Design for Simultaneous Sampling of Ecological Variables: From Concepts to Numerical Solutions

OIKOS 55: 30-42. Copenhagen 1989

Design for simultaneous sampling of ecological variables: from concepts to numerical solutions

Pierre Legendre, Marc Troussellier, Vincent Jarry and Marie-Josée Fortin

Legendre, P., Troussellier, M., Jarry, V. and Fortin, M.-J. 1989. Design for simultaneous sampling of ecological variables: from concepts to numerical solutions. – Oikos 55: 30–42.

A multidisciplinary ecological study is in progress in the Thau marine lagoon, on the Mediterranean coast of France. Sampling is being conducted in two phases. Phase 1 is a pre-sampling program (pilot study), space- and time-intensive, bearing on 10 variables only; it was conducted in 1986 and 1987. During phase 2, that began in 1988, more variables will be studied at fewer stations, and at the most appropriate time scales; the purpose is to increase our understanding of ecological processes through modelling. This paper examines the results of the pre-sampling program and attempts to determine how to distribute samples through space, and through time, in order to best sample the variability of the system. Through space, four methods are proposed to select 20 stations among 63. It is shown that none of the methods always performs better than all others, their power of reproducing the best part of the original variable's variability depending upon the shape of the spatial structure (gradient, patches, hole, etc.). It is also shown that all four methods are far more efficient at rendering the system's variability than either random or systematic sampling designs. Along the time axis, the hourly, daily and monthly sampling scales were compared as to their coefficients of variation for each variable, and the daily and monthly scales were selected as being, overall, the most informative for the processes under study.

P. Legendre, Département de sciences biologiques, Univ. de Montréal, C. P. 6128, Succ. A, Montréal, Québec, Canada H3C 3J7. M. Troussellier and V. Jarry, Lab. d'hydrobiologie marine, U.A. CNRS 694, Univ. des sciences et techniques du Languedoc, Place E. Bataillon, F-34060 Montpellier Cedex, France. M.-J. Fortin, Dept of Ecology and Evolution, State Univ. of New York, Stony Brook, NY 11794-5245, USA.

Introduction

The spatial heterogeneity of natural populations has been known for a long time by naturalists and ecologists. Already 2000 years ago, authors had noted the existence of spatial heterogeneity among plants (Luc 8: 5-8 – Sower's parable). At first, last century's ecologists ignored this reality. Indeed, most 19th century quantitative ecological studies assumed living organisms to be uniformly distributed in their geographic distribution areas (Darwin 1881, Hensen 1884). Near the end of the 19th century, the uniformity hypothesis began to be questioned (Haeckel 1891), but clear demonstrations of the spatial heterogeneity of natural populations did not appear until the beginning of the 20th century (Marsh 1897, Moberg 1918, Hanson 1934). Even now, ecological variability is often considered as a nuisance by biologists obsessed by the prototype paradigm (Conrad 1983).

The importance of spatial and temporal variability comes from its central role in ecological theories and its practical role in population sampling theory. Actually, several ecological theories and models make implicit or explicit assumptions as to the causes and the importance of spatio-temporal variability. Such is the case, for instance, for models of epidemics and other catastrophes (Bartlett 1960, Taylor 1984), for the theories of competition (Bartlett 1960, Watanabe 1984), succession (Mar-

Accepted 2 November 1988

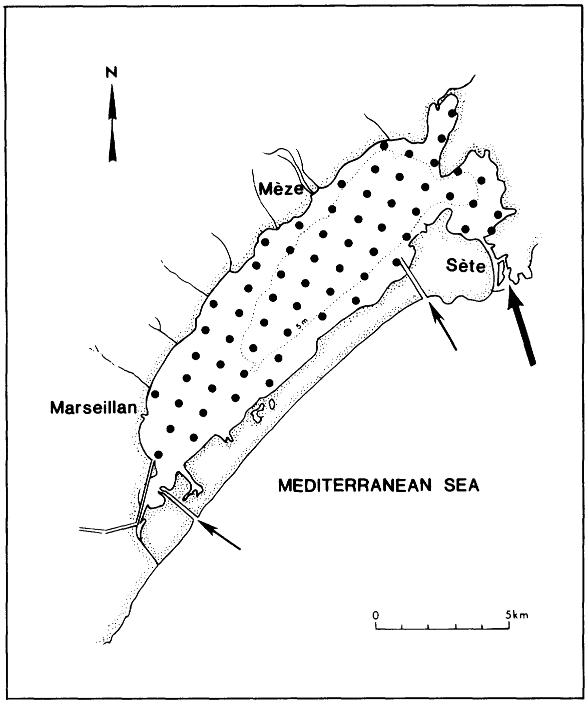

Fig. 1. Map of the 63 sampling stations (black dots) in the Thau brackish lagoon. Dotted line is the 5-m isobath. Arrows represent marine water inputs.

