منظمة الأغذية والزراعة للأمم المتحدة

联 合 国 粮 食 及 农 业 组 织

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS

ORGANISATION DES NATIONS UNIES POUR L'ALIMENTATION ET L'AGRICULTURE

ORGANIZACION DE LAS NACIONES UNIDAS PARA LA AGRICULTURA Y LA ALIMENTACION

FIR/MEDPOL/ALE/4

17 February 1992

# REPORT OF THE FAO/IOC/UNEP TRAINING WORKSHOP ON THE STATISTICAL TREATMENT AND INTERPRETATION OF MARINE COMMUNITY DATA

## Alexandria, Egypt, 9-19 December 1991

(Organized in the framework of the Long-term Programme for Pollution Monitoring and Research in the Mediterranean (MED POL - Phase II)

In cooperation with:

IOC UNEP

Athens, February 1992

TABLE OF CONTENTS

## 1. INTRODUCTION

In 1988 the first regional FAO/IOC/UNEP Training Workshop on the Statistical Treatment and Interpretation of Marine Community Data took place at the Marine Biological Station in Piran, Yugoslavia, from 14-24 June, in the framework of the Long-term Programme for Pollution Monitoring and Research in the Mediterranean Sea (MED POL - Phase II). Owing to the nature of the workshop (lectures and practical sessions with personal computers), the number of participants that could be accepted was limited. In view of the large number of applications for participation in the workshop which could not be satisfied, it was decided by the relevant UN Agencies to hold two workshops at a national level in the countries from which the majority of applications came from namely, Greece and Yugoslavia, in 1989 and 1990 respectively. The final workshop in the series was then planned for 1991, at a regional level, and it is this workshop that is the subject of this report.

The present workshop took place at the Arab Maritime Transport Academy (AMTA), Alexandria, Egypt, from 9-19 December 1991. It was attended by 21 participants from a diverse range of countries bordering the Mediterranean (Egypt, Italy, Libya, Morocco, Spain, Tunisia and Turkey). A full list of participants appears in Annex I. The lectures were principally given by Professor J.S. Gray (University of Oslo, Norway) and Drs. K.R. Clarke and R.M. Warwick (Plymouth Marine Laboratory (PML), UK). Mr. M.R. Carr (PML) was responsible for supervising the preparation of the multivariate statistical software; he and Dr. E. Papathanassiou (National Centre for Marine Research, Athens) acted as demonstrators for the practical sessions.

## 2. PROGRAMME OF THE WORKSHOP

The workshop covered the statistical treatment of population and community data (species abundances/biomass), arising in studies of the marine environment. In particular, the emphasis was on statistical analysis of the biological effects of pollutants, the workshop being conducted through lectures and practical computing sessions involving a range of data sets drawn from the literature.

The methods covered ranged from "classical" univariate statistics applied to, for example, population abundances and diversity indices, to multivariate clustering and ordination techniques, and other graphical methods, applied to large arrays of samples/species data. Practical work on univariate statistics used the STATGRAPHICS package, and multivariate analyses were undertaken using the package PRIMER (Plymouth Routines In Multivariate Ecological Research), a suite of PC programs written at the Plymouth Marine Laboratory, UK.

The lectures and practical sessions drew from the experience of the benthic community studies components of research workshops mounted by IOC/GEEP (the Intergovernmental Oceanographic Commission's Group of Experts on the Effects of Pollutants), in particular workshops held in Oslo, August 1986, Bermuda, September 1988 and Bremerhaven, March 1990.

The lectures and practical sessions (see Annex II) in the first half of the workshop were meant as an introductory (or refresher) course for participants on basic (univariate) statistics, so that all would be at comparable levels when undertaking the second (multivariate) part of the course. The emphasis given here was not to statistical theory but rather to the practicalities that marine ecologists face. Questions treated covered such problems as: how many samples? or how large a sample should one take?, etc. To answer these questions requires an understanding of some basic statistical terms such as variance, standard deviation and standard error, confidence limits, transformations etc. The introductory course covered these topics and then demonstrated how to compare univariate samples using Student's 't' test and the analysis of variance, and the relationships between samples using regression and correlation analyses.

The lecture material which followed covered:

a) the use of multivariate methods (clustering and ordination) to represent graphically the similarities between species abundances (or biomass) observed in a set of samples;

b) the demonstration of statistically significant differences in species composition between several sites (or the same site at several times) - this is a necessary pre-requisite to further analyses attempting to explain those differences;

c) the construction of univariate indices (eg diversity) and distributional plots (eg abundance-biomass comparisons) which indicate levels of disturbance or "stress" at sites;

d) the relation of both univariate and multivariate faunal descriptions to gradients of chemical contamination and background environmental variables.

The practical sessions allowed the participants to apply the methods described in the accompanying lectures, on published data sets chosen to illustrate changes in benthic community composition at macrofaunal and meiofaunal levels, resulting from contaminant impact by sewage sludge dumping, pulp mill effluent, oil spills etc.

In the event, the timing of breaks etc in the programme outlined in Annex II was not adhered to exactly, mainly due to the keenness of the participants to work for longer periods both in the middle and at the end of each day. However, the lecture content was exactly as programmed and the lecture notes are given in detail in Annex III.

## 3. EVALUATION OF THE WORKSHOP

At the end of the workshop, participants were asked to fill in the following questionnaire. This had exactly the same form as at previous workshops, in order to allow direct comparison. The questions asked are given in full below together with a summary of the replies (usually in the form of the percentage of replies that were a, b or c); a total of 18 questionnaires were returned.

QUESTIONNAIRE

Please respond to the following questions, which will help us run better courses in the future. Please be honest, the answers are anonymous!

Q 1. Did the course announcement describe the content of the course:
a) well, b) adequately, c) inaccurately (If c state why)
a) 94%, b) 6% c) 0%

Q 2. Were the levels of expertise expected and described in the course announcement:
a) accurate, b) acceptable, c) inaccurate (If c explain why)
a) 74%, b) 28%, c) 0%

Q 3. Was the information sent to you prior to the course:
a) good, b) average, c) poor (If c explain why)
a) 83%, b) 17%, c) 0%

Q 4. Were the basic institution facilities:
a) good, b) average, c) poor (If c explain why)
a) 50%, b) 50%, c) 0%

Q 5. Were the computing facilities:
a) good, b) average, c) poor (If c explain why)
a) 39%, b) 61%, c) 0%

Q 6. Was the overall course design:
a) fine, b) too detailed, c) inadequate (If b or c explain why)
a) 94%, b) 6%, c) 0%

Q 7. How important were the Introductory Statistics lectures as a refresher before the multivariate statistics?
a) very important, b) useful, c) could be omitted
a) 61%, b) 39%, c) 0%

Q 8. Is STATGRAPHICS an essential part of the course:
a) yes, b) perhaps, c) no (If c say why)
a) 89%, b) 11%, c) 0%

Q 9. Please fill out the spaces below with a score of 2 for good, 1 for average and 0 for poor for the content of:

| | Lectures | Practicals |
|---|---|---|
| Introductory statistics (In.1-In.5) | ________ | ________ |
| Multivariate/graphical statistics (1-14) | ________ | ________ |
| General lectures (no numbers) | ________ | |

Introductory statistics - content

| | | | |
|---|---|---|---|
| Lectures | 2: 100%, | 1: 0%, | 0: 0% |
| Practicals | 2: 78%, | 1: 22%, | 0: 0% |

Multivariate/graphical statistics - content

| | | | |
|---|---|---|---|
| Lectures | 2: 100%, | 1: 0%, | 0: 0% |
| Practicals | 2: 67%, | 1: 33%, | 0: 0% |

General lectures - content

| | | | |
|---|---|---|---|
| Lectures | 2: 94%, | 1: 6%, | 0: 0% |

Q 10. Please fill out the spaces below with a score of 2 for good, 1 for average and 0 for poor for the ease of understanding of:

| | Lectures | Manuals | Examples |
|---|---|---|---|
| Introductory statistics | ________ | ________ | ________ |
| Multivariate/ Graphical statistics | ________ | ________ | ________ |
| Computer programs | ________ | ________ | |

Introductory statistics - ease of understanding

| | | | |
|---|---|---|---|
| Lectures | 2: 83%, | 1: 17%, | 0: 0% |
| Manuals | 2: 76%, | 1: 24%, | 0: 0% |
| Examples | 2: 76%, | 1: 24%, | 0: 0% |

Multivariate/graphical statistics - ease of understanding

| | | | |
|---|---|---|---|
| Lectures | 2: 78%, | 1: 22%, | 0: 0% |
| Manuals | 2: 61%, | 1: 39%, | 0: 8% |
| Examples | 2: 72%, | 1: 28%, | 0: 0% |

Programs - ease of understanding

| | | | |
|---|---|---|---|
| Lectures | 2: 56%, | 1: 44%, | 0: 0% |
| Manuals | 2: 61%, | 1: 39%, | 0: 0% |

Q 11. Do the multivariate programs fulfil the demands that you have for data analyses:
a) well, b) averagely, c) poorly (If c say why)
a) 72%, b) 22%, c) 6%

Q 12. Were the programs:
a) easy to use, b) acceptable to use, c) difficult to use (If c say why)
a) 33%, b) 67%, c) 0%

Q 13. Was the progress through the course:
a) just right, b) too fast, c) too slow (If b or c say why)
a) 39%, b) 61%, c) 0%

Q 14. Which parts of the course did you find most useful?

All respondents replied to this question, identifying specific parts of the course. The following were mentioned by the given % of respondents:

| | |
|---|---|
| 53%: | Multivariate/graphical analysis |
| 29%: | All parts |
| 12%: | Introductory statistics and STATGRAPHICS |
| 6%: | Practicals |

Q 15. How does this course compare with other UN courses:
a) better than average, b) average, c) below average (If c say why)
a) 60%, b) 40%, c) 0% (there were 10 respondents to this question).

Q 16. How does this course compare with other courses (state which)
a) 67%, b) 33%, c) 0% (there were 16 respondents to this question).

Q 17. Any other comments you may wish to add?

The one comment that was common to 40% of the replies was to the effect that a longer workshop would be necessary to assimilate fully all the information presented (several verbal comments sugested a month!). There were no other critical comments apart from some congratulatory ones.

As in previous workshops in this series (Piran, Athens, Split) the overall response to the workshop was positive and enthusiastic. Course participants were generally very keen, prepared to work long and taxing days, and able to take on board some very complex concepts. Though the general level of previous experience of statistics and microcomputing, and the facility with spoken and written English, were even more variable than in the previous workshop, nonetheless all participants seemed to feel that they had acquired relevant knowledge and experience from the two-week period. This is evidenced by the rating of the overall course design (94% said "fine") and by the large number of replies to questions that gave the maximum possible rating. For all participants and across all questions that had a straight multiple choice rating for aspects of the course, between:

a or 2) good, fine, essential, easy to understand, easy to use, just right, better than average etc;

b or 1) adequate, acceptable, average and

c or 0) poor, below average, inaccurate, inadequate, difficult to use, 72% of the replies were a), 28% were b) with only 1 out of 454 replies of c). This is a very high level of satisfaction and indicates that the course was very worthwhile. Nonetheless, it is instructive to look more closely at the replies to specific questions, particularly where opinion was more evenly divided between a) and b) replies.

Questions on prior notification of the course content, levels of expertise expected etc attracted excellent ratings, very much better than in the previous (national) workshops in this series. The institution facilities were considered good or acceptable (equally split) by all participants, a rating very similar to previous workshops. There was somewhat less enthusiasm for the computing facilities than in the previous workshop (which in fact had attracted a rather higher rating than earlier workshops, because of the more powerful nature of the PCs). There was clearly a slight divergence of opinion here between the participants and the lecturers. The latter considered the computing facilities to be the best encountered in the entire series; all machines were basically of the same type and set up in the same way, there were ample machines (at least one for every two participants), printers were all set up with switching devices to allow hard-copy output from all machines, and the computing support from local AMTA staff, when occasional hardware or installation problems were encountered, was truly excellent. In comparison with previous workshops, installation and operation of the programs was straightforward and comparatively trouble-free. Being XT-compatibles, the PCs were rather slow in operation however, and some of the participants clearly found them slightly slower than the machines available in their home laboratories - though it should be emphasised that none of them considered the computing facilities to be "poor".

There is an interesting general point here about rising expectations in the area of personal computing. The other question to attract a notably lower rating than in previous questionnaires concerned the ease of use of the programs. Whilst all considered them acceptable to use (or better), only a third thought they were easy to use. This partly results from the fact that some participants had previously no experience of PCs but also, one suspects, reflects the fact that those with some previous experience (the majority) were more used to a Windows-type interface (a GUI - graphical user interface) which is becoming the "norm" in mass-market PC software. Expectations of hardware performance and software presentation are continually rising and it may be that the current "question and answer" user interface, essentially dating from the first workshop in 1988, is beginning to show its age.

The material covered by the PRIMER programs is continually evolving of course, and seemed very well received. Nearly three-quarters classed the multivariate programs as fulfilling their data requirements "well" (one-quarter said "averagely"). This is somewhat better than in the previous workshop, and, along with other answers, reflects the careful selection of participants for the Alexandria workshop - most had strong interest in analysing data sets of the type covered in the course. This is also clear from the selections given in answer to question 14 and the 100% rating given to the lecture content in question 9 (the first time that the main introductory and multivariate lectures have attracted the maximum rating for content from all participants!).

The Introductory Statistics lectures, covering univariate statistical techniques such as t-tests, ANOVA, regression etc were regarded as very important by nearly two-thirds of the respondents, and useful by all. The STATGRAPHICS package, used to perform these analyses, was also considered an essential part of the course by nearly all people, a marked increase on the response in the previous workshop. This must reflect the additional emphasis placed on univariate analyses following the experience of the Split workshop. There, the lectures and practicals on this aspect were condensed into (probably) too short a period, especially given the range of data sets brought to the workshop on that occasion, several of which were exclusively univariate. On this occasion, the univariate side was suitably expanded and this clearly met with the approval of participants (though the subsequent reduction of time available for multivariate methods may have contributed to the comments about lack of time to grasp fully all aspects of the material covered). In the discussions of "own data", several univariate analyses were important, particularly higher-way analysis of variance and regression/ correlation analyses.

Perhaps the only other point of significance to note is the slightly lower rating for ease of understanding (of lectures and manuals) than recorded for the Split workshop - the current levels were closer to those attracted at earlier workshops. Nonetheless, between two-thirds and three-quarters of respondents gave maximum rating for this attribute, with no-one classing "ease of understanding" as poor. Great efforts were made by the lecturers to simplify their presentations, in particular to lecture at a very slow speed, commensurate with the difficulties in understanding English that were expected with some participants. In general, this seems to have worked well, with many of the participants expressing gratitude (verbally) for the efforts made to maximise the effectiveness of communication. However, this again had a tendency to reduce the length of time actually spent in performing the practical examples or analysing "own data", adding to the feeling that progress through the course was too fast and that a much longer course was needed to do it justice (answers to questions 13 and 17). The latter was impossible of course, logistically, financially and from the viewpoint of availability of lecturers and demonstrators. Nonetheless, 40% of respondents felt that speed through the course was "just right", and there is always going to be difficulty in resolving the disparate needs of trainees with very variable competence in the workshop language, and in background experience (eg of PC use).

Fewer people answered question 15, though most answered 16. In general, comparisons with other courses were favourable. In summary, there can be no doubt that the workshop was very successful and it showed again the relevance of new methods of multivariate analysis, combined with basic univariate statistics, to the type of field and laboratory data being collected in the region.

## 4. CONCLUSION

The lecturers and demonstrators were most impressed with the enthusiasm and dedication of the trainees, the helpfulness of staff and the facilities provided at the Arab Maritime Transport Academy, and the smooth organization of the whole event; several interesting discussions of participants' own research data were initiated and it is anticipated that, as in previous workshops, this should result in an improved quality of reporting and publication of their work.

## ANNEX I

## LIST OF PARTICIPANTS

Daniela BASSO
Università degli Studi di Milano
Dipartimento di Scienze della Terra
Sezione di Geologia e Paleontologia
Via Mangiagalli, 34
20133 Milano
ITALY

Tel. 39-2-23698 ext. 226
Fax. 39-2-70638261

Amel CHAFFAI HAMZA
Laboratoire de l'Environnement et de la
Physiologie des poissons
Département de Biologie
Faculté des Sciences de Sfax
Route Sokra Km 4
3038 Sfax
TUNISIA

Tel. 216-4-74409/74088, ext. 423
Fax. 216-4-74437
Tlx. 40982

Sohier Abdel EISA
National Institute of Oceanography and Fisheries
Ministry of Scientific Research and Technology
Anfoshy
Kayet Bay
Alexandria
EGYPT

Tel. 20-3-801553/807138
Fax. 20-3-801174

Sawsan Abu EL-EZZ
National Institute of Oceanography and Fisheries
Ministry of Scientific Research and Technology
Anfoshy
Kayet Bay
Alexandria
EGYPT

Tel. 20-3-801553/807138
Fax. 20-3-801174

Mohamed EL-KOMI
National Institute of Oceanography and Fisheries
Ministry of Scientific Research and Technology
Anfoshy
Kayet Bay
Alexandria
EGYPT

Tel. 20-3-801553/807138
Fax. 20-3-801174

Hoda EL-RASHIDY
Alexandria University
Department of Oceanography
Maharram Bay
Alexandria
EGYPT

Tel. 20-3-4922918/919
Fax. 20-3-801174

Zeinab EL-SHERIF
National Institute of Oceanography and Fisheries
Ministry of Scientific Research and Technology
Anfoshy
Kayet Bay
Alexandria
EGYPT

Tel. 20-3-801553/807138
Fax. 20-3-801174

Ibrahim Saleh EL-ZIANI
Marine Biology Research Centre
P.O. Box 30830
Tajura, Tripoli
LIBYA

Tel. 218-21-690001-3
Tlx. 20523

Maria Luz FERNANDEZ de PUELLES
Spanish Institute of Oceanography
Baleares Laboratory
Muelle de Poniente
APDO 291
Palma de Mallorca
SPAIN

Tel. 34-71-401561
Fax. 34-71-404945
Tlx. 69493

Simonetta FRASCHETTI
Institute for Marine Environmental Sciences
University of Genoa
Corso Rainusso, 14
C P 79
16038 S. Margherita Ligure
Genova
ITALY

Tel. 39-185-286195/283415
Fax. 39-185-281089
Tlx. 271114

Ahmed Sanousi GHASEM
Marine Biology Research Centre
P.O. Box 30830
Tajura, Tripoli
LIBYA

Tel. 218-21-690001-3
Tlx. 20523

Amany MAHMOUD
Alexandria University
Department of Oceanography
Maharram Bay
Alexandria
EGYPT

Tel. 20-3-4922918/919
Fax. 20-3-801174

Erhan MUTLU
Institute of Marine Sciences
Middle East Technical University
P.O. Box 28
TR 33731, Erdemli
TURKEY

Tel. 90-7586-1406
Fax. 90-7586-1327
Tlx. 67796 DMS

Mohamed RAMDANI
Institut des Pêches Maritimes CASA
Rue de Tiznit
B.P. 21
Casablanca 01
MOROCCO

Tel. 212-276088
Tlx. 23823 IPEMAR

Paolo SARTOR
Department of Environmental and
Territorial Science
Section Marine Biology
Via Volta, 6
56100 Pisa
ITALY

Tel. 39-50-500943
Fax. 39-50-49694
Tlx. 590036 UNIVPI

Mario SBRANA
Department of Environmental and
Territorial Science
Section Marine Biology
Via Volta, 6
56100 Pisa
ITALY

Tel. 39-50-500943
Fax. 39-50-49694
Tlx. 590036 UNIVPI

Hanan SHOKRY
National Institute of Oceanography and Fisheries
Ministry of Scientific Research and Technology
Anfoshy
Kayet Bay
Alexandria
EGYPT

Tel. 20-3-801553/807138
Fax. 20-3-801174

Gamil Henen THOMAS
National Institute of Oceanography and Fisheries
Ministry of Scientific Research and Technology
Anfoshy
Kayet Bay
Alexandria
EGYPT

Tel. 20-3-801553/807138
Fax. 20-3-801174

Souad TURKI
Institut National Scientifique et Technique
d'Océanographie et de Pêche
Salammbo
Port de Pêche
2060 La Goulette
Tunis
TUNISIA

Tel. 216-1-731848
Fax. 216-1-730496
Tlx. 14739 MEDRAP

Fatma ZAGHLOUL
National Institute of Oceanography and Fisheries
Ministry of Scientific Research and Technology
Anfoshy
Kayet Bay
Alexandria
EGYPT

Tel. 20-3-801553/807138
Fax. 20-3-801174

Salem Wanis ZAGOUZI
Marine Biology Research Centre
P.O. Box 30830
Tajura, Tripoli
LIBYA

Tel. 218-21-690001-3
Tlx. 20523

## Lecturers, UN staff and Local Organizers

John GRAY
Professor of Marine Biology
University of Oslo
P.O.Box 1064
0316 Blindern
Oslo 3
NORWAY

Tel. 47-2-854510
Fax. 47-2-854438
Tlx. 72425

Robert CLARKE
Principal Research Scientist
Plymouth Marine Laboratory
Prospect Place, West Hoe
Plymouth Pl1 3DH
UNITED KINGDOM

Tel. 44-752-222772
Fax. 44-752-670637

Richard WARWICK
Principal Research Scientist
Plymouth Marine Laboratory
Prospect Place, West Hoe
Plymouth Pl1 3DH
UNITED KINGDOM

Tel. 44-752-222772
Fax. 44-752-670637

Martin CARR
Plymouth Marine Laboratory
Prospect Place, West Hoe
Plymouth Pl1 3DH
UNITED KINGDOM

Tel. 44-752-222772
Fax. 44-752-670637

Evangelos PAPATHANASSIOU
Head, Department of Marine Biology
National Centre for Marine Research
Aghios Kosmas
Hellinikon
166 04 Athens
GREECE

Tel. 30-1-9829237
Fax. 30-1-9833095
Tlx. 224135 NCMR

Gabriel P. GABRIELIDES
Senior Fishery Officer (Marine Pollution)
Food and Agriculture Organization of the
United Nations
Co-ordinating Unit for the Mediterranean
Action Plan
P.O. Box 18019
Leoforos Vassileos Konstantinou 48
116 10 Athens
GREECE

Tel. 30-1-7244536-9
Fax. 30-1-7291160
Tlx. 22564 MEDU GR

Youssri EL-GAMAL
Arab Maritime Transport Academy
Gamal Abdel Nasser Street
P.O. Box 1029
Miamy
Alexandria
EGYPT

Tel. 20-3-5601785
Fax. 20-3-5497882
Tlx. 54160 ACAD UN

Fouad M. ASSAL
Arab Maritime Transport Academy
Gamal Abdel Nasser Street
P.O. Box 1029
Miamy
Alexandria
EGYPT

Tel. 20-3-865429
Fax. 20-3-5497882
Tlx. 54160 ACAD UN

## ANNEX II

## PROGRAMME OF WORKSHOP

### Monday 9 December

| | | |
|---|---|---|
| 08:30-09:00 | | Registration of participants |
| 09:00-09:30 | | Opening of the Workshop |
| 09:30-10:30 | I.1 | The precision of the mean and confidence limits by J.S. Gray |
| 10:30-11:00 | | Coffee break |
| 11:00-12:30 | | Introduction to PC-DOS and STATGRAPHICS by M.R. Carr |
| 12:30-14:00 | | Lunch break |
| 14:00-16:00 | | Practical session on PC-DOS and introduction to STATGRAPHICS |

### Tuesday 10 December

| | | |
|---|---|---|
| 08:30-09:30 | I.2 | Spatial dispersion of populations and transformations of data by J.S. Gray |
| 09:30-10:30 | | Practical session on I.1 and I.2 |
| 10:30-11:00 | | Coffee break |
| 11:00-12:00 | | Practical session on I.2 |
| 12:00-12:30 | I.4 | Comparing samples: t-test and paired t-test by J.S. Gray |
| 12:30-14:00 | | Lunch break |
| 14:00-15:30 | | Practical session on I.4 |
| 15:30-16:00 | | Discussion of arrangements for analysing participants' own data sets |

### Wednesday 11 December

| | | |
|---|---|---|
| 08:30-09:30 | I.4 | Analysis of variance by J.S. Gray |
| 09:30-10:30 | | Practical session on I.4 |
| 10:30-11:00 | | Coffee break |

| | | |
|---|---|---|
| 11:00-11:30 | | Practical session on I.4 |
| 11:30-12:30 | I.5 | Regression and correlation by J.S. Gray |
| 12:30-14:00 | | Lunch break |
| 14:00-15:00 | | Practical session on I.5 |
| 15:00-15:30 | | Introduction to entering or reformatting own data sets by E. Papathanassiou |
| 15:30-16:30 | | Practical session on entering or reformatting own data |

Thursday 12 December

| | | |
|---|---|---|
| 08:30-09:30 | I.6 | Non-parametric methods by J.S. Gray |
| 09:30-10:30 | | Practical session on I.6 |
| 10:30-11:00 | | Coffee break |
| 11:00-12:30 | | General lecture on biological effects and monitoring of pollutants by J.S. Gray |
| 12:30-14:00 | | Lunch break |
| 14:00-15:30 | | Practical session on own data sets (univariate analysis only) |
| 15:30-16:30 | | Question and answer session on basic (univariate) statistics |

Saturday 14 December

| | | |
|---|---|---|
| 08:30-09:30 | II.1 | A framework for studying changes in community structure by R.M. Warwick |
| 09:30-10:30 | II.2 | Multivariate methods: measures of similarity of species abundance/biomass between samples by K.R. Clarke |
| 10:30-11:00 | | Coffee break |
| 11:00-11:30 | | Practical session on lecture II.2 (by hand) |
| 11:30-12:30 | II.3 | Multivariate methods: hierarchical clustering by K.R. Clarke |
| 12:30-14:00 | | Lunch break |
| 14:00-14:30 | | Practical session on lecture II.3 (by hand) |

| | | |
|---|---|---|
| 14:30-15:00 | | Introduction to multivariate software by M.R. Carr |
| 15:00-16:00 | | Practical session on lectures II.2 and II.3 (on computer) |

Sunday 15 December

| | | |
|---|---|---|
| 08:30-09:45 | II.4 | Multivariate methods: ordination of samples by Principal Components Analysis by K.R. Clarke |
| 09:45-10:30 | | Continued practical session on lectures II.2 and II.3 |
| 10:30-11:00 | | Coffee break |
| 11:00-12:30 | II.5 | Multivariate methods: ordination of samples by Multi-Dimensional Scaling (MDS) by K.R. Clarke |
| 12:30-14:00 | | Lunch break |
| 14:00-16:00 | | Practical session on lectures II.4 and II.5 |

Monday 16 December

| | | |
|---|---|---|
| 08:30-09:45 | II.8 | Univariate and distributional methods: diversity measures, dominance curves and other graphical analyses by R.M. Warwick |
| 09:45-10:30 | | Practical session on lecture II.8 |
| 10:30-11:00 | | Coffee break |
| 11:00-12:00 | | Continued practical session on lecture II.8 |
| 12:00-12:30 | II.10 | Species aggregation by R.M. Warwick |
| 12:30-14:00 | | Lunch break |
| 14:00-16:00 | | Practical session on own data sets |

Tuesday 17 December

| | | |
|---|---|---|
| 08:30-10:30 | II.6/7 | Testing for differences between groups of samples, and species contributions by K.R. Clarke |
| 10:30-11:00 | | Coffee break |
| 11:00-12:30 | | Practical session on lecture II.6/7 |
| 12:30-14:00 | | Lunch break |

| | | |
|---|---|---|
| 14:00-15:00 | II.9 | Transformations by K.R. Clarke |
| 15:00-16:30 | | Practical session on own data sets |

Wednesday 18 December

| | | |
|---|---|---|
| 08:30-09:30 | II.11 | Linking community analyses to environmental variables by K.R. Clarke |
| 09:30-10:30 | | Practical session on lecture II.11 |
| 10:30-11:00 | | Coffee break |
| 11:00-11:45 | II.12 | Causality: community experiments in the field and laboratory by R.M. Warwick |
| 11:45-12:30 | II.13 | Data requirements for biological effects studies: which components and attributes of the biota to examine? |
| 12:30-14:00 | | Lunch break |
| 14:00-16:00 | | Practical session on own data sets |

Thursday 19 December

| | | |
|---|---|---|
| 08:30-09:15 | II.14 | Relative sensitivities and merits of univariate, graphical/distributional and multivariate techniques by R.M. Warwick |
| 09:15-10:30 | | Practical session on own data sets |
| 10:30-11:00 | | Coffee break |
| 11:00-12:00 | | Practical session on own data sets |
| 12:00-12:30 | | Arrangements for obtaining and mounting multivariate software by M.R. Carr |
| 12:30-14:00 | | Lunch break |
| 14:00-15:30 | | Discussion of participants' own data results |
| 15:30-16:30 | | Question and answer session on multivariate/graphical methods |

Note: The numbering of lectures corresponds to the order of material presented in the Workshop notes: the Part I lecture notes (I.1 to I.6) by Prof. Gray deal with basic (univariate) statistics and the Part II notes (II.1 to II.14) by Drs Clarke and Warwick deal with multivariate community analyses. The other (unnumbered) lectures detailed above are not covered in the formal lecture notes but hand-outs will be available with the lecture, in some instances. A separate set of notes by Mr Carr covers the practical details of using the computer programs for multivariate analysis, developed at the Plymouth Marine Laboratory.

# ANNEX III

## LECTURE NOTES

# INTRODUCTION

This document is divided into two parts. Part I contains the material drawn upon by Professor J.S. Gray and Mr. M.R. Carr for the six introductory lectures and Part II consists of the lectures given by Drs K.R. Clarke and R.M. Warwick. No lecture notes are available for the informal lectures which do not constitute the core of the workshop but hand-outs will be available with the lecture, in some instances. A separate set of notes by Mr Carr covers the practical details of using the computer programs for multivariate analysis, developed at the Plymouth Marine Laboratory.

The following lectures which deal with some basic and important statistical concepts and their practical application are contained in Part I.

| | |
|---|---|
| Lecture 1: | The precision of the mean and confidence limits |
| Lecture 2: | Spatial dispersion of populations and transformations of data |
| Lecture 3: | Sampling and sub-sampling |
| Lecture 4: | Comparing samples: 't' test, paired 't' test and analysis of variance |
| Lecture 5: | Regression and correlation analyses |
| Lecture 6: | Non-parametric methods |

The following lectures which are principally concerned with graphical and multivariate statistical analysis of community data are contained in Part II.

| | |
|---|---|
| Lecture 1: | A framework for studying changes in community structure |
| Lecture 2: | Multivariate methods: measures of similarity of species abundance/biomass between samples |
| Lecture 3: | Multivariate methods: hierarchical clustering |
| Lecture 4: | Multivariate methods: ordination of samples by Principal Component Analysis (PCA) |
| Lecture 5: | Multivariate methods: ordination of samples by Multi-Dimensional Scaling (MDS). |
| Lecture 6: | Multivariate methods: testing for differences between groups of samples |
| Lecture 7: | Multivariate methods: species analyses |
| Lecture 8: | Univariate and distributional methods: diversity measures, dominance curves and other graphical analyses |
| Lecture 9: | Transformations |

Lecture 10: Species removal and aggregation

Lecture 11: Linking multivariate and univariate community analyses to environmental variables

Lecture 12: Causality: community experiments in the field and laboratory

Lecture 13: Data requirements for biological effects studies: which components and attributes of the biota to examine?

Lecture 14: Relative sensitivities and merits of univariate, graphical/distributional and multivariate techniques.

# PART I

## Lecture 1: THE PRECISION OF THE MEAN AND CONFIDENCE LIMITS

There are a few terms that must be understood from the beginning.

$\bar{x}$ - the sample mean
$\mu$ - the true population mean

The sample mean is simply the average of a series of samples. For example let us take 5 x 0.1 $m^2$ samples of a population of bivalves on a beach. We obtain the following data:

25,20,15,35,60 $\bar{x} = 31$ and the sample size (n)= 5

We first calculate the variance ($s^2$)

This is $s^2 = (\Sigma X^2 - (\Sigma X)^2/n) / (n - 1)$

$= (6075 - (155^2/5) / (4)$

$s^2 = 317.17$

This gives another term the STANDARD DEVIATION (s)

where $s = \sqrt{s^2}$

$s = 17.82$

Often we wish to estimate the total population in a given area and the precision with which we estimate that population.

e.g. $\bar{x} = 10.125$ taken from a sample area of 100 $cm^2$. The total area is 300,000 $cm^2$.

Therefore, the population estimate is

$300,000 * 10.125/100 = 30,375$

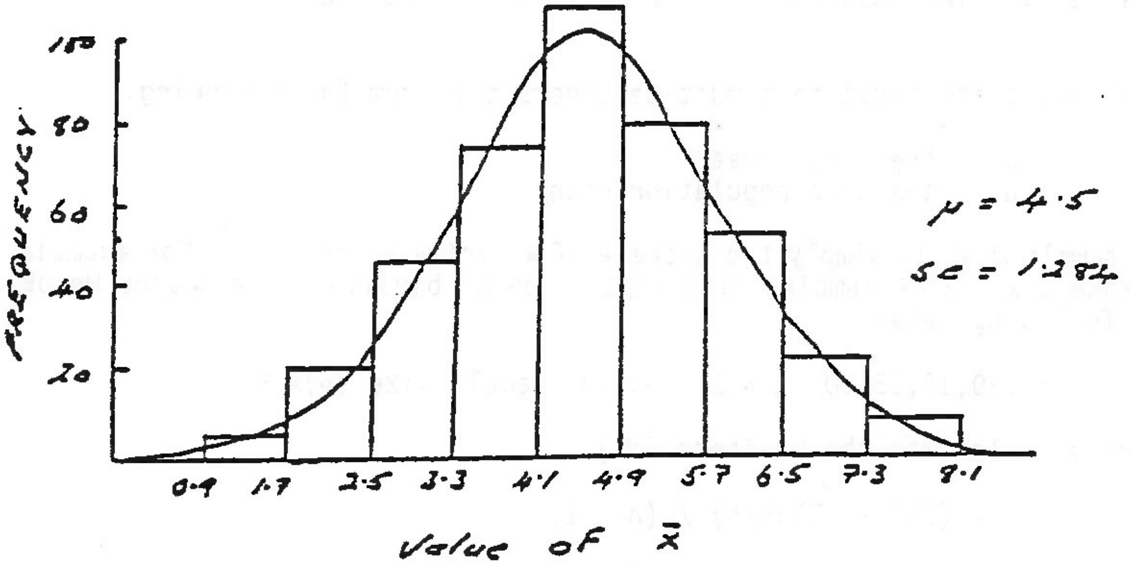
The true population mean ($\mu$) has only one value but the sample mean ($\bar{x}$) has many values

e.g. 9.51, 10.74, 9.82, 10.20, 10.125

and these could all be used to estimate the total population size. If we took enough samples all these estimates could be arranged in a frequency distribution and would give us a normal distribution centered around ($\mu$) the true population mean.

This gives us one of the major rules in statistics; the so-called Central Limit Theorem which states that:

"The means of large random samples from the same population are approximately normally distributed with mean equal to the population mean ($\mu$) and variance near the population variance ($\sigma^2$)."



In the above definition the term <u>large</u> is used. In statistical terms large means a sample size of over 30 samples, so-called large sample statistics. Often in marine biology it is impossible to take such large samples and we must be aware that in most cases we operate on <u>SMALL SAMPLE STATISTICS</u> which are NOT necessarily those that are in your P.C. or mainframe computer!

Above, we have used another term, variance. We have the <u>sample variance ($s^2$)</u> and the <u>population variance ($\sigma^2$)</u>. Often in statistics one uses the <u>standard deviation</u> called <u>(s)</u> for a sample and <u>$\sigma$</u> for a population.

But in studies of populations (and samples of populations) there is another term that is often confused with the standard deviation, namely, <u>standard error</u>. This term in relation to populations (and samples) refers strictly to the standard error of the mean.

In the above example we had a series of estimates of the true population mean ($\mu$) which gives :

Standard deviation of sample means called <u>standard error (s.e.)</u>

$$s.e. = \sqrt{\frac{\sigma^2}{n}}$$

More usually, this is estimated from samples as:

$$s.e. = \sqrt{\frac{s^2}{n}} \qquad or \qquad \frac{s}{\sqrt{n}}$$

This latter term estimates the error in $\bar{x}$ as an estimator of $\mu$. This is usually written as : $\bar{x} \pm$ s.e.

For example if $n = 80$, $\bar{x} = 10.125$, $s^2 = 8.5918$

s.e. = $\sqrt{(8.5918/80)}$

= 0.3277

For the first 8 counts only $\bar{x} = 10.5$, $s^2 = 12.857$

s.e. = 1.268 i.e. 4 x the s.e. of 80 counts.

The standard error thus estimates the precision which the sample mean ($\bar{x}$) is an estimate of the true population mean ($\mu$). Only when interested in this estimate should one use standard error. More often, one is interested in an estimate of the variability in the sample mean and here one should use <u>confidence limits.</u> These are usually written as 95 or 99% confidence limits. The limits show the range within which one can be 95 or 99% certain that the true population mean lies.

We will deal with both Large Sample and Small Sample statistics.

CONFIDENCE LIMITS FOR LARGE SAMPLES (n = > 30).

In a normal distribution 95% of the values lie within 1.96 s.d's of the true population mean. Therefore, 95% of the sample means lie within 1.96 s.e's of the population mean.

i.e. $\bar{x}$ -1.96 s.e. to $\bar{x}$ + 1.96 s.e.

or $\bar{x} - t\sqrt{s^2/n}$ to $\bar{x} + t\sqrt{s^2/n}$

Here we use the statistical table 't' when the population variance ($\sigma^2$) is unknown and is estimated by the sample variance ($s^2$). For 't', degrees of freedom (d.f. = n-1) where n = infinite 't'= 1.96. For d.f. = 30 't'= 2.04.

$\bar{x} = 10.125$, $s^2 = 8.5918$, $n = 80$, 't' for d.f. 79 = 1.99

s.e. = 0.3277

$\bar{x} = 10.125 \pm 0.3277$

95% c.l. = $\bar{x}$ - 1.99(0.3277) to $\bar{x}$ + 1.99(0.3277)

= 10.125 - 0.6251 to 10.125 + 0.6251

= 9.4730 to 10.7770

For total population estimate:

300,000(9.4730)/100 = 28,419

to 300,000(10.7770)/100 = 32,331

We can be 95% certain that the true population mean lies between 28,419 and 32,331.

SMALL SAMPLES (n < 30).

Here we cannot assume that a normal distribution holds. Instead we use an estimate based on the Poisson fraction:

$$\bar{x} \pm \frac{t_{0.05} \quad s}{\sqrt{n-1}}$$

e.g. $\bar{x} = 11.273$, $s^2 = 7.415$, $(s = 2.723)$, $n = 10$,

$t_{0.05}\ (9) = 2.262$

$\bar{x} \quad = \pm\ 2.262\ (2.723)\ /\ \sqrt{(9)}$

$\bar{x} \quad = \pm\ 2.0531$

$= 9.220 \quad \text{to} \quad 13.325$

## Lecture 2: SPATIAL DISPERSION OF POPULATIONS AND TRANSFORMATIONS OF DATA

There are three basic types of dispersal:

1. random

2. regular

3. contagious (or aggregated)

These groupings can overlap, for example a contagious distribution can result from randomly distributed individuals with regularly distributed individuals in each group. The investigation of patch structure and patch size becomes important.

The 3 types of dispersal can be characterized statistically and simply since in:

| | | |
|---|---|---|
| random distributions: | variance ($s^2$) = $\bar{x}$ | - Poisson |
| regular distribution: | $s^2 < \bar{x}$ | - Positive binomial |
| contagious distribution: | $s^2 > \bar{x}$ | - Negative binomial |

The mathematical distributions on the r.h.side are those which can be applied to the respective distributions. In fact the positive and negative binomial models are just one of many types that could be fitted.

1. Random distribution.

Statistically it is not in fact possible to test for this distribution! Yet one of the requirements before being able to do statistical tests is that one has a normal (i.e. random) distribution. One simply has to use a test and say that the hypothesis of randomness is not disproved.

A random distribution results from a) chance effects or b) the influence of a single environmental factor. Usually in nature environmental factors do not affect populations randomly and there are a multiplicity of factors which act in concert. So in nature random distributions are in fact rare.

Randomness is often produced by inefficient sampling (e.g. wrong sample size where quadrat is much larger or smaller than the average size of a patch and the patches are randomly distributed then the population will appear randomly distributed). Similarly, if the population density is low it is extremely hard to detect patches. As populations get older they tend to become more random as patches are split up. Here randomness may apply.

Tests for randomness.

The Index of Dispersion I = $s^2 (n-1) / \bar{x}$

This is tested for significance by reference to the Chi square table.

Example 1: Counts 14, 15, 12, 7, 8, 14, 11, 14, 10, 9, 10

$\bar{x} = 11.273 \quad s^2 = 7.415 \; n = 11$

$s^2 (n - 1) / \bar{x} = 7.415 (10) / 11.273$

$= 6.578$

From chi square table $p_{0.05}(10) = 18.387$

Conclusion: we cannot reject the hypothesis that the counts have a random distribution.

Example 2: Counts 98, 22, 72, 214, 67

$\bar{x} = 94.6 \qquad s^2 = 5202.8 \; n = 5$

$I = 5202.8 (4) / 94.6$

$= 219.99$

Chi square $p_{0.05}(4) = 9.488$

Conclusion: we reject the hypothesis that the counts come from a random distribution. As the variance / mean ratio is much greater than one a contagious distribution occurs.

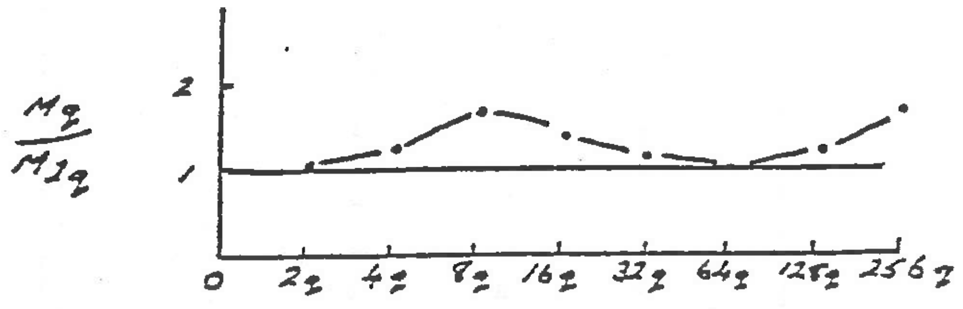
Often one wishes to know the size of the patches. One practical way to do this is to use quadrats of increasing sample size. The maximum variance will be found when patch size and quadrat size are equal.

Morisita's index (M).

Here one must keep doubling quadrat size and then compare the ratio of the index (M) of the smaller to the larger quadrat.

$$M = n (\Sigma(x^2) - \Sigma x) / (\Sigma x)^2 - \Sigma x$$

Ratio = M for quadrat size q / M for quadrat size 2q



In the above example patches are at 8q cm$^2$ with larger patches at 256q cm$^2$.

Many other indices of aggregation have been produced and Elliott (1971) has a good coverage of this topic.

Random distributions are rather rare in nature. Patchiness seems to be the rule. This has consequences for quantitative statistical analyses of most field data as there are a series of rules that data sets must conform to before one can apply statistical techniques. These are that:

1. The data follow a normal distribution

2. The variance of the sample is independent of the mean

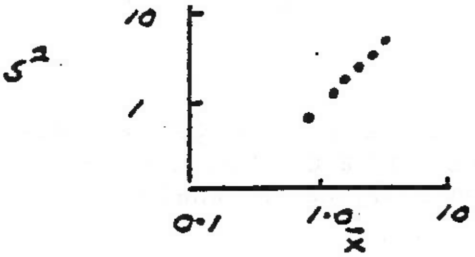
3. Components of the variance should be additive.

For most biological data patchiness is the general pattern so that rule 1 does not apply and usually the variance increases with the mean. Rule 3 is particularly important when applying the analysis of variance, which we will treat later.

One way of overcoming the problems above that the rules are broken is to TRANSFORM the raw data. There are a number of different transformations that are commonly applied such as the square root transformation or log transformation. These will be dealt with more fully under the multivariate analysis section. Probably the most widely used is the log transformation. Often in biological sampling there are frequently zeros in the data set. One cannot take the log of 0 so in such cases a transformation is used called $\log_{10}$ (n + 1), where 1 is added to every number in the set so that when transformed the log of 1 is 0.

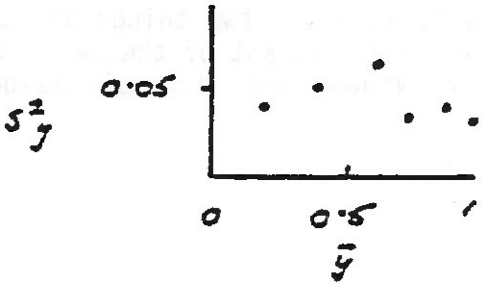
The log transformation usually achieves two things a) it normalizes the data and b) it renders the variance independant of the mean. That this is achieved should be tested, but is rarely done and is simply assumed. Let us examine an example.

| Samples | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| | 0 | 3 | 1 | 6 | 7 | 12 |
| | 2 | 1 | 5 | 1 | 2 | 7 |
| | 1 | 1 | 2 | 5 | 6 | 10 |
| | 0 | 1 | 0 | 7 | 9 | 15 |
| | 0 | 4 | 2 | 4 | 5 | 9 |
| | 1 | 0 | 5 | 1 | 2 | 6 |
| | 1 | 1 | 2 | 6 | 7 | 13 |
| | 0 | 4 | 1 | 5 | 6 | 11 |
| | 1 | 3 | 3 | 3 | 4 | 8 |
| | 1 | 3 | 4 | 3 | 3 | 7 |
| | 0 | 5 | 1 | 5 | 5 | 10 |
| | 2 | 3 | 3 | 3 | 3 | 8 |
| | 1 | 2 | 2 | 4 | 6 | 11 |
| | 0 | 2 | 4 | 3 | 4 | 8 |
| | 0 | 1 | 0 | 8 | 8 | 14 |
| | 2 | 1 | 3 | 4 | 5 | 9 |
| | 3 | 2 | 4 | 2 | 2 | 6 |
| | 0 | 2 | 4 | 2 | 3 | 7 |
| | 1 | 2 | 3 | 4 | 4 | 9 |
| | 1 | 0 | 6 | 2 | 1 | 5 |
| $\bar{x}$ | 0.85 | 2.05 | 2.75 | 3.90 | 4.60 | 9.25 |
| $s^2$ | 0.77 | 1.84 | 2.83 | 3.67 | 4.78 | 7.57 |



Here mean and variance are equal but variance increases with mean (rule 2 broken) so we must transform

| | | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|
| $\text{Log}_{10}$ (n +1) | $\bar{y}$ | 0.22 | 0.43 | 0.53 | 0.65 | 0.71 | 0.99 |
| | $s^2_y$ | 0.04 | 0.05 | 0.06 | 0.03 | 0.03 | 0.02 |



So variance and mean are no longer proportional.

Confidence limits where log transformation is used.

The mean of a transformed count (1$\bar{y}$) = $\Sigma \log x / n$ and the variance $s^2_y$ is calculated in the usual way but on the log-transformed data

95% confidence limits are $\bar{y} \pm t\ s^2_y / n$

But when one back transforms the confidence limits become *and / not + and -.

This gives a GEOMETRIC (or derived) MEAN which is always smaller than the arithmetic mean.

e.g. Counts 98, 22, 72, 214, 67

$\bar{x} = 94.60 \quad s2 = 5202.80$

Log transformed $\bar{y} = 1.8695 \quad s^2_y = 0.1268$

From the 't' table $t_{0.05}(4) = 2.776$

Confidence limits are $\bar{y} = \pm t\ (\sqrt{s^2_y} / n)$

$= 1.8695 \pm 2.776 \ \sqrt{(0.1268/5)}$

$= 1.8695 \pm 0.4419$

$= 1.4276$ to $2.3114$

Antilogs give $\bar{y} = 74.05$ with 95% c.l. 26.77 to 204.83

Alternatively 74.05 */ 2.77 = 27 to 205

---

e.g. $\text{Log}_{10}(\bar{x} + 1)$ Counts 0, 3, 9, 10 $\bar{x} = 5.5$

$\bar{y} = 0.6609 \quad s^2_y = 0.2333 \quad t_{0.05}(3) = 3.182$

$\bar{y} = \pm t \quad (s^2_y / n)$

$= 0.6609 \pm 3.182 \ \sqrt{(0.2333 / 4)}$

$= 0.6609 \pm 0.7685$

$= -0.1076$ to $1.4295$

Geometric mean = antilog $\bar{y}$ -1 = 4.58 - 1 = 3.58

95% c.l. (antilog -0.1076) -1 = < 0

(antilog 1.4295) -1 = 26.88 - 1 = 25.88

Geometric mean = 3.58 with c.l. from 0 to 25.88!

One should note that percentages are not normally distributed and where one has percentages between 0 and 30 or/and 70 and 100 a transformation is necessary. Physiological data often contains percentage data and only rarely have I seen transformations properly employed.

The transformation to be used here is the ANGULAR (OR ARCSINE) TRANSFORMATION.

The transformation involves replacing the percentage (p) by the angle whose sine is p. Tables are to found in most statistical books e.g. Snedecor and Cochran (1967), Rohlf and Sokal (1981).

For example 10% is transformed to 18.44, 15% to 22.79, 30% to 33.21 etc.

## Lecture 3: SAMPLING AND SUB-SAMPLING

Most sampling methods used in marine habitats are aimed to give an estimate of the size of a given population or populations, (e.g. the plankton or benthos where it is not possible to count all individuals). But how does one take the sample? The simplest method is to take a RANDOM sample but is this the most efficient method? The answer is "it is not" but let me try and demonstrate why.

Let us imagine we cover a whole area with 6 samples and find the following counts

| Sample | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|
| Nos. | 1 | 2 | 4 | 6 | 7 | 16 | Total = 36 |

If we draw only 3 samples from this what are our population estimates?

| Samples | Sample<br>Total(T) | Estimate<br>of popn.(x)<br>(T * 2) | Estimate<br>of error<br>(36 - x) |
|---|---|---|---|
| ABC | 7 | 14 | -22 |
| ABD | 9 | 18 | -18 |
| ABE | 10 | 20 | -16 |
| ABF | 19 | 38 | 2 |
| ACD | 11 | 22 | -14 |
| ACE | 12 | 24 | -12 |
| ACF | 21 | 42 | 6 |
| ADE | 14 | 28 | -8 |
| ADF | 23 | 46 | 10 |
| AEF | 24 | 48 | 12 |
| BCD | 12 | 24 | -12 |
| BCE | 13 | 26 | -10 |
| BCF | 22 | 44 | 8 |
| BDE | 15 | 30 | -6 |
| BDF | 24 | 48 | 12 |
| BEF | 25 | 50 | 14 |
| CDE | 17 | 34 | -2 |
| CDF | 26 | 52 | 16 |
| CEF | 27 | 54 | 18 |
| DEF | 29 | 58 | 22 |
| Mean | 18 | 36 | 0 |

As a measure of the accuracy of sampling we can use the MEAN SQUARE ERROR (M.S.E.).

This is M.S.E. $= \Sigma(\text{error estimate})^2$ / sample size (n)

$= (22^2 + 18^2 + 16^2 \ldots\ldots 22^2) / 20$

$= 3504 / 20 = 175.2$

This gives a standard error of $\sqrt{175.2} = 13.2$

i.e 36 ± 13.2 (or 37%)

The sampling plan adopted (3 random samples) is not therefore very efficient.

If we know something about the populations we can improve accuracy. Say we knew that F would give higher values than the other samples. Wherever F occurs in a sample we get much higher values. So the strategy is to divide the area into two STRATA, stratum 1 (S1) with F and stratum 2 (S2) without F. The tactic now is then to always include F but take the other two samples at random.

| Sample | | Sample total in Stratum 2(T2) | Estimate (16 + 2.5 T2) | Error of Estimate |
|---|---|---|---|---|
| S2 | S1 | | | |
| AB | F | 3 | 23.5 | -12.5 |
| AC | F | 5 | 28.5 | - 7.5 |
| AD | F | 7 | 33.5 | - 2.5 |
| AE | F | 8 | 36.0 | 0 |
| BC | F | 6 | 31.0 | - 5.0 |
| BD | F | 8 | 36.0 | 0 |
| BE | F | 9 | 38.5 | 2.5 |
| CD | F | 10 | 41.0 | 5.0 |
| CE | F | 11 | 43.5 | 7.5 |
| DE | F | 13 | 48.5 | 12.5 |
| Mean | | | 36.0 | 0 |

The estimate for S1 total is always correct: 16. The estimate for S2: there are 2 out of 5 samples therefore we multiply T2 by 2.5.

M.S.E. is now 487.5 / 10 = 48.75

standard error = 7.0 (or 19% of total)

So the s.e. is much improved from 37% to 19%. In general STRATIFIED RANDOM sampling is much the preferred strategy. It does require however that one know something about the area or populations being sampled. This implies that one should first do preliminary surveys before setting out on detailed quantitative sampling programmes.

An example showing how one plans a stratified random sampling programme for a benthic survey follows. We plan to sample an area of 200 $m^2$ with a grab taking an area of 0.05 $m^2$. Potentially therefore, there are 200/0.05 = 4,000 sampling units within the area. We do a preliminary survey and find that the bottom is very heterogeneous. Since we know nothing about the benthic fauna we suspect that grain size variations could be important so we map the sediment. Then we want to sample with equal intensity on each type of bottom. This is called PROPORTIONAL ALLOCATION of samples. Let us plan to give an even coverage of 10% to each area i.e. 40 samples total, a not unreasonable number.

We find gravel (n1) covers 1000 sampling units, coarse sand (n2) 500, sand (n3) 1500, fine sand (n4) 800 and mud (n5) 200, totalling 4000 sampling units.

We then allocate our 40 samples in proportion

n1 = 1000 * 40 / 4000 = 10 samples

n2 = 500 * 40 / 4000 = 5 samples

n3 = 1500 * 40 / 4000 = 15 samples

n4 = 800 * 40 / 4000 = 8 samples

n5 = 200 * 40 / 4000 = 2 samples

As to the placement of samples within the area ideally we divide up the whole area give each potential sampling unit a number and pick the numbers from random number tables from a book of statistical tables.

Another method of allocating samples within a stratified random approach is called OPTIMAL ALLOCATION where one takes more samples where there is high variability. As a simple method one can allocate samples according to the variability of the standard error. Let us take an example:

In this example the area was divided up first according to sediment types using methods similar to those shown above and a preliminary sampling done. The total number of animals found in this preliminary survey taking 7 replicates per station were:

| | | Replicates | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Stratum | Station | A | B | C | D | E | F | G |
| 1 | 1 | 1020 | 1180 | 1300 | 2100 | 980 | 900 | 1050 |
| | 2 | 390 | 490 | 210 | 360 | 220 | 310 | 150 |
| | 3 | 140 | 440 | 360 | 150 | 490 | 1070 | 920 |
| 2 | 4 | 140 | 150 | 190 | 160 | 150 | 180 | 140 |
| | 5 | 420 | 950 | 350 | 150 | 180 | 330 | 150 |
| 3 | 6 | 370 | 420 | 700 | 100 | 200 | 190 | 220 |
| | 7 | 620 | 1390 | 380 | 450 | 480 | 2600 | 870 |

| Stratum | Station | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|---|
| 4 | 8 | 390 | 430 | 110 | 440 | 110 | 180 | 160 |
| | 9 | 40 | 20 | 350 | 60 | 80 | 20 | 50 |
| | 10 | 150 | 140 | 660 | 320 | 240 | 880 | 1660 |
| | 11 | 730 | 670 | 470 | 340 | 930 | 370 | 410 |
| 5 | 12 | 1380 | 1410 | 1190 | 2710 | 1600 | 1290 | 530 |
| | 13 | 1620 | 320 | 1550 | 760 | 1250 | 1990 | 270 |
| | 14 | 1850 | 2060 | 1090 | 2410 | 1520 | 220 | 1620 |

Using Statgraphics calculate the means, standard deviation and standard errors for this data. This gives:

| Stratum | Samples size (n) | Mean | s.d | s.e |
|---|---|---|---|---|
| 1 | 21 | 677.62 | 504.62 | 110.12 |
| 2 | 14 | 260.00 | 218.32 | 58.35 |
| 3 | 14 | 642.14 | 654.90 | 175.03 |
| 4 | 21 | 309.05 | 381.59 | 83.27 |
| 5 | 28 | 1162.86 | 686.68 | 129.77 |
| | | | Total | 556.54 |

Calculate each s.e. as a proportion of the Total s.e. and use this proportion to calculate sampling allocation per stratum. Here it is assumed that a total of 65 samples can be taken in the next survey.

| Stratum | Proportion of Total s.e. | No. of samples/stratum |
|---|---|---|
| 1 | 0.198 | 13 |
| 2 | 0.104 | 7 |
| 3 | 0.315 | 20 |
| 4 | 0.149 | 10 |
| 5 | 0.232 | 15 |

We can then calculate how effective this sampling system has been, compared with random sampling. First run an analysis of variance (using Statg) on the data testing within strata variance compared with bewteen strata variance. This gives:

| Source of variation | Sum of Squares | d.f. | Mean Square | F ratio |
|---|---|---|---|---|
| Between strata | 11902293 | 4 | 2975573.3 | 10.27 |
| Within strata | 26931369 | 93 | 289584.6 | |
| Total | 38833662 | 97 | 400347.0 | |

The pooled standard deviation within strata $s_w$ is:

$$s_w = \sqrt{289584.6}$$
$$s_w = 538.13$$

Estimated s.e. of $Y_{st} = s(Y_{st})$

$s(Y_{st}) = s_w / \sqrt{n} \quad n = 98$

$= 538.13/\sqrt{98}$

$= 54.36$

With purely random sampling $S_y = s/\sqrt{n}$

$S_y = \sqrt{400347} / \sqrt{98}$

$= 63.91$

Therefore the stratified sampling reduces the s.e. by

((63.91 - 54.36) * 100) / 63.91 %

= 14.9%

This is a big change and again illustrates the advantages of stratified sampling.

Size of sample.

In general, small sample sizes are better than large ones because:

i) more small units can be taken with the same counting effort.

ii) more samples gives a greater number of degrees of freedom for statistical tests and therefore, a reduction in error.

iii) many small samples in a given area will cover more ground and be more representative than few large samples.

But size reduction must not go too far otherwise edge effects will occur where the population is underestimated due to the disturbance of the edge of the sampler. So a compromise is necessary. Do NOT assume however, that a given grab or plankton net is the appropriate size for a given population simply because it is available. Most grabs were developed as fractions of 1 $m^2$ and this may be a quite inappropriate size for a given population.

Number of samples.

As we have established most species are not randomly or regularly distributed but aggregated. If only a small number of samples is taken from an aggregated population then the population estimate will be highly inaccurate.

One of the simplest methods to determine the number of samples that should be taken is to take 5 samples and calculate the mean and variance, take 5 more and calculate the mean and variance for all ten samples and repeat until the mean and variance are stable. The minimum number if samples where this is achieved is the correct number.

An alternative is to decide on an acceptable error of one's estimate of the population mean and use this in the following equation:

Let us assume that 10% error is acceptable and call this proportion (0.1) D. The number of samples that should be taken (n) is:

$$n = s^2 / (0.1) \bar{x}^2$$

$$= 100 \; s^2 / \bar{x}^2 \text{ for a 10\% accepted error.}$$

Example: counts: 14,15,12,7,8,14,11,14,10,9,10

$s^2 = 7.42 \quad \bar{x} = 11.273$

$n = 100 * 7.42 / (11.273)^2$

$= 5.82$ i.e. 6 samples

For an aggregated distribution:

Counts: 98,22,72,214,67

$s^2 = 5202.8 \quad \bar{x} = 94.60$

$n = 100 * 5202.8 / (94.6)^2$

$= 58$

This is an enormous number of samples and is clearly impractical. So accept a lower error estimate e.g. 20%

$n = 25 * 5202.8 / (94.6)^2$

$= 14.53$

---

Sub-sampling.

Frequently with plankton samples one must sub-sample to reduce the amount of material obtained to reasonable numbers. There are a number of commercial plankton splitters on the market (and indeed there are some for splitting meiobenthos samples). One must test that these samplers are in fact making random splits of the sample.

Say that one has 4 l of concentrated plankton and that he takes 5 x 50ml subsamples and obtains the following data. Is it a random split?

20,25,25,30,40 $\quad s^2 = 57.5 \quad \Sigma x = 140 \; n = 5$

In a random distribution $s^2 / \bar{x} = 1$

We use the Index of Dispersion test $I = s^2 (n-1) / \bar{x}$

$= 57.5 * 4 / 28$

$= 8.2$

Chi square for $p_{0.05}$ 4 d.f. = 9.49

As the value is less than the tabulated one we cannot reject the hypothesis that the sample comes from a random distribution.

We have sampled 5 * 50 ml = 250 from 4000 ml i.e 1/16th

The estimated numbers of animals in the 4000ml is:
$16 * 140 = 2240$

To obtain 95% confidence limits we look up in a table of confidence limits for a Poisson variable (Biometrika 1959 46, 441-453 copy appended)

For 140 we find ± 23
So the population estimates are $23 * 16 = 368$
Giving 2240 ± 368.

If the sub-samples do NOT fit a random distribution then one cannot estimate the numbers in the original sample. In my experience many plankton splitters do not in fact give reliable splits. SO BE WARNED!.

Comparing efficiency of a sampler.

