048183	24.78239719	521.21037749	1298.53001096	824.62797616	771.22664678	450.30670911

7.3.1.6 Discrimination of Simulium sanctipauli 'Djodji' and S. squamosum

The squared Mahalanobis' distance between species using the 25 character set was 31.1 (unbiased D^2=23.68, e_{act}=0.0075) with no flies misallocated using resubstitution (e_{res}=0.0). A stepwise discriminant analysis produced a seven character subset which gave a Mahalanobis' squared distance of 24.66 (unbiased D^2=22.85,

e_{act}=0.00842) with two flies misallocated using resubstitution (e_{res}=0.018).

The dimension reduction technique described in section 7.2.2 resulted in a three character subset:

[V10,V18,V22]

i.e. one antennal, one wing and one leg character, with a Mahalanobis' squared distance of 15.5 (unbiased D^2=15.12, e_{act}=0.00259) and two misallocated flies using resubstitution (e_{res}=0.018), both misallocated into *S. sanctipauli* 'Djodji'.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in two misallocated flies (1.8%), no atypical flies, no overlapping flies and 109 correctly allocated flies (98.2%).

The null hypothesis of equal wing tuft colouration was rejected at $p<0.001$ using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.24. When the prior probabilities of species membership were adjusted for each wing tuft colouration category, the resubstituted error rate fell to one fly misallocated into *S. sanctipauli* 'Djodji' (e_{res}=0.009).

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.25) shows that the most important character in discrimination between these species is antennal length 2, which contrasts with the other two characters.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.55,0.58,0.59]

and accounted for 85.6% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 2.8403 to 2.8257 showing that size has negligible influence on discriminating between these species.

The mean vectors are shown as Table 7.26, the linear discriminant functions as Table 7.27 and the pooled within-species dispersion matrix as Table 7.28.

To conclude, there is significant multivariate morphometric differentiation between *S. sanctipauli* 'Djodji' and *S. squamosum*. The excellent discrimination between these species is encouraging as *S. sanctipauli* 'Djodji' has often been found in sympatry with *S. squamosum* (Table 2.1). Cheke and Denke (1988) found that *S. sanctipauli* 'Djodji' was a more efficient vector than *S. squamosum* so that successful identification of these two species is important. The four character subset correctly identifies over 98% of flies and is not influenced by size variation.

Adjusting the prior probabilities only slightly improves error rate, so it is recommended that this be done only in cases of doubt following typicality probability allocation.

Table 7.24

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. sanctipauli 'Djodji'	*S. squamosum*
1	0.0000	1.0000
2	0.0281	0.9719
3	0.9716	0.0284
4	1.0000	0.0000
5	1.0000	0.0000

Table 7.25

Standardised Canonical Variate for *S. sanctipauli* 'Djodji' and *S. squamosum*

Character	Canonical Variate
Antennal Length 2	-2.0029
Wing Width 1	0.7875
Femur Length 1	0.6119

Table 7.26

Mean Vectors for species *S. sanctipauli* 'Djodji' and *S. squamosum*

	S. sanctipauli 'Djodji'	*S. squamosum*
Antennal Length 2	476.10461538	425.89835294
Wing Width 1	1041.22711538	1087.65100000
Femur Length 1	660.41538462	678.32329412

Table 7.27

Linear Discriminant functions for species *S. sanctipauli* 'Djodji' and *S. squamosum*

	S. sanctipauli 'Djodji'	*S. squamosum*
CONSTANT	-208.20478421	-189.45004107
V10	0.66157179	0.41817336
V18	0.21633488	0.26616874
V22	-0.18749041	-0.13076059

Table 7.28

Pooled within-species dispersion matrix for species *S. sanctipauli* 'Djodji' and *S. squamosum*

Character	V10	V18	V22
V10	602.16406987	1039.78894941	785.17486436
V18	1039.78894941	3525.12490412	2182.91364838
V22	785.17486436	2182.91364838	1766.88772958

7.3.2 OVERALL DISCRIMINATION The matrix of Mahalanobis' squared distances between species using the full 25 character set excluding wing tuft colouration, abdominal setal colouration and basitarsal spine number is shown as Table 7.29. All of these distances are significant at $p<0.001$. Examining this matrix shows that the two members of the *S. sanctipauli* subcomplex are closer to each other than either is to either *S. damnosum* or *S. squamosum*. The table of re-substitutions is given in Table 7.30. The data set was too large to obtain an estimate of error rate using the 'leave-one-out' method. Because of this, the resubstituted error rate, e_{res}=0.0263 might be optimistic.

A stepwise discriminant analysis of the 25 character set resulted in an initial subset of 15 characters. The method described in section 7.2.3 was then used to reduce the number of characters from the 15 character subset to a nine character subset:

[V4,V6,V9,V18,V20,V23,V27,V28,V29]

i.e. one thorax, one head, one antennal, two wing and four leg characters. The matrix of Mahalanobis' squared distances resulting from this character subset is shown as Table 7.31. The same pattern of species relationships holds in this lower dimensional space as in the 25-character space, with *S. sanctipauli* 'Djodji' and *S. soubrense* 'Beffa' still closer to each other than either is to the other species.

Table 7.32 gives the table of reclassifications using resubstitution and the 'leave-one-out' method of error rate estimation. The resubstituted error rate using the nine character subset was 0.0842, higher than for the full character set, the estimate of error rate using the 'leave-one-out' method was 0.1105, showing that the resubstituted error rate is quite biased.

Allocation using the typicality probability method described in section 7.2.4 resulted in 154 (81.05%) correctly allocated flies, 10 (5.26%) misidentified flies, 20 (10.53%) flies lying on an overlap region and 6 (3.16%) flies being untypical of any of the reference species, with atypicality defined at $\alpha=0.01$. The larger than expected number of atypical flies was most likely due to the fact that *S. squamosum* showed considerable variation in size (see Chapter six), as five of the atypical flies were *S. squamosum*. Of the twenty flies which were overlapping 8 remained in an overlap region and 12 were correctly allocated when they were allocated using the relevant species-pair statistics described in section 7.3.1, bringing the number of flies correctly allocated up to 166 (87.37%).

The null hypothesis of equal wing tuft colouration was rejected using a Kruskal-Wallis test at $p<0.001$, the null hypothesis of equal abdominal setal colouration was not tested as all flies in these samples were character state one for this character (Chapter four). The prior probabilities of species membership according to a fly's wing tuft colouration were therefore calculated and are shown as Table 7.33. When the prior probabilities were adjusted in the way described in section 7.2.3, then 14 flies were classified incorrectly ($e_{res}=0.0737$).

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, legitimising the use of the pooled within-species dispersion matrix.

The standardised canonical variates (Table 7.34, with the species' means along each canonical variate shown as Table 7.35) shows the characters of importance in discriminating between these species. The first canonical variate with a canonical root of 2.91 is dominated by antennal length 1, with most of the other characters having small negative loadings. This vector discriminates predominately the species pair *S. soubrense* 'Beffa'/ *S. sanctipauli* 'Djodji' from the species pair *S. damnosum/S. squamosum*. The second canonical variate is a more complex vector but is basically a contrast between the two positively loading characters tibia length 1 and basitarsus length 2 and the two negatively loading characters thorax width and tibia length 2. The canonical root associated with this canonical variate was 2.173, which is high relative to the first canonical root (together they account for nearly 95% of total variance). This canonical variate discriminates mainly *S. damnosum* from the other species. The final canonical variate, with a canonical root of only 0.279, whilst being statistically significant is probably of no biological importance (Campbell 1982), because no single species is discriminated along its length.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.33,0.33,0.26,0.33,0.34,0.35,0.35,0.35,0.34]

and accounted for 83% of pooled within-species variation. When the scores computed along this vector were introduced into the model as a covariable the canonical roots fell from (2.9077,2.1729,0.2786) to (2.8209,1.2079,0.2279). Thus it is clear that the first canonical variate is very little influenced by size variation, whereas the second canonical variate is influenced by size variation to a considerable extent, although there is still significant discrimination along this vector in the absence of size.

The mean vectors are shown as Table 7.36, the pooled within-species dispersion matrix as Table 7.37, and the linear discriminant functions as Table 7.38.

Table 7.29

Matrix of Mahalanobis' distances between species, 25 character set

	S. soubrense	*S. damnosum s.s*	*S. sanctipauli*	*S. squamosum*
S. soubrense	0.0			
S. damnosum s.s	27.74	0.0		
S. sanctipauli	11.88	50.11	0.0	
S. squamosum	22.88	19.06	27.45	0.0

Table 7.30

Table of re-classifications, using resubstitution, 25 character set

	S. soubrense	*S. damnosum s.s*	*S. sanctipauli*	*S. squamosum*
S. soubrense	35	0	1	0
S. damnosum s.s	1	41	0	1
S. sanctipauli	0	0	26	0
S. squamosum	0	1	1	83

e_{res}=5/190=0.0263

Table 7.31

Matrix of Mahalanobis' distances between species, 9 character subset

	S. soubrense	*S. damnosum s.s*	*S. sanctipauli*	*S. squamosum*
S. soubrense	0.0			
S. damnosum s.s	16.03	0.0		
S. sanctipauli	6.31	25.43	0.0	
S. squamosum	16.66	12.64	14.95	0.0

Table 7.32

Table of re-classifications, using resubstitution (and 'leave-one-out'), 9 character subset

	S. soubrense	*S. damnosum s.s*	*S. sanctipauli*	*S. squamosum*
S. soubrense	31(30)	0(0)	5(6)	0(0)
S. damnosum s.s	2(2)	39(38)	0(0)	2(3)
S. sanctipauli	0(2)	0(0)	26(24)	0(0)
S. squamosum	2(2)	3(4)	2(4)	78(75)

e_{res}=16/190=0.0842 e_c=21/190=0.1105

Table 7.33

Prior probability of species membership for each wing tuft colour category

Wing Tuft Colour	*S. soubrense*	*S. damnosum s.s*	*S. sanctipauli*	*S. squamosum*
1	0.014	0.5102	0.001	0.4882
2	0.1293	0.3904	0.0274	0.4530
3	0.6038	0.0155	0.3588	0.0218
4	0.3747	0.0001	0.6251	0.0001
5	0.176	0.0	0.824	0.0

Table 7.34

Standardised Canonical Variates

Character	CV I	CV II	CV III
Thorax Width	-0.1382	-1.1370	0.9800
Head Width	0.3349	-0.4896	-0.4974
Antennal Length 1	1.8284	0.5821	0.5528
Wing Width 1	-0.8490	0.6044	0.2498
Wing Length 3	-0.4708	0.2720	1.2369
Tibia Length 1	-0.6883	1.0006	-1.8473
Femur Length 2	0.2425	0.5240	-2.5364
Tibia Length 2	0.1017	-1.6807	2.6348
Basitarsus Length	0.0327	1.6664	-0.0822

Table 7.35

Species Means on Canonical Variates

Species	CV I	CV II	CV III
S. soubrense	2.3471	-0.7959	-0.7467
S. damnosum s.s.	-1.1964	-2.3222	0.3222
S. sanctipauli	2.5281	1.0122	0.9882
S. squamosum	-1.1621	1.2022	-0.1490

Table 7.36

Mean Vectors

Character	*S. soubrense*	*S. damnosum s.s*	*S. sanctipauli*	*S. squamosum*
Thorax Width	872.15600000	907.70187907	914.47961538	927.77268706
Head Width	815.04746667	812.94496744	838.51038462	839.65947529
Antennal Length 1	294.17916667	259.07790698	317.62500000	282.91058824
Wing Width 1	961.54361111	996.17825581	1041.22711538	1087.65100000
Wing Length 3	1409.6376388	1452.37523256	1534.97576923	1574.30000000
Tibia Length 1	685.52000000	689.37209302	730.24153846	749.57647059
Femur Length 2	658.46000000	653.67348837	697.88307692	708.19058824
Tibia Length 2	611.72000000	619.80558140	659.46923077	662.40564706
Basitarsus Length	320.77500000	320.16279070	350.32500000	357.69882353

Table 7.37

Pooled within-species dispersion matrix

Character	V4	V6	V9	V18	V20	V23	V27	V28	V29
V4	2703.98603061	1785.83050939	443.98971197	2193.77677851	3177.78803740	1790.24030011	1686.60616466	1620.27957036	849.82461540
V6	1785.83050939	1657.19398678	356.99024657	1641.83501162	2437.04611390	1353.68431045	1291.23179439	1233.27917780	643.94857901
V9	443.98971197	356.99024657	218.59357585	459.42494772	692.30285860	385.62531003	370.05397465	362.93150118	181.67222156
V18	2193.77677851	1641.83501162	459.42494772	2882.78154798	3578.93882920	1801.24154958	1714.13777818	1667.97585554	890.86306685
V20	3177.78803740	2437.04611390	692.30285860	3578.93882920	5995.15954680	2675.27774004	2520.76676937	2486.64200623	1316.36304299
V23	1790.24030011	1353.68431045	385.62531003	1801.24154958	2675.27774004	1564.32903759	1379.39450559	1361.36720442	699.11470965
V27	1686.60616466	1291.23179439	370.05397465	1714.13777818	2520.76676937	1379.39450559	1402.44246022	1303.00588678	668.48042280
V28	1620.27957036	1233.27917780	362.93150118	1667.97585554	2486.64200623	1361.36720442	1303.00588678	1338.85239212	655.23029298
V29	849.82461540	643.94857901	181.67222156	890.86306685	1316.36304299	699.11470965	668.48042280	655.23029298	387.41117794

Table 7.38

Linear Discriminant functions

	S. soubrense	*S. damnosum s.s*	*S. sanctipauli*	*S. squamosum*
CONSTANT	-280.89972742	-260.34685396	-314.02902408	-292.11819108
V4	-0.01999528	0.03898369	-0.02685458	-0.04163823
V6	0.42873824	0.40578646	0.38885047	0.37081618
V9	0.98508531	0.69130814	1.08663561	0.77167116
V18	0.15593835	0.18779172	0.17451851	0.21464193
V20	0.11933462	0.14444330	0.14422052	0.14794828
V23	-0.09037097	-0.11213230	-0.12149516	-0.02259459
V27	-0.02194982	-0.11961641	-0.09813606	-0.05145069
V28	-0.34703442	-0.22974707	-0.31081350	-0.39701843
V29	-0.27889690	-0.38307706	-0.16981360	-0.15883762

7.3.3 DISCUSSION

The discriminant statistics presented in this section reveal that there is extensive multivariate morphometric differentiation between adult females of the four species of the *S. damnosum* complex for which samples were available from Togo and Benin.

The species which was most successfully discriminated relative to the other taxa, both using the species-pair statistics and in the overall analysis was *S. damnosum s.s.*. The maximum overlap of this species with any other was with *S. squamosum*, which was about 9% phenetic overlap. The influence of size variation on discrimination of *S. damnosum s.s.* was greater than for the other species, but even so, the size-free canonical roots were still larger when *S. damnosum s.s.* was involved than for most other species-pair discriminant analyses. The samples of *S. damnosum s.s.* used were smaller flies than the other species, but also they were a different shape, independent of size. Generally, the relative size of the antenna was the main morphological feature characterising *S. damnosum s.s.*, a finding in concordance with previous morphological studies of the *S. damnosum* complex (e.g. Garms 1978, Dang and Peterson 1980). The species also had consistently paler wing tufts than the other species, with the important exception of *S. squamosum*, to which it is closest morphologically.

Simulium squamosum is phenetically the next most isolated species, being closest to *S. damnosum s.s.*. The ability to discriminate between this species and the members of the *S. sanctipauli* subcomplex will be an important aid in the further understanding of the relative vectorial importance of the different species in Togo and Benin (see e.g., Cheke and Denke 1988). Previous morphological methods of identification using thorax/antennal ratios found considerable overlap between these groups (Garms and Cheke 1985).

Simulium soubrense 'Beffa' and *S. sanctipauli* 'Djodji' show the greatest phenetic overlap, over 10%. The species-pair discriminant analysis also indicated that size differences were important in the between-species variation which had been found so that more extensive sampling is likely to reveal greater overlap than was found in this analysis. However, the discrimination between these species is greater than previous methods, which have not distinguished between the two, therefore it may be possible to collect more information as to the relative vectorial importance of these species. The restricted geographic range of *S. sanctipauli* 'Djodji' (Chapter two, Surtees *et al.* 1988), also means that the phenetic overlap between it and *S. soubrense* 'Beffa' may not be of critical importance, because they tend not to be found sympatrically.

Table 7.39 summarises the four methods of allocation used in the overall discriminant analysis: forced allocation with and without adjusting the prior probabilities of species membership using the fly's wing tuft colour, and typicality probability allocation with and without subsequent allocation of overlapping flies using the appropriate species-pair statistics. The relatively small sample sizes of the four species accounts for over 10% of flies overlapping using typicality probability allocation, because the approximate confidence intervals for each fly's distanmce to each of the species are broad. The subsequent species-pair allocation improves this. The largest number of correct allocations was using forced allocation with adjusted prior probabilities, althought the effect of adjusting the prior probabilities is very slight. The smallest number of incorrect allocations was obtained using typicality probability allocation. This method is more conservative than forced allocation, but considering the small sample sizes used to calculate the discriminant statistics this caution is well justified.

To conclude, the discriminant analyses presented in this section show that the four species of the *S. damnosum* complex which were examined from Togo and Benin can be successfully identified, although the rate of correct classification varies according to species. The ability to identify *S. damnosum s.s.*, vector of the more debilitating form of onchocerciasis is clearly important, as is the ability to identify *S. squamosum*. The characters used in these analysis, and the use of multivariate statistical methods is without doubt an improvement over current morphological methods, and will be futher refined once larger samples have been obtained of each species, so that a wider range of variation, including seasonal size variation is sampled. The major limitation of the Togo and Benin statistics is that no *S. yahense* were available for analysis, despite its presence in Togo (Table 2.1). If account is to be taken of *S. yahense*, then the 'global' statistics developed in Chapter eight should be used, even though this assumes that *S. yahense* is the same morphologically in the east as in the west.

Table 7.39

Comparison of four methods of allocation for Togo and Benin

	Forced[1]	Forced[2]	Typicality[3]	Typicality[4]
Correct	174	176	154	166
Incorrect	16	14	10	10
Overlapping	na	na	20	8
Atypical	na	na	6	6

[1] Forced allocation without adjusted priors
[2] Forced allocation with adjusted priors
[3] Typicality probability without subsequent species pair allocation of overlapping flies
[4] Typicality probability with subsequent species pair allocation of overlapping flies

7.4 RESULTS AND DISCUSSION, WESTERN AREA

7.4.1 SPECIES PAIR DISCRIMINATION

Samples from seven taxa were available from the area west of the Volta Lake, making the situation more complex than in Togo and Benin. For the purposes of the species pair statistics the single sample of *S. damnosum s.s.* was pooled with the five *S. sirbanum* samples into an artificial category called 'Savanna', because they were morphometriclally very close (see section 7.4.1.1). Also OCP regards both species as dangerous vectors of onchocerciasis, to be controlled whenever either is found.

The single sample of *S. soubrense* 'B' was not pooled with the samples of *S. soubrense*even though it was for the overall analysis (see section 7.4.2). This was justified because of the chromosomal distinctiveness of this new species (Chapter three, Post 1986), even though morphometrically it was not distinctive (Chapter six).

7.4.1.1 Discrimination of Simulium damnosum and S. sirbanum

The squared Mahalanobis' distance between species using the 25 character set was 9.01 (unbiased D^2=7.88, e_{act}=0.0802) with 12 mis-allocated flies using resubstitution (e_{res}=0.0583) A stepwise discriminant analysis produced a nine character subset which gave a Mahalanobis' squared distance of 7.7 (unbiased D^2=7.32, e_{act}=0.088) with 18 flies misallocated using resubstitution (e_{res}=0.087).

The dimension reduction technique described in section 7.2.2 resulted in a seven character subset:

[V4,V9,V14,V16,V17,V23,V29]

i.e. one thorax, two antennal, two wing and two leg characters with a Mahalanobis' squared distance of 7.17 (unbiased D^2=6.89, e_{act}=0.095) and 17 misallocated flies using resubstitution (e_{res}=0.083), one *S.*

damnosum s.s. classified as *S. sirbanum* and 16 *S. sirbanum*classified as *S. damnosum s.s.*

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in 12 misallocated flies (5.83%), one atypical fly (0.49%), 15 overlapping flies (7.28%) and 178 correctly allocated flies (86.4%).

The null hypothesis of equal wing tuft colouration was not rejected at p<0.001 using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were therefore not calculated.

The null hypothesis of equal dispersion was not rejected at p<0.001 using the likelihood ratio test, legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.40) shows that the characters of most importance in discrimination are the positively loading characters: thorax width, tibia length 1, and basitarsus length 2 , and the negatively loading characters: wing length 2 and antennal length 1.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.41,0.33,0.17,0.41,0.41,0.42,0.42]

and accounted for 73.0% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 0.8498 to 0.5814 showing that size has considerable influence on the already relatively poor discrimination between these species.

The mean vectors are shown as Table 7.41, the pooled within-species dispersion matrix as Table 7.42 and the linear discriminant functions as Table 7.43.

To conclude, there is significant multivariate morphometric differentiation between *S. damnosum s.l.* and *S. sirbanum,* however this differentiation is relatively small, and is influenced to a considerable extent by size variation. Therefore, although the seven character subset classified correctly in over 86% of cases, this is probably an optimistic estimate, and can be expected to perform less well once a greater range of temporal and geographic variation within each species has been sampled. Operationally it is not important that these species cannot be differentiated very well as both are regarded as dangerous vectors of the more debilitating savanna form of onchocerciasis, and OCP controls both species whenever they are found.

Table 7.40

Standardised Canonical Variate for *S. damnosum s.s.* and *S. sirbanum*

Character	Canonical Variate
Thorax Width	0.8793
Antennal Length 1	-0.8197
Antennal Segment 7	0.2192
Wing Length 1	-0.5568
Wing Length 2	-0.9491
Tibia Length 1	0.8777
Basitarsus Length	1.0627

Table 7.41

Mean Vectors for species *S. damnosum s.s.* and *S. sirbanum*

	S. damnosum s.s.	*S. sirbanum*
Thorax Width	938.33485714	846.98794719
Antennal Length 1	252.66428571	253.14943820
Antennal Segment 7	36.89000000	35.70921348
Wing Length 1	751.35428571	698.86112360
Wing Length 2	457.56000000	432.46247191
Tibia Length 1	716.56285714	657.28988764
Basitarsus Length 2	330.73392857	301.03398876

Table 7.42

Pooled within-species dispersion matrix for species *S. damnosum s.s.* and *S. sirbanum*

Character	V4	V9	V14	V16	V17	V23	V29
V4	3334.50882078	532.59237158	45.66197091	2220.92410122	1481.87749590	1894.67718138	953.28978365
V9	532.59237158	210.61198152	20.20367118	385.07727646	272.14810905	335.59108748	177.40024425
V14	45.66197091	20.20367118	8.44149260	30.64210469	21.90185856	30.47224600	14.60045960
V16	2220.92410122	385.07727646	30.64210469	2005.02023363	1191.90521719	1499.65516118	754.20342956
V17	1481.87749590	272.14810905	21.90185856	1191.90521719	922.23770839	978.54598652	495.96160806
V23	1894.67718138	335.59108748	30.47224600	1499.65516118	978.54598652	1386.32722338	637.02149738
V29	953.28978365	177.40024425	14.60045960	754.20342956	495.96160806	637.02149738	357.89091733

Table 7.43

Linear Discriminant functions for species *S. damnosum s.s.* and *S. sirbanum*

	S. damnosum s.s.	*S. sirbanum*
CONSTANT	-234.18007423	-211.54025531
V4	-0.08108991	-0.11697217
V9	0.52340679	0.67496271
V14	1.91770284	1.71718024
V16	0.03924445	0.07018309
V17	-0.27948955	-0.19882914
V23	0.57646803	0.52100165
V29	0.08097309	-0.05162960

7.4.1.2 Discrimination of 'Savanna' and S. sanctipauli

The squared Mahalanobis' distance between species using the 25 character set was 77.16 (unbiased D^2=68.77, e_{act}<0.0001) with no misallocated flies using resubstitution (e_{res}=0.0). A stepwise discriminant analysis produced a 12 character set, but the dimension reduction technique described in section 7.2.2 reduced this to a three character subset:

[V9,V17,V29]

i.e. one antennal, one wing and one leg character with a Mahalanobis' squared distance of 48.28 (unbiased D^2=47.47, e_{act}=0.0003) and no misallocated flies using resubstitution (e_{res}=0.0).

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in no misallocated flies, four atypical flies (1.66%), no overlapping flies and 237 correctly allocated flies (98.3%).

The null hypothesis of equal wing tuft colouration was rejected at p<0.001 using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.44. From this table it is very clear that this character alone would classify most flies correctly, and when the prior probabilities were adjusted according to each fly's wing tuft colouration, the number of flies misallocated remained zero.

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.45) shows that the characters of most importance in discrimination are the positively loading character antennal length 1, and the negatively loading character wing length 2, with *S. sanctipauli* lying at the positive end of this vectór having relatively larger antennae.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.51, 0.61, 0.60]$$

and accounted for 77.4% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 6.0438 to 3.2476 showing that size has some influence on the discrimination between these species, but that shape differences are important in the absence of size variation.

The mean vectors are shown as Table 7.46, the pooled within-species dispersion matrix as Table 7.47 and the discriminant functions as Table 7.48.

To conclude, there is significant multivariate morphometric differentiation between 'savanna' and *S. sanctipauli*. This differentiation is a combination of size and shape variation but in the absence of size variation, shape differences discriminate between this species pair very well. Therefore, the three character set can be expected to classify correctly in over 98% of cases.

Adjusting the prior probabilities according to a fly's wing tuft colouration did not improve error rate, and so need only be used in cases of doubt following typicality probability allocation.

Table 7.44

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	'Savanna'	*S. sanctipauli*
1	1.0000	0.0000
2	1.0000	0.0000
3	0.0004	0.9996
4	0.0000	1.0000
5	0.0000	1.0000

Table 7.45

Standardised Canonical Variate for 'Savanna' and *S. sanctipauli*

Character	Canonical Variate
Antennal Length 1	2.6164
Wing Length 2	-1.0412
Basitarsus Length	0.6315

Table 7.46

Mean Vectors for species *S. sanctipauli* and *S. squamosum*

	S. sanctipauli	*S. squamosum*
Antennal Length 1	253.08349515	336.06857143
Wing Length 2	435.87378641	457.41942857
Basitarsus Length 2	305.07087379	344.64857143

Table 7.47

Pooled within-species dispersion matrix for species *S. sanctipauli* and *S. squamosum*

Character	V9	V17	V29
V9	203.87903478	252.59119768	157.58476736
V17	252.59119768	920.92884199	533.82209912
V29	157.58476736	533.82209912	427.91009486

Table 7.48

Linear Discriminant functions for species 'Savanna' and *S. sanctipauli*

	'Savanna'	*S. sanctipauli*
CONSTANT	-176.65358555	-294.73381294
V9	0.96290252	1.52092797
V17	0.00537202	-0.22632739
V29	0.35162620	0.52766237

7.4.1.3 Discrimination of 'Savanna' and S. soubrense 'B' -

The squared Mahalanobis' distance between species using the 25 character set was 33.65 (unbiased D^2=29.88, e_{act}=0.0031) with no misallocated flies using resubstitution (e_{res}=0.0). Stepwise discriminant analysis appled to this character set produced an eight character subset with a D^2 of 32.15 (unbiased D^2 =30.7, e_{act}=0.0027) and two misallocated flies using resubstitution (e_{res}= 0.0086). The dimension reduction technique described in section 7.2.2 further reduced this to a five character subset:

[V3,V9,V13,V19,V29]

i.e. one thorax, two antennal, one wing and one leg character with a Mahalanobis' squared distance between species of 28.34 (unbiased D^2=27.61, e_{act}=0.0043) and two misallocated flies using resubstitution (e_{res}=0.0086).

Allocation using typicality probability of species membership with atypicality defined at $\alpha=0.01$ resulted in one misallocated fly (0.43%), one atypical fly (0.43%), two overlapping flies (0.86%) and 230 correctly allocated flies (98.3%).

The null hypothesis of equal wing tuft colouration was rejected at $p<0.001$ using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.49. When the prior probabilities were adjusted according to each fly's wing tuft colouration, two flies were misallocated ($e_{res}=0.0086$).

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.50) shows that the contrast between the positively loading character antennal length 1, and the negatively loading character wing width 2, discriminates between these species, with *S. soubrense* 'B' lying at the positive end of this vector having relatively larger antennae.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.47,0.41,0.35,0.49,0.50]$$

and accounted for 67.4% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 3.0111 to 3.0097 showing that size has virtually no influence on the discrimination between these species.

The mean vectors are shown as Table 7.51, the pooled within-species dispersion matrix as Table 7.52 and the linear discriminant functions as Table 7.53.

To conclude, there is significant multivariate morphometric differentiation between 'savanna' and *S. soubrense* 'B'. This differentiation is entirely shape variation involving the relative length of the antenna. Therefore, the three character set can be expected to classify correctly in over 98% of cases.

Adjusting the prior probabilities according to a fly's wing tuft colouration did not improve error rate, and so it is recommended that this should not be done except in cases of doubt.

Table 7.49

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	'Savanna'	*S. soubrense* 'B'
1	1.0000	0.0000
2	0.9450	0.0550
3	0.0003	0.9997
4	0.0000	1.0000
5	0.0000	1.0000

Table 7.50

Standardised Canonical Variate for 'Savanna' and *S. soubrense* 'B'

Character	Canonical Variate
Thorax Length	-0.7091
Antennal Length 1	1.0633
Antennal Segment 6	0.7730
Wing Width 2	-0.9619
Basitarsus Length	0.8611

Table 7.51

Mean Vectors for species 'Savanna' and *S. soubrense* 'B'

	'Savanna'	*S. soubrense* 'B'
Thorax Length	618.73154272	586.83394286
Antennal Length 1	253.08349515	299.88214286
Antennal Segment 6	35.88174757	44.95000000
Wing Width 2	1383.34075243	1349.70892857
Basitarsus Length 2	305.07087379	315.69107143

Table 7.52

Pooled within-species dispersion matrix for species 'Savanna' and *S. soubrense* 'B'

Character	V3	V9	V13	V19	V29
V3	2083.38014492	337.07104861	57.99456940	2892.10382875	763.69444208
V9	337.07104861	216.20681015	22.08221096	702.64485893	171.88295908
V13	57.99456940	22.08221096	8.86393005	105.74630357	25.70746545
V19	2892.10382875	702.64485893	105.74630357	7074.88590766	1563.10241270
V29	763.69444208	171.88295908	25.70746545	1563.10241270	444.75467892

Table 7.53

Linear Discriminant functions for species 'Savanna' and *S. soubrense* 'B'

	'Savanna'	*S. soubrense* 'B'
CONSTANT	-185.24947818	-234.34607400
V3	0.02935689	-0.05145444
V9	0.72768806	0.99544177
V13	0.93113997	1.91417183
V19	0.13848980	0.07799207
V29	-0.18625321	0.02871013

7.4.1.4 Discrimination of 'Savanna' and <u>S. soubrense</u>

The squared Mahalanobis' distance between species using the 25 character set was 25.89 (unbiased D^2=24.03, e_{act}=0.0071) with six misallocated flies using resubstitution (e_{res}=0.0165). A stepwise discriminant analysis produced an eight character subset with a D^2 of 25.37 (unbiased D^2 =24.25, e_{act}=0.069). The dimension reduction technique described in section 7.2.2 reduced this to a five character subset:

[V4,V10,V13,V20,V29]

i.e. one thorax, two antennal, one wing and one leg character with a Mahalanobis' squared distance of 17.82 (unbiased D^2=17.53, e_{act}=0.0182) between species and seven misallocated flies using resubstitution (e_{res}=0.0192), all seven of which were *S. soubrense* classified into *S. sirbanum.*

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in six misallocated flies (1.65%), two atypical flies (0.55%), one overlapping fly (0.28%) and 355 correctly allocated flies (97.5%).

The null hypothesis of equal wing tuft colouration was rejected at $p<0.001$ using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.54. When the prior probabilities were adjusted according to each fly's wing tuft colouration, eight flies were misallocated (e_{res}=0.022).

The null hypothesis of equal dispersion was rejected at $p<0.001$ using the likelihood ratio test, but for the statistical and practical reasons given in section 7.2.5, the dispersion matrices were still pooled.

The standardised canonical variate (Table 7.55) shows that the most important contrast in discrimination is between antennal length

2. and thorax width. *S. soubrense* is at the positive side of this vector, having relatively larger antenna than 'savanna'.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.48,0.44,0.33,0.48,0.49]

and accounted for 74.6% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 4.4028 to 3.1025 showing that size has little influence on the discrimination between these species.

The mean vectors are shown as Table 7.56, the pooled within-species dispersion matrix as Table 7.57, and the linear discriminant functions as Table 7.58.

To conclude, there is significant multivariate morphometric differentiation between savanna and *S. soubrense*. This differentiation is mainly shape variation, so the five character set can be expected to classify correctly in over 97% of cases.

Adjusting the prior probabilities according to a fly's wing tuft colouration did not improve error rate, and so it is recommended that this should not be done except in cases of doubt.

Table 7.54

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	'Savanna'	*S. soubrense*
1	0.8887	0.1113
2	0.4913	0.5087
3	0.1046	0.8954
4	0.0139	0.9861
5	0.0017	0.9983

Table 7.55

Standardised Canonical Variate for 'Savanna' and *S. soubrense*

Character	Canonical Variate
Thorax Width	-1.1602
Antennal Length 2	1.3912
Antennal Segment 6	0.9875
Wing Length 3	0.7299
Basitarsus Length	-0.1533

Table 7.56

Mean Vectors for species 'Savanna' and *S. soubrense*

	'Savanna'	*S. soubrense*
Thorax Width	859.40403204	876.48680506
Antennal Length 2	380.12970874	456.40784810
Antennal Segment 6	35.88174757	46.11544304
Wing Length 3	1407.11145631	1504.37734177
Basitarsus Length 2	305.07087379	328.69841772

Table 7.57

Pooled within-species dispersion matrix for species 'Savanna' and *S. soubrense*

Character	V4	V10	V13	V20	V29
V4	4116.58167146	1154.38579563	87.36060217	5130.77823361	1292.57184921
V10	1154.38579563	703.06753937	53.53254960	1597.06880493	426.39566358
V13	87.36060217	53.53254960	10.43394721	123.04797006	34.96686560
V20	5130.77823361	1597.06880493	123.04797006	8131.90646643	1860.02049606
V29	1292.57184921	426.39566358	34.96686560	1860.02049606	507.57891532

Table 7.58

Linear Discriminant functions for species 'Savanna' and *S. soubrense*

	'Savanna'	*S. soubrense*
CONSTANT	-148.15174068	-199.77811381
V4	-0.02663396	-0.10241882
V10	0.25651160	0.38364221
V13	1.21771013	1.91060765
V20	0.22571626	0.25587704
V29	-0.45765225	-0.48316777

7.4.1.5 Discrimination of 'Savanna' and S. squamosum

The squared Mahalanobis' distance between species using the 25 character set was 18.14 (unbiased D^2=16.52, e_{act}=0.0211) with five misallocated flies using resubstitution (e_{res}=0.0171). Stepwise discriminant analysis produced an eight character subset with a D^2 between species of 17.17 (unbiased D^2 =16.16, e_{act}=0.022) and ten misallocated flies (e_{res}=0.034). The dimension reduction technique described in section 7.2.2 resulted in a seven character subset:

[V4,V6,V9,V20,V27,V28,V29]

i.e. one thorax, one head, one antennal, one wing and three leg characters with a Mahalanobis' squared distance of 13.28 (unbiased D^2=12.91, e_{act}=0.0362) and ten misallocated flies using resubstitution (e_{res}=0.034), eight 'savanna' classified into *S. squamosum* and two *S. squamosum* classified into 'savanna'.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in five misallocated flies (1.71%), four atypical flies (1.37%), six overlapping fly (2.05%) and 278 correctly allocated flies (94.9%).