galef 1968, 1974, Taylor and Littler 1982), evolution and adaptations (Williams 1975, Conrad 1983), maintenance of species diversity (Levin and Paine 1974, Bach 1984), parasitism (Wcislo 1984), population genetics (Levene 1953, Felsenstein 1976, Spieth 1979, Arnold and Anderson 1983), population growth (Bartlett 1960, Vance 1984), predator-prey interactions (Anscombe 1950, Cowie and Krebs 1979, Levin 1984, Watanabe 1984) and social behaviour (Cole 1946, Stanton 1982). Other theories assume that discontinuities between homogeneous zones are important for the structure of ecosystems (succession, species-environment relationships) or for ecosystem dynamics (ergoclines: Legendre and Demers 1985). Finally, the importance of spatial heterogeneity for the maintenance of ecological stability now seems well established (Huffaker 1958, May 1974, Hassell and May 1974, Levin 1984). This shows clearly that the spatial or temporal structure of ecosystems is an important element of most ecological theories.

It is therefore easy to understand why ecological research programs, that very often attempt to elucidate the dynamics of a system and to identify the factors responsible for it, should try to describe and analyze, and then to take into account, the spatio-temporal heterogeneity of a system and that of its sub-systems. Addicott et al. (1987) discuss the difficulties of scaling environmental patterns. In the present study, we will look for one or several scales of observation pertinent to the study of some environmental processes of interest, to define our sampling strategy; defining a sampling strategy on scientific grounds is a difficult but essential preliminary step of ecological research.

In June 1986 was launched a multidisciplinary research program on a brackish lagoon ecosystem (program ECOTHAU); the objectives of this program have been described in a previous paper (Amanieu et al. in press). Phase 1 of this program (1986 and 1987) consisted in a spatio-temporal pre-sampling program (pilot study) with the aim of identifying the significant scales of variability of the most important ecological processes occurring in this lagoon. Knowledge of these scales of variability will make it possible, in phase 2 of the program (that began in 1988), to sample more efficiently the variables involved in these processes. The data will then be used to test models (i.e., sets of hypotheses) explaining the spatial and temporal behaviour of various components of this ecosystem.

The overall objective of the ECOTHAU research program is to use predictive modelling (sensu Gold 1977) to explain the variability of some important target variables. The principles of the pre-sampling program were elaborated with the following sub-objectives in mind. In the space domain, we intend to study the spatial distribution of the target variables and find a single, compromise scale of observation for all these variables, to be used in an optimal sampling design during phase 2 of the program. On the other hand, we are also looking for one or several scales of observation which would be the most appropriate to study these same target variables along the time axis; indeed, the study includes several biological compartments of the trophic chain, that have different demographic capabilities. To our knowledge, no one in aquatic ecology has yet tried to define a single sampling strategy that takes into account the spatial and temporal distributions of several ecological variables, although this problem has been looked into in other disciplines (Burkhart et al. 1978).

Materials and methods

Study area and pre-sampling design

The Thau marine lagoon is located on the shore of the French Mediterranean coast, from 3° 32′ to 3° 42′E, and from 43° 20′ to 43° 28′ N. This brackish lagoon ecosystem has been described in a previous paper (Amanieu et al. in press).

The spatial pre-sampling program is based upon a regular sampling grid with square mesh, and nodes located 1 km apart. This grid defines a total of 63 sampling stations (dots in Fig. 1). Sampling was repeated four times: in June and October 1986, and in February and May 1987. Water samples were taken 50 cm under the surface and brought back to the laboratory within less than four hours, for analysis or inoculation into

Tab. 1. Pre-sampling design for the three time scales.

Scale	Symbols	Duration	Replicates	
			No. of stations Station numbers	No. of times Dates
2 hours	H1, H2	48 h	2* 16, 35	2 1987.01.28–30, 1987.05.13–15
1 day	D1, D2	15 d	3 16, 27, 35	2 1987.01.26–02.09, 1987.05.11–25
1 month	M	12 mo	3 16, 27, 35	1 July 1986 – June 1987

*For logistic reasons, station 27 was not sampled during the 48-h program.

cultures. Sampling was carried out by three teams of scientists working from three boats.

The temporal pre-sampling program implied three temporal scales, as described in Tab. 1. The same variables were measured as in the spatial campaigns. The temporal sampling campaigns involved three sampling stations, chosen among the 63 stations of the spatial program to represent different zones of the marine lagoon.

Biotic and abiotic variables

Tab. 2 lists the ten variables used in this paper. They are the most important target variables to be modelled during phase 2 of the ECOTHAU program. A complete list of variables measured during phase 1 can be found in Amanieu et al. (in press). References are given in Tab. 2 to the methods used when measuring these variables.