Often one wants to know whether the observed catches of a sampler are equally efficient within acceptable limits.
The $H_0$ is that the samplers are equally efficient.

| Example: | Samplers | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|---|
| | Counts | 6, | 8, | 16, | 5, | 18 | $\Sigma = 53; n = 5$ |

Expected count = 53 / 5 = 10.6

$$Chi^2 = (Observed - Expected)^2 / Expected$$

$$= (6-10.6)^2 /10.6 +..... (18-10.6)^2 / 10.6$$

$$= 13.51$$

d.f. = n-1 = 4, $p_{0.05}$ (4)= 9.49

Conclusion: As the calculated value is greater than the Table value we reject $H_0$ that the samplers are equally efficient. This type of test has many variants and is widely used.

## Lecture 4: COMPARING SAMPLES: 't' TEST, PAIRED 't' TEST AND ANALYSIS OF VARIANCE

Often one wants to compare the variability between two samples. A simple and illustrative test is the Coefficient of Variation (C).

Where $C = s\ (100) / \bar{x}\%$

The coefficient of variation is scaled for differences in mean and is a widely used descriptive parameter.

Another commonly used test is that of comparing two means. The null hypothesis ($H_o$) is that the two means come from the same population and that the means are within the error for that population. It is usual to assume a 5% error due to chance.

In all statistical methods there are two types of error that one can make Type I and Type II errors.

TYPE I ERROR - where one rejects $H_o$ when it is true

TYPE II ERROR - where one accepts $H_o$ when it was false

All statistical tests are prone to both types of error and there is a greater chance of making one type of error than another in each test. The ideal test is one where the probability of rejecting $H_o$ when true is small and the likelihood of rejecting $H_o$ when false is large. Both errors are reduced by increasing the number of degrees of freedom in a test.

Before doing any quantitative statistical test we must make sure that our three primary rules hold. When comparing two means with the 't' test there should be similar variances. If the variances are significantly different then we cannot validly test if the means are significantly different or not.

Here we use the variance ratio test 'F'

where $F = s_1^2 / s_2^2$ where $s_1$ is always the largest of the two.

Let us test $s_1^2 = 8.865$ $n_1 = 60$, $s_2^2 = 7.465$ $n_2 = 80$

$F = 8.855 / 7.465 = 1.11862$

Look up the F ratio table for n-1 = 59 and 79 d.f.

$p_{0.05}\ 60,120 = 1.48$

Conclusion: Since our value is < the table value we cannot reject the hypothesis that the variances come from the same population and so we can test for differences between means using the 't'test.

Student's 't'test.

a) For large samples from normal distributions.

Here 't' = $\bar{x}_1 - \bar{x}_2 /\surd(s_1^2 / n^1 + s_2^2 / n_2)$

e.g. $\bar{x}_1 = 10.125$ $s_1^2 = 7.465$ $n_1 = 80$

$\bar{x}_2 = 12.245$ $s_2^2 = 8.855$ $n_2 = 60$

't' = $12.245 - 10.125 /\surd(7.465/80 + 8.855/60)$

= 4.3194

d.f = $n_1 + n_2 - 2 = 80 + 60 - 2 = 138$

't' $p_{0.05}$ (138) = 1.96

Conclusion: Since our calculated value is greater than the table value we reject the hypothesis that the two means come from the same populations.

Small samples from contagious distributions.

Example:

$\bar{x}_1$ = 4, 5, 8, 14, 14, 15, 15, 19, 28, 36

$\bar{x}_2$ = 2, 4, 5, 7, 12

$\bar{x}_1 = 15.80$ $s_1^2 = 99.07$ $n_1 = 10$

$\bar{x}_2 = 6.00$ $s_2^2 = 14.50$ $n_2 = 5$

Clearly the variance increases with the mean so that we must transform the data. Let us assume that a $\log_n$ transformation is adequate. Now we obtain:

$\bar{y}_1$ = 0.602, 0.699, 0.903, 1.146, 1.146, 1.176, 1.176, 1.279, 1.447, 1.556

$\bar{y}_2$ = 0.301, 0.602, 0.699, 0.845, 1.079.

$\bar{y}_1 = 1.0638$ $s_1^2(y) = 0.2747$

$\bar{y}_2 = 0.7052$ $s_2^2(y) = 0.2887$

Firstly test the variances:

F = 0.2887 / 0.2747 = 1.050

$p_{0.05}$ 9,4 = 8.90

Conclusion: Since the calculated value is less than the table value we conclude that the variances are similar and we can proceed with a 't' test.

$t \quad = 1.113 - 0.705 / \sqrt{(0.0914/10) + (0.0833/5)}$

$\quad = 2.497$

$d.f = n_1 + n_2 - 2 = 13$

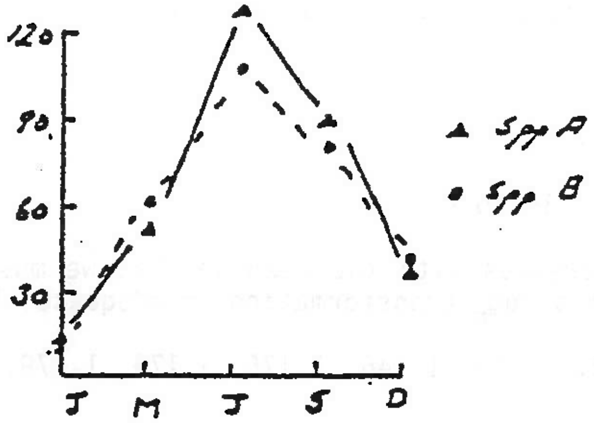
$p_{0.05}(13) = 2.16$

Conclusion: Since the calculated value is greater than the table value at $p_{0.05}$ we conclude that we reject $H_0$ that the means come from the same population and the means are therefore, significantly different.

Making paired comparisons.

Often two sets of data vary over seasons and one is interested not in comparing the overall means but in seeing if there is a significant overall difference, where the null hypothesis is that there is no significant difference between pairs.

The figure illustrates a typical data set.



The data are shown below in tabular form for calculating the Paired 't' test.

| Month | No. of individuals | | Difference | |
|---|---|---|---|---|
| | | | (D) | $D^2$ |
| | Spp A | Spp B | | |
| Jan | 12 | 11 | 1 | 1 |
| Mar | 56 | 63 | -7 | 49 |
| Jun | 125 | 107 | 18 | 324 |
| Sept | 87 | 78 | 9 | 81 |
| Dec | 34 | 36 | -2 | 4 |
| | | Total | 19 | 459 |

$\bar{D} = 19/5 = 3.8$

$$S_D = \surd((\Sigma D^2 - (\Sigma\ D)^2/\ n)\ /\ (n\ -1))$$

$$= \surd((459\ -\ 19^2\ /\ 5)\ /\ 4)$$

$$= 9.83$$

$$\overline{S}_D = S_D\ /\surd n$$

$$= 9.83\ /\surd 5$$

$$= 4.396$$

$$t = \overline{D}\ /\ \overline{S}_D$$

$$= 3.8\ /\ 4.396$$

$$= 0.8644$$

$$p_{0.05}\,(4) = 2.776$$

Conclusion: Since the calculated value is greater than the tabular value we reject $H_0$ that the means come from the same population.

---

Analysis of variance.

More often than not one is interested in comparing more than two means and here one should use the analysis of variance rather than test two and two means by themselves. The analysis of variance (anova) is one of the most used and robust statistical tests devised. But it requires that the data sets comply to the three rules a) samples normally distributed b) variance independent of the mean and c) components of the variance additive. This latter criterion has probably been a bit of a mystery but now all will be revealed in that the ANOVA test breaks down the sources of variance into their components on the assumption that the components are additive.

Example

| Sample | A | B | C | D |
|---|---|---|---|---|
| | 98 | 12 | 86 | 2 |
| | 22 | 13 | 12 | 5 |
| | 72 | 46 | 49 | 12 |
| | 214 | 38 | 33 | 3 |
| | 67 | 49 | 72 | 19 |
| $\bar{x}$ | 94.60 | 31.60 | 50.40 | 8.20 |
| $s^2$ | 5202.80 | 320.30 | 878.30 | 51.70 |

Clearly the variance increases with the mean so that we must transform. Use the $\log_{10}$ transformation.

| Sample | A | B | C | D |
|---|---|---|---|---|
| | 1.991 | 1.079 | 1.935 | 0.301 |
| | 1.342 | 1.114 | 1.079 | 0.699 |
| | 1.857 | 1.663 | 1.690 | 1.079 |
| | 2.330 | 1.580 | 1.519 | 0.477 |
| | 1.826 | 1.690 | 1.857 | 1.277 |
| Total | 9.346 | 7.126 | 8.080 | 3.835 |
| $\bar{x}_2$ | 1.8692 | 1.4252 | 1.6160 | 0.7670 |
| $s^2$ | 0.1268 | 0.0918 | 0.1158 | 0.1163 |

Now the variance is independent of the mean.

Calculate:

1. The Grand Total Y = 9.346 + 7.126....3.835 = 28.387

2. Sum of squared obs. = $1.991^2 + 1.342^2 ... 1.279^2$ = 45.625

3. Sum of squared group totals / n
   = $(9.346^2 + 7.126^2 ... 3.835^2)$ / n
   = 43.626

4. Grand Total squared / Total Sample Size (Correction term)
   = $(28.387)^2$ / 20 = 40.291

5. Sum of squares (Total) = (2) - (CT) =45.625-40.291=5.3340

6. S.S. (Groups) = (3) - (CT) = 43.626 - 40.291 =3.3350

7. S.S. (Within) = SS (Total) - SS (Groups) = 5.3340 - 3.3350
   =1.9990

ANOVA TABLE

| Source of Variation | d.f. | SS | MS | F |
|---|---|---|---|---|
| Between groups | 3 | 3.3350 | 1.1117 | 8.9007 |
| Within groups | 16 | 1.9990 | 0.1249 | |
| Total | 19 | 5.3340 | | |

$p_{0.05}$ 3,16 = 3.24

$p_{0.01}$ 3,16 = 5.29

Conclusion: Since the calculated value for F (8.9007) is greater than the table value for $p_{0.01}$ we conclude that $H_o$ must be rejected and there are significant differences between samples.

This design can be easily extended to a TWO or THREE-WAY ANOVA. Both are easy to calculate but take time. To illustrate the point let us use data on the oxygen consumption of two species of limpet in three concentrations of sea-water. The question is there a significant difference in oxygen consumption between species and how does this vary with salinity.

Oxygen figures are in $\mu$g $O_2$ /mg dry wt./min @ 22°C, n = 8.

FACTOR A SPECIES

| FACTOR B SEAWATER | *Acmaea scabra* | | *Acmaea digitalis* | | Total |
|---|---|---|---|---|---|
| 100% | 7.16 | 8.26 | 6.14 | 6.14 | |
| | 6.78 | 14.00 | 3.86 | 10.00 | |
| | 13.60 | 16.10 | 10.40 | 11.60 | |
| | 8.93 | 9.66 | 5.49 | 5.80 | |
| | 84.49 | | 59.43 | | 143.92 |
| 75% | 5.20 | 13.20 | 4.47 | 4.95 | |
| | 5.20 | 8.39 | 9.90 | 6.49 | |
| | 7.18 | 10.40 | 5.75 | 5.44 | |
| | 6.37 | 7.18 | 11.80 | 9.90 | |
| | 63.12 | | 58.70 | | 121.82 |
| 50% | 11.11 | 10.50 | 9.63 | 14.50 | |
| | 9.74 | 14.60 | 6.38 | 10.20 | |
| | 18.80 | 11.10 | 13.40 | 17.70 | |
| | 9.74 | 11.80 | 14.50 | 12.30 | |
| | 97.39 | | 98.61 | | 196.00 |
| | 245.00 | | 216.74 | | 461.74 |

1) Grand Total = 461.74

2) Sum of obs.squared = $(7.16)^2 + (6.78)^2 \ldots (12.30)^2 = 5065.1530$

3) Sum of squared group totals / sample size of groups
$= ((84.49)^2 + (59.43)^2 \ldots . (98.61)^2) / 8 = 4663.6317$

4) Sum of squared column totals / sample size of column
$= ((245.00)^2 + (216.74)^2) / 24 = 4458.3844$

5) Sum of squared row totals / sample size of row
$= ((143.92)^2 + (121.82)^2 + (196.00)^2) / 16 = 4623.0674$

6) Grand Total squared / total sample size = C.T.
$= (461.74)^2 / 48 = 4441.7464$

7) $SS_{total}$ (2) - (6) = 5065.1530 - 4441.7464 = 623.4066

8) $SS_{subgroupd}$(3) - (6) = 4663.6317 - 4441.7464 = 221.8853

9) $SS_A$ (4) - (6) = 4458.3844 - 4441.7464 = 16.6380

10) $SS_B$ (5) - (6) = 4623.0674 - 4441.7464 = 181.3210

11) $SS_{AB}$ (8)-(9)-(6) = 2221.8853-16.6380-181.3210 = 23.9263

12) $SS_{error}$ = (7) - (8) = 623.4066 - 221.8853 = 401.5213

| Source of variation | df | SS | MS | F |
|---|---|---|---|---|
| Subgroups | 5 | 221.8853 | 44.377 | |
| Between species (A) | 1 | 16.6380 | 16.638 | 1.74 n.s |
| Between salinities(B) | 2 | 181.3210 | 90.660 | 9.48 *** |
| Species X salinities (AB) | 2 | 23.9263 | 11.963 | 1.25 n.s |
| Error (within group) | 42 | 401.5213 | 9.560 | |
| Total | 47 | 623.4066 | | |

F 0.05 $_{1,42}$ = 4.07, $_{2,42}$ 0= 3.22, 0.001 $_{2,42}$ = 8.18

Conclusion: oxygen consumption does not differ significantly between species, but does between salinities for both species.

This test does <u>not</u> however, tell which means are significantly different from each other. One can use the <u>Least Significant Range</u> (LSR) test.

LSR = Q $0.05_{(df)}$ (MS within / n)

Q is a factor from Tables called the STUDENTIZED RANGE

Q 0.05 $_{(2,42)}$ = 2.858

LSR = 2.858 (9.560 / 16)

= 2.2092

Arrange means in ascending (or descending) order

| | 100% | 75% | 50% |
|---|---|---|---|
| $\bar{x}$ | 71.96 | 60.91 | 98.00 |

As all the difference between means are greater than the calculated significance level of 2.209 we conclude that all are significantly different.

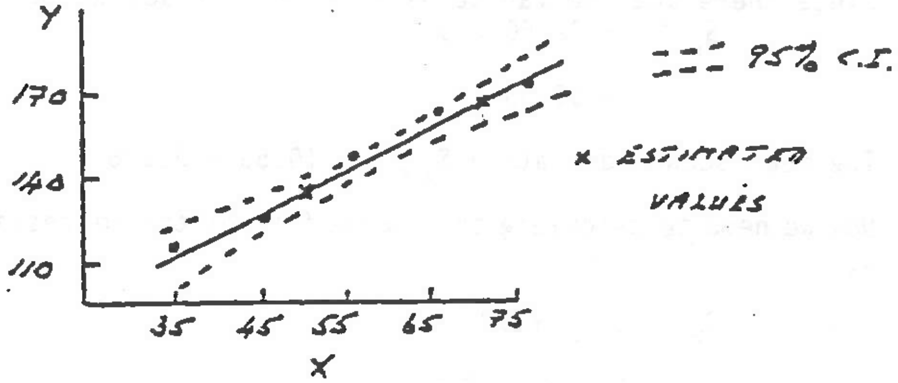
## Lecture 5: REGRESSION AND CORRELATION ANALYSES

In regression analysis we relate one variable, the dependent (Y) to another, the independent (X). Such techniques are used when one wishes to know

a) if Y depends on X
b) to predict Y knowing X
c) the shape of the relationship between Y and X

Example: Linear Regression analysis

|  |  | X<br>Age<br>(months) |  | Y<br>Length |
|---|---|---|---|---|
|  |  | 35 |  | 114 |
|  |  | 45 |  | 124 |
|  |  | 55 |  | 143 |
|  |  | 65 |  | 158 |
|  |  | 75 |  | 166 |
| Total |  | 275 |  | 705 |
|  | $\bar{x}$ | 55 | $\bar{y}$ | 141 |



$\Sigma X^2 = 16125 \qquad \Sigma Y^2 = 101341 \qquad \Sigma XY = 40155$

$(\Sigma X)^2 / n = 15125; \quad (\Sigma Y)^2 / n = 99405; \quad (\Sigma X)(\Sigma Y) / n = 38755$

$\Sigma x^2 = (\Sigma X)^2 - (\Sigma X)^2/n = 16125 - 15125 = 1000$

$\Sigma y^2 = (\Sigma Y)^2 - (\Sigma Y)^2/n = 101341 - 99405 = 1936$

$\Sigma xy = \Sigma XY - (\Sigma X)(\Sigma Y)/n = 40155 - 38755 = 1380$

$b = \Sigma xy / \Sigma x^2 = 1380 / 1000 = 1.38$

$\hat{Y} = \overline{Y} + b(X-\bar{x}) = 141 + 1.38(X - 55)$

$= 65.1 + 1.38x$

Substitute two values in the equation

Let $X = 50$, $Y = 65.1 + 1.38(50)$
$= 134.1$

Let $X = 70$, $Y = 65.1 + 1.38(70)$
$= 161.7$

But how good a fit is this line to the data points?

Here we will calculate the 95% confidence intervals for the line.

Firstly, calculate the deviations from the line

$d_{y.x} \quad = Y - \overline{Y}$

For all points this is:

$$d_{y.x} = \Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2$$

$$= 1936 - (1380)^2 / 1000$$

$$= 31.60$$

Since there are two variables X and Y this deviation has n-2 d.f. = 3.

$$S_{y.x}^2 = 31.60 / 3$$

$$= 10.53$$

The mean square deviation $S_{y.x} = \sqrt{10.53} = 3.245$

Now we need to calculate the sample S.D. of the regression coefficient b, i.e. $S_b$

$$S_b = S_{y.x} / \sqrt{\Sigma x^2}$$

$$= 3.245 / \sqrt{1000}$$

$$= 0.1026$$

The significance of the regression can be tested using a 't' test in the usual way where $t = b / S_b$

$$= 1.38 / 0.1026$$

$$= 13.5$$

$p_{0.05}$ (3) = 3.182

Conclusion: Since the calculated value is greater than the table value we conclude that since the regression coefficient is significantly greater than the error term there is a significant fit to the data.

Now we must calculate the confidence limits:

For b 95% = $t0.05_{(3)}\ S_b$

$$= 3.182 (0.1026)$$

$$= 0.3265$$

$$b = 1.38 \pm 0.3265$$

The s.e. of Y = $S_{yx} \surd((1/n) + (x^2 / \Sigma x^2))$

$= 3.245 \surd(1/5 + x^2 / 1000)$

$= \surd 10.30 \surd 0.2 + 0.001x^2$

$= 2.06 + 0.0103 x^2$

---

For X = 35 $x = X - \bar{x} = 35 - 55 = -20$

$Sy = \surd(2.06 + 0.0103 (-20)^2)$

$= 2.4859$

---

For X = 55, x = 55-55 = 0

$Sy = \surd(2.06 + 0.0103 (0)^2)$

$= 1.4352$

---

For X = 75, x = 75-55 = 20

$Sy = \surd(2.06 + 0.0103 (20)^2)$

$Sy = 2.4859$

---

95% confidence limits are:

Y - t0.05 Sy to Y +t0.05 Sy

X = 35

$\hat{Y}$ = Y + bx
= 141 + 1.38(-20)
= 113.4

t0.05 Sy = 3.182 (2.4869)
= 7.913
= 113.4 ± 7.913

---

X = 55

$\hat{Y}$ = Y + bx
= 141 + 1.38 (0)
= 141

t0.05 Sy = 3.182(1.4352)
= 4.567
= 141 ± 4.567

X = 75
$\hat{Y}$ = Y + bx
= 141 + 1.38 (20)
= 168.6
= 168.6 ± 7.913

So the confidence limits are asymmetrical, with greater chance of error the further one is from the mean value.

Curvilinear regressions.

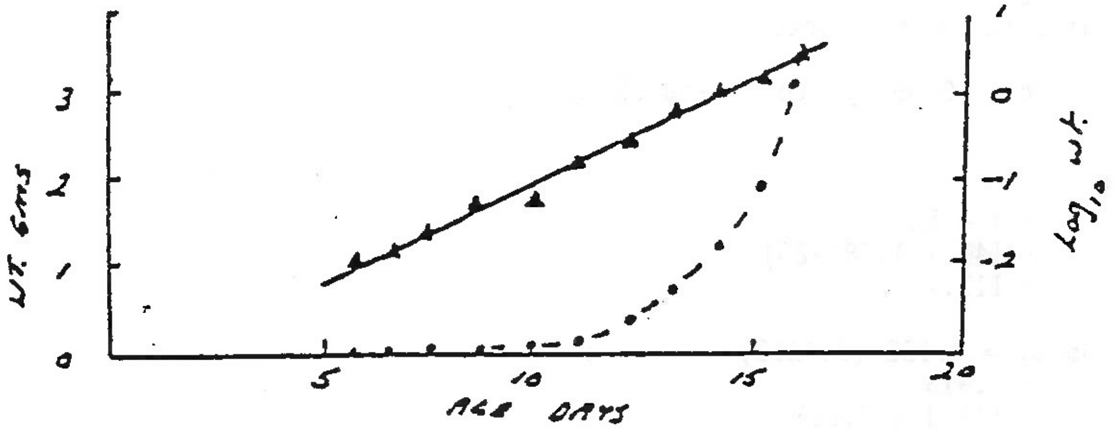
For many biological phenomena the relationship between the two variables is not linear but curvilinear. For example the initial phases of growth of populations of bacteria doubling in size at each time interval. Or data on weight increases over time. The regression now is:

Weight (W) = (A)($B^x$), where A and B are constants.

One can use a log transformation to obtain a linear relationship where

LogW = logA + (logB)X

The plots are shown below.



The data are:

| Age in days<br>X | Dry weight W<br>(gm) | $\text{Log}_{10}$ W<br>Y |
|---|---|---|
| 6 | 0.029 | -1.538 |
| 7 | 0.052 | -1.284 |
| 8 | 0.079 | -1.102 |
| 9 | 0.125 | -0.903 |
| 10 | 0.181 | -0.742 |
| 11 | 0.261 | -0.583 |
| 12 | 0.425 | -0.372 |
| 13 | 0.738 | -0.132 |
| 14 | 1.130 | 0.053 |
| 15 | 1.882 | 0.275 |
| 16 | 2.812 | 0.449 |

The calculations are as in the example for the linear regression using here X and Y, the log transformed data.

We obtain $Y = 0.1959X - 2.689$

One of the most interesting parameters from a relationship such as that above is the relative rate of increase

$$W = Ae^{cX}$$

where e = 2.718 the base of the natural logarithm series.

This gives us $\text{Log}_e W = (\log_{10} W)(\log_e 10)$

$$= 2.3026 \log_{10} W$$

In terms of the equation above we get:

$$W = (2.3026)(0.1959)$$

$$= 0.451 \text{ gm per day per gm}$$

We must also backtransform the equation

$$\text{Log } W = 0.159X - 2.689$$

antilog -2.689 = antilog (0.311 - 3) = 0.00205

Giving: $W = (0.00205)e^{0.451X}$

---

Correlation

This is related to regression analysis and shows the degree to which two variables are correlated together. Here there is no dependent and independent variable rather $X_1$ and $X_2$ instead of X and Y.

Example

Brothers height $X_1$ 71 68 66 67 70 71 70 73 72 65 66 $\bar{X}_1 = 69$

Sisters height $X_2$ 69 64 65 63 65 62 65 64 66 59 62 $\bar{X}_2 = 64$

$n = 11$, $\Sigma x_1^2 = 74$, $\Sigma x_2^2 = 66$, $\Sigma x_1 x_2 = 39$

Remember:

$$\Sigma x_{1,2} = \Sigma X_1^2 - (\Sigma X_1)^2 / n \quad \text{etc}$$

$$r = \Sigma x_1 x_2 / \sqrt{(\Sigma x_1^2)(\Sigma x_2^2)}$$

$$= 39 / \sqrt{(74)(66)}$$

$$= 0.558$$

d.f. $= n-2 = 11-2 = 9$ $p_{0.05}(9) = 0.602$

Conclusion: Since the value is less than the table value we cannot reject the hypothesis that there is no correlation between brothers height and sisters height.

## Lecture 6: NON-PARAMETRIC METHODS

So far we have been dealing with analysis of quantitative data. I believe that as far as possible one should use the parametric methods discussed before. There is a tendency within biology today to use the argument that we do not know if such and such a data set is normally distributed nor do we know the relationship between the variance and mean, particularly with small samples with which we usually work. So the decision is taken almost exclusively to use non-parametric statistics as these do not require distributions to be normal, nor variance be independant of the mean. I believe that this is often a mistaken belief in that one misses much of the potential of the data that can only be revealed by a proper parametric analysis.

It must be said, however, that non-parametric methods are in some cases as efficient as parametric ones, save when the analyses are complex such as ANOVA and regression analyses.

Let us examine a few of the most commonly used:

### A non-parametric 't' test the MANN-WHITNEY 'U' test.

This is a highly useful alternative to the 't' test and with almost equal precision. The null hypothesis is that two independant random samples come from the same populations having the same parent distribution and the same means.

Example: Sample 1 $n_1 = 5$ Counts 2,4,5,7,12

Sample 2 $n_2 = 10$ Counts 4,5,8,14,14,15,15,19,28,36

The counts are arranged in rank order from lowest to highest.
Now substitute ranks for each count giving an average rank for equal numbers.

This gives:

$n_1 = 1,\ 2.5,\ 4.5,\ 6,\ 8$

$n_2 = 2.5,\ 4.5,\ 7,\ 9.5,\ 9.5,\ 11.5,\ 11.5,\ 13,\ 14,\ 15$

Sum the ranks for each sample:

$R_1 = 22$

$R_2 = 98$ Check that $R_1 + R_2 = (n_1 + n_2)(n_1 + n_2 + 1)/2$

Calculate test statistics $U_1$ and $U_2$:

$U_1 = n_1n_2 + (n_2(n_2 + 1)/2) - R_2$

$U_2 = n_1n_2 + (n_1(n_1 + 1)/2) - R_1$

$U_1 = 50 + (110/2) - 98 = 7$

$U_2 = 50 + (30/2) - 22 = 43$ Check $U_1 + U_2 = 50 = n_1n_2$

Refer to Table of U for $n_1 = 5$ $n_2 = 10$ at $p_{0.05} = 8$

Conclusion: Here we use an unusual decision technique in that the calculated value must be SMALLER than the table value to be significant.

We find that the smallest (U) value = 7 and since this is smaller than the table value 8, we reject the hypothesis that the two means come from the same population and can conclude that the mean of sample 2 is significantly higher than that of sample 1.

Non-parametric ANOVA: Kruskall-Wallis test.

Here the null hypothesis is that the means come from the same population.

Example: Let us take the same data as used in the parametric test. Firstly, we arrange the data in ascending order within samples.

| | | |
|---|---|---|
| Samples | stn. 1) | 98, 22, 72, 214, 67 |
| | stn. 2) | 12, 13, 46, 38, 49 |
| | stn. 3) | 86, 12, 49, 33, 72 |
| | stn. 4) | 2, 5, 12, 3, 19 |

Now rank all the above in ascending order:

| | | Total | $n_i$ | $(R_i^2/n_i)$ |
|---|---|---|---|---|
| stn. 1) 19, 9, 16.5, 20, 15 | $R_1$ | 79.5 | 5 | 1264.05 |
| stn. 2) 5, 7, 12, 11, 13.5 | $R_2$ | 48.5 | 5 | 470.45 |
| stn. 3) 18, 5, 13.5, 10, 16.5 | $R_3$ | 63 | 5 | 793.80 |
| stn. 4) 1, 3, 5, 2, 8 | $R_4$ | 19 | 5 | 72.20 |
| Total | | R= 210 | N= 20 | 2600.50 |

Calculate K statistic

$$K = (12 / (N(N + 1))(\Sigma(R_i)^2 / n_i) - 3(N + 1)$$

$$= (12 / 20(21))(2600.5) - 3(21)$$

$$= 11.3$$

Refer to Tables of $chi^2$ for v = i - 1 d.f. = 3
$p_{0.05} = 7.81$; $p_{0.01} = 11.2$

Conclusion: Since the calculated value is greater than the table value at p0.01 we reject the hypothesis (with 99% certainty) that the means are from the same population.

## Non-parametric Two-Way ANOVA: The Friedman test.

This test is used in COMBINATION with the Kruskal-Wallis test under defined conditions. The number of counts in each sample must be the same and each count must belong to one sample and one group, where the group can represent different bottom types, different samplers or different workers. In the previous example let us assume that the four samples were each counted by five different workers and we are therefore, interested to know if there is also a significant difference between workers. The $H_0$ is that there is no significant difference between workers.

Example: 4 sample i = 4, 5 workers n = 5

| | | | Worker | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Sample 1 | 98 | 22 | 72 | 214 | 67 |
| Sample 2 | 12 | 13 | 46 | 38 | 49 |
| Sample 3 | 86 | 12 | 49 | 33 | 72 |
| Sample 4 | 2 | 5 | 12 | 3 | 19 |

Firstly arrange samples by rank in each row:

| | | | | | |
|---|---|---|---|---|---|
| Sample 1 | 4 | 1 | 3 | 5 | 2 |
| Sample 2 | 1 | 2 | 4 | 3 | 5 |
| Sample 3 | 5 | 1 | 3 | 2 | 4 |
| Sample 4 | 1 | 3 | 4 | 2 | 5 |
| Total | 11 | 7 | 14 | 12 | 16 |
| | $R_1$ | $R_2$ | $R_3$ | $R_4$ | $R_5$ |

$\Sigma R_n = 60 \qquad \Sigma R_n^2 = 766$

$$S = \Sigma R_n^2 - ((\Sigma R_n)^2/n)$$

$$= 766 - (60^2)/5 = 46$$

n = 5, i = 4

So we must calculate a chi$^2$ value where:

$$chi^2 = 12S/ (in(n + 1))$$

$$= 12(46) /4*5(6)$$

$$= 4.6$$

chi$^2$ for v = n-1 d.f. = 4 for $p_{0.05}$ = 9.49

Conclusion: As the calculated value is less than that of the Table value we cannot reject the null hypothesis that there was no significant difference between workers.

## Non-parametric Correlation tests.

### 1. Spearman's rank correlation.

Here we first simply rank the two sets of data.

| Example: | Sample 1 | Sample 2 | $(\text{Difference})^2$ |
|---|---|---|---|
| | 4 | 4 | 0 |
| | 1 | 2 | 1 |
| | 6 | 5 | 1 |
| | 5 | 6 | 1 |
| | 3 | 1 | 4 |
| | 2 | 3 | 1 |
| | 7 | 7 | 0 |
| | | Total | 8 |

Calculate $r_s$ $= 1- (6\Sigma d^2) / (n(n^2 - 1)$

$= 1 - 6(8) / 7(49-1)$

$= 0.857$

This value is tested against the Table value for the Correlation Coefficient (r). d.f. = n - 1 = 6 $p_{0.05} = 0.707$.

Conclusion: As the calculated value is greater than the Table value we reject the $H_0$ that there is no correlation between the two data sets.

This test is calculated to be 90% as efficient as the parametric test.

### 2. Kendall's tau.

Here the two ranks are set alongside each other:

| Sample 1 | Sample 2 | No. of ranks < pivotal rank |
|---|---|---|
| 1 | 2 | 1 |
| 2 | 3 | 1 |
| 3 | 1 | 0 |
| 4 | 4 | 0 |
| 5 | 6 | 1 |
| 6 | 5 | 0 |
| 7 | 7 | 0 |
| | Total (Q) | 3 |

$$\text{Tau} = 1 - (4(Q)) / (n(n - 1))$$
$$= 1 - (4*3 / 42)$$
$$= 0.714$$

We look up the Table of the Correlation Coefficient (r) for d.f. 6 and $p_{0.05}$ = 0.707

Conclusion: Since the calculated value is greater than the Table value we reject $H_0$ that there is no correlation between the two data sets.

References:

**General texts.** <u>One</u> of the following three is required:

Snedecor, G.W. and W.G. Cochran (1980), <u>Statistical Methods</u>. 7th ed. Iowa State Univ. Press, Ames, Iowa 507 pp.

Sokal, R.R. and F.J. Rohlf (1981), <u>BIOMETRY</u>, 2nd ed. W.H. Freeman & Co., San Francisco, 859 pp.

Zar, J.R. (1984), <u>Biostatistical Analysis</u>, Prentice-Hall.

**Statistical tables: One set is needed**

Snedecor & Cochran and Zar include tables in their books. Rohlf and Sokal has separate (and expensive) tables.

Rohlf, F.J. and R.R. Sokal (1984), Statistical Tables. W.H. Freeman

Neave, H.R. (1978), Statistical Tables Allen & Unwin. (Cheap!)

**Excellent book on analysis of field samples:**

Elliott, J.M. (1971), Statistical Analysis of samples of Benthic Invertebrates. Fresh Water Biological Association, Windermere, U.K.

# PART II

# LECTURE 1

## A FRAMEWORK FOR STUDYING CHANGES IN COMMUNITY STRUCTURE

### STAGES

1) REPRESENTING COMMUNITIES (graphical description of faunal relations).

2) DISCRIMINATING SITES on the basis of faunal composition (e.g. spatial: control v. impacted, temporal: before v. after impact).

3) DETERMINING LEVELS OF "STRESS" or disturbance in communities.

4) LINKING WITH ENVIRONMENTAL VARIABLES (e.g. correlating to contaminants)

5) ESTABLISHING CAUSALITY of link to contaminants.

### TECHNIQUES

UNIVARIATE - diversity indices
- indicator species abundance

DISTRIBUTIONAL - "ABC" curves (k-dominance)
- distn. of individuals amongst species

MULTIVARIATE - triangular matrix of similarities between samples, leading to:
- hierarchical classification (CLUSTER)
- multidimensional scaling (MDS)
- principal component analysis (PCA)

### UNIVARIATE TECHNIQUES

| STAGES | | EXAMPLES | |
|---|---|---|---|
| | | Diversity indices | Indicator species |
| 1) | REPRESENTING COMMUNITIES | Means ± confidence intervals (CIs for each site) | |
| 2) | DISCRIMINATING SITES | One-way analysis of variance (ANOVA) | |
| 3) | DETERMINING STRESS LEVELS | By reference to historical data, e.g. | |
| | | ultimately a decrease in diversity | initial increase in "opportunist" species |
| 4) | LINKING TO ENVIRONMENT | Regression techniques | |

5) ESTABLISHING CAUSALITY — Mesocosm or field <u>experiments</u> with controlled dosing of contaminants. All entries above apply, e.g. now significant discrimination of "sites" (=treatments) demonstrates that contaminant <u>causes</u> biological effect.

## IOC/GEEP WORKSHOP ON BIOLOGICAL EFFECTS OF POLLUTANTS

OSLO 1986: MACROFAUNAL DATA, Gray et al. (1988)

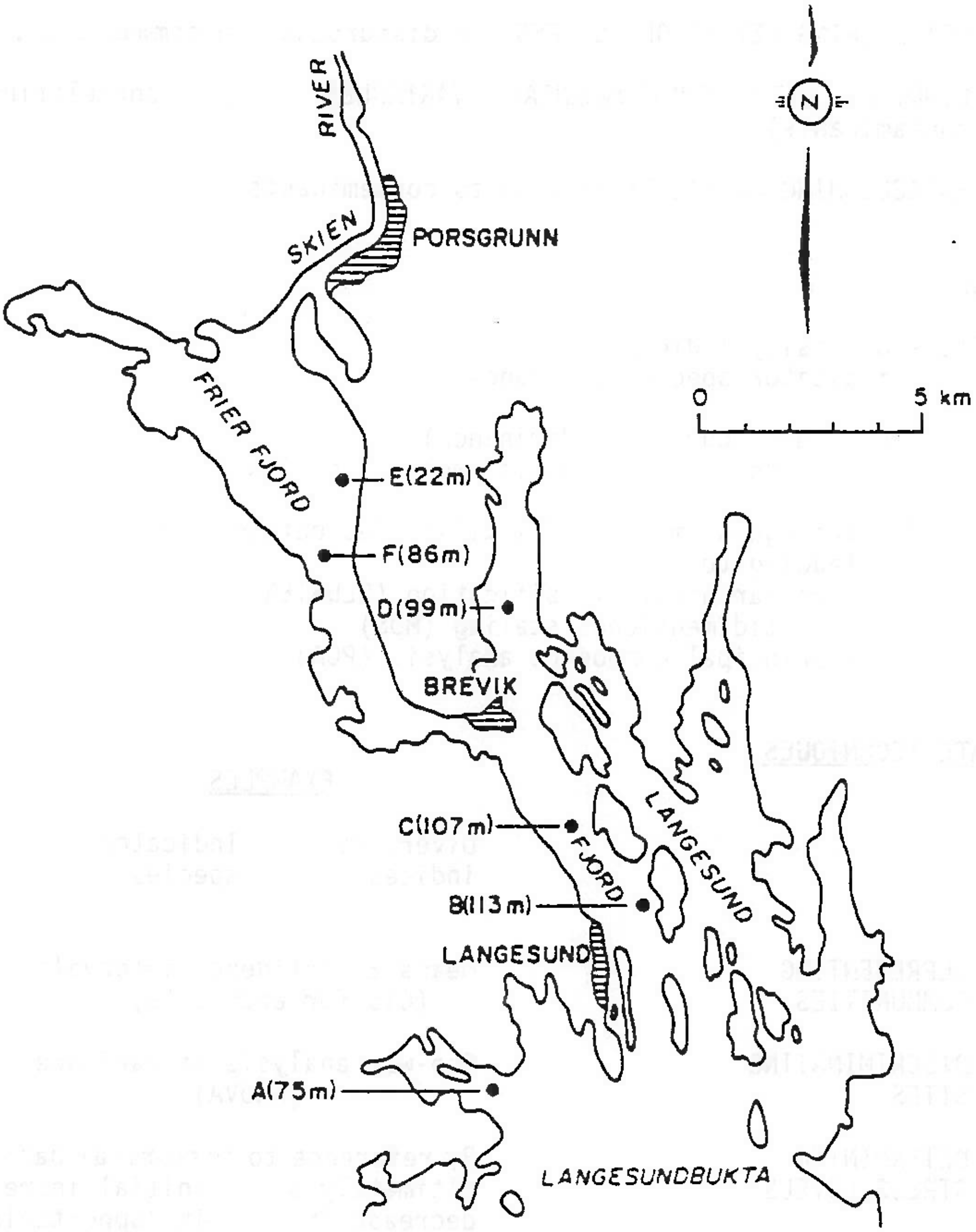


Fig. 1.1 Frierfjord and Langesundfjor, Norway. Benthic community sampling sites (A-G) for the Oslo Workshop

Four 0.1 m$^2$ Day grab samples taken at 6 sites (A-E, G), sieved at 1 mm, and counts/biomass recorded of 110 species identified.

Table 1.1

Macrofaunal abundance matrix (part), numbers per 0.1 m$^2$.

| Species | A | | | | B | | | |
|---|---|---|---|---|---|---|---|---|
| *Cerianthus lloydi* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Halicryptus sp.* | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| *Onchnesoma* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Phascolion strombi* | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| *Golfingia sp.* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Holothuroidea | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nemertina, indet. | 12 | 6 | 8 | 6 | 40 | 6 | 19 | 7 |
| Polychaeta, indet. | 5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| *Amaena trilobata* | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| *Amphicteis gunneri* | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 |
| Ampharetidae | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| *Anaitides groenlandica* | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| *Anaitides sp.* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 1.2

Macrofaunal biomass matrix (part), mg per 0.1 m$^2$.

| Species | A | | | | B | | | |
|---|---|---|---|---|---|---|---|---|
| *Cerianthus lloydi*/10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Halicryptus sp.* | 0 | 0 | 0 | 26 | 0 | 0 | 0 | 0 |
| *Onchnesoma* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Phascolion strombi* | 0 | 0 | 0 | 6 | 0 | 0 | 2 | 0 |
| *Golfingia sp.* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | C |
| Holothuroidea | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nemertina, indet./10 | 1 | 41 | 391 | 1 | 5 | 1 | 2 | 1 |
| Polychaeta, indet. | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Amaena trilobata* | 144 | 14 | 234 | 0 | 0 | 0 | 0 | 0 |
| *Amphicteis gunneri* | 0 | 0 | 0 | 0 | 45 | 0 | 0 | 0 |
| Ampharetidae | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *Anaitides groenlandica* | 0 | 0 | 0 | 7 | 11 | 0 | 0 | 0 |
| *Anaitides sp.* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

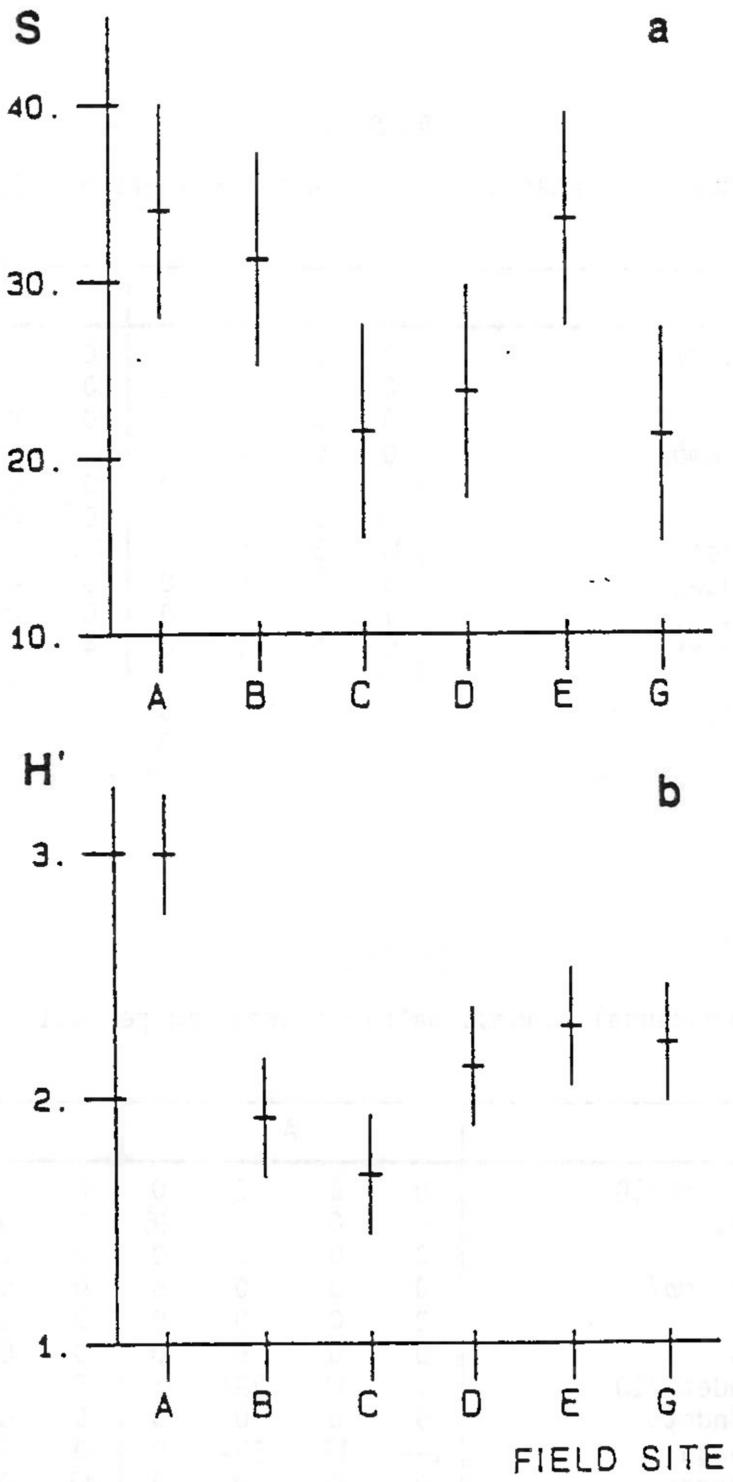


Fig. 1.2 Frierfjord macrofauna. Means and 95% confidence intervals for two indices. a) Number of species (S); b) Shannon diversity (H')

DISTRIBUTIONAL TECHNIQUES

| STAGES | | EXAMPLES | |
|---|---|---|---|
| | | "ABC" curves (k-dominance curves) | Distribution of individuals amongst species |
| 1) | REPRESENTING COMMUNITIES | Curves for each site (or preferably replicate) | |
| 2) | DISCRIMINATING SITES | ANOSIM (Analysis of Similarities) test on "distances" between every pair of curves | |
| 3) | DETERMINING STRESS LEVELS | Biomass curve drops below numbers curve when subject to disturbance | Species abundance distribution is less "smooth" with disturbance |
| 4) | LINKING TO ENVIRONMENT | Possible for univariate summary statistics by regression | |
| 5) | ESTABLISHING CAUSALITY | Mesocosm or field dosing experiments. Entries above apply | |

ORGANIC ENRICHMENT OF BENTHOS - Pearson (1975)

LOCH LINNHE (SCOTLAND) MACROFAUNA - discharges started in 1966, increased 1970, decreased 1972.

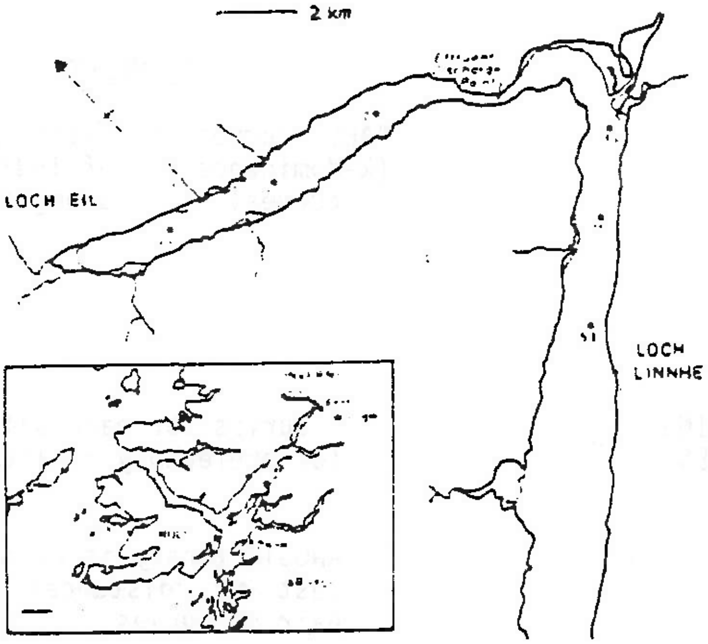


Fig. 1.3 Loch Linnhe and Loch Eil, showing site 34, sampled over 1963-1973

Table 1.3

Numbers/biomass matrix (part) for site 34 - one (pooled) set of values per year (1963-1973).

| Species | 1963 | | 1964 | | 1965 | | 1966 | |
|---|---|---|---|---|---|---|---|---|
| | No. | Wt. | No. | Wt. | No. | Wt. | No. | Wt. |
| Mollusca | | | | | | | | |
| *Scutopus ventrolineatus* Salvini-Plawen | - | - | - | - | 11 | 0.05 | - | - |
| *Nucula tenuis* (Montagu) | 2 | 0.01 | 13 | 0.07 | 16 | 0.10 | 6 | 0.064 |
| *Mytilus edulis* L. | - | - | - | - | 5 | 0.09 | - | - |
| *Modiolus sp.* indet. | - | - | - | - | - | - | - | - |
| *Thyasira flexuosa* (Montagu) | 93 | 3.57 | 210 | 7.98 | 28 | 1.06 | 137 | 5.17 |
| *Myrtea spinifera* (Montagu) | 214 | 27.39 | 136 | 17.41 | 2 | 0.26 | 282 | 36.10 |
| *Lucinoma borealis* (L.) | 12 | 0.39 | 26 | 1.72 | - | - | 22 | 0.73 |
| *Montacuta ferruginosa* (Montagu) | 1 | 0.00 | - | - | 4 | 0.02 | - | - |
| *Mysella bidentata* (Montagu) | - | - | - | - | - | - | - | - |
| *Abra sp.* indet. | - | - | - | - | 12 | 0.26 | - | - |
| *Corbula gibba* (Olivi) | 2 | 0.13 | 8 | 0.54 | 9 | 0.27 | 2 | 0.13 |
| *Thracia sp.* indet. | - | - | - | - | - | - | - | - |

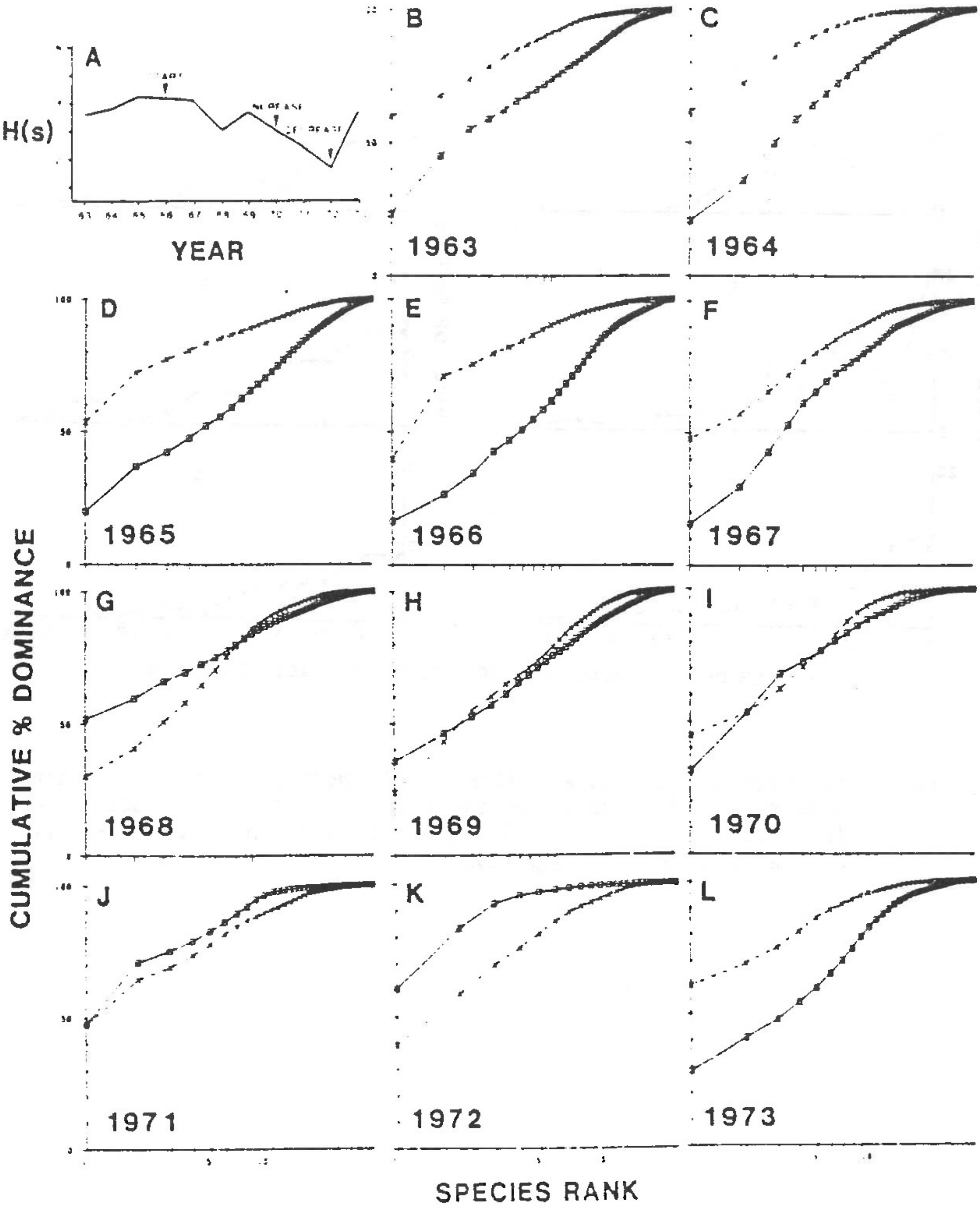


Fig. 1.4 Loch Linnhe site 34. (A) Shannon diversity. (B)-(L) ABC curves for 1963-73: biomass (x), numbers (□). Warwick (1986)

DISTRIBUTIONAL: REPRESENTATION AND STRESS DETERMINATION

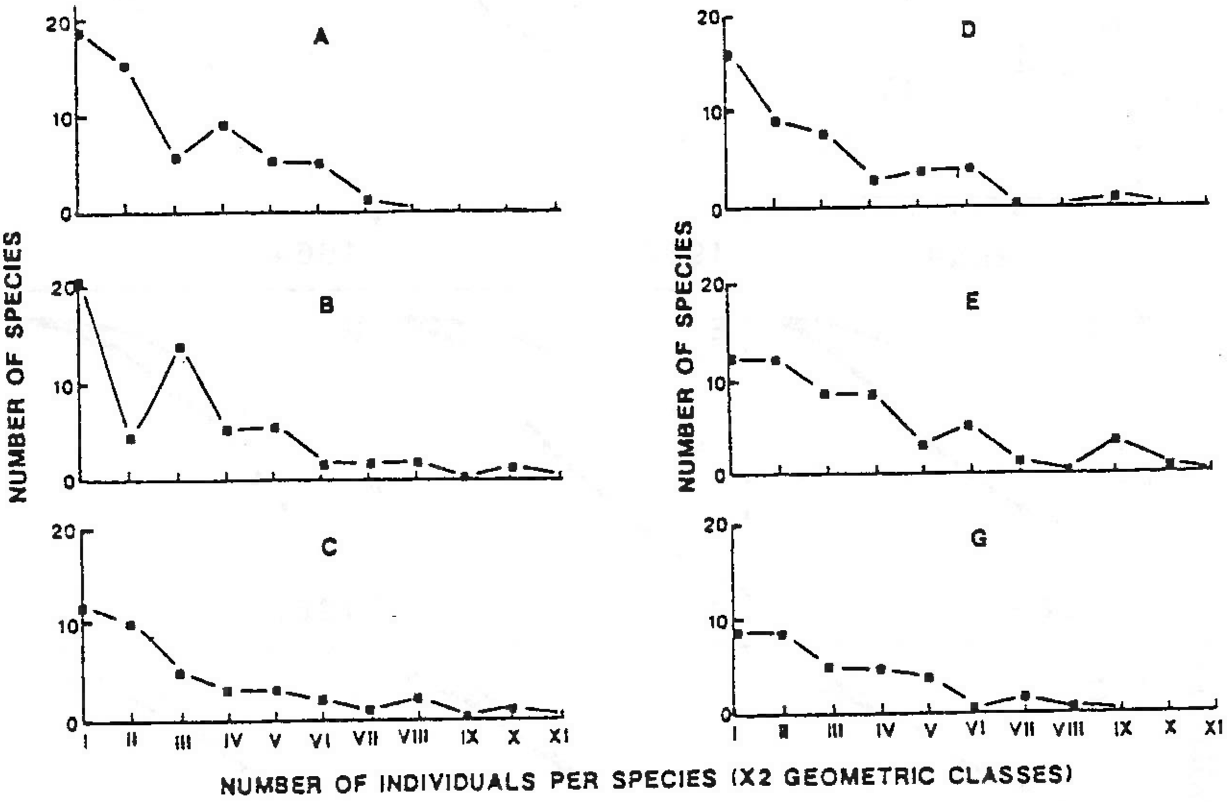


Fig. 1.5 Frierfjord macrofauna, sites A-E,G. Number of species against number of individuals per species in geometric classes (I = 1 individual per species, II = 2-3 ind. per spp., III = 4-7, IV = 8-15 etc.). Gray et al. (1988)

## MULTIVARIATE TECHNIQUES

| | EXAMPLES | | |
|---|---|---|---|
| STAGES | Hierarchical clustering | MDS ordination | PCA ordination |
| REPRESENTING COMMUNITIES | Dendrogram of replicates | Configuration of replicates (often 2-D) | |
| DISCRIMINATING SITES | ANOSIM test on triangular matrix of similarities<br>Similarity percentage breakdown (SIMPER) gives species responsible | | Multinormal tests (e.g. Wilks' Λ), but often invalid |
| DETERMINING STRESS LEVELS | Not appropriate | | |
| LINKING TO ENVIRONMENT | Visual (superimposing environmental variables on faunal ordinations). Finding subset of environmental variables whose ordination "best" matches the faunal ordination. | | |
| ESTABLISHING CAUSALITY | Mesocosm or field dosing experiments. Use above techniques - significance in discriminating "sites" (=treatments) establishes causality. | | |

## Table 1.4

Frierfjord macrofauna counts. Similarities (Bray-Curtis coefficient, after √√ transformation) between every pair of replicates (sites A-C only).

| | A1 | A2 | A3 | A4 | B1 | B2 | B3 | B4 | C1 | C2 | C3 | C4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A1 | - | | | | | | | | | | | |
| A2 | 61 | - | | | | | | | | | | |
| A3 | 69 | 60 | - | | | | | | | | | |
| A4 | 65 | 61 | 66 | - | | | | | | | | |
| B1 | 37 | 28 | 37 | 35 | - | | | | | | | |
| B2 | 42 | 34 | 31 | 32 | 55 | - | | | | | | |
| B3 | 45 | 39 | 39 | 44 | 66 | 66 | - | | | | | |
| B4 | 37 | 29 | 29 | 37 | 59 | 63 | 60 | - | | | | |
| C1 | 35 | 31 | 27 | 25 | 28 | 56 | 40 | 34 | - | | | |
| C2 | 40 | 34 | 26 | 29 | 48 | 69 | 62 | 56 | 56 | - | | |
| C3 | 40 | 31 | 37 | 39 | 59 | 61 | 67 | 53 | 40 | 66 | - | |
| C4 | 36 | 28 | 34 | 37 | 65 | 55 | 69 | 55 | 38 | 64 | 74 | - |

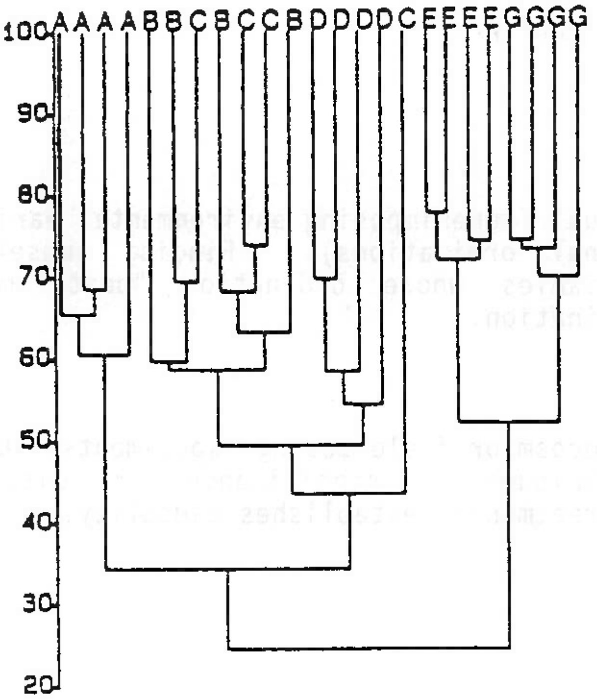


Fig. 1.6 Frierfjord macrofauna. Dendrogram for hierarchical clustering (group-average link) of 4 replicates from 6 sites, using above similarities

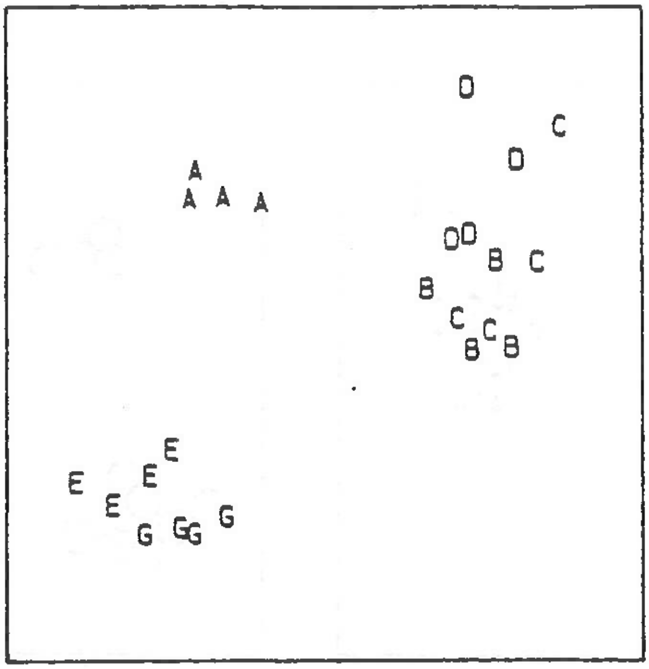


Fig. 1.7 Frierfjord macrofauna. Non-metric MDS ordination (in 2-D) of the 4 replicates from each of sites A-E and G, from Table 1.4 similarities

