COMPUTING OPTIMAL STRATIFIED RANDOM SAMPLING IN MARINE ENVIRONMENT

J.-P. Chanut, B. Pelletier and G. Bégin, Université du Québec à Rimouski

INTRODUCTION

The study of the distribution of any variable in marine environment generally begins by a sampling experiment. Usually, the collection of samples follows a procedure, often fixed in advance, that takes into account the general distribution of the water masses for a given region. However, much attention may be devoted to the actual structure of the water column present below the ship at a given experimental station, and about previous informations on its spatial variability.

The most significant information available in marine environment is related to the stratified structure of the vertical water column. This structure can be identified by measuring some environmental variables such as temperature or conductivity of the water. These variables, well known for its conservative properties, can be measured quickly with a high precision and recorded on magnetic tape or disk. The present study suggests a mini-computer program that permits the construction of strata on the basis of the environmental variables just after their recording.

During an oceanographic survey, many other variables of chemical of biological nature must be evaluated for a given area. To do this, water samples are usually collected at fixed depths by means of sampling bottles. In contrast to this procedure, the second objective of the present program is to define an optimal allocation of a sample into the strata identified previously on the basis of the environmental variables.

A more relevant allocation of a sample in the strata will be defined if previous estimations on the variability of the variables of interest have been entered in the program instead of the variability of the environmental variables. A compromise allocation is then provided.

The last operation of the program is to select randomly the sampling depths in agreement with the limits of strata and with the number of sampling units allocated to each stratum. Hence, the application of the entire procedure involved in the program provides an optimal stratified random sample for estimating oceanographic parameters.

Because the marine environment and the experimental location are continuously changing during an oceanographic cruise, it is suggested to use this sampling strategy on board of the ship with the support of a micro-computer. An illustrative example based on experimental data is provided in order to evaluate the program. A specific study is made on the vertical distribution of the oxygen and on the estimation of the associated parameters.

METHODS

Data processing

The environmental variables such as temperature and conductivity of the water, pressure and sometimes dissolved oxygen are first measured from the surface to bottom of the ocean by means of a special probe interfaced with a mini-computer and automatically recorded on magnetic tape or disk. These data can be displayed on a monitor, a deckwriter or a plotter and also, by means of an inter-connection cable or a solf disk, they can be entered in a separated micro-computer for special calculations such as sampling strategies. The first step in data processing is illustrated in Figure 1a.

The program "STRATE"

The various steps, functions, tests, etc. involved in the program "STRATE" are briefly described in the following paragraphs (Figure 1b):

a) *The construction of strata.* The procedure used to define intermediate stratum boundaries is derivated from equations given by Dalenius (1957, p. 170) or by Cochran (1976, p. 129) in the special case of a minimum variance allocation. The weight of each stratum and the variance of the stratification variable in each stratum are then deducted.

b) *Optimal allocation of a sample.* In case of a minimum variance allocation, the allocation of a sample of size n to the strata is made in accordance with the expression given by Dalenius (1957, p. 189).

c) *The optimal number of strata.* In a stratified sampling, the variance decreases generally as the number of strata increases. In the present program, the optimal number of strata is computed for a given gain in precision following similar expressions proposed by Dalenius (1957, p. 192) or by Cochran (1976, p. 133).

d) *Multiparametric stratified sampling and compromise allocation.* When the survey contains several variables of major interest, a compromise allocation of a sample can be found using the expression given by Dalenius (1957, p. 200). The relative efficiency (Dalenius, p. 198) of the compromise sampling for a given variable can also be computed.

The "STRATE" program has been developped in interactive mode first on a mini-computer HP-3000 and later on micro-computers IBM-PC and TI-PC.

RESULTS AND DISCUSSION

The "STRATE" program was performed on data collected in marine environment. These data correspond to the vertical profiles of some environmental variables such as water temperature and conductivity measured at a fixed station (Lat. 63^{o} 43'S; Long. 79^{o}52'E) in the southern part of the Indian Ocean during the "SIBEX" (1) scientific mission (January-February, 1985) on board of the French research ship "Marion-Dufresne". A graphical representation of the data is given in Figure 2.

a) *Delimitation of the strata for the whole vertical profiles.* The first objective of the computation with this kind of data is to identify the structure of the water column in term of stratification from the surface to the bottom (3600 m). This has been done on board the ship by running the "STRATE" program on micro-computer just after the collection of the data. Computations for four strata have been completed on mini-computer at the laboratory to avoid excessive computation time. The results are presented in Table 1 and shown on Figure 2.