Statistical methods, spatial sampling

Several methods are available for describing and analyzing spatial structures mathematically (point pattern analysis, widely used in vegetation science: Pielou 1977, Cicéri et al. 1977, Ripley 1981, 1987; surface pattern analysis: Cliff and Ord 1981, Legendre and Troussellier 1988, Legendre and Fortin in press). We have chosen the following methods to describe spatial structures: experimental variograms (Matheron 1962) were first computed, that describe the relationship between the variance of observations and the geographic distance among points; then, for each variable, interpolated maps were obtained by kriging (David 1977). Variograms as well as kriged maps were obtained using the UNIMAP computer package (European Software Contractors A/S, Nørregade, DK-2800 Lyngby, Denmark). When variograms presented difficulties of interpretation, spatial correlograms were also computed (Sokal and Oden 1978, Cliff and Ord 1981, Legendre and Legendre 1984a, Legendre and Fortin in press) using Moran's *I* index of spatial autocorrelation (1950). The spatial correlograms were computed using the "R package" (Legendre 1985).

Four different methods are proposed here to select a smaller number of stations, out of the 63 of the pre-sampling program, while taking into account the spatial distribution of each variable included in the computation. The logic behind each of these methods is the following: since the objective of the ECOTHAU program is to model ecological processes, rather than simply estimate resources, the selected sampling stations should represent as fully as possible the variability and the homogeneous zones found in the lagoon. Build-

Tab. 2. Variables studied during the pre-sampling campaigns. N = number of bacteria.

Variable	Unit	Code name	Reference
Bacteria growing on bioMérieux nutrient agar	$N \cdot ml^{-1}$	Bna	Legendre and Troussellier 1988
Bacteria growing on Difco marine agar	$N \cdot ml^{-1}$	Ma	Legendre and Troussellier 1988
Chlorophyll *a*	$\mu g \cdot L^{-1}$	Chl a	Neveux and Panouze 1987
Chlorophyll *b*/chlorophyll *a*	–	b/a	Neveux and Panouze 1987
Chlorophyll *c*/chlorophyll *a*	–	c/a	Neveux and Panouze 1987
Dissolved organic carbon	$mg \cdot L^{-1}$	Doc	Cauwet 1984
Fecal coliforms	$N \cdot 100\ ml^{-1}$	Fc	Legendre et al. 1984
NH_4^+	$mg \cdot L^{-1}$	NH_4	Aminot and Chaussepied 1983
NO_2^-	$mg \cdot L^{-1}$	NO_2	Aminot and Chaussepied 1983
NO_3^-	$mg \cdot L^{-1}$	NO_3	Aminot and Chaussepied 1983

Tab. 3. Decision-making process of selection method I (fictitious example). Stations are first ranked at random. The selected stations are in bold, as well as the group(s) that determined the choice of that station.

Station (random order)	Variable 1 (4 groups)	Variable 2 (4 groups)
8	**A**	*F*
7	*A*	*G*
3	**B**	*G*
10	*A*	*F*
14	*B*	**H**
11	*C*	*G*
12	**C**	**I**
5	*A*	*G*
6	*C*	*I*
4	**D**	*I*
15	*A*	*H*
1	*D*	*I*
9	*D*	*I*
13	*A*	*H*
2	*D*	*I*

ing coherent models implies furthermore that all variables be spatially sampled at the same stations, and temporally at the same moments.

All four methods proposed below are based upon a similarity matrix among stations, followed by clustering with spatial contiguity constraint, in order to delimit homogeneous zones in the marine lagoon. The term "homogeneous zone" is used here to mean a region of space, made of adjacent (contiguous) stations that are also similar considering the variable(s) under study. Similarities were computed using Gower's (1971) similarity coefficient; this coefficient was chosen because it is appropriate to the data at hand, according to the criteria given by Legendre and Legendre (1984a). This coefficient makes it possible to measure the degree of resemblance for all pairs of stations, based either on the values of a single variable for all four sampling campaigns (required in method I below), or else of all variables simultaneously and all four sampling campaigns (methods II to IV).

In all four methods, homogeneous zones are delimited using clustering with spatial contiguity constraint (Legendre 1987), computed from the Gower (1971) similarity matrix. In a first step, proportional-link linkage agglomerative clustering (with 50% connectedness) is computed, with spatial contiguity constraint, using the method described by Legendre and Legendre (1984b). Then the groups obtained from this first step are refined using a non-hierarchical "K-means" clustering algorithm, that minimizes within-group variances (MacQueen 1967, Anderberg 1973), also with spatial contiguity constraint. Both programs used for these computations are part of the "R package" (Legendre 1985).

Among the 63 stations of the pre-sampling program, the number of stations to be selected using the methods described below has been set to 20. This number is a compromise between logistic constraints (20 stations being the largest sampling effort that can be sustained during phase 2 of the program given the larger number of variables to be measured) and statistical constraints, since modelling requires one to have clearly more stations than variables in each model.