The null hypothesis of equal wing tuft colouration was rejected at $p<0.001$ using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.59. When the prior probabilities were adjusted according to each fly's wing tuft colouration, six flies were misallocated (e_{res}=0.021).

The null hypothesis of equal dispersion was rejected at $p<0.001$ using the likelihood ratio test, but for the statistical and practical reasons given in section 7.2.5, the dispersion matrices were still pooled.

The standardised canonical variate (Table 7.60) shows that the most important contrast in discrimination is between the positively loading characters: wing length 3, and basitarsus length 2 and the negatively loading characters: thorax width and tibia length 2.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.39,0.38,0.29,0.39,0.40,0.40,0.39]

and accounted for 85.8% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 2.7919 to 2.4606 showing that size has little influence on the discrimination between these species.

The mean vectors are shown as Table 7.61, the pooled within-species dispersion matrix as Table 7.62 and the linear discriminant functions as Table 7.63.

To conclude, there is significant multivariate morphometric differentiation between 'savanna' and *S. squamosum*. This differentiation is mainly shape variation, so the seven character set can be expected to classify correctly in nearly 95% of cases.

Adjusting the prior probabilities according to a fly's wing tuft colouration improved error rate, and so it is recommended that this should be done in cases of doubt following typicality probability allocation.

Table 7.59

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	'Savanna'	*S. squamosum*
1	0.9013	0.0987
2	0.1873	0.8127
3	0.0058	0.9942
4	0.0001	0.9999
5	0.0000	1.0000

Table 7.60

Standardised Canonical Variate for 'Savanna' and *S. squamosum*

Character	Canonical Variate
Thorax Width	-1.5111
Head Width	-0.2167
Antennal Length 1	0.5834
Wing Length 3	1.4802
Femur Length 2	0.9878
Tibia Length 2	-2.0050
Basitarsus Length	1.4755

Table 7.61

Mean Vectors for species 'Savanna' and *S. squamosum*

	'Savanna'	*S. squamosum*
Thorax Width	859.40403204	855.21798621
Head Width	775.71054175	795.31917241
Antennal Length 1	253.08349515	276.18275862
Wing Length 3	1407.11145631	1506.67091954
Femur Length 2	631.47961165	658.14896552
Tibia Length 2	597.82776699	612.96413793
Basitarsus Length 2	305.07087379	329.43793103

Table 7.62

Pooled within-species dispersion matrix for species 'Savanna' and *S. squamosum*

Character	V4	V6	V9	V20	V27	V28	V29
V4	4081.94577108	2731.89999834	589.35997597	4759.88112753	2431.88941630	2324.16176789	1210.31479711
V6	2731.89999834	2324.38954083	451.37162330	3534.85810881	1775.76543610	1714.74046939	901.31965086
V9	589.35997597	451.37162330	228.47447774	799.74397450	392.03597295	378.45810520	201.69990676
V20	4759.88112753	3534.85810881	799.74397450	7084.34530099	3224.99834846	3113.27041767	1637.40522063
V27	2431.88941630	1775.76543610	392.03597295	3224.99834846	1708.94166521	1611.01703297	827.81038517
V28	2324.16176789	1714.74046939	378.45810520	3113.27041767	1611.01703297	1604.65839616	796.96262868
V29	1210.31479711	901.31965086	201.69990676	1637.40522063	827.81038517	796.96262868	455.92922705

Table 7:63

Linear Discriminant functions for species 'Savanna' and *S. squamosum*

	'Savanna'	*S. squamosum*
CONSTANT	-196.45847171	-238.41977587
V4	-0.20494826	-0.29125591
V6	0.24599641	0.22986491
V9	0.70930704	0.82471657
V20	0.21088635	0.26732438
V27	0.23528375	0.31893918
V28	-0.17417372	-0.35421631
V29	-0.46702989	-0.24350292

7.4.1.6 Discrimination of 'Savanna' and S, yahense

The squared Mahalanobis' distance between species using the 25 character set was 46.59 (unbiased D^2=42.55, e_{act}=0.0006) with one misallocated fly using resubstitution (e_{res}=0.0033). A stepwise discriminant analysis produced a 12 character subset with a D^2 of 45.21 (unbiased D^2 =43.25, e_{act}=0.0005) and one misallocated fly

(e_{res}=0.0033). The dimension reduction technique described in section 7.2.2 resulted in a four character subset:

[V4,V9,V20,V29]

i.e. one thorax, one antennal, one wing and one leg character with a Mahalanobis' squared distance of 30.1 (unbiased D^2=29.6, e_{act}=0.0033) and one misallocated flies using resubstitution (e_{res}=0.0033), which was a 'savanna' fly classified into *S. yahense*.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in no misallocated flies, four atypical flies (1.32%), no overlapping flies and 298 correctly allocated flies (98.68%).

The null hypotheses of equal wing tuft colouration and equal abdominal setal colouration were both rejected at p<0.001 using Wilcoxon two-sample rank sum tests. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.64. When the prior probabilities were adjusted according to each fly's wing tuft colouration, no flies were misallocated (e_{res}=0.0). However, when the prior probabilities were adjusted for each abdominal setal colouration category, four flies were misallocated (e_{res}=0.0132), the four being *S. yahense* flies with abdominal setal colouration category one (Chapter four).

The null hypothesis of equal dispersion was rejected at p<0.001 using the likelihood ratio test, but for the statistical and practical reasons given in section 7.2.5, the dispersion matrices were still pooled.

The standardised canonical variate (Table 7.65) shows that the most important characters in discrimination are antennal length 1 and the negatively loading character thorax width.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.52,0.43,0.52,0.52]

and accounted for 83.1% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 6.5698 to 3.4387 showing that size has some influence on the discrimination between these species, but that shape differences are important in its absence.

The mean vectors are shown as Table 7.66, the pooled within-species dispersion matrix as Table 7.67 and the linear discriminant functions as Table 7.68.

To conclude, there is significant multivariate morphometric differentiation between 'savanna' and *S. yahense*. This differentiation is a combination of size and shape variation, with shape variation being more important, so the four character set can be expected to classify correctly in nearly 99% of cases.

Adjusting the prior probabilities according to a fly's wing tuft colouration improved error rate slightly, but since the fly which was allocated correctly as a result was actually atypical, it is recommended that this should be done only in cases of doubt (i.e. only on 'typical' overlapping flies). Adjusting prior probabilities for each abdominal setal colouration category did not improve error rate, so it is not recommended that this should be used.

Table 7.64

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	'Savanna'	*S. yahense*
1	1.0000	0.0000
2	1.0000	0.0000
3	0.0712	0.9288
4	0.0000	1.0000
5	0.0000	1.0000

Table 7.65

Standardised Canonical Variate for 'Savanna' and *S. yahense*

Character	Canonical Variate
Thorax Width	-1.5496
Antennal Length 1	2.2583
Wing Length 3	0.5025
Basitarsus Length	0.9848

Table 7.66

Mean Vectors for species 'Savanna' and *S. yahense*

	'Savanna'	*S. yahense*
Thorax Width	859.40403204	906.62228125
Antennal Length 1	253.08349515	319.98281250
Wing Length 3	1407.11145631	1576.95807292
Basitarsus Length 2	305.07087379	352.44218750

Table 7.67

Pooled within-species dispersion matrix for species 'Savanna' and *S. yahense*

Character	V4	V9	V20	V29
V4	4344.69214529	669.65652051	5096.98270101	1285.79911384
V9	669.65652051	251.12671008	890.65435174	225.75435193
V20	5096.98270101	890.65435174	7566.57036711	1739.46501848
V29	1285.79911384	225.75435193	1739.46501848	485.64772294

Table 7.68

Linear Discriminant functions for species 'Savanna' and *S. yahense*

	'Savanna'	*S. yahense*
CONSTANT	-168.33314123	-253.58470268
V4	-0.12202774	-0.24453803
V9	0.68764496	1.04177862
V20	0.23796683	0.26141911
V29	-0.22073570	-0.04745514

7.4.1.7 Discrimination of Simulium sanctipauli and S. soubrense 'B'

The squared Mahalanobis' distance between species using the 25 character set was 25.49 (unbiased D^2=14.63, e_{act}=0.0279) with no misallocated flies using resubstitution (e_{res}=0.0). A stepwise discriminant analysis produced an eleven character subset with a D^2 of 22.68 (unbiased D^2 =18.22, e_{act}=0.0164) and no misallocated flies (e_{res}=0.0). The dimension reduction technique described in section 7.2.2 resulted in a six character subset:

[V4,V10,V15,V16,V20,V24]

i.e. one thorax, two antennal, two wing and one leg character with a Mahalanobis' squared distance of 12.76 (unbiased D^2=11.3, e_{act}=0.0464) and no misallocated flies using resubstitution (e_{res}=0.0).

Allocation using typicality probability of species membership with atypicality defined at $\alpha=0.01$ resulted in no misallocated flies, one atypical fly (1.59%), eight overlapping fly (12.7%) and 54 correctly allocated flies (85.7%).

The null hypothesis of equal wing tuft colouration was not rejected at $p<0.001$ using a Wilcoxon two-sample rank sum test, so the prior probabilities of species membership for each wing tuft colouration category were not calculated.

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.69) shows that the most important character in discrimination is the positively loading character: basitarsus length 1.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.44,0.37,0.21,0.45,0.45,0.46]$$

and accounted for 66.2% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 3.2546 to 1.1461 indicating that size is important in discriminating between these two species.

The mean vectors are shown as Table 7.70, the pooled within-species dispersion matrix as Table 7.71 and the linear discriminant functions as Table 7.72.

To conclude, there is significant multivariate morphometric differentiation between *S. sanctipauli* and *S. soubrense* 'B', however much of this differentiation is size variation, although there is still significant discrimination in its absence. The six character subset classified nearly 86% of flies correctly, but once a wider range of temporal and geographic variation has been sampled, this figure may be optimistic.

Table 7.69

Standardised Canonical Variate for *S. sanctipauli* and *S. soubrense* 'B'

Character	Canonical Variate
Thorax Width	-0.7089
Antennal Length 2	0.8290
Antennal Segment 8	0.8452
Wing Length 1	0.4008
Wing Length 3	-0.6764
Basitarsus Length	1.2457

Table 7.70

Mean Vectors for species *S. sanctipauli* and *S. soubrense* 'B'

	S. sanctipauli	*S. soubrense* 'B'
Thorax Width	906.40608000	854.03575714
Antennal Length 2	501.27771429	441.74571429
Antennal Segment 8	50.20228571	42.91285714
Wing Length 1	771.73714286	707.60142857
Wing Length 3	1512.93442857	1422.18357143
Basitarsus Length 1	459.10628571	416.44285714

Table 7.71

Pooled within-species dispersion matrix for species *S. sanctipauli* and *S. soubrense* 'B'

Character	V4	V10	V15	V16	V20	V24
V4	2521.06007214	673.54413359	39.50850533	1511.12786324	2730.06872634	801.37675347
V10	673.54413359	499.30943775	34.56310370	437.44422857	879.84030464	266.71257780
V15	39.50850533	34.56310370	8.80771129	22.64939859	63.61045508	11.87441424
V16	1511.12786324	437.44422857	22.64939859	1512.06383044	2378.01545902	708.64661171
V20	2730.06872634	879.84030464	63.61045508	2378.01545902	5435.19611691	1248.19125639
V24	801.37675347	266.71257780	11.87441424	708.64661171	1248.19125639	430.94194735

Table 7.72

Linear Discriminant functions for species *S. sanctipauli* and *S. soubrense* 'B'

	S. sanctipauli	*S. soubrense* 'B'
CONSTANT	-351.35623515	-277.83430877
V4	-0.10026713	-0.05527611
V10	0.44030726	0.36060156
V15	3.06561061	2.42182736
V16	0.03974933	0.01122305
V20	0.04088571	0.06894794
V24	0.71104427	0.56107719

7.4.1.8 Discrimination of Simulium sanctipauli and S. soubrense

The squared Mahalanobis' distance between species using the 25 character set was 11.57 (unbiased D^2=10.0, e_{act}=0.0569) with nine misallocated flies using resubstitution (e_{res}=0.0415). A stepwise discriminant analysis produced an eleven character subset with a D^2 of 11.05 (unbiased D^2 =10.36, e_{act}=0.0538) and 12 misallocated flies

(e_{res}=0.062). The dimension reduction technique described in section 7.2.2 resulted in a five character subset:

[V11,V15,V19,V20,V27]

i.e. two antennal, two wing and one leg character with a Mahalanobis' squared distance of 8.2 (unbiased D^2=7.94, e_{act}=0.0794) and 12 misallocated flies using resubstitution (e_{res}=0.06), of which 10 were *S. soubrense* misclassified as *S. sanctipauli* and two were *S. sanctipauli* misclassified as *S. soubrense*.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in eight misallocated flies (4.15%), two atypical flies (1.04%), ten overlapping flies (5.18%) and 173 correctly allocated flies (89.6%).

The null hypothesis of equal wing tuft colouration was rejected at p<0.001 using a Wilcoxon two-sample rank sum test, so the prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.73. When the prior probabilities of species membership were adjusted according to a fly's wing tuft colouration, the number of flies misallocated rose to 18 (e_{res}=0.093).

The null hypothesis of equal dispersion was not rejected at p<0.001 using the likelihood ratio test, legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.74) shows that the most important character in discrimination is the negatively loading character wing length 3, with femur length 2 having some opposite influence

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.34,0.38,0.48,0.50,0.50]

and accounted for 68.8% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 1.23 to 1.1461 indicating that size is not important in discriminating between these species.

The mean vectors are shown as Table 7.75, the pooled within-species dispersion matrix as Table 7.77 and the linear discriminant functions as Table 7.78.

To conclude, there is significant multivariate morphometric differentiation between *S. sanctipauli* and *S. soubrense*, although this differentiation is not great. The five character subset can be expected to classify nearly 90% of flies correctly when it is known *a priori* that just this species pair can be expected.

Adjusting prior probabilities according to a fly's wing tuft colouration did not improve error rate, so it is not recommended that this should be done.

Table 7.73

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. sanctipauli	*S. soubrense*
1	0.1022	0.8978
2	0.2105	0.7895
3	0.3843	0.6157
4	0.5937	0.4063
5	0.7738	0.2262

Table 7.74

Standardised Canonical Variate for *S. sanctipauli* and *S. soubrense*

Character	Canonical Variate
Antennal Segment 4	0.5997
Antennal Segment 8	0.5237
Wing Width 2	0.7372
Wing Length 3	-2.1817
Femur Length 2	1.2937

Table 7.75

Mean Vectors for species *S. sanctipauli* and *S. soubrense*

	S. sanctipauli	*S. soubrense*
Antennal Segment 4	49.91885714	42.41113924
Antennal Segment 8	50.20228571	44.09848101
Wing Width 2	1488.56614286	1423.85350633
Wing Length 3	1512.93442857	1504.37734177
Femur Length 2	692.31428571	663.60835443

Table 7.76

Pooled within-species dispersion matrix for species *S. sanctipauli* and *S. soubrense*

Character	V11	V15	V19	V20	V27
V11	17.42241439	8.62668673	136.96191809	165.28271757	71.62593229
V15	8.62668673	12.15275525	141.42825524	162.66671418	72.04564640
V19	136.96191809	141.42825524	7812.75567753	7280.48278354	3060.16200252
V20	165.28271757	162.66671418	7280.48278354	8833.97372629	3535.85005757
V27	71.62593229	72.04564640	3060.16200252	3535.85005757	1612.55914885

Table 7.77

Linear Discriminant functions for species *S. sanctipauli* and *S. soubrense*

	S. sanctipauli	*S. soubrense*
CONSTANT	-181.19268736	-155.23678949
V11	0.73449596	0.39603997
V15	1.69442424	1.33750211
V19	0.10332047	0.08028381
V20	-0.07219432	-0.00559610
V27	0.28322704	0.19409267

7.4.1.9 Discrimination of Simulium sanctipauli and S. squamosum

The squared Mahalanobis' distance between species using the 25 character set was 80.49 (unbiased D^2=63.4, e_{act}<0.0001) with no misallocated flies using resubstitution (e_{res}=0.0). A stepwise discriminant analysis produced a nine character subset which gave a D^2 of 72.48 (unbiased D^2 =66.44, e_{act}<0.0001) and no misallocated flies (e_{res}=0.0). The dimension reduction technique described in section 7.2.2 resulted in a two character subset:

[V10,V20]

i.e. one antennal and one wing character with a Mahalanobis' squared distance of 41.42 (unbiased D^2=40.73, e_{act}=0.0007) and no misallocated flies using resubstitution (e_{res}=0.0).

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in no misallocated flies, one atypical fly (0.82%), no overlapping flies and 121 correctly allocated flies (99.2%).

The null hypothesis of equal wing tuft colouration was rejected at p<0.001 using a Wilcoxon two-sample rank sum test, so the prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.78. When the prior probabilities of species membership were adjusted according to a fly's wing tuft colouration, no flies were misallocated.

The null hypothesis of equal dispersion was not rejected at p<0.001 using the likelihood ratio test legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.79) shows that most discrimination is due to antennal length 2 with an opposite effect from wing length 3.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.71,0.71]$$

and accounted for 88.9% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 8.62 to 6.9401 indicating that size has some influence in discrimination, but that shape variation is much more important.

The mean vectors are shown as Table 7.80, the pooled within-species dispersion matrix as Table 7.81 and the linear discriminant functions as Table 7.82.

To conclude, there is significant multivariate morphometric differentiation between *S. sanctipauli* and *S. squamosum*, and the two character subset derived in this analysis can be expected to correctly classify over 99% of flies when it is known *a priori* that just this species pair can be expected.

Adjusting prior probabilities according to a fly's wing tuft colouration should be done in cases of doubt following typicality probability allocation.

Table 7.78

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. sanctipauli	*S. squamosum*
1	0.0008	0.9992
2	0.0207	0.9793
3	0.3623	0.6377
4	0.9386	0.0614
5	0.9976	0.0024

Table 7.79

Standardised Canonical Variate for *S. sanctipauli* and *S. squamosum*

Character	Canonical Variate
Antennal Length 2	3.3654
Wing Length 3	-1.2291

Table 7.80

Mean Vectors for species *S. sanctipauli* and *S. squamosum*

	S. sanctipauli	*S. squamosum*
Antennal Length 2	501.27771429	405.58896552
Wing Length 3	1512.93442857	1506.67091954

Table 7.81

Pooled within-species dispersion matrix for species *S. sanctipauli* and *S. squamosum*

Character	V10	V20
V10	547.65873687	1494.64465864
V20	1494.64465864	6709.02769555

Table 7.82

Linear Discriminant functions for species *S. sanctipauli* and *S. squamosum*

	S. sanctipauli	*S. squamosum*
CONSTANT	-233.40249585	-180.56918886
V10	0.76497296	0.32574574
V20	0.05508573	0.15200366

7.4.1.10 Discrimination of Simulium sanctipauli and S. yahense

The squared Mahalanobis' distance between species using the 25 character set was 14.58 (unbiased D^2=11.64 e_{act}=0.044) with four misallocated flies using resubstitution (e_{res}=0.0305). A stepwise discriminant analysis produced an eleven character subset with a D^2 of 12.91 (unbiased D^2 =11.71, e_{act}=0.0435) and five misallocated flies (e_{res}=0.0382). The dimension reduction technique described in section 7.2.2 resulted in a six character subset:

[V3,V9,V11,V17,V20,V22]

i.e. one thorax, two antennal, two wing and one leg character with a Mahalanobis' squared distance of 10.07 (unbiased D^2=9.52, e_{act}=0.0615) and six misallocated flies using resubstitution (e_{res}=0.046), all six of which were *S. yahense* misclassified as *S. sanctipauli*.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in five misallocated flies (3.82%), four atypical flies (3.05%), six overlapping flies (4.58%) and 116 correctly allocated flies (88.55%).

The null hypotheses of equal wing tuft colouration and equal abdominal setal colouration were both rejected at $p<0.001$ using Wilcoxon two-sample rank sum tests. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.83. When the prior probabilities of species membership were adjusted according to a fly's wing tuft colouration, the number of flies misallocated was seven (e_{res}=0.053). When the prior probabilities of species membership were adjusted according to a fly's abdominal setal colouration category, the number of flies misallocated was four (the four *S. yahense* with abdominal setal colouration category one).

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.84) shows that the most important contrast of characters is between femur length 1 and wing length 3.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.40,0.39,0.37,0.43,0.43,0.44]$$

and accounted for 77.3% of pooled within-species variation. When the scores along this vectoi were introduced into the model as a covariable the canonical root fell from 1.7747 to 1.7412 indicating that size is not important in discriminating between these species.

The mean vectors are shown as Table 7.85, the pooled within-species dispersion matrix as Table 7.86 and the linear discriminant functions as Table 7.87.

To conclude, there is significant multivariate morphometric differentiation between *S. sanctipauli* and *S. yahense*, mainly involving shape variation. The six character subset can be expected to classify nearly 89% of flies correctly when it is known *a priori* that just this species pair can be expected.

Adjusting prior probabilities according to a fly's wing tuft colouration did not improve greatly improve error rate, so it is recommended that this should be done only in cases of doubt following typicality probability allocation, likewise with abdominal setal colouration.

Table 7.83

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. sanctipauli	*S. yahense*
1	0.9991	0.0009
2	0.9934	0.0066
3	0.9524	0.0476
4	0.7265	0.2735
5	0.2609	0.7391

Table 7.84

Standardised Canonical Variate for *S. sanctipauli* and *S. yahense*

Character	Canonical Variate
Thorax Length	0.4583
Antennal Length 1	0.6552
Antennal Segment 4	0.7283
Wing Length 2	-0.8580
Wing Length 3	-1.8563
Femur Length 1	1.1017

Table 7.85

Mean Vectors for species *S. sanctipauli* and *S. yahense*

	S. sanctipauli	*S. yahense*
Thorax Length	657.81216000	650.18355000
Antennal Length 1	336.06857143	319.98281250
Antennal Segment 4	49.91885714	45.00166667
Wing Length 2	457.41942857	471.03875000
Wing Length 3	1512.93442857	1576.95807292
Femur Length 1	657.87428571	657.64000000

Table 7.86

Pooled within-species dispersion matrix for species *S. sanctipauli* and *S. yahense*

Character	V3	V9	V11	V17	V20	V22
V3	2101.72893060	476.38310941	130.43457686	1090.34712382	3119.57653853	1493.70925395
V9	476.38310941	295.53507418	56.21923638	392.28948301	1071.46097048	559.31001329
V11	130.43457686	56.21923638	22.27158828	93.66614091	279.49498206	130.80799823
V17	1090.34712382	392.28948301	93.66614091	994.05007937	2279.75222448	1170.04748438
V20	3119.57653853	1071.46097048	279.49498206	2279.75222448	7856.66628688	3479.19583477
V22	1493.70925395	559.31001329	130.80799823	1170.04748438	3479.19583477	1884.81091827

Table 7.87

Linear Discriminant functions for species *S. sanctipauli* and *S. yahense*

	S. sanctipauli	*S. yahense*
CONSTANT	-239.03455218	-244.46837753
V3	0.09601691	0.06611952
V9	1.31761714	1.21212558
V11	-2.13424984	-2.55398160
V17	-0.09305999	-0.01291173
V20	0.21241801	0.27220438
V22	-0.30426702	-0.38037825

7.4.1.11 Discrimination of Simulium soubrense 'B' and S. soubrense

The squared Mahalanobis' distance between species using the 25 character set was 5.67 (unbiased D^2=4.87 e_{act}=0.135) with 19 misallocated flies using resubstitution (e_{res}=0.102). A stepwise discriminant analysis produced a ten character subset with a D^2 of 4.56 (unbiased D^2 =4.29, e_{act}=0.1509) and 27 misallocated flies (e_{res}=0.145). The dimension reduction technique described in section 7.2.2 resulted in an eight character subset:

[V3,V4,V16,V17,V20,V21,V22,V28]

i.e. two thorax, four wing and two leg character with a Mahalanobis' squared distance of 4.09 (unbiased D^2=3.89, e_{act}=0.162) and 26 misallocated flies using resubstitution (e_{res}=0.1398), of which two were *S. soubrense* 'B' misclassified as *S. soubrense*, and 24 *S. soubrense*misclassified as *S. soubrense* 'B'.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in 17 misallocated flies (9.14%), seven atypical flies (3.76%), 29 overlapping flies (15.6%) and 133 correctly allocated flies (71.51%).

The null hypothesis of equal wing tuft colouration was not rejected at p<0.001 using a Wilcoxon two-sample rank sum test, therefore the prior probabilities of species membership for each wing tuft colouration category were not calculated.

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, legitimising the use of the pooled dispersion matrix.

The standardised canonical variate (Table 7.88) shows that the most important contrast of characters is between wing length 3, femur length 1, wing length 2, and the negatively loading characters tibia length 2, :thorax width and wing length 1.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.33,0.37,0.37,0.37,0.37,0.25,0.37,0.38]$$

and accounted for 82.2% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 0.5287 to 0.4281 indicating that size is important in discriminating between these species. Such a small canonical root suggests that the between species variation is in reality very small, and perhaps biologically unmeaningful (Campbell 1982).

The mean vectors are shown as Table 7.89, the pooled within-species dispersion matrix as Table 7.90 and the linear discriminant functions as Table 7.91.

To conclude, there is significant multivariate morphometric differentiation between *S. soubrense* 'B' and *S. soubrense*, but this is very slight in comparison to other species-pair discriminant analyses. It is not recommended that this species pair be separated using this seven character subset.

Table 7.88

Standardised Canonical Variate for *S. soubrense* and *S. soubrense* 'B'

Character	Canonical Variate
Thorax Length	0.6167
Thorax Width	-1.5308
Wing Length 1	-1.2355
Wing Length 2	1.2174
Wing Length 3	1.9824
Radial Hair Number	0.4037
Femur Length 1	1.5214
Tibia Length 2	-2.2162

Table 7.89

Mean Vectors for species *S. soubrense* and *S. soubrense* 'B'

	S. soubrense	*S. soubrense* 'B'
Thorax Length	586.83394286	628.90365570
Thorax Width	854.03575714	876.48680506
Wing Length 1	707.60142857	738.09341772
Wing Length 2	419.78142857	448.12481013
Wing Length 3	1422.18357143	1504.37734177
Radial Hair Number	13.07142857	14.82278481
Femur Length 1	607.09285714	640.53417722
Tibia Length 2	600.24000000	622.53569620

Table 7.90

Pooled within-species dispersion matrix for species *S. soubrense* and *S. soubrense* 'B'

Characte	V3	V4	V16	V17	V20	V21	V22	V28
V3	2753.67086144	2540.46154488	2109.08913798	1346.92465004	3902.89663612	64.50138627	1705.81167474	1655.36329864
V4	2540.46154488	3623.86125944	2712.06032687	1715.20579338	5142.61995803	101.31871863	2203.91076242	2128.43777672
V16	2109.08913798	2712.06032687	2802.81829727	1600.95871383	4687.19056899	88.23615673	1906.70059038	1889.07338437
V17	1346.92465004	1715.20579338	1600.95871383	1101.65924939	2830.88356890	53.73658601	1187.42862995	1179.67898734
V20	3902.89663612	5142.61995803	4687.19056899	2830.88356890	9271.31758194	150.65254580	3552.95675473	3530.32897496
V21	64.50138627	101.31871863	88.23615673	53.73658601	150.65254580	8.70051694	69.42897417	68.36891029
V22	1705.81167474	2203.91076242	1906.70059038	1187.42862995	3552.95675473	69.42897417	1674.31753269	1521.45783500
V28	1655.36329864	2128.43777672	1889.07338437	1179.67898734	3530.32897496	68.36891029	1521.45783500	1538.07225801

Table 7.91

Linear Discriminant functions for species *S. soubrense* and *S. soubrense* 'B'

	S. soubrense	*S. soubrense* 'B'
CONSTANT	-145.13327813	-150.85930038
V3	-0.10416705	-0.08126909
V4	0.09896614	0.04785658
V16	-0.00356864	-0.04991006
V17	-0.21647810	-0.14537616
V20	0.02961330	0.06952539
V21	-2.63733323	-2.36590739
V22	0.10765730	0.17999523
V28	0.47859779	0.36632780

7.4.1.12 Discrimination of Simulium soubrense 'B' and S. squamosum

The squared Mahalanobis' distance between species using the 25 character set was 34.67 (unbiased D^2=26.69 e_{act}=0.0049) with no mis-allocated flies using resubstitution (e_{res}=0.0). A stepwise discriminant analysis produced a 12 character subset with a D^2 of 31.57 (unbiased D^2 =27.94, e_{act}=0.0041) and no misallocated flies (e_{res}=0.0). The dimension reduction technique described in section 7.2.2 resulted in an four character subset:

[V4,V10,V17,V20]

i.e. one thorax, one antennal, and two wing characters with a Mahalanobis' squared distance of 20.75 (unbiased D^2=19.83, e_{act}=0.01299) and one misallocated fly using resubstitution (e_{res}=0.0087), which was a *S. soubrense* 'B' misclassified as *S. squamosum*.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in one misallocated fly (0.87%), one atypical fly (0.87%), one overlapping flies (0.87%) and 112 correctly allocated flies (97.39%).

The null hypothesis of equal wing tuft colouration was rejected at p<Q.001 using a Wilcoxon two-sample rank sum test, therefore the prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.92. Adjusting the prior probability of species membership according to a fly's wing tuft colouration resulted in a single misallocated fly (e_{res}=0.0087).

The null hypothesis of equal dispersion was not rejected at p<0.001 using the likelihood ratio test so the pooled dispersion matrix was used.

The standardised canonical variate (Table 7.93) shows that the most important contrast of characters is between the positively loading character antennal length 2 and the negatively loading characters wing length 2, and wing length 3.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.51,0.48,0.51,0.51]

and accounted for 86.2% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 3.8905 to 3.7107 indicating that size is of little importance in discriminating between these species.

The mean vectors are shown as Table 7.94, the pooled within-species dispersion matrix as Table 7.95 and the linear discriminant functions as Table 7.96.

To conclude, there is significant multivariate morphometric differentiation between *S. soubrense* 'B' and *S. squamosum*. The four character subset can be expected to allocate correctly in over 97% of cases when it is known *a priori* that just this species pair can be expected. The lack of influence of size variation in discrimination implies that once a wider range of temporal and geographic variation is sampled, the error rate should not be substantially worse.

Adjusting the prior probabilities of species membership according to the fly's wing tuft colouration did not improve error rate, so it is not recommended that this should be done.

Table 7.92

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. soubrense 'B'	*S. squamosum*
1	0.0570	0.9430
2	0.2275	0.7725
3	0.5894	0.4106
4	0.8750	0.1250
5	0.9715	0.0285

Table 7.93

Standardised Canonical Variate for *S. soubrense* 'B' and *S. squamosum*

Character	Canonical Variate
Thorax Width	0.6841
Antennal Length 2	1.8070
Wing Length 2	-1.2721
Wing Length 3	-1.0838

Table 7.94

Mean Vectors for species *S. soubrense* 'B' and *S. squamosum*

	S. soubrense 'B'	*S. squamosum*
Thorax Width	854.03575714	855.21798621
Antennal Length 2	441.74571429	405.58896552
Wing Length 2	419.78142857	450.71724138
Wing Length 3	1422.18357143	1506.67091954

Table 7.95

Pooled within-species dispersion matrix for species *S. soubrense* 'B' and *S. squamosum*

Characte	V4	V10	V17	V20
V4	3225.08076155	1016.36998911	1400.14442037	4194.91822849
V10	1016.36998911	539.42068401	531.85084180	1540.33015320
V17	1400.14442037	531.85084180	842.04421310	2093.11773874
V20	4194.91822849	1540.33015320	2093.11773874	7289.52899841

Table 7.96

Linear Discriminant functions for species *S. soubrense* 'B' and *S. squamosum*

	S. soubrense 'B'	*S. squamosum*
CONSTANT	-191.16952958	-176.45339324
V4	-0.02693330	-0.08204850
V10	0.76679613	0.47157782
V17	-0.21570523	-0.03357885
V20	0.11050698	0.16390038

7.4.1.13 Discrimination of Simulium soubrense 'B' and S. yahense

The squared Mahalanobis' distance between species using the 25 character set was 13.05 (unbiased $D^2=10.27$ $e_{act}=0.0055$) with six misallocated flies using resubstitution ($e_{res}=0.0484$). A stepwise discriminant analysis produced an 11 character subset with a D^2 of 11.65 (unbiased D^2 =10.5, $e_{act}=0.0053$) and six misallocated flies ($e_{res}=0.0484$). The dimension reduction technique described in section 7.2.2 resulted in an eight character subset:

[V4,V10,V17,V19,V20,V27,V28,V29]

i.e. one thorax, one antennal, three wing and three leg characters with a Mahalanobis' squared distance of 8.11 (unbiased $D^2=7.51$, $e_{act}=0.0853$) and eight misallocated flies using resubstitution ($e_{res}=0.065$), two *S. soubrense* 'B' into *S. yahense*, and six *S. yahense* into *S. soubrense* 'B'.

Allocation using typicality probability of species membership with atypicality defined at $\alpha=0.01$ resulted in six misallocated flies (4.84%), five atypical flies (4.03%), 10 overlapping flies (8.07%) and 103 correctly allocated flies (83.06%).

The null hypotheses of equal wing tuft colouration and equal abdominal setal colouration were both rejected at $p<0.001$ using Wilcoxon two-sample rank sum tests. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.97. Adjusting the prior probability of species membership according to a fly's wing tuft colouration resulted in two misallocated flies (e_{res}=0.0161). Adjusting the prior probabilities according to a fly's abdominal setal colouration resulted in four misallocated flies (e_{res}=0.0323), these being the four *S. yahense* with abdominal setal colouration category one (Chapter four).

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test so the pooled dispersion matrix was used, but the multivariate statistical test for skewness (Mardia 1970) showed some departure from normality, which could account for the relatively large proportion of atypical flies.

The standardised canonical variate (Table 7.98) shows that there are only two negatively loading characters, thorax width and tibia length·2, indicating that this discriminant vector may have a considerable size component.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.36,0.32,0.35,0.35,0.36,0.37,0.37,0.36]$$

and accounted for 87.4% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 1.4407 to 0.755 confirming the fact that size is of importance in discriminating between these species.

The mean vectors are shown as Table 7.99, the pooled within-species dispersion matrix as Table 7.100 and the linear discriminant functions as Table 7.101.

To conclude, there is significant multivariate morphometric differentiation between *S. soubrense* 'B' and *S. yahense*, but size variation contributes significantly to this discrimination. Therefore the eight character subset can be expected to allocate correctly about 83% of the time, but once a wider range of size variation has been sampled, it is likely that this estimate is optimistic.

Adjusting the prior probabilities of species membership according to the fly's wing tuft colouration improved error rate, so it is recommended that this should be done. Adjusting the prior probabilities according to abdominal setal colouration is subject to the extreme influence that this character imposes (Chapter nine) but the better performance of this character over the subset derived in this analysis means that its use is recommended.