Limits of strata and variances of the sample mean are given in cases of 2, 3 and 4 strata for water temperature and conductivity, and also for dissolved oxygen. Four strata are clearly identified and have quite similar limits for the three environmental variables. In agreement with the sampling theory, the total variance, S_T^2, decreases as the number of strata increases regardless the involved variable. The temperature-conductivity diagram (Figure 3), often used in the identification of oceanic water masses, shows the specific characteristics of the water masses involved in each stratum in terms of temperature and conductivity values.

It is of great interest to compare this optimal stratification method with the oceanographic standard methods of identification of the water layers in terms of extremum values. A stratified structure of the water masses encountered in Antarctic Ocean was proposed first by Deacon (1937) in terms of the location of the minimum and maximum values of temperature, conductivity and dissolved oxygen in function of the depth. These extremum values correspond to the kernel of these different water masses. Compared with this stratified structure, the actual limits of strata (for L=4) tend to make a frame to the extremum values and to draw some intermediate lines between these extremum values. The present method gives with more precision the limits of the different water masses. It appears as a complementary method to the classical one, with the advantage to be performed on an automatic and optimal ways, but with the disadvantage to need a computer machine. Some quick approximate methods are yet suggested by several workers.

Instant stratification appears as a relevant experimental procedure in oceanographic studies in following situations, for example: (1) to define, as it is the case in the present works, the optimal and compromise allocations of a sample of size n and its random distribution in a stratified population just before the immersion of the sampling bottles, in order to estimate with more accuracy and precision the parameters of the population; (2) to choose the more appropriate depth for moored current meters and other sensors to measure the chlorophyll, for example, at a given station during few days in relation to the environmental variables (temperature, conductivity and currents); (3) to identify the depth and the thickness of the cold water masses, generally present in the subsurface layer of cold seas, and to track them over a large area of several hundred kilometers, in order to identify their origin and to study their dynamic and heat exchanges with the surrounding water masses; (4) to identify the appropriate level for trawling a net in order to catch some living organisms generally confined in water masses with specific environmental characteristics.

b) *Optimal allocation of a sample into strata.* Attached to the probe that measures the environmental variables, sampling bottles (n=36) were used to collect water samples at fixed depths, as it is usually the case in oceanographic survey. These water samples are used to measure the concentrations of some chemical variables such as nutrients, carbon, nitrogen, dissolved oxygen, heavy metals, etc. or the concentrations of biological variables such as chlorophyll *a*, phaeopigments, suspended particles, etc.

In contrast to this type of sample allocation, the sample of size n=36 has been optimally allocated into the strata by using the STRATE program (Table 1). In the present study, fixed and optimal allocations of samples along the water column are very different (2), especially in the subsurface and bottom layers. In the fixed allocation procedure, 8 samples are allocated to the surface layer (0 to 150 m), 6 in the subsurface or intermediate layers (150 to 1000 m), 6 in the deep layer (1000 to 2500 m) and 16 in the bottom layer (2500 to 3600 m). The reason for taking as much samples in the bottom layer was to evaluate with precision the content of dissolved oxygen and nutrient elements in these waters, in addition to the temperature and conductivity measurements, in order to identify the origin of these specific water masses. But, the low variability of these water masses, specially in terms of oxygen of conductivity (Figure 2) does not justify such a sample allocation. In contrast, the subsurface layer was under-sampled, with respect to its variability. Consequently, the precision in the estimation of parameters will be very different from one layer to another and for the entire water column, compared with those resulting from optimal allocation in a stratified population. The effects of this fixed sampling procedure on the estimations of the parameters of the population will be pointed out in the presentation of the Table 3. It shows that only the optimal allocation, based on an objective criterion such as the variability of the stratification variables, guarantees minimum variance estimations for the environmental parameters.

c) *Optimal number of strata.* For the successive gains in precision such as 20, 15, 10, 5, 4, 3, 2 and 1%, the optimal numbers of strata are 3, 4, 4, 5, 6, 6, 7 and 9 respectively.

d) *Multiparametric stratified sampling and compromise allocation.* In a oceanographic survey, the environmental variables are not the only variables of interest. In the present case, eight variables of chemical or biological origins have been included in the study. Some variables correspond to the nutrient elements and others to phytoplankton. It is generally admitted, as a working hypothesis, that the spatial distribution of these variables depends more or less strongly on the physical structure of the water masses identified by the environmental variables. So, in this part of the program, the environmental variables are considered as the stratification variables.