Method I: Univariate clustering/Consensus. In this method, one first defines homogeneous geographic zones for each variable separately. A Gower similarity matrix is computed from all 4 realizations of a given variable (4 sampling campaigns), followed by clustering in order to obtain a partition of the lagoon into 5 to 9 homogeneous zones. Repeating this process on all 10 variables in Tab. 2, one obtains 10 maps of homogeneous zones. Finally, a consensus is sought among these 10 partitions. Delimiting homogeneous zones by clustering reduces the problem of the sampling design to that of selecting a single station per zone, for each variable. To optimize this choice of stations, that is, to choose those and only those that make it possible to sample every one of the homogeneous zones computed for all variables, we will use a method related to the notion of strict consensus (Sokal and Rohlf 1981) developed for the comparison of classifications. The stations are first ranked in a random order, in order to avoid favouring certain stations; indeed, in this method the stations at the beginning of the list are more likely to be chosen. A fictitious example is presented in Tab. 3 to illustrate the method. Given n stations classified into k groups, one follows the list of stations (rewritten in random order) and selects stations that represent a new group, not yet included in the list, for the first variable; while doing so, one tries also to maximize the number of different groups represented for the other variables. When the end of the list is reached, the process is repeated for variable 2, and so on, making sure that all groups of each new variable are represented in the subsample. In the present case, 16 stations were sufficient to represent all groups of all 10 variables; following the random list of stations a little further, 4 supplementary stations were chosen that represented different groups a second time.

In methods II to IV, that follow, the first step is also a clustering with spatial contiguity constraint, from which 20 spatially contiguous groups of stations are formed; the clustering is not computed from the values of a single variable, as it was the case in method I, but from the values of all 10 variables observed during the 4 sampling campaigns ($10 \times 4 = 40$ data columns). The methods differ in the way of selecting the representative station in each zone.

Method II: Multivariate clustering/Random selection. The method consists in selecting a single station at random within each of the 20 geographic regions produced by clustering.

Method III: Multivariate clustering/Centroid. The selection criterion consists in choosing the most "central" station in each of the 20 groups delineated by clustering.

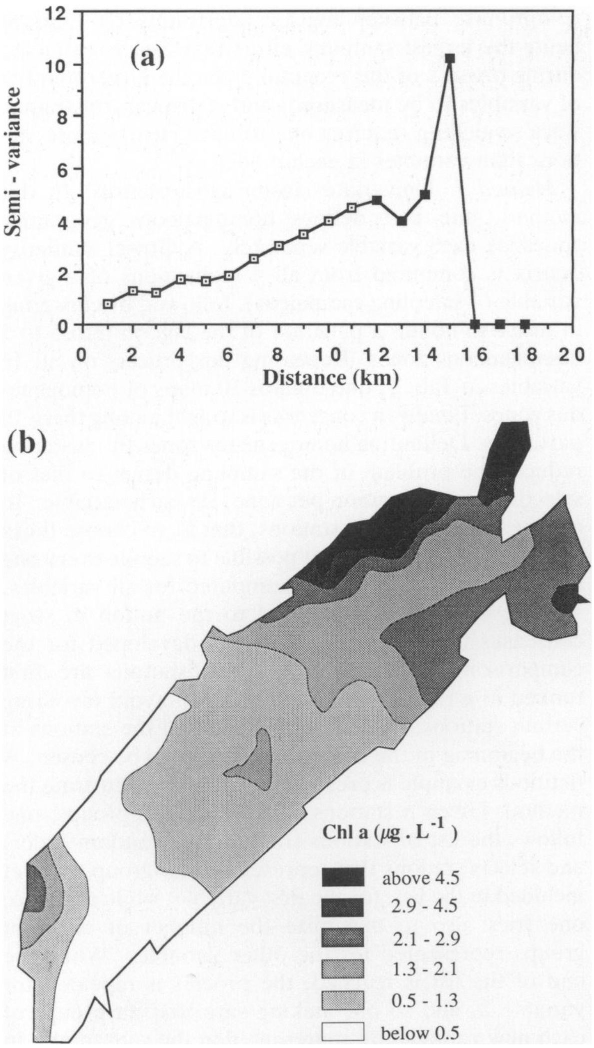

Fig. 2. Chlorophyll *a*, 17 June 1986. (a) Experimental variogram; values estimated from 45 pairs of points or fewer are represented by black squares. (b) Interpolated map obtained by kriging.

For groups containing three or more stations, a principal coordinate ordination (Gower 1966) is computed for the stations in that group and the station located closest to the center of the swarm of points is selected. For single-station groups, that station is selected. In groups with two stations, one of them is selected by random draw.

Method IV: Multivariate clustering/Maximizing the variance. In this method, we examined all 282000 possible combinations of 20 stations (one from each of the homogeneous groups obtained by multivariate clustering) and selected the combination that maximized the variance. The selection criterion is the sum of squares of the deviations from the mean of the 20 selected stations (SS); the SS values are summed over all 40 variables (TSS: Total Sum of Squares). Before this computation, the 40 variables had been standardized over all 63 stations, to avoid weighting the variances by the physical scales of measurement.