Table 7.97

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. soubrense 'B'	*S. yahense*
1	1.0000	0.0000
2	0.9988	0.0012
3	0.9711	0.0289
4	0.5677	0.4323
5	0.0489	0.9511

Table 7.98

Standardised Canonical Variate for *S. soubrense* 'B' and *S. yahense*

Character	Canonical Variate
Thorax Width	-1.6063
Antennal Length 2	0.2239
Wing Length 2	0.8375
Wing Width 2	0.9903
Wing Length 3	1.1733
Femur Length 2	0.2859
Tibia Length 2	-1.7129
Basitarsus Length	0.9400

Table 7.99

Mean Vectors for species *S. soubrense* 'B' and *S. yahense*

	S. soubrense 'B'	*S. yahense*
Thorax Width	854.03575714	906.62228125
Antennal Length 2	441.74571429	475.44625000
Wing Length 2	419.78142857	471.03875000
Wing Width 2	1349.70892857	1495.92739583
Wing Length 3	1422.18357143	1576.95807292
Femur Length 2	633.80142857	690.33750000
Tibia Length 2	600.24000000	648.41500000
Basitarsus Length 2	315.69107143	352.44218750

Table 7.100

Pooled within-species dispersion matrix for species *S. soubrense* 'B' and *S. yahense*

Characte	V4	V10	V17	V19
V4	3934.38975621	1599.54679697	1726.76572239	4858.91600003
V10	1599.54679697	987.09609292	740.83766329	2055.55129618
V17	1726.76572239	740.83766329	1052.89838519	2414.28330107
V19	4858.91600003	2055.55129618	2414.28330107	7656.33734523
V20	5065.53411486	2156.62348348	2468.48739038	7126.09981005
V27	2589.69969903	1134.30120386	1296.31614133	3440.14839687
V28	2376.82672350	1019.78553934	1208.88977213	3229.93999139
V29	1183.15732295	511.09617308	614.36624934	1661.39299409

Characte	V20	V27	V28	V29
V4	5065.53411486	2589.69969903	2376.82672350	1183.15732295
V10	2156.62348348	1134.30120386	1019.78553934	511.09617308
V17	2468.48739038	1296.31614133	1208.88977213	614.36624934
V19	7126.09981005	3440.14839687	3229.93999139	1661.39299409
V20	8460.19183903	3700.08764012	3485.06270144	1808.42069014
V27	3700.08764012	2023.16538478	1825.05343279	935.82996297
V28	3485.06270144	1825.05343279	1758.70078033	861.32035451
V29	1808.42069014	935.82996297	861.32035451	493.23935991

Table 7.101

Linear Discriminant functions for species *S. soubrense* 'B' and *S. yahense*

	S. soubrense 'B'	*S. yahense*
CONSTANT	-138.09186312	-173.61086040
V4	-0.12195821	-0.19099451
V10	0.23775813	0.25632349
V17	-0.09095115	-0.02952965
V19	0.11991277	0.14636684
V20	0.10357705	0.13332578
V27	-0.09082624	-0.07476517
V28	0.05108723	-0.05401815
V29	0.09895643	0.19821470

7.4.1.14 Discrimination of Simulium soubrense and S. squamosum

The squared Mahalanobis' distance between species using the 25 character set was 13.35 (unbiased D^2=11.92 e_{act}=0.0421) with nine misallocated flies using resubstitution (e_{res}=0.0367). A stepwise discriminant analysis produced a ten character subset with a D^2 of 12.49 (unbiased D^2 =11.93 e_{act}=0.0421) and eight misallocated flies (e_{res}=0.0327). The dimension reduction technique described in section 7.2.2 resulted in an eight character subset:

[V3,V4,V6,V10,V13,V24,V28,V29]

i.e. two thorax, one head, two antennal, and three leg characters with a Mahalanobis' squared distance of 11.92 (unbiased D^2=11.48, e_{act}=0.04152) and ten misallocated flies using resubstitution (e_{res}=0.0408), nine *S. soubrense* into *S. squamosum*, and one *S. squamosum* into *S. soubrense*.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in four misallocated flies (1.63%) , three atypical flies (1.22%), six overlapping flies (2.45%) and 232 correctly allocated flies (94.69%).

The null hypothesis of equal wing tuft colouration was rejected at $p<0.001$ using a Wilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.102. Adjusting the prior probability of species membership according to a fly's wing tuft colouration resulted in a 11 misallocated flies (e_{res}=0.0449).

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, so the pooled dispersion matrix was used.

The standardised canonical variate (Table 7.103) shows that antennal length 2 is important in discrimination, as well as thorax shape and leg length.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.34,0.37,0.37,0.34,0.25,0.38,0.38,0.38]

and accounted for 78.6% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 2.7528 to 2.6439 showing that size variation is of little importance relative to shape variation in discriminating between these species.

The mean vectors are shown as Table 7.105, the pooled within-species dispersion matrix as Table 7.106 and the linear discriminant functions as Table 7.107.

To conclude, there is significant multivariate morphometric differentiation between *S. soubrense* and *S. squamosum*, which is mainly shape variation. The eight character subset can be expected to allocate correctly in nearly 95% of the cases.

Adjusting the prior probabilities of species membership according to the fly's wing tuft colouration made the error rate worse, so it is not recommended that this should be done.

Table 7.102

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. soubrense	*S. squamosum*
1	0.3118	0.6882
2	0.4346	0.5654
3	0.5660	0.4340
4	0.6888	0.3112
5	0.7897	0.2103

Table 7.103

Standardised Canonical Variate for *S. soubrense* and *S. squamosum*

Character	Canonical Variate
Thorax Length	-0.4038
Thorax Width	0.6438
Head Width	-0.4703
Antennal Length 2	1.6163
Antennal Segment 6	0.6236
Basitarsus Length	-0.9714
Tibia Length 2	0.8248
Basitarsus Length	-0.8872

Table 7.104

Mean Vectors for species *S. soubrense* and *S. squamosum*

	S. soubrense	*S. squamosum*
Thorax Length	628.90365570	631.97175172
Thorax Width	876.48680506	855.21798621
Head Width	809.42156962	795.31917241
Antennal Length 2	456.40784810	405.58896552
Antennal Segment 6	46.11544304	39.08137931
Basitarsus Length 1	435.45113924	436.35310345
Tibia Length 2	622.53569620	612.96413793
Basitarsus Length 2	328.69841772	329.43793103

Table 7.105

Pooled within-species dispersion matrix for species *S. soubrense* and *S. squamosum*

Character	V3	V4	V6	V10	V13	V24	V28	V29
V3	2907.60441486	2664.11298272	1845.18036992	994.99138316	79.61103673	1243.10401386	1736.66642653	946.07944730
V4	2664.11298272	3760.51158467	2433.10407715	1251.91824591	101.85343706	1623.72048579	2191.73641037	1234.12716620
V6	1845.18036992	2433.10407715	2052.78305406	997.75622128	83.68869484	1162.11681398	1549.97195216	891.09574800
V10	994.99138316	1251.91824591	997.75622128	796.09547521	67.46957932	660.69797682	824.25232802	498.13886574
V13	79.61103673	101.85343706	83.68869484	67.46957932	12.79452738	57.26154752	69.73669632	43.90896423
V24	1243.10401386	1623.72048579	1162.11681398	660.69797682	57.26154752	906.30697719	1100.17549489	649.40660717
V28	1736.66642653	2191.73641037	1549.97195216	824.25232802	69.73669632	1100.17549489	1583.32221722	844.21730152
V29	946.07944730	1234.12716620	891.09574800	498.13886574	43.90896423	649.40660717	844.21730152	525.58342359

Table 7.106

Linear Discriminant functions for species *S. soubrense* and *S. squamosum*

	S. soubrense	*S. squamosum*
CONSTANT	-190.76465438	-170.00160277
V3	-0.09902805	-0.07312541
V4	-0.09027153	-0.12610120
V6	0.41966966	0.45519018
V10	0.20337566	0.05350385
V13	1.11867749	0.68023840
V24	-0.14500433	-0.03337458
V28	0.39341413	0.32216414
V29	-0.43487297	-0.30099220

7.4.1.15 Discrimination of Simulium soubrense and S. yahense

The squared Mahalanobis' distance between species using the 25 character set was 5.03 (unbiased D^2=4.51 e_{act}=0.1442) with 34 misallocated flies using resubstitution (e_{res}=0.134). A stepwise discriminant analysis produced a nine character subset with a D^2 of 4.28 (unbiased D^2 =4.12, e_{act}=0.155) and 35 misallocated flies (e_{res}=0.1378). The dimension reduction technique described in section 7.2.2 resulted in an eight character subset:

[V6,V18,V19,V22,V24,V25,V26,V29]

i.e. one head, two wing and five leg characters with a Mahalanobis' squared distance of 4.2 (unbiased D^2=4.05, e_{act}=0.1572) and 33 misallocated flies using resubstitution (e_{res}=0.1299), 22 *S. soubrense* into *S. yahense*, and 11 *S. yahense* into *S. soubrense*.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in 23 misallocated flies (9.06%), 10 atypical flies (3.94%), 24 overlapping flies (9.45%) and 197 correctly allocated flies (77.56%).

The null hypotheses of equal wing tuft colouration and equal abdominal setal colouration were both rejected at $p<0.001$ using Wilcoxon two-sample rank sum tests. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.108. Adjusting the prior probability of species membership according to a fly's wing tuft colouration resulted in 34 misallocated flies (e_{res}=0.1339). Adjusting the prior probabilities according to a fly's abdominal setal colouration resulted in four misallocated flies (e_{res}=0.0157), these being the four *S. yahense* with abdominal setal colouration category one (Chapter four).

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test legitimising the use of the pooled dispersion matrix

The standardised canonical variate (Table 7.108) shows that the main characters discriminating between this species pair are femur length 1, basitarsus length 1, and basitarsus length 2.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.35,0.35,0.35,0.36,0.37,0.36,0.34,0.36]

and accounted for 86.3% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 0.9954 to 0.807 showing that size is of some importance in discriminating between these species.

The mean vectors are shown as Table 7.109, the pooled within-species dispersion matrix as Table 7.110 and the linear discriminant functions as Table 7.111

To conclude, the eight character subset offers very poor discrimination between *S. soubrense* and *S. yahense*, and it is not recommended that they should be identified in this way.

Using the abdominal setal colour character gives a more satisfactory error rate, so it is recommended that this character be used on its own.

Table 7.107

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. soubrense	*S. yahense*
1	0.9864	0.0136
2	0.9405	0.0595
3	0.7747	0.2253
4	0.4279	0.5721
5	0.1399	0.8601

Table 7.108

Standardised Canonical Variate for *S. soubrense* and *S. yahense*

Character	Canonical Variate
Head Width	-0.5451
Wing Width 1	-0.6656
Wing Width 2	0.7507
Femur Length 1	-1.3216
Basitarsus Length	0.6647
Tarsus Segment 2	-1.3443
Tarsus Segment 3	0.6487
Basitarsus Length	2.4347

Table 7.109

Mean Vectors for species *S. soubrense* and *S. yahense*

	S. soubrense	*S. yahense*
Head Width	809.42156962	828.68378125
Wing Width 1	1009.80537975	1039.91921875
Wing Width 2	1423.85350633	1495.92739583
Femur Length 1	640.53417722	657.64000000
Basitarsus Length 1	435.45113924	462.27500000
Tarsus Segment 2	167.56481013	172.72166667
Tarsus Segment 3	125.46759494	131.40125000
Basitarsus Length 2	328.69841772	352.44218750

Table 7.110

Pooled within-species dispersion matrix for species *S. soubrense* and *S. yahense*

Character	V6	V18	V19	V22	V24	V25	V26	V29
V6	2124.90433379	2443.20135162	3350.45289437	1770.44352484	1224.49148122	403.85449661	318.43274511	937.62187253
V18	2443.20135162	4506.90502776	5215.71907686	2601.06808710	1771.16381253	580.51226840	492.37409063	1344.89810519
V19	3350.45289437	5215.71907686	8374.65043572	3474.59832884	2419.45455484	795.27150171	633.02918931	1807.00430882
V22	1770.44352484	2601.06808710	3474.59832884	1967.02331525	1242.25677083	410.31138581	319.55345630	954.17537914
V24	1224.49148122	1771.16381253	2419.45455484	1242.25677083	956.85661109	296.57958863	228.08255886	687.79443268
V25	403.85449661	580.51226840	795.27150171	410.31138581	296.57958863	116.88226380	85.98369376	228.53422402
V26	318.43274511	492.37409063	633.02918931	319.55345630	228.08255886	85.98369376	81.78923625	174.71857078
V29	937.62187253	1344.89810519	1807.00430882	954.17537914	687.79443268	228.53422402	174.71857078	558.47493550

Table 7.111

Linear Discriminant functions for species *S. soubrense* and *S. yahense*

	S. soubrense	*S. yahense*
CONSTANT	-186.57383097	-189.29742336
V6	0.44733902	0.42354367
V18	0.11935613	0.09946490
V19	0.07931896	0.09504821
V22	-0.17732607	-0.23746916
V24	-0.28476846	-0.24411637
V25	1.35795592	1.10923261
V26	-0.54475594	-0.40445037
V29	-0.43813258	-0.24808245

7.4.1.16 Discrimination of Simulium squamosum and S. yahense

The squared Mahalanobis' distance between species using the 25 character set was 20.64 (unbiased D^2=17.67 e_{act}=0.0178) with no mis-allocated flies using resubstitution (e_{res}=0.0). A stepwise discriminant analysis produced an 11 character subset with a D^2 of

19.73 (unbiased D^2 =18.42, e_{act}=0.016) and no misallocated flies (e_{res}=0.0). The dimension reduction technique described in section 7.2.2 resulted in an seven character subset:

[V9,V11,V15,V16,V22,V24,V29]

i.e. one head, two wing and five leg characters with a Mahalanobis' squared distance of 16.03 (unbiased D^2=15.32, e_{act}=0.0252) and one misallocated fly using resubstitution (e_{res}=0.0055), which was a *S. squamosum* misidentified as *S. yahense*.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in one misallocated fly (0.55%), four atypical flies (2.19%), two overlapping flies (1.09%) and 176 correctly allocated flies (96.17%).

The null hypotheses of equal wing tuft colouration and equal abdominal setal colouration were both rejected at p<0.001 using Wilcoxon two-sample rank sum tests. The prior probabilities of species membership for each wing tuft colouration category were calculated and are shown as Table 7.112. Adjusting the prior probability of species membership according to a fly's wing tuft colouration resulted in no misallocated flies (e_{res}=0.0). Adjusting the prior probabilities according to a fly's abdominal setal colouration resulted in four misallocated flies (e_{res}=0.0323), these being the four *S. yahense* with abdominal setal colouration category one (Chapter four).

The null hypothesis of equal dispersion was not rejected at p<0.001 using the likelihood ratio test and so the pooled dispersion matrix was used.

The standardised canonical variate (Table 7.113) shows that antennal length 1 and basitarsus length 2 contrast with wing length 1 and the other leg measurements.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.38,0.32,0.34,0.40,0.40,0.40,0.40]$$

and accounted for 79.6% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 4.0427 to 3.0228 implying that size has some influence in discrimination, but that shape differences are more important.

The mean vectors are shown as Table 7.114, the pooled within-species dispersion matrix as Table 7.115 and the linear discriminant functions as Table 7.116.

To conclude, there is significant multivariate morphometric differentiation between *S. squamosum* and *S. yahense*, mainly involving shape differences. The seven character subset can be expected to allocate correctly over 96% of the time.

Adjusting the prior probabilities of species membership according to the fly's wing tuft colouration improved error rate, so it is recommended that this should be done. However, adjusting the prior probabilities according to abdominal setal colouration is subject to the extreme influence that this character imposes (Chapter nine) so it is recommended that it should be used only in cases of doubt, otherwise all *S. yahense* with pale abdominal setal colouration will be incorrectly allocated.

Table 7.112

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. squamosum	*S. yahense*
1	1.0000	0.0000
2	0.9998	0.0002
3	0.9238	0.0762
4	0.0341	0.9659
5	0.0001	0.9999

Table 7.113

Standardised Canonical Variate for *S. squamosum* and *S. yahense*

Character	Canonical Variate
Antennal Length 1	1.7646
Antennal Segment 4	0.7048
Antennal Segment 8	0.4806
Wing Length 1	-0.8171
Femur Length 1	-0.8805
Basitarsus Length	-0.6487
Basitarsus Length	1.0703

Table 7.114

Mean Vectors for species *S. squamosum* and *S. yahense*

	S. squamosum	*S. yahense*
Antennal Length 1	276.18275862	319.98281250
Antennal Segment 4	36.13103448	45.00166667
Antennal Segment 8	37.96965517	46.06083333
Wing Length 1	734.04137931	770.69750000
Femur Length 1	631.34344828	657.64000000
Basitarsus Length 1	436.35310345	462.27500000
Basitarsus Length 2	329.43793103	352.44218750

Table 7.115

Pooled within-species dispersion matrix for species *S. squamosum* and *S. yahense*

Characte	V9	V11	V15	V16	V22	V24	V29
V9	308.74590209	46.21971990	45.24130861	792.98599417	634.17394018	408.80214036	302.42489782
V11	46.21971990	16.27798641	8.43954142	147.95524793	115.96479055	76.17205426	55.11946318
V15	45.24130861	8.43954142	10.92835261	129.31021791	100.58862046	61.18805632	48.38336886
V16	792.98599417	147.95524793	129.31021791	3270.62386317	2402.34416125	1474.16430291	1158.70728493
V22	634.17394018	115.96479055	100.58862046	2402.34416125	2072.96679870	1195.38910756	927.55830177
V24	408.80214036	76.17205426	61.18805632	1474.16430291	1195.38910756	836.71949537	600.89218237
V29	302.42489782	55.11946318	48.38336886	1158.70728493	927.55830177	600.89218237	495.40067109

Table 7.116

Linear Discriminant functions for species *S. squamosum* and *S. yahense*

	S. squamosum	*S. yahense*
CONSTANT	-145.00160013	-186.98928625
V9	0.75681927	1.00850247
V11	-0.80731730	-0.33643093
V15	-0.22155901	0.14686223
V16	-0.17568891	-0.23029386
V22	0.00051998	-0.07405566
V24	0.23273406	0.15061693
V29	0.44210122	0.61347053

7.4.2. OVERALL DISCRIMINATION

Samples of seven cytospecies from the area west of Togo were available for analysis, from four countries, Côte d'Ivoire, Guinea, Mali, and Sierra Leone (see Appendix one).

For this analysis, the single *S. damnosum s.s.* sample was included with the five *S. sirbanum* samples in a single artificial category designated 'Savanna'. This step was justified because the WHO Onchocerciasis Control Programme within this area regards both species to be dangerous vectors to be controlled wherever they are found, and morphometrically they are not distinctive (section 7.4.1.1). The single sample of *S. soubrense* 'B' was included within the rest of *S. soubrense*, because morphometrically this species is not distinctive enough to justify the extra cost involved in estimating more parameters (see Chapter six, and section 7.4.1.11), despite the chromosomal distinctiveness of this taxon (Chapter three, Post 1986).

The 25 character set (excluding wing tuft colouration and abdominal setal colouration) resulted in the matrix of Mahalanobis' squared distances shown as Table 7.117. This matrix shows some unexpected features, for example, while *S. yahense* is chromosomally close to *S. squamosum* (Vajime and Dunbar 1975), it is morphometrically closer to *S. soubrense* and *S. sanctipauli* than it is to *S. squamosum*. All of these distances are significant at $p<0.001$. The number of flies misclassified using resubstitution was 74 (e_{res}=0.1213, Table 7.118). The data set was too large to obtain an estimate of error rate using the 'leave-one-out' method of Lachenbruch and Mickey (1968), but the relatively large sample sizes probably means that the resubstituted error rate is not seriously biased.

Examining the table of resubstitution (Table 7.119) shows that the species with by far the greatest number of misidentifications is *S. soubrense*, with over 25% of this species being misallocated into other species.

A stepwise discriminant analysis on the 25 character set resulted in the rejection of only four characters. Therefore, the method described in section 7.2.2 was applied until an eleven character subset was obtained:

[V4,V6,V9,V11,V16,V17,V19,V20,V22,V28,V29]

i.e. one thorax, one head, two antennal, four wing and three leg characters.

The matrix of Mahalanobis' squared distances between species using this 11 character subset is shown as Table 7.119. This matrix shows essentially the same features as the full character set. The number of flies misallocated using this subset and resubstitution was 98 (e_{res}=0.1607, Table 7.120), while the 'leave-one-out' method resulted in 110 misclassifications (e_c=0.1803).

Allocating the flies using the typicality probability approach described in section 7.2.4, with a fly being defined as atypical if the none of the probabilities associated with the distance from each species' mean was greater than 0.01, then 479 (78.52%) were allocated correctly, 14 were atypical (2.29%), 38 (6.23%) were overlapping and 79 (12.95%) were incorrectly allocated. When the flies which were on the overlap of two species were allocated using the appropriate pairwise discriminant statistics, the number of correctly allocated flies rose to 502 (82.295%), the number overlapping fell to 11 (1.8%) and the number wrongly allocated rose to 83 (13.61%).

The null hypotheses of equal wing tuft colouration and equal abdominal setal colouration were both rejected at p<0.001, using a Kruskal-Wallis non-parametric one-way analysis of variance. Two sets of prior probabilities of species membership were calculated, one set from the wing tuft colour alone, the other set from both colour characters, these are shown as Tables 7.121 and 7.122. The method used to calculate these probabilities was described in section 7.2.3. These two sets of priors were calculated because of the extreme influence that abdominal setal colouration had on the analysis; any *S. yahense* flies with abdominal setal colouration category one were automatically classified into another species.

Adjusting the prior probability of a fly's species membership using the priors given in Table 7.121 resulted in 84 flies being wrongly allocated (e_{res}=0.1377), while adjusting the priors using both colour characters (Table 7.122) resulted in 68 flies being incorrectly allocated (e_{res}=0.1115).

Table 7.123 gives the standardised canonical variates, showing which of the eleven characters are important in discrimination, also the means of each species on each canonical variate are shown as Table 7.124 to aid interpretation.

The first canonical variate, with a canonical root of 5.2451, is clearly dominated by antennal length 1, such that flies at the positive end of this vector can be expected to have relatively larger antennae than those at the negative end. Thus, *S. sanctipauli* is at the positive end of the vector and 'Savanna' at the negative end. This confirms previous work on the morphology of the *S. damnosum* complex (see e.g. Garms 1977). Other characters with some influence along this vector are basitarsus length 2 and thorax width, the former is positively loading, the latter negative.

The second canonical variate has a canonical root of 0.9791, which is not very large (Campbell 1982). However, two species are quite well discriminated along this vector (*S. sanctipauli* and *S. squamosum*), so the vector is of importance. The most influential character is wing length 3, which has a large positive loading, so that wing size in *S. squamosum* can be expected to be relatively larger than in *S. sanctipauli*. A relationship between the two mid-leg characters, tibia length 2 and basitarsus length 2 is also of importance along this vector, with tibia length 2 loading positively and basitarsus length 2 loading negatively.

The other canonical variates have canonical roots of only 0.3187 and 0.0799 respectively, and so are not of any importance, especially as neither discriminates one species particularly well.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.32,0.31,0.25,0.2,0.32,0.31,0.31,0.32,0.32,0.32,0.32]

and accounted for 80.31% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable, effectively controlling for size variation, the first canonical root fell to 4.035, showing that although size has some influence along this vector, it is not as important as shape variation. The second canonical root fell to 0.9674 showing that size is not important along this vector.

The mean vectors are shown as Table 7.125, the pooled within-species dispersion matrix as Table 7.126 and the linear discriminant functions as Table 7.127.

To conclude, the overall discriminant analysis of western flies has revealed a great deal of multivariate morphological differentiation between the five taxa examined in this area. The statistics derived in this analysis should be of considerable assistance to attempts to understand more fully the relative vectorial importance of the different taxa in this area.

Table 7.117

Matrix of Mahalanobis' distances between species, 25 character set

	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
Savanna	0.0				
S. sanctipauli	47.92	0.0			
S. soubrense	25.43	11.42	0.0		
S. squamosum	13.89	36.66	12.69	0.0	
S. yahense	35.08	12.43	5.76	14.9	0.0

Table 7.118

Table of re-classifications, using resubstitution, 25 character set

	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
Savanna	202	0	1	3	0
S. sanctipauli	0	32	0	0	3
S. soubrense	5	10	139	8	24
S. squamosum	2	0	1	84	0
S. yahense	0	7	10	0	79

e_{res}=74/610=0.1213

Table 7.119

Prior probability of species membership for each wing tuft colour category

Wing Tuft Colour	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
1	0.664	0.0005	0.0476	0.2854	0.0
2	0.2568	0.0156	0.2493	0.4666	0.0027
3	0.0379	0.1655	0.4659	0.2724	0.0584
4	0.0013	0.4347	0.2152	0.0393	0.3095
5	0.0	0.3955	0.0344	0.002	0.5681

Table 7.120

Prior probability of species membership for each wing tuft and abdominal setal colour category

A[1]	B[2]	Savanna	*S. sanctipaul*	*S. soubrense*	*S. squamosum*	*S. yahense*
1	1	0.6665	0.0006	0.0476	0.2853	0.0
1	2	0.2662	0.0167	0.2497	0.4674	0.0
1	3	0.0399	0.1813	0.4915	0.2873	0.0
1	4	0.0019	0.6312	0.3103	0.0566	0.0
1	5	0.0	0.9139	0.0814	0.0046	0.0
2	1-5	0.0	0.0	0.0	0.0	1.0

A[1] Abdominal setal colour.
B[2] Wing tuft colour.

Table 7.121

Matrix of Mahalanobis' distances between species, 11 character subset

	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
Savanna	0.0				
S. sanctipauli	43.46	0.0			
S. soubrense	20.23	9.5	0.0		
S. squamosum	12.27	32.68	9.45	0.0	
S. yahense	30.73	9.65	4.35	12.41	0.0

Table 7.122

Table of re-classifications, using resubstitution (and 'leave-one-out'), 11 character subset

	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
Savanna	200(197	0(0)	2(2)	4(7)	0(0)
S. sanctipauli	0(0)	34(31)	1(3)	0(0)	0(1)
S. soubrense	5(5)	15(15)	121(119)	13(15)	32(32)
S. squamosum	4(4)	0(0)	4(4)	78(77)	1(2)
S. yahense	0(0)	4(5)	13(14)	0(1)	79(76)

e_{res}=98/610=0.1607 e_c=110/610=0.1803

Table 7.123

Standardised Canonical Variates

Character	CV I	CV II	CV III	CV IV
Thorax Width	-0.7428	-0.8216	-0.0093	-1.2503
Head Width	-0.3497	-0.1707	-0.4332	0.3250
Antennal Length 1	2.1559	-0.5048	-0.7147	-0.8025
Antennal Segment 4	0.3230	-0.7823	0.5356	0.2273
Wing Length 1	-0.3358	-0.9190	-0.0470	1.6083
Wing Length 2	-0.5758	0.6049	0.4459	-0.6043
Wing Width 2	-0.0215	-0.5004	1.0569	0.8070
Wing Length 3	0.5966	2.1948	-0.5248	-2.0623
Femur Length 1	0.2295	0.5962	-2.3321	2.0908
Tibia Length 2	-0.2896	-1.5068	0.6115	-1.5701
Basitarsus Length 2	0.9069	1.5298	1.8634	1.0686

Table 7.124

Species Means on Canonical Variates

Species	CV I	CV II	CV III	CV IV
Savanna	-2.9028	-0.4728	0.1692	-0.0403
S. sanctipauli	3.4005	-2.1675	0.4904	0.8259
S. soubrense	1.5073	-0.2749	-0.6867	-0.1472
S. squamosum	-0.4572	1.9737	-0.1852	0.3831
S. yahense	2.4832	0.5487	0.9564	-0.2766

Table 7.125

Mean Vectors

Character	Savanna	*S. sanctipauli*	*S. soubrense*
Thorax Width	859.40403204	906.40608000	873.10707742
Head Width	775.71054175	831.90817714	805.30381935
Antennal Length 1	253.08349515	336.06857143	307.02016129
Antennal Segment 4	35.08718447	49.91885714	42.24000000
Wing Length 1	705.99611650	771.73714286	733.50322581
Wing Length 2	435.87378641	457.41942857	443.85806452
Wing Width 2	1383.34075243	1488.56614286	1412.69195699
Wing Length 3	1407.11145631	1512.93442857	1492.00408602
Femur Length 1	597.99495146	657.87428571	635.50000000
Tibia Length 2	597.82776699	644.66057143	619.17935484
Basitarsus Length 2	305.07087379	344.64857143	326.74032258

Table 7.125 (continued)

Mean Vectors

Character	*S. squamosum*	*S. yahense*
Thorax Width	855.21798621	906.62228125
Head Width	795.31917241	828.68378125
Antennal Length 1	276.18275862	319.98281250
Antennal Segment 4	36.13103448	45.00166667
Wing Length 1	734.04137931	770.69750000
Wing Length 2	450.71724138	471.03875000
Wing Width 2	1411.51097701	1495.92739583
Wing Length 3	1506.67091954	1576.95807292
Femur Length 1	631.34344828	657.64000000
Tibia Length 2	612.96413793	648.41500000
Basitarsus Length 2	329.43793103	352.44218750

Table 7.126

Pooled within-species dispersion matrix

Character	V4	V6	V9	V11
V4	3938.81886706	2547.86417923	700.67149127	122.06071806
V6	2547.86417923	2170.96003419	537.16930604	88.17262381
V9	700.67149127	537.16930604	292.41436279	37.76227516
V11	122.06071806	88.17262381	37.76227516	14.47383944
V16	2830.31111161	2016.41932163	590.33707686	94.59387621
V17	1707.85388944	1227.64588087	365.56136579	56.38481987
V19	4660.59315723	3289.08156046	933.90669487	150.50287360
V20	4978.36830465	3572.66929859	1008.58724661	169.08697429
V22	2354.74558447	1703.22821304	485.32662599	80.14613543
V28	2274.59173427	1622.90781954	451.03237589	74.96679100
V29	1208.31487019	882.61668440	253.18259454	43.06575188

Character	V16	V17	V19	V20
V4	2830.31111161	1707.85388944	4660.59315723	4978.36830465
V6	2016.41932163	1227.64588087	3289.08156046	3572.66929859
V9	590.33707686	365.56136579	933.90669487	1008.58724661
V11	94.59387621	56.38481987	150.50287360	169.08697429
V16	2734.16453406	1507.75055961	3817.39933797	4281.72764426
V17	1507.75055961	1044.00879161	2383.22635347	2514.37284250
V19	3817.39933797	2383.22635347	7629.48029291	6894.47454785
V20	4281.72764426	2514.37284250	6894.47454785	8168.03898552
V22	1973.86763070	1181.72550851	3123.83941523	3457.55908978
V28	1910.00548457	1149.80753341	3058.94929888	3339.27956167
V29	1035.06316022	613.92763037	1619.15046300	1803.27475551

Character	V22	V28	V29
V4	2354.74558447	2274.59173427	1208.31487019
V6	1703.22821304	1622.90781954	882.61668440
V9	485.32662599	451.03237589	253.18259454
V11	80.14613543	74.96679100	43.06575188
V16	1973.86763070	1910.00548457	1035.06316022
V17	1181.72550851	1149.80753341	613.92763037
V19	3123.83941523	3058.94929888	1619.15046300
V20	3457.55908978	3339.27956167	1803.27475551
V22	1753.43979004	1600.05442471	842.56492900
V28	1600.05442471	1620.39537057	819.79925300
V29	842.56492900	819.79925300	486.25898631

Table 7.127

Linear Discriminant functions

	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
CONSTANT	-182.5282093	-254.03615624	-224.90401261	-205.64371422	-244.71672460
V4	-0.15031241	-0.21733740	-0.20081153	-0.21698026	-0.22003282
V6	0.34089754	0.30583341	0.31636906	0.32147748	0.29192484
V9	0.47002013	0.87270319	0.77099716	0.58770168	0.78939291
V11	-0.11107483	0.51258684	0.02045405	-0.31465100	0.10764204
V16	-0.09101005	-0.07669257	-0.12237300	-0..13247048	-0.14632048
V17	-0.12683996	-0.27332493	-0.20646178	-0.13679257	-0.18471890
V19	0.11571448	0.13393345	0.10339051	0.10209010	0.11586675
V20	0.12504147	0.10727173	0.15971974	0.18206171	0.17637932
V22	-0.19746609	-0.16590277	-0.13631686	-0.11886083	-0.20767140
V28	0.21674973	0.20685222	0.17287595	0.09688152	0.16570609
V29	-0.37117770	-0.20279346	-0.27723383	-0.16278160	-0.09361895

7.4.3 DISCUSSION

The discriminant analyses presented in this section have revealed that there is considerable multivariate morphometric variation within the *S. damnosum* complex in the area west or the Volta lake.

The two 'savanna' species, *S. damnosum s.s.* and *S. sirbanum* are phenetically similar, with over 10% phenetic overlap, and a considerable proportion of the differences between them being due to size variation. However, the pooled taxon 'savanna' shows very little morphological overlap with any of the other species in this area, the largest being with *S. squamosum*, at about 5%. The differentiation of the 'savanna' flies from the other species involves the relative length of the antenna, and also the relative length of the basitarsal segment of the mid-leg. Size also influences the differentiation of 'savanna' from the other species. but the size-free canonical roots are also the highest when this species is involved than for most other analyses. The ability to identity adult females of this species is of considerable practical importance considering the dangerous vectorial role of *S. damnosum s.s.*and *S. sirbanum*.

Simulium squamosum is also a well isolated species morphometrically, with the maximum overlap being about 5%. This rate of correct identification compares with that using enzyme electrophoresis (Meredith and Townson 1981, Garms and Zillman 1984), so that the two methods together should provide unequivocal identification of members of this species. This contrasts with larval cytotaxonomy, as *S. squamosum* and *S. yahense* are currently only distinguishable as samples in parts of the western area (Surtees and Post unpublished data).

Members of the *S. sanctipauli* subcomplex are morphologically very similar, and *S. yahense* is also very close to these species. Therefore it appears that the chromosomal evolution of the *S. sanctipauli* subcomplex and *S. yahense* has not been parallelled in morphology, or that these species have converged to a common morphology. *Simulium yahense* can be distinguished from the *S. sanctipauli* subcomplex with nearly 96% accuracy using the colour of the abdominal setae, although scoring this character is less objective, and hence more prone to error, than taking measurements.

Table 7.128 summaries the performance of the different allocation methods used in the overall analysis. The method which resulted in the largest proportion of correct identifications was the forecd allocation method with prior probabilities adjusted according to both wing tuft colour and abdominal setal colour. The typicality probability method resulted in over 6% of flies lying on an overlap of two or more species. This was a smaller proportion than the equivalent analysis in Togo and Benin because of the larger sample sizes giving narrower approximate confidence intervals. The proportion overlapping fell once the flies had been allocated using typicality probabilities calculated from the appropriate species-pair statistics.

To conclude, there is considerable morphological differentiation between members of the *S. damnosum* complex in the western area. In particular, 'savanna' and *S. squamosum* can be very successfully identified, as can *S. yahense* if abdominal setal colour is used. Members of the *S. sanctipauli* subcomplex cannot be very well distinguished, although the practical importance of this may not be great.

Table 7.128

Comparison of five methods of allocation for the Western area

	Forced[1]	Forced[2]	Forced[3]	Typicality	Typicality
Correct	512	526	542	479	502
Incorrect	98	84	68	79	83
Overlapping	na	na	na	38	11
Atypical	na	na	na	14	14

[1] Forced allocation without adjusted priors.

[2] Forced allocation with prior probabilities adjusted for wing tuft colour

[3] Forced allocation with prior probabilities adjusted for wing tuft colour and abdominal setal colour

[4] Typicality probability without subsequent species pair allocation of overlapping flies

[5] Typicality probability with subsequent species pair allocation of overlapping flies

CHAPTER EIGHT: GLOBAL DISCRIMINATION OF SPECIES.

8.1 INTRODUCTION

Chapter seven has developed the statistics for allocation assuming prior knowledge of the geographic origin of the fly to be allocated, and hence of the possible range of reference species. The purpose of this chapter is to develop the statistics necessary for allocation without prior geographic knowledge.