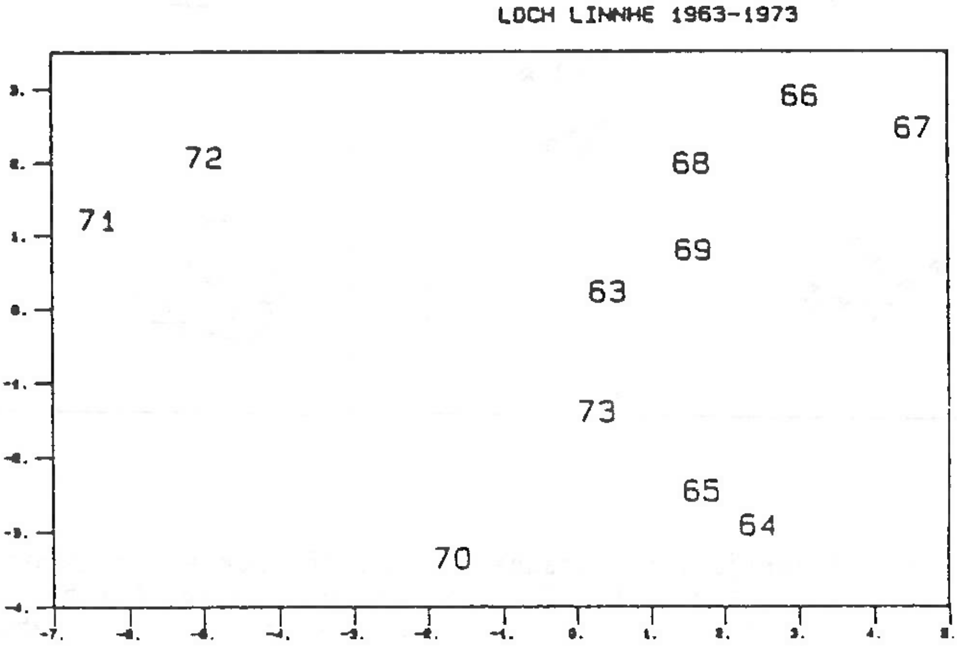


Fig. 1.8 Loch Linnhe macrofauna. PCA ordination (in 2-D) of the 11 years abundance data, omitting the less-common species

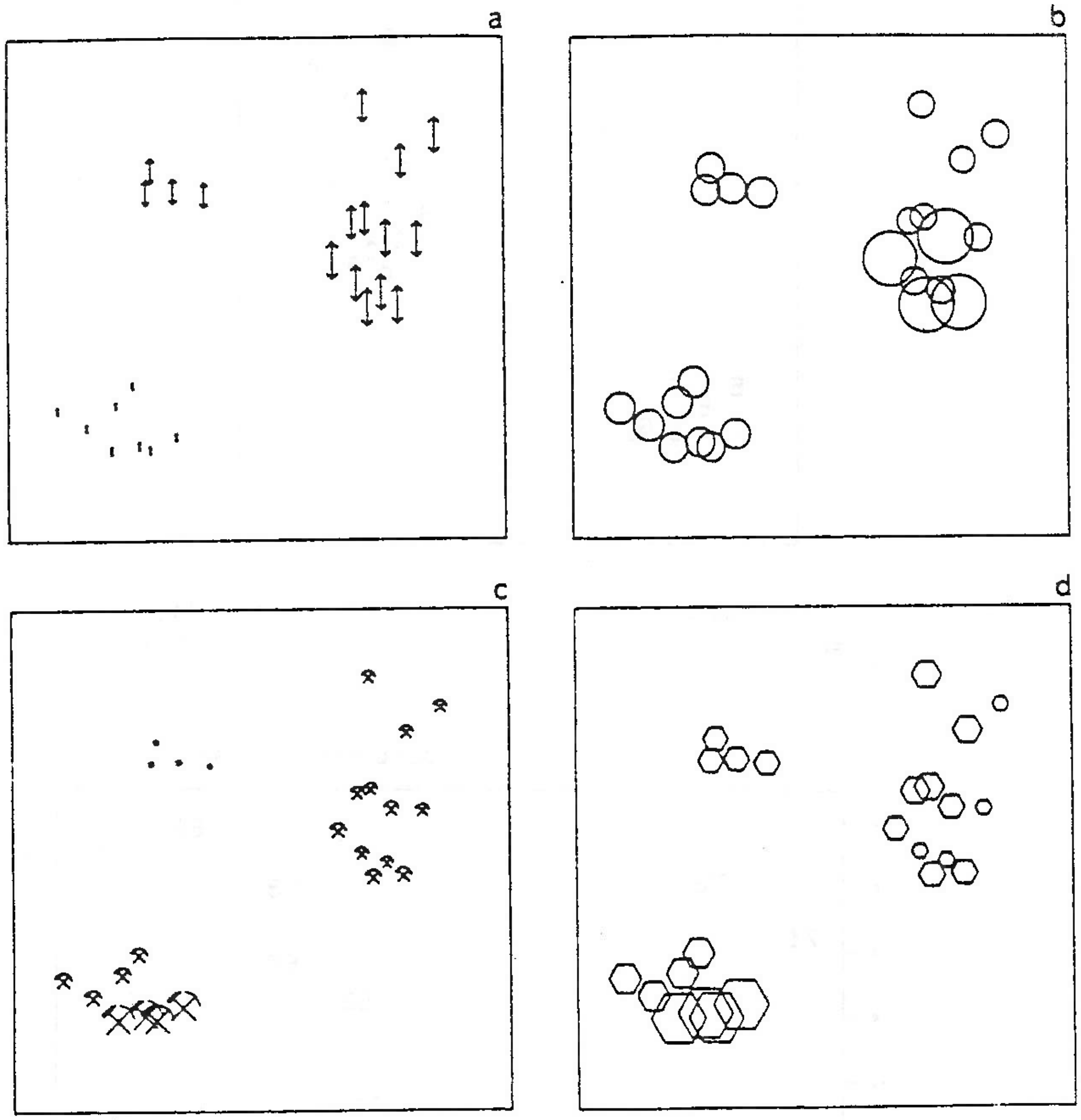


Fig. 1.9 Frierfjord macrofauna. Values of four environmental variables: (a) water depth, (b) sediment grain size, (c) metal and (d) PAH concentrations in sediment, superimposed on the abundance-based MDS

## NUTRIENT ENRICHMENT MESOCOSM EXPERIMENT

- Gee et al. (1985)

Meiofaunal abundances under 2 dosing regimes, Solbergstrand facility (NIVA), Norway

### Table 1.5

Copepod numbers (nematodes not shown) from 4 boxes for each treatment (high, low and no additions of powdered *Ascophyllum nodosum*).

| | Control | | | | Low dose | | | | High dose | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C4 | L1 | L2 | L3 | L4 | H1 | H2 | H3 | H4 |
| Copepoda, Harpacticoida | | | | | | | | | | | | |
| Ectinosomidae | | | | | | | | | | | | |
| Halectinosoma gothiceps | - | - | 1 | 1 | 16 | 23 | 8 | 16 | - | 1 | - | - |
| Tachidiidae | | | | | | | | | | | | |
| Danielssania fusiformis | 1 | 1 | 1 | 1 | 1 | 3 | 8 | 5 | 1 | - | - | 3 |
| Tisbidae | | | | | | | | | | | | |
| Tisbe sp. 1 (gracilis group) | - | - | - | - | - | - | - | - | 2 | 27 | 119 | 31 |
| Tisbe sp. 2 (graciloides?) | - | - | - | - | 45 | 22 | 39 | 25 | 6 | - | 3 | 32 |
| Tisbe sp. 3 | - | - | - | - | 86 | 83 | 88 | - | 5 | 29 | - | 20 |
| Tisbe sp. 4 | - | - | - | - | 151 | 249 | 264 | 87 | 8 | - | - | 34 |
| Tisbe sp. 5 | - | - | - | - | 129 | - | - | 115 | 4 | - | 1 | 40 |
| Diosaccidae | | | | | | | | | | | | |
| Typhlamphiascus typhlops | 4 | 2 | 2 | 4 | 5 | 8 | 4 | 3 | - | - | - | - |
| Bulbamphiascus imus | 1 | - | - | 2 | - | - | - | - | - | - | - | - |
| Stenhelia reflexa | 3 | 1 | - | 1 | 2 | - | - | - | - | - | - | - |
| Amphiascus tenuiremis | 1 | - | - | - | - | - | 2 | 6 | - | - | - | - |
| Ameiridae | | | | | | | | | | | | |
| Ameira parvula | - | - | - | - | 4 | 2 | 3 | 2 | 2 | - | 1 | 2 |
| Proameira simplex | - | - | - | - | - | 2 | - | 5 | - | - | - | - |
| Paramesochridae | | | | | | | | | | | | |
| Leptopsyllus paratypicus | - | - | 1 | - | - | - | - | - | - | - | - | - |
| Cletodidae | | | | | | | | | | | | |
| Enhydrosoma longifurcatum | 2 | 2 | 1 | 2 | 3 | 1 | - | - | - | - | - | - |
| Laophontidae | | | | | | | | | | | | |
| Unidentified copepodite | - | - | - | - | - | - | 1 | - | - | - | - | - |
| Ancorabolidae | | | | | | | | | | | | |
| Ancorabolis mirabilis | 3 | - | 4 | 4 | 2 | 18 | 3 | 3 | 27 | 3 | 1 | - |
| Unidentified Copepodites | - | - | 1 | - | 1 | 1 | 1 | 3 | - | - | - | - |

MULTIVARIATE: ESTABLISHING CAUSALITY

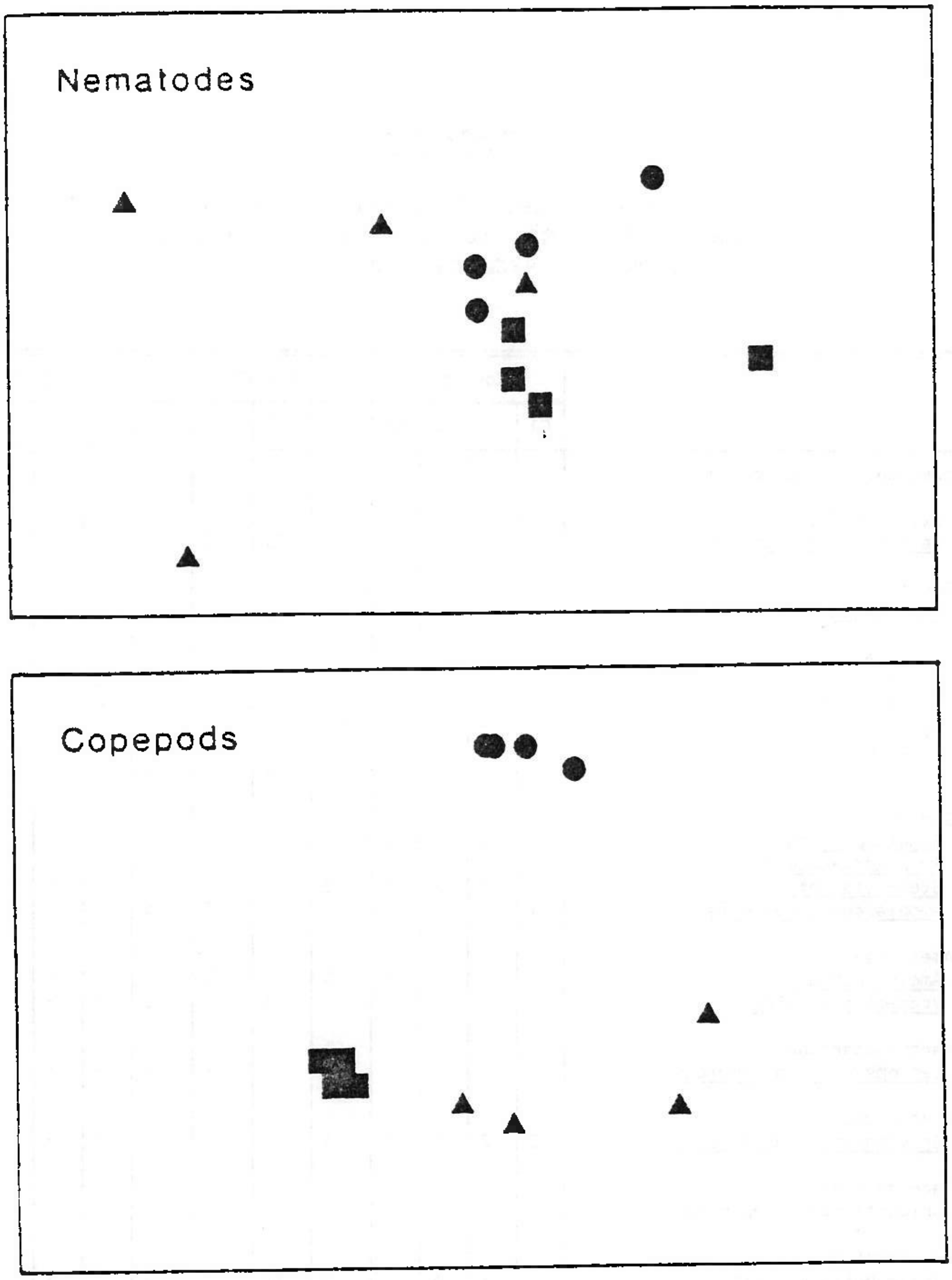


Fig. 1.10 Mesocosm meiofauna (nutrient enrichment). MDS ordination of abundances from 4 replicate boxes from 3 treatments: circles = control, squares = low dose, triangles = high dose. (Gee et al., 1985)

## DATA TRANSFORMATION AND SPECIES SELECTION/AGGREGATION

Some techniques may need TRANSFORMATION of the raw abundances/biomass (or derived statistics) for:

a) validity of assumptions for statistical analysis (e.g. normality, constant variance);

b) balancing contributions of rare/abundant species.

Some techniques may be possible with data on SELECTED (more dominant) species or data AGGREGATED to higher taxonomic levels, thus minimising identification time.

| TECHNIQUE | EXAMPLES | TRANSFORMATION | SELECTION/ AGGREGATION |
|---|---|---|---|
| UNIVARIATE | Diversity indices | Counts: No<br>Index: Possibly | No |
| | Indicator species | Yes (on counts/ biomass) | Yes |
| DISTRIBUTIONAL | ABC curves | Possible but not usual | Possible |
| | Ind. among species | No | No |
| MULTIVARIATE | Cluster | Usual (log or 4th root) on counts/biomass MDS transforms similarities also, to ranks. | Possible |
| | MDS | | Possible |
| | PCA | | Needed |

## LECTURE 2

## MULTIVARIATE METHODS: MEASURES OF SIMILARITY OF SPECIES ABUNDANCE/BIOMASS BETWEEN SAMPLES

DATA MATRIX: A p (species) x n (samples) array of scores (counts or biomass). The n samples might consist of a number of replicates (possibly only one) at each of a number of sites or times.

SIMILARITY COEFFICIENT: Measures the similarity (S) of the community structure between any pair of samples (thus SAMPLE SIMILARITIES), using:

a) absolute numbers (or biomass) of each species,

b) relative numbers (or biomass), i.e. STANDARDISE the scores, to reflect only species COMPOSITION (%),

c) only presence or absence of each species.

S is usually defined in the range (0, 1) or (0, 100%).

S = 1 (or 100%) means samples are totally similar,
S = 0 means samples are totally dissimilar.

SIMILARITY MATRIX: This is a set of similarity coefficients, calculated between every pair of samples and laid out in a lower triangular array.

Similarity matrices are the basis for many clustering and ordination techniques (REPRESENTATION) and associated tests (DISCRIMINATION), which:

a) discriminate sites or times (similarities between replicates at a site > similarities between sites)

b) cluster sites (similarities within groups of sites > similarities between groups)

c) allow gradation of sites (site A has similarities with B, and B has with C, but A and C less similar).

SPECIES SIMILARITY MATRIX: A matching triangular array of similarities between every pair of species, in terms of patterns of occurrence across the samples.

Many different ways to assess similarity (because data is multi-species). One of the most useful in ecology is:

BRAY-CURTIS COEFFICIENT: (Bray and Curtis, 1957).
Similarity between jth and kth samples is:

$$S_{jk} = 100 \{1 - \frac{\sum_{i=1}^{p} |y_{ij} - y_{ik}|}{\sum_{i=1}^{p} (y_{ij} + y_{ik})}\} \qquad (2.1)$$

$$= 100 \frac{\sum_{i=1}^{p} 2 \min(y_{ij}, y_{ik})}{\sum_{i=1}^{p} (y_{ij} + y_{ik})}$$

where $y_{ij}$ = score (count or biomass) for ith species in jth sample ($i=1,2,\ldots,p$; $j = 1,2,\ldots,n$).

Example: Loch Linnhe macrofauna (Pearson, 1975).

Table 2.1

(a) Abundance (untransformed) for some selected species and years from site 34 data. (b) Resulting Bray-Curtis similarity matrix.

| (a) Year: | 64 | 68 | 71 | 73 |
|---|---|---|---|---|
| (Sample: | 1 | 2 | 3 | 4) |
| Species | | | | |
| *Echinoca* | 9 | 0 | 0 | 0 |
| *Myrioche* | 19 | 0 | 0 | 3 |
| *Labidopl.* | 9 | 37 | 0 | 10 |
| *Amaeana* | 0 | 12 | 144 | 9 |
| *Capitella* | 0 | 128 | 344 | 2 |
| *Mytilus* | 0 | 0 | 0 | 0 |

| (b)<br>Sample | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | - | | | |
| 2 | 8 | - | | |
| 3 | 0 | 42 | - | |
| 4 | 39 | 21 | 4 | - |

1) Note S = 0 if the two samples have no species in common (e.g. 1 and 3 above).

2) A scale change in y (e.g. biomass changed from mg per $m^2$ to per $cm^2$) does not change S.

3) "Joint absences" also have no effect on S (as is desirable), e.g. can omit species 6 in the table.

With "raw" counts (or biomass), S gives too much weight to large scores, so a log(1+y) or $\sqrt{\sqrt{y}}$ transform is often applied, before computing S.

<u>Example:</u> Loch Linnhe macrofauna, √√ transformation

Table 2.2

(a) √√- transformed abundances for 4 years.
(b) Resulting Bray-Curtis similarity matrix.

| (a) Year: | 64 | 68 | 71 | 73 |
|---|---|---|---|---|
| (Sample: | 1 | 2 | 3 | 4) |
| Species | | | | |
| 1 | 1.7 | 0 | 0 | 0 |
| 2 | 2.1 | 0 | 0 | 1.3 |
| 3 | 1.7 | 2.5 | 0 | 1.8 |
| 4 | 0 | 1.9 | 3.5 | 1.7 |
| 5 | 0 | 3.4 | 4.3 | 1.2 |
| 6 | 0 | 0 | 0 | 0 |

(b)

| Sample | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | - | | | |
| 2 | 26 | - | | |
| 3 | 0 | 68 | - | |
| 4 | 52 | 68 | 42 | - |

CANBERRA COEFFICIENT: Lance and Williams, 1967.
Similarity between samples j and k is:

$$S_{jk} = 100 \{1 - p^{-1} \sum_{i=1}^{p} \frac{|y_{ij} - y_{ik}|}{(y_{ij} + y_{ik})}\} \qquad (2.2)$$

It gives a more equal contribution from each species (so tends to be <u>overdominated</u> by rarer ones).

CORRELATION COEFFICIENT: Product-moment correlation

$$r_{jk} = \frac{\sum_i (y_{ij} - \bar{y}_{.j})\ (y_{ik} - \bar{y}_{.k})}{\sqrt{[\sum_i (y_{ij} - \bar{y}_{.j})^2 . \sum_i (y_{ik} - \bar{y}_{.k})^2]}} \qquad (2.3)$$

is not a similarity (it can be <0). Use:

$$S_{jk} = 50\ (1 + r_{jk}) \qquad (2.4),$$

<u>but</u> note that S increases with more joint absences.

PRESENCE/ABSENCE DATA

Many similarity coefficients have been proposed based on (0,1) data arrays (Sneath and Sokal, 1973). For comparing samples j and k let:

a = number of species present in both samples,
b+c = number present in one sample and not the other,
d = number absent from both samples.

"SIMPLE MATCHING" COEFFICIENT:

$$S_{jk} = 100.(a+d)/(a+b+c+d) \qquad (2.5)$$

Note that this _is_ a function of joint absences (d).

JACCARD'S COEFFICIENT:

$$S_{jk} = 100.a/(a+b+c) \qquad (2.6)$$

SORENSEN (OR DICE) COEFFICIENT:

$$S_{jk} = 100.2a/(2a+b+c) \qquad (2.7)$$

This is simply BRAY-CURTIS applied to (0,1) data.

McCONNAUGHEY COEFFICIENT (McConnaughey, 1964):

$$S_{jk} = 100[a(2a+b+c)]/[2(a+b)(a+b)] \qquad (2.8)$$

**RECOMMENDATION:**

1) Use coefficient not dependent on joint absences.

2) Similarities from raw counts (or biomass) are too dominated by common (or large) species, but

3) Reduction to presence/absence loses too much useful information, so recommend use:

4) BRAY-CURTIS on $\sqrt{\sqrt{y}}$ or log(1+y) transformed data.

5) Standardise scores if non-comparable sample volumes used, or if "patchiness" makes compositional change more relevant than fluctuations in absolute counts.

_SPECIES SIMILARITIES:_ These are computed from the same data array but between any pair of _species_ (rows i,l say) across all samples (columns).

BRAY-CURTIS:

$$S'_{il} = 100\ \{1 - \frac{\sum_{j=1}^{n} |y_{ij} - y_{lj}|}{\sum_{j=1}^{n} (y_{ij} + y_{lj})}\} \qquad (2.9)$$

However:

1) Similarities between rare species have little meaning (S′ usually 0) and should be omitted from any <u>species</u> clustering or ordination.

2) Standardisation (not transformation) of y needed:

$$y^*_{ij} = 100\, y_{ij} / (\textstyle\sum_{k=1}^{n} y_{ik}) \qquad (2.10),$$

(before computing S′), so two species in strict ratio across samples are "perfectly similar".

<u>Example</u> <u>Counts</u> <u>Similarities</u>

| Sample: | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Species | | | | | |
| 1 | 2 | 0 | 0 | 4 | 4 |
| 2 | 10 | 0 | 0 | 20 | 20 |
| 3 | 0 | 4 | 4 | 1 | 1 |

——>

| Species | 1 | 2 | 3 |
|---|---|---|---|
| 1 | - | | |
| 2 | 33 | - | |
| 3 | 20 | 7 | - |

| Standardise

V

| Species | | | | | |
|---|---|---|---|---|---|
| 1 | 20 | 0 | 0 | 40 | 40 |
| 2 | 20 | 0 | 0 | 40 | 40 |
| 3 | 0 | 40 | 40 | 10 | 10 |

——>

| | | | |
|---|---|---|---|
| 1 | - | | |
| 2 | 100 | - | |
| 3 | 20 | 20 | - |

CORRELATION coefficients are more appropriate for species similarity, since they incorporate scale changes, but the location changes are undesirable.

<u>**RECOMMENDATION:**</u> For species similarities, use BRAY-CURTIS on standardised scores. Remove rarer species (never >3%, say, of total score in any sample).

## <u>DISSIMILARITY COEFFICIENTS</u>

These are important in constructing <u>ordinations</u>, in which <u>dissimilarities</u> (δ) between pairs of samples are turned into <u>distances</u> (d) between sample locations on a "map". (δ therefore >0, of course).

Similarities can easily become dissimilarities, by:

$$\delta = 100 - S \qquad (2.11),$$

e.g. for BRAY-CURTIS:

$$\delta_{jk} = 100 . \frac{\sum_{i=1}^{p} |y_{ij} - y_{ik}|}{\sum_{i=1}^{p} (y_{ij} + y_{ik})} \qquad (2.12)$$

so δ=0: no dissimilarity, δ=100: total dissimilarity.

Other dissimilarity measures, based on distances:

EUCLIDEAN DISTANCE:

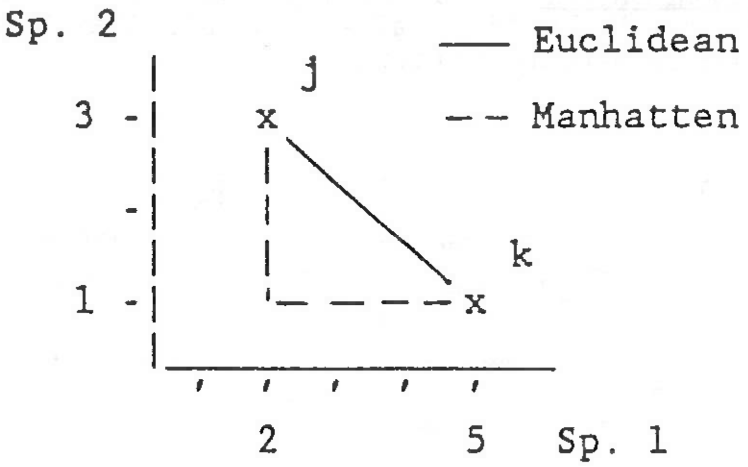
$$d_{jk} = \sqrt{[\sum_{i-1}^{p} (y_{ij} - y_{ik})^2]} \qquad (2.13)$$

MANHATTEN (or CITY-BLOCK) DISTANCE:

$$d_{jk} = \sum_{i=1}^{p} |y_{ij} - y_{ik}| \qquad (2.14)$$

<u>Example:</u>

| | Sample: | j | k |
|---|---|---|---|
| Sp. | 1 | 2 | 5 |
| | 2 | 3 | 1 |



[<u>METRICS</u>: Euclidean and Manhatten measures, (2.13) and (2.14), are called distances or <u>metrics</u> because they obey the triangle inequality, i.e. for any three samples, j,k,r:

$$d_{jk} + d_{kr} \geq d_{jr} \qquad (2.15)$$

Note: Bray-Curtis dissimilarity does not satisfy the triangle inequality, so should not be called a "metric". However, many useful dissimilarities are also not metrics (e.g. squared Euclidean distance, giving dissimilarities of the same rank order as Euclidean distance, i.e. identical MDS ordinations).

CONCLUDE: Unnecessary to insist that dissimilarities are true "metrics".]

Where necessary (e.g. for input to clustering), distance (d) can be conveniently converted to similarity (S) by:

$$S = 100/(1 + d) \qquad (2.16),$$

and, using (2.11), distance (d) turned to dissimilarity ($\delta$) by

$$\delta = 100d/(1 + d) \qquad (2.17).$$

So, $d = 0$ gives $\delta = 0$, $S = 100$, and $d \to \infty$ gives $\delta \to 100$, $S \to 0$.

However, note that EUCLIDEAN (or MANHATTEN) distance is the same if a species is absent in both samples or is present in both at the same abundance; this is undesirable. (Same problem as that of similarities based on correlation being dependent on joint absences.) So:

RECOMMENDATION: For clustering or MDS of species counts/biomass, use Bray-Curtis dissimilarities, after suitable transformation, rather than Euclidean (or Manhatten) distances.

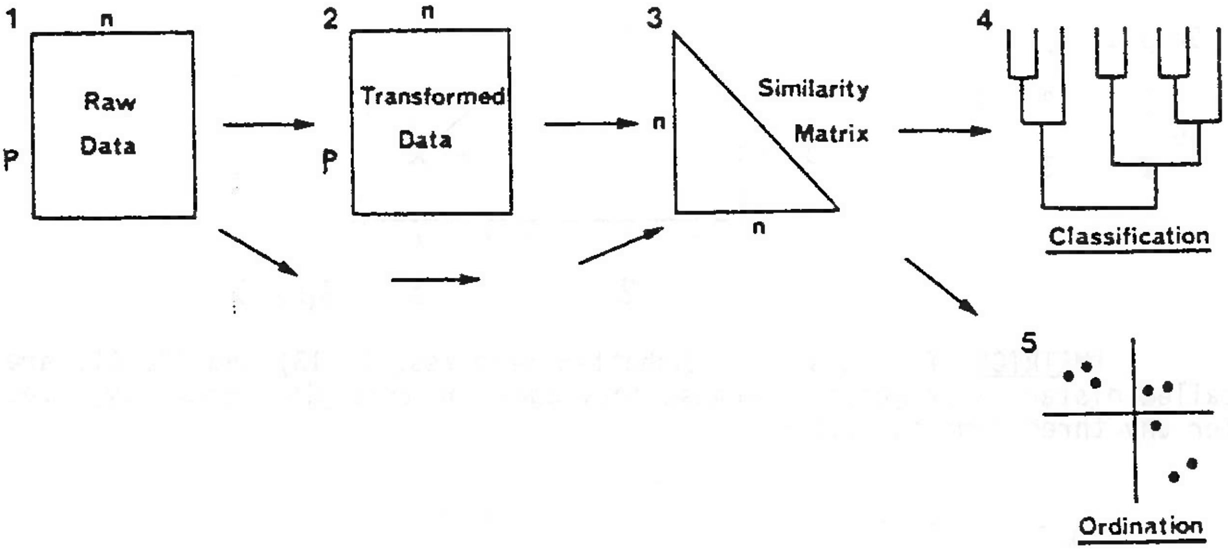


Fig. 2.1 Stages in a multivariate analysis based on (dis)similarity coefficients

# LECTURE 3

## MULTIVARIATE METHODS: HIERARCHICAL CLUSTERING

### Table 3.1

Frierfjord macrofauna counts. Similarities (Bray-Curtis coefficient, after √√ transformation) between every pair of replicates (sites A-C only).

| | A1 | A2 | A3 | A4 | B1 | B2 | B3 | B4 | C1 | C2 | C3 | C4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A1 | - | | | | | | | | | | | |
| A2 | 61 | - | | | | | | | | | | |
| A3 | 69 | 60 | - | | | | | | | | | |
| A4 | 65 | 61 | 66 | - | | | | | | | | |
| B1 | 37 | 28 | 37 | 35 | - | | | | | | | |
| B2 | 42 | 34 | 31 | 32 | 55 | - | | | | | | |
| B3 | 45 | 39 | 39 | 44 | 66 | 66 | - | | | | | |
| B4 | 37 | 29 | 29 | 37 | 59 | 63 | 60 | - | | | | |
| C1 | 35 | 31 | 27 | 25 | 28 | 56 | 40 | 34 | - | | | |
| C2 | 40 | 34 | 26 | 29 | 48 | 69 | 62 | 56 | 56 | - | | |
| C3 | 40 | 31 | 37 | 39 | 59 | 61 | 67 | 53 | 40 | 66 | - | |
| C4 | 36 | 28 | 34 | 37 | 65 | 55 | 69 | 55 | 38 | 64 | 74 | - |

Seeing structure in a similarity matrix is difficult - a graphic representation of relations is needed:

**CLUSTER ANALYSIS** Clustering (or classification) aims to find "natural groupings" of samples such that samples within a group are more similar than samples in different groups. Use clustering to:

1) Distinguish sites (or times) - replicates within sites fall in the same cluster;

2) Partition sites (or times) into groups;

3) Define species assemblages (spp. co-occur at sites)

Hundreds of clustering methods exist (Everitt, 1980), some operating on (dis)similarities, some on raw data. Cormack (1971) warns against indiscriminate use: "availability of ... classification techniques has led to the waste of more valuable scientific time than any other 'statistical' innovation".

Five classes of clustering methods can be defined:

1) Hierarchical, 2) Optimising, 3) Mode seeking, 4) Clumping and 5) Miscellaneous techniques.

Here consider only one (sub)class, which recognises that clustering can occur at several levels.

**<u>HIERARCHICAL AGGLOMERATIVE CLUSTERING:</u>** The n samples are successively <u>fused</u> into groups, starting with samples with the highest mutual similarities then gradually lowering the similarity level at which groups are fused, and ending in a single cluster. (DIVISIVE clustering is the opposite sequence). Process represented by a tree diagram or DENDROGRAM.

DISTINGUISHING SITES: Frierfjord macrofauna counts.

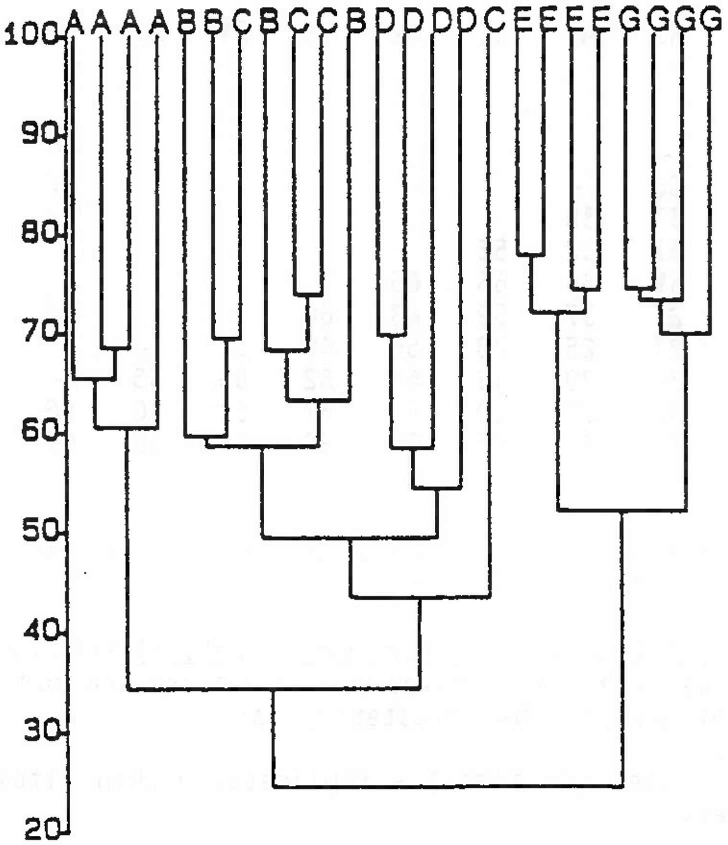


Fig. 3.1 Frierfjord macrofauna counts. Dendrogram for hierarchical clustering (using group-average linking) of 4 replicates from each of sites A-E,G, using Bray-Curtis similarity matrix (Table 3.1)

GROUPING TIMES: Loch Linnhe macrofauna - subset. After √√ transformation, data array and bray-Curtis similarity matrix are:

| Year: | 64 | 68 | 71 | 73 |
|---|---|---|---|---|
| Sample: | 1 | 2 | 3 | 4 |
| Species | | | | |
| *Echin.* | 1.7 | 0 | 0 | 0 |
| *Myrio.* | 2.1 | 0 | 0 | 1.3 |
| *Labid.* | 1.7 | 2.5 | 0 | 1.8 |
| *Amaea.* | 0 | 1.9 | 3.5 | 1.7 |
| *Capit.* | 0 | 3.4 | 4.3 | 1.2 |
| *Mytil.* | 0 | 0 | 0 | 0 |

| Sample | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | - | | | |
| 2 | 25.6 | - | | |
| 3 | 0.0 | 67.9 | - | |
| 4 | 52.2 | 68.1 | 42.0 | |

| 2 & 4 fused
V

| Sample | 1 | 2&4 | 3 |
|---|---|---|---|
| 1 | - | | |
| 2&4 | 38.9 | - | |
| 3 | 0.0 | 55.0 | - |

| (2&4) & 3 fused
V

| Sample | 1 | 2&3&4 |
|---|---|---|
| 1 | - | |
| 2&3&4 | 25.9 | - |

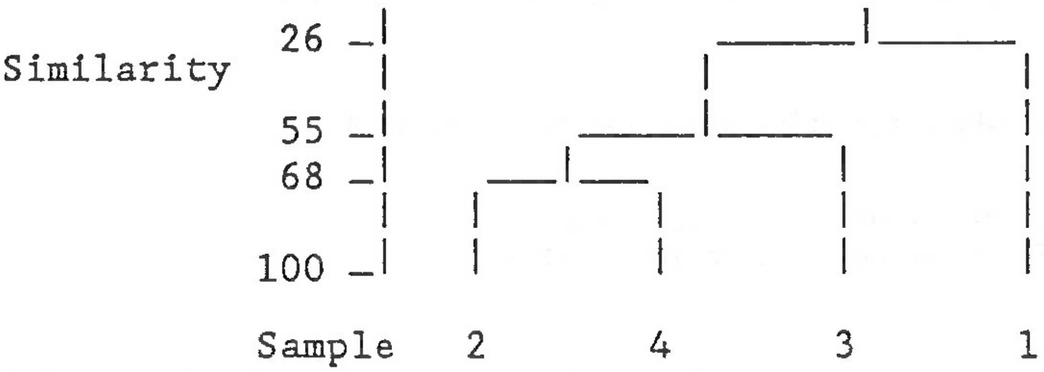
Samples 2 and 4 have the highest similarity, S(2,4), so they form the first group.

Their similarity to (say) sample 1 defined in one of 3 ways:

a) SINGLE LINKAGE: max{S(1,2), S(1,4)} (=52.2)

b) COMPLETE LINKAGE: min{S(1,2), S(1,4)} (=25.6)

c) GROUP AVERAGE LINK: [S(1,2) + S(1,4)]/2 (=38.9)
(Average weighted by number of samples in groups fused, e.g. S(1,2&3&4) = (2x38.9 + 1x0)/3 = 25.9).



Note:

1) Samples need to be reordered for clear presentation of the dendrogram (so there are no crossing lines).

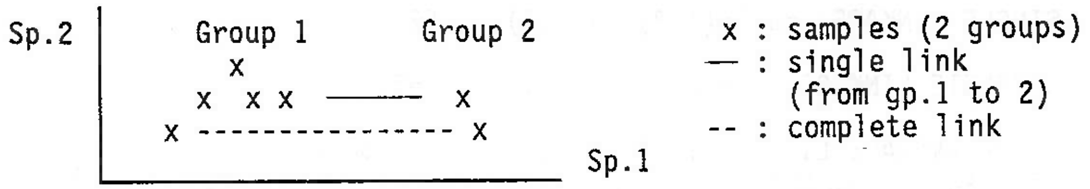
2) The order of samples on the x axis is not very meaningful (think of a dendrogram as a "mobile").

3) Here clustering imposes a (somewhat arbitrary) grouping on what is essentially a continuum (clean (1), impacted (2 and 3) and some recovery (4)), so:

4) Small changes in similarities can have larger effects on picture (e.g. reverse S(2,3) & S(2,4)).

DISSIMILARITIES: Exactly converse operations needed for a dissimilarity matrix, i.e. fuse samples with lowest dissimilarity, take minimum dissimilarity in single linkage, maximum in complete linkage.

LINKAGES: These three options are best visualised for an example with only 2 species and dissimilarity defined simply from Euclidean distance.



Group average is mean of all 12 intergroup distances.

Explains why alternative names for the linkages are:

"NEAREST NEIGHBOUR" = single linkage
"FURTHEST NEIGHBOUR" = complete linkage

Note: Though single linkage has some nice theoretical properties (e.g. clustering only a function of rank order of similarities), it has a tendency to give chains of linked samples rather than clear groups; group average linking is usually preferable.

Example: Bristol Channel (UK) zooplankton, April 1974, 57 sites X 24 species, Collins and Williams (1982).

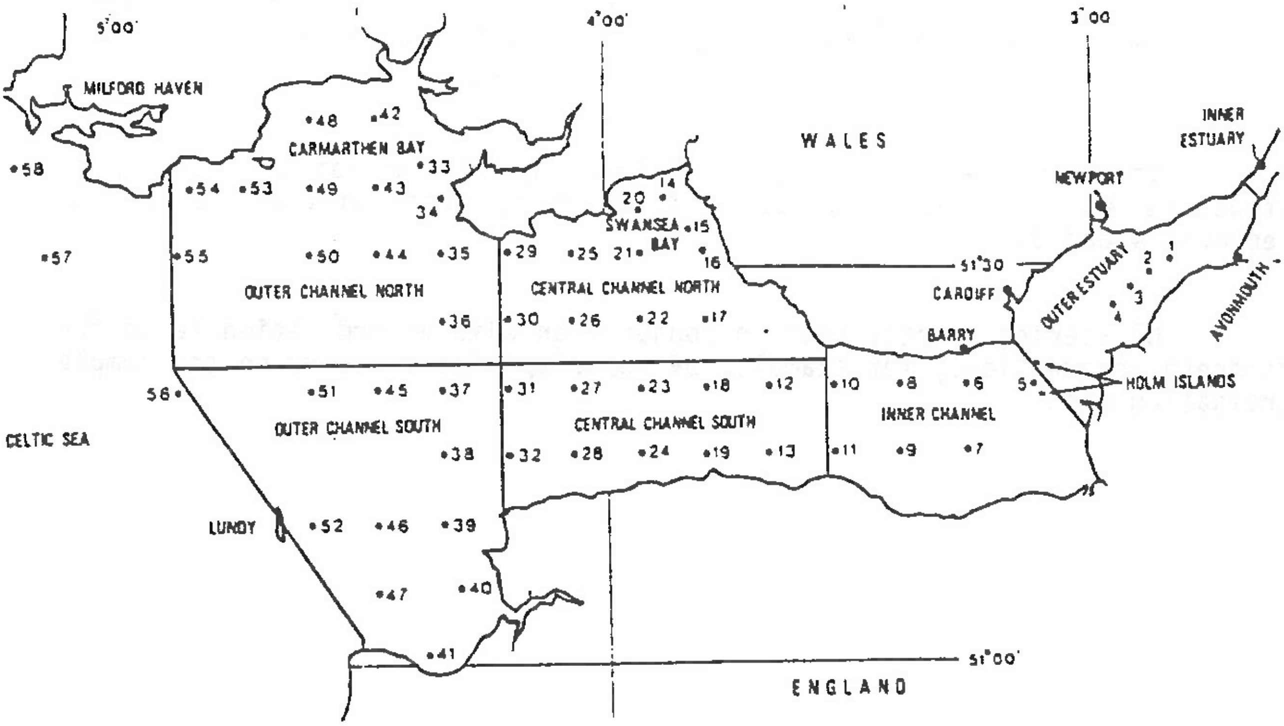


Fig. 3.2 Bristol Channel sampling sites 1-29, 31-58

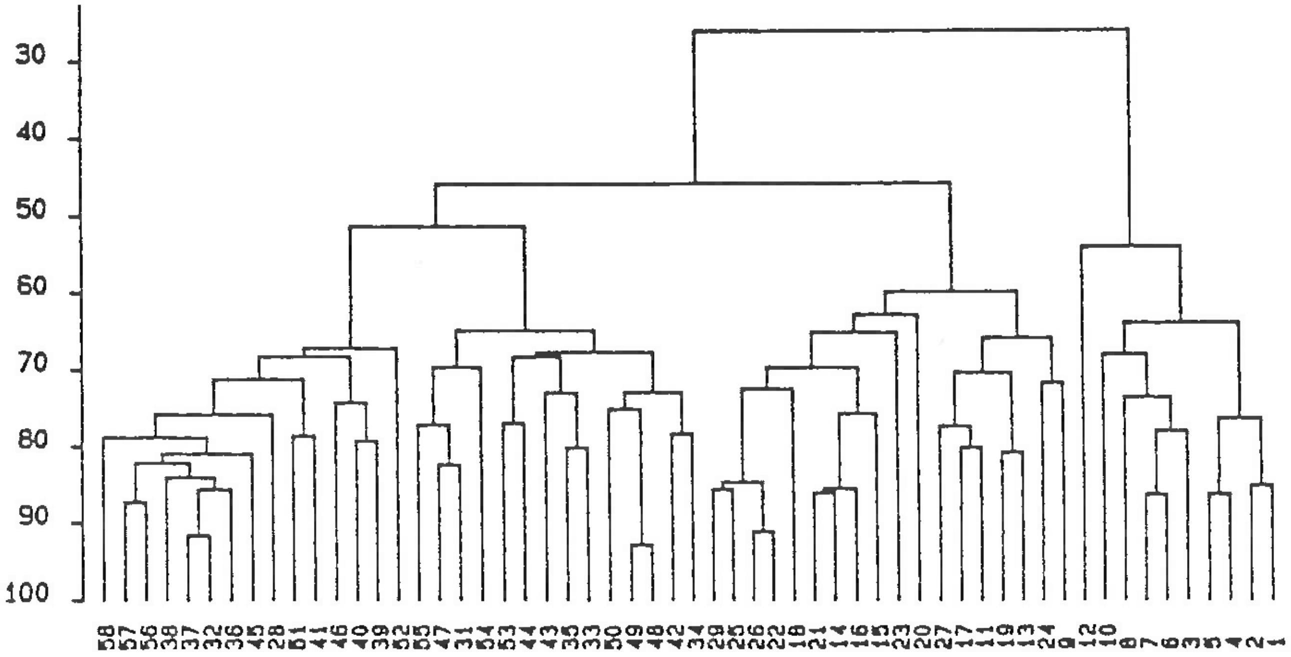


Fig. 3.3 Bristol channel. Dendrogram for hierarchical clustering of 57 sites (group average linking of Bray-Curtis similarities on √√-abundance)

## RECOMMENDATIONS

Hierarchical clustering (with group average linking) on sample (dis)similarity matrices can be useful, especially to delineate discrete communities at differing sites (or groups of sites).

It is less useful (and can be misleading) for a gradation in community structure across sites or times; ordination is preferable for this (see lectures 4 and 5).

Clustering is best used in conjunction with an ordination (even for discrete communities), for example, by superimposing clusters on the sample ordination plot.

## LECTURE 4

## MULTIVARIATE METHODS: ORDINATION OF SAMPLES BY PRINCIPAL COMPONENTS ANALYSIS (PCA)

**ORDINATIONS:** These are techniques for MAPPING the SAMPLES in a low number of dimensions (usually 2) such that the DISTANCE between samples attempts to reflect (DIS)SIMILARITY in community structure. (No guarantee that the attempt will succeed, if the relationships between the samples are complex, i.e. the structure is essentially "high-dimensional".)

Again there are many methods, for example:

PRINCIPAL COMPONENTS ANALYSIS (PCA, e.g. Chatfield & Collins, 1980),

PRINCIPAL CO-ORDINATES ANALYSIS (PCoA, Gower, 1966),

DETRENDED CORRESPONDENCE ANALYSIS (DECORANA, Hill & Gauch, 1980),

NON-METRIC MULTIDIMENSIONAL SCALING (MDS, e.g. Kruskal & Wish, 1978).

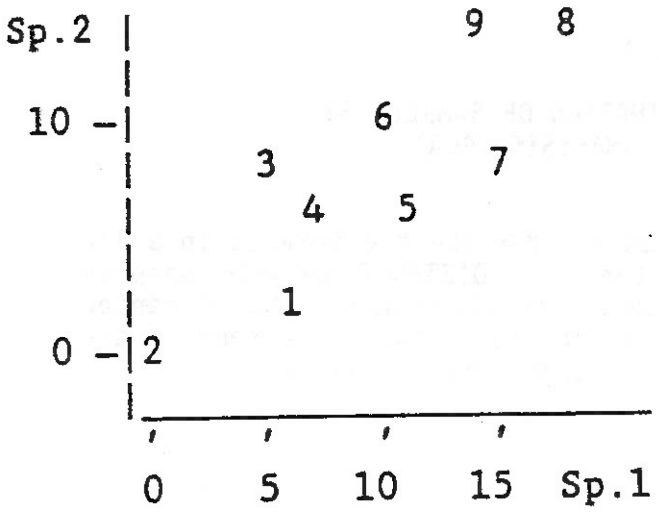
Here we consider only PCA (a simple but rather limited method) and MDS (a more complex algorithm but simple in concept and very generally applicable).

### PRINCIPAL COMPONENTS ANALYSIS

STARTING POINT is the original DATA MATRIX (rather than a similarity matrix). The data array is thought of as defining the positions of samples in relation to axes representing the full set of species (one axis for each species). The samples are thus POINTS in a very HIGH-DIMENSIONAL SPACE, so it helps to visualise the process by considering an example in which there are only two species, i.e. each sample is a point in 2-dimensions.

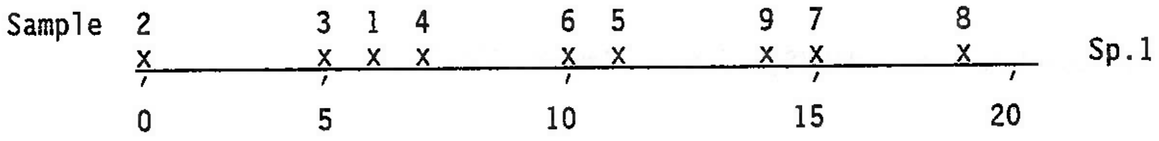
Example:

|  | Sample: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Abundance | Sp.1: | 6 | 0 | 5 | 7 | 11 | 10 | 15 | 18 | 14 |
|  | Sp.2: | 2 | 0 | 8 | 6 | 6 | 10 | 8 | 14 | 14 |

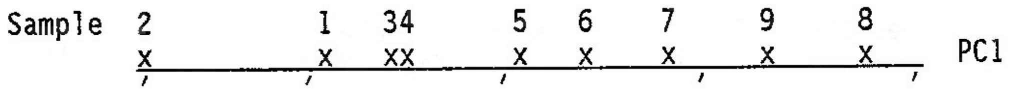


(This _is_ an ORDINATION already - of 2-d data in 2-d, thus perfectly summarising all the relationships between samples).

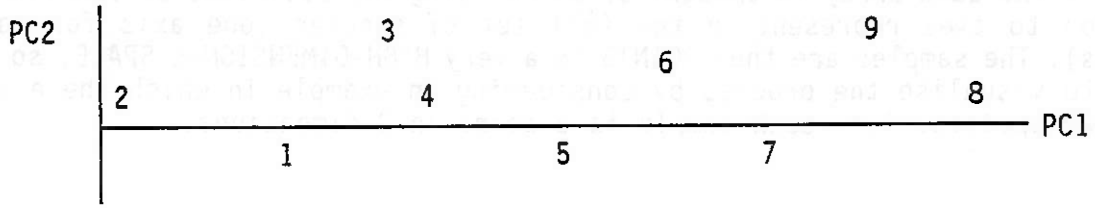
For a 1-d ordination (i.e. a genuine _ordering_ of samples) could take just one variable (Sp.1, say):



but this is poorer approximation to relations between samples than given by a (perpendicular) PROJECTION onto the line of "best fit" in the 2-d plot:



This is 1st PC AXIS; PC2 AXIS is PERPENDICULAR to this



PC AXES (full set) are simply a ROTATION of original species axes. Refer samples to (PC1, PC2) rather than (Sp.1, Sp.2) axes because may be able to DISPENSE WITH PC2, giving an ordination in 1-d.

Biggest differences between samples take place along PC1, and this is an equivalent definition of PC1 - the axis along which VARIANCE IS MAXIMISED.

Example: Add a third species to previous example.

| | Sample: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Abundance | Sp.1: | 6 | 0 | 5 | 7 | 11 | 10 | 15 | 18 | 14 |
| | Sp.2: | 2 | 0 | 8 | 6 | 6 | 10 | 8 | 14 | 14 |
| | Sp.3: | 3 | 1 | 6 | 6 | 9 | 11 | 10 | 16 | 15 |

Samples are now points in 3-d and there are 3 PC axes, again a rotation of the 3 species axes, such that:

PC1: Axis which MAXIMISES VARIANCE of points PROJECTED PERPENDICULARLY onto it.

PC2: Constrained to be perpendicular to PC1, again chosen to maximise variance along this axis.

PC3: Perpendicular to PC1 and PC2.

The new variables (PCs) are then just LINEAR COMBINATIONS of the old ones (species), such that PC1, PC2, PC3 are UNCORRELATED.

Here, the three PCs are:

$$\begin{aligned} PC1 &= \ \ 0.62 \times Sp.1 + 0.52 \times Sp.2 + 0.58 \times Sp.3 \\ PC2 &= -0.73 \times Sp.1 + 0.65 \text{ X } Sp.2 + 0.20 \times Sp.3 \qquad (4.1) \\ PC3 &= \ \ 0.28 \times Sp.1 + 0.55 \times Sp.2 - 0.79 \times Sp.3 \end{aligned}$$

Letting var(PCi) = variance of samples on ith PC axis, var(Sp.i) = variance on ith species axis (i-1,2,3):

$$\Sigma_i \text{ var(PCi)} = \Sigma_i \text{ var(Sp.i)} \qquad (4.2)$$

so % OF (original) VARIANCE EXPLAINED by ith PC is:

$$\text{var(PCi)} / \Sigma_i \text{ var(PCi)} \qquad (4.3).$$

Here PC1 explains 93%, PC2 6% and PC3 1% of variance. Little variability (information) in PC3. Ignore it, so

PCA ORDINATION: The PC1 and PC2 axes give a 2-d ordination plane (of "best fit" to the sample points) and points are projected perpendicularly onto this from the higher PCs (just PC3 here). In this case, the 2-d ordination is almost a perfect summary of the 3-d data (the sample points lie near to a plane in the original 3-d species space).

HIGHER-DIMENSIONAL DATA: Typically, there are many more species (say 30+) but the approach is identical. Samples are points in the 30-d (say) species space; the "best-fit" 2-d plane is found and samples projected onto it to get the 2-d PCA ordination. Success is measured by the % of the variability explained by the first 2 of the 30 PCs.

COMPUTATION: Construction of PCs requires derivation of eigenvalues and vectors of a pxp matrix (p = no. of species), e.g. Chatfield and Collins, 1980 (note: knowledge of matrix algebra essential). Problems if p is large (compared with no. of samples), so:

EXCLUDE LESS-COMMON SPECIES: These distort ordination badly (even if the matrix operations are possible). E.g. for Loch Linnhe data, the PCA ordination (Fig. 4.1) excludes species making up <3% of total counts at any site, leaving 29 species from 115.

TRANSFORM REMAINING ABUNDANCES (/BIOMASS) before applying PCA, to avoid over-domination by the very common species. E.g. in Loch Linnhe data, Capitella counts go over 4000; Fig. 4.1 uses √√ transform.

Example: Loch Linnhe macrofauna (site 34, 1963-1973).

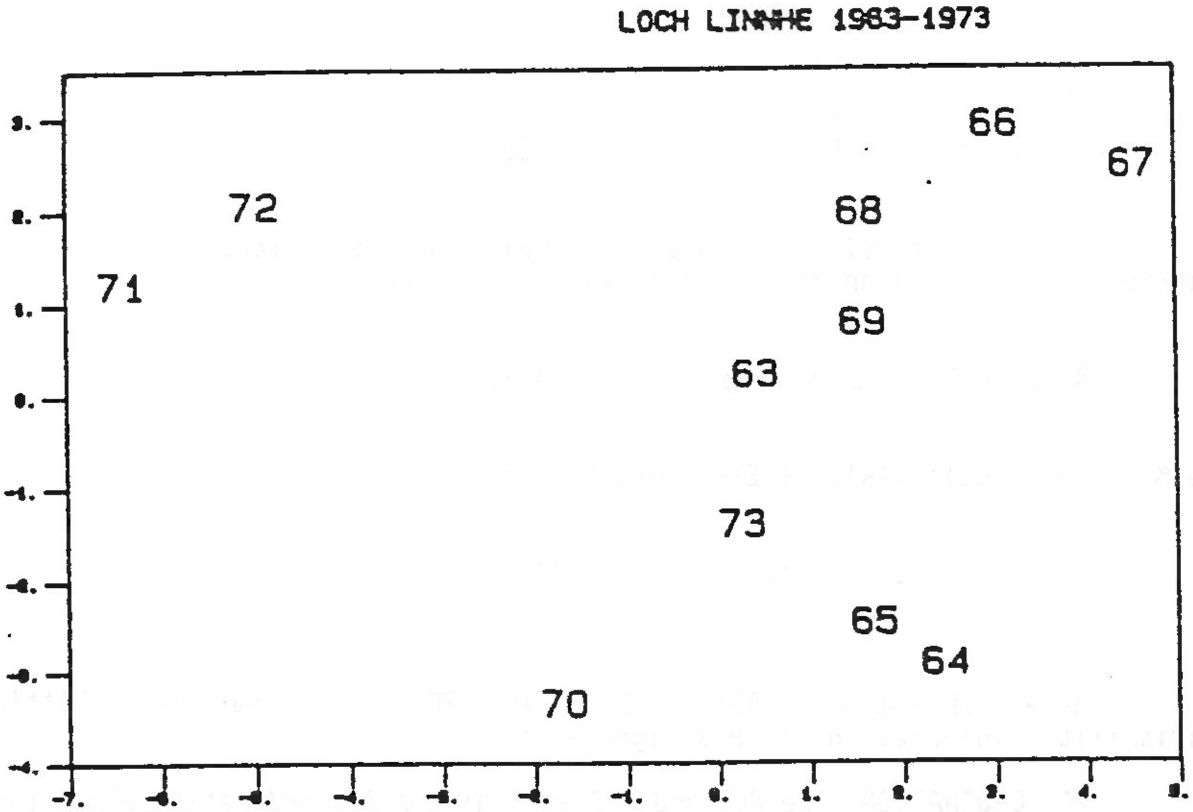


Fig. 4.1 Loch Linnhe abundances. 2-d PCA ordination of samples from 11 years; PC1 (x axis) and PC2 (y axis) account for 57% of total sample variability

SCALE AND LOCATION CHANGES: Data often NORMALISED (after any transform). For each species subtract the mean (across sites) and divide by the standard deviation. Equivalently, extract eigenvalues of the correlation rather than the covariance matrix, i.e. CORRELATION-BASED PCA rather than COVARIANCE-BASED PCA. Essential if variables have different scales (units) or widely differing ranges. Not the case here (after transform at least) so less necessary (but <u>was</u> done in Fig. 4.1).

## <u>PCA STRENGTHS</u>

1) CONCEPTUALLY SIMPLE.

2) COMPUTATIONALLY STRAIGHTFORWARD, <u>provided</u> the number of species is reduced (usually drastically), and it can then cope with an unlimited number of samples.

3) ORDINATION AXES potentially have some meaning, as simple LINEAR COMBINATIONS of the species (though these are rarely readily interpretable in practice).

## <u>PCA WEAKNESSES</u>

1) LITTLE FLEXIBILITY in defining relations between samples - in effect "dissimilarities" are simple Euclidean distances in the species space. The only flexibility comes from transformation of the species axes.

2) Does NOT do a very good job of PRESERVING these DISTANCES (dissimilarities) in the 2-d ordination - samples that are far apart in the full space can end up coincident on the 2-d "best fit" plane, e.g. projected onto it "from opposite sides".

<u>Example:</u> Nematodes from Solbergstrand mesocosm experiment, GEEP Workshop (Warwick <u>et</u> <u>al.</u>, 1988).

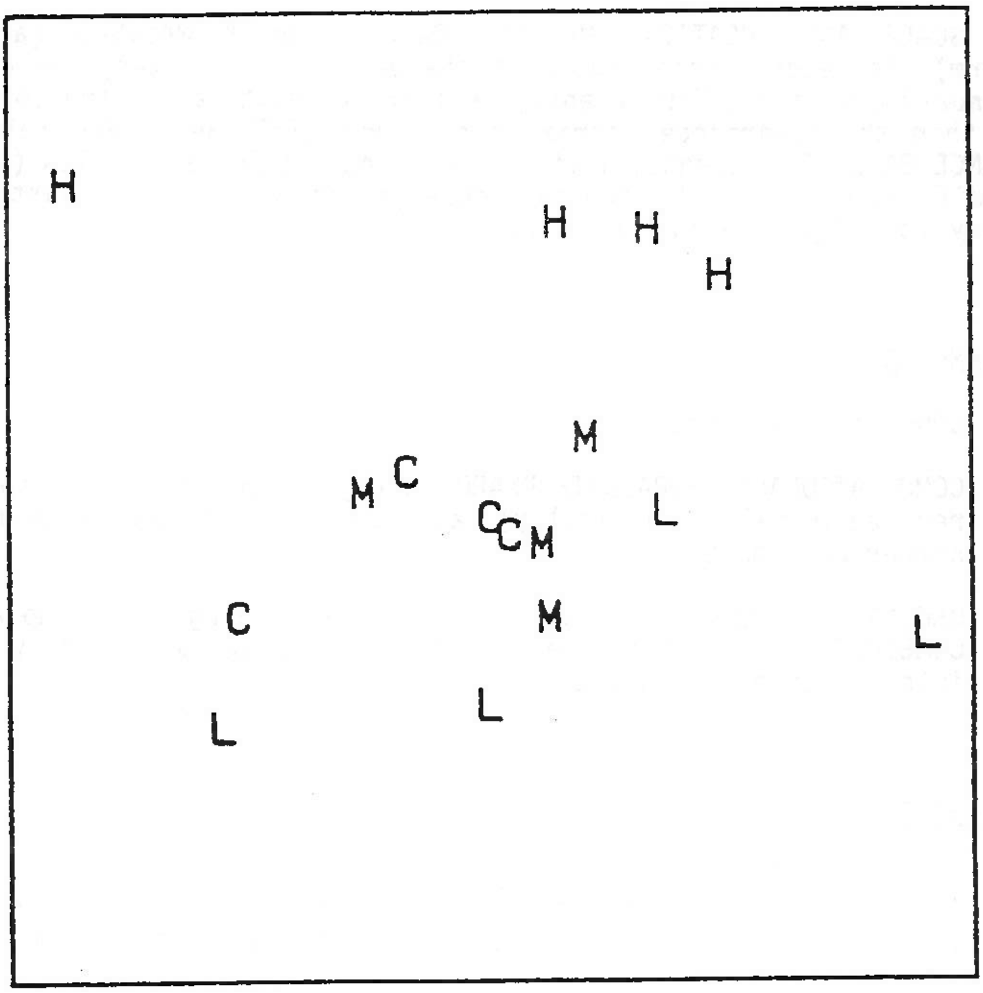


Fig. 4.2 Mesocosm nematodes. Correlation-based PCA of 16 samples: 4 replicate boxes from each of 4 treatments. (C=control, L=low, M=medium and H=high levels of diesel oil and Cu, water dosed for 11 weeks). 26 species retained (usual >3% dominance criterion) - log(1+count) transform applied. PC1 accounts for 23% of variability, PC2 15%

Strong suggestion of H replicates separating out but note low % of variability explained, so ORDINATION UNRELIABLE. (MDS gives more realistic picture - see Fig. 5.5).