In general, the chemical and biological variables cannot be measured automatically, in continuous, along the water column and must be evaluated from water samples collected at different depths by means of a set of sampling bottles. In the present survey, the number of sampling bottles, was fixed to 36. At what depths the sampling bottles must be placed in order to provide good estimations of the parameters of the chemical and biological variables? The answer to this question is given by computing the compromise allocation of a sample of size n=36 into the strata delimitated in function of the distribution of the environmental variables. However, previous informations about the variability of the chemical and biological variables of interest are needed to permit the computation of the compromise allocations of the sample in the strata. In the present case, these informations have been extracted from previous oceanographic surveys in this oceanic region (Post-FIBEX (1), 1981; Jacques, 1978, 1982). It is introduced in the STRATE program under the form of an estimated variances matrix.

The compromise allocation of samples and the estimated variances are given in Table 2 in cases of 2, 3 and 4 strata when the temperature acts as the stratification variable. The compromise allocation of the sample for multivariate surveys differs from the optimal allocation obtained for the temperature only, specially in the first strata near the water surface, the limits of strata remaining the sames. Some biological variables such as nutrients elements present a greater variability in the upper water layer than the temperature.

The estimated values of the variance are relatively similar for a given stratification, but decrease remarkably as the number of strata increases. The relative efficiency changes from a case to another but, in the present study, it tends to decrease when the number of strata increases. The increase of variance caused by the use of compromise allocation instead of the separate optimal allocation is modered, specially in case of L=2 strata (Table 2). This reflects the flatness of the optimum (Cochran, 1977) due, in the present study, to the low variability of the variables of interest in the deeper layer. When the number of strata increases, some subpopulations (portions of the vertical profiles) showing high variability are better delimited; the sample is allocated in a more appropriated manner depending on the variability of each new stratum, and the total variance quickly decreases. The optimum is more pronounced. In these conditions, small changes from the optimal design due, for example, to a near-optimum delimitation of strata or to a compromise allocation of a sample for a given variable of interest, cause a rapid increase of the total variance. This is the case for the silicates and, at the lesser range, for the oxygen, amoniac and chlorophyll (Table 2). In the case of oxygen, the loss of efficiency is due any more to the discrepancy between optimal and compromise allocations of the sample (Tables 1 and 2, respectively, for L=4 strata), rather than some sligh differences in the delimitation of strata given ty temperature and oxygen for L=4 strata (Table 1).

e) *Sampling strategies and estimations*. The measurement of the dissolved oxygen in water samples has two objectives: 1) to calibrate the measurements made by the probe; 2) to estimate the parameters of the distribution of this variable along the water column. The same two objectives are reached with the measures of conductivity in the water samples collected at various depths, while estimation only is made on the parameters of the other chemical and biological variables. In the present case, one of the chemical variables, the dissolved oxygen, was measured in two ways: 1) continuously by a probe along the water column, and 2) at fixed sampling depths by a set of water sampling bottles. We will suppose, for the rest of the study, that the whole water column represents the population while the water column of one meter height represents a sampling unit. In these conditions, it will be possible to compare the estimated values of the parameters obtained by means of various sampling plans with the true values in the population.

The parameter values of the dissolved oxygen distribution are given in the Table 3 for the unstratified population corresponding to the whole water column. These parameters are also estimated from 36 water samples. In the case of fixed sampling depths, if the set of the single values is only used in the computation of the parameters, the values seem to be surestimated. This is due to the presence of many observations taken in the surface layer where the values of dissolved oxygen are high. If the difference of depth between two observations is used as a weighing function of the observations, the values (also called the "integrated" values) of the parameters are very similar to those found in the unstratified population. But, in the case of non-random selection of the sampling depths along the water column, sampling-error formulas cannot be applied, and no confidence interval or precision can be given for the results of this kind of samples (Cochran, 1977).