For this analysis *S. soubrense* and *S. soubrense* 'B' were treated as a single category,*S. soubrense* for the reasons given in section 7.1, as were *S. damnosum* and *S. sirbanum*, which were pooled into the category 'savanna'. Thus there were five reference groups for allocation, *S. soubrense*, *S. sanctipauli*,*S. squamosum*, *S. yahense*, and 'savanna'.

8.2 MATERIALS AND METHODS

The full list of samples is given in Appendix one, and the methods used were the same as described in Chapter seven.

8.3 RESULTS AND DISCUSSION

8.3.1 PAIRWISE DISCRIMINATION

8.3.1.1 Discrimination of 'savanna' and S. sanctipauli

A total of 249 'savanna' flies from eight samples, three *S. damnosum s.s.*, and three *S. sirbanum*, from five West African countries, and 61 *S. sanctipauli*, from two samples in two West African countries were examined (Appendix one).

The 25 character set described in Chapter four resulted in a Mahalanobis' squared distance of 55.28 (unbiased $D^2=50.61$, $e_{act}<0.001$), with no misallocations using resubstitution, $e_{res}=0.0$, showing that there is considerable morphological divergence between these species.

A 12 character set derived using stepwise discriminant analysis resulted in a squared distance of 53.69 (unbiased $D^2 =51.43, e_{act}<0.001$) with no misallocated flies. This 12 character subset was further reduced using the method described in section 7.2.2 to the four character set:

[V4,V9,V13,V29]

i.e. one thorax, two antennal and one leg measurement (Chapter four), which gave a Mahalanobis' squared distance of 44.29 (unbiased $D^2=43.57$, $e_{act}=0.0005$) between the species, and one fly misallocated using resubstitution (a 'savanna' fly misidentified as *S. sanctipauli*), $e_{res}=0.003$.

Allocation using typicality probability of species membership with atypicality defined at $\alpha=0.01$ resulted in one misallocated fly (0.323%), four atypical flies (1.29%) and 305 correctly allocated flies (98.39%).

The null hypothesis of equal wing tuft colour was rejected at $p<0.001$ (Wilcoxon two sample rank-sum test), so the adjusted prior probabilities of species membership according to a fly's wing tuft colour are given as Table 8.1. This shows that *S. sanctipauli* flies have darker wing tufts than 'savanna'. The effect of adjusting the prior probabilities of species membership according to wing tuft colour was to reduce the resubstituted error rate to zero.

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, legitimising the pooling of the individual species' dispersion matrices.

The standardised canonical variate (Table 8.2) shows that three of the characters: the two antennal characters and the leg character have positive loadings on the vector, while thorax width has a negative loading. Thus, *S. sanctipauli*, which is at the positive side of this vector has a relatively larger antenna than 'savannna'.

The first eigenvector of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.55,0.48,0.41,0.55]$$

and accounted for 64.5% of pooled within-species variance. When the scores along this vector were introduced into the model as a covariable, the canonical root fell from 7.0458 to 4.3566, indicating that size has some influence on discrimination, but that shape differences are important.

The mean vectors are shown as Table 8.3, the pooled within-species dispersion matrix as Table 8.4 and the linear discriminant functions as Table 8.5.

To conclude, there is significant multivariate morphometric differentiation between 'savanna' and *S. sanctipauli*. This variation includes both size and shape components, but discrimination is still very successful when size variation is controlled for. The four character subset can be expected to classify flies to their correct species in over 98% of cases when it is known *a priori* that just this species pair can be expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration improves allocation rate, so it is recommended that this should be done for flies of doubtful affinity following typicality probability allocation.

Table 8.1

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	'Savanna'	*S. sanctipauli*
1	1.0000	0.0000
2	0.9998	0.0002
3	0.0012	0.9988
4	0.0000	1.0000
5	0.0000	1.0000

Table 8.2

Standardised Canonical Variate for 'Savanna' and *S. sanctipauli*

Character	Canonical Variate
Thorax Width	-1.0102
Antennal Length 1	1.5531
Antennal Segment 6	1.1886
Basitarsus Length 2	0.7587

Table 8.3

Mean Vectors for species 'Savanna' and *S. sanctipauli*

	'Savanna'	*S. sanctipauli*
Thorax Width	867.74462410	909.84725902
Antennal Length 1	254.11867470	328.20737705
Antennal Segment 6	36.08449799	50.53508197
Basitarsus Length 2	307.67710843	347.06803279

Table 8.4

Pooled within-species dispersion matrix for species 'Savanna' and *S. sanctipauli*

Characte	V4	V9	V13	V29
V4	3793.50722634	451.12961410	71.96460652	1069.40693381
V9	451.12961410	212.63157417	21.20256590	145.85851731
V13	71.96460652	21.20256590	9.01493600	22.06170470
V29	1069.40693381	145.85851731	22.06170470	399.42636260

Table 8.5

Linear Discriminant functions for species 'Savanna' and *S. sanctipauli*

	'Savanna'	*S. sanctipauli*
CONSTANT	-185.82239913	-301.48495611
V4	-0.02075579	-0.12623403
V9	0.80451694	1.11869762
V13	1.12638086	2.34553278
V29	0.46986876	0.66882302

8.3.1.2 Discrimination of 'savanna' and S. soubrense

Two hundred and twenty-two *S. soubrense* from eight samples taken in three West African countries were examined in relation to the 249 'savanna' flies (Appendix one).

The 25 character set resulted in a Mahalanobis' squared distance between species of 20.15 (unbiased D^2=19.036, e_{act}=0.0146) with nine individuals misallocated using resubstitution, e_{res}=0.019. Stepwise discriminant analysis resulted in an initial subset of 13 characters with a D^2 of 19.63 (unbiased D^2=19.044, e_{act}=0.0146) and ten misallocated flies (e_{res}=0.021). Applying the method described in section 7.2.2 a five character resulted:

[V9,V13,V17,V19,V22]

i.e two antennal, two wing and one leg character, which had a Mahalanobis' squared distance of 17.09 between species (unbiased D^2 =16.87, e_{res}=0.02) and 11 misallocated flies using resubstitution (e_{res}=0.023), these being three 'savanna' flies and eight *S. soubrense* misallocated.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in 11 misallocated flies (2.34%), five atypical flies (1.06%), three overlapping flies (0.64%) and 452 correctly allocated flies (95.97%).

The null hypothesis of equal wing tuft colour was rejected at p<0.001 using a Wilcoxon two-sample rank sum test, so the prior probabilities of species membership according to a fly's wing tuft colour were calculated (Table 8.6). This shows that 'savanna' flies have lighter wing tufts than *S. soubrense*. When the prior probabilities were adjusted, however, the resubstituted error rate rose to 15/471 (e_{res}=0.038), due to the pale wing tufted *S. soubrense* from sample V1=19 (Chapter six) being heavily penalised against 'own-group' membership. Including wing tuft colouration in the linear discriminant function resulted in eleven misallocations, nine *S. soubrense* into 'savanna', for the same reasons. Therefore, for 'global' discrimination of these species it is not recommended that wing tuft colouration be used for adjusting prior probabilities.

The likelihood ratio test for equality of dispersion was rejected at p<0.0001, bringing into question the validity of pooling the dispersion matrices. However, the pooled dispersion matrix was used for the practical reasons discussed in section 7.2.5.

The standardised canonical variate (Table 8.7) shows that the most important character in discriminating these two species is antennal length 1. The antennal segment and the femur length of the first leg also load positively on the variable, whist the wing characters both load negatively. Thus, the relative size of the antenna, which is larger in *S. soubrense*, is the main feature discriminating between this species pair.

The first eigenvector of the pooled within-species correlation matrix was a size vector, had coefficients:

[0.43,0.34,0.48,0.48,0.49]

and accounted for 71% of pooled within-species variation. When the scores along this vector were introduced as a covariable into the model, the canonical root fell from 4.2758 to 3.8441, indicating that size was not an important discriminant relative to shape differences between these species.

The mean vectors are shown as Table 8.8, the linear discriminant function as Table 8.9 and the pooled within-species dispersion matrix as Table 8.10

To conclude, there is significant multivariate morphometric differentiation between 'savanna' and *S. soubrense*. This variation is mostly shape variation involving relative antennal size. The five character subset can be expected to classify flies to their correct species in over 95% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration is detrimental to error rate, so its use is not recommended.

Table 8.6

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	'Savanna'	*S. soubrense*
1	0.9181	0.0819
2	0.5336	0.4664
3	0.1045	0.8955
4	0.0118	0.9882
5	0.0012	0.9988

Table 8.7

Standardised Canonical Variate for 'Savanna' and *S. soubrense*

Character	Canonical Variate
Antennal Length 1	1.6258
Antennal Segment 6	0.7682
Wing Length 2	-0.6028
Wing Width 2	-0.5660
Femur Length 1	0.4483

Table 8.8

Mean Vectors for species 'Savanna' and *S. soubrense*

	'Savanna'	*S. soubrense*
Antennal Length 1	254.11867470	304.93783784
Antennal Segment 6	36.08449799	45.50576577
Wing Length 2	439.08530120	443.02162162
Wing Width 2	1389.53391566	1401.13479279
Femur Length 1	602.69012048	633.77135135

Table 8.9

Pooled within-species dispersion matrix for species 'Savanna' and *S. soubrense*

Character	V9	V13	V17	V19	V22
V9	289.78578427	33.96758774	340.38985869	928.19454943	435.82477951
V13	33.96758774	11.24929548	44.42128441	130.29886217	59.87858877
V17	340.38985869	44.42128441	1055.46127339	2433.44880791	1141.13146814
V19	928.19454943	130.29886217	2433.44880791	7989.75793638	3045.75962135
V22	435.82477951	59.87858877	1141.13146814	3045.75962135	1610.71537461

Table 8.10

Linear Discriminant functions for species 'Savanna' and *S. soubrense*

	'Savanna'	*S. soubrense*
CONSTANT	-148.26573187	-188.51546669
V9	0.44074801	0.66063765
V13	0.48058134	1.03014963
V17	-0.12677299	-0.20340642
V19	0.10345703	0.07731114
V22	0.13123615	0.17433713

8.3.1.3 Discrimination of 'savanna' and S.squamosum

The 25 character set resulted in a Mahalanobis' squared distance of 15.09 between species (unbiased D^2=14.15, e_{act}=0.03) with 13 misallocated flies (e_{res}=0.031). Stepwise discriminant analysis produced an initial subset of 14 characters with a squared distance of 14.78 (unbiased D^2=14.25,e_{act}=0.0295) and 15 flies misallocated, e_{res}=0.036. Applying the dimension reduction technique described in section 7.2.2 resulted in a six character subset:

[V4,V9,V12,V20,V28,V29]

i.e. one thorax, two antennal, one wing and two leg characters, giving a Mahalanobis' squared distance between species of 11.85 (unbiased D^2=11.65, e_{act}=0.044) and 19 flies misallocated (e_{res}=0.045), ten 'savanna' flies into *S. squamosum*, nine *S. squamosum* flies into 'savanna'.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in 13 misallocated flies (3.09%), three atypical flies (0.71%) six overlapping flies (1.43%) and 399 correctly allocated flies (94.77%).

The null hypothesis of equal wing tuft colour was rejected at p<0.001 using a Wilcoxon two sample rank sum test. The adjusted prior probabilities of species membership were therefore calculated and are shown as Table 8.11. When the prior probabilities were adjusted according to a fly's wing tuft colour, the number misallocated remained 19, as it did when wing tuft colouration was included in the linear discriminant function. Therefore no extra discriminatory power was provided by including this character for discriminating between these species.

The null hypothesis of equal dispersion was not rejected at p=0.0001, legitimising the use of the pooled within-species dispersion matrix (see section 7.2.5)

The standardised canonical variate (Table 8.12) shows the most important characters in discrimination are the length of the basitarsus of the second leg, length of the tibia of the second leg, wing length and thorax width. Antennal measurements have only a minor impact on discrimination. Wing length and basitarsal length load positively, while thorax width and tibia length load negatively.

The first eigenvector of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.44, 0.35, 0.30, 0.44, 0.45, 0.45]$$

and accounted for 76% of pooled-within species variation. When the scores along this vector were introduced into the model as a covariable, the canonical root fell from 2.8767 to 2.0056, showing that size has some influence as a discriminatory factor, but that shape variation is more important than pure size.

The mean vectors are shown as Table 8.13, the pooled within-species dispersion matrix as Table 8.14 and the linear discriminant functions as Table 8.15.

To conclude, there is significant multivariate morphometric differentiation between 'savanna' and *S. squamosum*. This variation is mostly shape variation involving the thorax width, wing length and the mid-leg. The six character subset can be expected to classify flies to their correct species in nearly 95% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration does not improve allocation rate, so it is not recommended that this should be done.

Table 8.11

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	'Savanna'	*S. squamosum*
1	0.6691	0.3309
2	0.2589	0.7411
3	0.0569	0.9431
4	0.0103	0.9897
5	0.0018	0.9982

Table 8.12

Standardised Canonical Variate for 'Savanna' and *S. squamosum*

Character	Canonical Variate
Thorax Width	-1.5329
Antennal Length 1	0.5197
Antennal Segment 5	0.2363
Wing Length 3	1.2627
Tibia Length 2	-1.4962
Basitarsus Length 2	2.2498

Table 8.13

Mean Vectors for species 'Savanna' and *S. squamosum*

	'Savanna'	*S. squamosum*
Thorax Width	867.74462410	891.07350698
Antennal Length 1	254.11867470	279.50755814
Antennal Segment 5	34.29670683	37.76953488
Wing Length 3	1414.92809237	1540.09226744
Tibia Length 2	601.62313253	637.39744186
Basitarsus Length 2	307.67710843	343.40406977

Table 8.14

Pooled within-species dispersion matrix for species 'Savanna' and *S. squamosum*

Character	V4	V9	V12
V4	4488.85965498	633.61858811	117.91589590
V9	633.61858811	238.80768696	21.98895923
V12	117.91589590	21.98895923	10.80613189
V20	5081.60222426	853.06093478	140.29902677
V28	2648.98664488	430.14450062	75.35378894
V29	1413.40302741	230.21063521	41.07309918

Character	V20	V28	V29
V4	5081.60222426	2648.98664488	1413.40302741
V9	853.06093478	430.14450062	230.21063521
V12	140.29902677	75.35378894	41.07309918
V20	7492.71565724	3441.29744197	1837.01409675
V28	3441.29744197	1879.71363826	956.95545023
V29	1837.01409675	956.95545023	553.35364360

Table 8.15

Linear Discriminant functions for species 'Savanna' and *S. squamosum*

	'Savanna'	*S. squamosum*
CONSTANT	-183.65011147	-225.40730245
V4	-0.06024785	-0.13796231
V9	0.71732731	0.80740560
V12	1.02865257	1.24842656
V20	0.26325429	0.30419872
V28	-0.06744587	-0.17762073
V29	-0.42217823	-0.15829280

8.3.1.4 Discrimination of 'savanna' and S. yahense

The 25 character set resulted in a Mahalanobis' squared distance between species of 44.28 (unbiased D^2=40.92, e_{act}=0.0007) with no flies misallocated using resubstitution (e_{res}=0.0). Stepwise discriminant analysis reduced this to a fourteen character subset with a D^2 of 43.58 (unbiased D^2=41.674, e_{act}=0.0006) and a single fly misallocated (e_{res}=0.003). Applying the method for dimension reduction described in section 7.2.2, a four character subset was derived,

[V4,V9,V20,V29]

i.e., one thorax, one antennal, one wing and one leg character, with a Mahalanobis' squared distance of 29.6 between species (unbiased $D^2=29.17, e_{act}=0.0035$) and a single fly misallocated as *S. yahense* ($e_{res}=0.003$). This fly had a posterior probability of species membership (0.4999,0.5001) for 'savanna' and *S. yahense* respectively, and the typicality probability confidence intervals were (0.0118,0.0067) for 'savanna' and (0.0129,0.0062) for *S. yahense*, showing that the fly was in reality atypical of both species at $p<0.01$.

Allocation using typicality probability of species membership with atypicality defined at $\alpha=0.01$ resulted in no misallocated flies, six atypical flies (1.74%) one overlapping fly (0.29%) and 338 correctly allocated flies (97.97%).

The null hypotheses of equal wing tuft colouration and equal abdominal setal colouration were both rejected at $p<0.001$ using a Wilcoxon two-samples rank sum test. The prior probabilities of species membership for each wing tuft category are shown as Table 8.16. No flies were misallocated when the wing tuft colouration was included in the linear discriminant function or when the prior probabilities were adjusted according to wing tuft colouration. By contrast, four *S. yahense* flies were misallocated when prior probabilities were adjusted according to abdominal setal colouration; these flies were *S. yahense* with colour category one (Chapter four). The problems with using the character objectively are discussed in Chapter nine.

The null hypothesis of equal dispersion was rejected at $p<0.001$ using the likelihood ratio test. However for the practical and statistical reasons discussed in section 7.2.5 the dispersion matrices were pooled.

The standardised canonical variate (Table 8.17) shows that the most important contrast in characters responsible for discrimination is between thorax width and antennal length 1, with the other two characters having only a minor influence on discrimination. *Simulium yahense* is at the positive end of this variable, having a larger antenna relative to 'savanna'.

The first eigenvector of the pooled within-species correlation matrix was a size vector, with coefficients

$$[0.52,0.43,0.52,0.52]$$

and accounted for 82% of pooled within-spècies variation. When the scores along this vector were introduced as a covariable, controlling for size, the canonical root fell from 5.98 to 3.535, indicating that size has some influence on discrimination, but that shape differences are more important, especially when it is realised that only 18% of pooled within-species variation is involved in discrimination once size is controlled for.

The mean vectors are shown as Table 8.18, the linear discriminant functions as Table 8.19 and the pooled within-species dispersion matrix as Table 8.20.

To conclude, there is significant multivariate morphometric differentiation between 'savanna' and *S. yahense*. This differentiation involves both size and shape variation but in the absence of size there is still considerable discrimination. The shape variation involves a relationship between thorax width and antennal length 1.

The four character subset can be expected to classify flies to their correct species in nearly 98% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration improves allocation rate, so it is recommended that this should be done in cases of doubt following typicality probability allocation. Adjusting the prior probabilities of species membership according to a fly's abdominal setal colouration does not improve allocation rate because of the problem that abdominal setal colouration is not 100% diagnostic for *S. yahense*. Therefore it is not recommended that the character be used routinely but instead it should be used in cases of doubt.

Table 8.16

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	'Savanna'	*S. yahense*
1	1.0000	0.0000
2	1.0000	0.0000
3	0.0978	0.9022
4	0.0000	1.0000
5	0.0000	1.0000

Table 8.17

Standardised Canonical Variate for 'Savanna' and *S. yahense*

Character	Canonical Variate
Thorax Width	-1.4570
Antennal Length 1	2.1645
Wing Length 3	0.5252
Basitarsus Length 2	0.8300

Table 8.18

Mean Vectors for species 'Savanna' and *S. yahense*

Character	'Savanna'	*S. yahense*
Thorax Width	867.74462410	906.62228125
Antennal Length 1	254.11867470	319.98281250
Wing Length 3	1414.92809237	1576.95807292
Basitarsus Length 2	307.67710843	352.44218750

Table 8.19

Linear Discriminant functions for species 'Savanna' and *S. yahense*

	'Savanna'	*S. yahense*
CONSTANT	-178.12879395	-263.32515247
V4	-0.10363023	-0.22159312
V9	0.70408309	1.05637118
V20	0.24719283	0.27282018
V29	-0.26813365	-0.11546584

Table 8.20

Pooled within-species dispersion matrix for species 'Savanna' and *S. yahense*

Character	V4	V9	V20	V29
V4	4224.07247057	635.18361138	4820.91533268	1243.52743755
V9	635.18361138	244.49179243	834.35947942	217.97832669
V20	4820.91533268	834.35947942	7166.92779778	1657.94539485
V29	1243.52743755	217.97832669	1657.94539485	472.76026577

8.3.1.5 Discrimination of Simulium sanctipauli and S. soubrense

The 25 character set resulted in a Mahalanobis' squared distance of 6.81 between species, (unbiased $D^2=6.18$, $e_{act}=0.1069$) with 25 flies misallocated using resubstitution ($e_{res}=0.088$). A stepwise discriminant analysis resulted in an initial 14 character subset with a squared distance of 6.43 (unbiased distance=6.09, $e_{act}=0.1086$), with 27 misallocated flies. The method for dimension reduction given in section 7.2.2 resulted in a nine character subset:

[V6,V14,V15,V16,V17,V19,V20,V24,V27]

i.e. one head, two antennal, four wing and two leg characters, having a Mahalanobis' squared distance of 5.37 (unbiased $D^2=5.18$, $e_{act}=0.1276$) and 29 misallocated flies ($e_{res}=0.1025$), seven of which were misallocated *S. sanctipauli*, 22 were misallocated *S. soubrense*.

Allocation using typicality probability of species membership with atypicality defined at $\alpha=0.01$ resulted in 19 misallocated flies (6.7%), nine atypical flies (3.18%), 25 overlapping flies (8.83%) and 230 correctly allocated flies (81.27%).

The null hypothesis of equal wing tuft colouration was rejected at $p<0.001$ using a Wilcoxon two-sample rank sum test. Therefore the prior probabilities of species membership were calculated according to a fly's wing tuft colouration using the method described in Chapter seven; these are shown as Table 8.21. When the prior probabilities were altered in this way 28 flies were misallocated ($e_{res}=0.0989$), while 27 were misallocated when wing tuft colouration was included in the linear discriminant function ($e_{res}=0.095$), so that this character is having only minor influence on discrimination.

The null hypothesis of equal dispersion was rejected using the likelihood ratio test at $p<0.0001$. The dispersion matrices were still pooled for the practical and statistical reasons given in section 7.2.5.

The standardised canonical variate (Table 8.22) shows that the most important contrast of characters in discrimination of these two species is between wing length 3 and basitarsus length 1. The canonical root was only 0.915 reflecting the relatively poor discrimination between these species.

The first eigenvector of the pooled within-species correlation matrix was mainly a size vector with coefficients:

[0.34,0.21,0.23,0.36,0.36,0.35,0.36,0.37,0.37]

and accounted for 73% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 0.915 to 0.7255, indicating that size variation is a significant proportion of the between-species variation. The importance of size variation in discrimination is discussed in greater detail in Chapter nine.

The mean vectors are shown as Table 8.23, the linear discriminant functions as Table 8.24 and the pooled within-species dispersion matrix as Table 8.25.

To conclude, there is significant multivariate morphometric differentiation between *S. sanctipauli* and *S. soubrense*, although this differentiation is not large. The differentiation involves both size and shape variation but size variation is a considerable component of discrimination. The shape variation involves a relationship between wing length 3 and basitarsus length 1. The nine character subset can be expected to classify flies to their correct species in over 80% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration does not significantly improve allocation rate, so it is not recommended that the character be used routinely.

Table 8.21

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. sanctipauli	*S. soubrense*
1	0.1453	0.8547
2	0.2559	0.7441
3	0.4103	0.5897
4	0.5846	0.4154
5	0.7401	0.2599

Table 8.22

Standardised Canonical Variate for *S. sanctipauli* and *S. soubrense*

Character	Canonical Variate
Head Width	-0.6145
Antennal Segment 7	0.4369
Antennal Segment 8	0.3857
Wing Length 1	0.8474
Wing Length 2	-0.9158
Wing Width 2	0.7277
Wing Length 3	-1.7100
Basitarsus Length 1	1.0585
Femur Length 2	0.7642

Table 8.23

Mean Vectors for species *S. sanctipauli* and *S. soubrense*

Character	*S. sanctipauli*	*S. soubrense*
Head Width	834.72223279	806.88387027
Antennal Segment 7	49.57967213	44.30486486
Antennal Segment 8	48.38032787	43.55639640
Wing Length 1	774.45639344	730.55351351
Wing Length 2	461.91540984	443.02162162
Wing Width 2	1479.41942623	1401.13479279
Wing Length 3	1522.32909836	1478.64736486
Basitarsus Length 1	461.10885246	432.29513514
Femur Length 2	694.68786885	659.01405405

Table 8.24

Linear Discriminant functions for species *S. sanctipauli* and *S. soubrense*

	S. sanctipauli	*S. soubrense*
CONSTANT	-221.66933357	-198.01356901
V6	0.28345358	0.31523435
V14	1.68731319	1.44047546
V15	0.69683414	0.47216568
V16	0.02301751	-0.01450255
V17	-0.24375687	-0.17865843
V19	0.05061038	0.03318706
V20	-0.02458801	0.01642916
V24	-0.12400258	-0.20545171
V27	0.29346838	0.25045209

Table 8.25

Pooled within-species dispersion matrix for species *S. sanctipauli* and *S. soubrense*

Character	V6	V14	V15	V16	V17	V19	V20	V24	V27
V6	1883.76682660	62.88650931	70.53286981	1647.19040184	1054.11490471	2883.48344565	3179.23135670	1001.93177282	1434.61892441
V14	62.88650931	12.15178626	8.45178754	66.53560759	40.37357351	127.20979513	126.00207483	39.59252467	51.93177857
V15	70.53286981	8.45178754	11.92571356	77.22545795	46.69724907	149.01901390	155.47490537	45.24714709	58.99575191
V16	1647.19040184	66.53560759	77.22545795	2421.69212195	1399.57821112	3610.91211572	4154.26366871	1198.99741830	1686.60200843
V17	1054.11490471	40.37357351	46.69724907	1399.57821112	1006.13740260	2345.59235201	2542.01867686	764.64662573	1070.32927098
V19	2883.48344565	127.20979513	149.01901390	3610.91211572	2345.59235201	8361.24114879	7626.77650072	2027.15060422	2889.49389325
V20	3179.23135670	126.00207483	155.47490537	4154.26366871	2542.01867686	7626.77650072	9045.06737298	2304.48047417	3224.88183294
V24	1001.93177282	39.59252467	45.24714709	1198.99741830	764.64662573	2027.15060422	2304.48047417	769.10953724	977.96635349
V27	1434.61892441	51.93177857	58.99575191	1686.60200843	1070.32927098	2889.49389325	3224.88183294	977.96635349	1484.68711308

8.3.1.6 Discrimination of Simulium sanctipauli and S. squamosum

The squared Mahalanobis' distance between species using the full 25 character set was 32.18 (unbiased D^2=28.558, e_{act}=0.0038) with no flies misallocated using resubstitution. A stepwise discriminant analysis produced a 13 character subset which gave a Mahalanobis' squared distance of 30.93 (unbiased D^2=29.06, e_{act}=0.0038) and no misallocated flies using resubstitution (e_{res}=0.0).

The dimension reduction technique given in section 7.2.2 resulted in a six character subset:

[V10,V13,V16,V18,V20,V22]

i.e. two antennal, three wing and one leg character, with a Mahalanobis' squared distance of 25.16 (unbiased D^2=24.4, e_{act}=0.0068) and no misallocated flies using resubstitution (e_{res}=0.0).

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in no misallocated flies, six atypical flies (2.58%), no overlapping flies and 227 correctly allocated flies (97.43%).

The null hypothesis of equal wing tuft colouration was rejected at p<0.001 using a Wilcoxon two-sample rank sum test, therefore the prior probabilities of species membership for each wing tuft colouration category were calculated using the method described in Chapter seven and are shown as Table 8.26. When the prior probabilities were adjusted no flies were misallocated using resubstitution, as was also the case when wing tuft colouration was included in the linear discriminant function.

The null hypothesis of equal dispersion was rejected at p<0.001 using the likelihood ratio test. For the reasons given in section 7.2.5 the pooled dispersion matrix was still used.

The standardised canonical variate (Table 8.27) shows that most discrimination between these species is achieved by antennal length 2 with respect to the other characters. Wing length 3 also has some opposite influence. *Simulium sanctipauli* is at the positive end of this variable indicating that this species has a relatively longer antenna than *S. squamosum*.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.37,0.32,0.43,0.43,0.43,0.44]

and accounted for 79.7% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 4.9 to 4.82 showing that size variation is negligible in discriminating between these species, despite the fact that nearly 80% of within-species variation is size variation.

The mean vectors are shown as Table 8.28, the linear discriminant functions as Table 8.29 and the pooled within-species dispersion matrix as Table 8.30.

To conclude, there is significant multivariate morphometric differentiation between *S. sanctipauli* and *S. squamosum*. The differentiation involves only shape variation which is the relative length of the antenna. The six character subset can be expected to classify flies to their correct species in over 97% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration does not alter an already very good allocation rate, so it is recommended that the character be used only in cases of doubt, following typicality probability allocation.

Table 8.26

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. sanctipauli	*S. squamosum*
1	0.0013	0.9987
2	0.0441	0.9559
3	0.6255	0.3745
4	0.9837	0.0163
5	0.9995	0.0005

Table 8.27

Standardised Canonical Variate for *S. sanctipauli* and *S. squamosum*

Character	Canonical Variate
Antennal Length 2	1.9016
Antennal Segment 6	0.7835
Wing Length 1	0.2548
Wing Width 1	-0.3820
Wing Length 3	-0.7374
Femur Length 1	-0.3777

Table 8.28

Mean Vectors for species *S. sanctipauli* and *S. squamosum*

Character	*S. sanctipauli*	*S. squamosum*
Antennal Length 2	490.54819672	415.62558140
Antennal Segment 6	50.53508197	40.36488372
Wing Length 1	774.45639344	758.70976744
Wing Width 1	1041.64434426	1048.84267442
Wing Length 3	1522.32909836	1540.09226744
Femur Length 1	658.95737705	654.56023256

Table 8.29

Linear Discriminant functions for species *S. sanctipauli* and *S. squamosum*

	S. sanctipauli	*S. squamosum*
CONSTANT	-218.41358418	-188.35870422
V10	0.53891100	0.31238504
V13	1.16221940	0.46913288
V16	-0.20335053	-0.22676802
V18	0.09262248	0.12133480
V20	0.24567002	0.28775022
V22	-0.30237415	-0.26036749

Table 8.30

Pooled within-species dispersion matrix for species *S. sanctipauli* and *S. squamosum*

Character	V10	V13	V16
V10	686.32956079	60.19675399	1066.78068583
V13	60.19675399	12.12667542	101.81901954
V16	1066.78068583	101.81901954	2942.82959999
V18	1274.19712450	127.45080241	3162.78558166
V20	1724.89487348	162.55085422	4307.87779348
V22	906.58547657	89.40313341	2221.13883910
Character	**V18**	**V20**	**V22**
V10	1274.19712450	1724.89487348	906.58547657
V13	127.45080241	162.55085422	89.40313341
V16	3162.78558166	4307.87779348	2221.13883910
V18	4461.52969955	5128.11500962	2722.01904411
V20	5128.11500962	7697.39292593	3592.05175656
V22	2722.01904411	3592.05175656	2038.62453148

8.3.1.7 Discrimination of Simulium sanctipauli and S. yahense

The squared Mahalanobis' distance between species using the 25 character set was 8.27 (unbiased D^2=6.88, e_{act}=0.095) with 14 misallocated flies using resubstitution (e_{res}=0.089). A stepwise discriminant analysis resulted in an initial subset of 10 characters giving a Mahalanobis' squared distanceof 7.33 (unbiased D^2=6.86, e_{act}=0.0952) and 14 misallocated flies using resubstitution

(e_{res}=0.089). Applying the method described in section 7.2.2 for dimension reduction in the context of discrimination, a six character subset resulted:

[V3,V11,V16,V17,V20,V28]

i.e. one thorax, one antennal, three wing and one leg measurement. This character subset resulted in a Mahalanobis' squared distance of 6.07 (unbiased D^2 =5.796, e_{act}=0.1143) and 14 misallocated flies (e_{res}=0.089), seven of each species into the other.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in 9 misallocated flies (5.73%), three atypical flies (1.9%), 11 overlapping flies (7.0%) and 134 correctly allocated flies (85.35%).

The null hypotheses of equal wing tuft colouration and equal abdominal setal colouration were both rejected at $p<0.001$ using Wilcoxon two-sample rank sum tests. The prior probabilities of species membership according to a fly's wing tuft colouration were calculated and are shown as Table 8.31. When the prior probabilities were adjusted for wing tuft colouration the number of flies misallocated fell to eight (e_{res}=0.051), with four from each species misallocated. Including wing tuft colouration in the linear discriminant function resulted in 14 misallocated flies (e_{res}=0.089), nine *S. yahense* classified as *S. sanctipauli*, and five *S. sanctipauli* classified as *S. yahense*. Including abdominal setal colouration either in the linear discriminant function or by adjusting the prior probabilities resulted in four flies being misallocated (e_{res}=0.026), these flies being the four *S. yahense* with abdominal setal colouration character state one. The special nature of this character is discussed in further detail in Chapter nine.

The null hypothesis of equal dispersion was rejected at $p<0.001$ using the likelihood ratio test. For the reasons given in section 7.2.5 the dispersion matrices were still pooled.

The standardised canonical variate (Table 8.32) shows that the main discrimination between these species is effected by a contrast between wing length 3 and tibia length 2.

The first eigenvector of the pooled within-species correlation matrix was a size vector with coefficients:

$$[0.39,0.32,0.43,0.43,0.43,0.43]$$

and accounted for 77.3% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 1.4606 to 1.4386 showing that size variation is negligible in discriminating between these species.

The mean vectors are shown as Table 8.33, the linear discriminant functions as Table 8.34 and the pooled within-species dispersion matrix as Table 8.35.

To conclude, there is significant multivariate morphometric differentiation between *S. sanctipauli* and *S. yahense*. This differentiation involves mainly shape variation as in the absence of size there is still quite good discrimination. The shape variation involves a relationship between wing length 3 and femur length 2. The six character subset can be expected to classify flies to their correct species in over 85% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration does not greatly improve allocation rate, so it is not recommended that this should be done. Adjusting the prior probabilities of species membership according to a fly's abdominal setal colouration improves allocation rate but because of the problem that abdominal setal colouration is not 100% diagnostic for *S. yahense*, it is recommended that the character be used only in cases of doubt following typicality probability allocation.

Table 8.31

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. sanctipauli	*S. yahense*
1	0.9994	0.0006
2	0.9944	0.0056
3	0.9526	0.0474
4	0.6949	0.3051
5	0.2051	0.7949

Table 8.32

Standardised Canonical Variate for *S. sanctipauli* and *S. yahense*

Character	Canonical Variate
Thorax Length	-0.3837
Antennal Segment 4	-0.7846
Wing Length 1	-0.7147
Wing Length 2	1.0096
Wing Length 3	2.2319
Tibia Length 2	-1.4841

Table 8.33

Mean Vectors for species *S. sanctipauli* and *S. yahense*

Character	*S. sanctipauli*	*S. yahense*
Thorax Length	660.35632131	650.18355000
Antennal Segment 4	48.33967213	45.00166667
Wing Length 1	774.45639344	770.69750000
Wing Length 2	461.91540984	471.03875000
Wing Length 3	1522.32909836	1576.95807292
Tibia Length 2	650.97245902	648.41500000

Table 8.34

Linear Discriminant functions for species *S. sanctipauli* and *S. yahense*

	S. sanctipauli	*S. yahense*
CONSTANT	-174.22329721	-189.77089675
V3	0.06924040	0.04793217
V11	-0.06203030	-0.45841296
V16	-0.17491616	-0.20837892
V17	0.15615325	0.23834071
V20	0.20421750	0.26670128
V28	0.08935998	-0.00499819

Table 8.35

Pooled within-species dispersion matrix for species *S. sanctipauli* and *S. yahense*

Character	V3	V11	V16	V17	V20	V28
V3	1955.75824597	108.85775937	1687.92559940	940.42740710	2775.62344008	1267.40300477
V11	108.85775937	21.25674404	142.22053115	77.90040328	228.90582445	90.49433967
V16	1687.92559940	142.22053115	2783.37874972	1345.25345219	3836.82413251	1786.60875833
V17	940.42740710	77.90040328	1345.25345219	901.83462300	1994.79639946	993.08270767
V20	2775.62344008	228.90582445	3836.82413251	1994.79639946	7076.81380738	2895.72912168
V28	1267.40300477	90.49433967	1786.60875833	993.08270767	2895.72912168	1509.68528084

8.3.1.8 Discrimination of Simulium soubrense and S. squamosum

The squared Mahalanobis' distance between species using the 25 character set was 13.15 (unbiased D^2=12.28, e_{act}=0.0399) with 13 flies misallocated using resubstitution (e_{res}=0.033). A stepwise discriminant analysis produced a 12 character subset which gave a Mahalanobis' squared distance of 12.78 (unbiased D^2=12.356, e_{act}=0.0394) and 15 misallocated flies using resubstitution (e_{res}=0.038).