# LECTURE 5

# MULTIVARIATE METHODS: ORDINATION OF SAMPLES BY MULTI-DIMENSIONAL SCALING (MDS)

## OTHER ORDINATION METHODS

PRINCIPAL CO-ODRINATES ANALYSIS (PCoA; Gower, 1966; Everitt, 1978): Also referred to as "CLASSICAL SCALING". Overcomes inflexibility of PCA by allowing WIDER RANGE of DISSIMILARITY definitions; essentially converts these to distance and does a PCA (so still subject to same PCA weakness of poor distance preservation). PCA thus a special case of PCoA, with dissimilarity = Euclidean distance.

DETRENDED CORRESPONDENCE ANALYSIS (DECORANA; Hill and Gauch, 1980): Relaxes another constraint of PCA, that of linear combinations of species. Allows CURVILINEAR COMPONENT AXES and can have effect of straightening out "horseshoe" ordinations. But:

MDS offers arguably the GREATEST FLEXIBILITY, in the sense of (lack of) assumptions made about the data.

## NON-METRIC MULTIDIMENSIONAL SCALING (MDS, e.g. Kruskal and Wish, 1978)

STARTING POINT is the (DIS)SIMILARITY MATRIX between samples (i.e. the relevant sample relationships). In fact, the ordination depends only on the RANKS of similarities in the triangular matrix, so is conceptually simple:

MDS attempts to construct a SAMPLE "MAP" (in a given number of dimensions, e.g. 2-d) using information of the form "Sample 1 is closer to Sample 4 (in species composition) than it is to Samples 2 or 3".

Example: Loch Linnhe macrofauna - subset (√√ counts)

| Year: | 64 | 68 | 71 | 73 |
|---|---|---|---|---|
| Sample: | 1 | 2 | 3 | 4 |
| Species | | | | |
| *Echin.* | 1.7 | 0 | 0 | 0 |
| *Myrio.* | 2.1 | 0 | 0 | 1.3 |
| *Labid.* | 1.7 | 2.5 | 0 | 1.8 |
| *Amaea.* | 0 | 1.9 | 3.5 | 1.7 |
| *Capit.* | 0 | 3.4 | 4.3 | 1.2 |
| *Mytil.* | 0 | 0 | 0 | 0 |

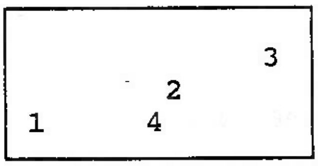
Similarities

| Sample | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | - | | | |
| 2 | 25.6 | - | | |
| 3 | 0.0 | 67.9 | - | |
| 4 | 52.2 | 68.1 | 42.0 | - |

Rank dissimilarities

| Sample | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | - | | | |
| 2 | 5 | - | | |
| 3 | 6 | 2 | - | |
| 4 | 3 | 1 | 1 | - |

MDS plot



NOTE:

1) MDS plot can be arbitrarily SCALED, LOCATED, ROTATED or INVERTED; it gives positions of samples <u>relative</u> to each other.

2) Not difficult here to place 4 points <u>in 2-d</u> with interpoint distances preserving the rank order dissimilarities exactly. Usually not possible and there will be some distortion or STRESS between (ranked) dissimilarities and corresponding distances in the plot (even in a higher-dimensional ordination).

Example: R. Exe nematodes (Field *et al.*, 1982)

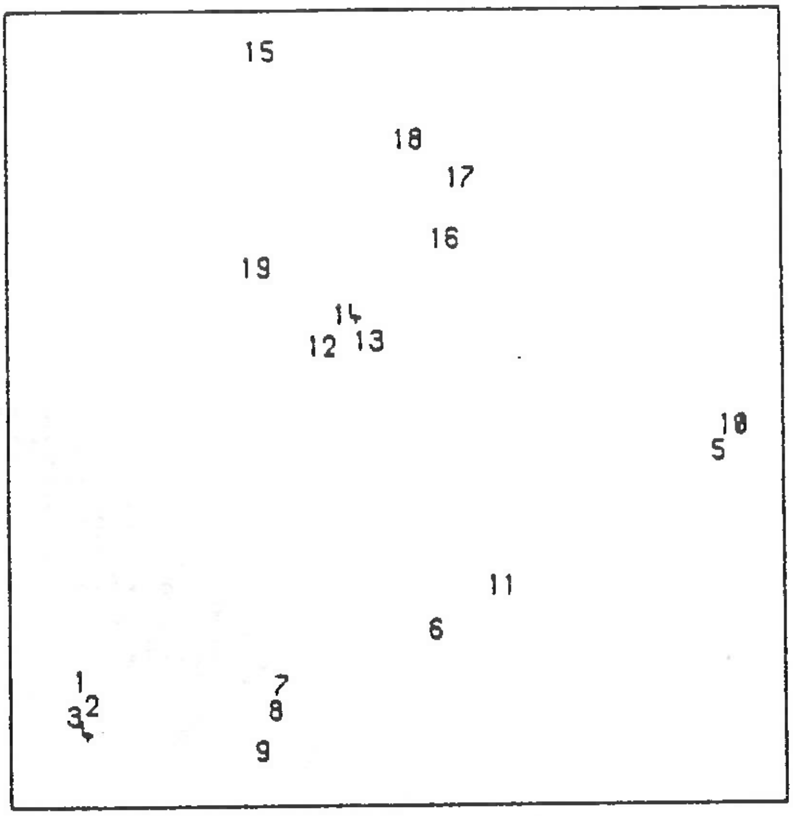


Fig. 5.1 Exe nematodes. 2-d MDS ordination of 19 sites, from bray-Curtis similarities on √√ transformed abundances (182 species)

**MDS ALGORITHM** - an iterative process

1) SPECIFY NUMBER OF DIMENSIONS for MDS plot (= m).

2) CONSTRUCT STARTING "MAP" of n samples; this could be result of (say) a PCA ordination or simply a random set of points (in m-dimensions).

3) REGRESS INTERPOINT DISTANCES $\{d_{jk}\}$ from this map on the corresponding dissimilarities $\{\delta_{jk}\}$. Can be

a) LINEAR (or CURVILINEAR) regression - METRIC MDS; or, more usually

b) MONOTONIC (increasing) regression - NON-METRIC MDS (Fig. 5.2).

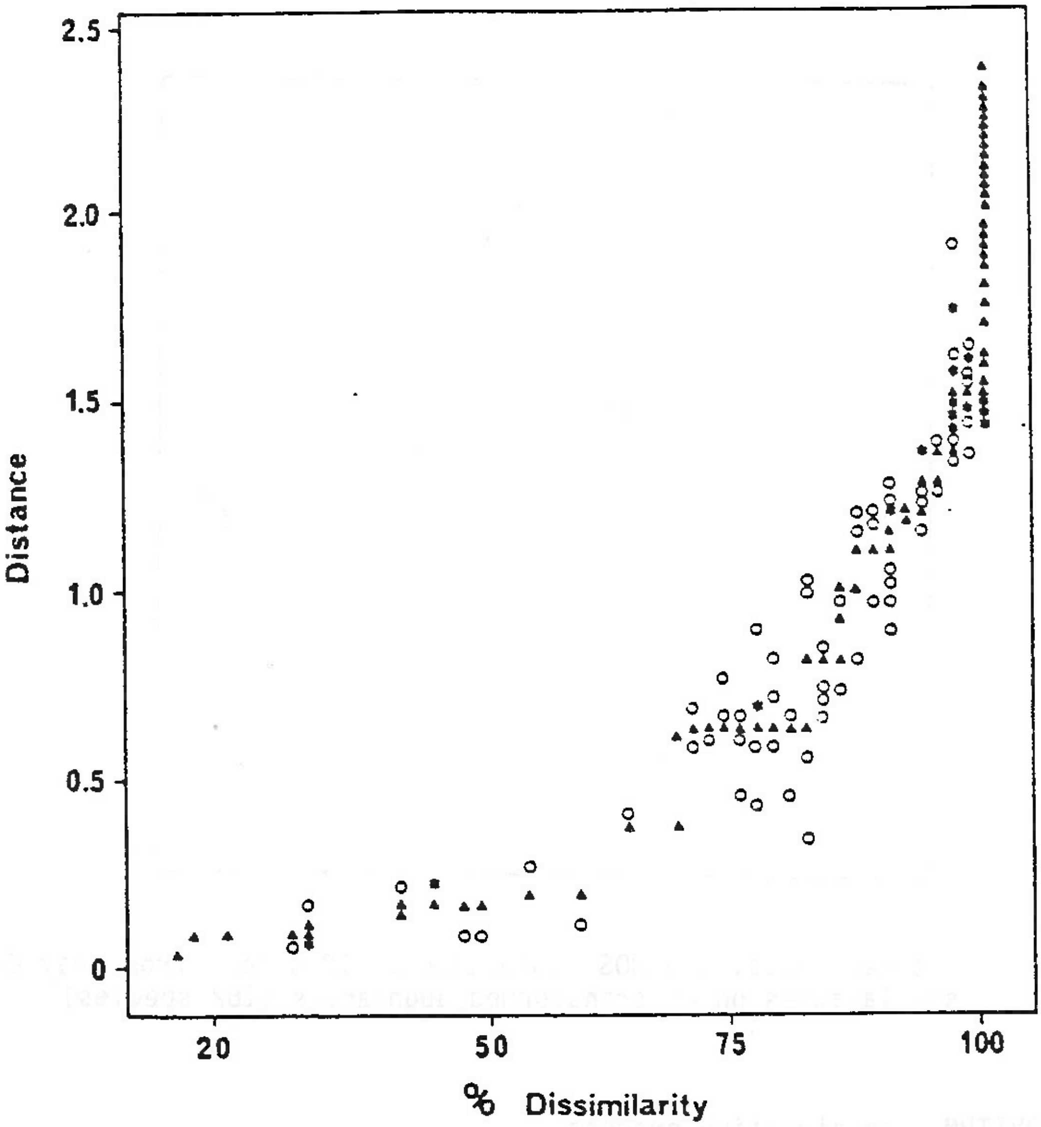


Fig. 5.2 Exe nematodes. "Shepard diagram" of distance (d) in MDS plot (Fig. 5.1) against dissimilarity ($\delta$) in Bray-Curtis matrix.
o = actual distance ($d_{jk}$),
(* = ≥2 coincident points),
▵ = fitted monotonic regression {$\hat{d}_{jk}$}.
Stress (= 0.053) is a measure of scatter about the regression line

4) MEASURE GOODNESS-OF-FIT of the regression by:

$$\text{STRESS} = \Sigma_j\Sigma_k \,(d_{jk} - \hat{d}_{jk})^2 \,/\, \Sigma_j\Sigma_k \, d^2_{jk} \qquad (5.1)$$

where $\hat{d}_{jk}$ = distance given by the fitted regression line for dissimilarity $\delta_{jk}$.

Stress = 0 if the distances preserve the rank order of the dissimilarities $\{\delta\}$.

Stress is large if the current map is poorly related to the dissimilarities $\{\delta\}$.

5) PERTURB CURRENT SAMPLE POSITIONS on the map, in direction decreasing the stress, using a STEEPEST DESCENT algorithm.

6) REPEAT STEPS 3 TO 5 (regress d on $\delta$, measure stress, perturb points) until no further reduction in stress is possible.

NOTE:

a) The algorithm is an ITERATIVE PROCEDURE so could converge to a LOCAL MINIMUM rather than a global minimum of the stress function.

Also possible to get DEGENERATE SOLUTIONS where most samples collapse to the same point, or to the vertices of a triangle, or are strung out round a circle.

REPEAT FOR DIFFERENT RANDOM STARTING CONFIGURATIONS to confirm that gives same solution (with lowest stress value) several times - this is then very likely the GLOBAL MINIMUM (though not guaranteed).

b) Unlike PCA, a 2-d MDS plot is NOT A PROJECTION of the 3-d plot. Still useful to do the 3-d MDS and use first 2 axes as the start for 2-d MDS - also useful to compare 2-d and 3-d stress values.

## ADEQUACY OF MDS REPRESENTATION

1) STRESS VALUE: This increases with increasing number of samples and decreasing dimension of the plot, but roughly speaking, in 2-d:

STRESS < 0.05 implies excellent representation,
< 0.1 good,
< 0.2 still useful, but
> 0.3 little better than random points.

(An alternative formula with a different denominator, "STRESS2", is preferred by some, but it increases the likelihood of finding local minima and is not recommended for routine use).

2) SHEPARD DIAGRAM: Scatter in this is measured by the stress value (low in Fig. 5.2, stress = 0.053, implying good MDS representation). Diagram also aids detection of "OUTLYING" POINTS and ERRORS in individual dissimilarities.

3) CONNECTION OF SIMILAR SAMPLES: Distortion in an MDS plot seen by connecting points whose similarities are in the top 10% or 20% (say) of values in the similarity matrix.

4) MINIMUM SPANNING TREE (MST): A similar idea - all points in the MDS plot are joined by a SINGLE CONNECTED LINE (which branches but is not allowed to form a closed loop) such that the sum of dissimilarities along this line is minimised; distortion is indicated by connections which look out of keeping with the distances in the plot (see Gower and Ross, 1969, for MST algorithm).

5) SUPERIMPOSITION OF GROUPS FROM CLUSTER ANALYSIS: The combination of clustering and ordination can be very effective.

<u>Example:</u> Exe nematodes, 19 sites (182 species)

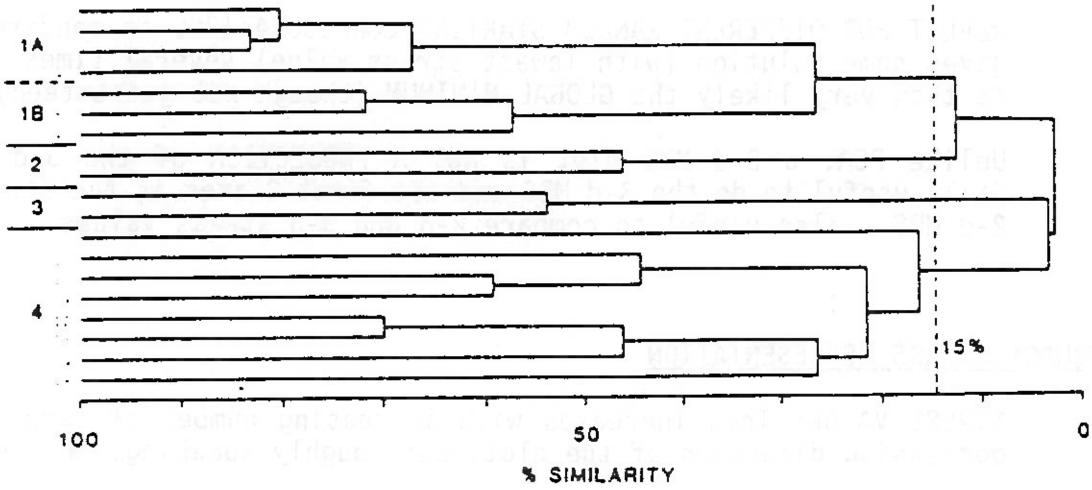


Fig. 5.3 Exe nematodes. Dendrogram (group average linking, Bray-Curtis similarities on √√ - abundance). 4 groups of sites separated by 15% similarity cut-off; 8 groups by a 30% (to 45%) threshold

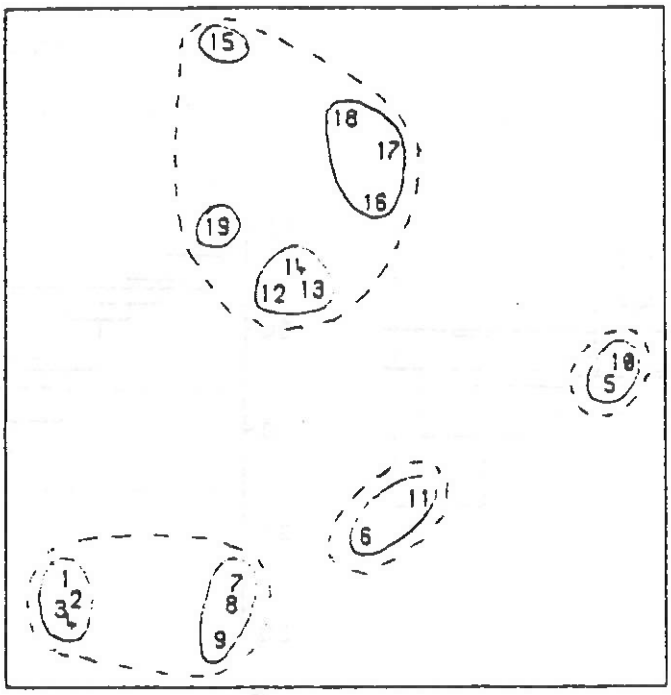


Fig. 5.4 Exe nematodes. MDS (as Fig. 5.1) with clusters indicated at: --- 15%, —— 30% similarity

Agreement clearly excellent (because clusters are sharp and MDS stress low). More revealing example provided by the data of Fig. 4.2:

<u>Example:</u> Mesocosm nematodes, GEEP Workshop.

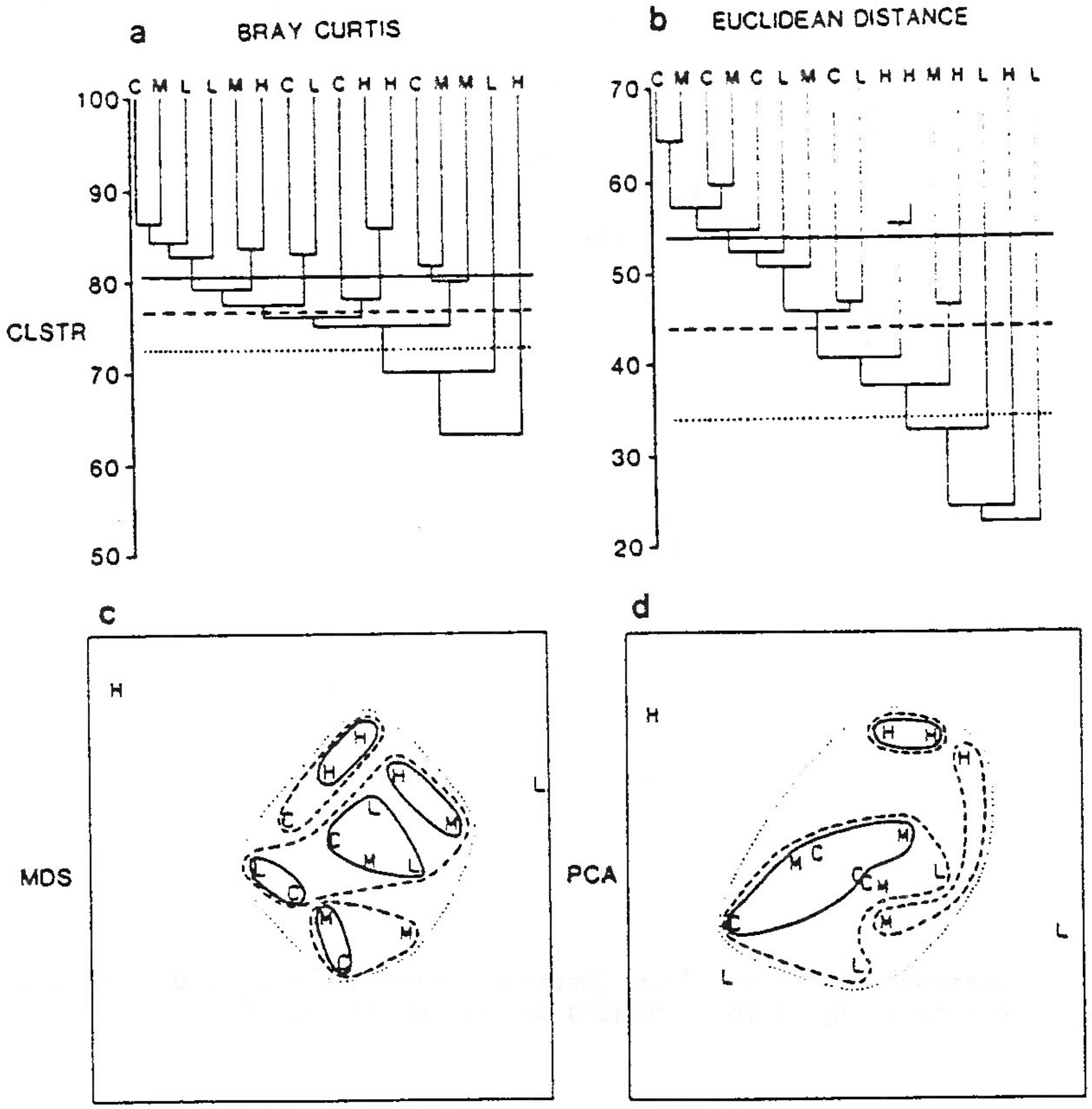


Fig. 5.5 Mesocosm: 4 replicates from 4 treatments (reduced species and log transform, as Fig. 4.2).
a), c) Group-average clustering from Bray-Curtis similarities; clusters formed at 3 (arbitrary) levels superimposed on the MDS obtained from the same similarities (stress = 0.19).
b), d) Group average clustering from "Euclidean distance" (dis)similarities superimposed on the PCA (Fig. 4.2). (Euclidean distance is the dissimilarity measure implicit in a PCA ordination)

NOTE:

1) Though no natural groupings are apparent from the MDS, the Bray-Curtis cluster and MDS analyses (a and c) are not really inconsistent.

2) The PCA and its corresponding cluster analysis (d and b) are in disagreement, indicating that the 2-d PC axis is a distorted representation of the true "distances" between samples.

ORDINATION v CLUSTERING: Strength of ordination is in displaying GRADATION (rather than categorisation) of community composition in a set of samples.

Example: Celtic Sea zooplankton (Collins, pers. comm.)

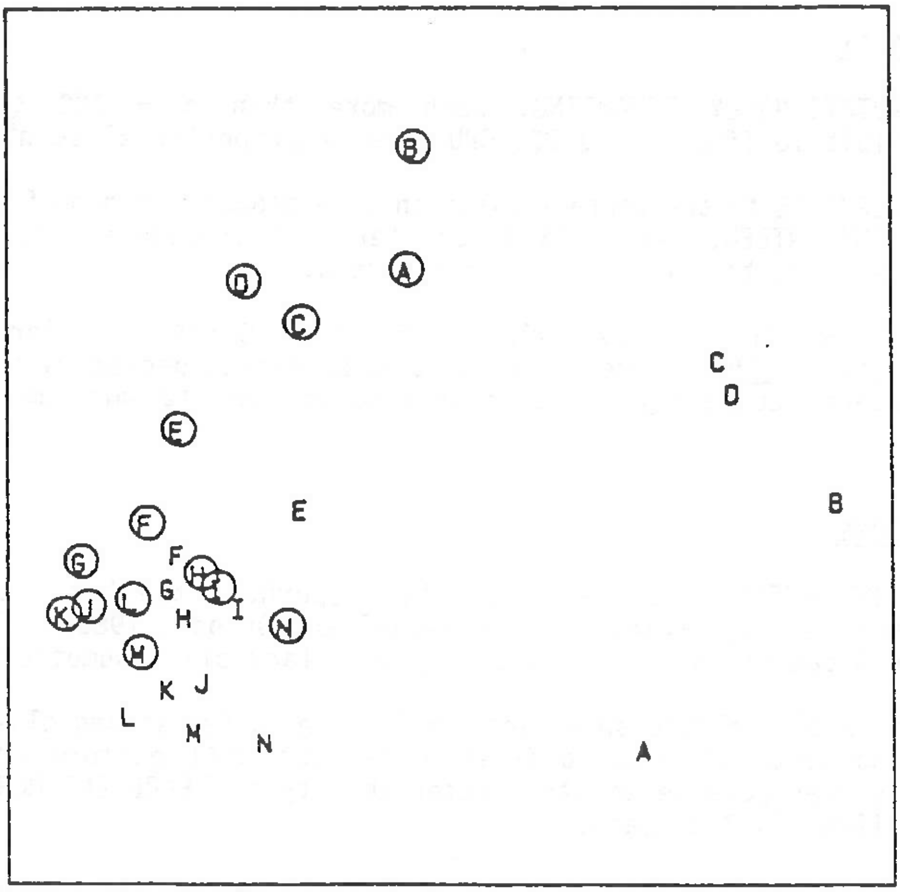


Fig. 5.6 MDS of zooplankton samples at a single site (22/9/78), from 14 depths (5 m to 70 m, denoted A,B,..,N) for night (circles) and day-time hauls

**MDS STRENGTHS:**

1) SIMPLE in concept.

2) BASED ON RELEVANT INFORMATION. It can be used with the most appropriate measure of (dis)similarity for the particular data.

3) SPECIES DELETIONS UNNECESSARY for an ordination of samples (any exclusion dividing line is inevitably arbitrary). The similarity measure can automatically weight rarer species appropriately (and can be chosen to ignore joint absences).

4) GENERALLY APPLICABLE. Since MDS uses only rank order of dissimilarities it makes the weakest possible assumptions about quality of the data.

5) SIMILARITIES CAN BE GIVEN UNEQUAL WEIGHT in constructing the MDS plot (e.g. some samples may be more reliable, perhaps because they are based on combining more replicates).

**MDS WEAKNESSES:**

1) COMPUTATIONALLY DEMANDING; much more than n = 100 samples is prohibitive (fewer on a PC; CPU time is proportional to $n^2$).

2) CONVERGENGE to the correct solution (the global minimum of stress) is NOT GUARANTEED, since MDS is an iterative procedure; the necessary repeats add to the computational burden.

3) ALGORITHM PLACES MOST WEIGHT ON LARGE DISTANCES. For detailed structure within large clusters it is sometimes necessary to ordinate clusters separately (same constraint applies to most methods, eg. PCA).

**RECOMMENDATIONS:**

1) MDS RECOMMENDED as one of the best (perhaps the best) ordination technique (e.g. Everitt, 1978; Kenkel and Orloci, 1986). Preferable to PCA because of its flexibility and (lack of) assumptions.

2) When sample relationships are simple (e.g. a few strong clusters; one strong gradient) most ordination methods will perform adequately. MDS scores because of its greater ability to REPRESENT MORE COMPLEX RELATIONS in 2-d space.

3) If stress is low (say, <0.1), an MDS ordination is probably a more useful representation than a cluster analysis, even when the samples are strongly grouped. However, the techniques complement each other, so PERFORM BOTH, AND VIEW THEM IN COMBINATION, especially for higher stress. (In the latter case also try a higher-dimensional ordination).

## LECTURE 6

## MULTIVARIATE METHODS: TESTING FOR DIFFERENCES BETWEEN GROUPS OF SAMPLES

DISTINGUISHING SITES (or TIMES) by formal significance tests is a necessary first step to INTERPRETING differences (e.g. control v. impacted site) but usually overlooked for multivariate methods (because of unavailability of suitable tests).

(Note: Cluster analysis will always find clusters, even from random data points!)

### UNIVARIATE TESTS

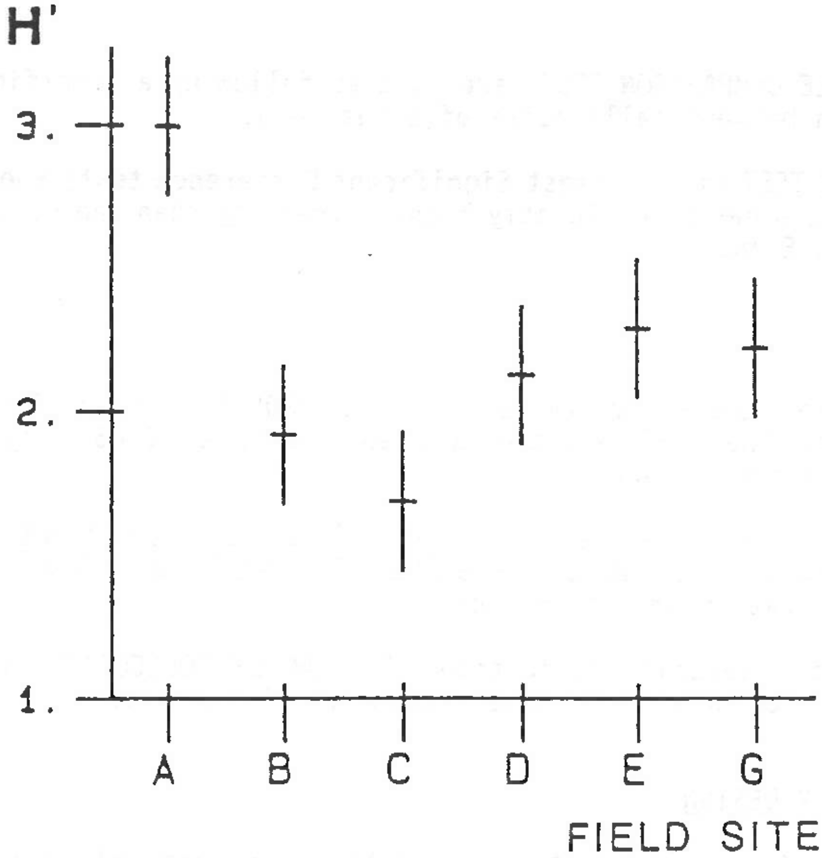


Fig. 6.1 Frierfjord macrofauna. Means and 95% confidence intervals for Shannon diversity (H′) at 6 field sites

ONE-WAY ANOVA provides a test of the (null) hypothesis:

$H_o$: No difference between sites

It assumes normality of H′ and constant variance across sites (hence the confidence intervals in Fig. 6.1 use a pooled variance estimate and are of the same widths).

Table 6.1

Frierfjord macrofauna diversity H′; ANOVA.

| | Sum of squares | Deg. of freedom | Mean Square | F ratio | Sig. level |
|---|---|---|---|---|---|
| Treatments | 3.938 | 5 | 0.788 | 15.1 | <0.1% |
| Residual | 0.937 | 18 | 0.052 | | |
| Total | 4.874 | 23 | | | |

MULTIPLE COMPARISON TESTS are used to follow up a significant F-test with comparison between (all) pairs of sites, e.g. ``

TUKEY T TEST (i.e. a Least Significant Difference test) shows that the "reference" site A has significantly higher diversity than the rest, and C has a lower H′ than E and G.

NOTE:

1) Multiple comparison tests FIX the PROBABILITY of TYPE I ERROR ("reject the null hypothesis when true") at 0.05 (say) over all pairwise comparisons.

2) Global F-test is best thought of as a "red light" - unless significant it BARS PROGRESS TO PAIRWISE COMPARISONS and interpretation of differences.

3) There are several implications for SAMPLE COLLECTION, which apply equally to the multivariate testing which follows:

IMPLICATIONS FOR DESIGN

1) CONTROL (reference) site(s) essential - impact only established by reference to similar unimpacted site(s), or to same site pre-impact. (Preferable to have both spatial and temporal controls).

2) REPLICATION at each site essential - should be over appropriate spatial scale (i.e. genuinely representative of that location).

3) "BLIND" ANALYSIS desirable - avoids (unconscious) biases, e.g. tendency to uniformity of replicates.

MULTIVARIATE TESTS

INFORMAL: CLUSTER, MDS, etc. assume no knowledge of how samples are divided into sites. So, plots can be inspected for evidence of REPLICATE GROUPING.

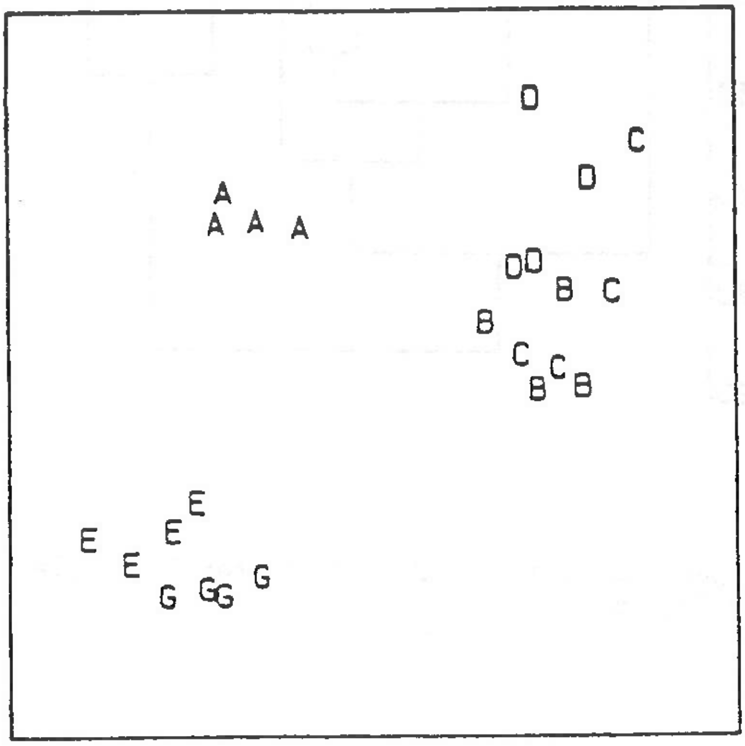


Fig. 6.2 Frierfjord macrofauna. MDS plot (Bray-Curtis similarities, √√ transform), for 24 samples, 4 replicates from each of sites A-E,G

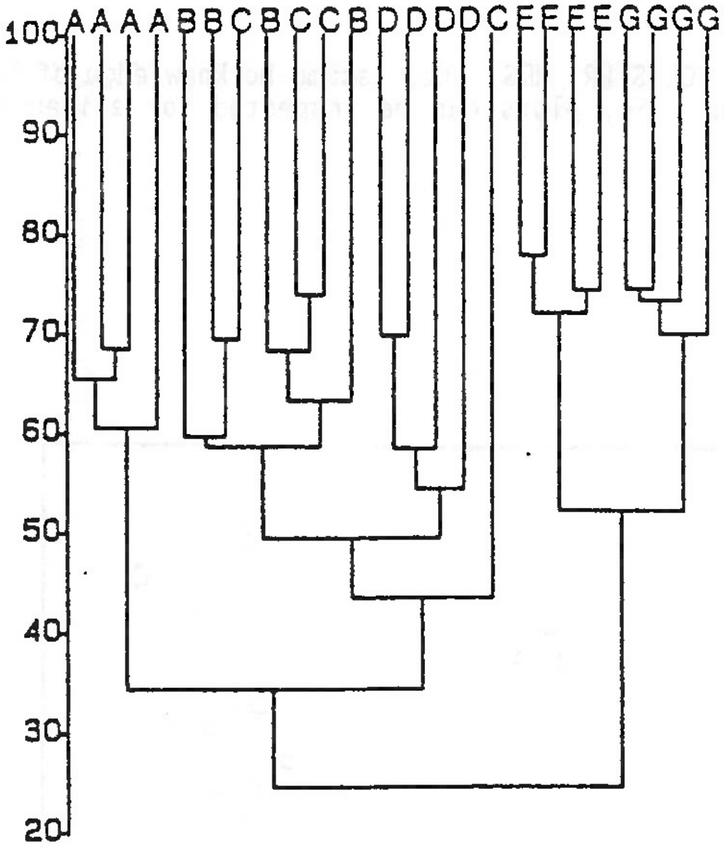


Fig. 6.3 Frierfjord macrofauna. Dendrogram for 24 samples (similarities as for Fig. 6.2)

PARAMETRIC TESTS

EXACT ANALOGUE OF ONE-WAY ANOVA is multivariate analysis of variance (MANOVA), the F-test being replaced by WILKS' Λ test (e.g. Mardia, 1979). Pairwise differences can be tested by MAHALANOBIS' DISTANCES (e.g. Seber, 1984); but

ASSUMPTIONS RARELY SATISFIED: Tests require multivariate normality of abundances and "large samples" (at each site). For Frierfjord macrofauna, even after reduction to 30 species:

a) 50% of abundances are zero - normality impossible (even with transform),

b) ratio of observations to parameters needing estimation is 1.1 - hardly large!

RANDOMISATION/PERMUTATION TESTS:

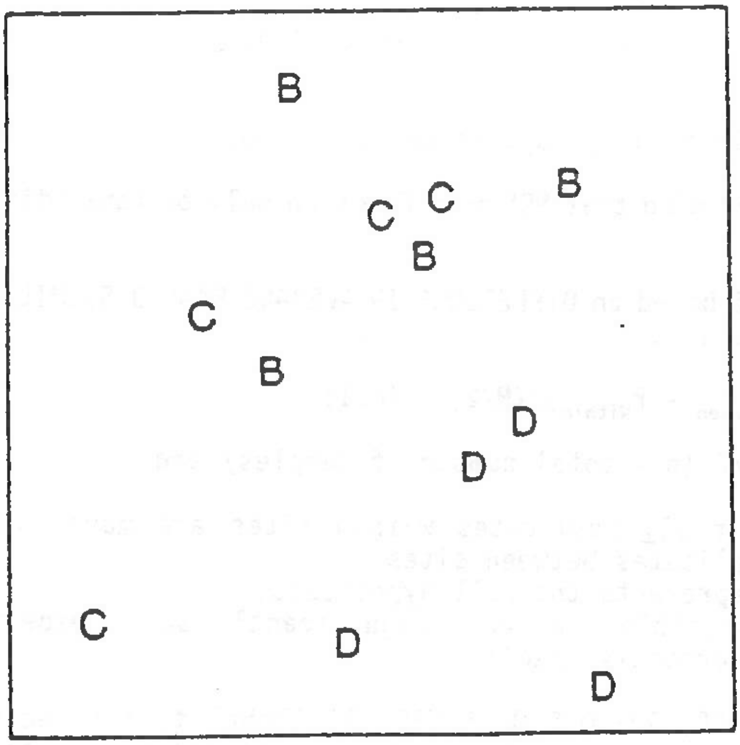


Fig. 6.4 Frierfjord macrofauna. MDS plot (Bray-Curtis, $\sqrt{\sqrt{}}$ transform) of 4 replicates from B,C,D

NULL HYPOTHESIS $H_o$: no difference between sites. If $H_o$ false, distances between replicates within sites are less than distances across sites. So:

1) COMPUTE STATISTIC reflecting this difference. To derive its sampling distribution, note that when $H_o$ true, the 12 labels (4 B's, 4 C's, 4 D's) could be allocated at random to the 12 MDS points. So:

2) RECOMPUTE STATISTIC under ALL POSSIBLE PERMUTATIONS of the 12 labels between the 12 MDS points, or (since that is prohibitive) under a LARGE NUMBER OF RANDOM ALLOCATIONS of the 12 labels to the points.

3) RANDOMISATION/PERMUTATION TEST will reject $H_o$ AT 5% SIGNIFICANCE LEVEL if observed statistic greater than its value for 95% of the random relabellings.

FORM OF DISPLAY SHOULD BE IRRELEVANT: Desirable that the statistic has exactly the same value whether the representation is:

a) a dendrogram (Fig. 6.3)

b) an MDS for all 6 sites (Fig. 6.2) or just a subset of sites (Fig. 6.4)

c) an MDS in 3-d, say, rather than 2-d.

Bearing in mind that MDS is a function only of rank (dis)similarities, this suggests:

STATISTIC based on DIFFERENCE IN AVERAGE RANK DISSIMILARITIES between and within sites, i.e.

$$R = (\bar{r}_{\text{Between}} - \bar{r}_{\text{Within}})/(M/2) \qquad (6.1)$$

where $M = n(n-1)/2$ ($n$ = total number of samples) and:

R = 1 if <u>all</u> replicates within sites are more similar than <u>any</u> replicates between sites.
R = 0 represents the null hypothesis.
(R < 0 possible, but only significantly so if experimental design incorrectly specified).

PAIRWISE COMPARISONS OF SITES: If global test rejects $H_o$ then same type of test can be carried out on each pair of sites, though note:

a) These tests must be treated with some caution since NOT true "MULTIPLE COMPARISON" TESTS; overall Type I error not controlled.

b) Minimum of 4 replicates per site needed for pairwise tests. Can be fewer for global test since NUMBER OF DISTINCT PERMUTATIONS is:

$$(\Sigma_i n_i)!/(n_i!n_2!\ldots n_k!k!) \qquad (6.2)$$

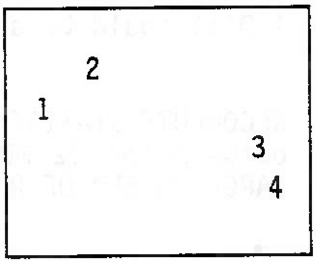
where $\{n_i\}$ replicates at site i (i=1,2,..,k).

<u>Example:</u> 2 replicates at each of 2 sites (A,B)

| | A | A | B | B |
|---|---|---|---|---|
| Sample | 1 | 2 | 3 | 4 |
| A 1 | - | | | |
| A 2 | 2 | - | | |
| B 3 | 4 | 3 | - | |
| B 4 | 6 | 5 | 1 | - |

Rank dissimilarities

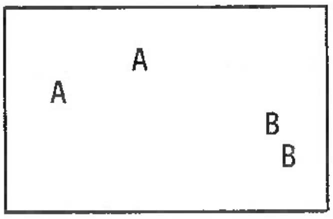
——>

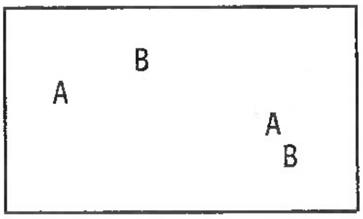


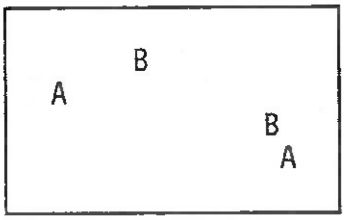
MDS (1,2 = A; 3,4 = B)

$\bar{r}_{Between} = 4.5$, $\bar{r}_{Within} = 1.5$, M = 6, so R = 1.

Only three possible distinct PERMUTATIONS OF LABELS:







|  | A | A | B | B |
|---|---|---|---|---|
| Smp. | 1 | 2 | 3 | 4 |
| A 1 | - |  |  |  |
| A 2 | 2 | - |  |  |
| B 3 | 4 | 3 | - |  |
| B 4 | 6 | 5 | 1 | - |

R = 1

|  | A | A | B | B |
|---|---|---|---|---|
| Smp. | 1 | 3 | 2 | 4 |
| A 1 | - |  |  |  |
| A 3 | 4 | - |  |  |
| B 2 | 2 | 3 | - |  |
| B 4 | 6 | 1 | 5 | - |

R = -0.5

|  | A | A | B | B |
|---|---|---|---|---|
| Smp. | 1 | 4 | 2 | 3 |
| A 1 | - |  |  |  |
| A 4 | 6 | - |  |  |
| B 2 | 2 | 5 | - |  |
| B 3 | 4 | 1 | 3 | - |

R = -0.5

Observed case (R = 1) has 33% probability of occurring by chance, so could not reject the hypothesis of "no difference between sites" (even though the observed case is the most extreme possible, here)

A more realistic example, where there are 12 samples divided between 3 sites (and thus 12!/(4!4!4!3!) = 5775 possible permutations) is given by Fig. 6.4:

Example: Frierfjord macrofauna abundances.

Table 6.2

Frierfjord macrofauna. Ranked dissimilarity matrix (Bray-Curtis, √√ transform) between the 12 replicates from sites B,C,D.

|  | B1 | B2 | B3 | B4 | C1 | C2 | C3 | C4 | D1 | D2 | D3 | D4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B1 | - |  |  |  |  |  |  |  |  |  |  |  |
| B2 | 33 | - |  |  |  |  |  |  |  |  |  |  |
| B3 | 8 | 7 | - |  |  |  |  |  |  |  |  |  |
| B4 | 22 | 11 | 19 | - |  |  |  |  |  |  |  |  |
| C1 | 66 | 30 | 58 | 65 | - |  |  |  |  |  |  |  |
| C2 | 44 | 3 | 15 | 28 | 29 | - |  |  |  |  |  |  |
| C3 | 23 | 16 | 5 | 38 | 57 | 6 | - |  |  |  |  |  |
| C4 | 9 | 34 | 4 | 32 | 61 | 10 | 1 | - |  |  |  |  |
| D1 | 48 | 17 | 42 | 56 | 37 | 55 | 51 | 62 | - |  |  |  |
| D2 | 14 | 20 | 24 | 39 | 52 | 46 | 35 | 36 | 21 | - |  |  |
| D3 | 59 | 49 | 50 | 64 | 54 | 53 | 63 | 60 | 43 | 41 | - |  |
| D4 | 40 | 12 | 18 | 45 | 4 | 27 | 26 | 31 | 25 | 2 | 13 | - |

GLOBAL TEST:

$\bar{r}_{Between} = 37.54$, $\bar{r}_{Within} = 22.72$, $M = 66$, so $R = 0.45$.

In 500 random relabellings, none of them gave R>0.45, so $H_o$ rejected at significance level p<.002 (0.2%).

PAIRWISE TESTS:

For each pair of sites, the corresponding subset of the above triangular matrix is extracted, re-ranked and R computed as above, e.g. for B v C, R = 0.23. This time, R can be re-evaluated for all possible relabellings, giving p<12%, so B & C not significantly different (only 35 distinct permutations, so the maximum attainable significance level is 3%).

However, D does differ from B and C (B v D: R = 0.54, p<3%, C v D: R = 0.57, p<3%).

FURTHER FEATURES AND EXTENSIONS:

1) PERMUTATION TEST CONCEPT dates to Mantel (1967) and general Monte Carlo (randomisation) tests discussed by Hope (1968). Practicals use a FORTRAN program called ANOSIM (Analysis of Similarities).

2) ANOSIM test makes NO ASSUMPTION OF "EQUAL VARIANCE"

Example: Coral communities at South Tikus, Thousand Is., Indonesia (Warwick, Clarke & Suharsono, 1990).

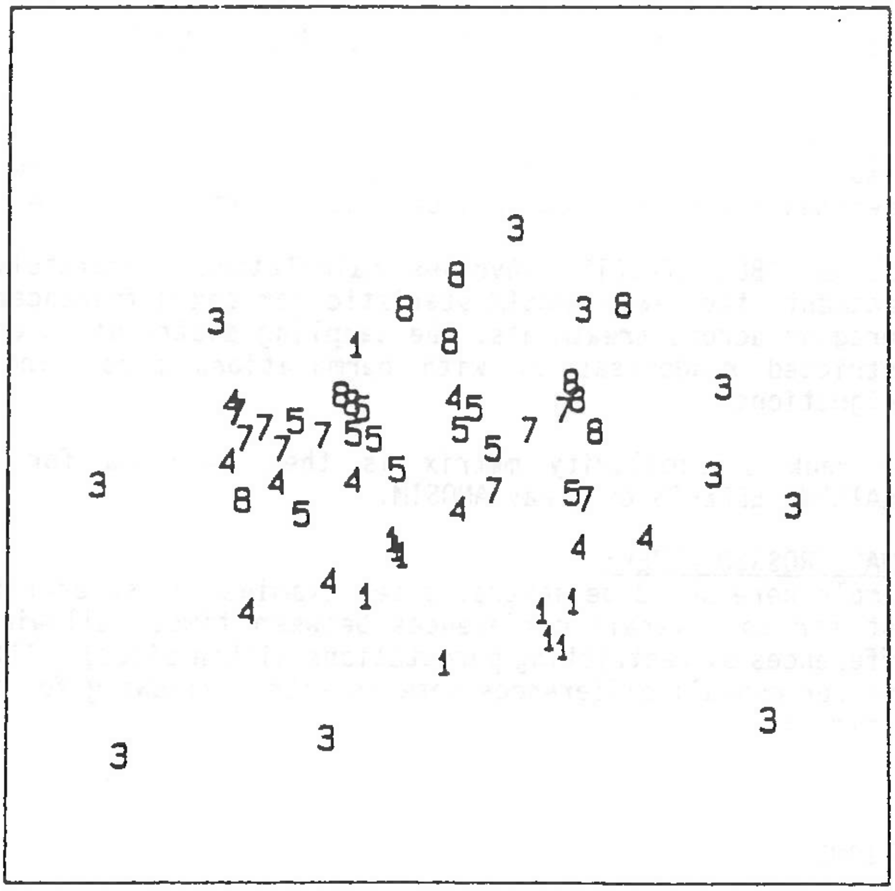


Fig. 6.5 MDS for % cover of coral species (Bray-Curtis, no transform) for 10 replicates in each of 5 years: 1 = 1981 (pre-El Niño), 3 = 1983 etc

ANOSIM test distinguishes the clear difference in initial and impacted conditions (1 and 3), though change is largely in variance rather than location.

3) ANOSIM TEST NOT RESTRICTED TO BALANCED REPLICATION at sites (or times); some sites can even have only one replicate provided enough replicates overall to generate sufficient permutations (eqt. (6.2)).

4) WIDE APPLICABILITY in that ANOSIM can be used with any (dis)similarity matrix; e.g. for a Euclidean distance matrix (appropriate to a PCA) ANOSIM can be seen as a non-parametric alternative to the parametric Wilks' Λ test for a MANOVA, though it:

5) LACKS SENSITIVITY (as with many non-parametric tests) in the (unlikely) event that the data is genuinely multivariate normal.

6) ANOSIM PROGRAM EXTENDS TO ANALOGUE OF 2-WAY ANOVA:

<u>2-WAY NESTED MODEL:</u>
Example is Oslo Workshop macrofauna data from the mesocosm experiment: 2 cores from each of 4 boxes from each of 4 treatments.

TEST OF "BOX EFFECTS" involves calculating, separately for each treatment, the 1-way ANOSIM statistic for box differences, and then averaging across treatments. The sampling distribution comes from a restricted randomisation, with permutations preserving treatment designations.

The rank dissimilarity matrix is then reformed for a TEST OF TREATMENT EFFECTS by 1-way ANOSIM.

<u>2-WAY CROSSED MODEL:</u>
Example here would be several sites examined at several times. Can test for any overall differences between times (allowing for site differences by restricting permutations within sites). Alternatively test for overall differences between sites (allowing for differences in times).

## <u>RECOMMENDATIONS</u>

1) USE RANDOMISATION/PERMUTATION TEST (ANOSIM) rather than parametric methods for testing of multivariate differences between <u>previously-defined</u> groups of samples (i.e. sites, times, treatments etc.); its ROBUSTNESS (lack of assumptions) more than makes up for its CONSERVATISM - latter is not so bad anyway. (Note: <u>cannot</u> test if differences between groups of samples are 'significant', if the grouping came from multivariate analysis of that <u>same</u> data).

2) USE NORMALITY-BASED TESTS for univariate INDICES, after any necessary transform (see lecture 9).

# LECTURE 7

# MULTIVARIATE METHODS: SPECIES ANALYSES

## SPECIES CLUSTERING

Clustering methods can be applied to SPECIES SIMILARITY matrices (latter defined on pages 85-86).

Example: R. Exe (UK) nematodes, Field *et al.* (1982)

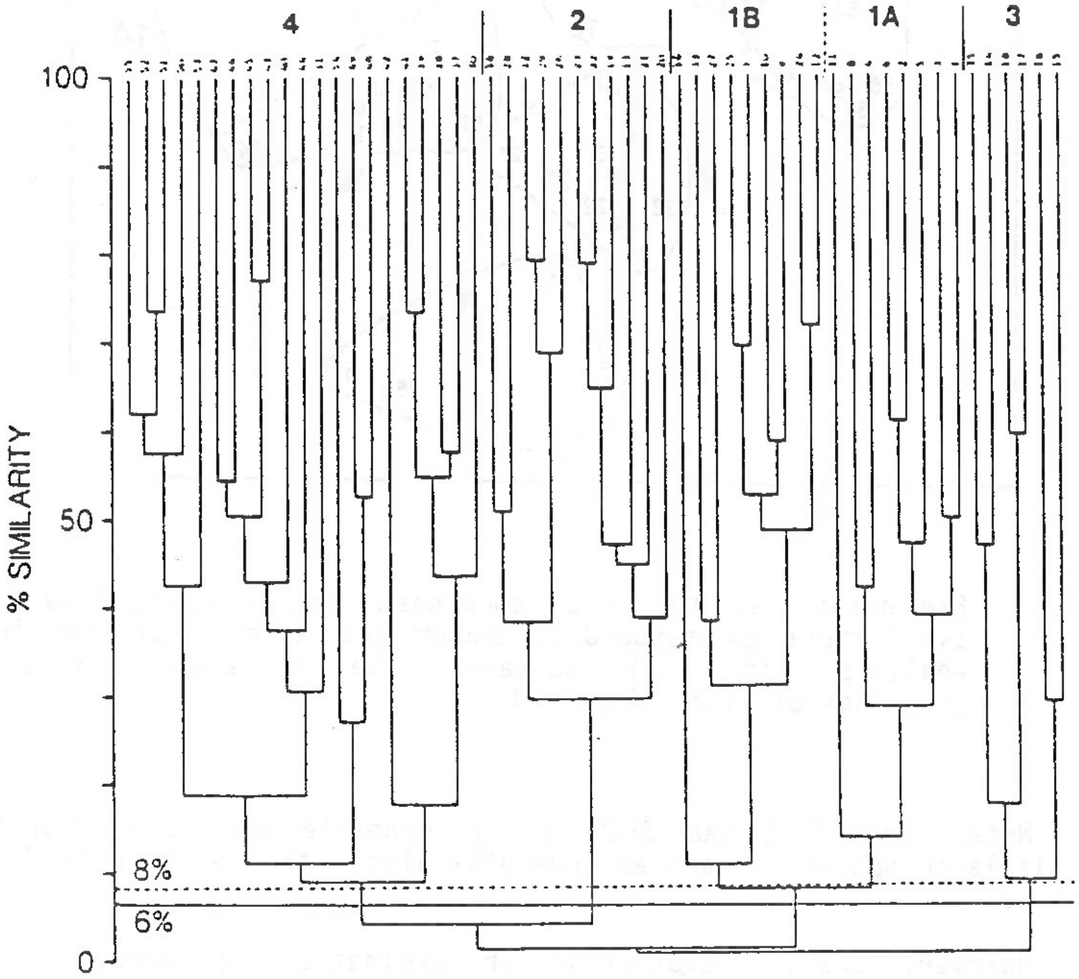


Fig. 7.1 Exe nematodes. Dendrogram (group average link) from Bray-Curtis similarities (standardised abundance data) for 55 species from 19 sites - reduced from 182 species by including those with counts >4% of total at any one site. The 4 to 5 groups indicated correspond closely with sharply defined clusters in the sites analysis (Fig. 5.3)

## SPECIES MDS

A species similarity matrix can also be input to an MDS, in the same way as for samples. In practice, often gives high 2-d stress. As with clustering, works best when samples form strong groups, arising from species sets which tend to be exclusive.

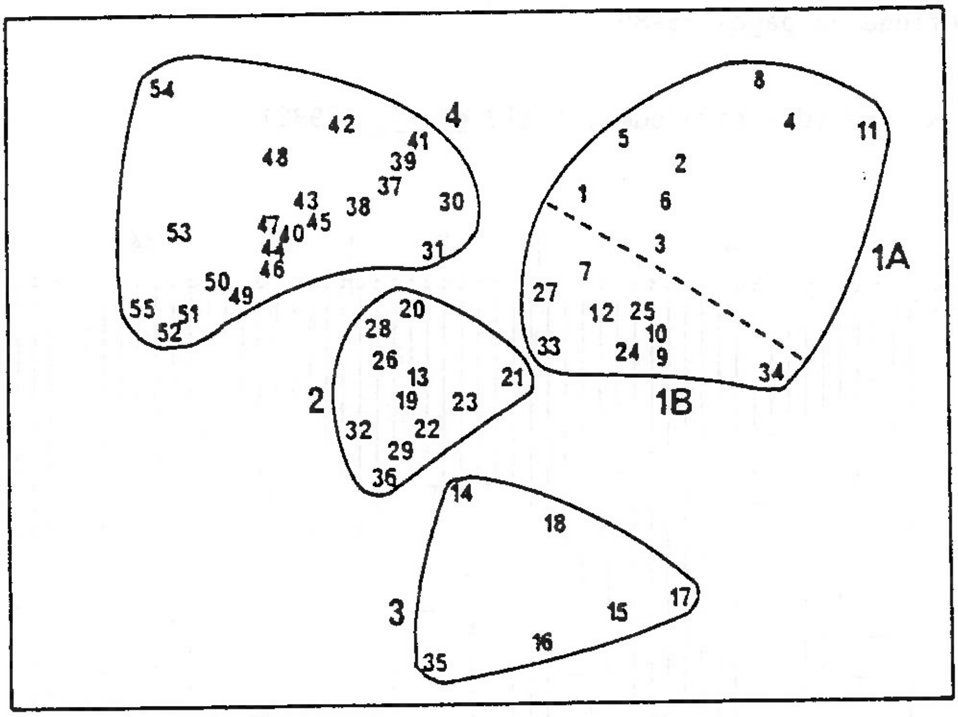


Fig. 7.2 Exe nematodes. MDS of 55 commonest species using Bray-Curtis similarities on standardised abundances. Main groups from cluster analysis (Fig. 7.1) indicated; they correspond closely to groupings of sites (Fig. 5.4)

Note: The LESS-COMMON SPECIES will generate erratic similarities, giving isolated MDS points and an unhelpful plot - they need to be REMOVED initially.

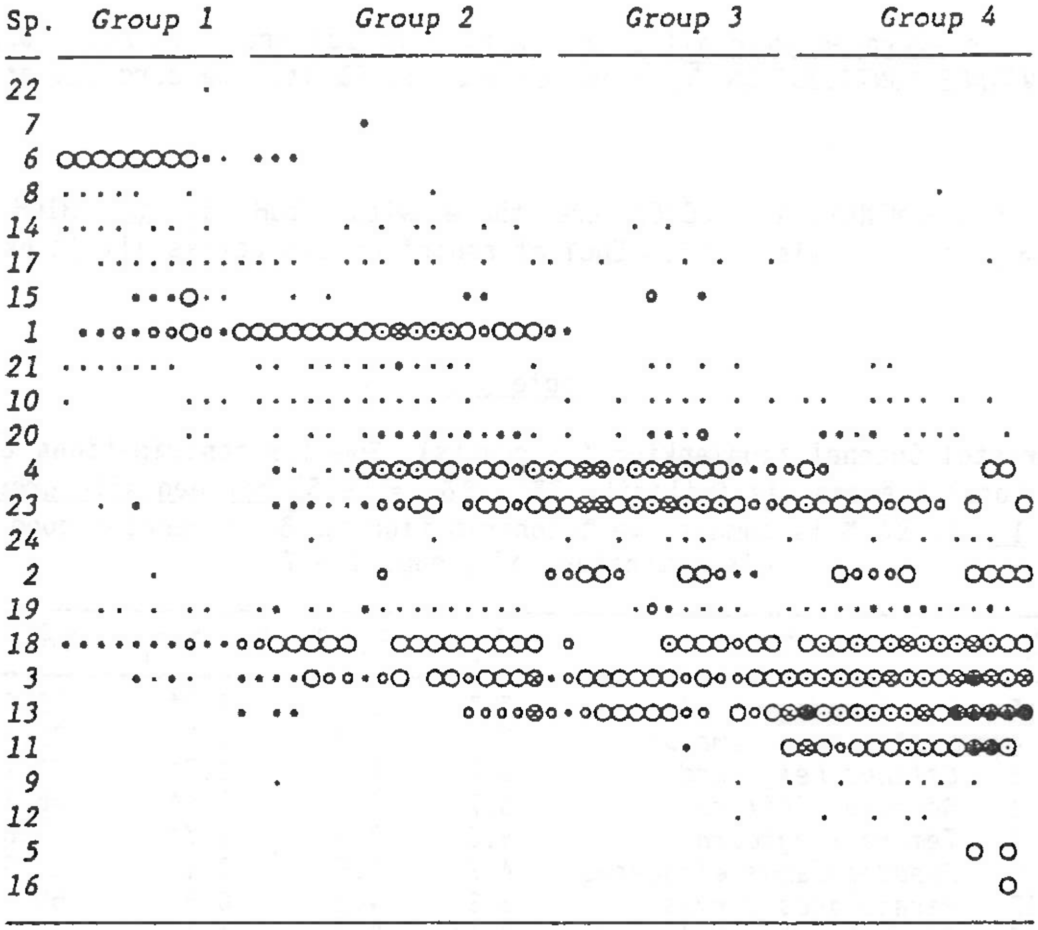
However, SPECIES clustering or ordination is generally less informative than methods which HIGHLIGHT SPECIES contributing to pattern of SAMPLE clustering or ordination:

## DETERMINING DISCRIMINATING SPECIES

Given clear CLUSTERING of SAMPLES, what methods will determine SPECIES RESPONSIBLE for groupings? Hard to see patterns in the original data matrix, so:

RE-ORDER COLUMNS (samples) and ROWS (species) to match groupings from site and species clustering and MDS. CATEGORISE counts/biomass and represent by symbols of increasing size & density, to give SHADE MATRIX.

Example: Bristol Channel zooplankton, April 1978.



*Group 1: 1,2,4,5,3,6,7,8,10,12*
*Group 2: 9,24,13,19,27,17,11,20,15,16,14,21,18,25,29,22,26,23*
*Group 3: 42,34,48,49,50,53,44,43,33,35,54,55,47,31*
*Group 4: 51,41,45,37,32,36,38,57,56,58,28,39,40,46,52*

Fig. 7.3 SHADE matrix for 24 species X 57 sites. Site groups determined by clustering of Fig. 3.3; symbols denote increasing (√√ - transformed) counts

Alternative is to BREAK DOWN average DISSIMILARITY ($\bar{\delta}$) between two groups of samples into CONTRIBUTIONS from each SPECIES - revealing GOOD DISCRIMINATORS.

From (2.11), contribution to $\delta_{jk}$ from ith species is:

$$\delta_{jk}(i) = 100.|y_{ij} - y_{ik}| / \textstyle\sum_{i=1}^{p} (y_{ij} + y_{ik}) \qquad (7.1)$$

$\delta_{jk}(i)$ then averaged over all pairs (with j in 1st and k in 2nd group), to give AVERAGE CONTRIBUTION $\bar{\delta}_i$ from ith species (& its standard deviation SD ($\delta_i$)).