The four selection methods have been compared, for some selected variables, as to their capacity of restituting, with fewer stations, the same spatial structure that had been found with the full set of sampling stations. First, the variograms and the kriged maps obtained from 20 stations were visually compared with those obtained with all 63 stations; to make the comparison more objective, the values of the variables were interpolated by kriging at 441 geopraphic points, from 20 stations on the one hand and from all 63 stations on the other, and the resemblance between these two sets of interpolated points was measured using Kendall's tau.

The stations chosen by the four selection methods were also compared with six systematic sub-sampling designs, as well as with 10000 sets of 20 stations selected among the 63 by random sub-sampling desing, to determine if the methods proposed in this paper are any better than systematic or random sampling at rendering as large as possible a fraction of the variables' variance.

Statistical methods, temporal sampling

In the time domain, coefficients of variation were computed for each variable, at each temporal scale and for each replicate sampling. The purpose is to identify the temporal scale that offers more variability for each variable, since the objective of phase 2 of the program is to explain, through modelling, the variability of the measured variables. The choice of the most pertinent scale of variability will be constrained, however, by the logistic obligation of using the same time scale(s) for all variables; this will allow using the same variables in several models during the analysis of the ECOTHAU program results, either as independent (explanatory) variables, or as dependent variables (to be explained).

Results

Examples of spatial structures

The variograms obtained for the various variables during the ECOTHAU pre-sampling campaigns show different types of spatial structures. The theoretical variogram models mentioned below in the discussion of particular cases are described by Journel and Huijbregts (1978), David (1977), and Legendre and Fortin (in press).

Regular structures

Longitudinal gradient. This type of structure corresponds to the progressive evolution of the values of the variable along the long axis of the lagoon; in the variogram, the variance increases with distance. This is the case most commonly encountered in the Thau la-

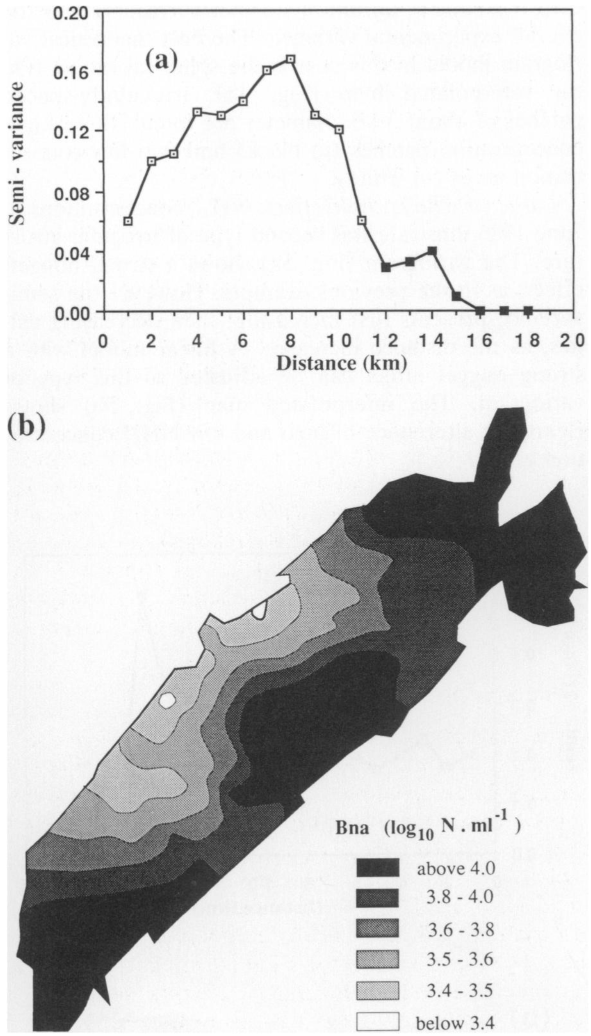

Fig. 3. Heterotrophic bacteria growing on bioMérieux nutrient agar, 6 February 1987. (a) and (b) as in Fig. 2.

goon: about 60% of the cases in our variables. The highest values may be located at the north-east end of the lagoon, or else at the south-west; one has to look at the map to decide, since a variogram does not carry this information.

Variable chlorophyll *a* in the June 1986 sampling is an example of a longitudinal gradient (Fig. 2). The experimental variogram, to which one can fit a linear model (not shown in the figure), indicates the presence of a gradient. The interpolated map shows that the highest values of chlorophyll *a* are found at the north-east end of the lagoon. The so-called "nugget effect", which is the value of semi-variance (ordinate) at distance zero, is small; if large, it would indicate that the sampling scale is inadequate to describe that variable's spatial structure. The standard deviations of the values interpolated by kriging (not illustrated) are small (average of 15%), except near the south-eastern shore; this is due to the fact that few stations are present in that part of the lagoon, compared to shore length.