The dimension reduction technique described in Chapter seven resulted in a seven character subset:

[V10,V13,V18,V19,V23,V28,V29]

i.e. two antennal, two wing and three leg characters, with a Mahalanobis' squared distance of 11.56 (unbiased D^2=11.296, e_{act}=0.0464) and 18 misallocated flies using resubstitution (e_{res}=0.046), of which 13 were *S. soubrense* classified as *S. squamosum* and five were *S. squamosum* classified as *S. soubrense*.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in 16 misallocated flies (4.06%), nine atypical flies (2.3%), three overlapping flies (0.76%) and 366 correctly allocated flies (92.89%).

The null hypothesis of equal wing tuft colouration was rejected at p<0.001 using aWilcoxon two-sample rank sum test. The prior probabilities of species membership for each wing tuft colouration category were calculated using the method described in Chapter seven and are shown as Table 8.36. When the prior probabilities were adjusted the number of flies misallocated rose to 22 (e_{res}=0.056) whilst including wing tuft colouration in the linear discriminant function resulted in 18 misallocated flies (e_{res}=0.046). Therefore wing tuft colouration does not improve discrimination between these species.

The null hypothesis of equal dispersion was rejected at p<0.001 using the likelihood ratio test. For the reasons given in section 7.2.5 the pooled dispersion matrix was still used.

The standardised canonical variate (Table 8.37) shows that the main contrast in discriminating these species involves tibia length 1, and basitarsus length 2 on the positive side, and antennal length 2 andtibia length 2 on the negative side. *Simulium squamosum* is at the positive end of this vector.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.36,0.28,0.39,0.38,0.40,0.41,0.40]

and accounted for 80.2% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 2.8576 to 2.6199 showing the negligible effect of size in discriminating between these species, despite the fact 80% of within-species variation is size variation.

The mean vectors are shown as Table 8.38, the linear discriminant functions as Table 8.39 and the pooled within-species dispersion matrix as Table 8.40.

To conclude, there is significant multivariate morphometric differentiation between *S. soubrense* and *S. squamosum*. This differentiation involves mainly shape variation as in the absence of size there is still quite good discrimination. The seven character subset can be expected to classify flies to their correct species in over 92% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration does not improve allocation rate, so it is not recommended that this should be done.

Table 8.36

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. soubrense	*S. squamosum*
1	0.1905	0.8095
2	0.4062	0.5938
3	0.6655	0.3345
4	0.8526	0.1474
5	0.9439	0.0561

Table 8.37

Standardised Canonical Variate for *S. soubrense* and *S. squamosum*

Character	Canonical Variate
Antennal Length 2	-1.3938
Antennal Segment 6	-0.5038
Wing Width 1	0.2542
Wing Width 2	0.2544
Tibia Length 1	0.9448
Tibia Length 2	-1.0839
Basitarsus Length 2	1.1753

Table 8.38

Mean Vectors for species *S. soubrense* and *S. squamosum*

Character	*S. soubrense*	*S. squamosum*
Antennal Length 2	452.30756757	415.62558140
Antennal Segment 6	45.50576577	40.36488372
Wing Width 1	997.15797297	1048.84267442
Wing Width 2	1401.13479279	1466.39627907
Tibia Length 1	687.64756757	721.09465116
Tibia Length 2	617.96972973	637.39744186
Basitarsus Length 2	325.77297297	343.40406977

Table 8.39

Linear Discriminant functions for species *S. soubrense* and *S. squamosum*

	S. soubrense	*S. squamosum*
CONSTANT	-150.38773459	-140.32654738
V10	0.41885412	0.27668607
V13	0.74360050	0.36064597
V18	0.10132735	0.11313158
V19	0.03829657	0.04664547
V23	0.11749550	0.18100888
V28	-0.01654648	-0.09953965
V29	-0.45363441	-0.30004984

Table 8.40

Pooled within-species dispersion matrix for species *S. soubrense* and *S. squamosum*

Character	V10	V13	V18	V19	V23	V28	V29
V10	781.20558043	69.05704496	1353.66897306	1898.06680633	1022.42660720	933.35206049	530.81165688
V13	69.05704496	13.52189112	131.75112334	187.14544559	94.10687451	87.13826351	52.03809484
V18	1353.66897306	131.75112334	4712.98607757	5870.89534929	2935.76237058	2693.58657081	1478.19712810
V19	1898.06680633	187.14544559	5870.89534929	9708.31406014	3942.77347389	3612.26322503	1991.04171904
V23	1022.42660720	94.10687451	2935.76237058	3942.77347389	2287.88248971	1995.10209720	1090.85562028
V28	933.35206049	87.13826351	2693.58657081	3612.26322503	1995.10209720	1883.38427413	997.03216370
V29	530.81165688	52.03809484	1478.19712810	1991.04171904	1090.85562028	997.03216370	601.71878824

8.3.1.9 Discrimination of Simulium soubrense and S. yahense

The 25 character set resulted in a Mahalanobis' squared distance of 5.25 (unbiased D^2=4.82, e_{act}=0.136) with 36 misallocated flies using resubstitution (e_{res}=0.113). A stepwise discriminant analysis resulted in an initial subset of 15 characters giving a Mahalanobis' squared distanceof 5.02 (unbiased D^2=4.77, e_{act}=0.137) and 41 misallocated flies using resubstitution (e_{res}=0.129). Applying the dimension reduction method described in section 7.2.2 resulted in an eight character subset:

[V6,V10,V18,V22,V24,V25,V26,V29]

i.e. one head, one antennal, one wing and five leg measurement. This character subset resulted in a Mahalanobis' squared distance of 3.86 (unbiased D^2 =3.75,e_{act}=0.166) and 41 misallocated flies (e_{res}=0.129), 31 *S. soubrense* classified as *S. yahense* and 10 *S. yahense* classified as *S. soubrense*.

Allocation using typicality probability of species membership with atypicality defined at $\alpha=0.01$ resulted in 33 misallocated flies (10.38%), 12 atypical flies (3.8%), 25 overlapping flies (7.86%) and 248 correctly allocated flies (77.99%).

The null hypothesis of equal wing tuft colouration and equal abdominal setal colouration were both rejected at $p<0.001$ using Wilcoxon two-sample rank sum tests. The prior probabilities of species membership according to a fly's wing tuft colouration were calculated and are shown as Table 8.41. When the prior probabilities were adjusted for wing tuft colouration the number of flies misallocated only fell to 34 ($e_{res}=0.1069$), with 29 *S. soubrense* classified as *S. yahense* and five *S. yahense* classified as *S. soubrense*. Including wing tuft colouration in the linear discriminant function resulted in 29 misallocated flies ($e_{res}=0.091$), 23 *S. soubrense* classified as *S. yahense*, and six *S. yahense* classified as *S. soubrense*. Including abdominal setal colouration either in the linear discriminant function or in the prior probabilities resulted in four flies being misallocated ($e_{res}=0.013$), these flies being the four *S. yahense* with abdominal setal colouration character state one. The special nature of this character is discussed in further detail in Chapter nine.

The null hypothesis of equal dispersion was not rejected at $p<0.001$ using the likelihood ratio test, therefore the dispersion matrices were pooled.

The standardised canonical variate (Table 8.42) shows that the main character involved in discrimination between these species is basitarsus length 2, with femur length 1 and tarsus segment 2 having a significant opposite influence.

The first eigenvector of thepooled within-species correlation matrix was a size vector with coefficients:

[0.35,0.32,0.35,0.37,0.37,0.36,0.34,0.37],

and accounted for 83.7% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 0.8193 to 0.6148 showing that size variation has some influence of the already relatively poor discrimination between these species.

The mean vectors are shown as Table 8.43, the linear discriminant functions as Table 8.44 and the pooled within-species dispersion matrix as Table 8.45.

To conclude, there is significant multivariate morphometric differentiation between *S. soubrense* and *S. yahense*, although this differentiation is not great. This differentiation involves both size and shape variation as in the absence of size the discrimination deteriorates. The eight character subset can be expected to classify flies to their correct species in nearly 80% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration improves allocation rate, so it is recommended that this should be done in cases of doubt.

Table 8.41

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. soubrense	*S. yahense*
1	0.9832	0.0168
2	0.9343	0.0657
3	0.7756	0.2244
4	0.4564	0.5436
5	0.1694	0.8306

Table 8.42

Standardised Canonical Variate for *S. soubrense* and *S. yahense*

Character	Canonical Variate
Head Width	-0.7411
Antennal Length 2	0.2887
Wing Width 1	-0.2564
Femur Length 1	-1.2047
Basitarsus Length 1	0.6478
Tarsus Segment 2	-1.1176
Tarsus Segment 3	0.5920
Basitarsus Length 2	2.5188

Table 8.43

Mean Vectors for species *S. soubrense* and *S. yahense*

Character	*S. soubrense*	*S. yahense*
Head Width	806.88387027	828.68378125
Antennal Length 2	452.30756757	475.44625000
Wing Width 1	997.15797297	1039.91921875
Femur Length 1	633.77135135	657.64000000
Basitarsus Length 1	432.29513514	462.27500000
Tarsus Segment 2	165.95333333	172.72166667
Tarsus Segment 3	124.53063063	131.40125000
Basitarsus Length 2	325.77297297	352.44218750

Table 8.44

Linear Discriminant functions for species *S. soubrense* and *S. yahense*

	S. soubrense	*S. yahense*
CONSTANT	-195.36976155	-199.46621622
V6	0.45053367	0.41909206
V10	0.18662705	0.20430887
V18	0.16447478	0.15714705
V22	-0.21177450	-0.26513282
V24	-0.27046140	-0.23166637
V25	1.49677445	1.29398832
V26	-0.57435456	-0.45030602
V29	-0.45106015	-0.26045124

Table 8.45

Pooled within-species dispersion matrix for species *S. soubrense* and *S. yahense*

Character	V6	V10	V18	V22
V6	2052.10239071	1005.73982336	2280.95932025	1675.78371658
V10	1005.73982336	919.86051752	1327.76129138	986.45393206
V18	2280.95932025	1327.76129138	4358.58590824	2476.85048452
V22	1675.78371658	986.45393206	2476.85048452	1854.60560568
V24	1150.43329966	694.13496194	1663.37658176	1158.69955525
V25	373.05311287	225.45368038	550.82719596	383.35612405
V26	306.66309748	172.83653605	462.67142999	305.37873041
V29	879.53423597	515.24732941	1284.70297311	897.80567756
VARIABLE	V24	V25	V26	V29
V6	1150.43329966	373.05311287	306.66309748	879.53423597
V10	694.13496194	225.45368038	172.83653605	515.24732941
V18	1663.37658176	550.82719596	462.67142999	1284.70297311
V22	1158.69955525	383.35612405	305.37873041	897.80567756
V24	889.95130489	272.87887342	212.26754456	642.00407298
V25	272.87887342	107.98558312	79.66183143	212.12306725
V26	212.26754456	79.66183143	78.24852773	164.49460988
V29	642.00407298	212.12306725	164.49460988	525.87139075

8.3.1.10 Discrimination of Simulium squamosum and S. yahense

The Mahalanobis' squared distance between species using the 25 character set was 20.75 (unbiased D^2=18.72, e_{act}=0.015) with no misallocated flies using resubstitution (e_{res}=0.0). A stepwise discriminant analysis resulted in an initial subset of 12 characters giving a Mahalanobis' squared distance between species of 20.88 (unbiased D^2=19.86, e_{act}=0.013) and 1 misallocated fly using resubstitution (e_{res}=0.004). Applying the method described in section 7.2.2 for dimension reduction in the context of discrimination, a six character subset resulted:

[V9,V11,V15,V22,V23,V29]

i.e. three antennal, and three wing characters. This character subset resulted in a Mahalanobis' squared distance of 16.6 (unbiased D^2 =16.2,e_{act}=0.022) and three misallocated flies (e_{res}=0.011), 2 *S. squamosum* classified as *S. yahense* and 1 *S. yahense* classified as *S. squamosum*.

Allocation using typicality probability of species membership with atypicality defined at α=0.01 resulted in three misallocated flies (1.12%), seven atypical flies (2.6%) three overlapping flies (1.12%) and 255 correctly allocated flies (95.15%).

The null hypothesis of equal wing tuft colouration and equal abdominal setal colouration were both rejected at p<0.001 using Wilcoxon two-sample rank sum tests. The prior probabilities of species membership according to a fly's wing tuft colouration were calculated and are shown as Table 8.46. When the prior probabilities were adjusted for wing tuft colouration 2 flies were misallocated using resubstitution (e_{res}=0.075), both of which were *S. squamosum* classified as *S. yahense*. Including wing tuft colouration in the linear discriminant function resulted in 1 misallocated fly (e_{res}=0.0037), a *S. squamosum* classified as *S. yahense*. Including abdominal setal colouration either in the linear discriminant function or in the prior probabilities resulted in four flies being misallocated (e_{res}=0.0149), these flies being the four *S. yahense* with abdominal setal colouration character state one. The special nature of this character is discussed in further detail in Chapter nine.

The null hypothesis of equal dispersion was rejected at p<0.001 using the likelihood ratio test. However, for the practical and statistical reasons discussed in section 7.2.5, the dispersion matrices were still pooled.

The standardised canonical variate (Table 8.47) shows that antennal length 1 is the most important character involved in discrimination between these species, with tibia length 1 having a significant opposite influence.

The first eigenvector of the pooled within-species correlation matrix was a size vector with coefficients:

[0.40,0.36,0.38,0.44,0.44,0.43],

and accounted for 78.1% of pooled within-species variation. When the scores along this vector were introduced into the model as a covariable the canonical root fell from 3.8454 to 3.6416 showing that size variation has negligible influence on discrimination between these species.

The mean vectors are shown as Table 8.48, the linear discriminant functions as Table 8.49 and the pooled within-species dispersion matrix as Table 8.50.

To conclude, there is significant multivariate morphometric differentiation between *S. squamosum* and *S. yahense*. This differentiation involves mainly shape variation because when size variation is controlled, discrimination is still effective. The six character subset can be expected to classify flies to their correct species in over 95% of cases when it is known *a priori* that just this species pair is expected.

Adjusting the prior probabilities of species membership according to a fly's wing tuft colouration slightly improves allocation rate, so it is recommended that this should be done in cases of doubt following typicality probability allocation

Adjusting the prior probabilities of species membership according to a fly's abdominal setal colouration improves allocation rate but because of the problem that abdominal setal colouration is not 100% diagnostic for *S. yahense*, it is recommended that the character be used only in cases of doubt.

Table 8.46

Prior probabilities of species membership for each wing tuft category.

Wing tuft category	Species	
	S. squamosum	*S. yahense*
1	1.0000	0.0000
2	0.9993	0.0007
3	0.7880	0.2120
4	0.0091	0.9909
5	0.0000	1.0000

Table 8.47

Standardised Canonical Variate for *S. squamosum* and *S. yahense*

Character	Canonical Variate
Antennal Length 1	1.3978
Antennal Segment 4	0.6915
Antennal Segment 8	0.5655
Femur Length 1	-0.7261
Tibia Length 1	-0.9980
Basitarsus Length	0.4308

Table 8.48

Mean Vectors for species *S. squamosum* and *S. yahense*

Character	*S. squamosum*	*S. yahense*
Antennal Length 1	279.50755814	319.98281250
Antennal Segment 4	37.69744186	45.00166667
Antennal Segment 8	39.01674419	46:06083333
Femur Length 1	654.56023256	657.64000000
Tibia Length 1	721.09465116	721.60000000
Basitarsus Length 2	343.40406977	352.44218750

Table 8.49

Linear Discriminant functions for species *S. squamosum* and *S. yahense*

	S. squamosum	*S. yahense*
CONSTANT	-134.68348826	-171.69187356
V9	0.80352471	1.02174383
V11	-0.47696294	0.06143848
V15	-0.20676689	0.28235314
V22	-0.10656273	-0.16655907
V23	0.12162377	-0.04427324
V29	0.15396694	0.22205447

Table 8.50

Pooled within-species dispersion matrix for species *S. squamosum* and *S. yahense*

Character	V9	V11	V15	V22	V23	V29
V9	304.26573051	41.86146645	42.54871657	642.09971315	680.31886073	321.28657639
V11	41.86146645	15.12832559	8.09739261	126.98207858	137.48406333	63.39587609
V15	42.54871657	8.09739261	10.77909064	109.91556440	116.57001280	54.90745077
V22	642.09971315	126.98207858	109.91556440	2438.84470974	2485.28616321	1166.02930014
V23	680.31886073	137.48406333	116.57001280	2485.28616321	2774.23815895	1244.06463814
V29	321.28657639	63.39587609	54.90745077	1166.02930014	1244.06463814	648.06330373

8.3.2 OVERALL DISCRIMINATION

The full 25 character set excluding wing tuft colouration, abdominal setal colouration, and basitarsal spine number for the reasons given in section 7.2.3 resulted in the matrix of Mahalanobis' squared distances shown as Table 8.50a. All of these distances were significant at p<0.001, but examining the matrix reveals that there are three morphometric subgroups within *S. damnosum s.l.*: *S. squamosum*, 'savanna', and *S. sanctipauli/S. soubrense/S. yahense*, with members of the last group being relatively close together.

The matrix of classifications using resubstitution is given as Table 8.51. The overall resubstituted error rate was 0.144. Of the individual species' error rates, *S. soubrense* is the highest with nearly 30% of flies being misallocated into other species.

A stepwise discriminant analysis on the 25 character set only rejected four characters,antennal segment 5,antennal segment 7, radial hair number and tarsus segment 3 (Chapter four). This was regarded as too large a character set for practical allocation purposes, and so the method for dimension reduction in the context of allocation described in Chapter seven was used. This resulted in a 13 character subset:

[V4,V9,V11,V13,V16,V17,V18,V19,V20,V22,V24,V28,V29]

i.e. one thorax, three antennal, five wing and four leg characters (Chapter four).

The matrix of Mahalanobis' squared distances between species resulting from these 13 characters is shown as Table 8.54. The table of resubstitutions (Table 8.55) shows that the resubstituted error rate for the 13 character subset was 0.15. The individual error rates range from 0.044 ('savanna') to 0.293 (*S. soubrense*).

Allocation using the typicality proability method given in section 7.2.4 resulted in 29 atypical flies (3.6%), 54 overlapping flies (6.75%), 94 incorrectly allocated flies (11.8%) and 623 correctly allocated flies (77.9%). Once the flies which lay on a species-pair overlap had been allocated using the appropriate species-pair statistics given in sections 8.3.1.1- 8.3.1.10, then 29 flies remained atypical (3.6%), 16 were still overlapping (2%), 104 were incorrect (13%) and 651 were correctly allocated (81.4%).

The null hypotheses of equal wing tuft colouration and equal abdominal setal colouration were both rejected at $p<0.001$ using a Kruskal-Wallis test. Therefore the prior probabilities of species membership for each wing tuft colouration category and each abdominal setal colouration category were calculated using the method given in Chapter seven. These are shown as Tables 8.52 and 8.53.

When the prior probabilities were adjusted according to a fly's wing tuft colour, then the overall error rate remained at 0.15. When the priors were adjusted using both wing tuft colouration and abdominal setal colouration, then 89/799 flies were incorrectly allocated (e_{res}=0.108).

The standardised canonical variates (Table 8.56) show the characters of importance in discrimination and also give each species' mean score along each of the canonical vectors as Table 8.57. The first canonical variate accounted for 76% of total variance and was influenced most by antennal length 1 and basitarsus length 2. The species best discriminated along this vector were *S. sanctipauli/S. yahense* at the positive end and 'savanna' at the negative end. The second canonical variate accounted for a further 17% of total variance, and was strongly positively influenced by basitarsus length 2 and wing length 3, and strongly negatively by thorax width and tibia length 2. Discrimination between *S. squamosum* at the positive end and *S. sanctipauli* at the negative end was the most important function of this vector. The third and fourth canonical vectors, whilst being statistically significant are probably unimportant biologically because both have small canonical roots (Campbell 1982).

The null hypothesis of equal dispersion was rejected at p<0.0001 using the likelihood ratio test. However, for the practical and statistical reasons given in section 7.2.5 the dispersion matrices were still pooled.

The first principal component of the pooled within-species correlation matrix was a size vector with coefficients:

[0.29,0.23,0.19,0.20,0.30,0.28,0.29,0.28 0.30,0.30,0.30,0.30,0.30]

and accounted for 77.4% of pooled within-species variation. When the scores along this vector were included in the model as a covariable the canonical roots fell from:

[4.6377,1.0643,0.2908,0.1459]

to

[4.052,0.8732,0.2502,0.1366]

showing that size has a small influence on discrimination, principally along the second canonical variate.

The mean vectors are given as Table 8.58, the pooled dispersion matrix as Table 8.59 and the linear discriminant functions as Table 8.60.

To conclude, the 13 character subset derived in this analysis has shown that there is significant multivariate morphometric differentiation between the species of the *S. damnosum* complex. There is one major axis of between species variation, the first canonical variate, which is a vector expressing antennal length and mid-leg basitarsus length in relation to the rest of the body. It is the relative size (i.e. shape) of these characters which is responsible for the discrimination between the species. Characters of secondary, but significant importance include thorax width, wing length, and the mid-leg tibia length. The other characters contribute to a lesser extent to between-species variation, probably through correlation with characters more important in species discriminantion (Lubischew 1962).

Table 8.50a

Matrix of Mahalanobis' distances between species, 25 character set

	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
Savanna	0.0				
S. sanctipauli	37.74	0.0			
S. soubrense	21.07	6.53	0.0		
S. squamosum	11.91	25.89	12.91	0.0	
S. yahense	31.89	7.02	6.01	14.6	0.0

Table 8.51

Table of re-classifications, using resubstitution, 25 character set

	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
Savanna	241	0	3	5	0
S. sanctipauli	0	51	5	0	5
S. soubrense	7	21	156	7	31
S. squamosum	9	0	4	158	1
S. yahense	0	10	7	0	79

e_{res}=115/800=0.1438

Table 8.52

Prior probability of species membership for each wing tuft colour category

Wing Tuft Colour	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
1	0.5644	0.0009	0.0320	0.4026	0.0000
2	0.2766	0.0273	0.2357	0.4580	0.0024
3	0.0395	0.2432	0.5063	0.1519	0.0591
4	0.0012	0.4534	0.2273	0.0105	0.3076
5	0.0000	0.3316	0.0400	0.0003	0.6280

Table 8.53

Prior probability of species membership for each wing tuft and abdominal setal colour category

A^1	B^2	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
1	1	0.5645	0.001	0.0319	0.4026	0.0
1	2	0.2768	0.0285	0.2359	0.4587	0.0
1	3	0.0417	0.263	0.535	0.1604	0.0
1	4	0.0017	0.6551	0.3281	0.0152	0.0
1	5	0.0	0.8895	0.1097	0.0008	0.0
2	1-5	0.0	0.0	0.0	0.0	1.0

A^1=Abdominal setal colour category (see Chapter four).
B^2=Wing tuft colour category (see Chapter four).

Table 8.54

Matrix of Mahalanobis' distances between species, 13 character subset

	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
Savanna	0.0				
S. sanctipauli	35.85	0.0			
S. soubrense	19.85	5.89	0.0		
S. squamosum	10.79	23.48	11.04	0.0	
S. yahense	29.47	6.08	4.91	12.58	0.0

Table 8.55

Table of re-classifications, using resubstitution, 13 character subset

	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
Savanna	238	0	2	9	0
S. sanctipauli	0	48	5	0	8
S. soubrense	6	19	157	10	30
S. squamosum	7	0	4	159	2
S. yahense	0	10	8	0	78

e_{res}=120/800=0.15

Table 8.56

Standardised Canonical Variates

Character	CV I	CV II	CV III	CV IV
Thorax Width	-0.6323	-1.1533	-0.0874	-0.7891
Antennal Length 1	1.6678	-0.4761	-0.1345	-0.5461
Antennal Segment 4	0.1762	-0.3361	0.8469	0.1080
Antennal Segment 6	0.6211	-0.3631	-0.6173	0.3769
Wing Length 1	-0.1310	-0.6600	0.4007	1.0422
Wing Length 2	-0.3995	0.1800	0.0109	-0.8774
Wing Width 1	-0.1248	0.3747	-0.5498	1.0485
Wing Width 2	-0.2206	-0.1079	1.2190	0.1655
Wing Length 3	0.2579	1.1661	-0.4448	-2.4018
Femur Length 1	0.0959	0.8137	-2.5897	1.0942
Basitarsus Length	-0.5418	0.1549	0.8505	1.1273
Tibia Length 2	-0.0902	-1.6104	0.8879	-0.0030
Basitarsus Length	1.1033	2.1201	0.6605	-0.1962

Table 8.57

Species Means on Canonical Variates

Species	CV I	CV II	CV III	CV IV
Savanna	-2.7399	-0.6702	0.1969	-0.0604
S. sanctipauli	3.1093	-1.0771	0.8367	0.9707
S. soubrense	1.6280	-0.6058	-0.6789	-0.1120
S. squamosum	-0.6053	1.7760	-0.1843	0.2615
S. yahense	2.4507	0.6415	0.8578	-0.6699

Table 8.58

Mean Vectors

Character	Savanna	*S. sanctipauli*	*S. soubrense*
Thorax Width	867.74462410	909.84725902	872.95284865
Antennal Length 1	254.11867470	328.20737705	304.93783784
Antennal Segment 4	35.31759036	48.33967213	42.19909910
Antennal Segment 6	36.08449799	50.53508197	45.50576577
Wing Length 1	708.12433735	774.45639344	730.55351351
Wing Length 2	439.08530120	461.91540984	443.02162162
Wing Width 1	978.44353414	1041.64434426	997.15797297
Wing Width 2	1389.53391566	1479.41942623	1401.13479279
Wing Length 3	1414.92809237	1522.32909836	1478.64736486
Femur Length 1	602.69012048	658.95737705	633.77135135
Basitarsus Length 1	419.64240964	461.10885246	432.29513514
Tibia Length 2	601.62313253	650.97245902	617.96972973
Basitarsus Length 2	307.67710843	347.06803279	325.77297297

Table 8.58 (continued)

Mean Vectors

Character	*S. squamosum*	*S. yahense*
Thorax Width	891.07350698	906.62228125
Antennal Length 1	279.50755814	319.98281250
Antennal Segment 4	37.69744186	45.00166667
Antennal Segment 6	40.36488372	47.19750000
Wing Length 1	758.70976744	770.69750000
Wing Length 2	463.22372093	471.03875000
Wing Width 1	1048.84267442	1039.91921875
Wing Width 2	1466.39627907	1576.95807292
Wing Length 3	1540.09226744	1495.92739583
Femur Length 1	654.56023256	657.64000000
Basitarsus Length 1	453.87000000	462.27500000
Tibia Length 2	637.39744186	648.41500000
Basitarsus Length 2	343.40406977	352.44218750

Table 8.59

Pooled within-species dispersion matrix

Character	V4	V9	V11	V13	V16	V17	V18	V19	V20	V22	V24	V28
V4	4020.42684612	674.11534770	123.93425781	114.81760517	2819.63025059	1705.33879012	3479.02054715	4779.96597811	4901.55169491	2430.36654352	1660.98659993	2354.07338731
V9	674.11534770	290.36670606	36.22041760	37.77400930	573.81903537	349.30148288	645.92634072	947.15089526	998.76130192	479.44726851	332.33085966	450.54402160
V11	123.93425781	36.22041760	14.12867695	9.20617799	94.33719941	55.26046145	109.26739900	156.22669013	163.94461750	81.94251823	58.78564509	76.77856170
V13	114.81760517	37.77400930	9.20617799	12.62327975	96.15314058	56.36812476	113.81280202	162.58573156	168.26927707	81.64578963	57.16285608	76.01723060
V16	2819.63025059	573.81903537	94.33719941	96.15314058	2754.94213103	1510.11756107	2949.81478407	3918.69270629	4243.34067646	1999.44104911	1380.11258810	1953.41437341
V17	1705.33879012	349.30148288	55.26046145	56.36812476	1510.11756107	1048.71143120	1678.25278978	2414.76598315	2465.36734418	1188.65965191	836.14006698	1168.40568665
V18	3479.02054715	645.92634072	109.26739900	113.81280202	2949.81478407	1678.25278978	4112.98487223	5067.96192179	5134.02817590	2423.13912677	1641.46026172	2362.22639952
V19	4779.96597811	947.15089526	156.22669013	162.58573156	3918.69270629	2414.76598315	5067.96192179	8195.58351299	7155.46508249	3276.69614499	2268.03800475	3208.22684180
V20	4901.55169491	998.76130192	163.94461750	168.26927707	4243.34067646	2465.36734418	5134.02817590	7155.46508249	8224.10690353	3472.63849913	2387.76530057	3377.77197174
V22	2430.36654352	479.44726851	81.94251823	81.64578963	1999.44104911	1188.65965191	2423.13912677	3276.69614499	3472.63849913	1816.16396259	1154.99263296	1670.50349394
V24	1660.98659993	332.33085966	58.78564509	57.16285608	1380.11258810	836.14006698	1641.46026172	2268.03800475	2387.76530057	1154.99263296	881.10960971	1134.99996767
V28	2354.07338731	450.54402160	76.77856170	76.01723060	1953.41437341	1168.40568665	2362.22639952	3208.22684180	3377.77197174	1670.50349394	1134.99996767	1706.41061326
V29	1262.94455076	250.45044613	43.66888838	43.47540170	1061.19383404	629.00740560	1268.70992201	1727.30826743	1835.20200128	887.63622455	635.37505416	870.84426109

Table 8.59(continued)

Character	V29
V4	1262.94455076
V9	250.45044613
V11	43.66888838
V13	43.47540170
V16	1061.19383404
V17	629.00740560
V18	1268.70992201
V19	1727.30826743
V20	1835.20200128
V22	887.63622455
V24	635.37505416
V28	870.84426109
V29	516.02553624

Table 8.60

Linear Discriminant functions

	Savanna	*S. sanctipauli*	*S. soubrense*	*S. squamosum*	*S. yahense*
CONS- -TANT	-159.733167	-221.51972253	-196.68139491	-187.19981965	-219.46907820
V4	-0.02705877	-0.09002968	-0.06882842	-0.09453154	-0.09422150
V9	0.55777538	0.85369529	0.79324788	0.63004027	0.82112976
V11	0.15679247	0.48019491	0.15689796	0.02635963	0.32868239
V13	-0.21406578	0.41275768	0.31915059	-0.08214442	0.13634649
V16	-0.14728705	-0.13290588	-0.16488123	-0.17682596	-0.18035777
V17	0.04504152	-0.05064176	-0.00406664	0.02482137	0.00756623
V18	0.15185719	0.14968480	0.15052823	0.16886344	0.13539866
V19	0.04640510	0.04342813	0.02582277	0.03487736	0.04055290
V20	0.10955495	0.09344627	0.12554969	0.13566331	0.14721648
V22	-0.13414412	-0.14034412	-0.07841564	-0.06072305	-0.15103736
V24	0.05064384	0.00576679	-0.04229086	0.02890527	-0.02977846
V28	0.14400142	0.15937790	0.11569554	0.04474148	0.09978495
V29	-0.34843142	-0.14215441	-0.19269392	-0.09191129	-0.02668580

8.4 DISCUSSION

The 'global' discriminant analyses presented in this chapter have shown that there is a large amount of multivariate morphometric variation between females of the five taxa examined within the *S. damnosum* complex form West Africa. These analyses also demonstrate that multivariate statistical methods contribute valuable information to the morphological study of the complex, and should provide an additional, powerful method for the field identification of flies.

The species which is most successfully discriminated is 'savanna', the pooled *S. damnosum s.s.*, *S. sirbanum* category. This is true of both the species-pair analyses and the overall analysis. The species with which it has most phenetic overlap is *S. squamosum*, but this is only about 5%. The ability to identify this species with accuracy is of considerable importance, because both *S. damnosum s.s.* and *S. sirbanum* are known to be dangerous vectors of the more debilitating savanna strain of *O. volvulus*. The rate of correct identification using the multivariate statistical method is an improvement over published methods currently in routine use for adult female identification.

The morphological characters which were most commonly derived in discriminant analyses involving 'savanna' were thorax width, antennal length and mid-basitarsus length Chapter four) for all of which it is relatively smaller than the other species. 'Savanna' flies are also pale, mostly having wing tuft colour category one and two, and always showing abdominal setal colour category two.

Size variation was a significant component of the between-species analyses involving 'savanna', but the size-free canonical roots were still higher than for most other analyses, demonstrating that both size and shape differences are important in defining the distinctiveness of this taxon. Whether the consistently smaller size of 'savanna' flies is a taxonomic feature, or whether it is an environmentally mediated feature which might disappear once a wider range of geographic and seasonal variation has been sampled cannot be stated on the available data.

Simulium squamosum is the next most isolated species, with phenetic overlap ranging from about 3% to 8%. The species to which it is phenetically closest is 'savanna', followed by *S. soubrense* and *S. yahense*, and it is morphologically most distant from *S. sanctipauli*. It is of interest that *S. yahense*, to which it is chromosomally very close (Vajime and Dunbar 1975) is morphologically so distant. Clearly, morphological and chromosomal evolution in the *S. damnosum* complex are not always closely correlated, as was emphasised in Chapter six.

Morphological characters which are taxonomically useful in distinguishing *S. sanctipauli* in relation to the other a species include having a relatively broader thorax, and longer wing, and the leg characters. The species is also paler than most, except for 'savanna'.

The ability to identify this species using multivariate morphometrics is an improvement over the currently used morphological methods (e.g. Garms and Cheke 1985), but the rate of correct identification is not as good as that achieved using enzyme electrophoresis (Meredith and Townson 1981). In combination, the two methods should be able to identify flies with great accuracy, assuming the rate of incorrect identification for each method is independent.

The three species *S. soubrense*, *S. sanctipauli* and *S. yahense* show a larger degree of phenetic overlap. The morphological similarity of *S. soubrense* and *S. sanctipauli* might have been expected considering their chromosomal relatedness, but the proximity of *S. yahense* to this pair of species is unusual considering their chromosomal distance. Wing tuft colour parallels this morphological similarity, with *S. yahense*and *S. sanctipauli* being the darkest and *S. soubrense* being very variable (Chapter six). Abdominal setal colour is 96% diagnostic for *S. yahense* on the basis of this data. The cause of the phenetic similarity of the three species may be due to ecological similarities between them, as all three are predominantly forest dwellers (although *S. soubrense* is variable).

Table 8.61 summaries the results of the overall discrimination using the five allocation schemes used in this analysis, forced allocation without adjusted prior probabilities, forced allocation with priors adjusted for wing tuft colour, forced allocation with priors adjusted for wing tuft and abdominal setal colour, and typicality probability allocation with and without subsequent allocation of overlapping flies using the appropriate species-pair statistics.