By means of the STRATE program, it has been possible to identify various strata in the population and to set boundaries to the corresponding subpopulations. As it is generally the case in this kind of experiment, only the environmental variables can be used to identify the stratified structure of the water column. In the present study, the temperature is used as the stratification variable, and a good determination of strata is given for L=4 (see Table 1 and Figure 1). Because the dissolved oxygen is one of the chemical variables of interest, the compromise allocation of the 36 sampling units for L=4 strata has been chosen (see Table 2). The selection of the sampling units, e.g. the sampling depths, has been performed in two ways: 1) by a simple random sampling without replacement in each stratum, and 2) by a systematic random sampling in each strata. The systematic sampling plan is purposed in addition to the simple random sampling because (a) it can give better estimation of parameters than a simple random sampling in presence of a linear tendancy in the data; (b) it is more easy to realize during the field experiment; (c) it warrants a quite regular distribution of the sampling units in the population; and (d) consequently, it is better accepted by the oceanographers. The random selection of the first sampling unit for the systematic sampling and of the n_h units (sampling depths) for the simple random sampling in each stratum has been performed by means of a subroutine associated to the STRATE program. These two sampling plans have been simulated on the actual experimental data.

The estimated values of the parameters provided by the two sampling plans are quite similar (see Table 3). Confidence intervals are relatively tight in spite of the little sample sizes. The true values of the parameters are also computed for each stratum. Compared with them, the values given by the random samples are good estimations of the corresponding parameters in the population. Furthermore, we can say on the basis of confidence intervals that the strata are significatively different from one to the others.

As expected, the valuations of the mean in each stratum and for the entire population are more precise when the population is stratified than in absence of stratification.

CONCLUSION

The STRATE program appears to be useful in the identification of strata in marine environment. It is adapted to provide a minimum variance allocation of a sample for estimating one specific parameter, and also to provide a compromise allocation in the multiparametric case. Some improvements can be made on the choice of the weights associated to the variables of interest. Different approaches proposed for allocating a stratified sample to estimate several population means or totals have been reviewed by Bean and Burmeister (1978). Some suggested weighing functions of the parameters of interest could be considered and applied in the experimental context of oceanographic studies.

The algorithm used in the actual program for determining the best set of strata boundaries is derivated from the basic formula proposed by Dalenius (1957). It appears as a big consumer of computation time. A more efficient algorithm such that proposed by Norland (1983) for selecting strata boundaries will be used with more convenience than the actual one.

ACKNOWLEDGMENTS

Thanks are due to Dr. Seymour Sudman, from the University of Illinois at Urbana-Champaign, for his very pertinent and constructive criticism of the paper. Thanks are also due to Dr. M.I. El-Sabh, from the Université du Québec à Rimouski, for reviewing the manuscript. One of the authors (J.P. Chanut) is grateful towards Mr. A. Pieri, chief administrator of the Terres Australes et Antarctiques Françaises for its invitation to participate at this international scientific mission. We thank Céline Lavoie for typing the manuscript, and the members of the UQAR Computer Service for their assistance. This study was supported by grants from Fonds Institutionnels de Recherche-UQAR, Rimouski, and from the National Sciences and Engineering Research Council of Canada, Ottawa (grant A-7708) to J.P. Chanut.

FOOTNOTES

(1) FIBEX and SIBEX: First (1981) and Second (1985) International BIOMASS Experiments; BIOMASS: Biological Investigations of Marine Antarctic Systems and Stocks; ANTIPROD: Antarctic PRODuction.
(2) The inverse statement was made, by mistake, in the previous version of this work, and pointed out by the discussant.

REFERENCES

BEAN, J.A. and BURMEISTER, L.F. (1978). A review of optimal sample allocation for multipurpose surveys. Biom. J., 20: 3-14.

COCHRAN, W.G. (1977). Sampling Techniques. John Wiley & Sons, New York, 3rd edition; 413 p.

DALENIUS, T. (1957). Sampling in Sweden. Almquist & Wiksell, Stockholm; 250 p.