Transversal gradient. The bacteria growing on bioMérieux nutrient agar (variable Bna), in February 1987, are an example of this case, less frequent than longitudinal gradients. The variogram (Fig. 3a) shows the variance increasing regularly in the first distance classes ($\leq$ 8 km), and decreasing afterwards; the "range" is the distance value where the variance stops increasing. The best theoretical variogram in this case is again the linear model. The interpolated map (Fig. 3b) clearly shows the transversal gradient, with the lowest concentrations found along the central northern shore of the lagoon. Notice that the nugget effect in the variogram is higher than in the previous example, but still lower than in the irregular structures below; it indicates that the observation scale is still adequate to describe the general trend, but the map may be less accurate than in the previous

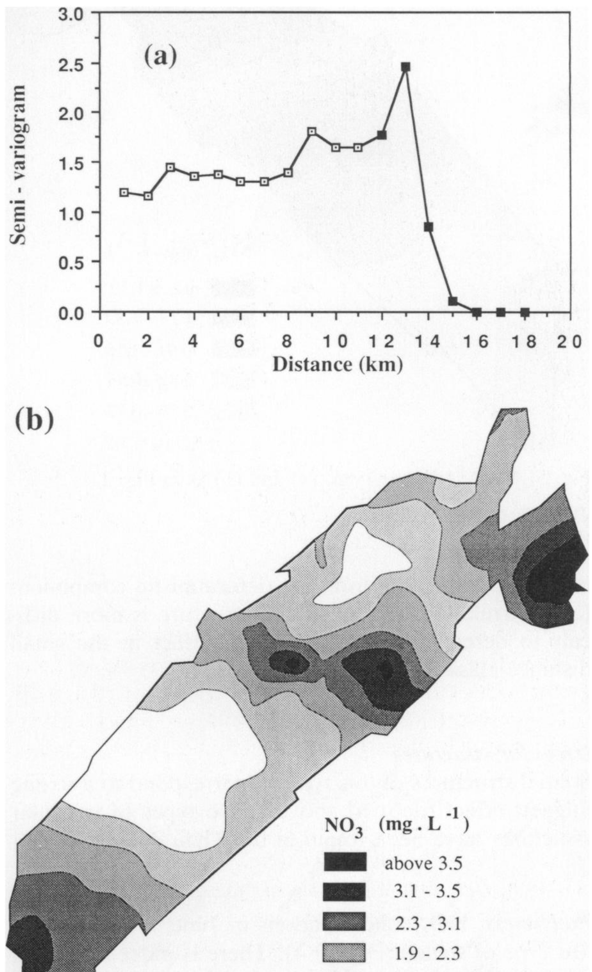

Fig. 4. NO_3^-, 17 June 1986. (a) and (b) as in Fig. 2.

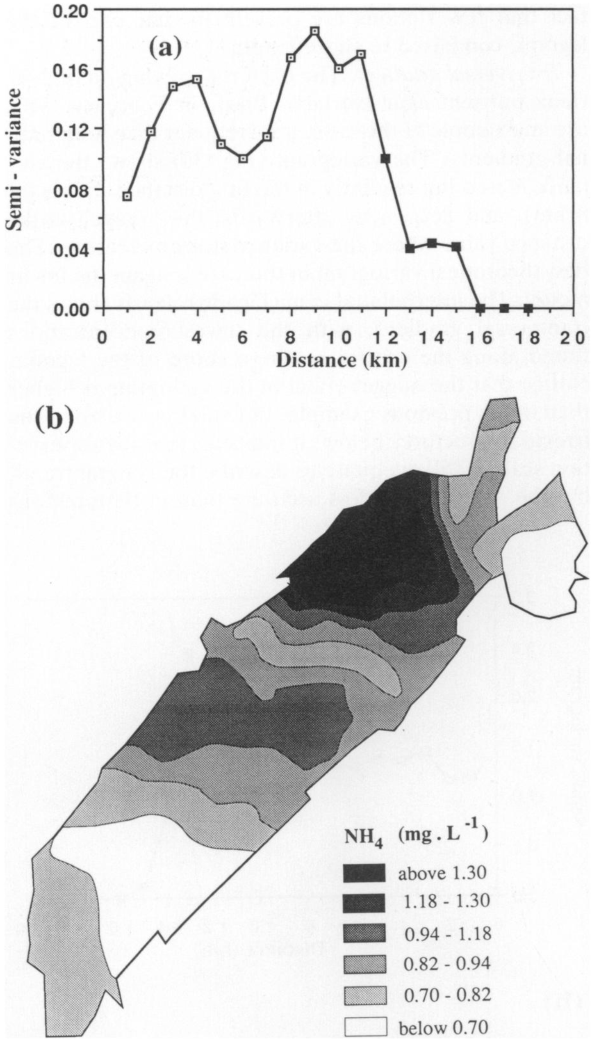

Fig. 5. NH_4^+, 17 June 1986. (a) and (b) as in Fig. 2.

example. In other words the deterministic component (autocorrelation) of the spatial structure is more difficult to detect due to a stochastic effect in the small distance classes.

Irregular structures

Spatial structures of this type do correspond to a strong nuggett effect (defined above). Two types of irregular structures have been found in the Thau data.