The method which results in the largest number of correct allocations is forced allocation with priors adjusted according to a fly's wing tuft and abdominal setal colour, although comparison of the first two columns reveals that this is due entirely to abdominal setal colour. By automatically classifying all category 2 flies into *S. yahense* and all category 1 flies into some other species, the character is acting as a diagnostic character in traditional taxonomic analysis.

Typicality probability allocation without subsequent allocation of overlapping flies results in considerably fewer correct allocations, although as was emphasised in Chapter seven, the purpose of using this method is to provide a biologically more meaningful method of allocation. Ignoring atypicality or overlapping flies will ultimately lead to poorer allocation rates.

Table 8.61

Comparison of five methods of allocation

	Forced[1]	Forced[2]	Forced[3]	Typicality	Typicality
Correct	680	680	713	623	651
Incorrect	120	120	86	94	104
Overlapping	na	na	na	54	16
Atypical	na	na	na	29	29

[1] Forced allocation without adjusted priors.
[2] Forced allocation with prior probabilities adjusted for wing tuft colour
[3] Forced allocation with prior probabilities adjusted for wing tuft colour and abdominal setal colour
[4] Typicality probability without subsequent species pair allocation of overlapping flies
[5] Typicality probability with subsequent species pair allocation of overlapping flies

CHAPTER NINE: GENERAL DISCUSSION

9.1 INTRODUCTION

The previous chapters of this project have examined variation within and between species of the *S. damnosum* complex in West Africa from three distinct viewpoints:

1. 'Classical' larval polytene chromosome analysis,
2. Multivariate statistical analysis of larval polytene chromosome variation,
3. Multivariate statistical analysis of adult female morphological variation.

The purpose of this chapter is to discuss issues arising from these three approaches to the analysis of variation, to summarise and discuss the adult morphological variation, to provide worked examples of the mathematics of the different allocation procedures for identifying females, and to suggest a protocol for the field identification of adult females. Suggestions for future study will also be made.

9.2 CHROMOSOMAL VARIATION

The study of larval polytene chromosome variation in the *S. damnosum* complex using the 'classical' methods (e.g. Vajime and Dunbar 1975, Quillévéré 1975, Post 1986) is justified because species which are responsible for more serious disease transmission (e.g. *S. sirbanum*, *S. damnosum s.s.*) can be identified, and also species can be associated with ecological and bionomic factors which can influence control of the vector. Vector control problems such as the evolution of insecticide resistance and reinvasion of flies from uncontrolled areas can be clarified by detailed cytotaxonomic study of the problem flies, which in turn helps to overcome such problems more quickly than if the species complex were regarded as unitary.

The identification of a new chromosomal form, *S. sanctipauli* 'Djodji' from Togo (Chapter two, Surtees *et al.* 1988) demonstrates the continuing value of the usual approach to polytene chromosome variation, and the epidemiological work arising from its recognition (Cheke and Denke 1988) justifies the detailed analysis of chromosomal variation within previously recognised species.

However, the multivariate analyses presented in Chapter three demonstrate clearly that chromosomal variation in the *S. damnosum* complex is much more subtle and involved than can be revealed simply by using the classical approach. Important substantive results arising from this analysis include the recognition that the insecticide resistant flies within the *S. sanctipauli* subcomplex are chromosomally distinctive in relation to other *S. sanctipauli* , and may represent a new chromosomal form, a finding which had been overlooked using classical cytotaxonomy. That the variation revealed within *S. soubrense* 'Menankaya/Konkoure' cannot simply be broken into classically defined taxa without considerable loss of information is another substantive result of this analysis, as is the recognition of the chromosomal distinctiveness of *S. soubrense* 'B'. Apart from these substantive findings, however, the importance of the analyses shown in Chapter three lies in the new approach taken to the analysis of polytene chromosome variation within the *S. damnosum* complex. As emphasised in that chapter, the methods used and the materials available were not entirely optimal, nevertheless, it is clear that the method applied routinely to the *S. damnosum* complex will reveal greater understanding of the population structure of the different species which will include factors of importance in vector control and disease transmission such as tracing insecticide resistance, more detailed and subtle analysis of relative rates of gene flow and migration, and the recognition of new sibling species.

For the approach adopted in Chapter three to be applied to other parts of the *S. damnosum* complex, and to the complex as a whole, the homologies of chromosomal sequences amongst different taxa must first be established. It is known that this work is not complete (Post personal communication), and until it is, the method will be restricted to better known parts of the complex such as the *S. sanctipauli* subcomplex and the *S. squamosum* subcomplex.

9.3 ADULT FEMALE MORPHOLOGICAL VARIATION

The adult female morphological characters described and figured in Chapter four were used to examine intra- and inter-specific variation in the *S. damnosum* complex, with the results of these analyses presented in Chapters six, seven and eight.

The purpose of this section is to examine which of the morphological characters were most important in discrimination, to discuss the significance of size variation in the *S. damnosum* complex, to comment on the two colour characters and to discuss the importance of the departure from the null assumption of equal dispersion.

9.3.1 MORPHOLOGICAL CHARACTERS OF IMPORTANCE

Table 9.1 summarises the characters derived for each of the analyses presented in Chapters six, seven and eight. The first column refers to the chapter heading for each analysis.

9.3.1.1 Thorax Characters

Of the two thorax characters measured in this project, thorax width (V4) was more important taxonomically than thorax length (V3). Neither character was important in describing intraspecific variation (analyses 6.3.1-6.3.6), but of the 35 between-species analyses, thorax length was derived in six analyses, and thorax width in 16. Of these, none of the standardised canonical variate coefficients were relatively large for thorax length whereas 10 were considered important for thorax width. Thus thorax width seems to be a useful taxonomic character in the *S. damnosum* complex.

9.3.1.2 Head Characters

Of the two head characters used in this project, vertex width (V5) was not derived in any of the discriminant analyses, and so can be discounted as a taxonomic character. Head width (V6) was selected in 12 of the analyses, but was only considered important in the within-species analysis for *S. sirbanum*, and in the species pair discrimination of *S. soubrense* 'Beffa' and *S. sanctipauli* 'Djodji'. Therefore, neither character on its own can be considered of great importance taxonomically, although head width can be useful in conjunction with other characters.

9.3.1.3 Antennal Characters

Seven antennal characters were measured in this project, and they contributed considerably to the discrimination between species.

Antennal length one (V9), which is the shorter of the two measurements which included more than one segment,was selected in 21 of the analyses, while antenna length two (V10) was selected in 12 of the analyses. In no analysis were both selected together. Of the analyses in which they were selected, antennal length one was considered important in 14 and antennal length two in nine.

Of the within-species analyses, only in *S. sanctipauli* was antennal length considered important, indicating that the character may be relatively invariant within species, but of considerable importance between species.

The individual antennal segments 4-8 (V11-V15) were sometimes derived in the between-species analyses, but the only segment considered important was antennal segment 6 (V13), which was chosen in analyses including 'savanna' species and members of the *S. sanctipauli* subcomplex.

Antennal length and the relative compaction of the antennal segments 4-8 were considered important in previous morphological studies of the *S. damnosum* complex (e.g. Garms 1978, Quillévéré *et al.* 1977) and this has been confirmed in these analyses.

9.3.1.4 Wing Characters

Excluding wing tuft colour, which is dealt with in section 9.3.3, six characters were measured on the wing . Five of these were linear measurements, and one a count. This character, the number of hairs on the radial vein of the wing (V21) was chosen initially because it had previously been considered taxonomically important by Quillévéré and Sechan (1978). However it was derived in the characters subsets of only one of the 35 analyses in chapters six, seven and eight, and it was not considered important in this analysis. Therefore, the character can be discounted as a taxonomic character in the *S. damnosum* complex, confirming the findings of Garms (1978) and Townson and Meredith (1979).

Of the linear measurements, the longest wing length measurement (V20) was chosen in 24 of the 35 analyses and considered particularly important on three of the six within-species analyses and 14 of the 29 between-species analyses. Thus wing length appears to be a taxonomically useful character.

Of the other characters, V17 and V18 were quite frequently present in the final character subsets of the between-species analyses.

9.3.1.5 Leg Characters

Eight leg character were measured (excluding basitarsus length 2 which was rejected at an early stage in the analysis), five on the fore-leg, three on the mid-leg. The least important characters on the fore-leg were the very small tarsal segments. This is possibly a reflection of their small size decreasing the accuracy of measurement.

Femur length, tibia length and basitarsal length of the fore-leg were sometimes considered important in between-species analyses, but none were of great taxonomic value.

Of the three mid-leg characters, the least important was femur length, but tibia length and basitarsal length were frequently considered important, and were often chosen together, indicating that the relative proportion of these segments might be of taxonomic significance in the *S. damnosum* complex.

9.3.2 INFLUENCE OF SIZE VARIATION

Table 9.2 summarises the influence of size variation on the within- and between-species analyses presented in Chapters six, seven and eight.

The proportion of variance along the first principal component of the pooled within-groups correlation indicates the amount of error variance due to size. This proportion ranges from 58.6% to 88.9% showing that within-groups variation is predominantly along the size axis, i.e. individuals within a sample, or within a species are more likely to differ in size than shape.

The size-free canonical root is an indication of the shape-only differentiation between samples (6.3.1.1-6.3.1.6) or between species (7.3.1.1-8.3.2). As an informal guideline based on extensive practical experience, Campbell (1978) suggests that a canonical root less than about 0.75 to 1.0 is unlikely to be of practical use, although this guideline was not for size-free canonical roots. Using this guideline strictly and conservatively, only one of the within-species analyses (*S. soubrense*) shows meaningful shape differentiation, reflecting the chromosomal heterogeneity of this taxon.

Six of the between-species analyses have size-free canonical roots less than one, these are *S. damnosum s.s./S. sirbanum*, *S. soubrense* 'B'/*S. soubrense*, *S. soubrense* 'B'/*S. yahense*, *S. soubrense/S. yahense* and *S. sanctipauli/S. soubrense*. All other analyses have shape-only canonical roots greater than one and so in principle all are useful.

The influence of size variation, as measured by the ratio of the size-free canonical root to the size-in canonical root, expressed as a percentage, gives the influence of the pooled within-groups scatter along the size vector on the canonical variate analysis. This influence varies greatly, from very heavy influence (33.2%) to negligible influence (97.9%).

There appears to be no obvious pattern as to the influence of within-groups size variation on between-groups discrimination, beyond seeing that 'savanna' flies tend to be smaller. It seems that size variation is a random component of within-species variation which does not influence the taxonomically more important shape differences between species.

To conclude, whilst size variation is the predominant mode of variation in *S. damnosum s.l.*, the extent of between-species shape differentiation is unaffected by size variation.

Table 9.1

Characters of Importance

Section	Thorax	Head	Antenna	Wing	Fore-leg	Mid-leg
6.3.1			V9	V16	V22 V25	V27
6.3.2	V3 V4	V6		V17 V18 V20	V23	
6.3.3			V9	V20		
6.3.4	V4		V10	V17 V18 V20	V24	V27-V29
6.3.5			V9 V11	V19 V20	V22	V28 V29
6.3.6		V6		V17V19V20	V22	V27 V28
7.3.1.1			V9 V14	V18 V19		V27 V28
7.3.1.2		V6	V10	V17 V20		V27 V29
7.3.1.3	V6 V10	V18 V	0			
7.3.1.4		V6	V9 V13			V29
7.3.1.5	V3 V4		V9 V12	V17 V19	V23	V28 V29
7.3.1.6			V10	V18	V22	
7.3.2	V4	V6	V9	V18 V20	V23	V27V28V29
7.4.1.1	V4		V9 V14	V16 V17	V23	V29
7.4.1.2			V9	V17		V29
7.4.1.3	V3		V9 V13	V19		V29
7.4.1.4	V4		V10 V13	V19		V29
7.4.1.5	V4	V6	V9	V20		V27V28V29
7.4.1.6	V4		V9	V20		V29
7.4.1.7	V4		V10 V15	V16 V20	V24	V29
7.4.1.8			V11 V15	V19 V20		V27
7.4.1.9			V11	V20		
7.4.1.10	V3		V9 V11	V17 V20	V22	
7.4.1.11	V3 V4			V16V17V20V21	V22	V28
7.4.1.12	V4		V10	V17 V20		
7.4.1.13	V4		V10	V17V19V20		V27V28V29
7.4.1.14	V3 V4	V6	V10 V13		V24	V28 V29
7.4.1.15		V6		V17 V18	V22V24V25V26	V29
7.4.1.16			V9V11V15	V16	V22 V24	V29
7.4.2	V4	V6	V9 V11	V16V17V19V20	V22	V28 V29
8.3.1.1	V4		V9 V13			V29
8.3.1.2			V9 V13	V17 V19	V22	
8.3.1.3	V4		V9 V12	V20		V28 V29
8.3.1.4	V4		V9	V20		V29
8.3.1.5		V6	V14 V15	V16V17V19V20	V24	V27
8.3.1.6			V10 V13	V16 V18 V20	V22	
8.3.1.7	V3		V11	V16 V17 V20		V28
8.3.1.8			V10 V13	V18 V19	V23	V28 V29
8.3.1.9		V6	V10	V18	V22V24V25V26	V29
8.3.1.10			V9 V11 V15		V22 V23	V29
8.3.2	V4		V9 V11 V13	V16-V20	V22 V24	V28 V29

Table 9.2
The influence of size variation

Analysis[1]	% Var Size[2]	Size-free Root[3]	Size Influence[4]	Species[5]
6.3.1.1	74	0.717, 0.122	79.6, 53.4	dam
6.3.1.2	85.0	0.811, 0.243	69.5, 95.8	sir
6.3.1.3	69.5	0.828	99.84	san
6.3.1.4	83.7	1.829, 1.206,	84.3, 76	soub
6.3.1.5	69.6	0.927, 0.348	39.5, 35.3	squ
6.3.1.6	86	0.453, 0.364	35.3, 35.3	yah
7.3.1.1	66.4	4.81	93.95	bef/dam
7.3.1.2	80.1	1.23	48.4	bef/djo
7.3.1.3	84.4	2.43	60.2	bef/squ
7.3.1.4	58.6	3.4	38.6	dam/djo
7.3.1.5	73.5	2.05	67.9	dam/squ
7.3.1.6	85.6	2.83	99.7	djo/squ
7.3.2	2.82, 1.21	96.9, 55.8	overall	
7.4.1.1	73	0.581	68.4	dam/sir
7.4.1.2	77.4	3.25	53.7	sav/san
7.4.1.3	67.4	3.01	99.9	sav/sob
7.4.1.4	74.6	3.1	70.5	sav/soub
7.4.1.5	85.8	2.46	88.1	sav/squ
7.4.1.6	83.1	3.44	52.3	sav/yah
7.4.1.7	66.2	1.15	35.2	san/sob
7.4.1.8	68.8	1.15	93.2	san/soub
7.4.1.9	88.9	6.94	80.51	san/squ
7.4.1.10	77.3	1.74	98.1	san/yah
7.4.1.11	82.2	0.43	81.0	sob/soub
7.4.1.12	86.2	3.71	95.4	sob/squ
7.4.1.13	87.4	0.76	52.4	sob/yah
7.4.1.14	78.6	2.64	96.0	soub/squ
7.4.1.15	86.3	0.807	81.0	soub/yah
7.4.1.16	79.6	3.02	74.8	squ/yah
7.4.2	4.04, 0.97	76.9, 98.8	overall	
8.3.1.1	64.5	4.36	61.8	sav/san
8.3.1.2	71	3.84	89.9	sav/soub
8.3.1.3	76	2.01	69.7	sav.squ
8.3.1.4	82	3.54	59.1	sav/yah
8.3.1.5	73	0.73	79.3	san/soub
8.3.1.6	79.7	4.82	98.4	san/squ
8.3.1.7	77.3	1.44	98.5	san/yah
8.3.1.8	80.2	2.62	91.7	soub/squ
8.3.1.9	83.7	0.62	75	soub/yah
8.3.1.10	78.1	3.64	94.7	squ/yah
8.3.2	77.4	4.05, 0.87	87.4, 82.0	overall

[1] Numbers refer to chapter headings

[2] Percentage of variation along the first principal component of the pooled within-groups correlation matrix

[3] Canonical root resulting from multivariate analysis of covariance with size as the covariable

[4] Ratio of the size-free canonical root to the usual canonical root expressed as a percentage.

[5] dam= *S. damnosum s.s.*, sir=*S. sirbanum*, san=*S. sanctipauli*, soub=*S. soubrense*, sob=*S. soubrense* 'B', bef=*S. soubrense* 'Beffa', djo=*S. sanctipauli* 'Djodji', squ=*S. squamosum*, yah=*S. yahense*, sav='savanna' (pooled *S. sirbanum* and *S. damnosum s.s.*)

9.3.3 COLOUR CHARACTERS

Two colour characters were included in this project (Chapter four) because both were considered to be of taxonomic importance in the *S. damnosum* complex, wing tuft colour (e.g. Garms 1978) and abdominal setal colour (Garms and Zillman 1984). The special nature of these characters was mentioned in Chapter seven, where a method was described which exploited the taxonomic potential of these characters without risking certain statistical assumptions being invalidated.

9.3.3.1 Wing Tuft Colour

Figure 9.1 gives the frequency histograms for each species combined as in Chapter eight. The taxonomic importance of this character is obvious from this figure, but it is also clear that there is considerable overlap between species for the five categories of the character.

The original five character state system for scoring this character of Kurtak *et al.* was used. This system probably obscures more subtle expression of this character, so it is recommended that in future studies, either more categories be defined (thus splitting the heterogeneous middle category), or the character be expressed differently (such as a proportion).

The overall 'global' discriminant analysis (Chapter eight) showed that, in practice, the influence of this character on helping allocation of flies to each of the five species was negligible. Whilst it is extremely useful for distinguishing between, for example, 'savanna' and *S. yahense*, these species are already well separated morphometrically, whereas those species which need extra information to aid discrimination, e.g *S. soubrense* and *S. yahense*, overlap considerably for this character.

To conclude, the taxonomic importance of wing tuft colour has been confirmed in this project, although the method of scoring it has been criticised. In practise, the character helps to identify species which are already well distinguished, so its use is likely to be restricted to allocating flies of doubtful affinity after typicality probability.

9.3.3.2 Abdominal Setal Colour

The colour of the setae on the ninth abdominal tergite was recorded as a two-state character (Garms and Zillman 1984). In this project, it was found to be 95.8% diagnostic for *S. yahense*, meaning that the effect of this character on adjusted prior probabilities was diagnostic. The main problem with such extreme influence of a single character is that those *S. yahense* with character state one (white setae) are automatically incorrectly allocated, as are flies of other species with character state two, should they exist.

This extreme influence may be justified, as the effect of adjusting prior probabilities of species membership using this character is beneficial in discriminating between *S. yahense* and the species to which it is morphologically closest, *S. sanctipauli* and *S. soubrense*. However, the objectivity of scoring this character is not certain, and it is recommended that a detailed, double-blind study be undertaken to establish the true taxonomic status of this character.

9.3.4 COMMENTS ON DEPARTURES FROM EQUAL DISPERSION

The null hypothesis of equal dispersion was tested on each of the discriminant analyses because genuine rejection of this assumption is potentially serious, more so than departure from normality on the allocation rate of a given character subset (Campbell 1978). The pooled dispersion matrix was used throughout the analyses because to calculate separate dispersion matrices for each group in an analysis involves calculating more parameters with less accuracy. The effect, generally would be to introduce optimistic bias into the estimate of error rate for a particular analysis.

In the between-species analyses given in Chapters seven and eight, the null hypothesis was tested 35 times and rejected 12 times at a significance level of $p<0.0001$. Clearly, even though the testing of each null hypothesis is not independent, its rejection is too frequent to be by chance alone. A larger proportion of the rejections occurred in the 'global' analyses than in the regional analyses, which may be due to pooling resulting in departures from normality, to which the likelihood ratio test is particularly sensitive.

To conclude, given the poor performance of the statistical test for equal dispersion (Seber 1984), it is not possible to state with certainty whether the larger than expected rejections of the null hypotheses is due to genuine difference in covariance structure between species, or due to other causes. Determining the true nature of between-species variation in covariance structure is very important, and will need to be investigated in greater detail once more comprehensive sampling has been obtained. If differences in covariance structure exist, then allocation should be by the Quadratic discriminant function, or by typicality probability allocation to each species' mean vector weighted by the inverse of its own, rather than the pooled, dispersion matrix (Ambergen and Schaafsma 1984).

9.4 WORKED EXAMPLE OF ALLOCATION PROCEDURES

Two basic allocation procedures were used in these analyses, forced allocation and typicality allocation. Forced allocation assumes that the fly belongs to one of the reference species with probability 1, whereas typicality probability allocation calculates the probability that a fly is sampled from each species without reference to the other species. Forced allocation can be achieved in a number of ways, calculating the fly's score on each species' LDF, calculating each fly's posterior probability of species membership derived from its Mahalanobis' distance,and adjusting the prior probability of species membership in a way which reflects prior belief about the probability of the fly being one or other species. In this analysis, prior probabilities were calculated on the basis of colour, but other statements of prior belief could be used to adjust the probabilities.

As an example of the calculations of the statistics necessary for the different allocation procedures, three flies will be allocated using the overall 'global' statistics presented in section 8.3.1. Two (A and B) were collected from Bioko, West Africa by Dr. J. Mas of the Universidad de Barcelona. These flies belong to the *S. squamosum* subcomplex (i.e., *S. squamosum*, *S. yahense*, Post unpublished cytotaxonomic results). The other fly (C), was collected by Dr. P.J. McCall, Liverpool School of Tropical Medicine biting at cattle at Cynwyd, North Wales, and belonged to the species *S. variegatum*.

The vector of observations for the three flies, for the 13 character set given in Chapter 8 is shown below:

Character	Fly		
	A	B	C
Thorax Width	968.95	993.792	1043.48
Antennal Length 1	323.7	331.5	304.2
Antennal Segment 4	47.12	42.16	44.64
Antennal Segment 6	42.16	47.12	44.64
Wing Length 1	777.36	787.2	1230.0
Wing Length 2	492.0	492.0	747.84
Wing Width 1	1029.35	1029.35	1499.91
Wing Width 2	1529.32	1544.02	1970.47
Wing Length 3	1646.96	1646.96	2088.11
Femur Length 1	688.8	678.96	797.04
Basitarsus Length 1	492.0	482.16	619.92
Tibia Length 2	678.96	669.12	797.04
Basitarsus Length 2	358.8	354.9	460.2

i). LDF Allocation

Computationally, this is the simplest method. The transposed vector of observations for each fly is postmultiplied by the vector part of each linear discriminant function given in Table 8.60, and the resultant score added to the constant. The three flies A, B and C score on each species' LDF in the following way,

Species	Fly		
	A	B	C
savanna	213.62	215.45	258.55
S. sanctipauli	217.16	220.88	241.29
S. soubrense	219.89	224.73	241.85
S. squamosum	218.07	219.1	272.56
S. yahense	223.13	225.96	255.38

Based on the highest score, both Bioko flies would be allocated into *S. yahense* and the *S. variegatum* would be allocated into *S. squamosum*.

Prior probabilities could be adjusted according to wing tuft and abdominal setal colour using the priors given in Tables 8.52 or 8.53 by adding $\ln\pi_i$ to each constant.

Although it is not necessary to calculate the discriminant functions each time they are to be used, this can be done if it is decided to write a general allocation program, rather than simply storing each species' LDF. The vector part of the LDF is calculated by $S^{-1}\bar{x}_i$ and the constant part by $\bar{x}_i'S^{-1}\bar{x}_i$.

ii). Posterior Probability Allocation

The first step in allocation is to calculate the Mahalanobis' squared distance to each of the five reference mean vectors shown in Table 8.58 using the inverse of the pooled within-species dispersion matrix shown in Table 8.59.

$$D_i^2 = (x-\bar{x})'S^{-1}(x-\bar{x})$$

These distances D_i^2, are given below:

Species	Fly		
	A	B	C
savanna	37.34	37.4	342.32
S. sanctipauli	30.27	26.54	376.85
S. soubrense	24.81	18.85	375.72
S. squamosum	28.44	30.11	314.3
S. yahense	18.31	16.38	348.7

If forced allocation without adjusted prior probabilities is being used, then this is equivalent to assigning the fly to the species to which it is closest, thus the two Bioko flies would both be allocated into *S. yahense* and the *S. variegatum* would be allocated into *S. squamosum*.

To calculate the posterior probabilities of species membership, then the following calculation is performed:

$$R=\pi_i\exp(-0.5\ D_i^2\ )/SUM\ \pi_i\exp(-0.5\ D_i^2\ ),$$

where summation is over i=1...g, the number of reference groups (5), and π_i is the prior probability of species membership taken either from Table 8.52 or Table 8.53.

The following table gives the posterior probabilities of species membership for each of the three flies, without adjusted prior probabilities:

Species	Fly		
	A	B	C
savanna	<0.001	<0.001	<0.0001
S. sanctipauli	0.0024	0.0048	<0.0001
S. soubrense	0.037	0.224	<0.0001
S. squamosum	0.006	0.0008	0.999
S. yahense	0.954	0.7703	<0.0001

Thus the posterior probabilities of species membership are highest for *S. yahense* for the two Bioko flies, and highest for *S. squamosum* for the *S. variegatum* fly.

If the prior probabilities are adjusted for wing tuft colour, then the probabilities in Table 8.52 are used. Fly A has wing tuft colour category 4 and therefore enters the table at the fourth row, fly B has colour category 5 and enters the table in the final row, while fly C has colour category 1 and enters the table in row 1.

The following table gives the posterior probability of species membership for each of the flies, with each fly's prior probability adjusted using the appropriate prior probabilities:

Species	Fly		
	A	B	C
savanna	0.0	0.0	<0.0001
S. sanctipauli	0.0036	<0.0001	<0.0001
S. soubrense	0.028	0.018	<0.0001
S. squamosum	0.0002	<0.0001	0.999
S. yahense	0.9684	0.979	<0.0001

Once again the three flies are allocated into *S. yahense*, *S. yahense* and *S. squamosum* respectively.

Adjusting the prior probabilities of species membership using abdominal setal colour and wing tuft colour involves using the prior probabilities shown in Table 8.53. The two Bioko flies had character state 2 for this character and so enter the table in the last row, and Fly C had character state one for both characters and so enters the table in the first row. The following table gives the posterior probabilities of species membership for each fly:

Species	Fly		
	A	B	C
savanna	0.0	0.0	<0.0001
S. sanctipauli	0.0	0.0	<0.0001
S. soubrense	0.0	0.0	<0.0001
S. squamosum	0.0	0.0	0.999
S. yahense	1.0	1.0	0.0

Once again, this shows that the two Bioko flies show a strong affinity for *S. yahense* whilst, apparently the *S. variegatum* shows strong affinity for *S. squamosum*.

iii). Typicality Probability Allocation

However, to obtain greater detail of the affinities of the three flies for each of the reference species, it is necessary to calculate the typicality probabilities of species membership. Using the method of Ambergen and Schaafsma (1984) first involves calculating an unbiased estimate of Mahalanobis' distance:

$$((n-g-p-1)/n)D_i^2-p/n_i$$

where n=total sample size = 800,

g = number of reference groups = 5,

p = number of characters = 13,

n_i = sample size of i-th reference species (i=1...5).

The following table gives the unbiased distances of the three flies to each of the reference species:

Species	Fly		
	A	B	C
savanna	36.4	36.5	334.14
S. sanctipauli	29.33	25.69	367.69
S. soubrense	24.16	18.35	366.73
S. squamosum	27.69	29.32	306.78
S. yahense	17.74	15.85	340.23

Using these distances an estimate of its variance is then calculated:

$$(n-g-p-3)^{-1}\ \{2D^4 + 4(n-g-1)n_i^{-1}\ D^2 +2p(n-g-1)n_i^{-2}\}$$

where n is the total sample size on which the dispersion matrix is based, g is the number of reference species, p is the number of characters and n_i is the sample size of the i-th species.

The unbiased distance plus and minus half the variance gives an approximate confidence interval for the distance of the fly to each of the reference species:

Species	Fly		
	A	B	C
savanna	36.4±2	36.5±2	334.14±146.1
S. sanctipauli	29.33±2.1	25.69±1.7	367.69±185.8
S. soubrense	24.16±0.97	18.35±0.6	366.73±176.0
S. squamosum	27.69±1.3	29.32±1.45	306.78±124.32
S. yahense	17.74±0.78	15.85±0.66	340.23±155.8

The upper and lower distances thus derived can then be referred to the χ^2 distribution with p-degrees of freedom (13), to give the approximate confidence intervals:

Species	Fly		
	A	B	C
savanna	0.001,0.0002	0.001,0.0002	0.0,0.0
S. sanctipauli	0.0115,0.0029	0.0313,0.011	0.0,0.0
S. soubrense	0.0395,0.0222	0.1673,0.1247	0.0,0.0
S. squamosum	0.0151,0.0065	0.0095,0.0036	0.0,0.0
S. yahense	0.2011,0.1386	0.2955,0.2225	0.0,0.0

Clearly, therefore, the *S. variegatum* shows no affinity for any of the reference species. This example is extreme, in that the Mahalanobis' distances of this fly to the five species were all very large, but the principle is clearly demonstrated: that forced allocation can sometimes lead to an unrelated or atypical fly being forced into a species to which it does not belong.

Flies A and B both show greater affinity for *S. yahense* than for any other species, and their confidence intervals do not overlap with any other species. Therefore, especially when considering the forced allocation results, both flies should be allocated to *S. yahense*.

9.5 FIELD PROTOCOL

The exact statistics to be used in identifying adult females of the *S. damnosum* complex in the field clearly depends on the specific aims of the individual project. The purpose of this section is to give a key to the analyses presented in Chapters seven and eight, and hence to the characters to be measured. Some suggestions will also be made for adjusting the methods to suit a specific situation which may arise in the field.

Key to Discriminant Analyses, West Africa.

1).	Geographical region taken into account	2
	No account taken of geography	5
2).	Togo and Benin	3
	Western Area	4
3).	Overall discrimination only	7.3.2
	Overall and species-pair discrimination	7.3.1,7.3.2
	Species-pair discrimination only	7.3.1.1-7.3.1.6
4).	Overall discrimination only	7.4.2
	Overall and species-pair discrimination	7.4.1,7.4.2
	Species-pair discrimination only	7.4.1.1-7.4.1.16
5).	Overall discrimination only	8.3.2
	Overall and species-pair discrimination	8.3.1,8.3.2
	Species-pair discrimination only	8.3.1.1-8.3.1.10

Examining the statistics in each section and referring to Table 9.1 will give the characters to be measured for each analysis. In addition, it is recommended that wing tuft colour and abdominal setal colour also be recorded.

There are advantages and disadvantages to each of the approaches which could be taken to identifying adult females using these statistics. The principal advantage of using the overall 'global' statistics to the species-pair 'global' statistics in Chapter eight is that sample sizes are larger, and so estimates of the parameters more accurate than the regional estimates. However, unlike samples are pooled, which may give unpredictable results in allocation. The advantage of the regional statistics is that only flies native to a region were used to derive the statistics for allocation in that region, but the disadvantage is that smaller sample sizes were used with consequent loss of resolving power.

The species-pair statistics have the advantage that the statistics were developed for just that species pair, but therefore the estimate of the pooled dispersion matrix is based on a smaller sample size than for the overall statistics. Also, if the assumption implicit in using only species-pair statistics is violated, that flies in an area only belong to those species, then results might be unpredictable. This does not occur if the species-pair statistics are used subsequent to overall allocation, but then there is the disadvantage that more characters need to be measured if all species-pair combinations are accounted for.

This last problem can be alleviated if certain of the species-pair statistics are discounted prior to the analysis, for example, if all combinations in section 7.4 were covered, then 22 characters would need to be measured on each fly as well as the two colour characters. However, if those analyses with size-free canonical roots less than 1.0 are discounted, then the number of characters falls to 16, plus the two colour characters. Likewise, the 'global' statistics require

23 characters, plus the two colour characters, but if species-pair analyses with size-free canonical roots less than 1.0 are discounted then this number falls to 18, plus the two colour characters.

In practise, in the field, it is unlikely that all of the species-pair statistics will be required, and these are more likely to be used in a laboratory where more sophisticated computers will be able to access the relevant statistics in a straightforward manner not likely to be practicable with the programmable calculator or portable computer envisaged as the field tool for this method. It is more likely that the overall statistics, whether 'global' or regional, together with certain species-pair statistics known to be of use in that area will be the only statistics used in the field method, therefore reducing the number of characters which need to be measured.

The statistics developed in Chapters seven and eight can be modified to meet the particular requirements of an area. For example, if it is known with some certainty from other evidence that it is extremely unlikely that 'savanna' flies will be found in an area, then the prior probabilities of species membership can be adjusted accordingly. If the 'global' overall statistics of Chapter eight are used, then the priors might be set at

0.0, 0.25, 0.25, 0.25, 0.25

for the five species 'savanna', *S. sanctipauli*, *S. soubrense*, *S. squamosum*, *S. yahense*. If the information in the wing tuft colour is also required, then this can simply be achieved by first multiplying the new priors and the appropriate colour character priors, summing the intermediate result, and dividing each intermediate result by this sum. For example, the first row of Table 8.52 is,

0.5644, 0.0009, 0.032, 0.4026, 0.0

multiplying this row by the new set of priors gives the intermediate result,

0.0, 0.000225, 0.008, 0.10065, 0.0

which sums to 0.1089. Dividing each of these by the sum gives,

0.0, 0.002, 0.0735, 0.9245, 0.0

the new set of prior probabilities. This would be done for each row of this table (except for the final one where the result is identical).

It must be emphasised, however, that objective justifications for altering the prior probability of a species must be established before this step can be taken.

A user of the statistics developed in this project may decide, on objective grounds that certain of the characters derived in any one of the discriminant analyses are unnecessary. If this is so, then the corresponding row and column of the pooled dispersion matrix and the relevant row of the mean vectors can simply be deleted, and the LDF also recalculated.

9.6 GENERAL CONCLUSIONS

The methods developed in this project for the identification of adult females of the *S. damnosum* complex in West Africa will clearly be of benefit to studies aimed at determining the relative vectorial importance of the different species in transmitting *O. volvulus*. The very successful identification of the 'savanna' species *S. damnosum s.s.* and *S. sirbanum* from all other species is of most benefit, as these species transmit the more debilitating strain of onchocerciasis. The ability to identify *S. squamosum* and *S. yahense* individually provides a more practical field identification method than enzyme electrophoresis. But the proximity of *S. sanctipauli S. soubrense* and *S. yahense* means that unequivocal identification of these species cannot be established in the absence of abdominal setal colour, but the statistics presented in this project represent the best available method for distinguishing between them.

Of the two methods of allocation, forced allocation and typicality probability allocation, it is recommended that both be used simultaneously, as both can easily be written into the same computer program. Typicality probability allocation allows a more realistic assessment of a fly's affinities to be made, while forced allocation allows the inclusion of additional prior knowledge to influence the specific identity of a fly.

The multivariate statistical method for identification of adult females as it is presented in this project could very easily be applied as a field method, without requiring any laboratory facilities, and after very short training of field workers. This facility is a great improvement over other methods of adult identification, with the exception of current morphological methods.

There are still outstanding problems within the morphology and morphometrics of the *S. damnosum* complex which will need to be investigated in further detail, although the method presented in this project can still be used in their absence.

1). Wider geographic and seasonal variation needs to be sampled for each species in the *S. damnosum* complex, with the ultimate aim of providing discriminant statistics for specific geographic regions and climatic seasons.

2). More intensive sampling of chromosomally known samples of new cytoforms needs to occur, such as *S. sanctipauli* 'Djodji', the OP-insecticide resistant *S. sanctipauli* , and the various forms of *S. soubrense*. The morphological status of *S. dieguerense* needs to be established in light of recent data on more extensive geographic range of this species (Boakye and Mosha 1988).