DISCRIMINATING SPECIES are those with HIGH $\bar{\delta}_i$ and HIGH ratio $\bar{\delta}_i$/SD($\delta_i$) (this implies CONSISTENCY of contributions across all jk pairs).

Table 7.1

Bristol Channel zooplankton ($\surd\surd$ counts). Species contributions $\bar{\delta}_i$ to total average dissimilarity ($\bar{\delta} = \Sigma\bar{\delta}_i = 59.5$) between site groups 1 & 2; $\Sigma\bar{\delta}_i$% is cumulative % contribution to $\bar{\delta}$. * denotes good discriminators of groups 1 & 2.

| Sp. | Name | $\bar{\delta}_i$ | SD($\delta_i$) | $\bar{\delta}_i$/SD($\delta_i$) | $\Sigma\bar{\delta}_i$ % |
|---|---|---|---|---|---|
| 6 | *Eurytemora affinis* | 7.7 | 2.8 | 2.7* | 13.0 |
| 4 | *Centropages hamatus* | 7.3 | 4.4 | 1.7* | 25.2 |
| 3 | *Calanus helgolandicus* | 6.8 | 4.0 | 1.7* | 36.7 |
| 1 | *Acartia bifilosa* | 5.7 | 4.0 | 1.4* | 46.3 |
| 23 | *Temora longicornis* | 5.6 | 3.3 | 1.7* | 55.6 |
| 18 | *Pseudocalanus elongatus* | 4.7 | 1.5 | 3.1* | 63.5 |
| 13 | *Paracalanus parvus* | 3.3 | 4.2 | 0.8 | 69.1 |
| 15 | *Pleurobrachia pileus* jv | 3.1 | 2.8 | 1.1 | 74.3 |
| 20 | *Sagitta elegans* jv | 2.9 | 1.9 | 1.6* | 79.1 |
| 19 | *Sagitta elegans* jv | 2.1 | 1.6 | 1.3 | 82.5 |
| 8 | *Gastrosaccus spinifer* | 2.0 | 1.8 | 1.1 | 85.9 |
| 14 | *Pleurobrachia pileus* | 1.9 | 1.6 | 1.2 | 89.0 |
| 10 | *Mesopodopsis slabberi* | 1.7 | 1.4 | 1.3 | 91.9 |
| 21 | *Schistomysis spiritus* | 1.6 | 1.4 | 1.1 | 94.5 |
| 17 | *Polychaete larvae* | 1.5 | 1.3 | 1.2 | 97.1 |
| 2 | *Acartia clausi* | 0.7 | 1.8 | 0.4 | 98.3 |
| . | ................. | ... | ... | ... | .... |

Can similarly compute the contribution of the ith species ($\bar{S}_i$) to the AVERAGE SIMILARITY WITHIN A GROUP ($\bar{S}$), using the 2nd form of (2.1). This highlights species consistently prominent in that group (i.e. HIGH $\bar{S}_i$, HIGH ratio $\bar{S}_i$/SD($S_i$)).

Table 7.2

Zooplankton. Species contribution ($\bar{S}_i$) to average similarity ($\bar{S}$ = 66.3) within site group 1.

| Sp. | Name | $\bar{S}_i$ | SD($S_i$) | $\bar{S}_i$/SD($S_i$) | $\Sigma\bar{S}_i$ % |
|---|---|---|---|---|---|
| 6 | *Eurytemora affinis* | 19.3 | 6.3 | 3.1* | 29.1 |
| 18 | *Pseudocalanus elongatus* | 14.7 | 2.7 | 5.4* | 51.3 |
| 1 | *Acartia bifilosa* | 12.2 | 6.4 | 1.9* | 69.6 |
| 17 | *Polychaete larvae* | 3.9 | 3.1 | 1.2 | 75.5 |
| 14 | *Pleurobrachia pileus* | 3.4 | 3.8 | 0.9 | 80.7 |
| 21 | *Schistomysis spiritus* | 3.3 | 3.6 | 0.9 | 85.7 |
| 15 | *Pleurobrachia pileus* jv | 3.3 | 4.7 | 0.7 | 90.7 |
| . | .................... | ... | ... | ... | .... |

**RECOMMENDATION**

USE SIMILARITY % BREAKDOWN (program SIMPER) or a SHADE MATRIX to INDICATE (not test) which species are mainly responsible for an observed clustering of the samples into groups (or for a confirmed difference between previously-defined groups).

# LECTURE 8

## UNIVARIATE AND DISTRIBUTIONAL METHODS: DIVERSITY MEASURES, DOMINANCE CURVES AND OTHER GRAPHICAL ANALYSES

### INDICES OF DIVERSITY AND EVENNESS

A single index of species (or higher taxon) diversity is commonly employed in community studies, and is amenable to simple statistical analysis. A bewildering variety of diversity indices has been used, and it is not appropriate here to discuss their relative merits and disadvantages. A good account can be found in Heip et al. (1988).

Two different aspects contribute to the concept of community diversity:

SPECIES RICHNESS - A measure related to the total number of species present.

EQUITABILITY - Expresses how evenly the individuals are distributed among different species.

The most commonly used diversity measure is the SHANNON-WIENER INDEX:

$$H' = - \Sigma_i \; p_i(\log p_i)$$

This incorporates both the species richness and equitability components. Note that logarithms to the base 2 are often used in the calculation, giving the diversity units as 'bits per individual'. $\text{Log}_e$ is also frequently used, so care should be exercised when comparing published indices.

SPECIES RICHNESS is often given simply as the total number of species (S), which is obviously very dependent on sample size, but more commonly as MARGALEF'S INDEX d, which also incorporates the total number of individuals (N):

$$d = (S-1) / \log N$$

EQUITABILITY is most commonly expressed as PIELOU'S EVENNESS INDEX:

$$J' = H'(\text{observed}) / H'_{\text{max}}$$

where $H'_{\text{max}}$ is the maximum possible diversity (log S).

### UNITS OF MEASUREMENT

Numbers of individuals belonging to each species are the most common units. For internal comparative purposes other units could be used, e.g. biomass or total cover of each species along a transect (e.g. for hard-bottom epifauna).

REPRESENTING COMMUNITIES

Data usually presented as plots of means and confidence intervals for each site or time.

Example 1: Benthos from Hamilton Harbour, Bermuda.

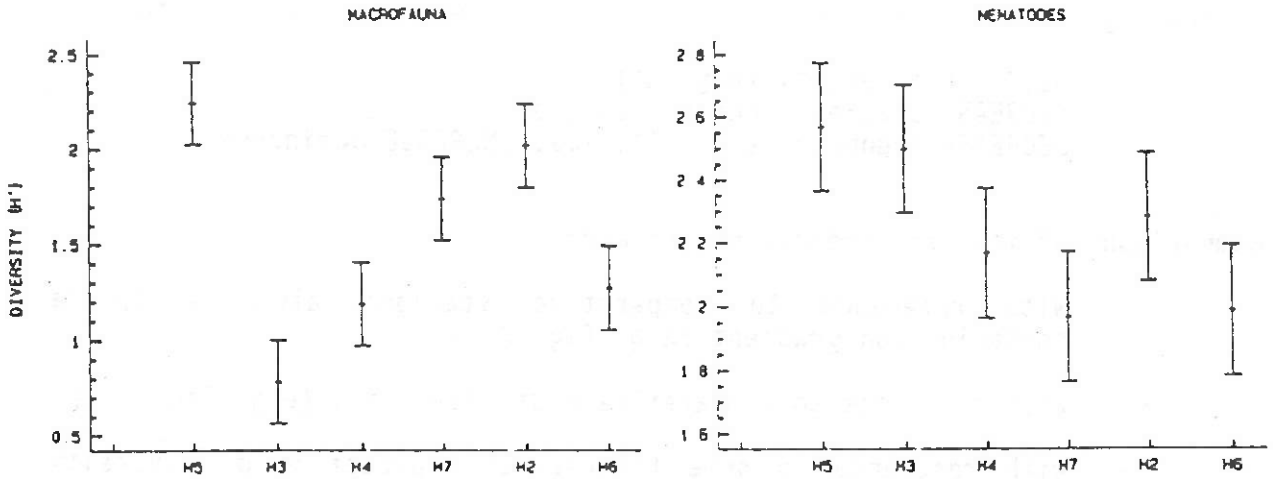


Fig. 8.1 Diversity (H′) and 95% confidence intervals for macrobenthos (left) and meiobenthic nematodes (right) at six stations

Example 2: Reef-corals from South Tikus Island, Indonesia.

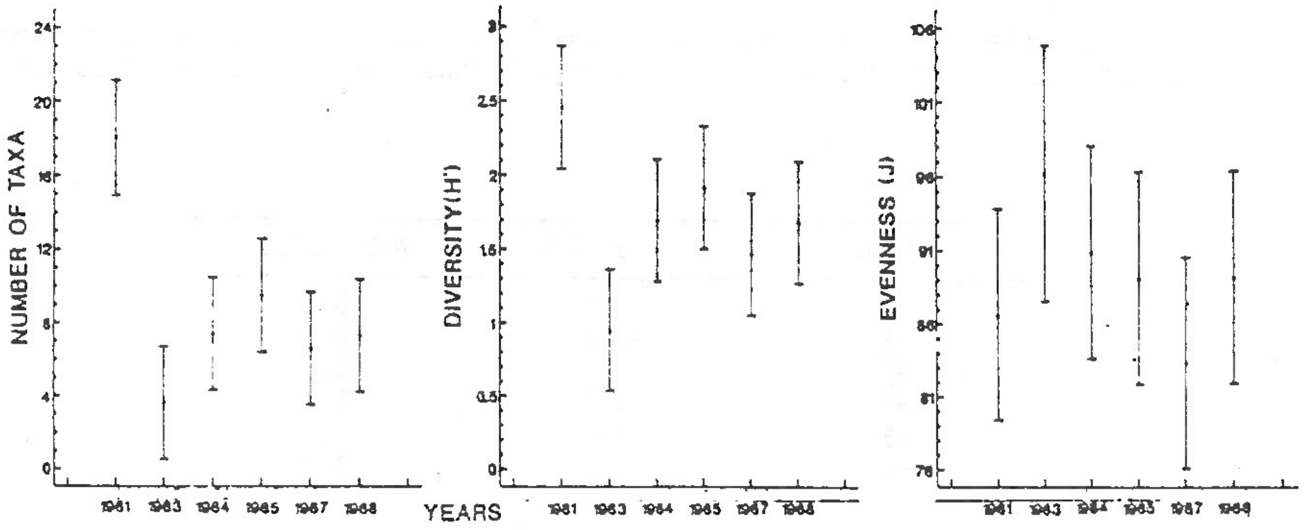


Fig. 8.2 Total number of species (S), Diversity (H′) and Evenness (J′) based on coral species cover data along transects, spanning the 1982-3 El Niño. Note dramatic decline and partial recovery of S and H′, but no obvious changes in J′

## DISCRIMINATING SITES OR TIMES

The significance of differences in diversity indices between sampling sites or times can be tested by one-way analysis of variance (ANOVA).

## DETERMINING STRESS LEVELS

Increasing levels of environmental stress are generally considered to:

- DECREASE diversity (e.g. H′)
- DECREASE species richness (e.g. d)
- DECREASE evenness (e.g. J′), i.e. INCREASE dominance

Comparisons of measured indices can be made:

- with reference to comparative stations along a spatial contamination gradient (e.g. Fig. 8.1).
- with reference to comparative historical data (e.g. Fig. 8.2).
- with reference to some theoretical expectation of diversity, given the number of individuals and species present. Comparisons of observed diversity have been compared with predictions from CASWELL'S NEUTRAL MODEL (Caswell, 1976), which assumes certain community assembly rules and no interactions between species. A value of zero for the V statistic indicates neutrality, positive values indicate greater diversity than predicted and negative values lower diversity. Values >+2 or <-2 indicate significant departures from neutrality. The computer program of Goldman & Lambshead (1989) is useful.

Example: V statistics for summed replicates of macrobenthos and meiobenthic nematode samples at six stations in Hamilton Harbour, Bermuda (cf. Fig. 8.1)

| STATION | MACROBENTHOS | NEMATODES |
|---|---|---|
| H6 | -1.3 | -0.4 |
| H2 | +0.5 | -0.1 |
| H7 | -0.2 | -0.4 |
| H4 | -4.5 | -0.5 |
| H3 | -5.4 | +0.4 |
| H5 | -1.9 | 0.0 |

Note diversity of macrobenthos at H4 and H3 is significantly below neutral model predictions, but nematodes are close to neutrality at all stations.

## GRAPHICAL DISTRIBUTION PLOTS

The purpose of graphical/distributional representations is to extract information on patterns of relative species abundances without reducing that information to a single summary statistic, such as a diversity index. This class of techniques can be thought of as intermediate between *univariate* summaries and full *multivariate* analyses. Unlike multivariate methods, these distributions may extract universal features of community structure which are not a function of the specific taxa present, and may therefore be related to levels of biological stress.

### RAREFACTION CURVES

Rarefaction curves (Sanders, 1968) were among the earliest to be used in marine studies. They are plots of the number of individuals on the x-axis against the number of species on the y-axis. The more diverse the community is, the steeper and more elevated is the rarefaction curve.

Example: Polychaete/bivalve fraction of macrobenthos.

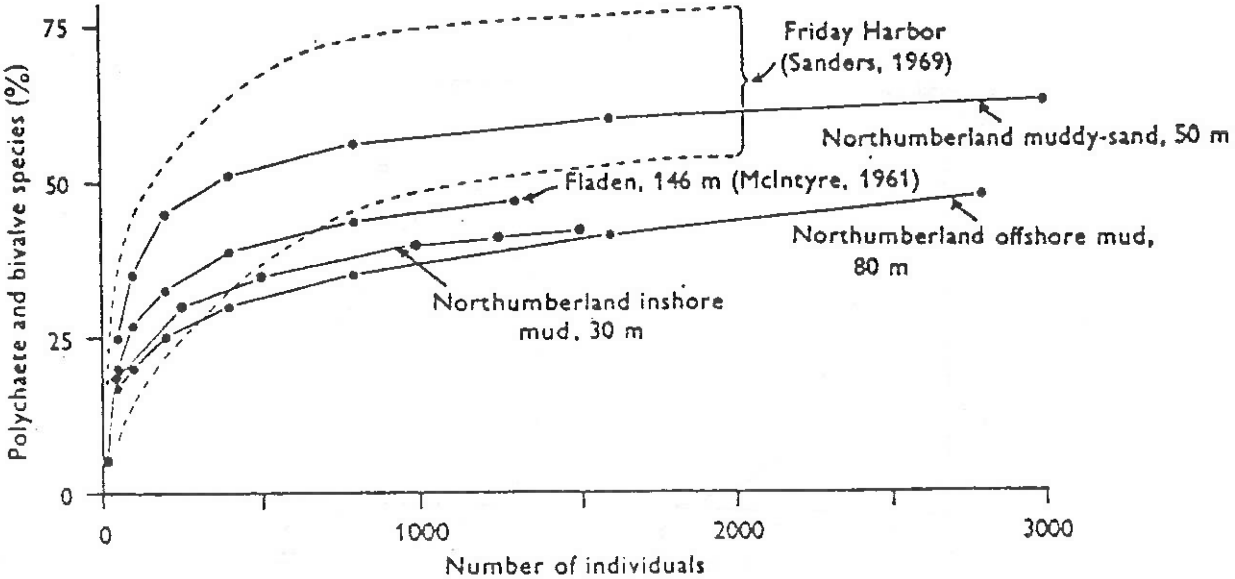


Fig. 8.3 Rarefaction curves comparing North Sea and Friday Harbor stations (from Buchanan & Warwick, 1974)

## RANKED SPECIES ABUNDANCE (DOMINANCE) CURVES

These are based on the ranking of species (or higher taxa) in decreasing order of their importance in terms of abundance or biomass. The ranked abundances, expressed as a percentage of the total abundance of all species, are plotted against the relevant species rank. Log transformations of one or both axes have frequently been used to emphasise or downweight different sections of the curves.

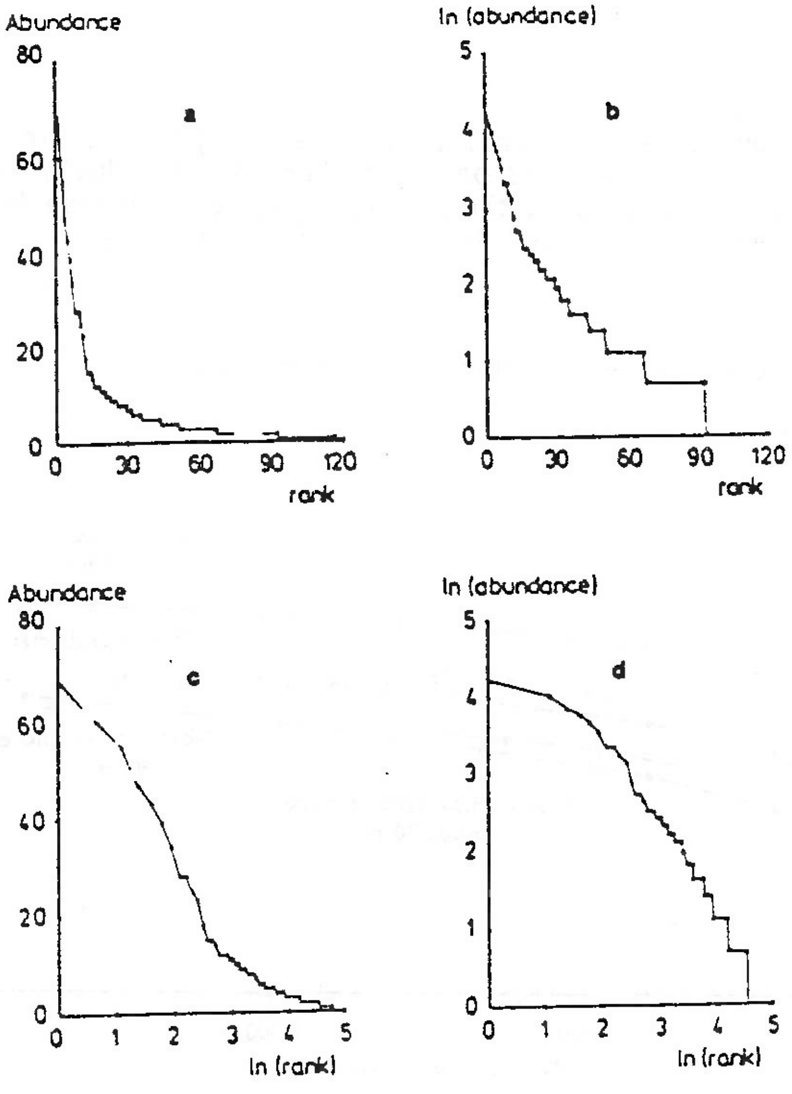


Fig. 8.4 The same (hypothetical) species abundance data plotted as ranked species abundance curves with none, one or both axes on a log scale (from Heip et al., 1988)

## k-DOMINANCE AND LORENZ CURVES

As an alternative to the simple dominance curves above, cumulative ranked abundances may be plotted against species rank, or log species rank, to produce k-DOMINANCE CURVES (Lambshead et al., 1983). This has a smoothing effect on the curves. Ordering of curves on a plot will obviously be the reverse of rarefaction curves, with the most elevated curve having the lowest diversity. To compare dominance separately from the number of species, the x-axis (species rank) can be rescaled from 0-100 (relative species rank), to produce LORENZ CURVES.

Example: Nematodes from Loch Ewe, Scotland.

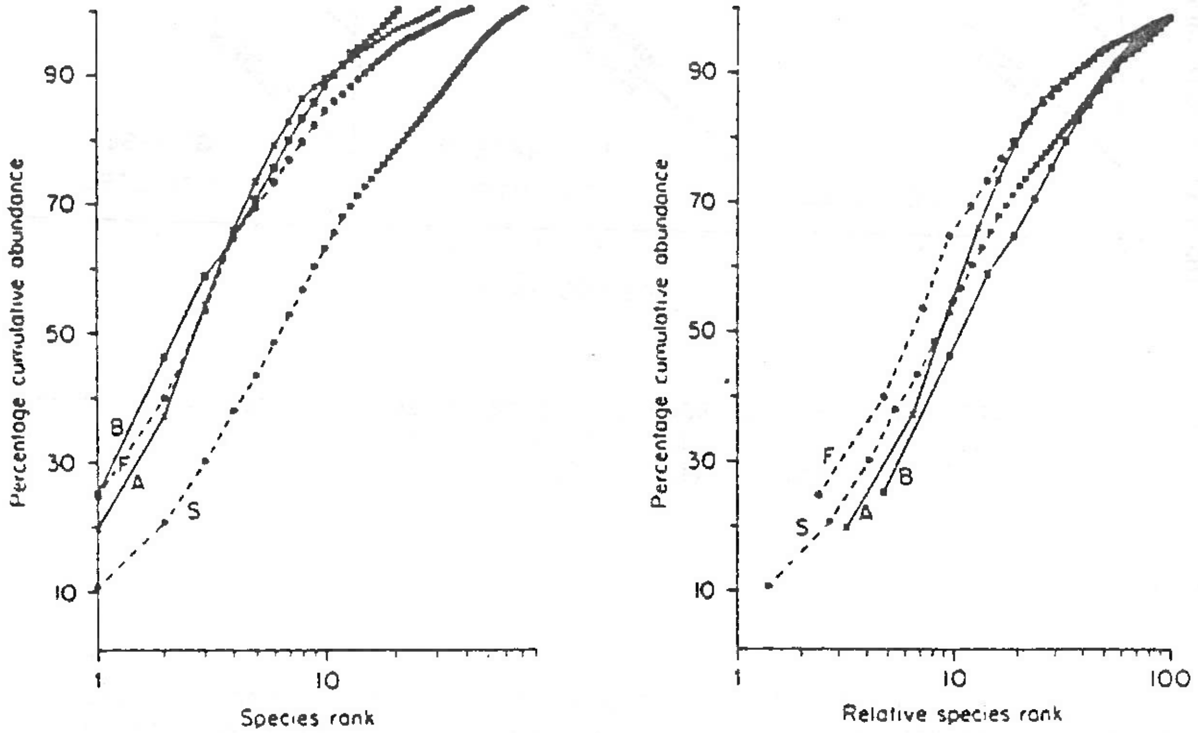


Fig. 8.5 k-dominance curves (left) and Lorenz curves (right) for 20 cm deep cores taken from experimental sand columns 20 days (A) and 77 days (B) after initial setup, and from intertidal (F) and subtidal (S) sand from the study site (from Lambshead et al., 1983)

## ABUNDANCE / BIOMASS COMPARISON (ABC) PLOTS

The advantage of distribution plots such as k-dominance curves is that the distribution of species abundances among individuals and the distribution of species biomasses among individuals can be compared on the same terms. Since the two have different units of measurement, this is not possible with diversity indices.

This is the basis of the ABUNDANCE / BIOMASS COMPARISON (ABC) method of determining levels of disturbance (pollution-induced or otherwise) on benthic macrofauna communities. Both empirical evidence and theoretical considerations suggest that the k-dominance curve for biomass will fall above the curve for abundance in undisturbed (or unpolluted) communities, and *vice versa* for disturbed (or polluted) communities.

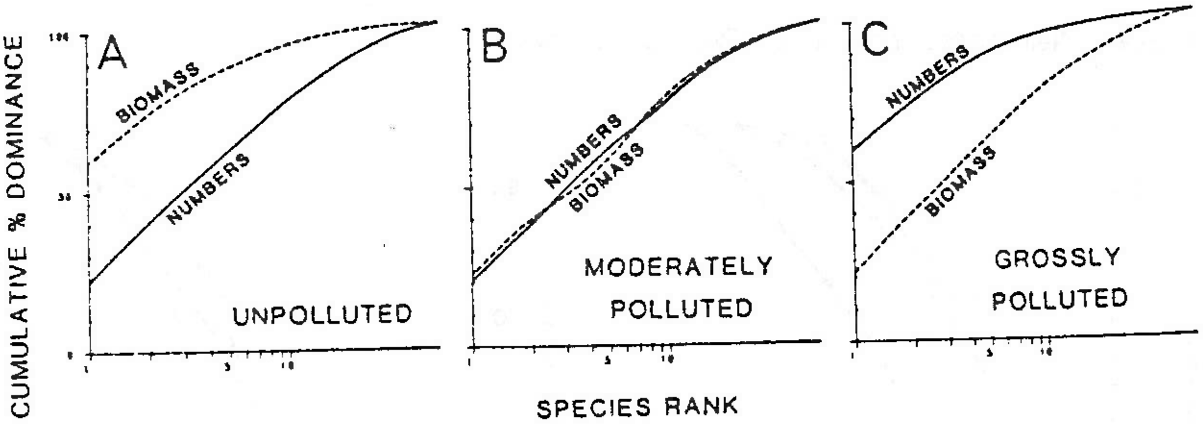


Fig. 8.6 Hypothetical k-dominance curves for species biomass and numbers, showing unpolluted, moderately polluted and grossly polluted conditions (from Warwick, 1986)

<u>Example 1:</u> Time series of macrobenthos in Loch Linnhe, Scotland in response to increasing and decreasing levels of organic enrichment (pulp-mill effluent). See Lecture 1, Figs. 1.3 and 1.4

<u>Example 2:</u> Transect across sewage-sludge dumping ground at Garroch Head, Firth of Clyde, Scotland.

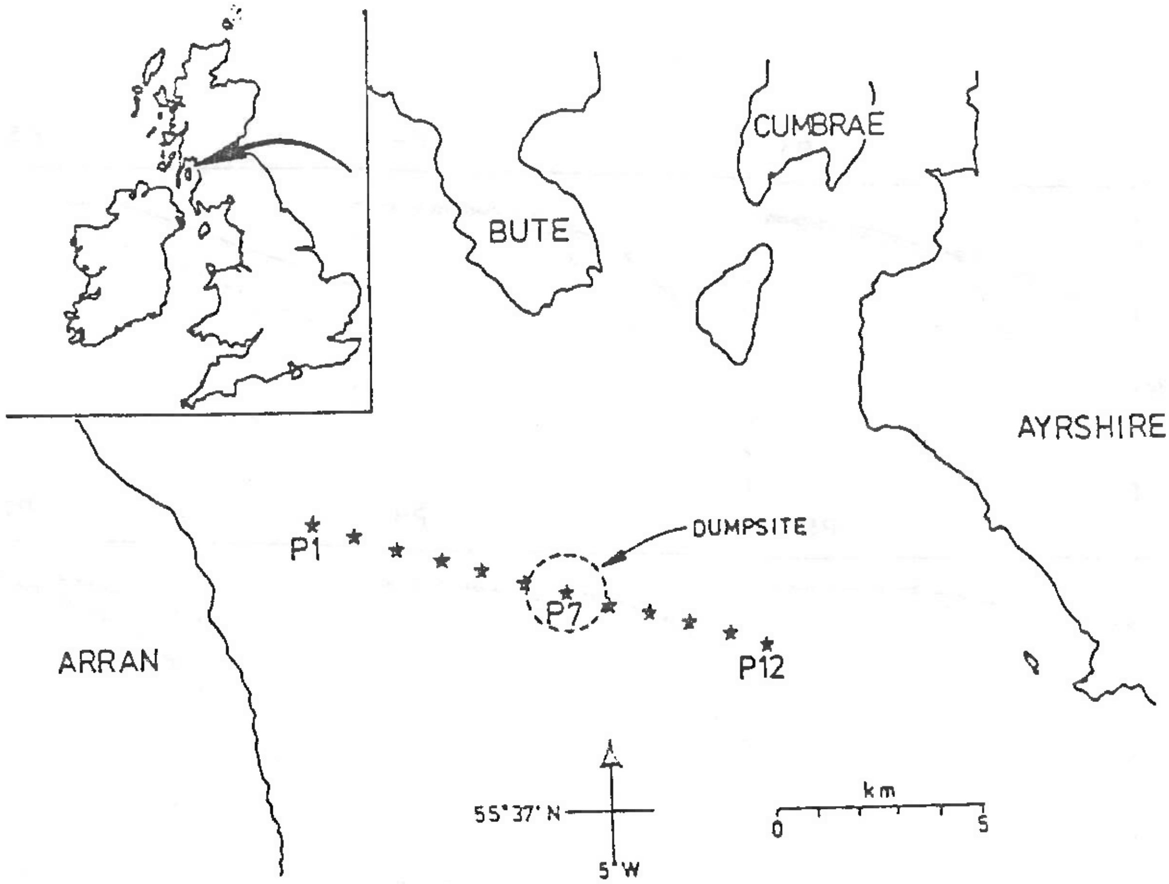


Fig. 8.7 Map showing location of dumping-ground. Centre of dump-site denoted by dashed circle: positions of sampling stations (P1 - P12) identified by asterisks

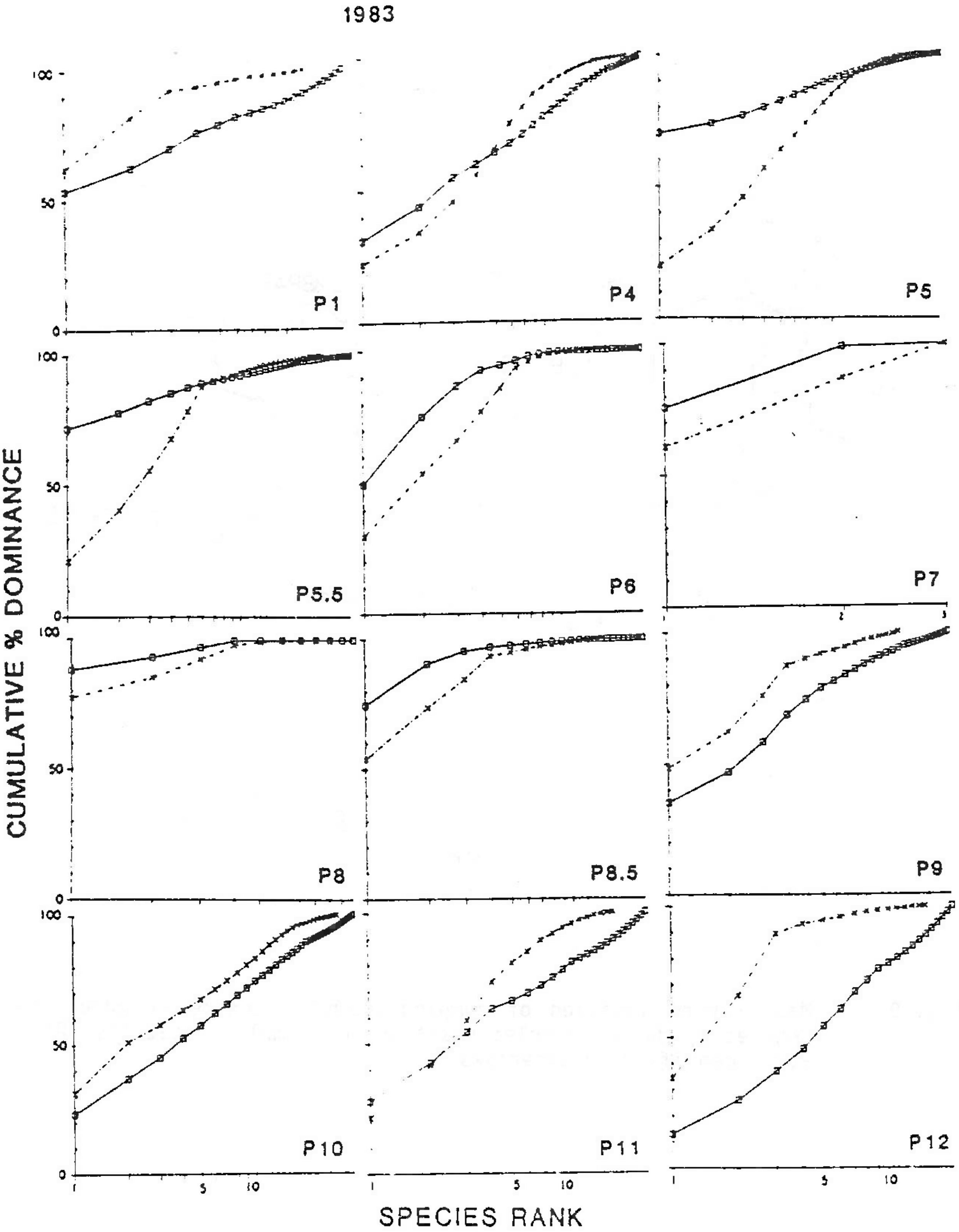


Fig. 8.8 ABC plots for macrobenthos on Garroch Head transect in 1983. Abundance = squares, biomass = crosses (From Warwick et al., 1987)

## TRANSFORMATIONS OF k-DOMINANCE CURVES

PROBLEM: It is difficult to distinguish differences between k-dominance curves when cumulative frequencies are near 100% (sometimes after the first 2 or 3 spp.).

SOLUTION: Transform y-axis so that cumulative values are close to linearity. Clarke (1990) suggests the modified *logistic* transformation:

$$y_i' = \log[(1 + y_i)/(101 - y_i)]$$

Example: Macrobenthos from Frierfjord / Langesundfjord, Norway (IOC/GEEP Oslo Workshop).

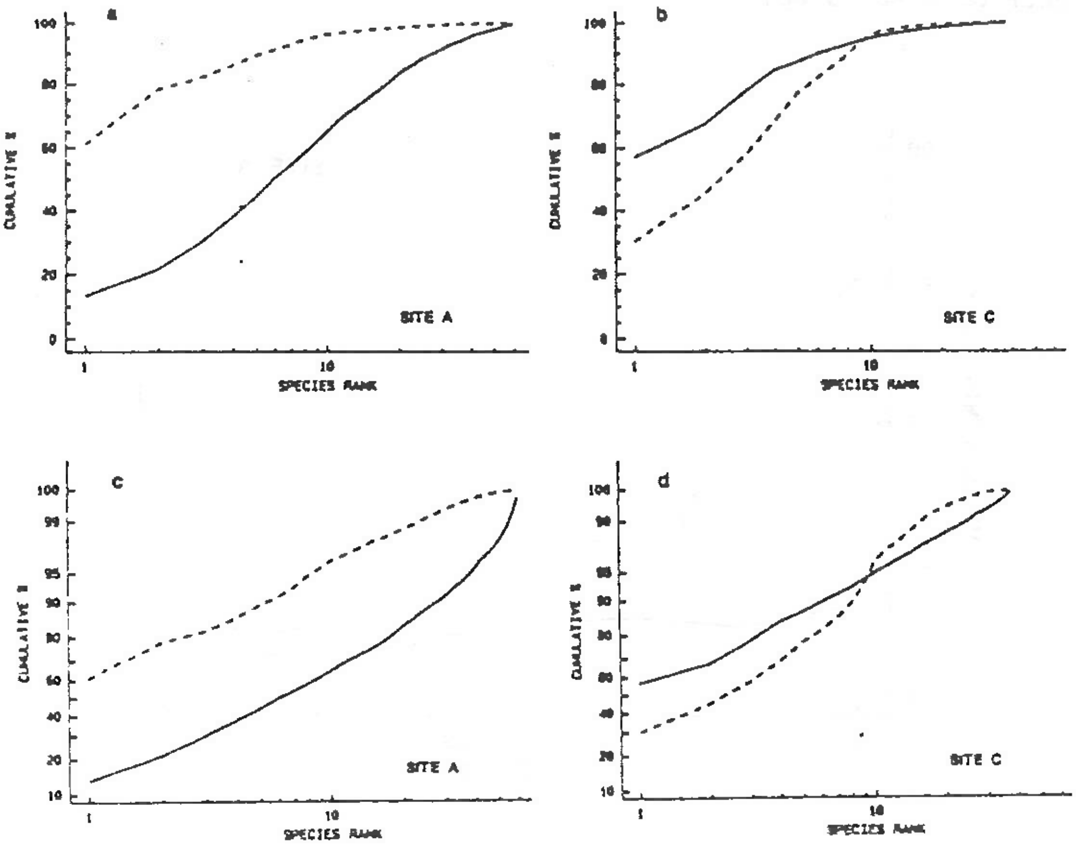


Fig. 8.9 a), b) Standard ABC plots for sites A (reference) and C (potentially impacted). c), d) ABC plots for sites A and C with the y-axis subjected to modified logistic transformation. Abundance = continuous line, biomass = dashed line

PARTIAL DOMINANCE CURVES

PROBLEM: Visual information presented by k-dominance (and ABC) curves is over dependent on single most dominant species. Unpredictable presence of large numbers of small biomass species, or heavy spatfall of young of one species, may give false impression of disturbance.

SOLUTION: With genuine disturbance, patterns of ABC curves should be unaffected by successive removal of most dominant species in terms of abundance or biomass. PARTIAL DOMINANCE CURVES (Clarke, 1990) compute the dominance of the second most dominant species over the remainder, the same with the third most dominant etc.

Example 1: Macrobenthos from Frierfjord/Langesundfjord, Norway (IOC/GEEP Oslo Workshop).

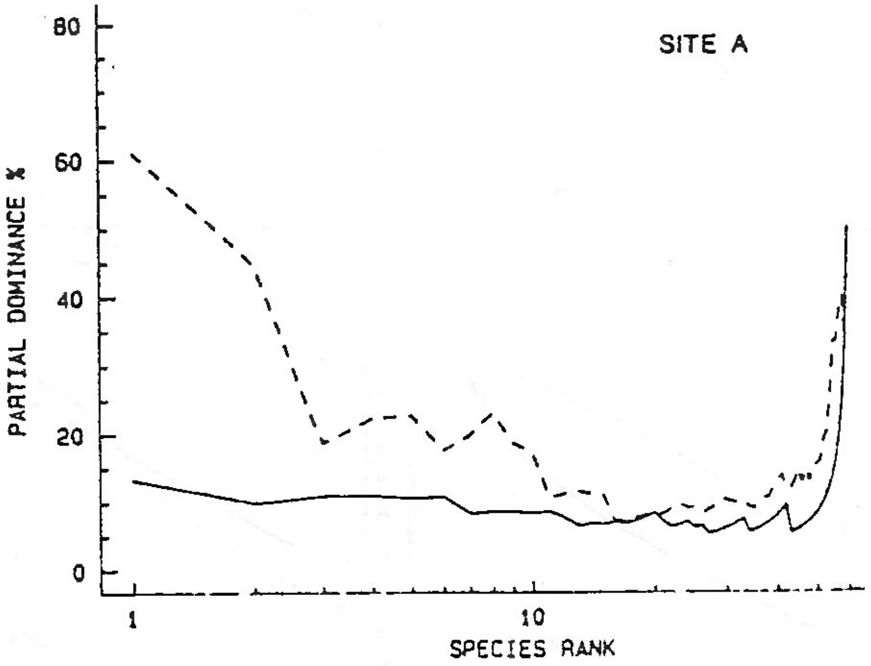


Fig. 8.10 Partial dominance curves (abundance/biomass comparison) for reference station A (c.f. Figs 8.9a and c for corresponding standard and transformed ABC plots). This illustrates the typically undisturbed condition)

Example 2: Loch Linnhe macrobenthos, 1966-68, 1970-72.

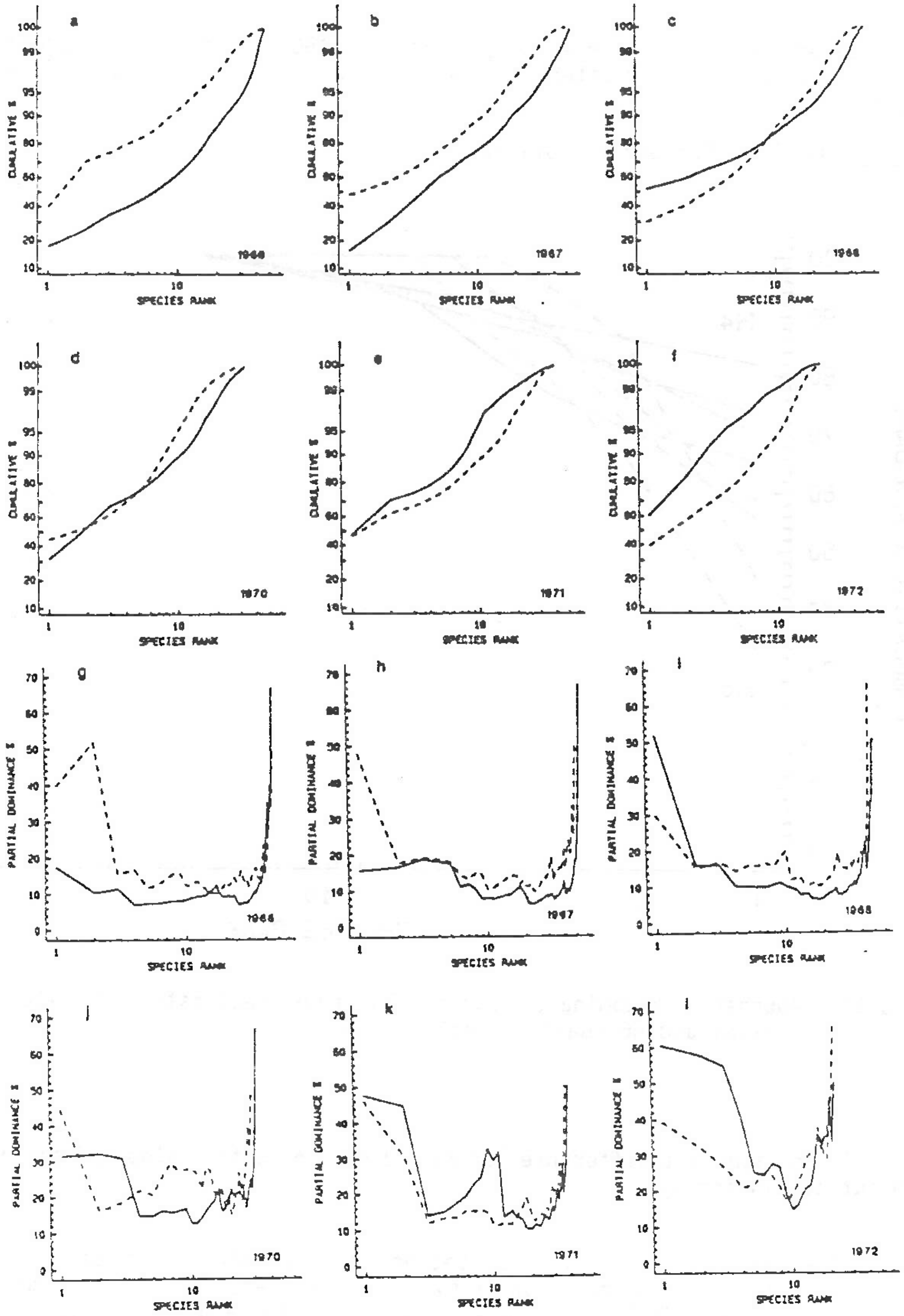


Fig. 8.11 a)-f) ABC curves (logistic transform). g)-l) Partial dominance curves for abundance (solid line) and biomass (dashed line) for the same years

## SIGNIFICANCE TESTING FOR GRAPHICAL METHODS

Given replicate curves (k-dominance, ABC, 'individuals amongst species' etc.) at 2 or more sites (or times etc.), need a TEST FOR SIGNIFICANT DIFFERENCE.

Example: Hamilton Harbour macrofauna.

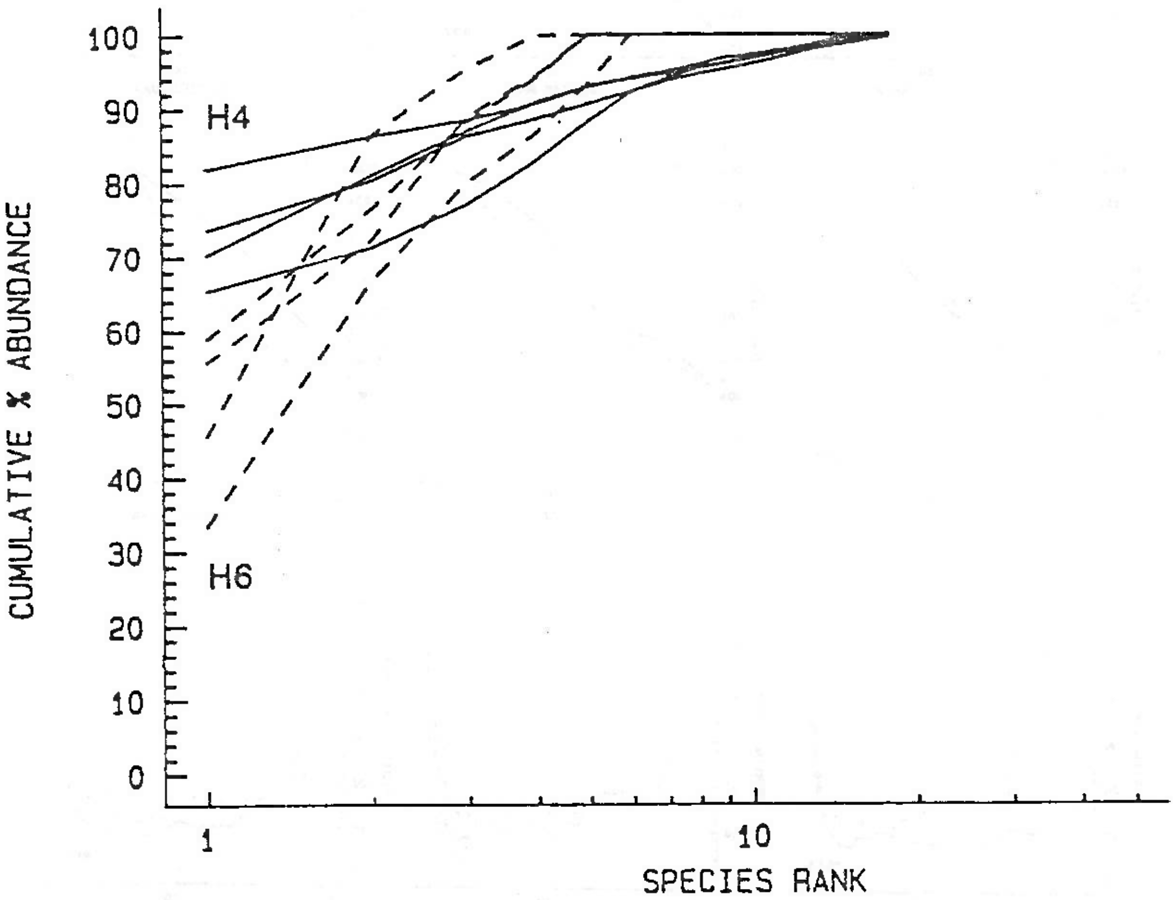


Fig. 8.12 Abundance k-dominance curves for four replicates at site H4 (solid) and H6 (dashed line)

Is the apparent difference for H4 and H6, in initial slope of curves, borne out statistically?

Also, testing for difference between sets of ABC CURVES at two (or more) sites reduces to a comparison of two (or more) sets of replicate curves by computing the DIFFERENCE CURVE B-A for each sample, e.g. Fig. 8.13.

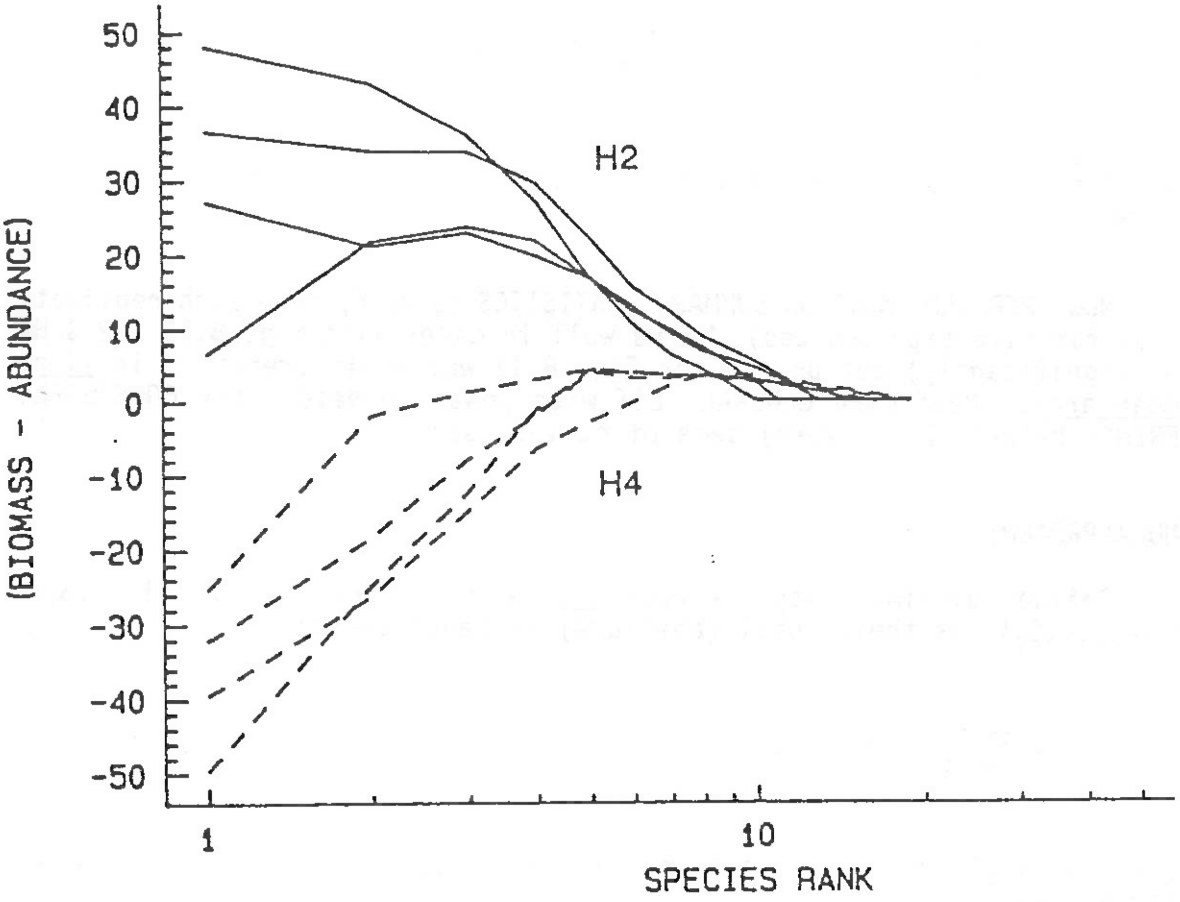


Fig. 8.13 Difference (B-A) between k-dominance curves for biomass and abundance for four replicate samples at H2 (solid) and H4 (dashed line)

FIRST APPROACH:

Reduce each replicate curve to a SINGLE SUMMARY STATISTIC. E.g. if $\{A_1\}$ and $\{B_1\}$ are the cumulative abundance and biomass values from an ABC plot ($i=1,..,$ S species), define:

$$W = \sum_{i=1}^{S} (B_i - A_i) / [50(S-1)]$$

W takes values in (-1,1), with W→1 for totally even abundances across species but biomass dominated by a single species, and W→-1 for the converse case.

Similarly, for k-dominance curves of cumulative $\{A_i\}$:

$$K_A = [(\textstyle\sum_{i=1}^{S} A_i) - 50(S+1)]/[50(S-1)]$$

where extremes are K→0 (evenness) and K→1 (dominance). $K_b$ defined similarly for biomass.

Now, PERFORM ANOVA on SUMMARY STATISTICS (W or K) from each replicate (e.g. as for diversity indices). Works well in cases like Fig. 8.13 (H2 & H4 differ significantly) but poorly for Fig. 8.12 where difference is in slope not mean area. Need more GENERAL TEST with power to detect any CONSISTENT DIFFERENCE between 2 (or more) sets of curves, so

SECOND APPROACH:

Define 'dissimilarity' between any pair of curves $\{A_{i1};\ i=1,..,S_1\}$, $\{A_{i2};\ i=1,..,S_2\}$, as their total (absolute) distance apart:

$$d = \textstyle\sum_{i=1}^{Smax} |A_{i1}-A_{i2}|$$

where $S_{max} = \max(S_1, S_2)$. Or better reflection of visual difference in two k-dominance curves is:

$$d' = \textstyle\sum_{i=1}^{Smax} |A_{i1}-A_{i2}| \log(1+i^{-1})$$

Compute d (or d') for every pair of replicate curves, to give lower triangular dissimilarity matrix, and CALCULATE ANOSIM STATISTIC R, eqt. (6.1).

PERMUTATION/RANDOMISATION TEST of difference between sites/times etc. then carried out exactly as in lecture 6. (ANOSIM on Fig. 8.12 distinguishes H4 and H6, whereas ANOVA on $K_A$ does not).

Details in Clarke (1990). Note that principle EXTENDS TO OTHER GRAPHICAL METHODS, e.g. partial dominance, 'individuals amongst species' curves etc.

## LECTURE 9

## TRANSFORMATIONS

There are two distinct roles for transformations in community analysis:

a) to validate assumptions for parametric analyses - applies to UNIVARIATE tests

b) to weight the contributions of common and rare species in a MULTIVARIATE representation.

### UNIVARIATE

Example: Frierfjord macrofauna. Indicator species.

Table 9.1

*Thyasira sp.* numbers in 4 replicate grabs at 6 sites.

| Site: | A | B | C | D | E | G |
|---|---|---|---|---|---|---|
| Replicate | | | | | | |
| 1 | 1 | 7 | 0 | 1 | 62 | 66 |
| 2 | 4 | 0 | 0 | 8 | 102 | 68 |
| 3 | 3 | 3 | 0 | 5 | 93 | 52 |
| 4 | 11 | 2 | 3 | 13 | 69 | 36 |
| Mean | 4.8 | 3.0 | 0.8 | 6.8 | 81.8 | 55.5 |
| Stand.dev. | 4.3 | 2.9 | 1.5 | 5.1 | 18.7 | 14.8 |

NOTE:

1) The replicates are not symmetrically distributed (they tend to be right-skewed), so normality assumptions are dubious.

2) More importantly (for test validity), the variance increases strongly with the mean - this invalidates "constant variance" assumptions of ANOVA.

Both problems can be tackled by:

## POWER TRANSFORMATION

Individual replicates y are transformed to y*, given by:

$$y^* = (y^\lambda - 1)/\lambda \qquad (9.1)$$

where, in order of INCREASING SEVERITY,

| | |
|---|---|
| $\lambda = 1$ | - no transform |
| $\lambda = 0.5$ | - square root ($\sqrt{}$) |
| $\lambda = 0.25$ | - 4th root ($\sqrt{}\sqrt{}$) |
| $\lambda \to 0$ | - log transform ($y^* = \log_e y$) |

Possible to determine best λ, anywhere in (0,1), for each separate data set (Box and Cox, 1964), but unnecessarily precise - better just to choose between above 4 cases, using:

TAYLOR'S POWER LAW:

If:

$$\mathrm{var}(y) \propto (\text{mean } y)^v \qquad (9.2)$$

then:

$$\mathrm{var}(y^\lambda) \propto (\text{mean})^{2(\lambda-1)+v} \text{ (approx.)} \qquad (9.3)$$

Choose $\lambda = 1-(v/2)$ to get var(y) = constant.

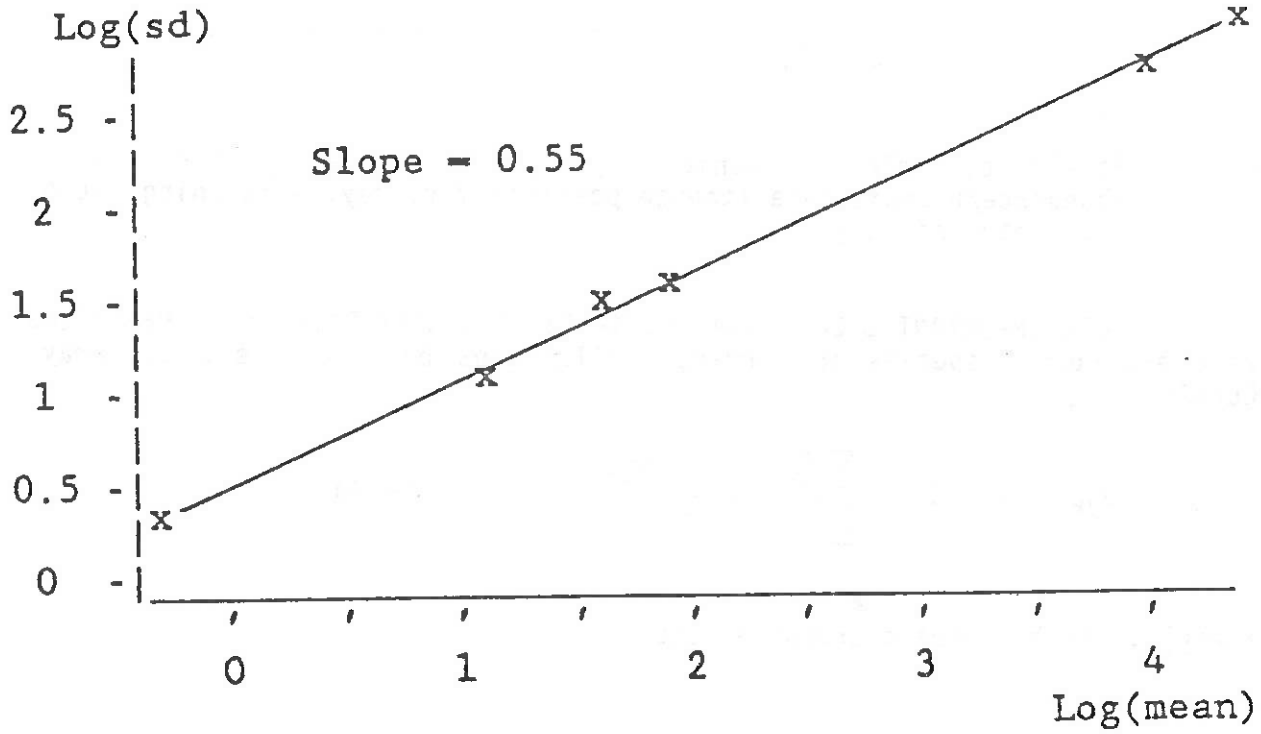
Find v by regressing log (stand.dev.) on log (mean), because:

$$\log(\mathrm{sd}(y)) = (v/2)\log(\text{mean } y) + \text{constant} \qquad (9.4)$$

So λ = 1 - (slope of regression), thus if:

$$\begin{aligned} \text{slope} &= 0 && \text{- no transform} \\ &= 0.5 && \text{- use } \sqrt{} \\ &= 0.75 && \text{- use } \sqrt{}\sqrt{} \\ &= 1 && \text{- use } \log_e \end{aligned} \qquad (9.5)$$

Example: *Thyasira* numbers at 6 sites



Plot indicates √ appropriate. After transform:

| Site | A | B | C | D | E | G |
|---|---|---|---|---|---|---|
| Mean(y*) | 2.01 | 1.45 | 0.43 | 2.42 | 9.00 | 7.40 |
| Sd (y*) | 0.97 | 1.10 | 0.87 | 1.10 | 1.04 | 1.04 |

VARIANCE STABILISED so ANOVA and follow-up tests VALID (show E,G different from the rest, clearly). Means and confidence intervals should be back-transformed to original scales (intervals not symmetric but then data was not symmetric).

CAUTION: Beware of doing multiple ANOVAs on a range of indicator species (each runs a 5% risk of error and this compounds). Alright if performed (at higher significance) on a few species selected a priori.

AVOID "SNOOPING" in a large data array for likely species to do an ANOVA on; certain to find some which are significant, even in a random array!

**MULTIVARIATE**

TRANSFORMS can be used for the same reason as in univariate analyses - to induce (multivariate) normality, eg. for MANOVA tests (lecture 6), but:

a) Insufficient to demonstrate univariate normality and constant variance (for each variable) to prove multivariate normality and constant covariance.

b) Rarely possible to achieve (marginal) normality for species abundance/biomass data (though possible for, say, a matching set of diversity indices).

MORE IMPORTANT USE OF TRANSFORMS IN COMMUNITY DATA is in WEIGHTING rare and common species in forming similarities between sites, eg. Bray-Curtis:

$$S_{jk} = 100 \{1 - \frac{\sum_{i=1}^{p} |y_{ij} - y_{ik}|}{\sum_{i=1}^{p} (y_{ij} + y_{ik})}\} \qquad (9.6)$$

Example: Loch Linnhe macrofauna, subset

| Sample: | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Species | | | | |
| *Echinoca.* | 9 | 0 | 0 | 0 |
| *Myrioche.* | 19 | 0 | 0 | 3 |
| *Labidopl.* | 9 | 37 | 0 | 10 |
| *Amaeana* | 0 | 12 | 144 | 9 |
| *Capitella* | 0 | 128 | 344 | 2 |
| *Mytilus* | 0 | 0 | 0 | 0 |

UNTRANSFORMED

| Sample | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | - | | | |
| 2 | 8 | - | | |
| 3 | 0 | 42 | - | |
| 4 | 39 | 21 | 4 | - |

| Sample: | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Species | | | | |
| *Echinoca.* | 1.7 | 0 | 0 | 0 |
| *Myrioche.* | 2.1 | 0 | 0 | 1.3 |
| *Labidopl.* | 1.7 | 2.5 | 0 | 1.8 |
| *Amaeana* | 0 | 1.9 | 3.5 | 1.7 |
| *Capitella* | 0 | 3.4 | 4.3 | 1.2 |
| *Mytilus* | 0 | 0 | 0 | 0 |

√√TRANSFORMED

| Sample | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | - | | | |
| 2 | 26 | - | | |
| 3 | 0 | 68 | - | |
| 4 | 52 | 68 | 42 | - |

Untransformed similarities are lower (unimportant in itself since MDS is only a function of ranks) but RANK SIMILARITIES ARE TOTALLY CHANGED by transform.