Patchiness. NO_3^- concentrations in June 1986 illustrate this type of structure (Fig. 4). There is indeed a strong nugget effect in the variogram, which is due to a strong stochastic component in the data. The semi-variance increases in the variogram up to distance class 3 km, then it stabilizes around a sill that corresponds to the overall experimental variance. The best theoretical variogram model in this case is the spherical model. On the interpolated map (Fig. 4b), irregularly-spaced patches of about 3 km diameter are found: three high-concentration patches (in black) and two low-concentration areas (in white).

Large structures: hole effect. NH_4^+ concentrations in June 1986 illustrate this second type of irregular structure. The variogram (Fig. 5a) shows a strong nuggett effect, as in the previous example. However the semi-variance presents first increasing, then decreasing values, as the distance increases. A linear model with a strong nugget effect can be adjusted to this type of variogram. The interpolated map (Fig. 5b) shows clearly an alternance of high and low NH_4^+ concentration zones.

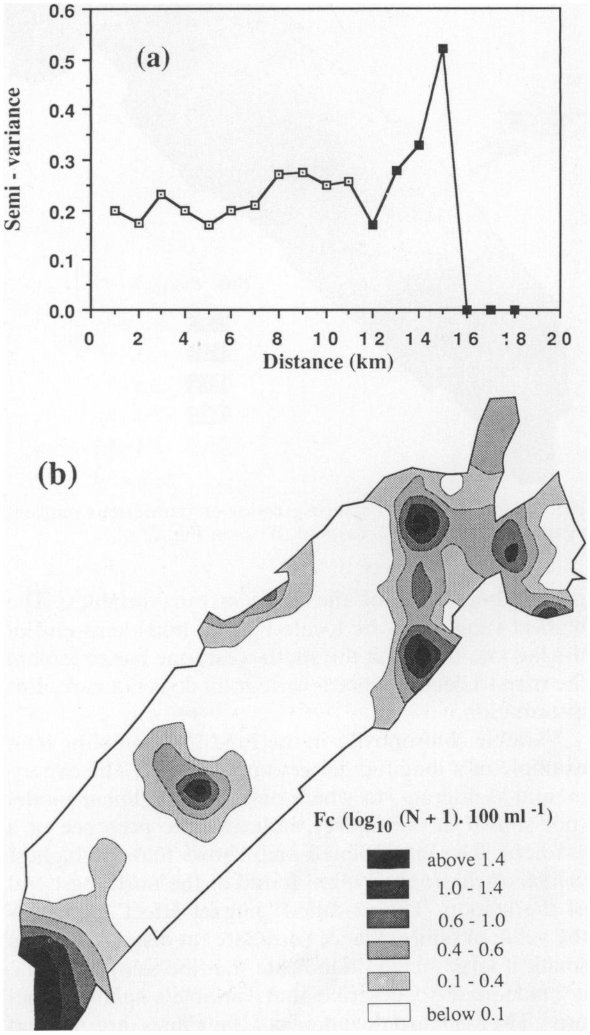

Fig. 6. Fecal coliforms, 17 June 1986. (a) as in Fig. 2. (b) Map obtained by bi-linear interpolation.