A more sensitive system for scoring wing tuft colour should be adopted which more accurately represents the variation of this character.

4). The reliability of abdominal setal colour as a taxonomic character for *S. yahense* (Garms and Zillman 1984) needs to e established using a double-blind trial.

5). A continuously updated data base should be established on computer of measurements on flies of known chromosomal identity containing the characters used in this project, those previously used in morphological studies of the *S. damnosum* complex, and any new characters as they are discovered. This will provide more reliable estimates of parameters such as mean vectors and dispersion matrix, and also allow an objective assessment of the error rates of the different morphological methods. The question of equal dispersion will also be more easily established once larger sample sizes of each species are available.

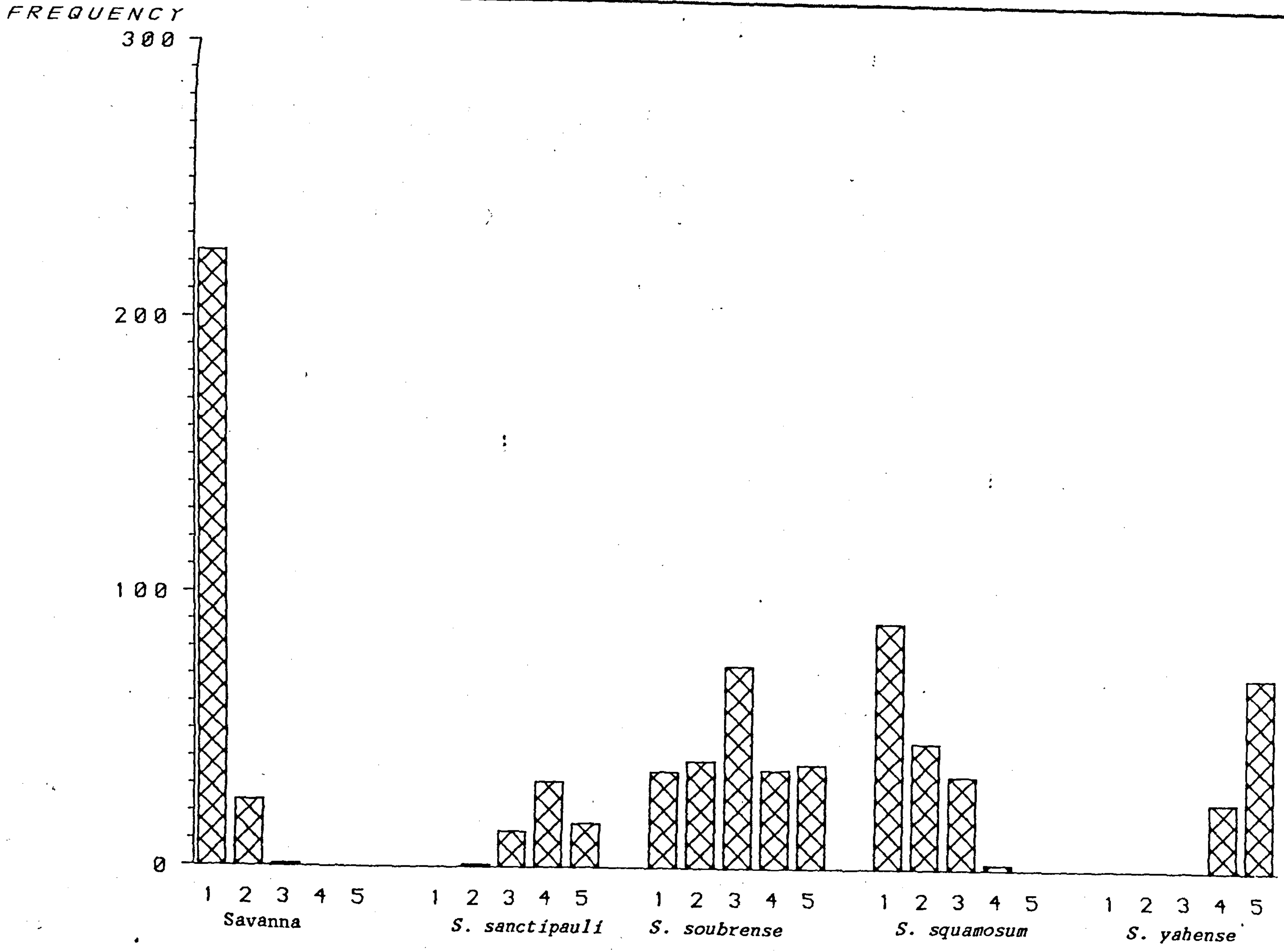


Figure 9.1
Histograms of Wing Tuft Colour for *S. damnosum s.l.*

References

Ambergen, A.W., Schaafsma, W. (1984). Interval estimates for posterior probabilities, applications to Border Cave. *In* van Vark, G.N., Howells, W.W. (eds.). *Multivariate Statistical Methods in Physical Anthropology*, pp. 115-134, Reidel: Dordrecht.

Anon., (1976). Species complexes in insect vectors of disease (blackflies, mosquitoes, tsetse flies). Report of a WHO informal consultation. *WHO/VBC/77.656*, Mimeogr. Doc.

Ashton, E.H., Healy, M.J.R., Oxnard, C.E., Spence, T.F. (1965). The combination of locomotor features of the primate shoulder girdle by canonical analysis. *J. Zool.*, 147, 406-429.

Atchley, W.R. (1971a). Analysis of geographic variation in the pupae of three species of *Culicoides* (Diptera: Ceratopogonidae). *Evolution*, 25, 51-74.

Atchley, W.R. (1971b). Study of the cause and significance of morphological variation in adults and pupae of *Culicoides*: A factor analysis and multiple regression study. *Evolution*, 25, 563-583.

Atchley, W.R. (1971c). Sexual dimorphism in *Chironomus* larvae (Diptera: Chironomidae).. *Am. Nat.*, 105, 455-466.

Atchley, W.R. (1973). Separation of the females of *Culicoides (Selfia) denningi*, *C. (S.) hieroglyphicus* and *C. (S.) jamesi* (Diptera: Ceratopogonidae). *J. Med. Ent.*, 10, 629-632.

Atchley, W.R. (1974). Morphometric differentiation in chromosomally characterised parapatric races of Morabine grasshoppers (Orthoptera: Eumastacidae). *Aust. J. Zool.*, 22, 25-37.

Atchley, W.R., Bryant, E.H. (1975) (eds.). *Multivariate Statistical Methods: Among-Groups Covariation.* Benchmark papers in systematic and evolutionary biology, 1. Halstead Press: Pennsylvania.

Atchley, W.R., Hensleigh, D.A. (1974). The congruence of morphometric shape in relation to genetic divergence in four races of Morabine grasshoppers (Orthoptera: Eumastacidae). *Evolution*, 28, 416-427.

Atchley, W.R., Martin, J. (1971). Morphometric analysis of differential sexual dimorphism in larvae of *Chironomus* (Diptera). *Can. Entom.*, 103, 319-327.

Atchley, W.R., Rutledge, J.J., Cowley, D.E. (1982). A multivariate statistical analysis of direct and correlated response to selection in the rat. *Evolution*, 36, 677-698.

Baker, R.J., Atchley, W.R., McDaniel, V.R. (1972). Karyology and morphometrics of Peters' Tent-Making bat, *Uroderma bilobatum* Peters (Chiroptera: Phyllostomatidae). *Systematic Zoology*, 21, 414-429.

Barnett, V., Lewis, T. (1984). *Outliers in Statistics.* Wiley: Chichester.

Bedo, D.G. (1976). Polytene chromosomes in pupal and adult black flies. *Chromosoma*, 57, 387-396.

Bedo, D.G. (1977). Cytogenetics and evolution of *Simulium ornatipes* Skuse (Diptera: Simuliidae) I. Sibling speciation. *Chromosoma*, 64, 37-65.

Beech-Garwood, P., Davies, J.B., McMahon, J.E. (in the press). A morphological technique to determine the presence of *Simulium squamosum* in mixed populations of the *S. damnosum* complex in Sierra Leone, West Africa. *Angew. Zool.*, (in the press).

Bird, J., Riska, B., Sokal, R.R. (1981). Geographic variation in variability of *Pemphigus populicaulus*. *Systematic Zoology*, 30, 58-70.

Blackith, R.E., Blackith, R.M. (1969). Variation of shape and of discrete anatomical characters in the Morabine grasshoppers. *Aust. J. Zool.*, 17, 697-718.

Blackith, R.E., Reyment, R.A. (1971). *Multivariate Morphometrics*. Academic Press: London.

Blacklock, D.B. (1926). The development of *Onchocerca volvulus* in *Simulium damnosum*. *Ann. Trop. Med. Parasit.*, 20, 1-48.

Blashfield, R.K. (1976). Questionnaire on cluster analysis software. *Classification Society Bulletin*, 3, 25-42.

Boakye, D.A., Mosha, F.W. (1988). The distribution and chromosome polymorphism of *Simulium dieguerense* (Diptera: Simuliidae). *Trop. Med. Parasit.*, 39, 117-119.

Brown, K.R. (1979). Multivariate assessment of phenetic relationships within the tribe Luciliini (Diptera: Calliphoridae). *Aust. J. Zool.*, 27, 465-477.

Brown, K.R., Shipp, E. (1977). Wing morphometrics of Australian Luciliini (Diptera: Calliphoridae). *Aust. J. Zool.*, 25, 765-777.

Brown, K.R., Shipp, E. (1978). Morphometric analysis of Australian Sarcophagidae (Diptera: Sarcophagidae). *Syst. Ent.*, 3, 179-188.

Bryant E.H., Turner, C.R. (1978). Morphometric adaptation of the housefly and the face fly in the United States. *Evolution*, 32, 759-770.

Campbell, N.A. (1978). Multivariate analysis in biological anthropology: Some further considerations. *J. Hum. Evol.*, 7, 197-203.

Campbell, N.A. (1980). Robust procedures in multivariate analysis. I, Robust covariance estimation. *Applied Statistics*, 29, 231-237.

Campbell, N.A. (1982). Robust procedures in multivariate analysis. II, Robust canonical variate analysis. *Applied Statistics*, 31, 1-8.

Campbell, N.A. (1984). Some aspects of allocation and discrimination. *In* van Vark, G.N., Howells, W.W. (eds.). *Multivariate Statistical Methods in Physical Anthropology*, pp. 177-192, Reidel: Dordrecht.

Campbell, N.A., Atchley, W.R. (1981). The geometry of canonical variate analysis. *Systematic Zoology*, 30, 268-280.

Campbell, N.A., Dearn, J.M. (1980). Altitudinal variation in, and morphological divergence between, three related species of grasshopper, *Praxibulus* sp., *Kosciuscola cognatus* and *K. usitatus* (Orthoptera: Acridae). *Aust. J. Zool.*, 28, 103-118.

Campbell, N.A., Kitchener, D.J. (1980). Morphological divergence in the genus *Eptesius* (Microchiroptera: Vespertilionidae) in Western Australia: a multivariate approach. *Aust. J. Zool.*, 28, 457-475.

Carlson, D.A., Service, M.W. (1979). Differentiation between species of the *Anopheles gambiae* complex (Diptera: Culicidae) by analysis of cuticular hydrocarbons. *Ann. Trop. Med. Parasit.*, 73, 589-592.

Carlson, D.A., Service, M.W. (1980). Identification of mosquitoes of *Anopheles gambiae* complex A and B by analysis of cuticular components. *Science*, 207, 1089-1091.

Carlson, D.A., Walsh, J.F. (1981). Identification of two West African black flies (Diptera: Simuliidae) of the *Simulium damnosum* species complex by analysis of cuticular paraffins. *Acta Tropica*, 38, 253-239.

Chang, W.C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 32, 267-275.

Cheke, R.A., Harris, J.R.W. (1980). Seasonal size variation in females of the *Simulium damnosum* complex in the Ivory Coast. *Trop. Med. Parasit.*, 31, 381-385.

Cheke, R.A., Denke, A.M. (1988). Anthropophily, zoophily and roles in onchocerciasis transmission of the Djodji form of *Simulium sanctipauli* and *S. squamosum* in a forest zone of Togo. *Trop. Med. Parasit.*, 39, 123-127.

Cheke, R.A., Garms, R. (1983). Reinfestation of the southeastern flank of the onchocerciasis control programme area by windborne vectors. *Phil. Trans. R. Soc. Lond.*, B 302, 471-484.

Cheke, R.A., Garms, R. (1986). Fecundities of different members of the *Simulium damnosum* complex in Togo. *Trans. R. Soc. Trop. Med. Hyg.*, 80, 489-490.

Cheke, R.A., Garms, R., Ouedraogo, J., Some, A., Sowah, S. (1987). The Beffa form of *Simulium soubrense* of the *S. damnosum* complex in Togo and Benin. *Med. Vet. Ent.*, 1, 29-35.

Corbet, G.B., Cummins, J., Hedges, S.R., Krzanowski, W. (1970). The taxonomic status of British water voles, genus *Arvicola J. Zool.*, 161, 301-316.

Corsten, L.C.A., Gabriel, K.R., (1976). Graphical exploration in comparing variance matrices. *Biometrics*, 32, 851-863.

Costanza, M.C., Afifi, A.A. (1979). Comparison of stopping rules in forward stepwise discriminant analysis. *J. Am. Stat. Assoc.*, 74, 777-785.

Crisp, G. (1956). Simulium *and onchocerciasis in the northern territories of the Gold Coast.* Lakeman: London.

Crosskey, R.W. (1957). Man-biting behaviour in *Simulium bovis* De Meillon in Northern Nigeria, and infection with developing filariae. *Ann. Trop. Med. Parasit.*, 51, 80-86.

Crosskey, R.W. (1960). A taxonomic study of the larvae of West African Simuliidae (Diptera: Simuliidae) with comments on the morphology of the larval blackfly head. *Bull. Br. Mus. nat. Hist.*(Ent.), 10,1 1-74.

Crosskey, R.W. (1973). Simuliidae. *In* Smith, K.V.G. (ed.), *Insects and Arthropods of Medical Importance*, pp. 109-153. British Museum (Natural History): London.

Crosskey, R.W. (1987a). An annotated checklist of the world blackflies (Diptera: Simuliidae). *In* Kim, K.C., Merrit, R.W. (eds.), *Blackflies. Ecology, Population Management and Annotated World List.*, pp. 425-520. Penn. State University.

Crosskey, R.W. (1987b). A taxa summary for the *Simulium damnosum* complex, with special reference to distribution outside the control areas of West Africa. *Ann. Trop. Med. Parasit.*, 81, 181-192.

Dahl, C., Wold, S., Nielsen, L.T. Nilson, C. (1984). A SIMCA pattern recognition study in taxonomy: Claw shape in mosquitoes (Culicidae, Insecta). *Systematic Zoology*, 33, 355-369.

Daley, H. (1985). Insect morphometrics. *Ann. Rev. Entom.*, 30, 415-438.

Dang, P.T., Peterson, P.V. (1980). Keys to the main species and species groups within the *Simulium damnosum* Theobald complex occurring in West Africa (Diptera: Simuliidae). *Tropenmed. Parasit.*. 31 117-120.

Davies, J.B., Thomson, M.C., Beech-Garwood, P. (1988). Morphological identification of *Simulium yahense* and *S. squamosum* in the south of Sierra Leone confirmed by enzyme electrophoresis. *Trans. Röy. Soc. Trop. Med. Hyg.*, 82.

Davies, R.G., Boryatinski, K.L. (1979). Character selection in relation to the numerical taxonomy of some male Diapsidae (Homoptera: Coccoidea). *Biol. J. Linn. Soc.*, 12, 95-165.

Davis, G.M. (1983). Relative roles of molecular genetics, anatomy, morphometrics and ecology in assessing relationships among North American Unionidae (Bivalvia). *In* Oxford, G.S., Rollinson, D. (eds.), *Protein polymorphisms: adaptive and taxonomic significance.*, pp. 193-222. Systematics Association Special Volume 24. Academic Press: London.

Duke, B.O.L., Lewis, D.J., Moore, P.J. (1966). *Onchocerca-Simulium* complexes I. Transmission of forest and Sudan-savanna strains of *Onchocerca volvulus*, from Cameroon, by *Simulium damnosum* from various West African bioclimatic zones. *Ann. Trop. Med. Parasit.*, 60, 318-336.

Dunbar, R.W. (1959). The salivary gland chromosomes of seven forms of blackflies included in *Eusimulium aureum* Fries. *Can. J. Zool.*, 37, 597-599.

Dunbar, R.W. (1966). Four sibling species included in *Simulium damnosum* Theobald (Diptera: Simuliidae) from Uganda. *Nature*, 209, 597-599.

Dunbar, R.W. (1969). Nine cytological segregates in the *Simulium damnosum* complex (Diptera: Simuliidae). *Bull. Wld. Hlth. Org.*, 40, 974-979.

Dunbar, R.W., Vajime, C.G. (1971). Cytotaxonomic analysis of the *Simulium damnosum* complex. *WHO/VBC/7.320*, mimeographed document.

Dunbar, R.W., Vajime, C.G. (1972). The *Simulium (Edwardsellum) damnosum* complex. A report on cytotaxonomic studies to April 1972. *WHO/ONCHO/72.100*, mimeographed document.

Dunbar, R.W., Vajime, C.G. (1981). Cytotaxonomy of the *Simulium damnosum* complex. *In* Laird M. (ed.), *Blackflies, the Future for Biological Methods in Integrated Control.* pp. 31-43. Academic Press: London.

DuPraw, E.J. (1965). Non-Linnean taxonomy and the systematics of honeybees. *Systematic Zoology*, 14, 1-24.

Endler, J.A. (1977). *Geographic Variation, Speciation, and Clines.* Princeton University Press: Princeton.

Fveritt, B.S. (1978). *Graphical Techniques for Multivariate Data.* Heinemann: London.

Everitt, B.S. (1980). *Cluster Analysis.* Heinemann: London.

Ferson, S., Rohlf, F.J., Koehn, R.K. (1985). Measuring shape variation in two dimensional outlines. *Systematic Zoology*, 34, 59-68.

Garms, R. (1978). Use of morphological characters in the study of *Simulium damnosum s.l.* populations in West Africa. *Tropenmed. Parasit.* 29, 483-491.

Garms, R. (1983). Studies of the transmission of *Onchocerca volvulus* by species of the *Simulium damnosum* complex occurring in Liberia. *Z. Angew. Zool.* 70, 101-117.

Garms R., Cheke, R.A., Vajime, C.G., Sowah, S. (1982). The occurrence and movements of different members of the *Simulium damnosum* complex in Togo and Benin. *Z. Angew. Zool.* 69, 219-236.

Garms R., Cheke, R.A., (1985). Infections with *Onchocerca volvulus* in different members of the *Simulium damnosum* complex in Togo and Benin. *Z. Angew. Zool.* 72, 479-495.

Garms R., Vajime, C.G. (1975). On the ecology of the species of the *Simulium damnosum* complex in different bioclimatic zones of Liberia and Guinea. *Tropenmed. Parasit.*, 26, 375-380.

Garms R., Walsh, J.F., Davies, J.B. (1979). Studies on the reinvasion of the onchocerciasis control programme in the Volta river basin by *S. damnosum s.l.* with emphasis on the south-western area. *Tropenmed. Parasit.*, 30, 345-326.

Garms R., Walsh, J.F. (1987). The migration and dispersal of blackflies: *Simulium damnosum s.l.*, the main vector of human onchocerciasis. *In* Kim, K.C., Merrit, R.W. (eds.), *Blackflies, Ecology, Population Management and Annotated World List.* pp.201-214. Penn. State Universiy.

Garms R., Zillman, U. (1984). Morphological identification of *Simulium sanctipauli* and *S. yahense* in Liberia and comparison of results with those of enzyme electrophoresis. *Tropenmed. Parasit.*, 35, 217-220.

Geisser, S. (1977). Discrimination, allocatory and separatory, linear aspects. *In* van Ryzin, J. (ed.), *Classification and Clustering*, pp 301-330. Academic Press: New York.

Gibbins, E.G. (1933). Ethiopean Simuliidae. *Simulium damnosum*, Theo.. *Trans r. Ent. Soc. Lond.*, 81, 37-51.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley: New York.

Gnanadesikan, R., Kettenring, J.R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics*, 28, 81-124.

Gordon, A.D. (1981). *Classification- Methods for the Exploratory Analysis of Multivariate Data*. Chapman and Hall: London.

Gordon, A.D. (1987). A review of hierarchical classification. *J. R. Statist. Soc.* A, 150, 119-137.

Gould, S.J. (1984). Covariance sets and ordered geographic variation in *Cerion* from Aruba, Bonaire and Curacao: A way of studying non-adaptation. *Systematic Zoology*, 33, 217-237.

Gould, S.J., Woodruff, D.S., Martin, J.P. (1975). Genetics and morphometrics of *Cerion* at Pongo Carpet: A new systematic approach to this enigmatic land snail. *Systematic Zoology*, 23, 518-528.

Gower, J.C., Digby, P.G.N. (1981). Expressing complex relationships in two dimensions. *In* Barnett, V. (ed.), *Interpreting Multivariate Data* , pp 83-118. Wiley, Chichester.

Gower, J.C., Ross, G.J.S. (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18, 54-64.

Grenier, P., Ovazza, M. (1951). Simuliidae de Moyen Congo. *Bull. Soc. Path. Exot.*, 14, 222-234.

Grenier, P., Ovazza, M., Valade, M. (1960). Notes biologique et faunistiques sur *Simulium damnosum* et les Simuliidae d'Afrique Occidentale (Haute-Volta, Côte d'Ivoire, Dahomey, Soudan). *Bull. Inst. fr. Afr. noire*, A, 14, 892-918.

Habemma, J.D.F., Hermans, J. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. *Technometrics*, 19, 487-493.

Hawkins, D.M. (1981). A new test for multivariate normality and homoscedacity. *Technometrics*, 23, 105-110.

Hensleigh, D.A., Atchley, W.R. (1977). Morphometric variability in natural and laboratory populations of *Culicoides variipennis* (Diptera: Ceratopogonidae). *J. Med. Ent.*, 14 379-386.

Hermans, J., Habbema, J.D.F., Schater, J.R. (1982). The ALLOC80 package for discriminant analysis. *Stat. Software Newsl.*, 8, 15-20.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24 417-441.

Huber, P.J. (1981). *Robust Statistics*. Wiley: New York.

Jardine, N. (1971). Patterns of local differentiation between human local populations. *Phil. Trans. Roy. Soc. Lond.*, B, 293, 1-33.

Jardine, N., Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. *Computer Journal*, 11, 117-184.

Jardine, N., Sibson R. (1971). *Mathematical Taxonomy*. Wiley: London.

Jeffers J.N.R. (1967). Two case studies in the application of principal component analysis. *Applied Statistics*, 16, 25-36.

Johnston, R.F., Selander, R.K. (1971). Evolution in the House Sparrow. II. Adaptive differentiation in North American populations. *Evolution*, 25, 1-28.

Jolicoeur, P. (1959). Multivariate geographical variation in the Wolf *Canis lupus* L.. *Evolution*, 13, 283-299.

Jolicoeur, P., Mossiman, J.E. (1960). Size and shape variation in the painted turtle, a principal component analysis. *Growth*, 24, 339-354.

Kruskal, J.B. (1964). Multidimensional scaling by optimising goodness of fit to a non-metric hypothesis. *Psychometrica*, 29, 1-27.

Kruskal, J.B. (1977). The relationship between multidimensional scaling and clustering. *In* van Ryzin, J. (ed.), *Classification and Clustering* , pp 17-44. Academic Press: New York.

Kurtak, D.C., Raybould, J.N., Vajime, C. (1981). Wing tuft colours in the progeny of single individuals of *Simulium squamosum* (Enderlein). *Trans. Roy. Soc. Trop. Med. Hyg.*, 75, 126.

Lachenbruch, P.A. (1975). *Discriminant Analysis.* Hafner Press: New York.

Lachenbruch, P.A., Goldstein, M. (1979). Discriminant analysis. *Biometrics*, 35, 69-85.

Lachenbruch, P.A., Mickey, M.R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10, 1-11.

Lande, R. (1977). On comparing coefficients of variation. *Systematic Zoology*, 26, 214-217.

Lane, R.P. (1981). A quantitative analysis of wing pattern in the *Culicoides pulicaris* species group (Diptera: Ceratopogonidae). *Zoo. J. Linn. Soc.*, 72, 21-41.

Lane, R.P., Ready, P.D. (1985). Multivariate discrimination between *Lutzomyia wellcomei*, a vector of mucocutaneous leishmaniasis and *Lu. complexus* (Diptera: Phlebotominae). *Ann. Trop. Med. Parasit.*, 79, 469-472.

Lawes Agricultural Trust, (1984). *GENSTAT: A general Statistical Program.* Rothamsted Experimental Station.

Layard, M.W.J. (1974). A Monte Carlo comparison of tests for equality of covariance matrices. *Biometrika*, 61, 461-465.

Lessios, H.A. (1981). Divergence in allopatry: Molecular and morphological differentiation between sea urchins separated by the isthmus of Panama. *Evolution*, 35, 618-634.

Levins, R. (1965). Theory of fitness in a heterogeneous environment, V. Optimal genetic systems. *Genetics*, 52, 891-904.

Lewis, D.J. (1960). Observations on *Simulium damnosum* in the Southern Cameroons and Liberia. *Ann Trop. Med. Parasit.*, 54, 208-223.

Lewis, D.J., Duke, B.O.L. (1966). *Onchocerca-Simulium* complexes. II, Variation in West African female *S. damnosum*. *Ann Trop. Med. Parasit.*, 60, 337-346.

Lewis, D.J., Lyons, G.R.L, Marr, J.D.M (1961). Observations on *Simulium damnosum* from the Red Volta in Ghana. *Ann. Trop. Med. Parasit.*, 55, 202-210.

Lindenfelser, M.E. (1984). Allozymic congruence: evolution in the prawn *Macrobrachium rosenbergii* (Decapoda: Palaemonidae). *Systematic Zoology*, 33, 195-204.

Lubischew, A.A. (1962). On the use of discriminant functions in taxonomy. *Biometrics*, 18, 455-477.

Macgregor, H.C., Varley, J.M. (1983). *Working with Animal Chromosomes.* Wiley: Chichester.

McCrae, A.W.R. (1965). Recent studies on non-anthropophilic *Simulium damnosum* Theo. in Uganda. Proceedings of the 12th International Congress of Entomology, London, 1964, pp 823-824.

McCrae, A.W.R. (1967). The *Simulium damnosum* species complex. *Rep. E. Afr. Virus. Res. Inst.*, 17, 67-70.

McKay, R.J., Campbell, N.A. (1982a). Variable selection techniques in discriminant analysis, I, Description. *Br. J. Math. Stat. Psychol.*, 35, 1-29.

McKay, R.J., Campbell, N.A. (1982b). Variable selection techniques in discriminant analysis, II, Allocation. *Br. J. Math. Stat. Psychol.*, 35, 30-41.

Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India*, 2, 49-55.

Mahalanobis, P.C., Majumdar, D.N., Rao, C.R. (1949). Anthropometric survey of the United Provinces, 1941; a statistical survey. *Sankhya*, 9, 89-324.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57, 519-520.

Mardia, K.V., Kent, J.T., Bibby, J.M. (1979). *Multivariate Analysis*. Academic Press: London.

Marr, J.D.M., Lewis, D.J. (1963). Colour variation in *Simulium damnosum*. *Trans. R. Soc. Trop. Med. Hyg.*, 57, 7.

Marr, J.D.M., Lewis, D.J. (1964). Observations on the dry-season survival of *Simulium damnosum* Theo. in Ghana. *Bull. Ent. Res.*, 55, 547-564.

Matthews, J.N.S. (1984). Robust methods in the assessment of multivariate normality. *Applied Statistics*, 33, 272-277.

Mayr, E. (1970). *Populations, Species and Evolution*. Harvard University Press: Cambridge, Mass..

Mebrahtu, Y., Beach, R.F., Khamala, C.P.M., Hendricks, L.D. (1984). Characterisation of *Simulium (Edwardsellum) damnosum s.l.* from six river systems in Kenya by cellulose acetate electrophoresis. *Trans. Roy. Soc. Trop. Med. Parasit.*, 80, 914-922.

Meredith, S.E.O. (1982). Enzyme identification of *Simulium damnosum s.l.* caught biting man. *Ann. Trop. Med. Parasit.*, 76, 375-376.

Meredith, S.E.O., Cheke, R.A., Garms, R. (1983). Variation and distribution of *Simulium soubrense* and *Simulium sanctipauli* in West Africa. *Ann. Trop. Med. Parasit.*, 77, 627-640.

Meredith, S.E.O., Townson, H. (1981). Enzymes for species identification in the *Simulium damnosum* complex from West Africa. *Tropenmed. Parasit.*, 32, 123-129.

Milligan, G.W., Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, 159-179.

Mosteller, F., Tukey, J.W. (1977). *Data Analysis and Regression.* Addison-Wesley: Reading, Mass..

Nei, M. (1987). *Molecular Evolutionary Genetics.* Columbia University Press: New York.

Oxnard, C.E. (1984). Testing in multivariate statistical approaches to physical anthropology: The example of sexual dimorphism in the primates. *In* van Vark, G.N., Howells, W.W. (eds.). *Multivariate Statistical Methods in Physical Anthropology*, pp. 193-227, Reidel: Dordrecht.

Pearson, K. (1901). On the lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2, 559-572.

Pearson, K. (1926). On the coefficient of racial likeness. *Biometrika*, 18, 105-120.

Peterson, B.V., Dang, P.T. (1981). Morphological means of separating siblings of the *Simulium damnosum* complex (Diptera: Simuliidae). *In* Laird M. (ed.), *Blackflies, the Future for Biological Methods in Integrated Control*. pp. 45-56. Academic Press: London.

Phillipon, B. (1987). Problems in epidemiology and control of West African onchocerciasis. *In* Kim, K.C., Merrit, R.W. (eds.), *Blackflies, Ecology, Population Management and Annotated World List*. pp.363-373. Penn. State University.

Phillips, A., Walsh, J.F., Garms, R., Molyneux, D.H., Milligan, P., Ibrahim, G. (1985). Identification of adults of the *Simulium damnosum* complex using hydrocarbon analysis. *Trop. Med. Parasit.*, 36, 97-101.

Plowright, R.C., Stephen, W.P. (1973). A numerical taxonomic analysis of the evolutionary relationships of *Bombus* and *Psithyrus* (Apidae: Hymenoptera). *Can Ent.*, 105, 733-743.

Post, R.J. (1982). Sex-linked inversions in blackflies (Diptera: Simuliidae). *Heredity*, 48, 85-93.

Post, R.J. (1985). DNA probes for vector identification. *Parasitology Today*, 1, 89-90.

Post, R.J. (1986). The cytotaxonomy of *Simulium sanctipauli* and *Simulium soubrense* (Diptera: Simuliidae). *Genetica*, 69, 191-207.

Post, R.J., Crosskey, R.W. (1985). The distribution of the *Simulium damnosum* complex in Sierra Leone and its relation to onchocerciasis. *Ann. Trop. Med. Parasit.*, 79, 169-194.

Post, R.J., Kurtak, D. (1987). Identity of the OP-insecticide resistant species in the *Simulium sanctipauli* subcomplex. *Ann. Soc. Belge. Med Trop.*, 67, 71-73.

Procunier, W.S., Post, R.J. (1986). Development of a method for the cytological identification of man- biting sibling species within the *Simulium damnosum* complex. *Trop. Med. Parasit.*, 37, 49-53.

Puri, I.M. (1925). On the life history and structure of the early stages of Simuliidae (Diptera, Nematocera). *Parasitology*, 17, 295-334.

Quillévére, D. (1975). Etude du complexe *Simulium damnosum* en Afrique de l'Ouest. I. Techniques d'etude. Identification des cytotypes *Cah. O.R.S.T.O.M., Ser. Ent. Med. Parasit.*, 13, 85-98.

Quillévére, D. (1979). Contribution a l'etude des characteristiques taxonomiques, bioecologiques et vetrices ces membres du complexe *Simulium damnosum* present en Cote d'Ivoire. *Travaux et Documents de l'O.R.S.T.O.M.*, 109.

Quillévéré, D., Pendriez, B. (1975). Etude du complexe *Simulium damnosum* en Afrique de l'Ouest. II. Reparation geographique de cytotypes en Cote d'Ivoire *Cah. O.R.S.T.O.M., Ser. Ent. Med. Parasit.*, 13, 165-172.

Quillévére, D., Sechan, Y., Pendriez, B. (1977). Etude du complexe *Simulium damnosum* en Afrique de l'Ouest. V. Identification morphologique des femelles en Cote d'Ivoire *Tropenmed. Parasit.*, 28, 244-253.

Quillévéré, D., Sechan, Y. (1978). Morphological identification of females of the *Simulium damnosum* complex in West Africa: differentiation of *S. squamosum* and *S. yahense*. *Trans Roy. Soc. Trop. Med. Hyg.*, 72, 99-100.

Quillevere, D., Guillet, P. (1982). La reparation geographique des especes du complexe *Simulium damnosum* dans la zone du project Senegambiae. *Cah. O.R.S.T.O.M., Ser. Ent. Med. Parasit.*, 22, 303-311.

Rao, C.R. (1948). The utilisation of multiple measurements in problems of biological classification (with discussion). *J. Roy. Stat. Soc.*, B, 10, 159-203.

Rao, C.R. (1964). The use and interpretation of principal components in applied research. *Sankhya*, 26, 329-358.

Raybould, J.N., Vajime, C.G., Quillevere, D., Barro, T., Sawadogo, R. (1979). The laboratory maintenance of *Simulium damnosum* complex species as a research tool for the onchocerciasis control programme in the Volta river basin. *Tropenmed. Parasit.*, 30, 499-504.

Reyment, R.A. (1961). A note on geographic variation in European *Rana*. *Growth*, 25, 219-227.

Reyment, R.A. (1962). Observations on homogeneity of covariance matrices in paleontologic biometry. *Biometrics*, 18, 1-11.

Reyment, R.A. (1982). Evolution in a Cretaceous foraminifer. *Evolution*, 36, 1182-1199.

Ribiero, H. (1980). A biometric study of the taxonomy of the *Anopheles gambiae* Giles complex (Diptera, Culicidae). *Garcia de Orta, Ser. Zoo.*, 9, 139-154.

Rising, J.D. (1970). Morphological variation and evolution in some North American orioles. *Systematic Zoology*, 19, 315-351.

Riska, B. (1981). Morphological variation in the horseshoe crab *Limulus polyphemus Evolution*, 35, 647-658.

Rohlf, F.J. (1963). Congruence of larval and adult classifications in *Aedes* (Diptera: Culicidae). *Systematic Zoology*, 12, 97-117.

Rohlf, F.J. (1985). *NT-SYS, Numerical Taxonomy System of Multivariate Statistical Programs.*

Rohlf, F.J., Archie, J.W. (1984). A comparison of Fourier methods for the description of wing shape in mosquitoes (Diptera: Culicidae). *Systematic Zoology*, 33, 302-317.

Rohlf, F.J., Sokal, R.R. (1972). Comparative morphometrics by factor analysis in two species of Diptera. *Z. Morph. Tiere.*, 72, 36-45.

Rothfels, K.H. (1956). Blackflies: Siblings, sex and species groupings. *J. Heredity*, 47, 113-122.

Rothfels, K.H. (1979). Cytotaxonomy of black flies (Simuliidae). *Annu. Rev. Entom.*, 24, 507-539.

Rothfels, K.H. (1987). Cytological approaches to black fly taxonomy. *In* Kim, K.C., Merrit, R.W. (eds.), *Blackflies, Ecology, Population Management and Annotated World List.* pp.39-52. Penn. State University.