Untransformed similarities are DOMINATED BY THE COMMONEST SPECIES, eg. comparing samples 2 and 4 and omitting each species in turn:

| Species omitted: | None | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Bray-Curtis (S): | 21 | 21 | 21 | 14 | 13 | 54 | 21 |

By contrast, under a √√ transform, ALL (present) SPECIES MAKE SOME CONTRIBUTION to the similarity:

| Species omitted: | None | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Bray-Curtis (S): | 68 | 68 | 75 | 61 | 59 | 76 | 68 |

TRANSFORMATION SEQUENCE:

None ——> √ ——> √√ ——> log ——> Presence/absence

puts PROGRESSIVELY LESS WEIGHT on common species and increasingly takes account of rarer ones.

Logical end-point is REDUCTION of the data array to one of PRESENCE OR ABSENCE OF SPECIES (this is a transformation to the numbers 0 or 1), where all species contribute equally.

Example: Loch Linnhe macrofauna, subset.

| Sample: | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Species | | | | |
| *Echino.* | 1 | 0 | 0 | 0 |
| *Myrioc.* | 1 | 0 | 0 | 1 |
| *Labido.* | 1 | 1 | 0 | 1 |
| *Amaeana* | 0 | 1 | 1 | 1 |
| *Capite* | 0 | 1 | 1 | 1 |
| *Mytilus* | 0 | 0 | 0 | 0 |

PRESENCE/ABSENCE

| Sample | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | - | | | |
| 2 | 33 | - | | |
| 3 | 0 | 80 | - | |
| 4 | 57 | 86 | 67 | - |

NOTE: 1) NEED TO USE log(1+y) not log y which DISTORTS TRANSFORM SEQUENCE. log (1+y) intermediate between √√ and presence/absence for moderate or large counts but less severe than √√ for small counts.

2) √√y preferred to log(1+y) because Bray-Curtis is INVARIANT TO A SCALE CHANGE (eg. for biomass) if √√ is used. (Little difference in practice though).

3) As severity of transform increases, more species contribute, so sample relationships are expressed in higher-dimensional space, and ordination in 2-d is harder (eg. Fig. 9.1). So, WRONG to assume that TRANSFORMS GIVING LOWER STRESS ARE BETTER; the converse is true if added species are important.

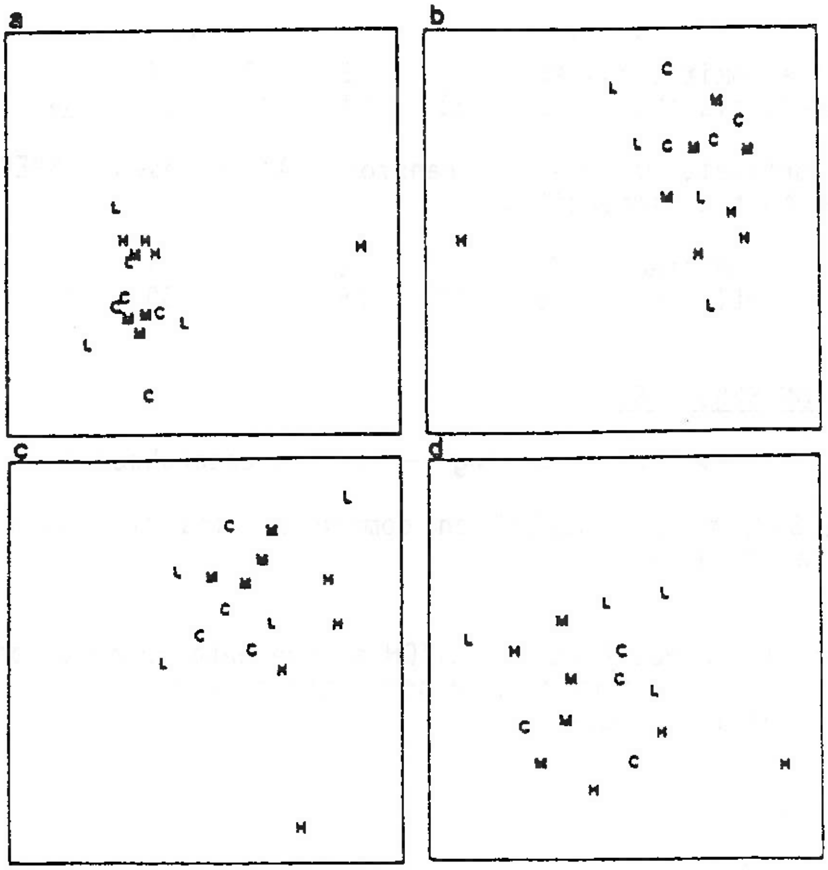


Fig. 9.1 GEEP mesocosm nematodes (Warwick et al., 1988). MDS of 4 boxes from 4 treatments (C,L,M,H). Bray-Curtis similarities from transformed counts: a) no transform, b) √, c) √√, d) presence/absence. Stress: a) 0.08, b) 0.14, c) 0.19, d) 0.19

4) SAME TRANSFORM SEQUENCE APPLIES TO PCA (and other ordinations) with much the same consequences.

5) Log (or √√) transforms effectively REDUCE DATA TO A 6 POINT SCALE, i.e. 0 = absent, 1 = one individual, 2 = handful, 3 = sizeable, 4 = abundant, 5 = very abundant; replacing data by this scale will make no real difference to the multivariate displays. This may appear crude but often genuinely reflects inherent variability, so greater accuracy in counting may be unnecessary.

CONCLUDE:

1) CHOICE OF TRANSFORM often has a bigger effect on conclusions than the CHOICE OF ORDINATION method.

2) "What is the RIGHT TRANSFORM for a multivariate analysis?" is largely a BIOLOGICAL rather than a STATISTICAL question (unlike the use of transforms for validating assumptions); the choice of transform determines how the similarity of two samples is defined.

RECOMMEND:

Use INTERMEDIATE transform (eg. √, √√ or LOG) rather than either of the two EXTREMES:

a) NO TRANSFORM - MDS reflects only 2 or 3 commonest species, so INTERPRETATION is likely to be SHALLOW.

b) PRESENCE/ABSENCE - CHANCE OCCURRENCES of rare species DOMINATE the SAMPLE RELATIONSHIPS in high dimensions and make it difficult to get an interpretable low-dimensional ordination.

# LECTURE 10

## SPECIES REMOVAL AND AGGREGATION

### SPECIES REMOVAL

Two reasons for ELIMINATING SPECIES discussed earlier:

a) For sample PCA (not MDS) ordination, must reduce to (say) <50 species, else problems with eigenvalues.

b) For species ordinations, though MDS and CLUSTER are possible for all species, rarer (chance) species must be excluded for an interpretable outcome.

RECOMMEND RETAINING SPECIES ACCOUNTING FOR >p% of total score (abundance or biomass) in ANY ONE SAMPLE (p chosen to reduce to required number, typically p = 3 or 4). Allows for high diversity/ low abundance samples which could have all species eliminated by simple selection of the top q% most abundant species over all samples.

SPECIES REDUNDANCY: Since sample relationships can often be well summarised in a 2-d ordination (from, say, a 100-d species space), many SPECIES MUST BE INTERCHANGEABLE in the way they characterise the samples. This can be seen by performing MDS on a randomly chosen subset (say 20%) of species:

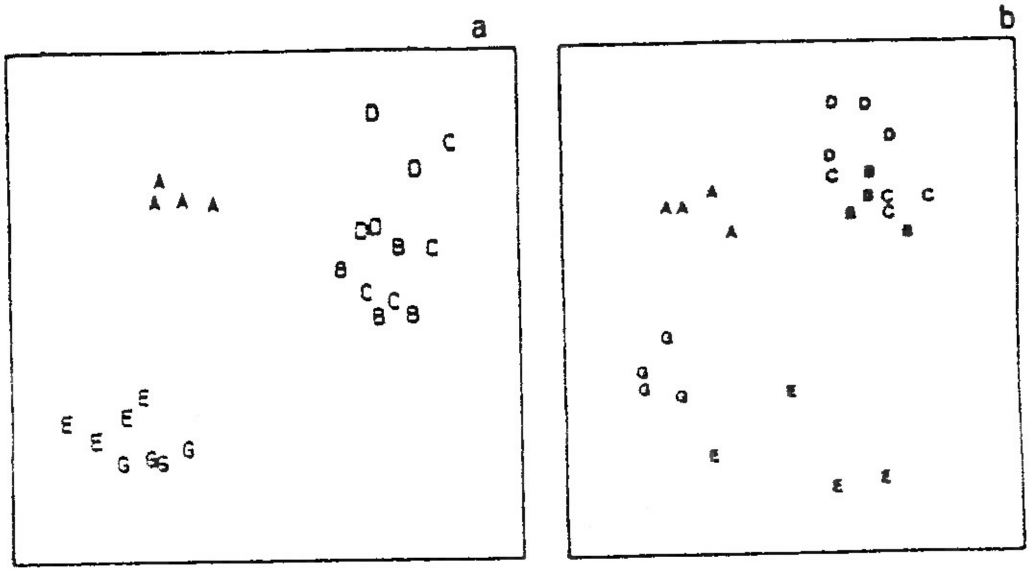


Fig. 10.1 Frierfjord macrofauna counts. Sample MDS (Bray-Curtis, √√) for: a) all 110 species, b) 19 random species. (Stress: a) 0.14, b) 0.13)

Above example of no practical interest, but suggests:

SPECIES AGGREGATION to higher taxonomic levels.

If results from identifications to higher taxonomic levels are comparable to a full species analysis:

a) a great deal of LABOUR CAN BE SAVED;

b) LESS FAUNAL EXPERTISE NEEDED - major factor in parts of the world where fauna is poorly described.

METHODS AMENABLE TO AGGREGATION:

1) MULTIVARIATE:

All ordination/clustering techniques.

Empirical evidence is increasing that identification only to family level makes little difference.

2) DISTRIBUTIONAL:

a) Aggregation for ABC curves is possible; family level analyses are often identical to species level analyses (see Figs. 10.6 and 10.7).

b) Untried for other methods (eg. Individuals amongst species curves).

3) UNIVARIATE:

a) Concept of "indicator groups" is well-established (eg. nematode/copepod ratios).

b) Can define diversity indices at hierarchical taxonomic levels (though not commonly used in practice).

Warwick (1988) hypothesises further motivation: that pollution may change community composition at higher taxonomic levels (eg. phyla) whereas natural variables (grain size, water depth etc.) modify it more by species replacement (within phyla). Thus, distribution of higher taxa may even relate more closely to the contamination gradient than species data, the latter being more complicated by effects of confounding natural variables.

## MULTIVARIATE EXAMPLES

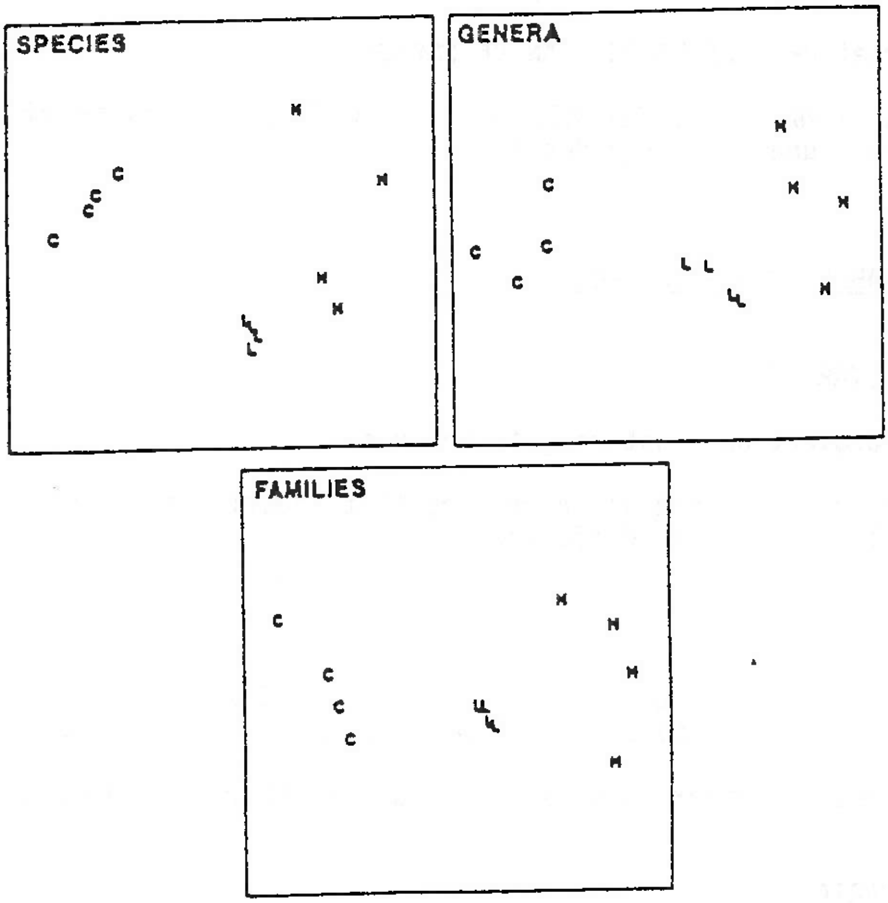


Fig. 10.2 Mesocosm copepod counts - 3 levels of nutrient enrichment (Gee et al., 1985). Sample MDS plot (Bray-Curtis, √√ transform); species data aggregated into genera and families (Warwick, 1988)

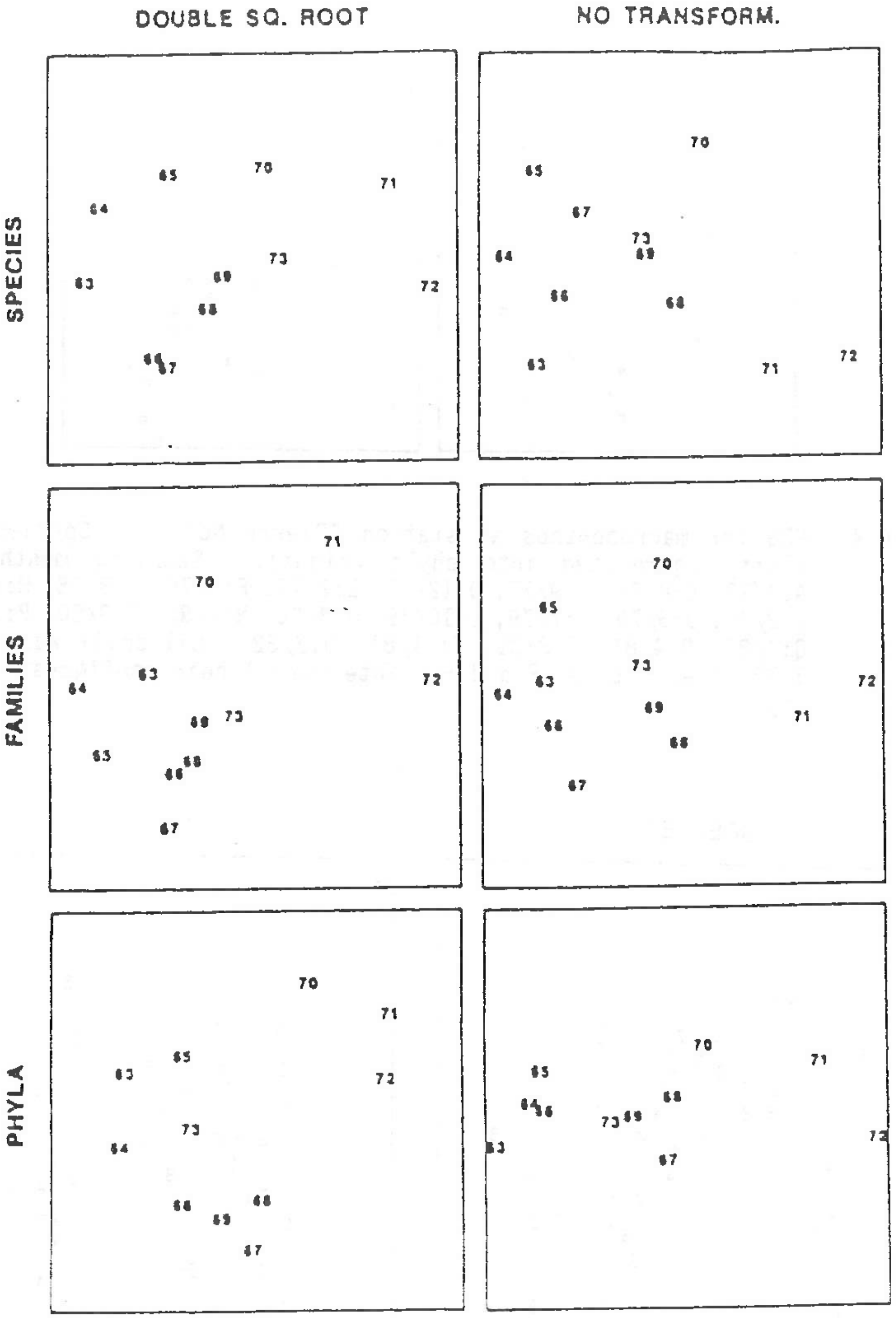


Fig. 10.3 Loch Linnhe macrofauna (Pearson, 1975). MDS (Bray-Curtis) of 11 years samples for √√ transform (left) and no transform (right), based on abundances from 115 species (top), aggregated into 45 families (middle) and 9 phyla (bottom), Warwick (1988). Note more linear configuration for phyla

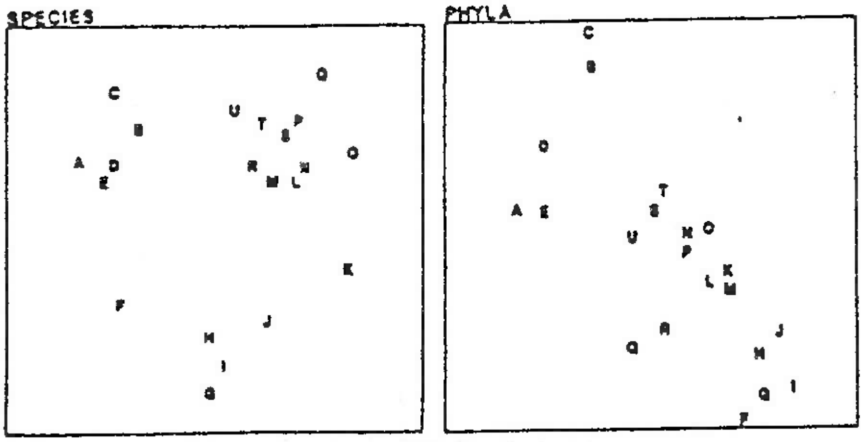


Fig. 10.4 MDS for macrobenthos at station "Pierre Noire". Species data (left) aggregated into phyla (right). Sampling months are A:4/77, B:8/77, C:9/77, D:12/77, E:2/78, F:4/78, G:8/78, H:11/78, I:2/79, J:5/79, K:7/79, L:10/79, M:2/80, N:4/80, O:8/80, P:10/80, Q:1/81, R:4/81, S:8/81, T:11/81, U:2/82. Oil-spill was during 3/78, i.e. between E and F. Note more linear configuration for phyla

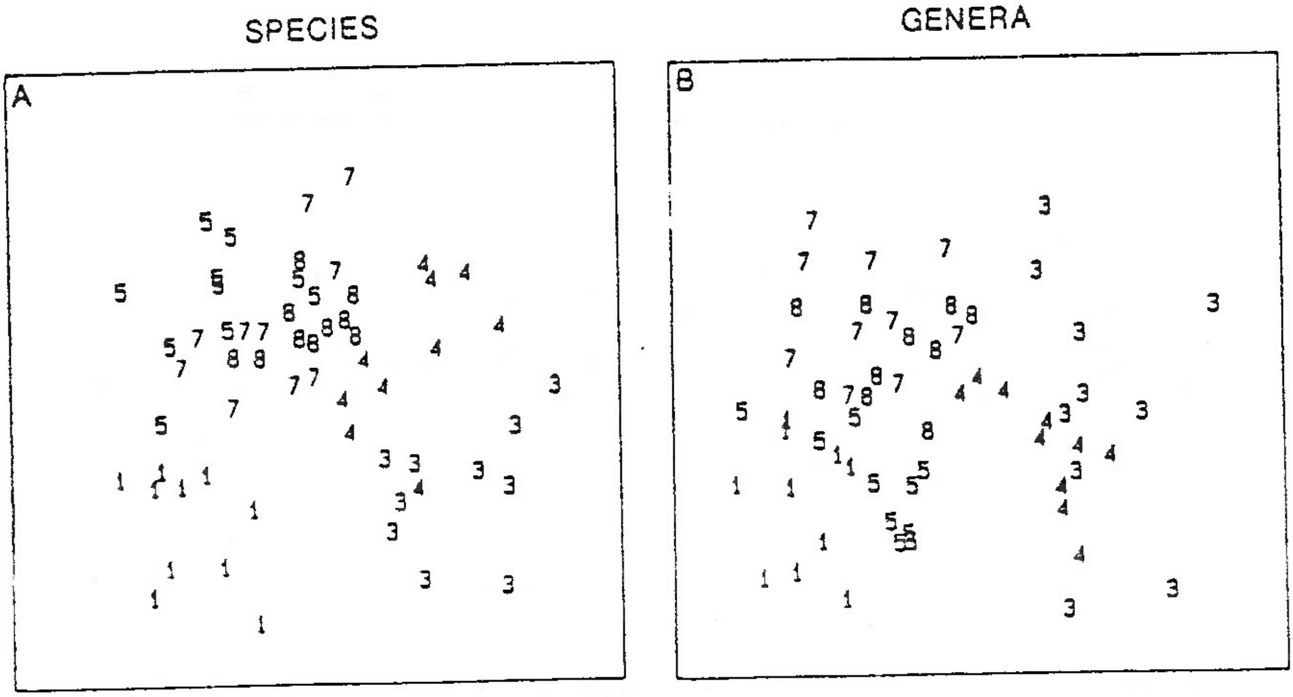


Fig. 10.5 MDS for coral species (n=75) and genus (n=24) cover data at South Pari Island, Indonesia. El Niño occurred in 1982-3. 1=1981, 3=1983, etc

GRAPHICAL/DISTRIBUTIONAL EXAMPLES

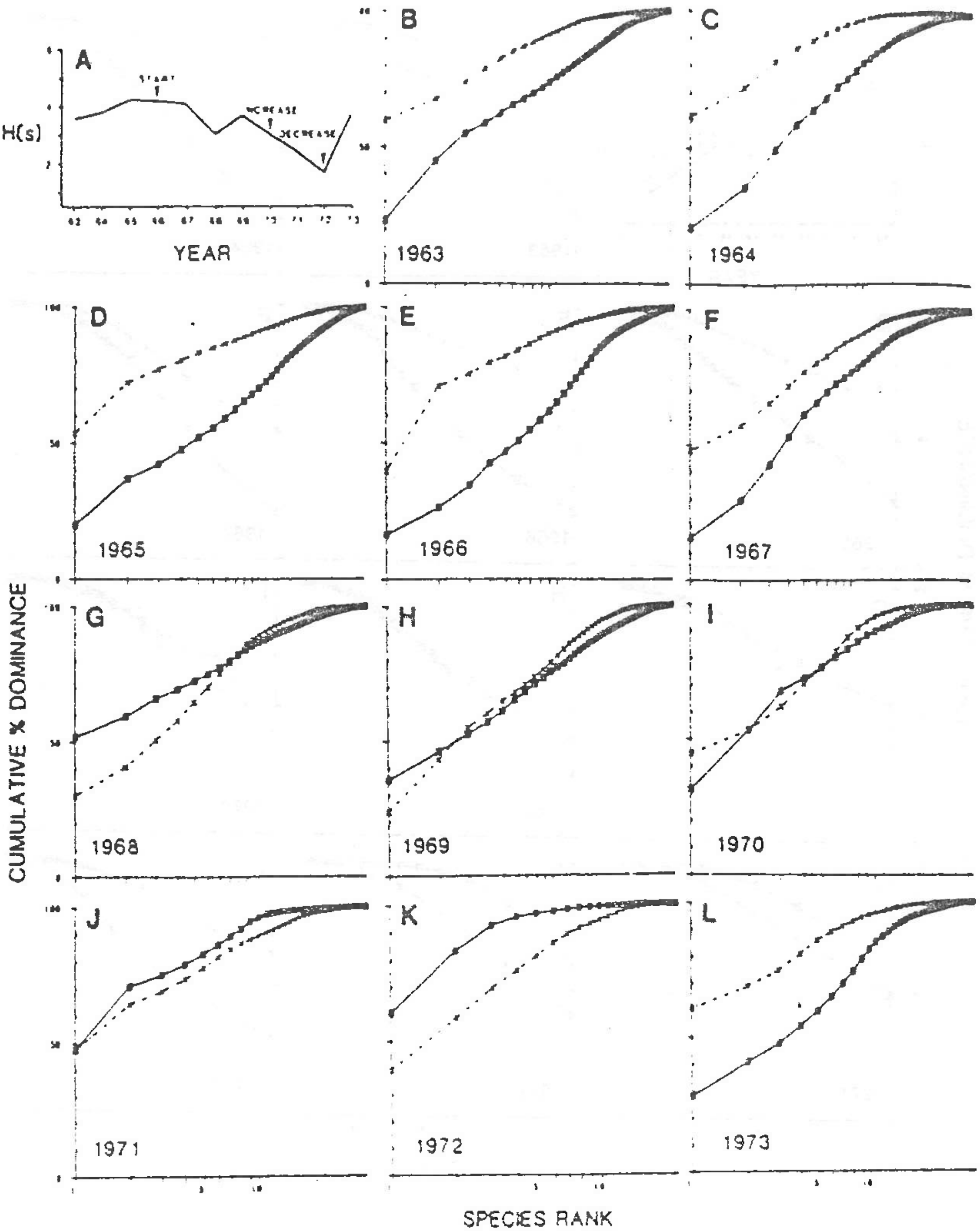


Fig. 10.6 Loch Linnhe macrofauna. (A) Diversity H′, (B)-(L) "ABC" curves for 11 years, of biomass (crosses) and abundance (squares). Analysis at species level, Warwick (1986)

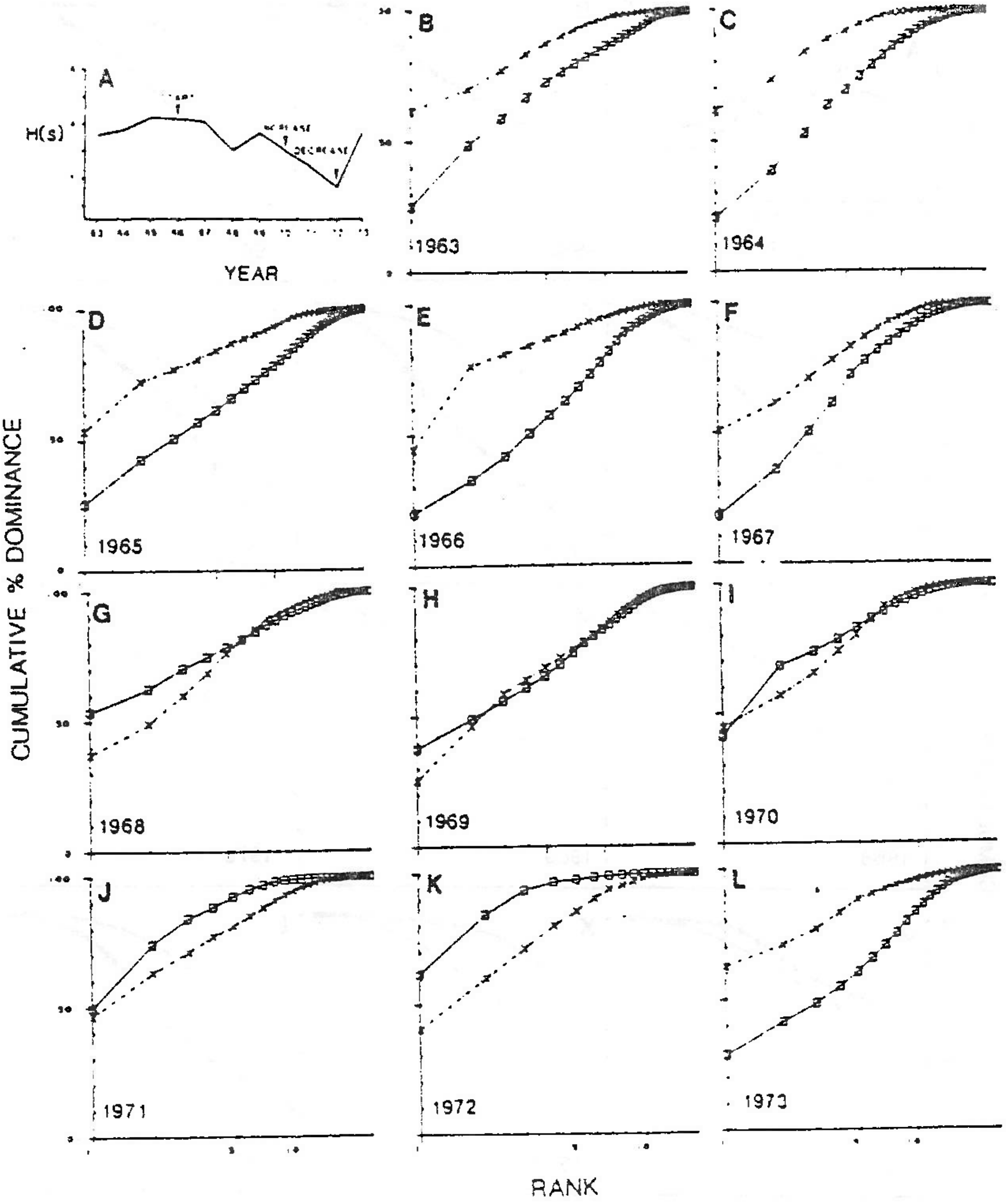


Fig. 10.7 Loch Linnhe macrofauna. (A) Diversity H', (B)-(L) "ABC" curves for 11 years, of biomass (crosses) and abundance (squares) for data aggregated to families, Warwick (1986)

UNIVARIATE EXAMPLE

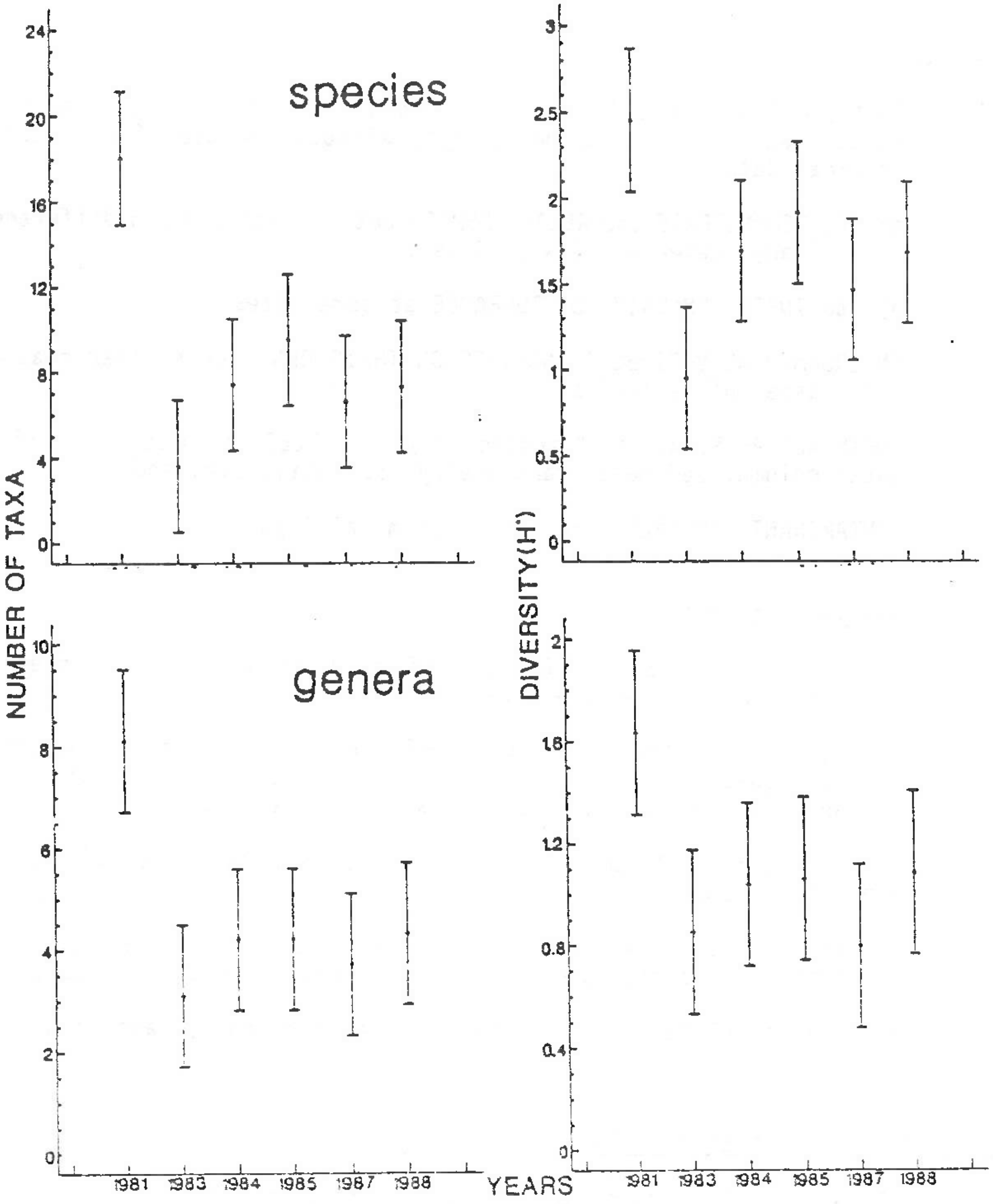


Fig. 10.8 Plots of number of taxa and Shannon diversity for reef corals at South Tikus Island, Indonesia, showing impact and partial recovery from 1982-3 El Niño. Species data (upper) have been aggregated into genera (lower). Note similarity of patterns

# LECTURE 11

## LINKING MULTIVARIATE AND UNIVARIATE COMMUNITY ANALYSES TO ENVIRONMENTAL VARIABLES

### APPROACH

1) FAUNAL AND ENVIRONMENTAL ANALYSIS SEPARATED initially, i.e. the biota is allowed to "tell its own story", without the use of physical or chemical data:

   a) to DEMONSTRATE the RELATIONSHIPS between samples and differences (if any) between sites (/times),

   b) to INFER COMMUNITY DISTURBANCE at some sites.

2) ENVIRONMENTAL VARIABLES ANALYSED ON THEIR OWN, for similar reasons. Two classes of variables:

   "NATURAL" PHYSICAL (or "background") VARIABLES, such as depth of the water column, sediment granulometry, salinity, etc. and

   CONTAMINANT VARIABLES, measuring chemical impact.

   Analysis attempts:

   a) to DEMONSTRATE DIFFERENCES (if any) in physical or chemical variables between the sites,

   b) to REDUCE the COMPLEXITY of the environmental measures, particularly the chemical data, so the nature of the impact (if any) can be summarised by a few key variables.

3) SUMMARY REPRESENTATIONS of both biological and environmental analyses are VIEWED TOGETHER:

   a) to examine whether changes between sites (/times) seem to be the product of differences in "natural" environmental variables, or

   b) are correlated with inferred or measured contaminant impact.

### ANALYSIS OF ENVIRONMENTAL DATA

UNIVARIATE: Background (physical) variables are typically univariate, with little variability between replicates within a site (e.g. water depth).

Where there is variability, and it is helpful to establish site differences, use ANOVA and confidence intervals (e.g. as for diversity).

MULTIVARIATE: Chemical measurements can often be highly multivariate (e.g. wide range of PAH compounds, PCB congeners, heavy metals etc.)

Example: Frierfjord sediment - heavy metals.

Table 11.1

Frierfjord sediments. Metal concentrations (μg $g^{-1}$ dry wt, Fe as %) in top 2 cm from 3 replicate cores at sites A-E,G. Abdullah & Steffenak (1988).

| Site | Cu | Zn | Pb | Ni | Cr | Cd | Mn | Fe |
|---|---|---|---|---|---|---|---|---|
| A | 28 | 141 | 73 | 33 | 40 | 0.8 | 454 | 3.5 |
| | 26 | 139 | 71 | 30 | 40 | (0.6) | 653 | 3.3 |
| | 27 | 147 | 67 | 29 | 35 | (0.6) | 503 | 3.1 |
| B | 48 | 238 | 134 | 33 | 50 | (0.6) | 1050 | 3.5 |
| | 47 | 228 | 130 | 32 | 50 | 1.1 | 2880 | 3.5 |
| | 64 | 297 | 167 | 32 | 40 | 1.1 | 664 | 3.1 |
| C | 44 | 228 | 135 | 35 | 51 | 0.8 | 1500 | 4.1 |
| | 42 | 216 | 126 | 35 | 60 | 0.8 | 3570 | 4.2 |
| | 42 | 208 | 117 | 33 | 45 | 1.1 | 5880 | 4.0 |
| D | 48 | 241 | 142 | 37 | 56 | 0.9 | 1720 | 4.3 |
| | 39 | 205 | 114 | 33 | 50 | 0.8 | 8480 | 4.4 |
| | 44 | 238 | 141 | 35 | 34 | 1.1 | 5440 | 4.1 |
| E | 38 | 199 | 160 | 22 | 40 | 0.8 | 484 | 2.2 |
| | 40 | 241 | 156 | 25 | 40 | 1.1 | 925 | 2.1 |
| | 107 | 275 | 184 | 28 | 45 | 1.1 | 1400 | 2.5 |
| F | 48 | 328 | 118 | 32 | 35 | 3.6 | 10380 | 3.1 |
| | 44 | 296 | 110 | 30 | 35 | 3.1 | 5880 | 3.0 |
| | 47 | 320 | 118 | 32 | 35 | 3.4 | 7430 | 3.0 |
| G | 67 | 349 | 212 | 35 | 61 | 2.2 | 1060 | 2.8 |
| | 70 | 357 | 229 | 35 | 66 | 2.5 | 638 | 2.7 |
| | 77 | 417 | 267 | 38 | 70 | 4.5 | 619 | 2.6 |

SAME RANGE OF MULTIVARIATE METHODS AVAILABLE as for faunal analyses (replace species by chemical "species"). However, type of data is different:

a) ZEROS do NOT predominate.
b) distribution NOT highly RIGHT-SKEWED.
c) REDUNDANCY can be very extreme, i.e. similar chemical compounds correlate very closely with each other along a spatial contaminant gradient.

So, possibly after (mild) TRANSFORMATION (e.g. √),

a) MULTIVARIATE NORMAL assumptions often justified;

b) PCA is useful, a 2-d ordination often giving a good representation of site chemistry,

c) TESTING of site differences can either be by MANOVA (e.g. Wilks' Λ) or by ANOSIM on a Euclidean distance dissimilarity matrix.

Example: Frierfjord sediment metals.

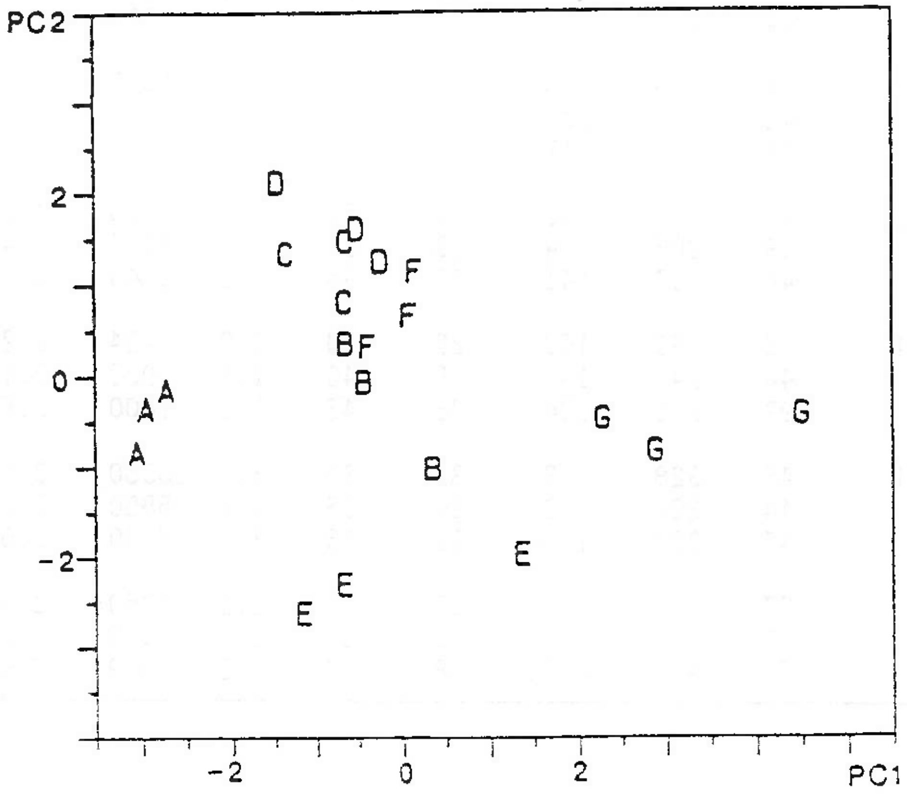


Fig. 11.1 Frierfjord sediments. 2-d PCA of metal data of Table 11.1 (√ - transformed and normalised)

NOTE, in Fig. 11.1:

1) First 2 PCs ACCOUNT FOR 69% OF VARIABILITY, so 2-d ordination is not too bad a representation.

2) Some DIFFERENCES BETWEEN SITES ($p<0.001$ in ANOSIM test), principally between A, G and the rest.

3) PC1 represents an AXIS OF INCREASING CONTAMINANT LOAD, the weights given to the (normalised) Cu, Zn, Pb, Ni, Cr, Cd, Mn, Fe levels being 0.41, 0.48, 0.46, 0.30, 0.35, 0.35, -0.05 and -0.21.

4) PC1 AXIS is thus a UNIVARIATE descriptor of the overall metal load, useful in relating this chemistry to faunal descriptions.

5) Though it exists, the CONTAMINANT GRADIENT is WEAK, no more than a factor of 2 or 3 between the extremes, A and G. (PAH gradient weaker still).

## RELATION TO FAUNAL ANALYSES - FIRST APPROACH

SELECT at most 2 or 3 DESCRIPTORS of the CONTAMINANT GRADIENT (eg. one for metals, one for hydrocarbons) - even 2 or 3 could be ambitious if the different classes of contaminants are well-correlated.

The two cases considered below are when the biological data are UNIVARIATE (eg. diversity indices) and when they are MULTIVARIATE (eg. ordinations).

### UNIVARIATE

REGRESSION is a possible technique: either SIMPLE LINEAR REGRESSION (1 environmental variable)

or MULTIPLE LINEAR REGRESSION (for 2 or more)

or NON-LINEAR REGRESSION (if there is a range of contaminant values and sufficient replicates to justify a more complex "dose-response" curve.)

Example: Frierfjord macrofauna.

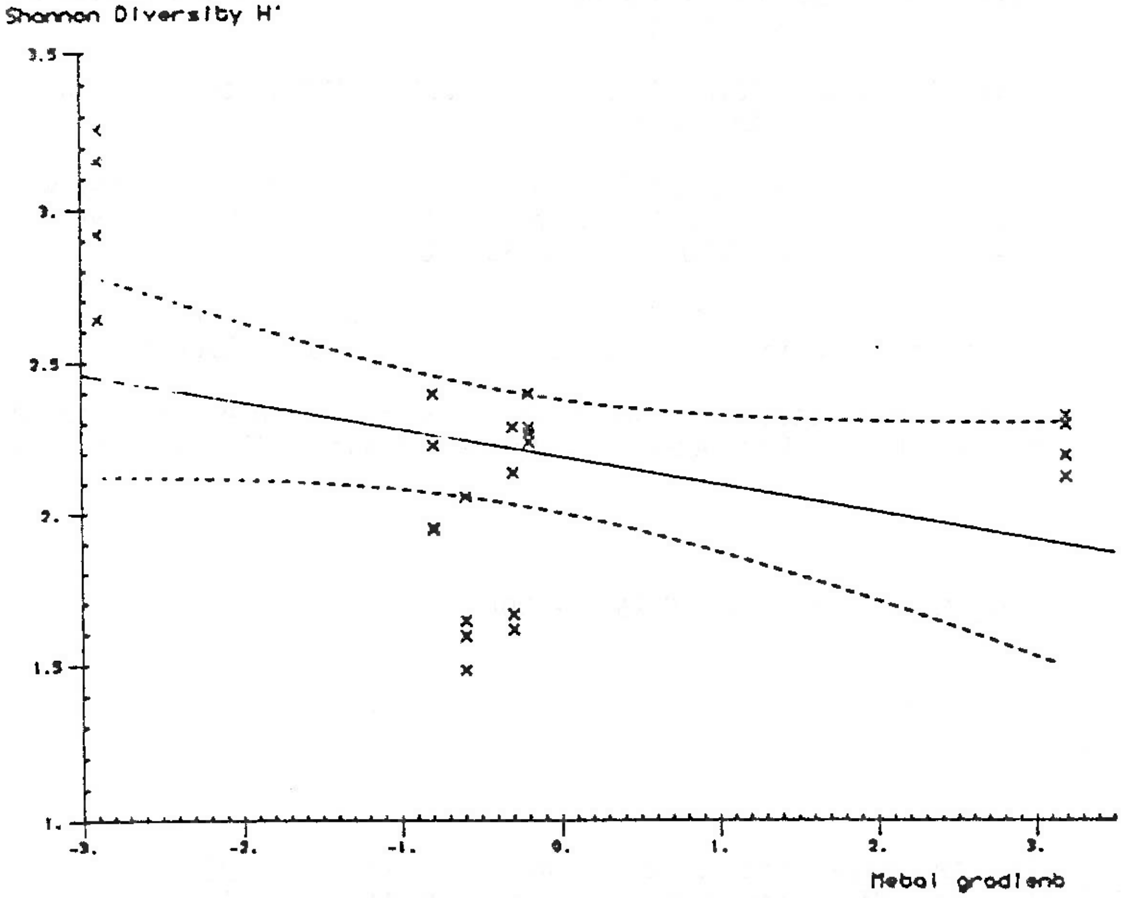


Fig. 11.2 Frierfjord macrofauna abundances. Shannon diversity H′ regressed on an overall measure of sediment metal concentration (latter is mean PC1 at each of the 6 sites, from the PCA of Fig. 11.1). x - replicate grabs, — fitted regression line, --- 95% confidence "funnel" for the mean H′ at any metal concentration

NOTE: Simple linear regression of H′ on metal levels is not convincing!

a) Slope just fails to differ significantly from zero, at 5%.

b) Linear relation is not adequate (but data does not justify more complex fit).

c) Most prominent feature (clear from the earlier ANOVA also - Fig. 6.1) is the general drop in diversity from the "reference′ site (A).

MULTIVARIATE

SUPERIMPOSITION OF ENVIRONMENTAL VARIABLES OF FAUNAL ORDINATION: an effective visual technique performed separately for each environmental variable.

This may allow a GRADIENT in the ENVIRONMENTAL VARIABLE to be matched visually to a GRADIENT of change in the COMMUNITY structure.

Example: Bristol Channel zooplankton, April 1978.

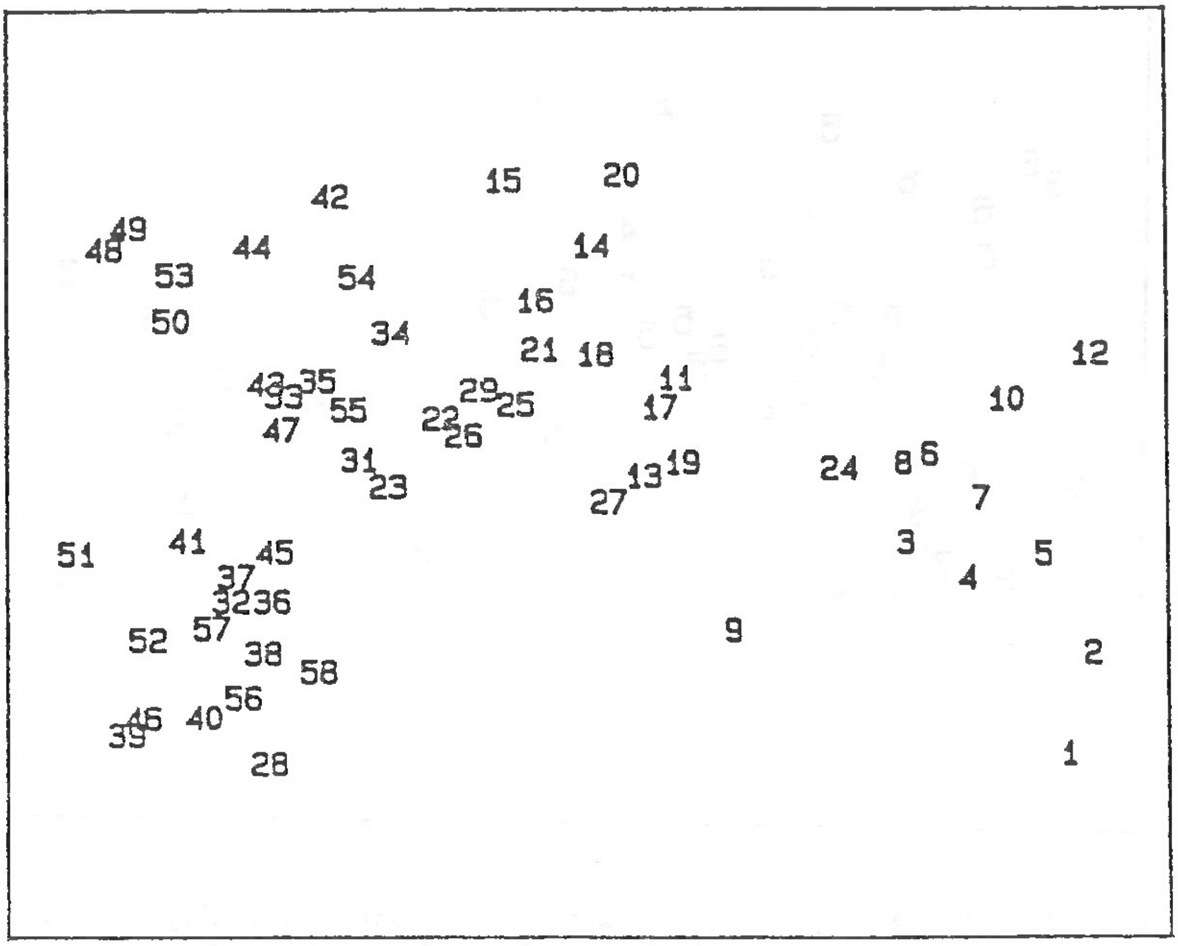


Fig. 11.3 MDS of 57 sites (from Bray-Curtis similarities, on √√ - transformed counts; stress = 0.11). For map of sites and corresponding cluster analysis, see Figs. 3.2 and 3.3

Though clear evidence of clusters (from Fig. 3.3), overall pattern is one of GRADATION of COMMUNITY STRUCTURE across the plot (note characteristic "arching", common for strong gradation).

Physical variable driving the structure is SALINITY s, ranging from 24.6‰ (site 1) to 35.1‰ (site 52). Non-linear TRANSFORMATION needed (36‰ → 35‰ is a more important change than 26‰ → 25‰); suggest

$$s^* = a - b.\log(36 - s) \qquad (11.1)$$

Choosing a = 8.33, b = 3 gives $1 \le s^* \le 9$, and can:

CATEGORISE (transformed) SALINITY into (say) 9 groups (s* to nearest integer), and SUPERIMPOSE on MDS.

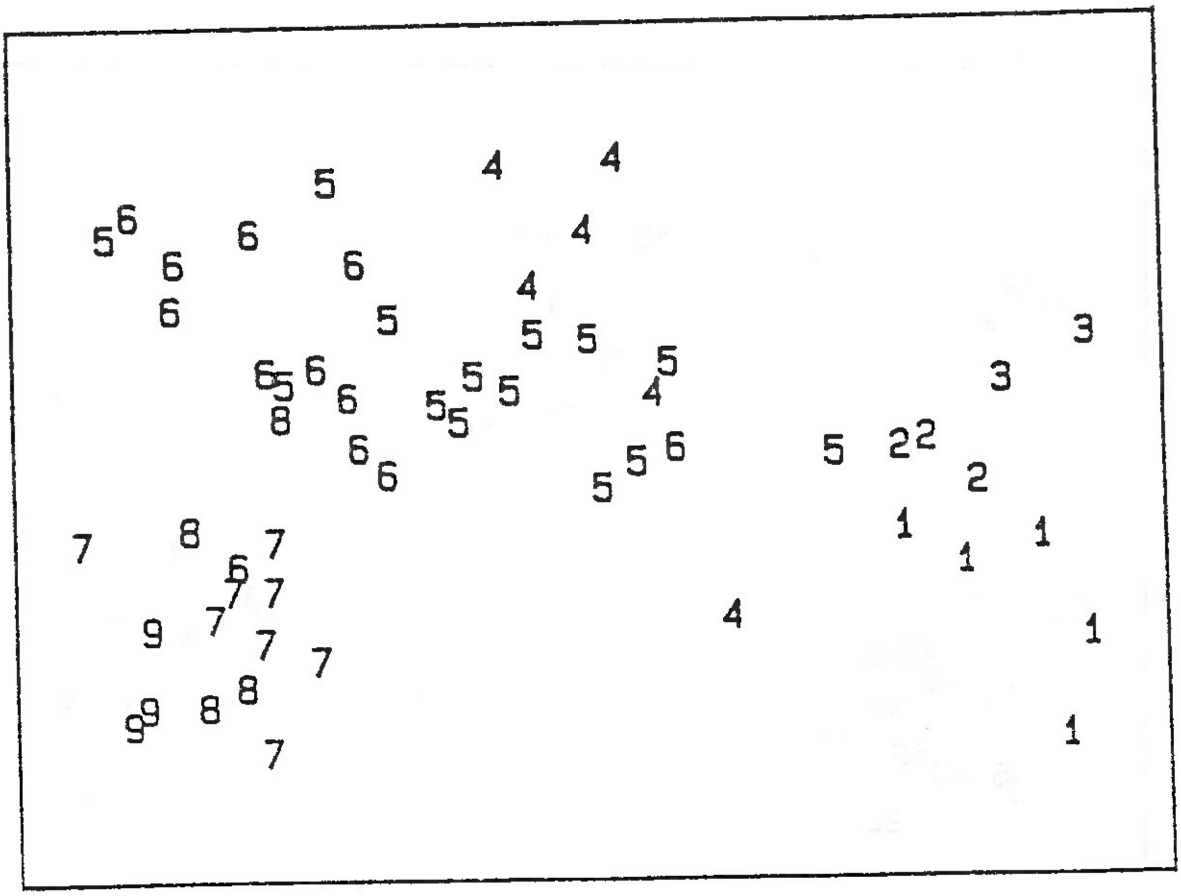


Fig. 11.4 MDS of 57 sites, with increasing salinity categories superimposed. 1: ≤26.3, 2: (26.3, 29.0), 3: (29.0, 31.0), ..., 8: (34.7, 35.1), 9: ≥35.1‰

Alternatively, at each sample point on the faunal MDS, draw a symbol (e.g. circle) with SIZE PROPORTIONAL to the ENVIRONMENTAL VARIABLE value for the sample.

Example: Frierfjord macrofauna counts (√√ - transformed)

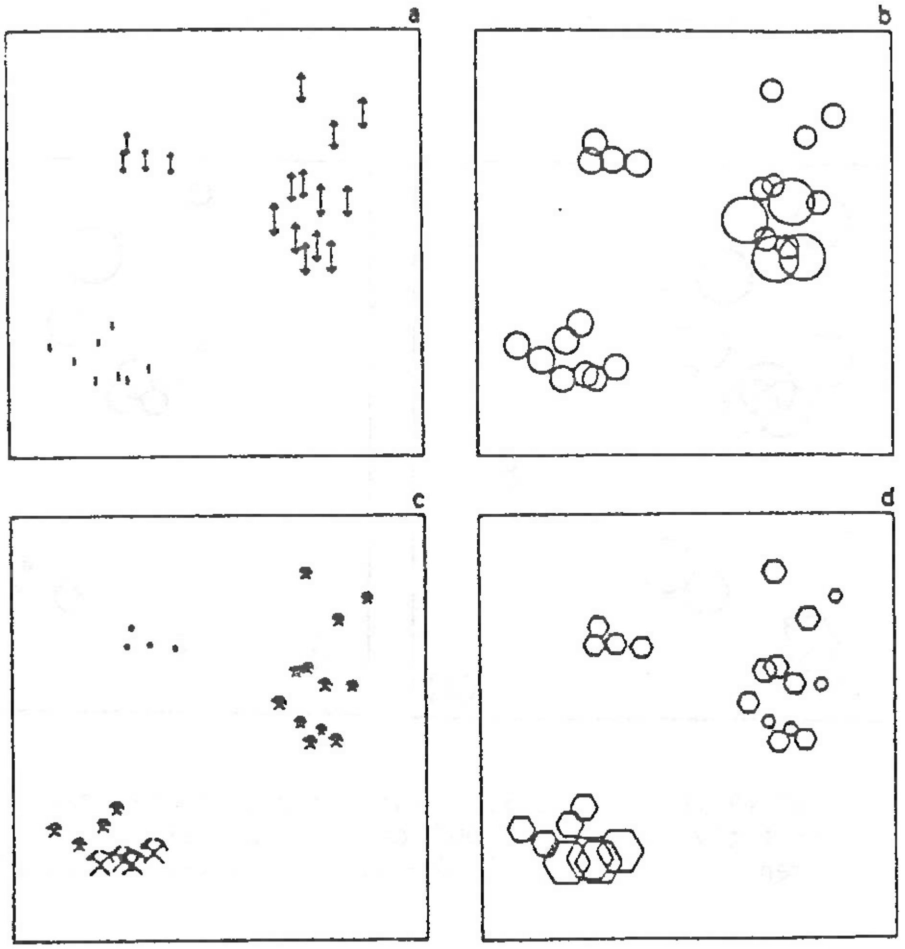


Fig. 11.5 MDS of sites A-E,G with superimposed values of (a) water depth (22-113 m), (b) sediment median grain size (7.8-16.5 µm), (c) metal levels (PC1 in Fig. 11.1) and (d) "total" PAH (4.4-14.8 µg $g^{-1}$)

1) Site grouping on the MDS bears LITTLE RELATION to the (weak) metal and PAH CONTAMINANT GRADIENTS.

2) Sediment granulometry is NOT A DETERMINANT of COMMUNITY DIFFERENCES here (B & C span the range of grain sizes but have the same communities).

3) DEPTH-RELATED differences between the sites appear to be the major CORRELATE of COMMUNITY DIFFERENCES. (Seasonal anoxia in the deeper parts of the fjord is likely to be a significant "stress" factor.)

Sometimes MORE THAN ONE AXIS OF CHANGE MAY BE SEEN, correlating with different environmental variables.

Example: Exe nematode abundances, Field et al. (1982)

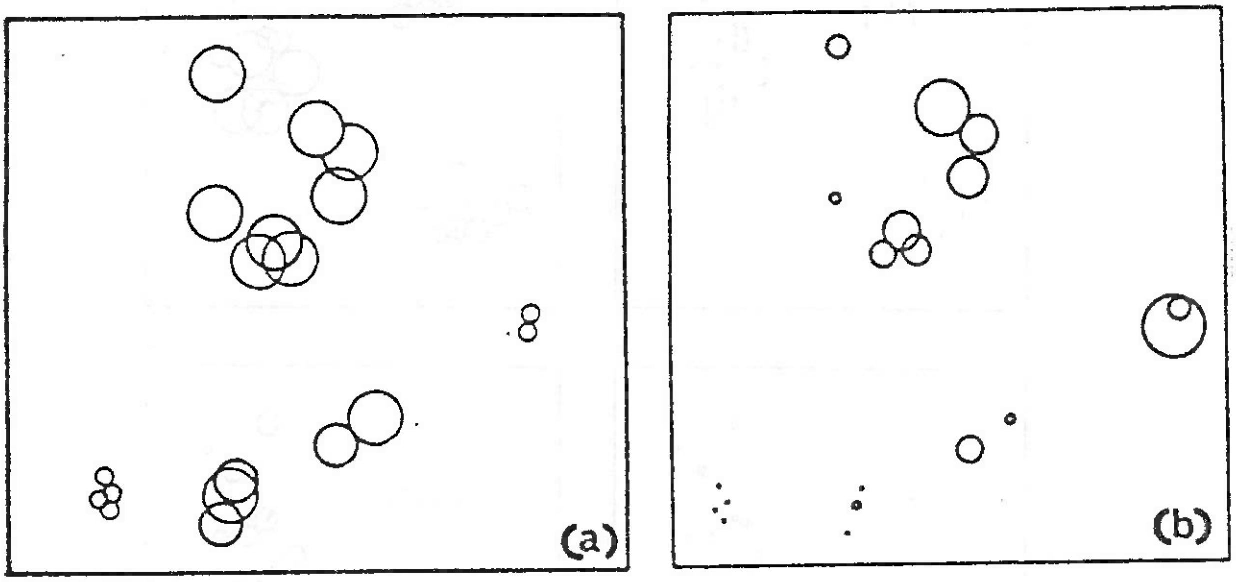


Fig. 11.6 MDS of 19 sites (Fig. 5.1), with values of: (a) mean salinity of interstitial water (10-90% of standard seawater), (b) median sediment particle size (0.06-1.14 mm), superimposed at each site

Grain size forms a gradient from bottom left to top right, whereas salinity distinguishes the "middle" from the "end" sites along the first MDS axis.