DEACON, G.E.R. (1937). The hydrology of the Southern Ocean. Discovery Reports, vol. 15, pp. 1-124.

JACQUES, G. (ed.) (1978). Campagne ANTIPROD I; Marion-Dufresne 1-28 mars 1977. Publications du CNEXO, Résultats des Campagnes à la Mer, no. 16, 149 p.

JACQUES, G. (ed.) (1982). Campagne océanographique MD 21/ANTIPROD II (mars 1980); Production pélagique dans le secteur Antarctique de l'Océan Indien. Publications du CNFRA, no. 53, 141 p.

NORLAND, R.E. Jr. (1983). An efficient algorithm for determining strata boundaries for discrete populations using Ekman's method. Proceedings of the Statistical Computing Section of the Annual Meeting of the ASA, pp. 174-176.

POST-FIBEX (1981). Post-FIBEX Data Interpretation Workshop, Hamburg, FRG, September 21-October 9, 1981; BIOMASS Report Series, no. 20; SCAR Secretariat, Cambridge, U.K.

Figure 1. Schematic representation of the collection and analysis of experimental data. a) Systems of measurement, registering and compilation of the environmental data; b) Flow chart of the calculations involved in the STRATE program.

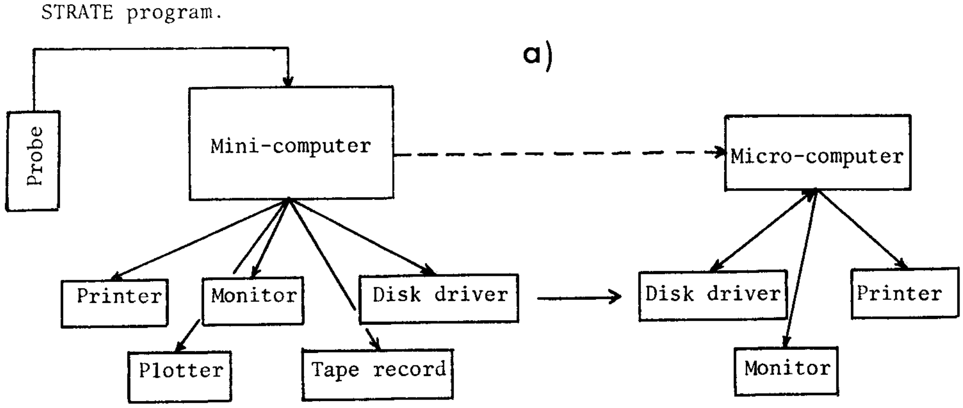

TABLE 1. Stratification of the entire water column (1 to 3600 m depth) and optimal allocation of 36 sampling units along the vertical profiles of some environmental variables measured in the Antarctic Ocean. Total variances (S_T^2), limit depths (m) of strata (italic numbers), and numbers of sampling units in each stratum (underlined numbers).

Number of strata	Stratification variable	Temperature (°C)	Conductivity (mmho)	Oxygen (µA)
L=1	y_{min1}	T_{min1} = −1.481 at 50 m	C_{min1} = 27.129 at 50 m	O_{min1} = 0.176 at 310 m
	y_{max1}	T_{max} = +1.880 at 330 m	C_{max} = 30.642 at 590 m	O_{max1} = 0.351 at 20 m
	y_{min2}	T_{min2} = −0.114 at 3600 m	C_{min2} = 30.115 at 2900 m	O_{min2} = 0.125 at 3600 m
	y_{max2}			O_{max2} = 0.184 at 750 m
	S_T^2	$(0.68)^2$	$(0.40)^2$	$(0.03)^2$
L=2	I	25	4	27
		1800	*110*	*1690*
	II	11	32	9
	S_T^2	$(0.39)^2$	$(0.22)^2$	$(0.02)^2$
L=3	I	25	8	5
		1430	*130*	*110*
	II	6	12	14
		2380	*1390*	*1720*
	III	5	16	17
	S_T^2	$(0.31)^2$	$(0.14)^2$	$(0.01)^2$
L=4	I	5	9	7
		110	*130*	*110*
	II	21	14	10
		1240	*1510*	*1390*
	III	5	8	10
		2220	*2590*	*2300*
	IV	5	5	9
	S_T^2	$(0.21)^2$	$(0.10)^2$	$(0.01)^2$

TABLE 2. Compromise allocations (n_{ch}) of sample of size n=36 and estimated variances in cases of L=2, 3 and 4 strata with temperature as the stratification variable. S^2_{min} and S^2_c correspond to the minimum variance allocation and compromise allocation respectively; their ratio, E_x, expressed in percent, for the variable of interest, X, represents the relative efficiency of the compromise allocation.