Rothfels, K.H., Nambiar, R. (1981). A cytological study of natural hybrids between *Prosimulium multidentatum* and *P. magnum*. *Chromosoma*, 82, 673-691.

Roy, S.G., Mukherjee, G.D. (1964). Use of Mahalanobis D^2-statistic in phase-discrimination in desert locust males. *Sankhya*, 26, 237-252.

SAS Institute (1984). *SAS User's Guide*, version 5, Cary: NC.

Schnell, G.D. (1970). A phenetic study of the sub-order Lari (Aves). II. Phenograms, discussion and conclusions. *Systematic Zoology*, 19, 264-302.

Seber, G.A.F. (1984). *Multivariate Observations.* Wiley: New York.

Simon, C. (1983). Morphological differentiation in wing venation among broods of 13- and 17-year periodical cicadas. *Evolution*, 37, 104-115.

Sneath, P.H., Sokal.R.R. (1973). *Numerical Taxonomy: the Principles and Practices of Numerical Classification.* Freeman: San Francisco.

Sokal, R.R., Rohlf, F.J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11, 33-40.

Sokal, R.R., Thomas, P.A. (1965). Geographic variation of *Pemphigus populi-transversus* in eastern North America: stem mothers and new data on alates. *Univ. Kansas Sci. Bull.*, 46, 201-252.

Soponis, A.R., Peterson, B.V. (1976). A preliminary investigation of some morphological characters in adult females of the *Simulium damnosum* complex from Togo. *WHO/VBC/SC/76*, Mimeogr. Doc.

SPSS inc. (1985). *SPSSX User's Guide.* McGraw-Hill: New York.

Surtees, D.P., Fiasorgbor, G., Post, R.J., Weber, E.A. (1988). The cytotaxonomy of the Djodji form of *Simulium sanctipauli* (Diptera: Simuliidae). *Trop. Med. Parasit.*, 39, 120-122.

Theobald, F.V. (1903). Report on a collection of mosquitoes and other flies from equatorial East Africa and the Nile provinces of Uganda. *Roy. Soc. Rept. Sleeping Sickness*, 3, 40.

Thorpe, R.S. (1976). Biometric analysis of geographic variation and racial affinities. *Biol. Rev.*, 51, 407-452.

Thorpe, R.S. (1980). A comparative study of ordination techniques in numerical taxonomy in relation to racial variation in the ringed snake *Natrix natrix* (L.). *Biol. J. Linn. Soc.*, 13, 7-40.

Thorpe, R.S., Leamy, L. (1983). Morphometric studies in inbred and hybrid house mice (*Mus* sp.): Multivariate analysis of size and shape. *J. Zool.*, 199, 421-432.

Thomson, M.C., Davies, J.B., Bockarie, M.J. (1988). Identification of species of the *Simulium damnosum* complex in Sierra Leone by enzyme electrophoresis and morphology. *Trop. Med. Parasit.*, 39, 84-85.

Titmus, G., Babcock, R.M. (1981). A morphometric study of the effects of a mermethid parasite on its host, *Einfeldia dissidens* (Walker) (Diptera). *Z. Pararsit.*, 65, 353-357.

Townson. H. Meredith, S.E.O. (1979). Identification of the Simuliidae in relation to onchocerciasis. *In* Taylor, A.E.R, Muller, R. (eds.), *Problems in the Identification of Parasites and their Vectors.* pp.145-174. 17&su./th/ Symposium of the British Society of Parasitology. Blackwell: Oxford.

Townson, H., Post, R.J., Philips, A. (1987). Biochemical approaches to blackfly taxonomy. *In* Kim, K.C., Merrit, R.W. (eds.), *Blackflies, Ecology, Population Management and Annotated World List.* pp.24-38. Penn. State University.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley: Reading, Mass..

Urbakh, U.Y. (1971). Linear discriminant analysis: loss of discriminating power when a variate is omitted. *Biometrics*, 27, 531-534.

Vajime, C.G. (1984). The population structure of *Simulium sirbanum* (Diptera: Simuliidae). XI International Congress for Tropical Medicine and Malaria, Calgary. Abstract and Poster Volume, 159.

Vajime, C.G., Dunbar, R.W. (1975). Chromosomal identification of eight species of the subgenus *Edwardsellum* near and including *Simulium (Edwardsellum) damnosum* Theobald (Diptera: Simuliidae). *Tropenmed. Parasit.*, 26, 111-138.

Vajime, C.G., Quillévéré, D. (1978). The distribution of the *Simulium damnosum* complex in West Africa with particular reference to the onchocerciasis control programme area. *Tropenmed. Parasit.*, 29, 473-482.

Van Ness, J., Simpson, C. (1976). On the effects of dimension in discriminant analysis. *Technometrics*, 8, 175-187.

van Vark, G.N. (1984). On the determination of hominid affinities. *In* van Vark, G.N., Howells, W.W. (eds.). *Multivariate Statistical Methods in Physical Anthropology*, pp. 323-349, Reidel: Dordrecht.

van Vark, G.N., Howells, W.W.(eds.)(1984).*Multivariate Statistical Methods in Physical Anthropology*, Reidel: Dordrecht.

Walsh, J.F., Davies, J.B., Garms, R. (1981). Further studies on the reinvasion of the onchocerciasis control programme by *Simulium damnosum s.l.*: The effects of an extension of control activities into southern Ivory Coast during 1979. *Tropenmed. Parasit.*, 32, 269-273.

Walsh, J.F., Philippon, B., Henderickx, J.E.E., Kurtak, D.C. (1987). Entomological aspects and results of the onchocerciasis control programme. *Tropenmed. Parasit.*, 38, 57-60.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Ass.*, 58, 236-244.

Weldon, W.F.R. (1893). On certain correlated variation in *Carcinus moenas*. *Proc. Roy. Soc. Lond.*, 54, 318-329.

WHO Technical Report Series, No. 597, 1976 (*Epidemiology of onchocerciasis*: report of a WHO Expert Committee).

WHO Technical Report Series, No. 752, 1987 (Third report of the WHO Expert Committee on Onchocerciasis).

Wirtz, C., Raybould, J.N. (1986). Artificial feeding of West African *Simulium damnosum* Theobald s.l. (Diptera: Simuliidae). through membranes and their subsequent fecundity. *Tropenmed. Parasit.*, 32, 269-273.

Wishart, D. (1978). *CLUSTAN User Manual*, 3rd edition. Report No. 47, Program Library Unit, Edinburgh University: Edinburgh.

Young, F.W., Takahane, Y., Lewyckyj, R. (1980). ALSCAL: a multidimensional scaling package with several individual differences options. *Am. Stat.*, 34, 117-118.

APPENDIX ONE: DETAILS OF ADULT FEMALE SAMPLES USED FOR MULTIVARIATE MORPHOMETRICS

Sample V1=1.

River Amou at Amou Oblo, Togo, 7° 24'N 0° 53'E, collected by Dr. R.J. Post, 19/10/1984, reared from pupae, preserved in 95% propanol.

Species identity, *S. squamosum*, from correlated larval cytotaxonomy, 36 *S. squamosum*, determined by D.P. Surtees.

Initial sample size = 40

Number rejected because of missing values = 2

Number rejected as outliers = 4

Final sample size = 34

Sample V1=2.

River Nanie at Nigbi, Côte d'Ivoire, 5° 38'N 6° 38'W, collected by Dr. R.J. Post, 10/07/1985, reared from pupae, preserved in 95% propanol.

Species identity, *S. yahense*, from correlated larval cytotaxonomy, 109 *S. yahense*, (determined by D.P. Surtees).

Initial sample size = 41

Number rejected because of missing values = 2

Number rejected as outliers = 2

Final sample size = 37

Sample V1=3.

River Niger at Tienfala, Mali, 12° 43'N 7° 44'W, collected by Dr. R.J. Post, 02/11/1984, reared from pupae, preserved in 95% propanol.

Species identity, *S. sirbanum*, from correlated larval cytotaxonomy, 7 *S. sirbanum*, determined by D.P. Surtees.

Initial sample size = 35

Number rejected because of missing values = 2

Number rejected as outliers = 4

Final sample size = 29

Sample V1=4.

River Sassandra at Soubre, Côte d'Ivoire, 5° 47'N 6° 37'W, collected by Dr. R.J. Post, 26/10/1984, reared from pupae, preserved in 95% propanol.

Species identity, *S. sanctipauli*, from correlated larval cytotaxonomy, 29 *S. sanctipauli*, determined by D.P. Surtees.

Initial sample size = 40

Number rejected because of missing values = 4

Number rejected as outliers = 1

Final sample size = 35

Sample V1=5.

River Gban-Houa at Djodji, Togo, 7° 42'N 0° 35'E, collected by Dr. R.J. Post, 15/10/1984, reared from pupae, preserved in 95% propanol.

Species identity, *S. sanctipauli* 'Djodji'/*S. squamosum*, from correlated larval cytotaxonomy, 23 *S. sanctipauli* 'Djodji': 11 *S. squamosum*, determined by D.P. Surtees. Mixture separated by external analysis, using LDFs derived from samples V1=1 and V1=4, and internal analysis using PCA and cluster analysis.

Initial sample size = 50

Number rejected because of missing values = 4

Number rejected as outliers (following separation of mixture) = 1

Final sample size, V1=1 (*S. sanctipauli* 'Djodji') = 26 Final sample size, V1=30 (*S. squamosum*) = 17

Sample V1=6.

River Bankasoka at Port Loko, Sierra Leone, 8° 45'N 12° 47'W, collected by MRC laboratory staff, Bo, 25/01/1986, reared from pupae, preserved in 95% propanol.

Species identity, *S. soubrense* 'B', from previous chromosomal identification (Post 1986).

Initial sample size = 37

Number rejected because of missing values = 7

Number rejected as outliers = 2

Final sample size = 28

Sample V1=7.

River Waanje at Kenema Waterfall, Sierra Leone, 7° 54'N 11° 14'W, collected by Dr. R.J. Post, 21/07/1985, reared from pupae, preserved in 95% propanol.

Species identity, *S. yahense*, from correlated larval cytotaxonomy, 38 *S. yahense*, determined by D.P. Surtees.

Initial sample size = 20

Number rejected because of missing values = 4

Number rejected as outliers = 2

Final sample size = 14

Sample V1=8.

River Bebeye at Gerihun, Sierra Leone, 7° 57'N 11° 35'W, collected by D.P. Surtees, Dr. J.B. Davies, M.C. Thomson, 20/01/1986, reared from pupae, preserved in 95% propanol.

Species identity, *S. yahense*, from previous chromosomal identifications (Post 1986).

Initial sample size = 40

Number rejected because of missing values = 12

Number rejected as outliers = 3

Final sample size = 25

Sample V1=9.

River Taia at Mongeri, Sierra Leone, 8° 19'N 11° 44'W, collected by D.P. Surtees, Dr. J.B. Davies, M.C. Thomson, 23/01/1986, reared from pupae, preserved in 95% propanol.

Species identity unknown.

Initial sample size = 7

Sample not used in subsequent analyses.

Sample V1=10.

River Seli at Yirafilaia, Sierra Leone, 9° 28'N 11° 20'W, collected by D.P. Surtees, Dr. J.B. Davies, M.C. Thomson, 09/02/1986, reared from pupae, preserved in 95% propanol.

Species identity, *S. damnosum s.s.*:(*S. squamosum*/*S. yahense*) , from previous larval cytotaxonomic identifications (4 *S. damnosum s.s.*, 7 *S. squamosum*) determined by D.P. Surtees. Also, correlated DNA probes identifications of same adults (Post pers. comm.), 27 *S. damnosum s.s.*, 6 *S. squamosum*/*S. yahense*, 3 unknown.

Initial sample size = 40

Number rejected because of missing values = 4

Number rejected as outliers/ contaminants = 8

Final sample size = 28

Sample V1=11.

River Kakatemadaru, Benin, 10° 07'N 03° 20'E, collected by Dr. R.A. Cheke, 10/09/1984, caught at human bait, preserved in 70% ethanol.

Species identity, *S. damnosum s.s.*, Dr. R.A. Cheke personal communication.

Initial sample size = 20

Number rejected because of missing values = 4

Number rejected as outliers = 0

Final sample size = 16

Sample V1=12.

River Amou at Amou Oblo, Togo, 7° 24'N 0° 53'E, collected by Dr. R.A. Cheke, 14/03/1985, reared from pupae, preserved in 95% propanol.

Species identity, *S. squamosum*, from correlated larval cytotaxonomy, 40 *S. squamosum*determined by D.P. Surtees.

Initial sample size = 18

Number rejected because of missing values = 0

Number rejected as outliers = 4

Final sample size = 14

Sample V1=13.

River Amoutchou at Idifiou, Togo, 7° 38'N 0° 58'E, collected by Dr. R.A. Cheke, A.M. Denke, 09/10/1985, reared from pupae, preserved in 95% propanol.

Species identity, *S. squamosum*, from correlated larval cytotaxonomy, 30 *S. squamosum* determined by D.P. Surtees.

Initial sample size = 29

Number rejected because of missing values = 6

Number rejected as outliers = 3

Final sample size = 20

Sample V1=14.

River Mono at T52, Togo, 6° 54'N 01° 36'E, collected by Dr. R.A. Cheke, 02/11/1981, caught bitong on man, preserved in 70% ethanol.

Species identity, *S. soubrense* 'Beffa', Dr. R.A. Cheke personal communication.

Initial sample size = 32

Number rejected because of missing values = 5

Number rejected as outliers = 2

Final sample size = 25

Sample V1=15.

River Baoule at Wandadou, Guinea, 9° 04'N 09° 20'W, collected by Dr. R. Baker, Sept. 1986, caught at human bait, preserved in 95% propanol.

Species identity, *S. squamosum*, Dr. R. Baker personal communication. Also, correlated DNA probes identifications, 43 *S. squamosum*, Dr. R.J. Post personal communication.

Initial sample size = 45

Number rejected because of missing values = 1

Number rejected as outliers = 4

Final sample size = 40

Sample V1=16.

River Milo at Balan, Guinea, 9° 46'N 09° 10'W, collected by Dr. R. Baker, Sept. 1986, caught at human bait, preserved in 95% propanol.

Species identity, *S. soubrense*, Dr. R. Baker personal communication.

Initial sample size = 39

Number rejected because of missing values = 1

Number rejected as outliers = 8

Final sample size = 30

Sample V1=17.

River Milo at Konsankoro, Guinea, 9° 02'N 09° 00'W, collected by Dr. R. Baker, Sept. 1986, caught at human bait, preserved in 95% propanol.

Species identity, *S. soubrense*, Dr. R. Baker personal communication.

Initial sample size = 38

Number rejected because of missing values = 2

Number rejected as outliers = 4

Final sample size = 32

Sample V1=18.

River Makona at Yalamba, Guinea, 8° 31'N 10° 11'W, collected by Dr. R. Baker, Sept. 1986, caught at human bait, preserved in 95% propanol.

Species identity, *S. soubrense*, Dr. R. Baker personal communication.

Initial sample size = 40

Number rejected because of missing values = 1

Number rejected as outliers = 1

Final sample size = 38

Sample V1=19.

River Bafing at Koukoutamba, Guinea, 11° 17'N 11° 20'W, collected by Dr. R. Baker, Sept. 1986, caught at human bait, preserved in 95% propanol.

Species identity, *S. soubrense*, Dr. R. Baker personal communication.

Initial sample size = 38

Number rejected because of missing values = 1

Number rejected as outliers = 6

Final sample size = 31

Sample V1=20.

River Koudeta at Bassi, Guinea, 10° 51'N 11° 14'W, collected by Dr. R. Baker, Sept. 1986, caught at human bait, preserved in 95% propanol.

Species identity, *S. squamosum*, Dr. R. Baker personal communication.

Initial sample size = 35

Number rejected because of missing values = 0

Number rejected as outliers = 1

Final sample size = 34

Sample V1=22.

River Bouka 2 at Sidakele, Guinea, 11° 30'N 10° 10'W, collected by Dr. R. Baker, Sept. 1986, caught at human bait, preserved in 95% propanol.

Species identity, *S. sirbanum*, Dr. R. Baker personal communication.

Initial sample size = 33

Number rejected because of missing values = 1

Number rejected as outliers = 3

Final sample size = 29

Sample V1=23.

River Mafou at Serekoroba, Guinea, 10° 24'N 10° 09'W, collected by Dr. R. Baker, Aug. 1985, caught at human bait, preserved in 70% ethanol.

Species identity, *S. sirbanum*, Dr. R. Baker personal communication.

Initial sample size = 50

Number rejected because of missing values = 1

Number rejected as outliers = 6

Final sample size = 43

Sample V1=24.

River Niger at Diaragbela, Guinea, 10° 36'N 09° 59'W, collected by Dr. R. Baker, Dec. 1985, caught at human bait, preserved in 70% ethanol.

Species identity, *S. sirbanum*, Dr. R. Baker personal communication.

Initial sample size = 35

Number rejected because of missing values = 0

Number rejected as outliers = 2

Final sample size = 33

Sample V1=25.

River Bale at Menankaya, Guinea, 09° 33'N 09° 35'W, collected by Dr. R. Baker, Aug. 1985, caught at human bait, preserved in 70% ethanol.

Species identity, *S. soubrense*, Dr. R. Baker personal communication.

Initial sample size = 27

Number rejected because of missing values = 0

Number rejected as outliers = 0

Final sample size = 27

Sample V1=26.

River Niger at Mamouria, Guinea, 09° 23'N 10° 34'W, collected by Dr. R. Baker, Sept. 1986, caught at human bait, preserved in 70% ethanol.

Species identity, *S. squamosum/S. yahense*, Dr. R. Baker personal communication.

Initial sample size = 20

Number rejected because of missing values = 0

Number rejected as outliers = 7

Final sample size = 13

Sample V1=27.

River Makona at Bofossou, Guinea, 08° 39'N 09° 41'W, collected by Dr. R. Baker, Sept. 1986, caught at human bait, preserved in 70% ethanol.

Species identity, *S. yahense/S. squamosum*, Dr. R. Baker personal communication.

Initial sample size = 25

Number rejected because of missing values = 1

Number rejected as outliers = 4

Final sample size = 20

Sample V1=28.

River Anie at Konogbe, Togo, 7° 48'N 1°. 5'E, collected by OCP insecticide team, 12/10/84, reared from pupae, preserved in 95% propanol. Species identity, *S. damnosum s.s.* and *S. soubrense* 'Beffa' mix, 14 *S. damnosum s.s.*, 4 *S. soubrense* 'Beffa' determined by D.P. Surtees. Mixture separated by external analysis, using LDFs, and internal analysis using PCA and cluster analysis.

Initial sample size = 43

Number rejected because of missing values = 2

Number rejected as outliers (following separation of mixture) = 4

Final sample size, V1=28 (*S. damnosum s.s.*) = 27 Final sample size, V1=29 (*S. soubrense* 'Beffa') = 11

Sample V1=29.

See V1=28.

Sample V1=30.

See V1=5.

Trop. Med. Parasit. 39 (1988) 120–122
© Georg Thieme Verlag Stuttgart - New York

The cytotaxonomy of the Djodji form of Simulium sanctipauli (Diptera: Simuliidae)

D. P. Surtees[1], G. Flasorgbor[2], R. J. Post[1], E. A. Weber[2]

[1]Department of Medical Entomology, Liverpool School of Tropical Medicine, Liverpool; [2]World Health Organisation, Onchocerciasis Control Programme, Ouagadougou, Burkina Faso

Summary

The Djodji form is described as a new cytotype of *Simulium sanctipauli*, within the *S. damnosum* complex, from the Ghana/Togo border area on the basis of sex chromosome differentiation.

Introduction

The potential importance of describing genetically distinct forms or geographic races within previously recognised cytospecies of vector complexes such as the *Simulium damnosum* complex comes from the possible correlation of the different forms with factors of epidemiological significance, such as anthropophily, together with their usefulness in tracing migration patterns or insecticide resistance distribution.

The purpose of this paper is to describe a new cytotaxonomic form within *S. sanctipauli* from Ghana and Togo, and to give criteria for its routine identification in any future studies of onchocerciasis transmission in that area.

Materials and methods

Breeding sites where the Djodji form of *S. sanctipauli* was collected are listed in Table 1. Larvae were fixed in 3:1 ethanol :acetic acid and stored in a refrigerator. For preparation of polytene chromosomes the larvae were split open ventrally and hydrolysed in hydrochloric acid. The silk glands were stained in feulgen and/or orcein following standard methods (Vajime and Dunbar 1975, Quillévéré 1975, Post 1986). Larval sex was determined according to the shape of the developing gonads (Puri 1925) after staining with feulgen. Inversions were scored from polytene chromosome preparations by comparison with the standard maps of Post (1986).

Chromosomal characteristics and cytotaxonomic key

All fixed and polymorphic inversions within Djodji form are indicated on the idiogram (Fig. 1), and frequencies of polymorphic inversions are listed in Table 2. The new form is homozygous for the fixed inversions 1L-P&Q, 2L-4&6&A and 3L-2, but there are no fixed inversions unique to Djodji form, and only one new rare polymorphic inversion (1S-P, see Figure 2). The presence of inversion 2L-A places Djodji form within *S. sanctipauli* (Post 1986). However, 1S-21 (Figure 3) is strongly Y-linked in Djodji form (Table 2), and this unique feature is the most important cytotaxonomic criterion for both description and routine identification.

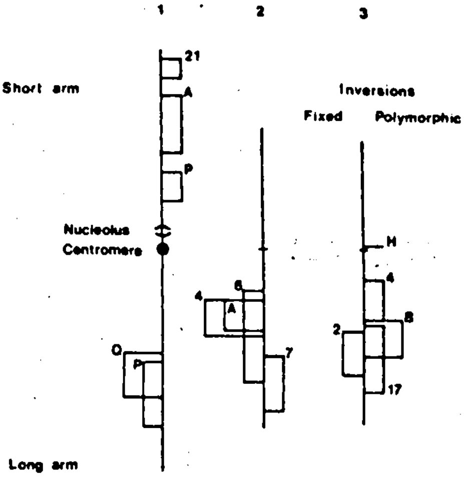


Fig. 1 Idiogram showing the relative positions of the breakpoints of all the inversions currently known from the Djodji form of *S. sanctipauli*. Those inversions plotted to the right are intraspecific polymorphisms, whilst those to the left are fixed inversions plotted on the idiogram to indicate the derivation of the Standard sequence (as seen in *S. squamosum*) from the basic Djodji sequence. The polymorphic inversion 3L-8 is based on the 3L-4.17.2 sequence, and not the 3L-2 sequence. Most of these inversions are illustrated by Post (1986), although 1S-21, 1S-A and the new inversion IS-P are also shown in Figures 2 and 3

Since 1S-21 is Y-linked in Djodji form there is no single inversion which is diagnostic of all individuals. However, samples in which there is strong Y-linkage of the inversion can be unequivocally identified as *S. sanctipauli* Djodji form, and mixed samples (should they exist) of the form with typical *S. sanctipauli*, will be recognised as such using standard population genetic analysis.

Accepted 21 January 1988

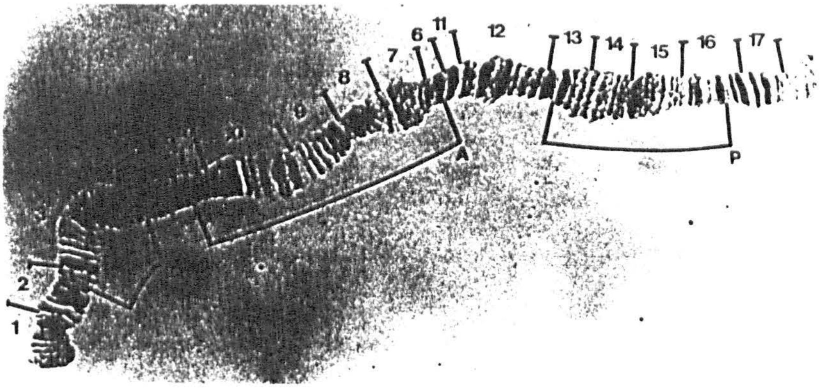


Fig. 2 *S. sanctipauli* (collected by R. J. Post from the river Sassandra at Soubré 26.10.84) showing the karyotype 1S-A/A with the breakpoints of inversions 1S-21 and 1S-P indicated

Table 1 List of larval samples from which the Djodji form of *S. sanctipauli* has been identified

Code: River: collection number, co-ordinates (N/E), date, cytospecies composition [1], collectors[2].

[1] sq = *S. squamosum*, ya = *S. yahense*, sa = Djodji form of *S. sanctipauli*, da = *S. damnosum* s.s. and si = *S. sirbanum*. Other cytospecies were not found in these samples.

[2] AKA = A. K. Adzah, AKO = A. K. Opoku, AMD = A. M. Denke, CKP = C. P. Kowal Post, EAW = E. A. Weber, HSA = H.S.K. Avissey, JEH = J. E. E. Henderickx, JFW = J. F. Walsh, MA = M. Ampah, MD = M. David, RAC = R. A. Cheke, RJP = R. J. Post, SS = S. Sowah, SSH = OCP subsector Hohoe, YY = Y. Yamagata.

Dayi: 1, 7°09′0°29′, 13.01.87, 22sq 5ya 3sa 1da, AKA. 2, 7°07′0°27′, 22.01.87, 20sq 74ya 8sa 1da, RAC EAW YY. 3, 7°06′0°26′, 17.03.86, 31sq 12ya 4sa AKA. 4, 7°06′0°26′, 29.04.86, 20sq 5ya 4sa, SSH. 5, 6°57′0°21′, 12.02.86, 17sq 15ya 21sa 55da 2si, AKA. 6, 6°57′0°21′, 21.03.86, 9sq 8sa 47da AKA. 7, 6°53′0°21′, 06.05.86, 2sq 4sa 8da 3si, SSH. 8, 6°53′0°21, 16.05.86, 5sq 2ya 17sa 16 da 8si, SSH. 9, 6°53′0°21′, 13.01.87, 2sq 4sa 8da 3si, RAC EAW YY. 10, 6°52′0°19′, 29.05.86, 2sq 8sa 19da 2si, AKA. 11, 6°52′0°19′, 22.01.87, 2sq 1sa 7da, RAC EAW YY.

Asukawkaw: 12, 7°54′0°37′, 05.02.87, 54sq 7sa JFW JEEH. 13, 7°54′0°37′, 27.05.86, 41sq 13sa, YY. 14, 7°54′0°37′, 23.01.87, 91sq 27sa, RAC EAW YY. 15, 7°52′0°36′, 18.03.86, 26sq 4sa, MD. 16, 7°52′0°29′, 23.01.87, 33sq 28sa, RAC EAW YY. 17, 7°41′0°26′, 28.05.86, 29sq 52sa, YY. 18, 7°41′0°25′, 06.02.86, 5sq 90sa. 19, 7°41′0°25′, 23.01.87, 2sq 74sa, RAC EAW YY.

Menou: 20, 7°37′0°39′, 28.05.86, 28sq 14sa, YY.

Gban-Houa: 21, 7°42′0°38′, 28.05.86, 3sq 25sa, YY. 22, 7°42′0°38′, 23.01.87, 12sq 29sa, RAC EAW YY. 23, 7°41′0°37′, 06.02.86, 12sq 18sa, YY. 24, 7°42′0°36′, 23.01.87, 11sq 22sa 1da, RAC EAW YY. 25, 7°42′0°35′, 15.10.84, 4sq 29sa, RJP CKP. 26, 7°42′0°35′, 15.03.85, 17sq 23sa, RAC AMD. 27, 7°42′0°35′, 21.03.85, 11sq 23sa, RAC AMD. 28, 7°42′0°35′, 26.03.85, 8sq 32sa, RAC AMD. 29, 7°42′0°35′, 29.03.85, 10sq 29 sa, RAC AMD. 30, 7°42′0°35′, 15.10.85, 13sq 21sa, RAC AMD. 31, 7°42′0°35′, 15.03.86, 45sq 22sa, YY. 32, 7°42′0°35′, 27.01.87, 28sq 54sa 1da, RAC HSA.

Wawa: 33, 7°43′0°33′, 20.03.86, 13sq 12sa, YY. 34, 7°43′0°33′, 23.01.87, 13sq 32sa, RAC EAW YY. 35, 7°41′0°30′, 27.03.86, 17sq 6sa, AKA.

Kpaza: 36, 8°33′0°41′, 15.10.87, JFW SS. 37, 8°33′0°41′, 22.10.87, JFW YY AKO. 38, 8°33′0°37′, 22.10.87, JFW YY AKO.

Niankpe: 39, 9°05′0°42′, 21.10.87, JFW YY AKO.

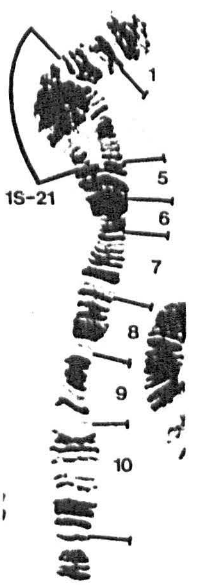


Fig. 3 The Djodji form of *S. sanctipauli* from the river Gban-Houa at Djodji showing the karyotype 1S-St/21

The following cytotaxonomic key can be used for the identification of *S. soubrense*, *S. soubrense* Beffa form, *S. sanctipauli* and the Djodji form of *S. sanctipauli*. The key should not be used west of Côte d'Ivoire, where *S. soubrense* form Konkouré and *S. soubrense* 'B' might also be encountered.

1) Larva homozygous for inversions 1L-P&Q, 2L-4&6 and 3L-2
.... *S. sanctipauli* subcomplex 2)
These inversions absent from larva
.... Other species of *S. damnosum* complex

Table 2 Inversion frequencies in the Djodji form of S. sanctipauli

		IS-21 karyotype frequency											
		Numbers of males			Numbers of females			Autosomal polymorphic inversion frequencies[2]					
River	Samples[1]	st/st	st/21	21/21	st/st	st/21	21/21	IS-A	2L-7	3L-B	3L-4.17	3L-24	IS-P
Dayi	1, 2, 4, 7 8, 9, 11	2	14	0	22	1	0	0.56	1.00	0	1.00	0	0
Asukaw-kaw	14, 16, 19	0	58	0	31	2	0	0.42	1.00	0	0.97	0	0
Gban-Houa	22, 24, 25 26, 27, 28, 29, 30, 32	.3	108	0	113	1	0	0.45	1.00	0	0.996	0.002	0.002
Wawa	34	1	12	0	8	0	0	0 50	1.00	0	1.00	0	0
Kpaza	36, 37	0	10	0	11	0	0	0.31	1.00	0	1.00	0	0

1Samples are as listed in Table 1. However, it was not always possible to score all inversions in every specimen, and hence sample sizes may be slightly smaller than those listed in Table 1.
2Inversions 2L-7 and 3L-B were noted heterozygously in a very few specimens from other samples which were not scored systematically for autosomal polymorphisms

2) Larva homozygous for inversion 2L-A........... *S. sanctipauli* 3)
   Inversion 2L-A absent from larva.................... *S. soubrense* 4)
3) Inversion 1S-21 Y-linked in population
   ....*S. sanctipauli* Djodji form
   Inversion 1S-21 not Y-linked in population
   ....*S. sanctipauli* typical form
4) Inversion 2S-6b absent from larva
   ....*S. soubrense* typical form,
   Inversion 2S-6c present in larva
   ....*S. soubrense* Beffa form

Discussion

The new cytotype seems to be largely confined to the Asukawkaw and Dayi river systems in the mountainous forest on the Ghana/Togo border (see Table 1). Within Togo and Benin, to the north and east, *S. soubrense* Beffa form appears to be the sole representative of the *S. sanctipauli* subcomplex with the exception of a few samples of the Djodji form identified from the rivers Kpaza and Niankpe in October 1987. To the west of the Volta lake *S. sanctipauli* typical form and *S. soubrense* are found (Meredith et al. 1983, Post 1986, and Fiasorgbor, Weber, Post and Surtees, unpublished data).

In view of the absence of any sympatric samples or unique fixed inversions, there is no evidence for Djodji form being a species distinct from *S. sanctipauli* elsewhere. However, the sex-linkage of 1S-21 indicates that Djodji populations are, by definition, genetically differentiated from other *S. sanctipauli* populations, and therefore it seems that Djodji form should be considered to be a geographic race within *S. sanctipauli*.

In a preliminary description this inversion was mistaken for a new inversion 1S-α (Surtees 1986), because 1S-21 is not just a simple reversal of the included bands, but is also associated with a consistent additional puff just outside the inversion and proximal to it. This gives 1S-21 the superficial appearance of being longer in Djodji form.

The taxonomic significance of sex-linked inversions in the Simuliidae has been discussed by Post (1982), and within the *S. damnosum* complex sex-linked inversions have been considered important in the cytotaxonomic description of several forms and species, such as *S. soubrense* s.n. Beffa form (Meredith et al. 1983), *S. yahense* (Vajime and Dunbar 1975), and Turiani form (Dunbar and Vajime 1981). In any case the importance of Djodji form, as with other forms described within the *S. damnosum* complex, lies not in its taxonomic level but rather in its possible epidemiological importance which is discussed by Garms and Cheke (1985) and Cheke and Denke (1988).

Acknowledgements
This work was supported by the World Health Organisation Onchocerciasis Control Programme (Director: E. M. Samba) and the British Medical Research Council. We should like to thank all those listed in Table 1 for the provision of larval samples, also R. A. Cheke and J. B. Davies for discussion, and K. B. Bougsere and T. Batascome for technical assistance in the laboratory.

References

Cheke, R. A., A. M. Denke: Anthropophily, zoophily and roles in onchocerciasis transmission of the Djodji form of *Simulium sanctipauli* and *Simulium squamosum.* Trop. Med. Parasit. 39 (1988) 123-127

Dunbar, R. W., C. G. Vajime: Cytotaxonomy of the *Simulium damnosum* complex. In: Laird, M. (ed.): Blackflies the future for biological methods of integrated control. Academic Press, London (1981) 31-43

Garms, R., R. A. Cheke: Infections with *Onchocerca volvulus* in different members of the *Simulium damnosum* complex in Togo and Benin. Z. angew. Zool. 72 (1985) 479-495

Meredith, S. E. O., R. A. Cheke, R. Garms: Variation and distribution of forms of *Simulium soubrense* and *S. sanctipauli* in West Africa. Ann. Trop. med. Parasit. 77 (1983) 627-640

Post, R. J.: Sex-linked inversions in blackflies. Heredity 48 (1982) 85-93

Post, R. J.: The cytotaxonomy of *Simulium sanctipauli* and *Simulium soubrense*,(Diptera: Simuliidae). Genetica 69 (1986) 191-207

Puri, I. M.: On the life history and structure of the early stages of Simuliidae (Diptera: Nematocera). Part I. Parasitology 17 (1925) 295-334

Quillévéré, D.: Étude du complexe *Simulium damnosum* en Afrique du l'Ouest. I. Techniques d'étude. Identification des cytotypes. Cah. O.R.S.T.O.M., ser. Ent. méd. et Parasitol. 13 (1975) 87-100

Surtees, D. P.: A new cytotype within *Simulium sanctipauli* (Diptera: Simuliidae) from Togo. Trans. R. Soc. trop. Med. Hyg. 80 (1986) 343

Vajime, C. G., R. W. Dunbar: Chromosomal identification of eight species of the subgenus *Edwardsellum* near and including *Simulium Edwarsellum damnosum* Theobald (Diptera: Simuliidae). Tropenmed. Parasit. 26 (1975) 111-138

D. P. Surtees, Dr. J. Post, Department of Medical Entomology, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, U.K.; G. Fiasorgbor, E. A. Weber, Organisation Mondiale de la Santé, Programme de lutte contre l'Onchocercose, B.P. 549 Ouagadougou, Burkina Faso.