Though the visual approach is generally more helpful, FORMAL TESTING of gradients can be performed by:

a) REGRESSING each environmental variable on the (x,y) CO-ORDINATES of the SAMPLE LOCATIONS on the MDS; this would be multiple linear regression (and not appropriate for a curvilinear gradient).

b) Using 2-WAY ANOSIM on sites (treated as replicates), which are categorised by, say, 2 environmental variables at 2 levels, e.g. deep/shallow, high/low contaminant loads. This would need a reasonable number of sites (with some in all 4 combinations).

## RELATION TO FAUNAL ANALYSES - SECOND APPROACH

First approach designed mainly to show COMMUNITY pattern related to ONE ENVIRONMENTAL VARIABLE at a time. Alternative considers ALL environmental variables together and COMPARES ordination of biota to ORDINATION of environmental variables.

Example: Exe nematode abundances.

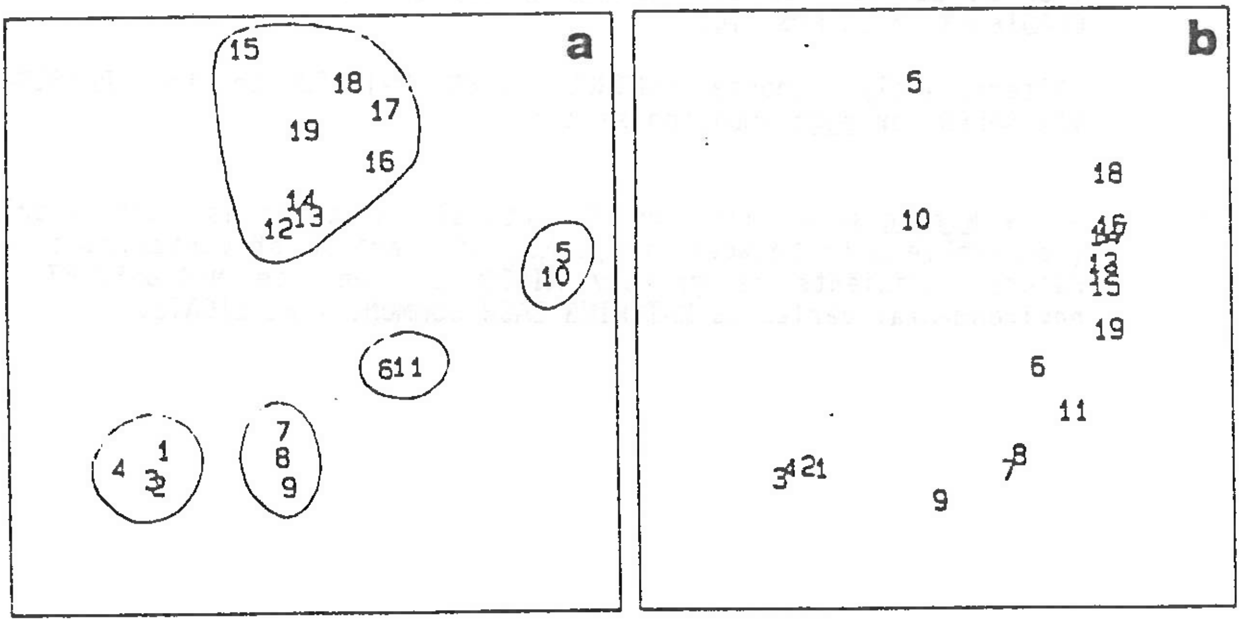


Fig. 11.7 (a) MDS of 19 sites (as in Fig. 5.1), (b) PCA of 4 environmental variables (salinity, median particle size, % organics, depth of $H_2S$ layer)

The close match of patterns shows these 4 variables "EXPLAIN" biota clusters (in Fig. 11.7a) well. Two questions: Would subset of environmental variables do as well? Would more variables do better? (e.g. height up shore, water table depth.)

Answer by DEFINING MATCH between two ordinations as some form of RANK CORRELATION ($\rho$) between underlying DISSIMILARITY MATRICES (Bray-Curtis and Euclidean distance, respectively). Then find subset of environmental variables which MAXIMISES $\rho$. Here, this is the 4 variables in Fig. 11.7b.

## IMPLICATIONS FOR DESIGN

1) SITE SELECTION: where there is choice, attempt to select sites such that VARIATION IN "NUISANCE" (physical) VARIABLES IS SMALL, (i.e. small enough not to have a significant affect on community structure).

2) Where between-site variation in natural variables is considerable, AVOID DESIGNS in which important physical variables are TOTALLY CONFOUNDED (i.e. run in parallel) with contaminant gradients. It may then be possible to DISTINGUISH SEPARATE PHYSICAL AND CONTAMINANT GRADIENTS in an MDS plot.

   (Alternatively, choose CONTROL SITE<u>S</u> MATCHED to the PHYSICAL VARIABLES for <u>each</u> impacted site.)

3) Where <u>within</u>-site variation in natural variables is considerable (comparable with between-site), MDS distinction of contaminant and natural gradients is greatly AIDED by separate MEASUREMENT of environmental variables MATCHING EACH COMMUNITY REPLICATE.

# LECTURE 12

## CAUSALITY: COMMUNITY EXPERIMENTS IN THE FIELD AND LABORATORY

In experimental situations we can investigate the effects of a single factor (the TREATMENT) on community structure, while other factors are held constant or controlled. There are three main categories of experiments that can be used:

1. 'NATURAL EXPERIMENTS' - Nature provides the treatment: i.e. we compare places or times which differ in the intensity of the environmental factor in question.

2. FIELD EXPERIMENTS - The experimenter provides the treatment: i.e. environmental factors are manipulated in the field.

3. LABORATORY EXPERIMENTS - Environmental factors are manipulated by the experimenter in laboratory mesocosms or microcosms.

The degree of 'naturalness' (hence realism) decreases from 1-3, but the degree of control which can be exerted over confounding environmental variables increases from 1-3.

In all cases care should be taken to avoid PSEUDOREPLICATION, i.e. the treatments should be replicated, rather than a series of 'replicate' samples taken from a single treatment (pseudoreplicates). This is because other confounding variables, often unknown, may also differ between the treatments. It is also important to run experiments long enough for community changes to occur: this favours components of the fauna with short generation times (see Lecture 13).

### NATURAL EXPERIMENTS

The obvious logical flaw with this approach is that its validity rests on the assumption that places or times differ only in the intensity of the selected environmental factor (treatment). Experimental design is often a problem, but statistical techniques such as TWO-WAY ANOVA or TWO-WAY ANOSIM, which enable us to examine the treatment effect allowing for differences between sites, are useful.

Example: The effects of disturbance by soldier crabs (*Mictyris platycheles*) on meiobenthic community structure.

LOCATION: Sand-flat at Eaglehawk Neck, S.E. Tasmania.

SAMPLING: Sediment disturbed by crabs in discrete patches. 4 x 5 $m^2$ blocks of 4 samples with each block including 2 disturbed and 2 undisturbed:

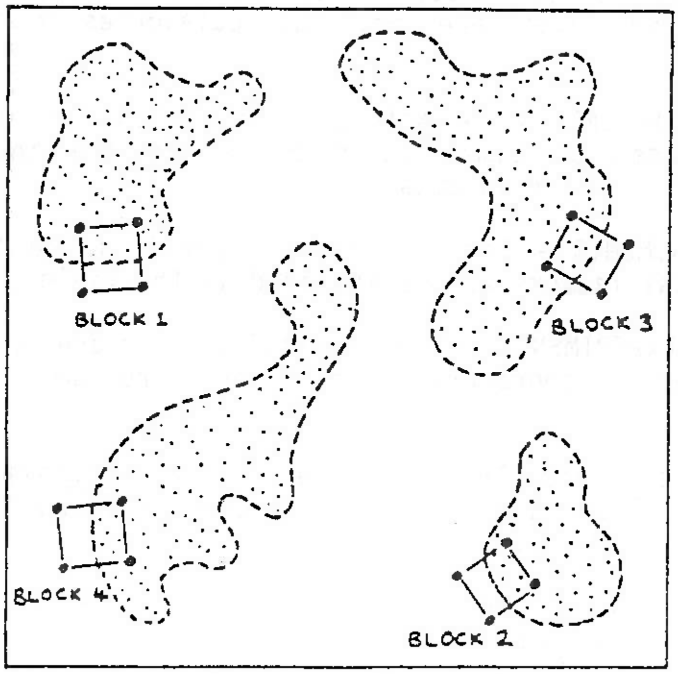


Fig. 12.1 Sketch showing the type of sample design. Sample positions (large dots) in relation to disturbed sediment patches (stippled)

UNIVARIATE INDICES:

Table 12.1

Mean values per core sample of univariate measures for nematodes, copepods and total meiofauna (nematodes + copepods) in the disturbed and undisturbed areas. The significance levels for differences are from a two-way ANOVA, i.e. they allow for differences between blocks, although these were not significant at the 5% level.

| | Tot.ind. | Tot.sp. | d | H′ | J′ |
|---|---|---|---|---|---|
| *Nematodes* | | | | | |
| Disturbed | 205 | 14.4 | 2.6 | 1.6 | 0.58 |
| Undisturbed | 200 | 20.1 | 3.7 | 2.2 | 0.74 |
| Significance (%) | 91 | 1 | 0.3 | 0.1 | 1 |
| *Copepods* | | | | | |
| Disturbed | 94 | 5.4 | 1.0 | 0.96 | 0.59 |
| Undisturbed | 146 | 5.7 | 1.0 | 0.84 | 0.49 |
| Significance (%) | 11 | 52 | 99 | 52 | 38 |
| *Total meiofauna* | | | | | |
| Disturbed | 299 | 19.8 | 3.4 | 2.0 | 0.66 |
| Undisturbed | 346 | 25.9 | 4.4 | 2.3 | 0.69 |
| Significance (%) | 48 | 1 | 3 | 3 | 16 |

For NEMATODES: significant reduction in total number of species, Species Richness, Shannon Diversity and Evenness in relation to disturbance.

For COPEPODS: no differences in any of these univariate measures.

GRAPHICAL/DISTRIBUTIONAL PLOTS

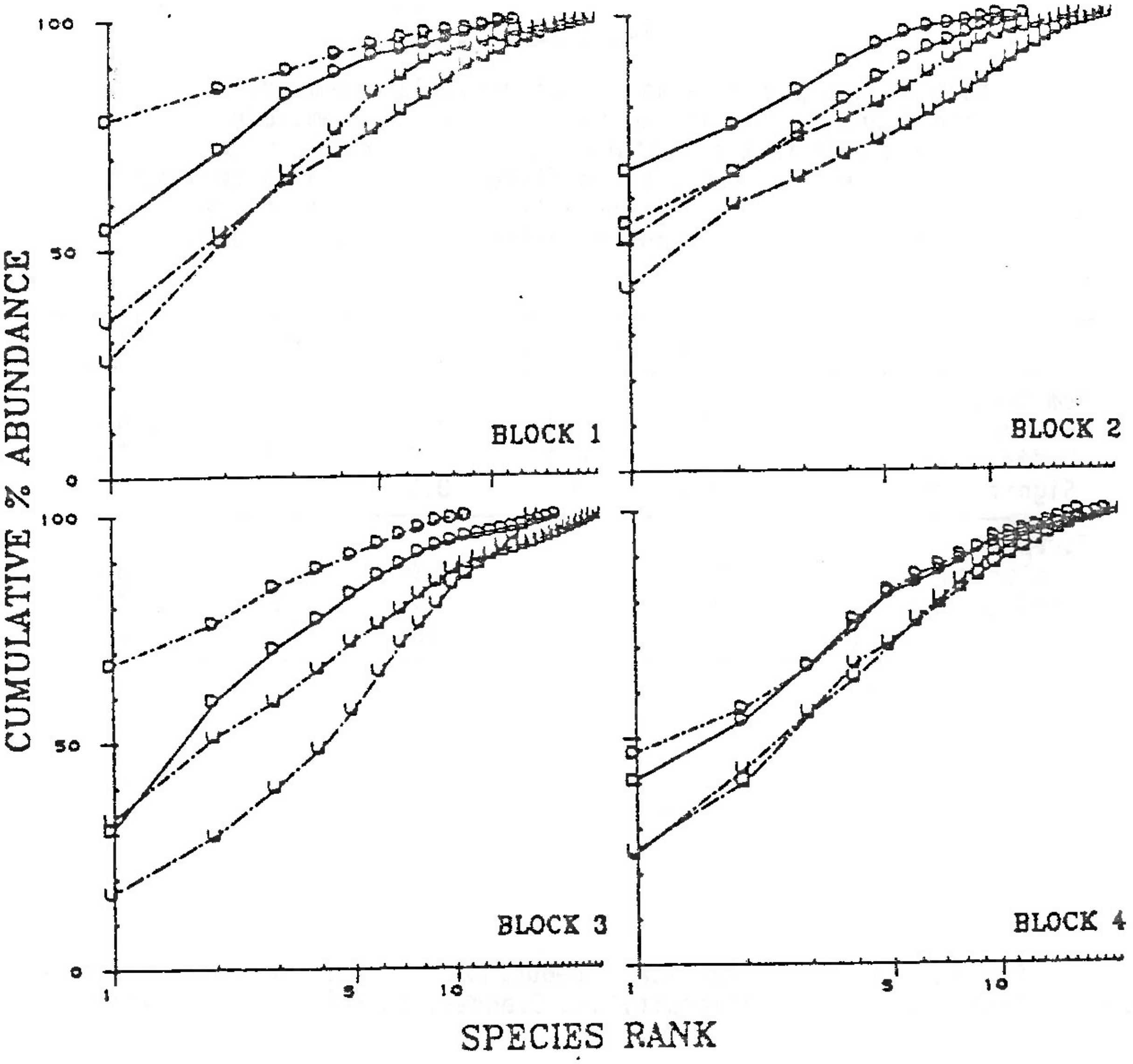


Fig. 12.2 Replicate k-dominance curves for NEMATODE abundance in each sampling block. D = disturbed, U = undisturbed

Summary statistics $K_A$ and R (see Lecture 8) both show significant treatment effect when tested with two-way ANOSIM.

For COPEPODS (figure not given here), k-dominance curves are intermingled and crossing, and there is no significant treatment effect on $K_A$ and R.

MULTIVARIATE ANALYSIS:

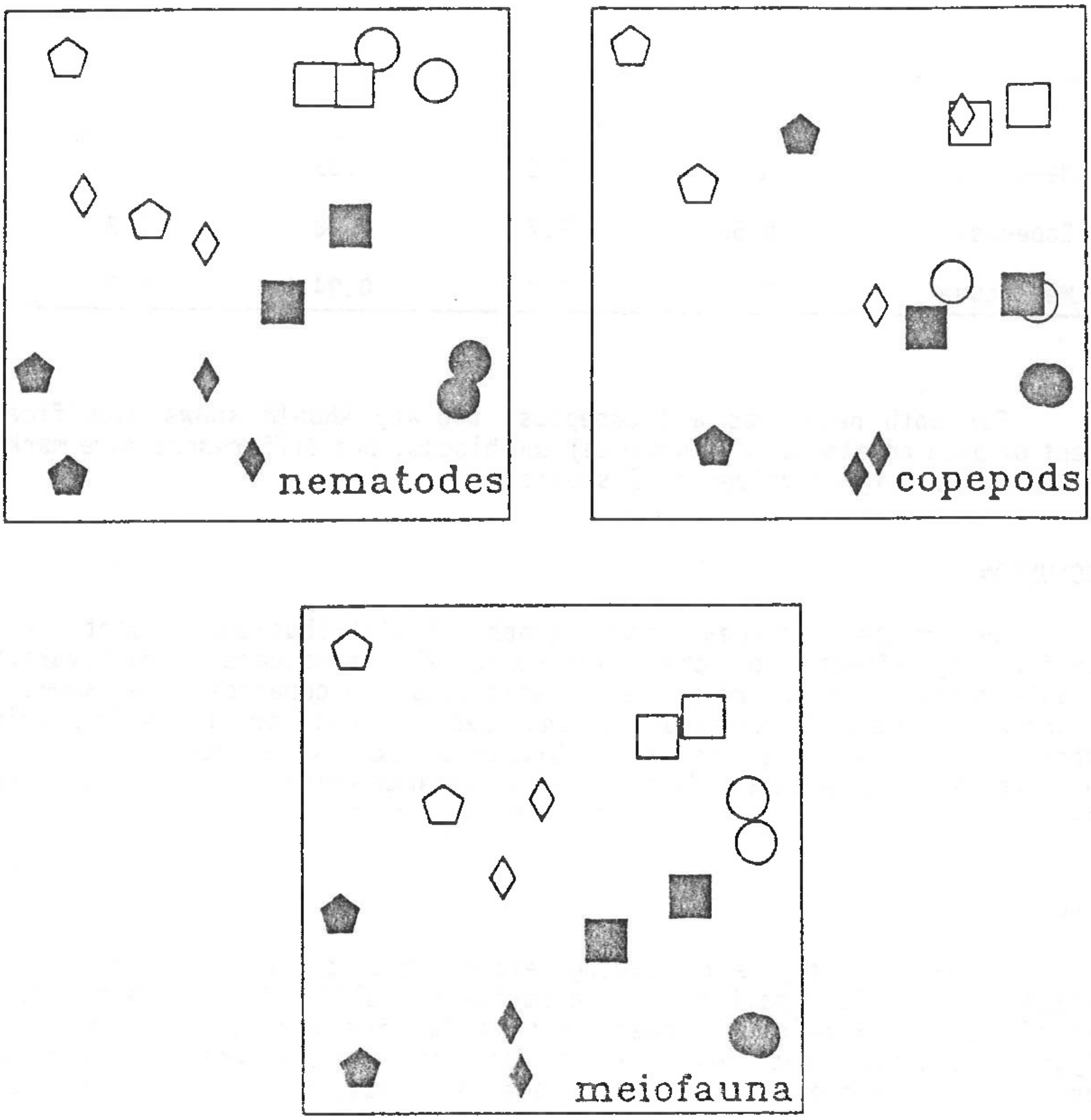


Fig. 12.3 MDS configurations for nematode, copepod and 'meiofauna' (nematode + copepod) abundance.
Circles = Block 1, Squares 2, Pentagons 3, Diamonds 4.
Open symbols = disturbed, shaded = undisturbed

Note similarities: both disturbed samples within each block are above both undisturbed; blocks arranged in sequence (left to right) 3,4,2,1.

Table 12.2

Results of the two-way ANOSIM test for treatment (disturbance/no disturbance) and block effects.

| | DISTURBANCE | | BLOCKS | |
|---|---|---|---|---|
| | R Statistic | Sig.(%) | R statistic | Sig. (%) |
| Nematodes | 1.0 | 1.2 | 0.99 | 0.2 |
| Copepods | 0.56 | 3.7 | 0.70 | 0.2 |
| Meiofauna | 0.94 | 1.2 | 0.94 | 0.2 |

For both nematodes and copepods, two-way ANOSIM shows significant effect of both treatment (disturbance) and blocks, but differences more marked for nematodes (higher values of R statistic).

CONCLUSIONS:

Univariate indices and graphical/distributional plot only significantly affected by crab disturbance for nematodes. Multivariate analysis reveals similar response for nematodes and copepods (i.e. seems to be more sensitive). In multivariate analyses, natural variations in species composition across the beach (i.e. between blocks) were about as great as those between treatments within blocks: disturbance effect would not have been clearly evidenced without this block sampling design.

## FIELD EXPERIMENTS

These include, e.g. caging experiments to exclude or include predators, controlled pollution of experimental plots, big-bag experiments with plankton. Have mostly been used so far for population rather than community studies: not possible to find an example where univariate, graphical/distributional and multivariate techniques have all been applied.

Example: Effect of sediment particle diameter on a harpacticoid copepod community (Hockin, 1982).

LOCATION: Sandy estuarine beach, Ythan estuary, Scotland.

SAMPLING: 2 replicates of 4 grades of glass beads deployed in plastic trays in randomised block design at two tide levels. Left in field for 14 wks, with core sample taken every 5 days.

UNIVARIATE INDICES:

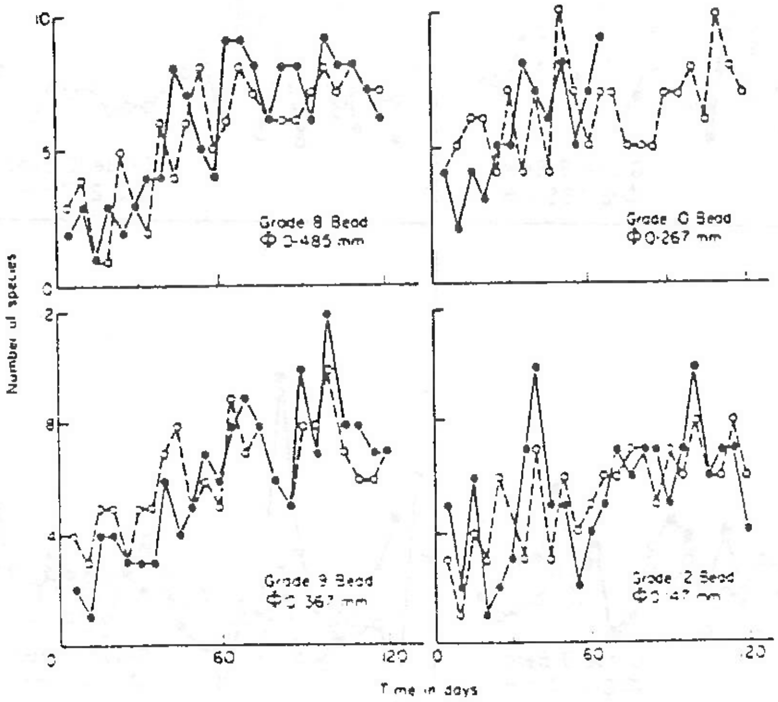


Fig. 12.4 Number of species at upper (solid circles) and lower (open circles) sites

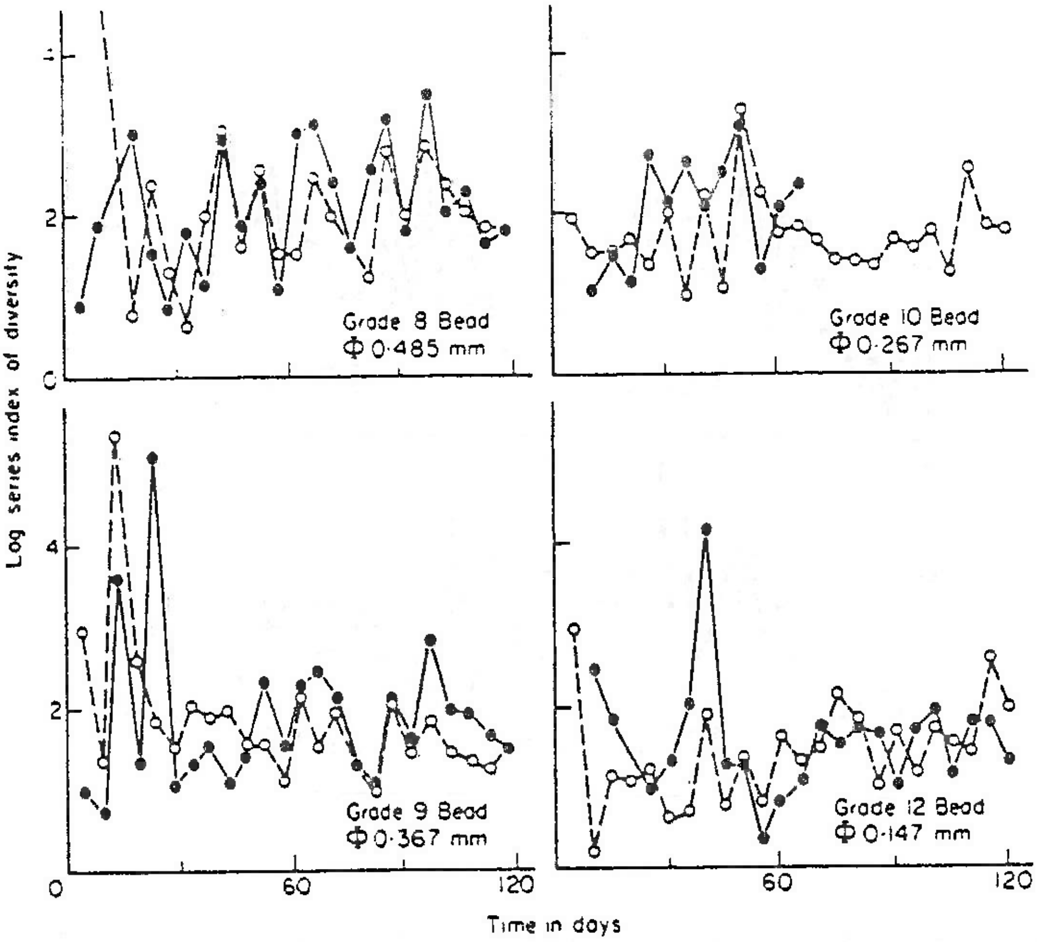


Fig. 12.5 The index of diversity (based on the log-series distribution) for upper (solid circles) and lower (open circles) sites

ANOVA on both the number of species and the species diversity revealed no significant differences with respect to the treatment (sediment particle size).

Table 12.3

Particle diameter of artificial monometric sediments in which the maximum population densities of the numerically dominant harpacticoid copepod species were found.

| COPEPOD SPECIES | PARTICLE DIAMETER (MM) |
|---|---|
| *Arenosetella germanica* | 0.267 |
| *Arenosetella tenuissima* | 0.367 |
| *Arenopontia subterranea* | 0.147 |
| *Evansula incerta* | 0.367 |
| *Stenocaris pygmea* | 0.267 |
| *Heterolaophonte minuta* | 0.485 |
| *Heterolaophonte littoralis* | 0.485 |
| *Esola typhlops* | 0.367 |
| *Paronychocamptus curticaudatus* | 0.485 |
| *Huntemannia jadensis* | 0.147 |
| *Nannopus palustris* | 0.147 |

Although no MULTIVARIATE ANALYSES were done, different species reached maximum abundance in different sediment grades. This suggests that a multivariate analysis may well have provided discrimination between treatments.

## LABORATORY EXPERIMENTS

More or less natural communities of some components of the biota can be maintained in laboratory mesocosms or microcosms (also in outdoor mesocosms), and subjected to a variety of manipulations.

Example: Effects of organic enrichment on meiofaunal community structure (Gee et al., 1985).

LOCATION: Sediment from Oslofjord; mesocosm at Solbergstrand, Norway.

SAMPLING: Undisturbed 0.25 $m^2$ box cores of sediment transferred to mesocosm basin. 4 replicate boxes dosed with high (200 g C $m^{-2}$) and low (50 g C $m^{-2}$) levels of powdered algae (*Ascophyllum*), with 4 undosed controls, in randomised block design. Meiofauna sampled 56 days after dosing: 5 cores from each box combined to give one sample.

UNIVARIATE INDICES:

Nematodes: No significant differences in species richness or diversity between treatments, but evenness significantly higher in enriched boxes than controls.

Copepods: Significant differences in species richness and evenness between treatments, but not in diversity.

Table 12.4

Univariate measures for all replicates at end of experiment, with F-ratio and significance levels from one-way ANOVA.

| Treatment | Sample number | Species richness | Shannon-Wiener index | Species evenness |
|---|---|---|---|---|
| Nematodes | | | | |
| Control | 1 | 3.023 | 2.245 | 0.750 |
| | 2 | 3.739 | 2.394 | 0.774 |
| | 3 | 3.357 | 2.470 | 0.824 |
| | 4 | 4.589 | 2.764 | 0.829 |
| | Total | 6.342 | 2.738 | 0.747 |
| Low dose | 1 | 4.386 | 2.856 | 0.877 |
| | 2 | 2.652 | 2.474 | 0.840 |
| | 3 | 4.669 | 2.885 | 0.875 |
| | 4 | 2.327 | 2.268 | 0.860 |
| | Total | 6.153 | 2.877 | 0.791 |
| High dose | 1 | 2.856 | 2.168 | 0.782 |
| | 2 | 2.824 | 2.388 | 0.843 |
| | 3 | 4.302 | 2.395 | 0.829 |
| | 4 | 4.088 | 2.466 | 0.853 |
| | Total | 5.508 | 2.677 | 0.759 |
| F-ratio | | 0.043 | 1.387 | 5.131 |
| Significance | | ns | ns | P<0.05 |
| Copepods | 1 | 2.525 | 1.927 | 0.927 |
| Control | 2 | 1.924 | 1.560 | 0.969 |
| | 3 | 2.502 | 1.768 | 0.908 |
| | 4 | 2.471 | 1.936 | 0.931 |
| | Total | 2.531 | 2.102 | 0.877 |
| Low dose | 1 | 1.804 | 1.597 | 0.643 |
| | 2 | 1.661 | 1.275 | 0.532 |
| | 3 | 1.655 | 1.160 | 0.484 |
| | 4 | 1.786 | 1.535 | 0.640 |
| | Total | 1.907 | 1.581 | 0.584 |
| High dose | 1 | 1.747 | 1.594 | 0.767 |
| | 2 | 0.973 | 0.997 | 0.620 |
| | 3 | 1.034 | 0.297 | 0.165 |
| | 4 | 1.179 | 1.696 | 0.872 |
| | Total | 1.666 | 1.683 | 0.702 |
| F-ratio | | 17.715 | 2.654 | 4.559 |
| Significance | | P<0.001 | ns | P<0.05 |

GRAPHICAL/DISTRIBUTIONAL PLOTS:

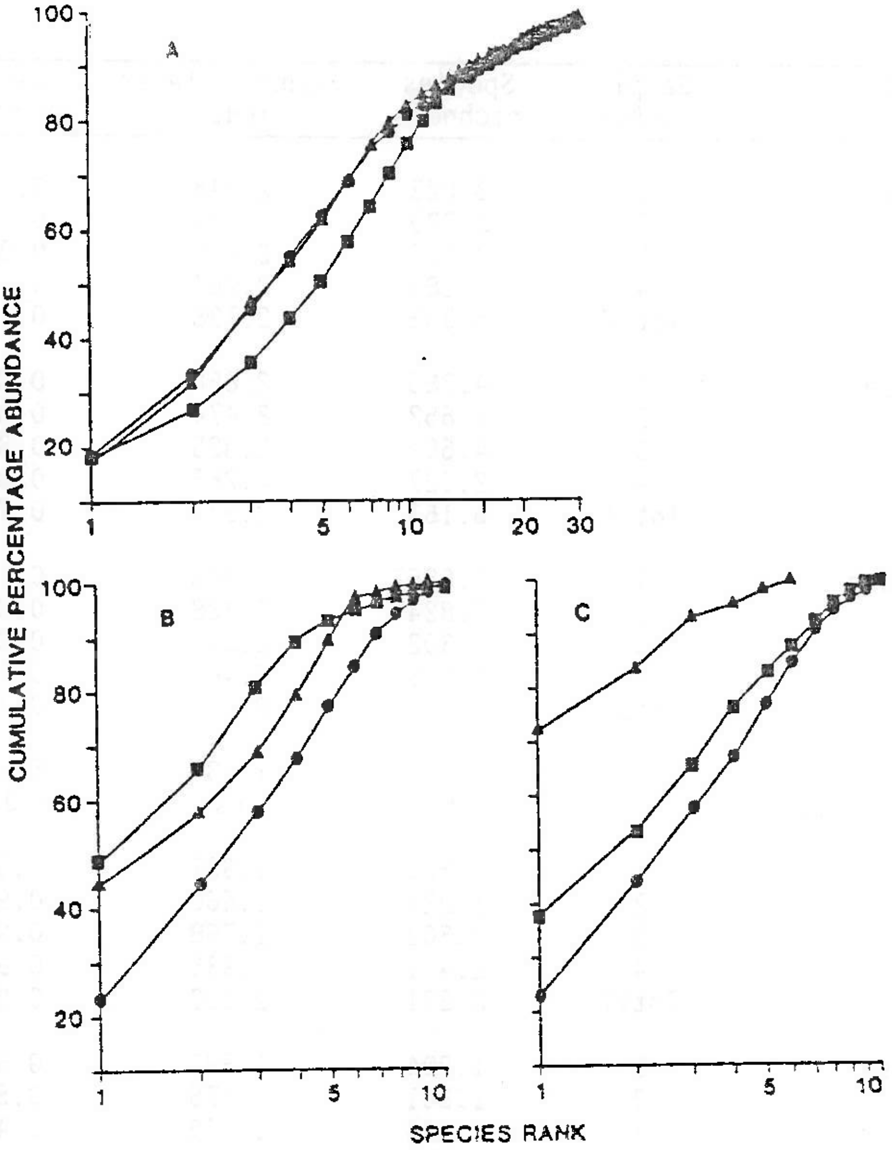


Fig. 12.6 k-dominance curves for A nematodes, b total copepods and C copepods omitting the 'weed' species of Tisbe for summed replicates of each treatment. Circles = control, squares = low dose, triangles = high dose

NEMATODES: No obvious treatment effect.

COPEPODS: Control with highest diversity; when *Tisbe spp.* omitted, sequence of increasing elevation of curves (decreasing diversity) from control to high dose.

MULTIVARIATE ANALYSES:

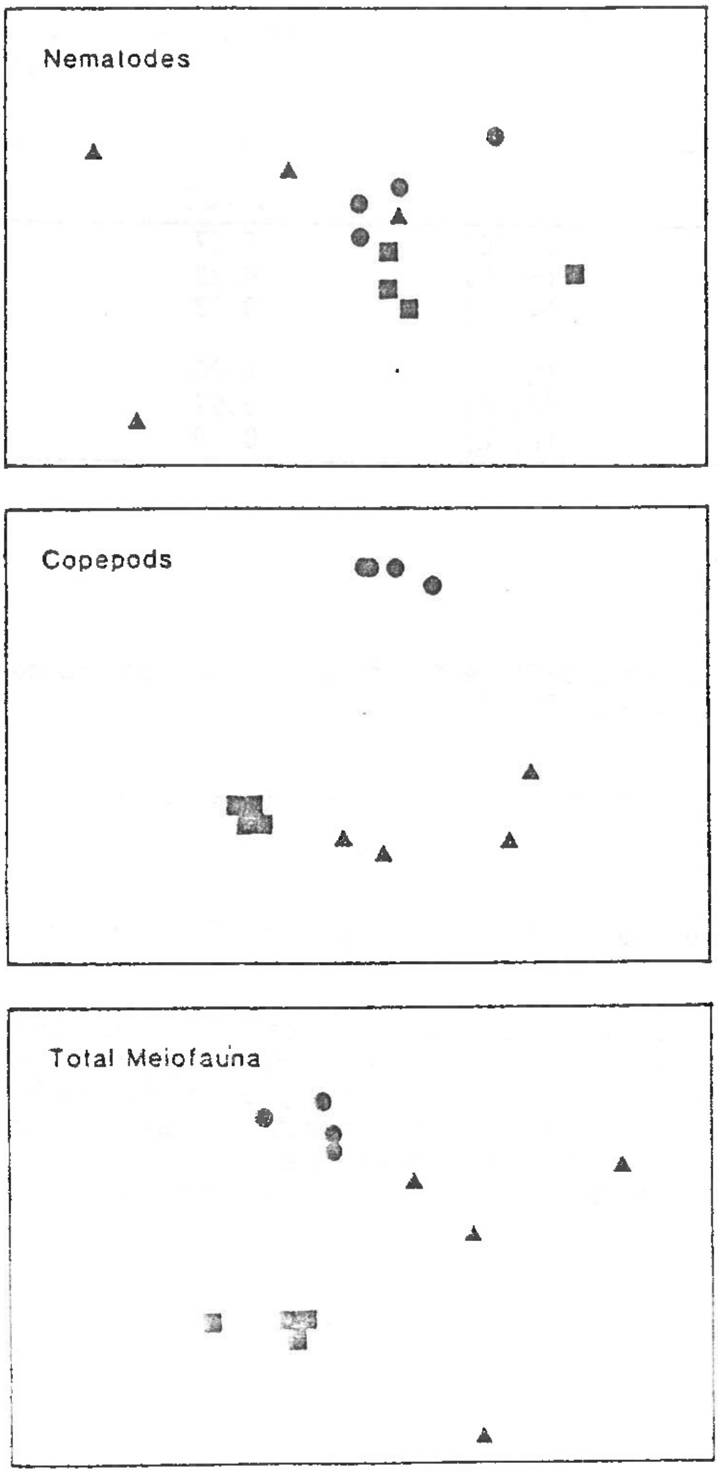


Fig. 12.7 MDS of double square root transformed abundances of nematodes, copepods and total meiofauna (nematodes + copepods). Circles = control, squares = low dose, triangles = high dose

Table 12.5

Values of the R statistic from the ANOSIM test, in pairwise comparisons between treatments, together with significance levels. C = control, L = low dose, H = high dose.

| | TREATMENT | STATISTIC VALUE | % SIG LEVEL |
|---|---|---|---|
| Nematodes | (L, C) | 0.27 | 2.86 |
| | (H, C) | 0.22 | 5.71 |
| | (H, L) | 0.28 | 8.57 |
| Copepods | (L, C) | 1.00 | 2.86 |
| | (H, C) | 0.97 | 2.86 |
| | (H, L) | 0.59 | 2.86 |

NEMATODES: Only differences between low dose and control treatments are significant at the 5% level.

COPEPODS: Differences between all treatments significant at the 5% level.

Note higher values of the R statistic for copepods in all cases.

CONCLUSIONS: Univariate and graphical/distributional techniques show lowered diversity with increasing dose for copepods, but no effect on nematodes. Multivariate techniques clearly discriminate between treatments for copepods, and still have some discriminating power for nematodes. Changes in nematode community may not have been detectable because of great variability in abundance of nematodes in the high dose boxes.

# LECTURE 13

## DATA REQUIREMENTS FOR BIOLOGICAL EFFECTS STUDIES: WHICH COMPONENTS AND ATTRIBUTES OF THE BIOTA TO EXAMINE

COMPONENTS:
- Pelagos
  - plankton
  - fish
- Benthos
  - soft-bottom
    - macrobenthos
    - meiobenthos
    - (microbenthos)
  - hard-bottom
    - epifauna
    - motile fauna
      - macrofauna
      - meiofauna

ATTRIBUTES:
- Abundance
  - species
  - higher taxa
- Biomass
  - species
  - higher taxa
- (Production)

<u>PLANKTON</u>

ADVANTAGES:

- Integrate ecological conditions over areas; useful in monitoring more global changes.
- Taxonomy moderately easy.

DISADVANTAGES:

- Not useful for monitoring local effects, due to mobility.

Example: Continuous Plankton Recorder Survey of NE Atlantic.

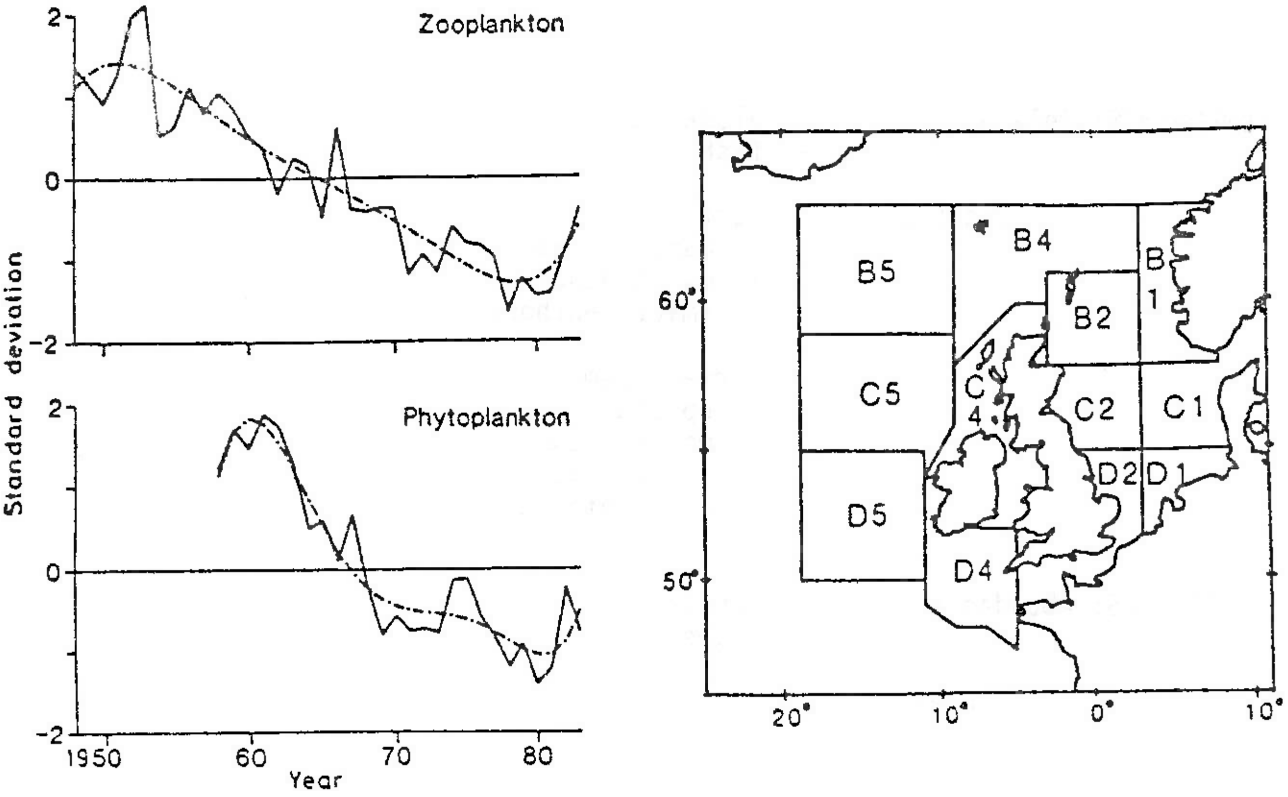


Fig. 13.1 First principal components for zooplankton and phytoplankton (left) in each of the 12 areas shown in the chart (right). Graphs scaled to zero mean and unit variance

## FISH

ADVANTAGES:

- Again more useful for general rather than local effects, but demersal spp. may have site-fidelity
- Taxonomy easy (at least in Europe)
- Of immediate commercial/public interest

DISADVANTAGES:

- Strictly quantitative sampling difficult
- Uncertainty about site-fidelity

Example: Effects of mining activity on coral-reef fish communities in the Maldives.

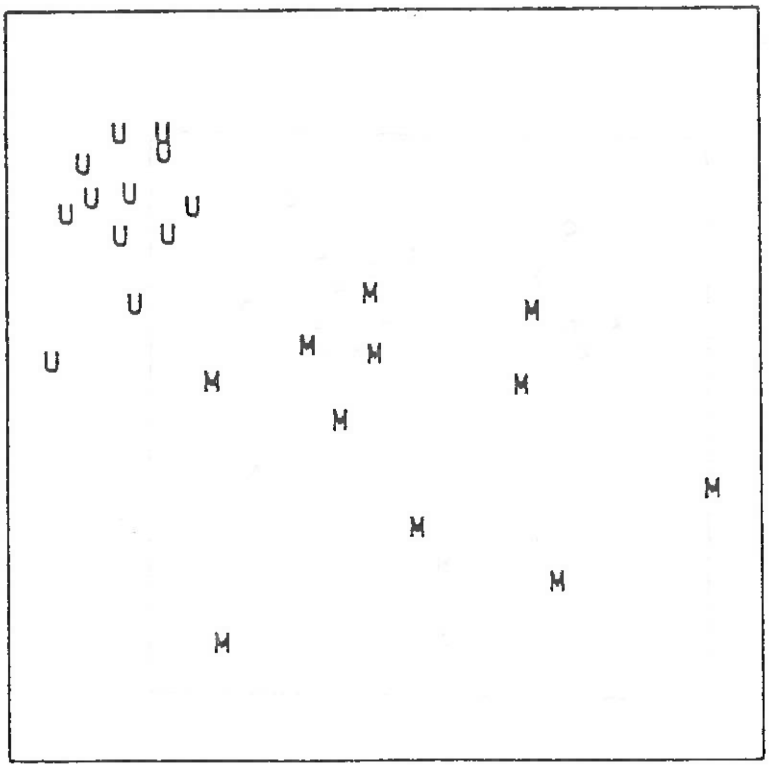


Fig. 13.2 MDS ordination of fish species abundance data from mined (M) and un-mined (U) reef-tops

MACROBENTHOS

ADVANTAGES:

- Non-mobile, therefore useful for local effects
- Taxonomy relatively easy
- Quantitative sampling easy
- Extensive research literature on community effects

DISADVANTAGES:

- Sampling requires relatively large ships
- Sample-processing at sea labour-intensive

- Response time relatively slow (long generation time)
- Unsuitable for causality experiments (slow response time, planktonic larvae).

Example: Amoco Cadiz oil-spill in the Bay of Morlaix.

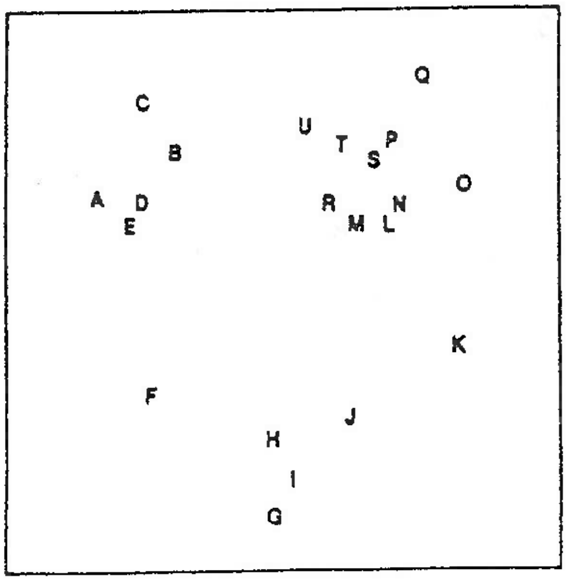


Fig. 13.3 MDS for macrobenthos at station "Pierre Noire". Sampling months are A:4/77, B:8/77, C:9/77, D:12/77, E:2/78, F:4/78, G:8/78, H:11/78, I:2/79, J:5/79, K:7/79, L:10/79, M:2/80, N:4/80, 0:8/80, P:10/80, Q:1/81, R:4/81, S:8/81, T:11/81, U:2/82. Oil-spill was during 3/78, i.e. between E and F

MEIOBENTHOS

ADVANTAGES:

- Useful for local effects studies
- Quantitative sampling easy from small ships
- Samples need not be processed on ship
- Potentially fast response (short generation time)
- Good for causality experiments (direct benthic development, fast response)

DISADVANTAGES:

- Taxonomy considered difficulty
- Community responses not well known or documented

<u>Example:</u> Effects of soldier crab disturbance on nematode assemblages at Eaglehawk Neck, Tasmania.

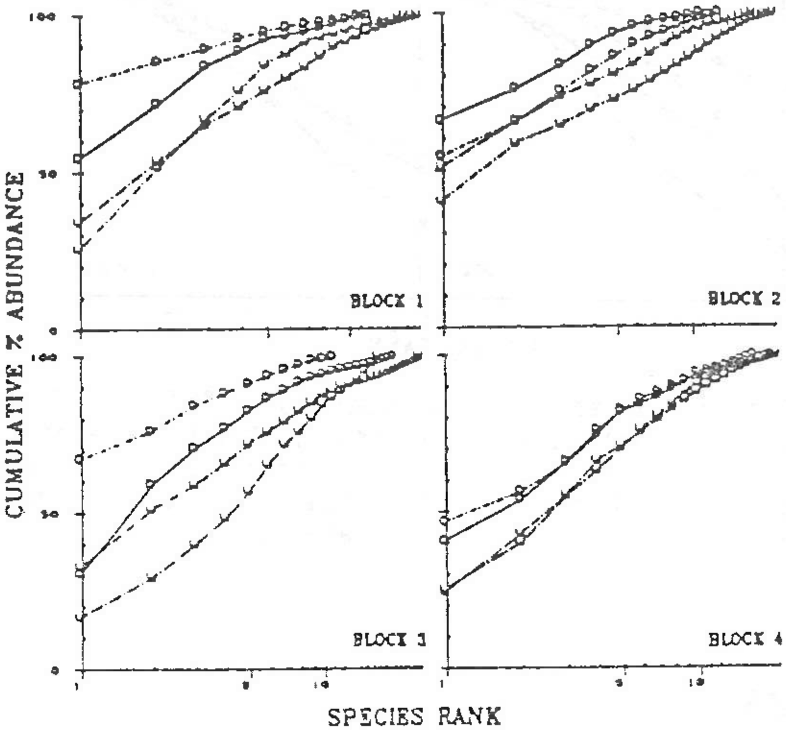


Fig. 13.4 k-dominance curves for disturbed (D) and undisturbed (U) samples in 4 separate sampling blocks

The macrobenthos & meiobenthos may RESPOND DIFFERENTLY to different kinds of perturbation (e.g. physical disturbance, "pollution") so that a comparative study of both may be indicative of the cause.

Example: Hamilton Harbour, Bermuda.

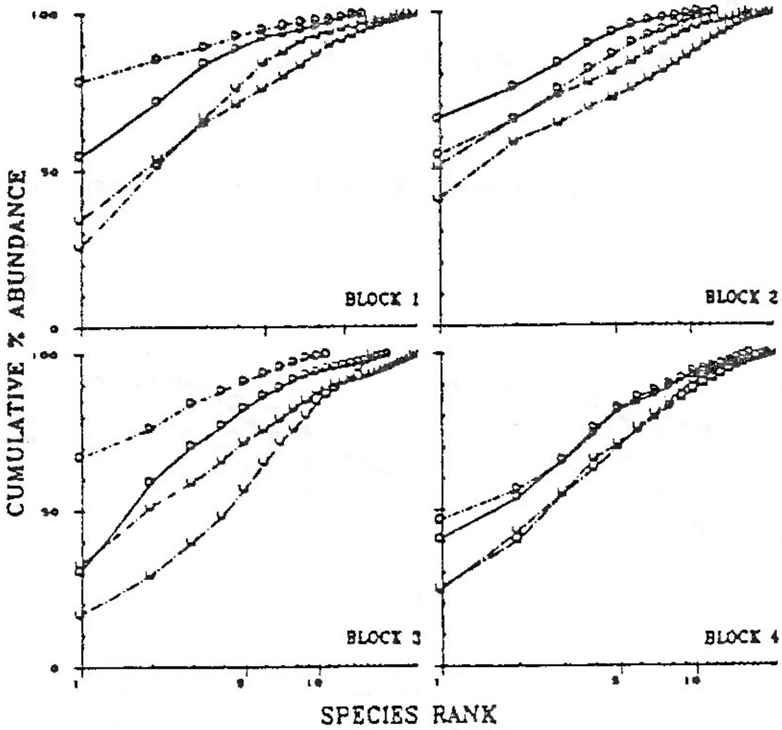


Fig. 13.5 k-dominance curves for macrobenthos (left) and meiobenthic nematodes (right) at six stations in Hamilton Harbour, Bermuda. Elevated macrofauna curves at stations 3 and 4 suggest that physical disturbance is the cause, since the corresponding meiofauna curves at these sites are not similarly affected

HARD-BOTTOM EPIFAUNA

ADVANTAGES:

- Immobile; good for local effects
- Two dimensional nature permits non-destructive (visual) sampling for determination of temporal changes

DISADVANTAGES:

- Remote sampling difficult
- Enumeration of colonial organisms difficult
- Biomass measurements difficult

Example: Effects of the 1982-3 El Niño on Indonesian reef corals.

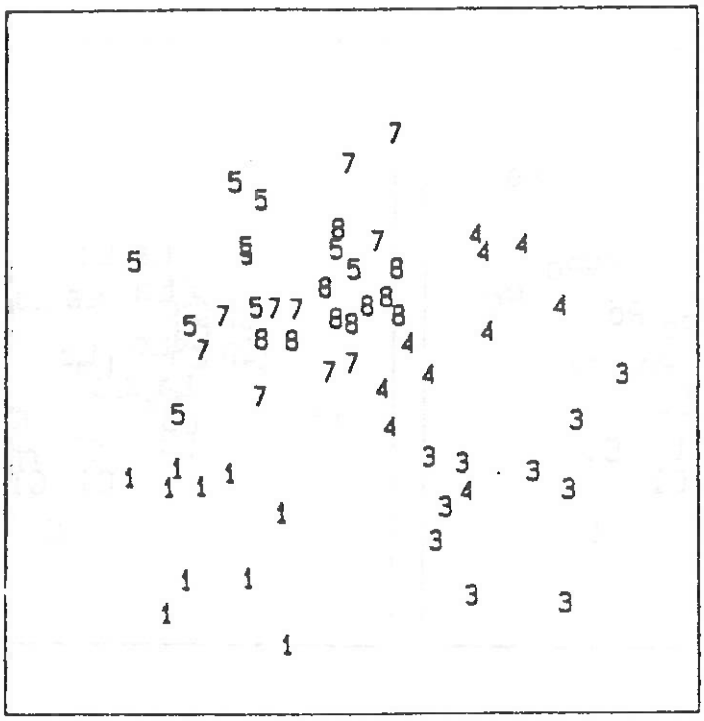


Fig. 13.6 MDS for coral species percentage cover data for South Pari Island. 1=1981, 3=1983 etc

## HARD-BOTTOM MOTILE FAUNA

DISADVANTAGES:

- Remote sampling difficult
- Quantification difficult
- Responses to perturbation not known
- Suitable habitat (e.g. algae) not always available

<u>Example:</u> Macrofauna and meiofauna of replicated intertidal seaweed samples from the Isles of Scilly.

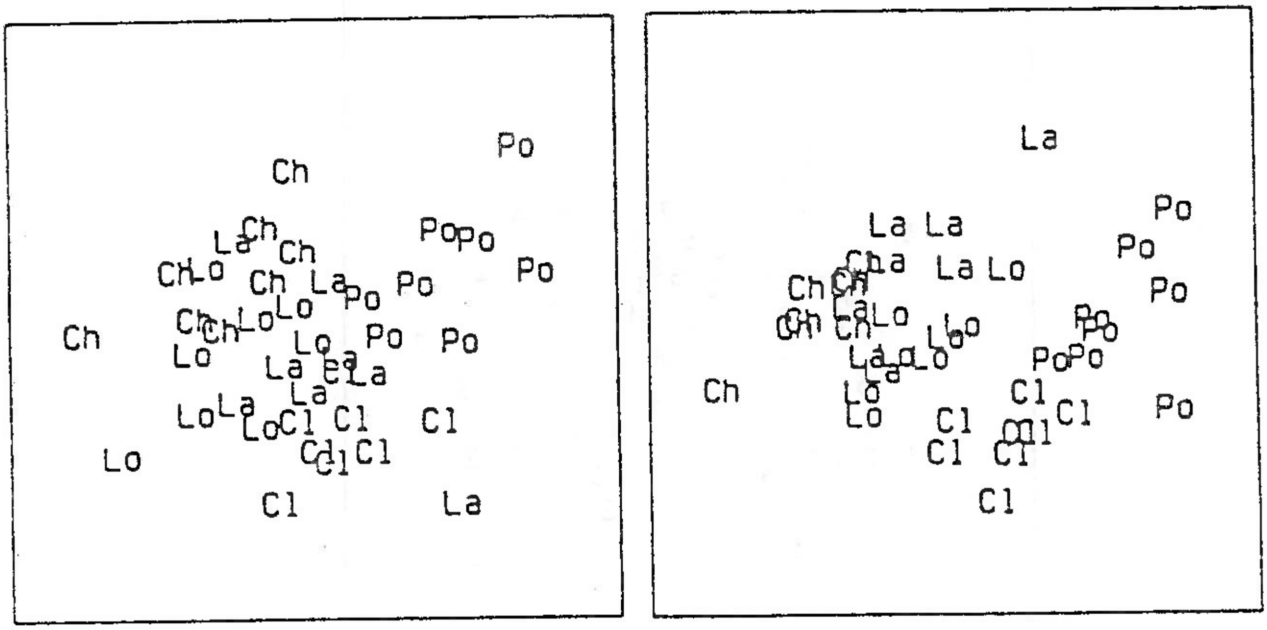


Fig. 13.7 MDS macrobenthos (left) and meiobenthos (right) from different species of seaweeds: Ch=Chondrus, Lo=Lomentaria, La=Laurencia, Cl=Cladophora, Po=Polysiphonia. Note similarity between the two configurations

## <u>ABUNDANCE, BIOMASS OR BOTH?</u>

Abundances are easier to measure, but biomass may be a better reflection of the ecological importance of a species within a community. In practice, multivariate analyses of abundance and biomass data give remarkably similar results, despite the fact that the species mainly responsible for discriminating between stations are different.

Example: Frierfjord macrofauna

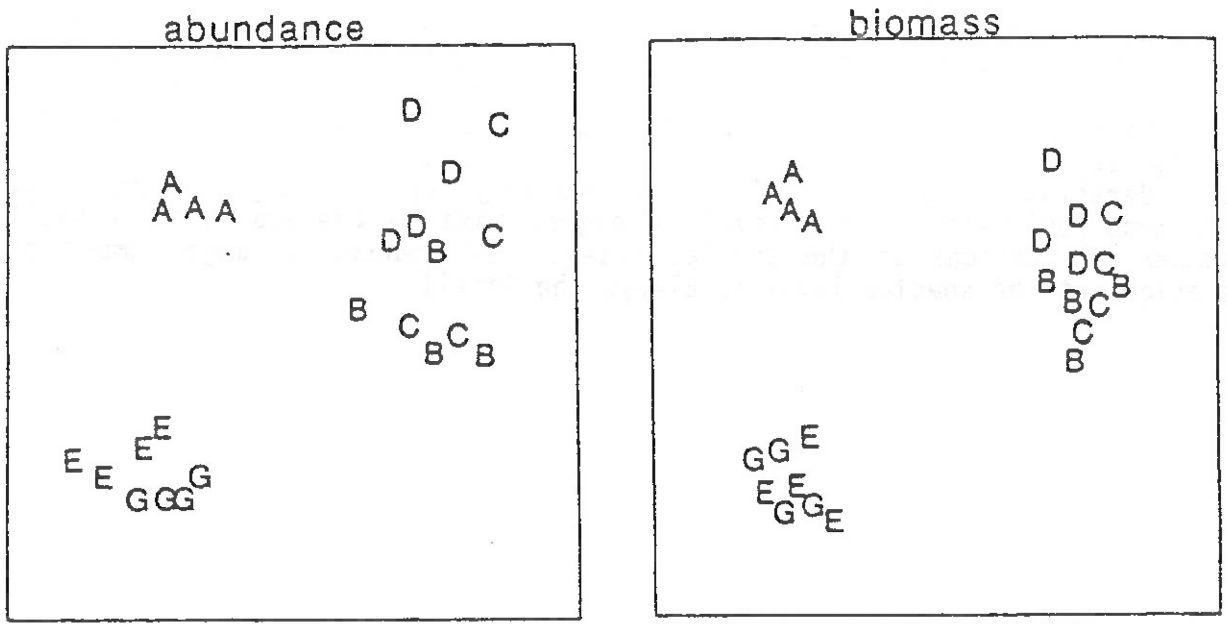


Fig. 13.8 MDS ordinations for macrofauna abundance and biomass. Note the close similarity

Perturbations of various kinds may affect the distribution of numbers of individuals among species differently from the distribution of biomass among species. This is the basis of the 'ABC' (Abundance Biomass Comparison) method for the assessment of disturbance, which was dealt with in Lecture 8.

## SPECIES OR HIGHER TAXA

In a wide variety of pollution-impact studies, it has been found for both graphical-distributional and multivariate analyses that there is surprisingly little loss of information when the species data are aggregated into higher taxa, e.g. genera, families or even phyla. Initial collection of data at the level of higher taxa would result in a considerable saving of time (and cost) in the analysis of samples. This was dealt with in more detail in Lecture 10.

## RECOMMENDATIONS

It is difficult to give firm recommendations as to which components or attributes of the biota should be studied, since this depends on the problem in hand and the expertise and funds available. In general, however, the wider the variety of components and attributes studied, the easier the results will be to interpret. A broad approach at the level of higher taxa is often preferable to a painstakingly detailed analysis of species abundances. If only one component of the fauna is to be studied, then consideration should be given to working up a larger number of stations/replicates at the level of higher taxa in preference to a small number of stations at the species level. Of course, a large number of stations at the species level is always the ideal!

# LECTURE 14

## RELATIVE SENSITIVITIES AND MERITS OF UNIVARIATE, GRAPHICAL/DISTRIBUTIONAL AND MULTIVARIATE TECHNIQUES

Two communities with a completely different taxonomic composition may have identical univariate or graphical/distributional structure, and conversely those comprising the same species may have very different univariate or graphical/distributional structure. Do species dependent and species independent attributes of community structure behave the same or differently in response to environmental changes, and which are the most sensitive? These questions will be addressed by reference to a number of case studies in which a variety of methods of data analysis has been employed.

Example 1: Macrobenthos from Frierfjord/Langesundfjord, Norway (IOC/GEEP Oslo Workshop).

MAP OF SITES: See Fig. 1.1.