L	L = 2			L = 3			L = 4		
n_{ch}	31	5		29	4	3	8 14	8	6
$S_{\bar{x}}^{2}$; E_x	S^2_{min} ($x10^3$)	S^2_c ($x10^3$)	E_x (%)	S^2_{min} ($x10^3$)	S^2_c ($x10^3$)	E_x (%)	S^2_{min} ($x10^3$)	S^2_c ($x10^3$)	E_x (%)
Oxygen (ml/l)	32.61	32.65	99.8	21.22	22.61	93.8	3.31	4.35	76.1
Amoniac (µg-at/l)	48.11	54.07	88.9	32.60	37.52	86.9	0.93	1.25	74.4
Nitrites (µg-at/l)	189.90	214.13	88.7	121.45	141.75	85.7	1.21	1.39	87.0
Nitrates (µg-at/l)	15.29	15.35	99.6	10.64	10.91	97.5	2.28	2.85	80.0
Phosphates (µg-at/l)	82.84	82.95	99.8	63.09	63.47	99.4	14.00	16.10	86.9
Silicates (µg-at/l)	128.17	143.00	89.6	97.94	121.86	80.4	25.57	40.78	62.7
Chlorophyll a (µg/l)	133.48	146.20	91.3	83.42	93.38	89.3	1.27	1.69	75.1
Phaeopigments (µg/l)	4.95	4.96	99.8	3.62	3.77	96.0	0.81	0.89	91.0

TABLE 3. Parameter values of the oxygen distribution in the stratified and unstratified population, and their estimation following various sampling strategies. Confidence intervals (C.I.) in function of Student-t values with (n_h-1) or n_e degrees of freedom and a 5% confidence level. Estimations from single values (*) and weighed values (+). Values given in round numbers.

Sampling mode \ Parameters	Limits of strata (m)	Population (N=3600) N_h	$\bar{Y}_h$ (µA)	S_{hy} (µA)	C.V. (%)	Sample (n=36) n_h	$\bar{y}_h$ (µA)	s_{hy} (µA)	c.v. (%)	$s_{\bar{y}}$ (µA)	C.I. (µA)
Whole wat. col. No stratⁿ	1-3600	3600	0.160	0.032	20.0	--	--	--	--	--	--
Fix. depths No stratⁿ	1-3600	--	--	--	--	36	0.181*	0.071*	39.2	--	--
							0.160+	0.030+	18.7	--	--
Stratⁿ by temper. L=4 a) Simple rdm sampling	1-110	110	0.293	0.048	16.4	8	0.278	0.046	16.5	0.0163	±0.038
	111-1390	1280	0.180	0.006	3.3	14	0.178	0.005	2.8	0.0013	±0.003
	1391-2300	910	0.154	0.007	4.5	8	0.157	0.009	5.7	0.0032	±0.007
	2301-3600	1300	0.133	0.005	3.7	6	0.132	0.006	4.5	0.0024	±0.006
	1-3600	3600	0.160	0.004	2.5	36	0.159	0.004	2.5	0.0014	±0.003
b) Systᶜ rdm sampling	1-110					8	0.285	0.055	19.4	0.0194	±0.045
	111-1390	Identical values as above				14	0.178	0.005	2.8	0.0013	±0.003
	1391-2300					8	0.156	0.008	5.1	0.0028	±0.007
	2301-3600					6	0.132	0.005	3.9	0.0020	±0.005
						36	0.159	0.004	2.5	0.0014	±0.003

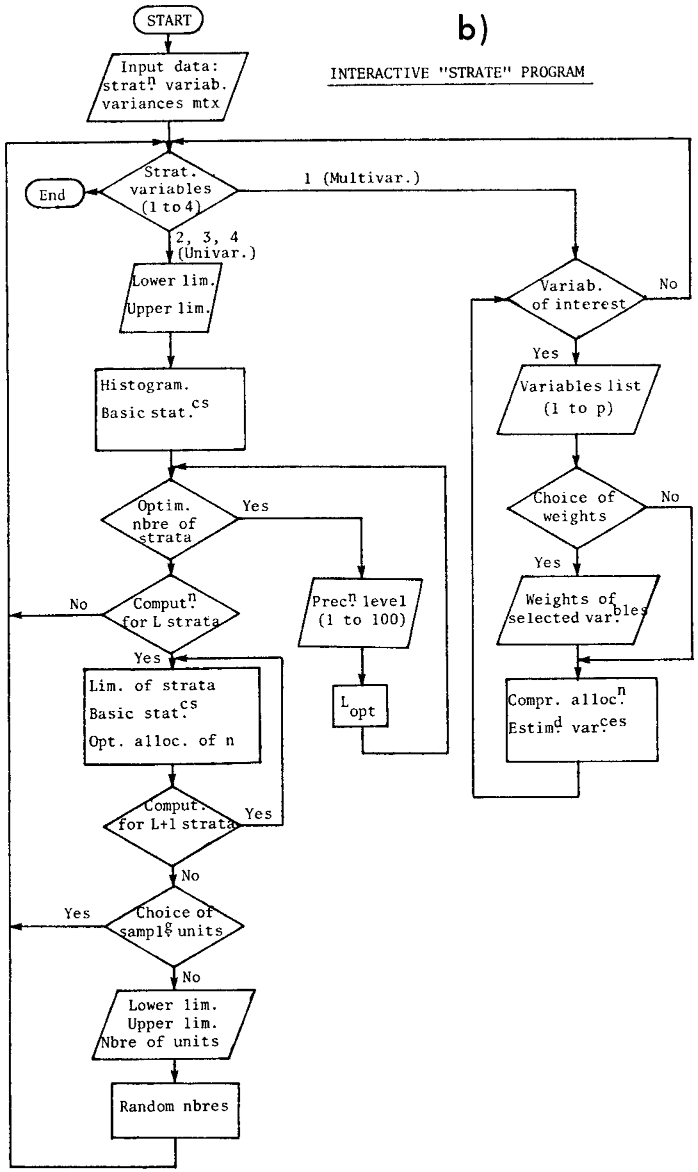

Figure 2. Vertical profiles of three environmental variables collected in the southern part of the Indian Ocean in January, 1985. Limits of strata shown by symbols in cases of L=2 (=), L=3 (Δ) and L=4 ( ) strata.

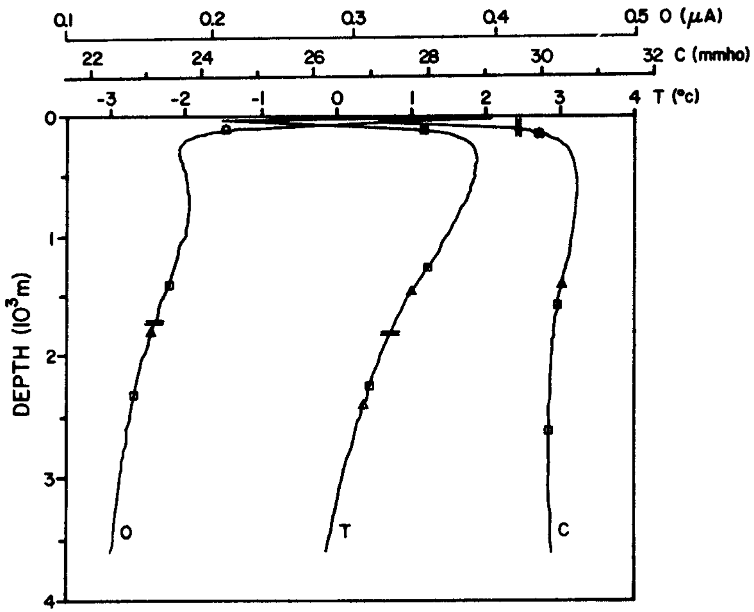

Figure 3. Temperature-conductivity diagram from data collected in the southern part of the Indian Ocean in January, 1985. Limits of strata shown by symbols in case of L=2 (=), L=3 (Δ) and L=4 (□) strata.