UNIVARIATE INDICES:

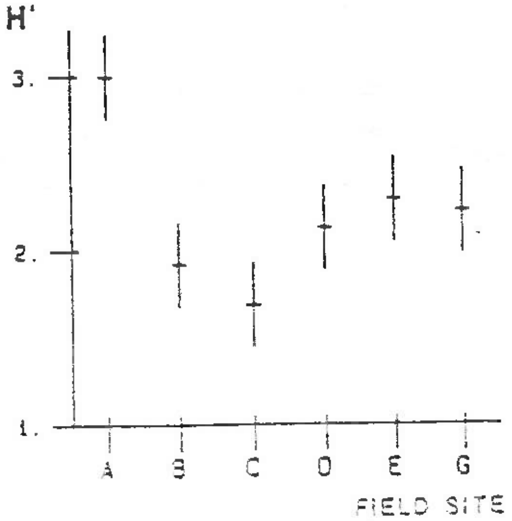


Fig. 14.1 Means (and 95% CIs) for diversity H'

Site A has higher species diversity (H') and site C the lowest: others not significantly different.

GRAPHICAL/DISTRIBUTIONAL PLOTS:

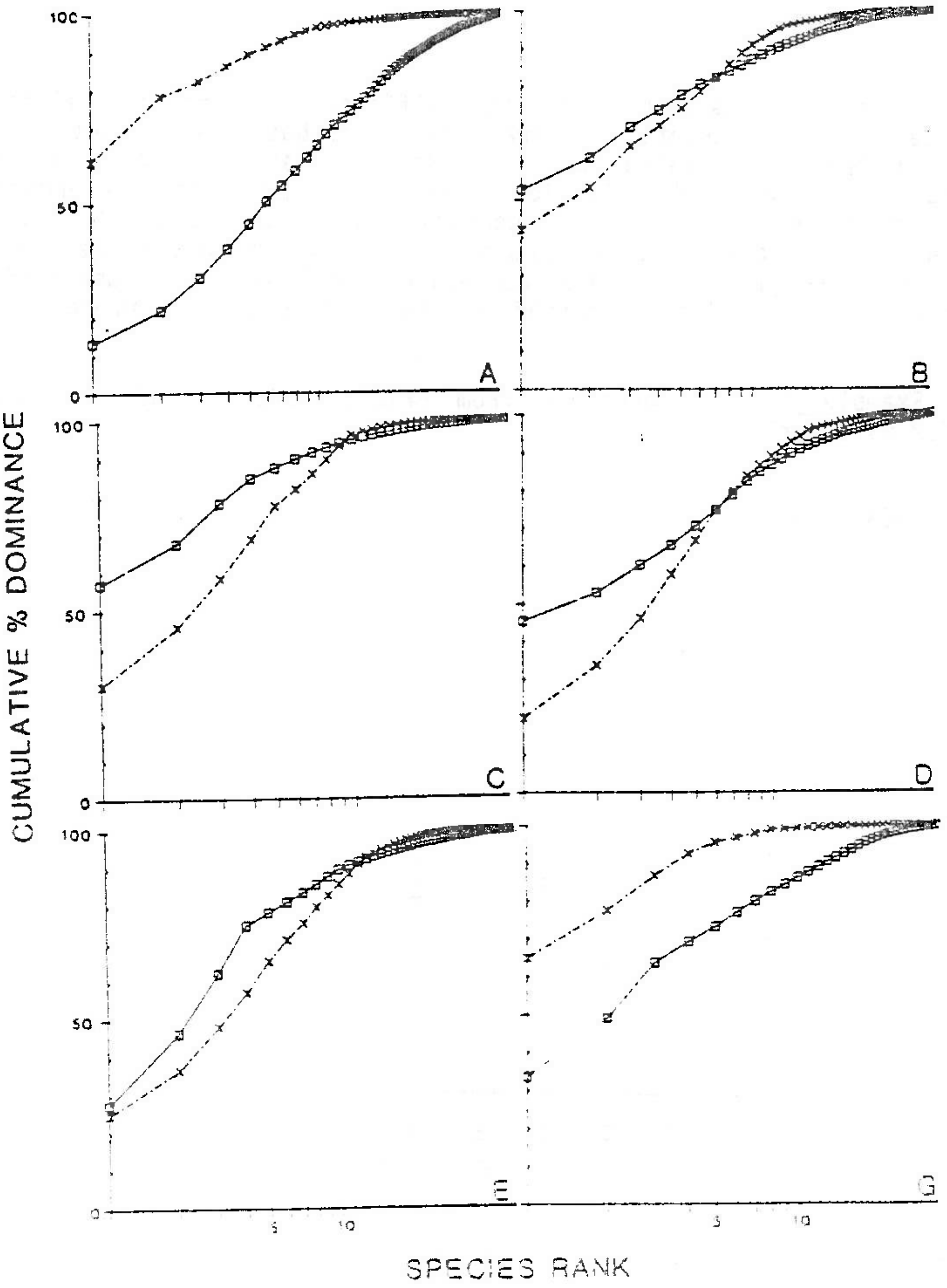


Fig. 14.2 ABC plots based on totals of 4 replicates. Squares = abundance, crosses = biomass

These indicate C, D and E most stressed, B moderately stressed, A and G unstressed. No tests have been done to determine significance of differences.

MULTIVARIATE ANALYSES:

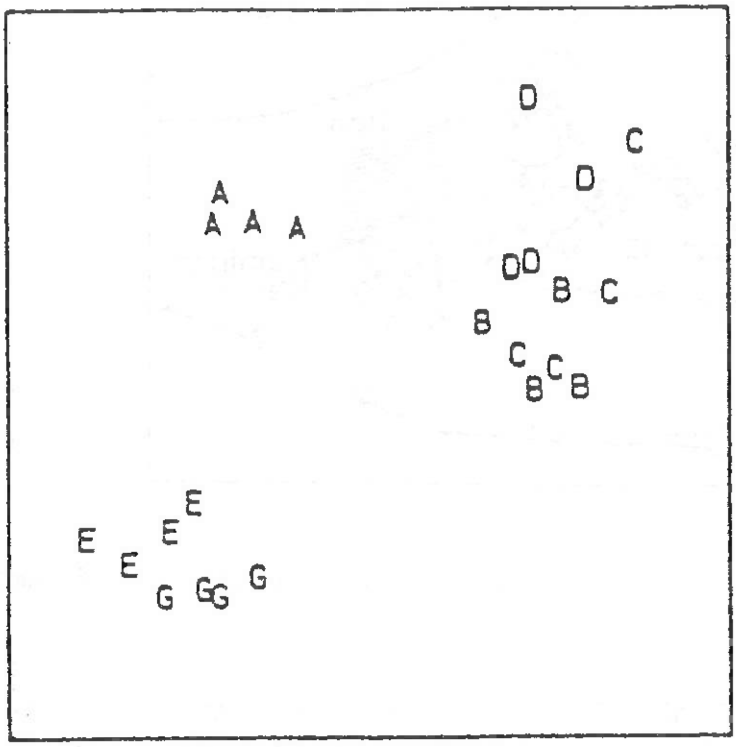


Fig. 14.3 MDS of 4 replicates at each of sites A-E,G (Bray-Curtis similarities on √√-transformed counts)

Stations B,C and D cluster together (ANOSIM separates B from A and C), E and G together (separated with ANOSIM), A on its own. Clusters correlate with water depth rather than measured levels of anthropogenic variables (see Fig. 11.5).

CONCLUSIONS: Multivariate analysis the most sensitive for discriminating stations (only B and C not significantly different). Univariate and graphical distributions conflict with this. For example, E & G have different ABC plots but cluster together; diversity at E is not significantly different from D, but they are the furthest apart on the MDS plots. However, B, C and D all have low diversity and ABC indicates disturbance. Most likely explanation is that these deep-water stations are affected by seasonal anoxia, rather than anthropogenic pollution.

Example 2: Macrobenthos from Hamilton Harbour, Bermuda (IOC/GEEP Bermuda Workshop).

MAP OF SITES:

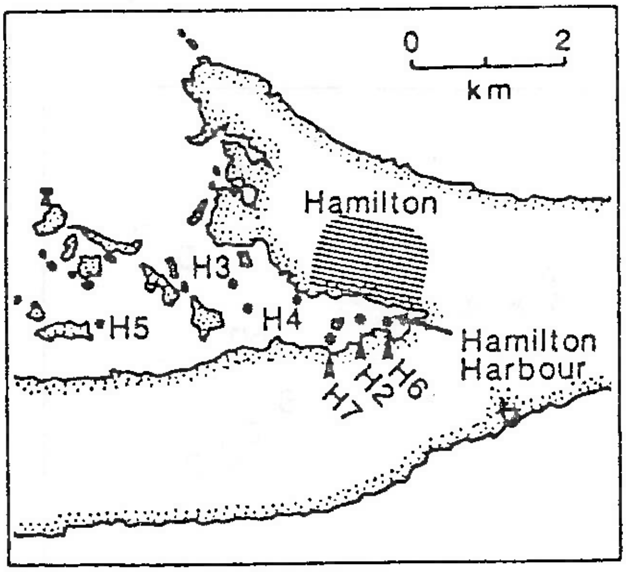


Fig. 14.4 Map of Hamilton Harbour showing locations of 6 sampling stations

UNIVARIATE INDICES: See Fig. 8.1. H5 with highest diversity, H3 and H4 with lowest diversity (significantly below neutral model prediction, see Table on page 128).

GRAPHICAL/DISTRIBUTIONAL PLOTS: ABC curves show H2, H6 and H7 undisturbed, H5 moderately disturbed, H3 and H4 moderately/grossly disturbed (Fig. 14.5).

MULTIVARIATE ANALYSES: On MDS (Fig. 14.6) stations ordered (left to right) 5,4,3,2,7,6. ANOSIM gives all sites significantly different from each other. Superimposing values of environmental variables shows close correlation with metals and TBT, not with water depth, sediment type or hydrocarbons.

CONCLUSIONS: MDS most sensitive in discriminating sites, and relates to pollution levels. Diversity not ordered in the same way. Stations with highest pollution levels not the most 'stressed'.

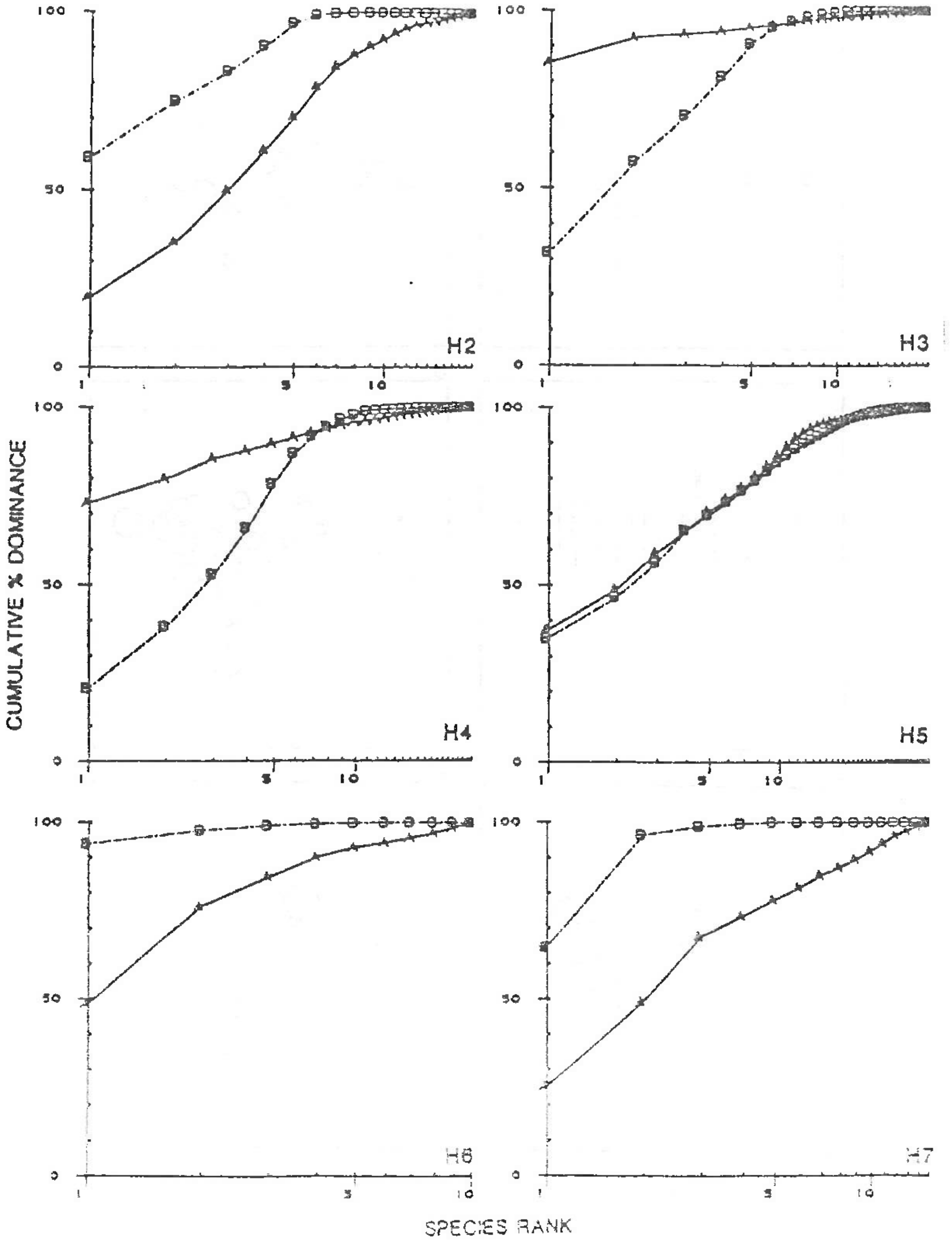


Fig. 14.5 ABC curves for Hamilton Harbour macrobenthos (sum of 4 replicates at each station); A = abundance, B = biomass

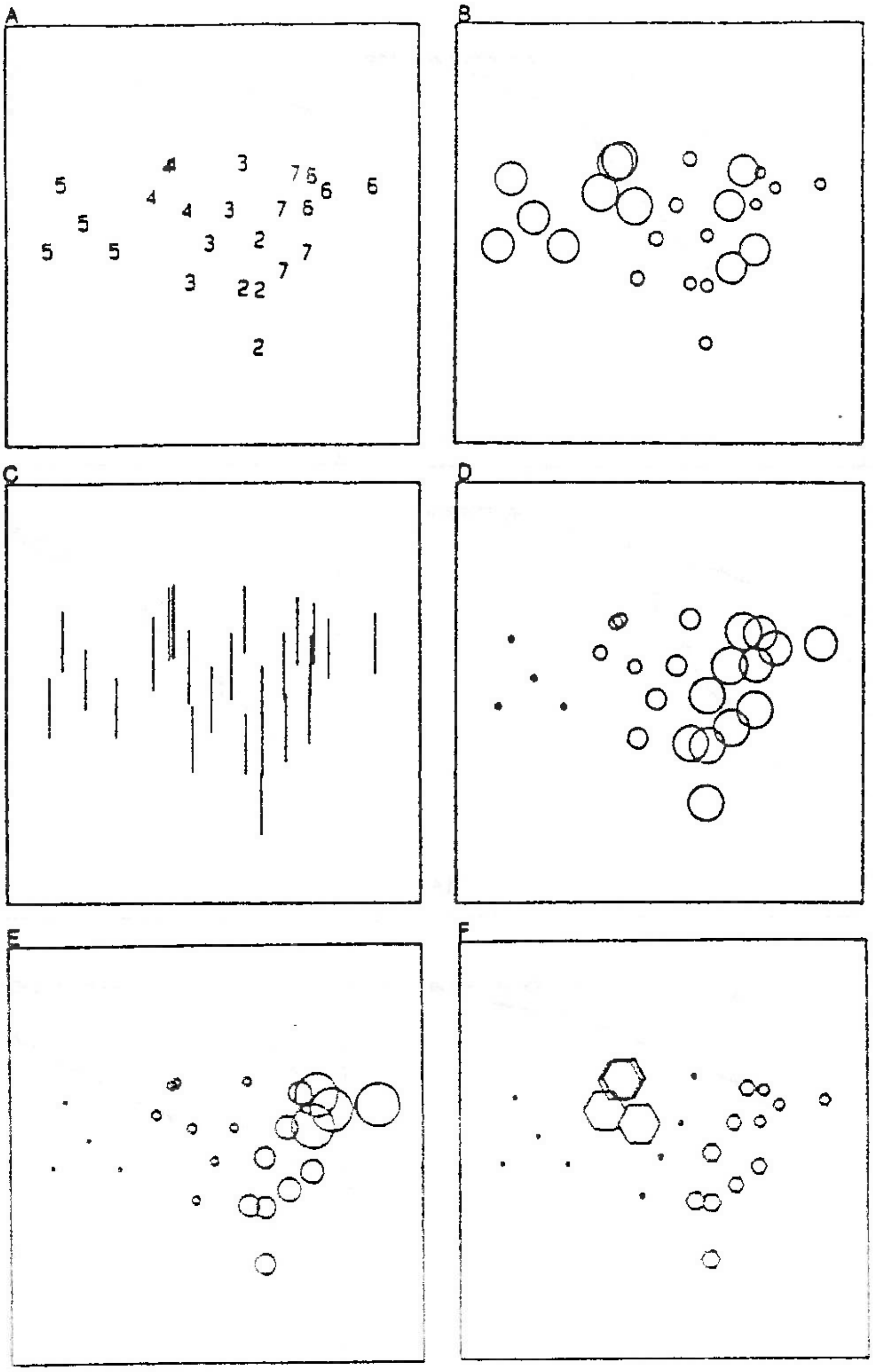


Fig. 14.6 A) 2-D MDS configuration for macrofauna standardised root-transformed abundance. B-F) same configuration with symbols representing values of environmental variables superimposed: B) grain size, C) water depth, D) sediment Pb concentration, E) TBT in water, F) sediment PAH

Example 3: Reef corals at South Tikus Island, Indonesia, before and after 1982-3 El Niño.

MAP OF SITES: Not available. Ten sets of 3 x 10 m transects across reef-flat in each year.

UNIVARIATE INDICES: See Fig. 8.2. Immediate post El Niño decline in number of species and H′, slight recovery in 1984 but no significant change after this. No significant changes in J′.

GRAPHICAL DISTRIBUTIONAL PLOTS: From 1984 onwards, k-dominance curves lie entirely above that of 1981, indicating no apparent recovery. With ANOSIM, few significant differences between years detectable after 1984.

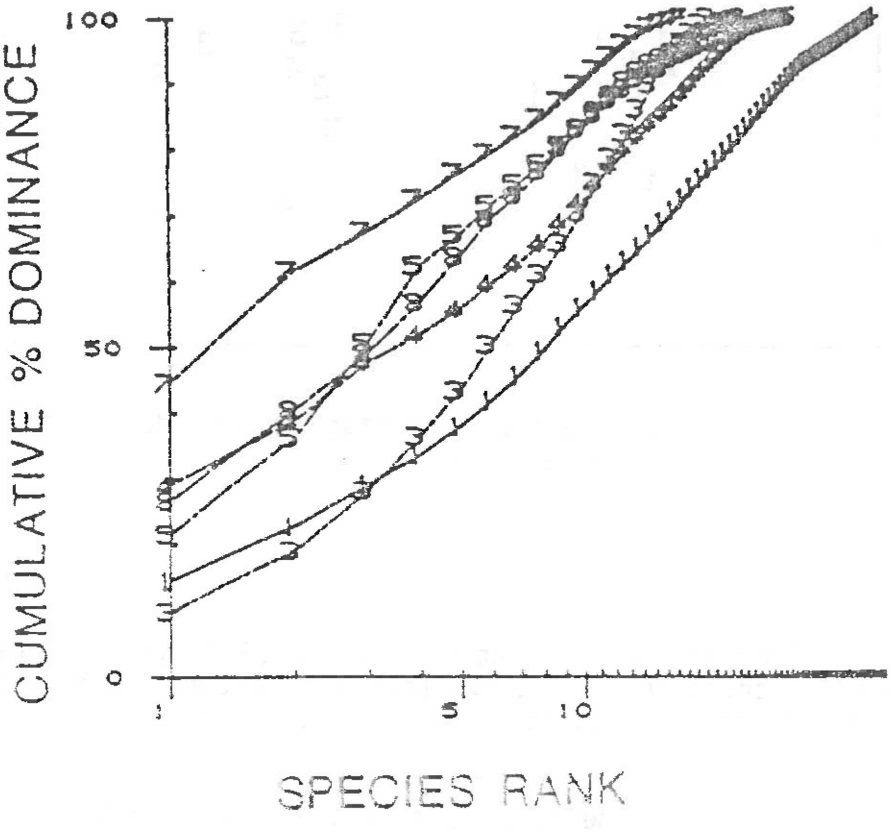


Fig. 14.7 k-dominance curves for totals of all ten replicates in each year. 1=1981, 2=1982 etc

MULTIVARIATE ANALYSES:

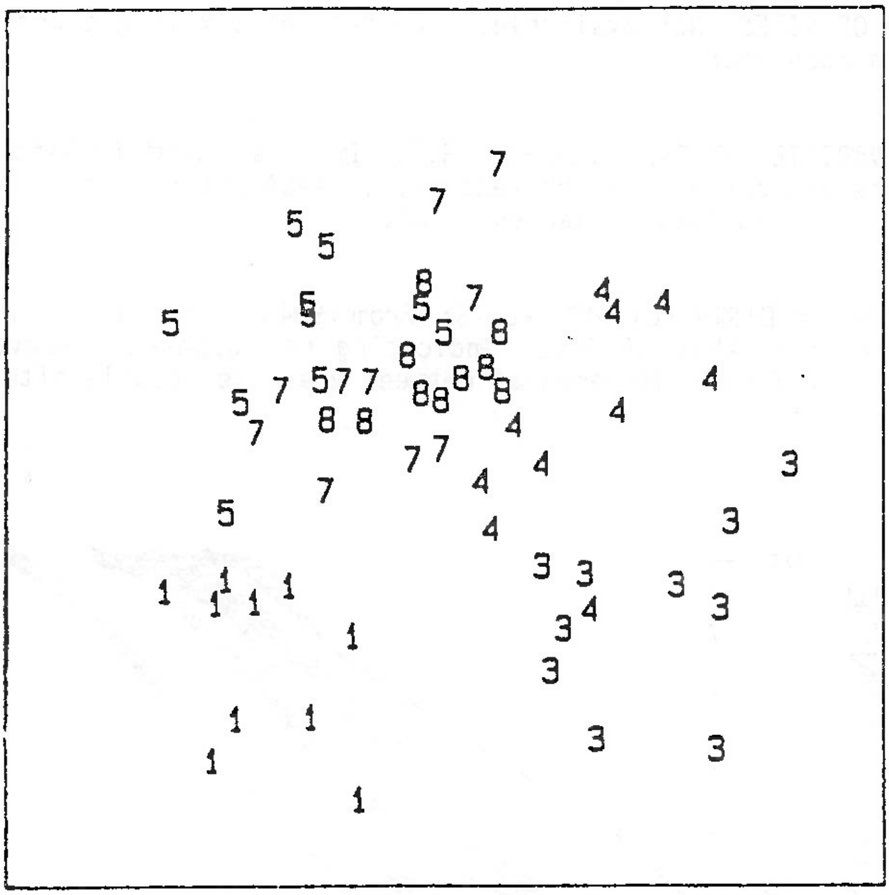


Fig. 14.8 MDS for coral species percentage cover data for South Pari Island. 1=1981, 3=1983 etc

El Niño location shift between 1981 and 1983, with gradual recovery towards the 1981 condition until 1985, then a slight move away again in 1987 and 1988. ANOSIM shows all pairs of years to be significantly different.

CONCLUSIONS: All methods demonstrate the dramatic post El Niño decline in species, though the multivariate techniques were seen to be more sensitive in monitoring the recovery phase in later years.

<u>Example 4:</u> Fish communities from mined and non-mined reef tops in the Maldives.

MAP OF SITES: Not available.

UNIVARIATE INDICES: ANOVA shows no significant effect of mining on H′ or J′.

GRAPHICAL/DISTRIBUTIONAL PLOTS. k-dominance curves for individual replicates given in Fig. 14.9. ANOSIM shows no significant difference between mined and non-mined sites.

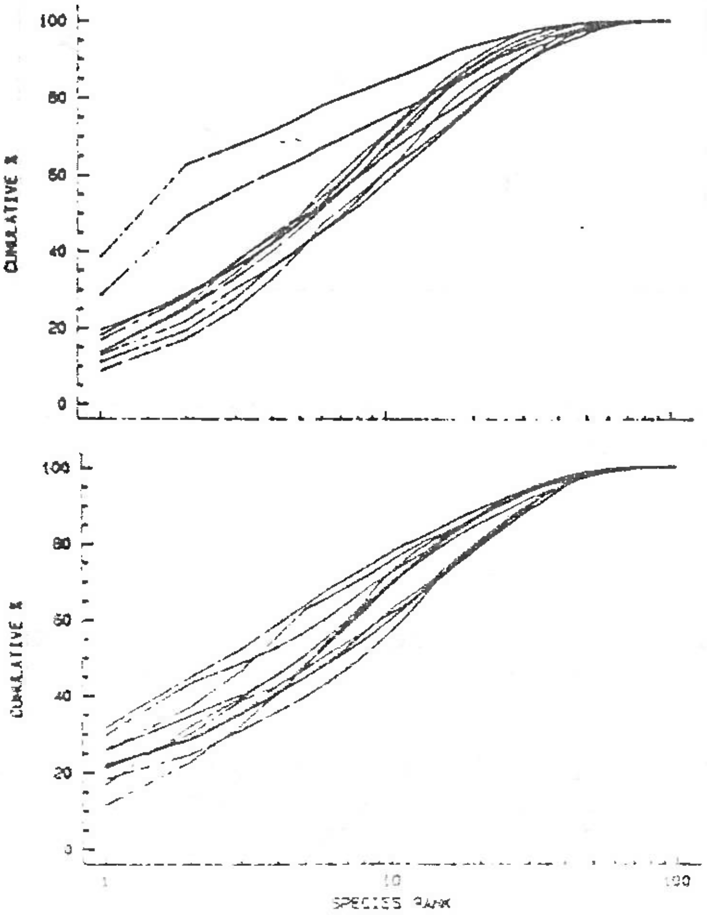


Fig. 14.9 Replicate k-dominance curves for fish communities from mined (top) and non-mined (bottom) reef-tops

MULTIVARIATE ANALYSES:

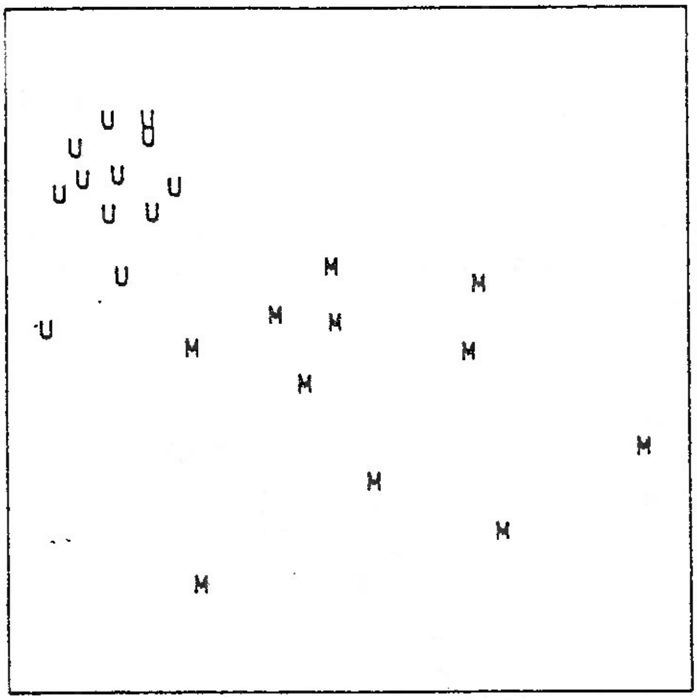


Fig. 14.10 MDS of fish species abundance data from mined (M) and un-mined (U) reef-tops

Clear separation of mined and non-mined sites, which ANOSIM shows to be significant (though test is unnecessary in such a clear-cut case).

CONCLUSIONS: Clear difference in community composition due to mining activity revealed by multivariate methods, but not detected at all by univariate or graphical/distributional techniques.

Example 5: Macro- and meiobenthos from different seaweed species on the Isles of Scilly.

MAP OF SITES:

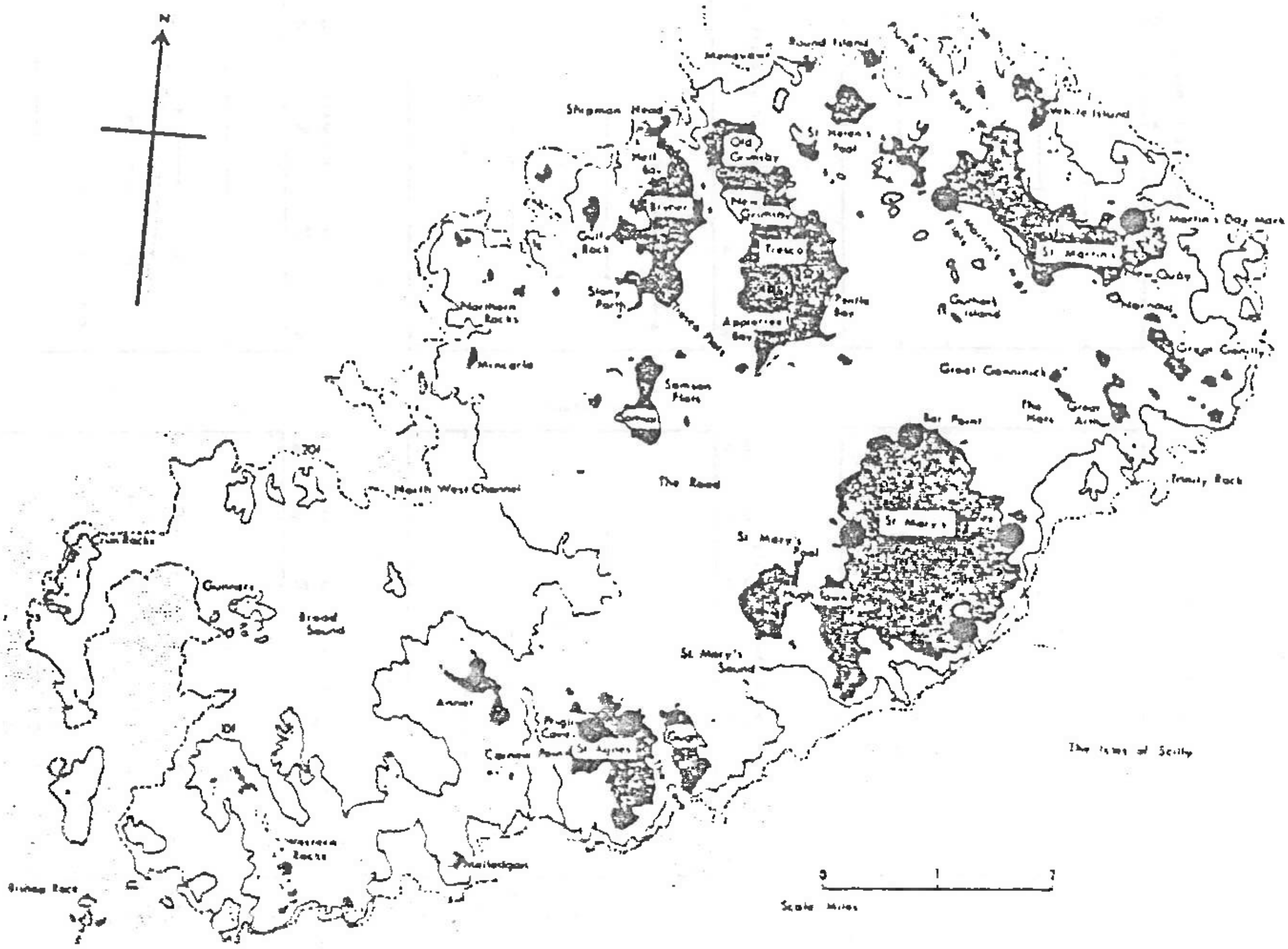


Fig. 14.11 Eight sites on the Isles of Scilly from each of which 5 seaweed species were collected

UNIVARIATE INDICES: Note that meiobenthos and macrobenthos show different trends, and for all indices many pairs of weeds are not significantly different from each other.

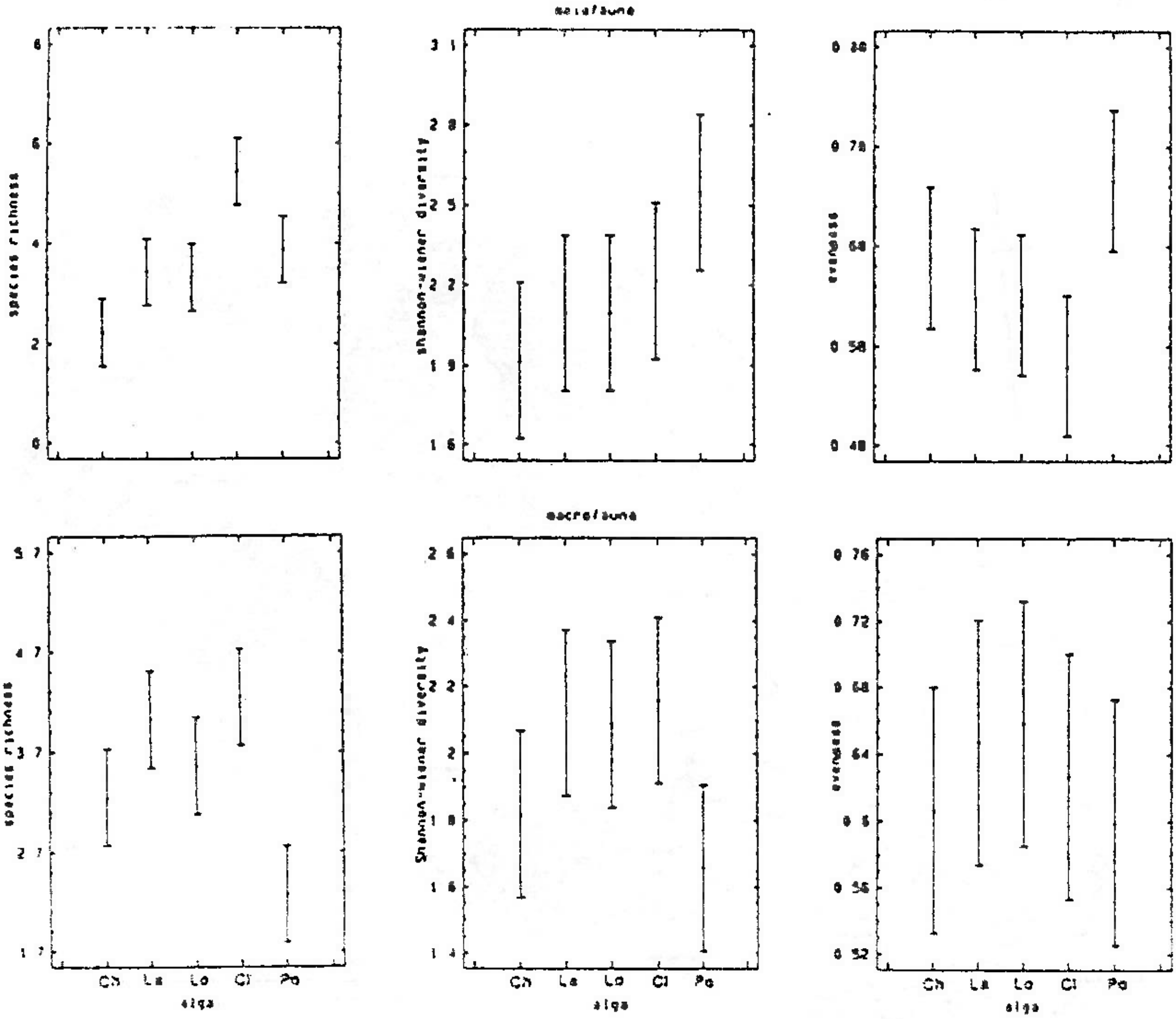


Fig. 14.12 Species richness (left), Shannon diversity (middle) and evenness (right) for meiofauna (top) and macrofauna (bottom), with 95% confidence intervals. Ch = Chondrus, La = Laurencia, Lo = Lomentaria, Cl = Cladophora, Po = Polysiphonia

GRAPHICAL/DISTRIBUTIONAL PLOTS: k-dominance curves for meiofauna show only *Polysiphonia* with a distinctly lower curve than the other species. For macrofauna, curves not clearly distinguishable from each other.

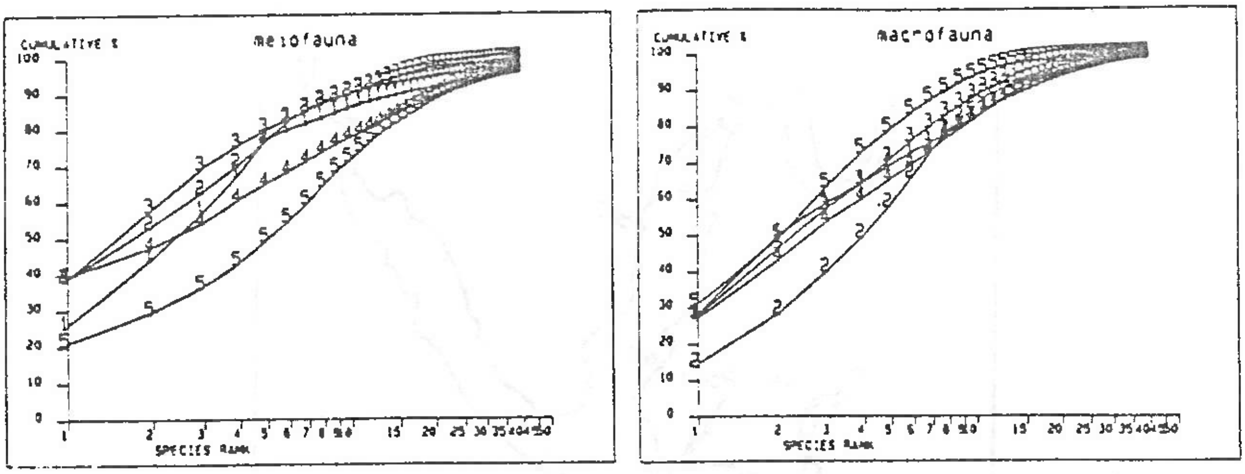


Fig. 14.13 k-dominance curves for meiofauna (left) and macrofauna (right). 1 = Chondrus, 2 = Laurencia, 3 = Lomentaria, 4 = Cladophora, 5 = Polysiphonia

MULTIVARIATE ANALYSES: See Fig. 13.7. Two-way ANOSIM (weed species/sites) shows all weed species significantly different for both meiofauna and macrofauna. Note similarity of macrofauna and meiofauna configurations.

CONCLUSIONS: Multivariate methods more sensitive than univariate or graphical/distributional methods for discriminating between weed species. Univariate and graphical/distributional methods give different results for macrobenthos and meiobenthos, whereas for the multivariate methods the results are similar for both.

<u>Example 6:</u> Meiobenthos (nematodes and copepods) from the Tamar estuary, S.W. England (Austen & Warwick, 1989).

MAP OF SITES:

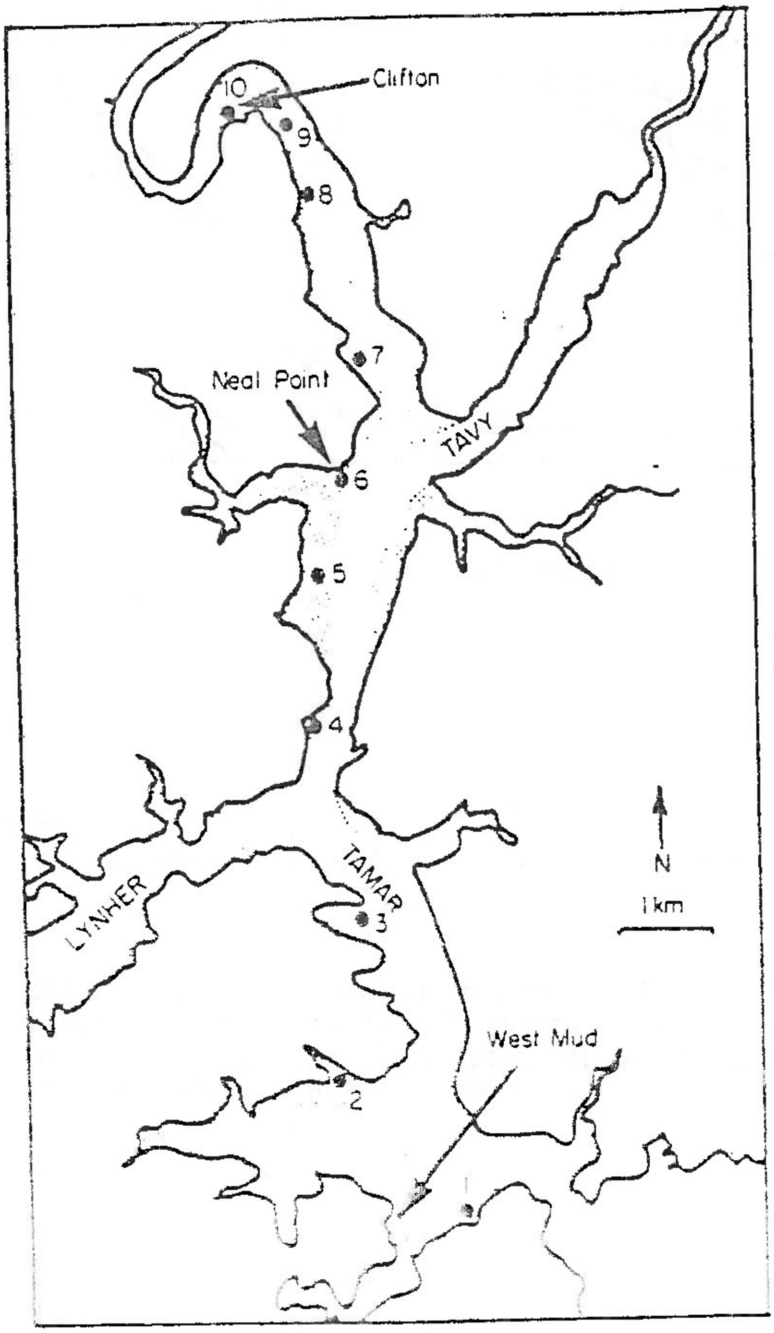


Fig. 14.14 Map of Tamar estuary showing locations of 10 intertidal mud-flat sites

UNIVARIATE INDICES: Not determined.

GRAPHICAL/DISTRIBUTIONAL PLOTS: k-dominance curves for nematodes and copepods do not show similar sequence. For nematodes, sequence does not correspond to the salinity gradient, but for the copepods the agreement with salinity is closer.

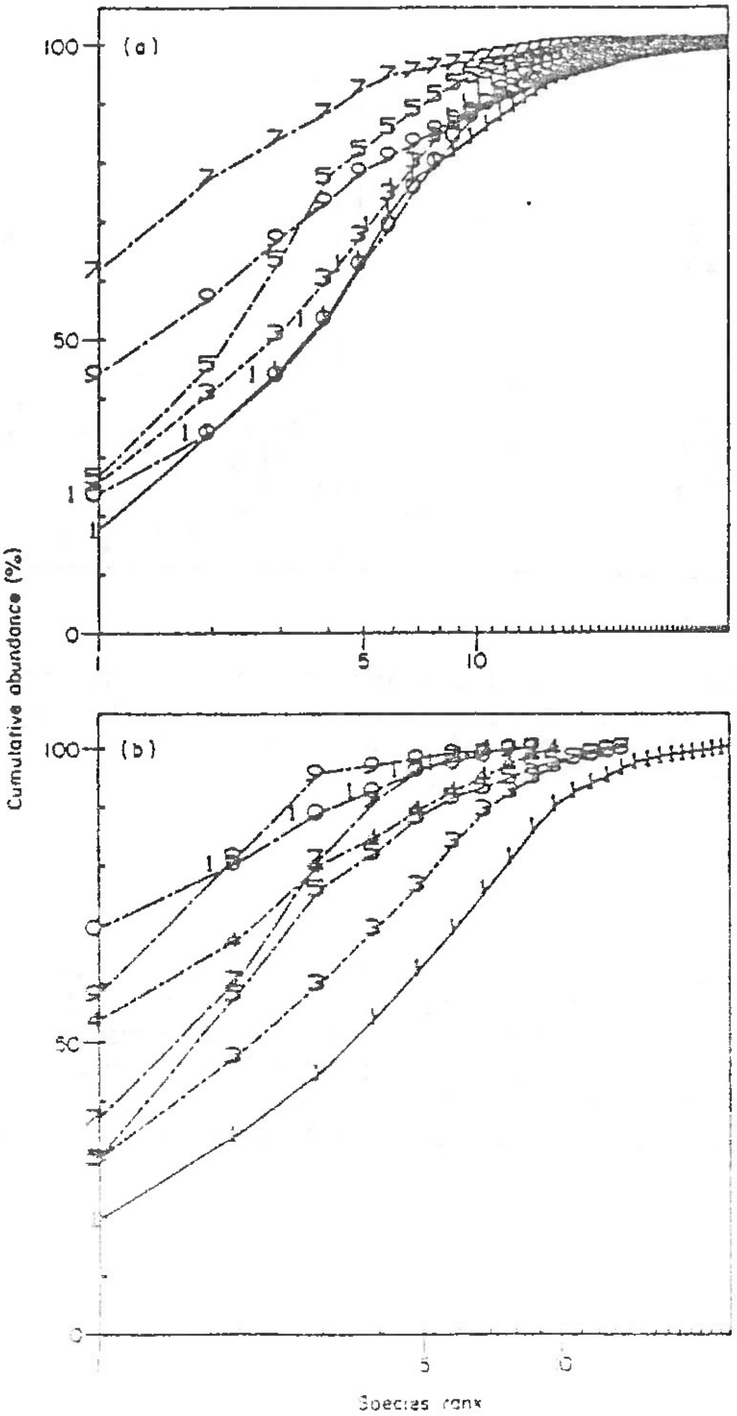


Fig. 14.15 k-dominance curves for amalgamated data from 6 replicate cores for nematodes (top) and copepods (bottom)

MULTIVARIATE ANALYSES: Sequence of sites ordered along the salinity gradient for both nematodes and copepods. ANOSIM shows copepod assemblages significantly different at all pairs of sites, nematodes at all pairs except 6/7 and 8/9.

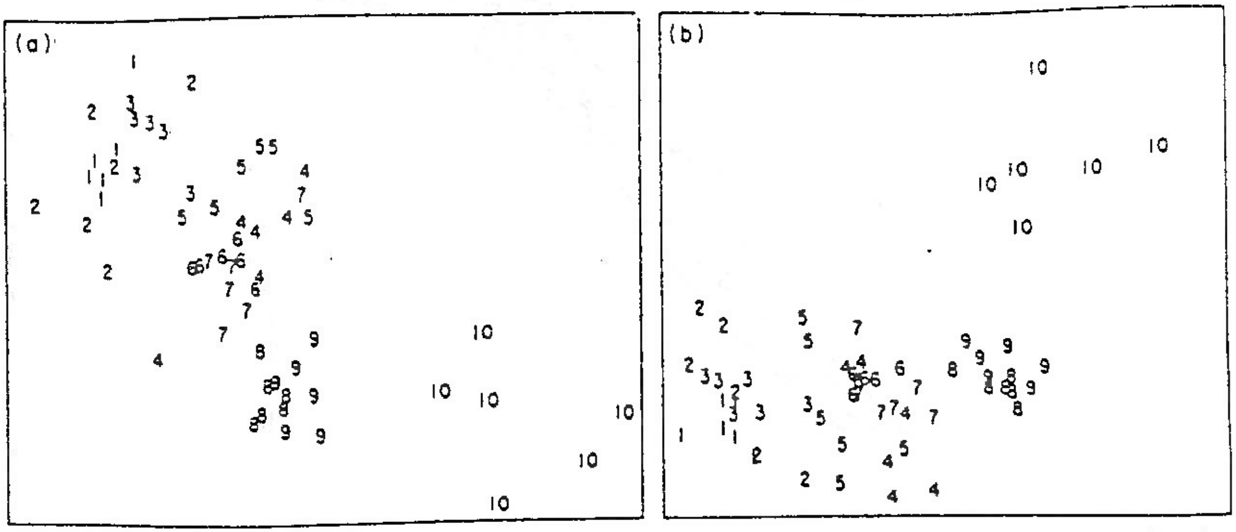


Fig. 14.16 MDS for nematodes (left) and copepods (right) for six replicate cores at each of 10 stations. Note that, allowing for the difference in orientation, the configurations are almost identical

CONCLUSIONS: Multivariate techniques more sensitive in discriminating sites (many sites indistinguishable on basis of k-dominance curves). Multivariate methods give similar patterns for nematodes and copepods; graphical/distributional methods give different patterns for the two taxa. For nematodes, factors other than salinity are more important in determining diversity profiles, but for copepods salinity correlates well with diversity.

Example 7: Meiofauna from Tasmanian sandflat, influenced by burrowing and feeding of soldier crabs.

MAP OF SITES:

UNIVARIATE INDICES:

GRAPHICAL/DISTRIBUTIONAL PLOTS:

See Lecture 12

MULTIVARIATE ANALYSES:

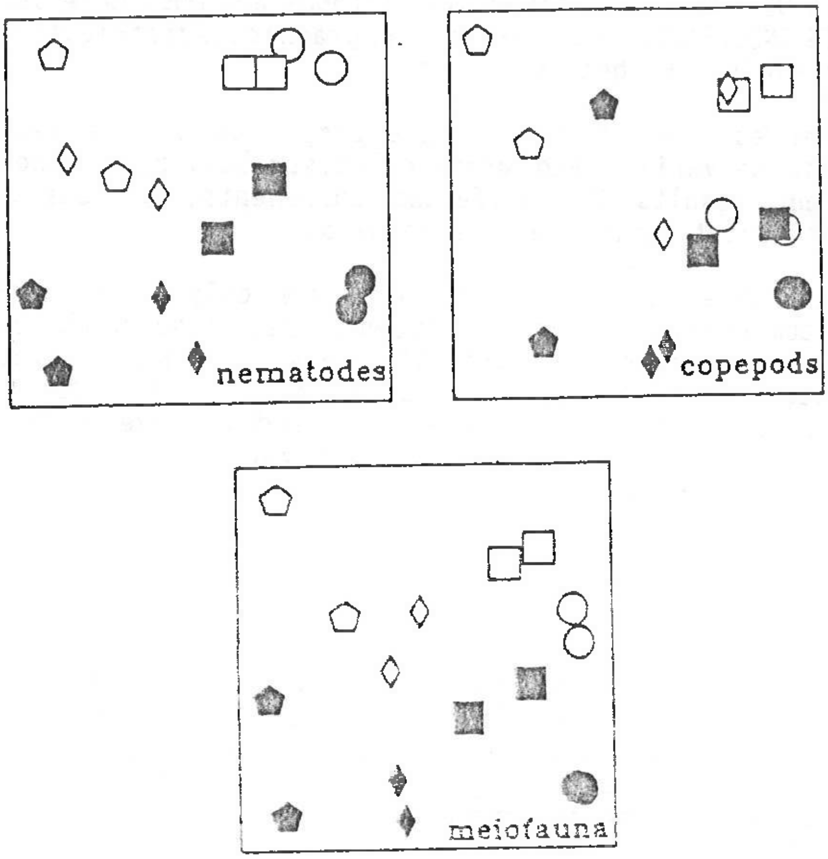


Fig. 14.17 MDS plots for nematode, copepod and 'meiofauna' (nematodes + copepods) abundance. Open symbols = disturbed samples, closed = undisturbed (different shapes denote the flour blocks)

CONCLUSIONS: For nematodes, univariate, graphical/distributional and multivariate methods all distinguish disturbed from undisturbed sites. For copepods only the multivariate methods do. Univariate and graphical/distributional methods indicate different responses for nematodes and copepods;, multivariate methods indicate a similar response.

## GENERAL CONCLUSIONS

Three general conclusions emerge from these examples:

1. Similarity between sites based on their univariate or graphical/distributional properties is usually different from their clustering in multivariate analyses.

2. SPECIES DEPENDENT (multivariate) methods are much more sensitive than SPECIES INDEPENDENT (univariate and graphical/distributional) methods in discriminating between sites.

3. In examples where more than one component of the fauna has been studied, univariate and graphical/distributional methods may give different results for different components, whereas multivariate methods tend to give the same results.

The sensitive multivariate methods are only capable of detecting differences in community composition between sites, although these differences can be correlated with measured levels of stressors such as pollutants. Only the species independent methods of data analysis can be used to determine deleterious (stress) responses. There is a need to develop techniques for determining stress which utilise the full multivariate information contained in a species/sites matrix.

## RECOMMENDATIONS

At present, it is important to apply a wide variety of classes of data analysis, as each will give different information and this will aid interpretation. Sensitive multivariate methods will give an 'early warning' that community changes are occurring, but indications that these changes are deleterious are required by environmental managers, and the less sensitive species independent methods must be used.

## FURTHER READING

For general texts on multivariate methods, the two books by Everitt (1978 and 1980) are useful introductions, and Chatfield and Collins (1980) can be recommended (though requires some knowledge of matrix algebra and statistical inference). A more detailed, but still approachable, exposition of MDS is the monograph by Kruskal and Wish (1978). (None of these texts is written from an ecological viewpoint).

Papers which reflect the approach taken in these lectures include Field *et al.* (1982), Warwick (1986), papers from the GEEP Oslo Workshop Proceedings (Mar.Ecol.Prog.Ser.Vol.46), e.g. Gray *et al.* (1988), Warwick *et al.* (1988), Clarke and Green (1988), and from the GEEP Bermuda Workshop Proceedings (to appear in J.Exp.Mar.Biol.Ecfol. in July 1990), viz. Clarke (1990) and Warwick *et al.* (1990).

## LITERATURE CITED

Austen, M.C. and R.M. Warwick (1989), Comparison of univariate and multivariate aspects of estuarine meiobenthic community structure. *Est.Cstl.Shelf.Sci.*, 29:23-42.

Bayne, B.L., K.R. Clarke, and J.S. Gray (eds.) (1988), Biological effects of pollutants: results of a practical workshop. *Mar.Ecol.Prog.Ser.*, 46 p.

Box, G.E.P. and D.R. Cox (1964), An analysis of transformations. *J.R.Statist.Soc.Ser.B*, 26:211-243.

Bray, J.R. and J.T. Curtis (1957), An ordination of the upland forest communities of Southern Wisconsin. *Ecol.Monogr.*, 27:325-349.

Buchanana, J.B. and R.M. Warwick (1974), An estimate of benthic macrofauna production in the offshore mud of the Northumberland coast. *J.Mar.Biol.Ass.U.K.*, 54:197-222.

Chatfield, C. and A.J. Collins (1980), Introduction to multivariate analysis. London, Chapman and Hall.

Clarke, K.R. (1990), Comparisons of dominance curves. *J.Exp.Mar.Biol.Ecol.*, (in press).

Clarke, K.R. and R.H. Green (1988), Statistical design and analysis for a 'biological effects' study. *Mar.Ecol.Prog.Ser.*, 46:213-226.

Collins, N.R. and R. Williams (1982), Zooplankton communities in the Bristol Channel and Severn Estuary. *Mar.Ecol.Prog.Ser.*, 9:1-11.

Cormack, R.M. (1971), A review of classification. *J.R.Statist.Soc.Ser.A*, 134:321-367.

Dawson-Shepherd, A., R.M. Warwick, K.R. Clarke and B.E. Brown, An analysis of fish community responses to coral mining in the Maldives. *Env.Biol.Fishes* (submitted).

Everitt, B. (1978), Graphical techniques for multivariate data. London, Heinemann.

Everitt, B. (1980), Cluster analysis, 2nd edn. London, Heinemann.

Field, J.G., K.R. Clarke and R.M. Warwick (1982), A practical strategy for analysing multispecies distribution patterns. *Mar.Ecol.Prog. Ser.*, 8:37-52.

Gee, J.M., R.M. Warwick, M. Schaanning, J.A. Berge, and W.G. Ambrose Jr. (1985), Effects of organic enrichment on meiofaunal abundance and community structure in sublittoral soft sediments. *J.Exp.Mar.Biol.Ecol.*, 91:247-262.

Gower, J.C. (1966), Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325-328.

Gower, J.C. and G.J.S. Ross (1969), Minimum spanning trees and single linkage cluster analysis. *Appl.Statist.*, 18:54-64.

Gray, J.S. and T.H. Pearson (1982), Objective selection of sensitive species indicative of pollution-induced change in benthic communities. I. Comparative methodology. *Mar.Ecol.Prog.Ser.*, 9:111-119.

Gray, J.S., M. Aschan, M.R. Carr, K.R. Clarke, R.H. Green, T.H. Pearson, R. Rosenberg and R.M. Warwick (1988), Analysis of community attributes of the benthic macrofauna of Frierfjord/Langesundfjord and in a mesocosm experiment. *Mar.Ecol.Prog.Ser.*, 46:151-165.

Gray, J.S., K.R. Clarke, R.M. Warwick and G. Hobbs (1990), Detection of initial effects of pollution on marine benthos: an example from the Ekofisk and Eldfisk oilfields, North Sea. *Mar.Ecol.Prog.Ser.*, (in press).

Heip, C., P.M.J. Herman and K. Soetaert (1988), Data processing, evaluation. and analysis. pp. 197-231 in R.P. Higgins and H. Thiel (eds.), Introduction to the study of meiofauna. Smithsonian Institution, Washington DC.

Hill, M.O. and H.G. Gauch (1980), Detrended correspondence analysis, an improved ordination technique. *Vegetatio*, 42:47-48.

Hockin, D.C. (1982), The effects of sediment particle diameter upon the meiobenthic copepod community of an intertidal beach: a field and laboratory experiment. *J.Anim.Ecol.*, 51:555-572.

Hope, A.C.A. (1968), A simplified Monte Carlo significance test procedure. *J.R.Statist.Soc.Ser.B*, 30:582-598.

Kenkel, N.C. and L. Orloci (1986), Applying metric and nonmetric multidimensional scaling to some ecological studies: some new results. *Ecology*, 67:919-928.

Kruskal, J.B. and M. Wish (1978), Multidimensional scaling. Beverley Hills, California, Sage Publications.

Lambshead, P.J.D., H.M. Platt and K.M. Shaw (1983), The detection of differences among assemblages of marine benthic species based on an assessment of dominance and diversity. *J.Nat.Hist.*, 17:859-874.

Lance, G.N. and W.T. Williams (1967), A general theory of classificatory sorting strategies: 1 Hierarchical Systems. *Comp.J.*, 9:373-380.

McConnaughey, B.H. (1964), The determination and analysis of plankton communities. Marine Research in Indonesia (Spec. No.): 1-25.

Mantel, N. (1967), The detection of disease clustering and a generalized regression approach. *Cancer Res.*, 27:209-220.

Mardia, K.V., J.T. Kent and J.M. Bibby (1979), Multivariate analysis. London, Academic Press.

Pearson, T.H. (1975), The benthic ecology of Loch Linnhe and Loch Eil, a sea-loch system on the west coast of Scotland. IV. Changes in the benthic fauna attributable to organic enrichment. *J.Exp.Mar.Biol.Ecol.*, 20:1-41.

Pearson, T.H., J.S. Gray and P.J. Johannessen (1983), Objective selection of sensitive species indicative of pollution-induced change in benthic communities. 2. Data analyses. *Mar.Ecol.Prog.Ser.*, 12:237-255.

Sanders, H.L. (1968), Marine benthic diversity: a comparative study. *Am.Nat.*, 102:243-282.

Seber, G.A.F. (1984), Multivariate observations. New York, Wiley.

Sneath, P.H.A. and R.R. Sokal (1973), Numerical taxonomy. W.H. Freeman, San Francisco.

Warwick, R.M. (1986), A new method for detecting pollution effects on marine macrobenthic communities. *Mar.Biol.*, 92:557-562.

Warwick, R.M. (1988), The level of taxonomic discrimination required to detect pollution effects on marine benthic communities. *Mar.Pollut.Bull.*, 19:259-268.

Warwick, R.M. and K.R. Clarke, A comparison of methods for analysing changes in benthic community structure. *J.Mar.Biol.Ass.U.K.*, (submitted).

Warwick, R.M. and T.H. Pearson, Ruswahyuni (1987), Detection of pollution effects on marine macrobenthos: further evaluation of the species abundance/biomass method. Mar.Biol., 95:193-200.

Warwick, R.M., M.R. Carr, K.R. Clarke, J.M. Gee and R.H. Green (1988), A mesocosm experiment on the effects of hydrocarbon and copper pollution on a sublittoral soft-sediment meiobenthic community. Mar.Ecol.Prog.Ser., 46:181-191.

Warwick, R.M. and K.R. Clarke, Suharsono. (1990), A statistical analysis of coral community responses to the 1982-1983 El Nino in the Thousand Islands, Indonesia. Coral Reefs 8:171-179.

Warwick, R.M., K.R. Clarke and J.M. Gee (1990), The effect of disturbance by soldier crabs, Mictyris platycheles H. Milne Edwards, on meiobenthic community structure. J.Exp.Mar.Biol.Ecol., 135:19-33.

Warwick, R.M., H.M. Platt, K.R. Clarke, J. Agard and J. Gobin (1990), Analysis of macrobenthic and meiobenthic community structure in relation to pollution and disturbance in Hamilton Harbour, Bermuda. J.Exp.Mar.Biol.Ecol., (in press).

## ACKNOWLEDGEMENTS

We are grateful to Martin Carr, Jane Addy, John Hall, Ralph Bruno, Martin Budge, Mike Ainsworth, Robert Pritchard, Roger Carter, Mike Gee and Mel Austen who in various ways contributed to the programs and analyses featured in this series of lectures.

K.R. Clarke
R.M. Warwick