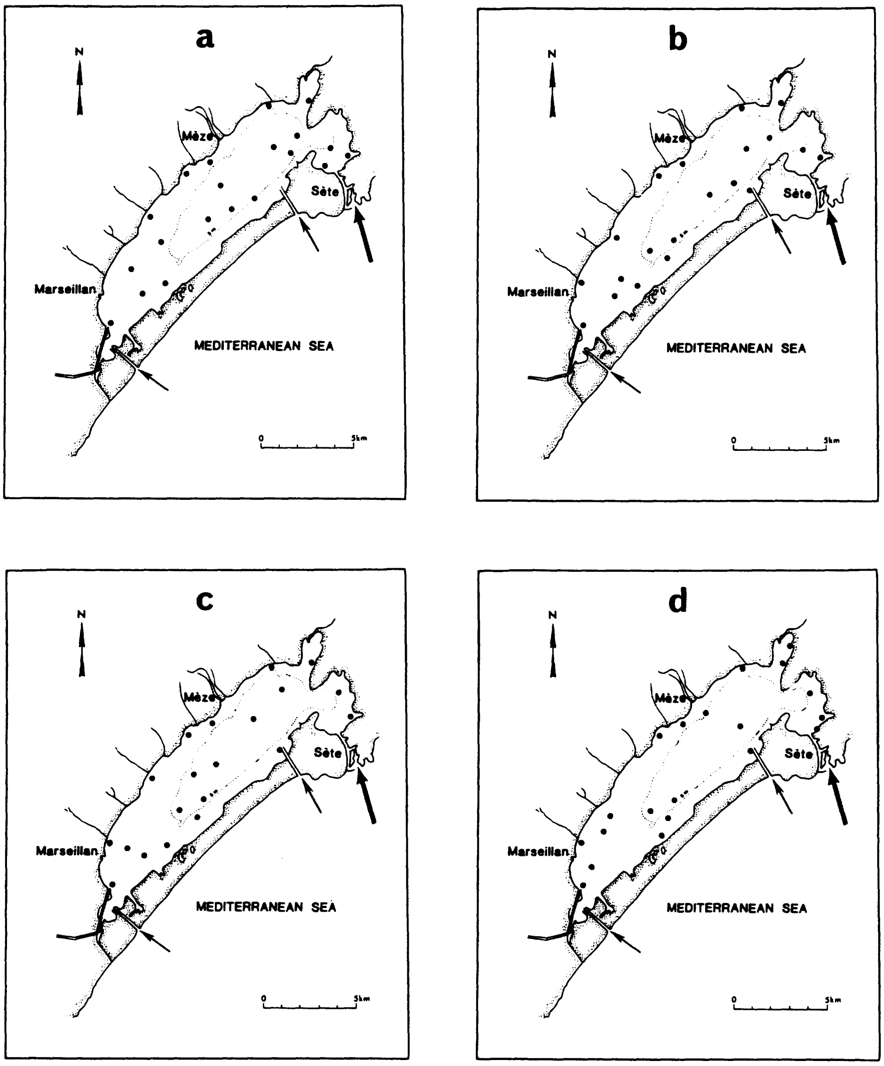

Fig. 7. Position of the 20 stations selected (a) by method I: univariate clustering/consensus; (b) by method II: multivariate clustering/random selection; (c) by method III: multivariate clustering/centroid; and (d) by method IV: multivariate clustering/maximizing the variance.

With these irregular structures, the standard deviations associated with the interpolated values (not shown here) are higher than with regular structures; this is partly due to the higher nugget values, and partly to the fact that several smaller structures (patches) are more difficult to predict accurately than a single large one (gradient). More precision could only be obtained from a sampling grid with more observation points.

Absence of a spatial structure

The absence of a spatial structure at the scale of observation considered here (grid with 1 km mesh) corresponds to a flat (horizontal) variogram, displaying only a nugget effect. This was obtained only rarely in the present data set (10 variables × 4 sampling campaigns). An example is the distribution of fecal coliforms in June 1986 (Fig. 6a). The interpolated map for this variable could not be obtained by kriging, since no model could be fitted to the experimental variogram. So the map shown in Figure 6b was obtained using the bi-linear interpolation method available in the UNIMAP package. The map shows a structure made of very small patches, each containing a single station; these patches display no particular geographic arrangement.

Comparison of methods for selecting sampling stations

Even though the most common spatial structure in the ECOTHAU data is the gradient, different spatial structures can be found for the same variable at different moments, or for different variables at the same time. In the case of the fecal coliforms for instance, variograms and spatial correlograms indicate the absence of a spatial structure detectable at our observation scale in June and October 1986, a hole effect in February 1987, and a gradient in May 1987.

The first criterion that we will examine to determine the efficiency of a sub-sampling method is the capacity

Tab. 4. Resemblance (measured by Kendall's rank-order correlation coefficient) between the maps interpolated from 63-station data on the one hand, and the maps reconstructed from the 20-station data as chosen by each of the four methods. Correlations are based on 441 interpolated geographic locations in each case. Lines are examples of variables corresponding to the various types of spatial structures found in the marine lagoon. The two best results in each line are in italics.

Type of spatial structure	Variable	Method I Consensus	Method II Clust/Random	Method III Clust/Centroid	Method IV Clust/Max TSS
Longitudinal gradient	Chl a	0.72	*0.73*	0.69	*0.74*
Transversal gradient	Bna	0.59	*0.69*	0.55	*0.68*
Patches	NO_3	*0.46*	0.44	*0.46*	0.31
Hole	NH_4	*0.55*	0.45	*0.52*	0.50
No spatial structure	Fc	0.43	*0.54*	*0.54*	0.52

of the selected 20 stations of recovering different types of spatial structures, as displayed by different variables or by the same variable at different moments. Fig. 7 shows the position of the 20 stations selected by methods I to IV proposed above. These four selections have 35 to 65% of their stations in common. Their restitution power is compared, in Tab. 4, for selected variables representing different types of spatial structures, using the method explained above.

The second criterion consists of comparing the 20 stations selected by each method to 10000 sets of 20 stations selected at random among the 63 available stations, and to six systematic sub-sampling designs. The comparison uses the variance criterion of Method IV. The 40 data vectors (10 variables × 4 sampling dates) were first standardized to prevent the physical dimensions of the individual variables from influencing the results; physical dimensions, arbitrary as they are, cause the variables to have different ranges of dispersion. Then the sum of squares (SS) of the deviations from the mean of the 20 stations selected by each sampling design was computed for each variable, and these values were summed over all 40 variables (TSS20: Total SS, 20 stations) for each of the proposed sampling designs; this statistic was then divided by the TSS value obtained for all 63 sampling stations (TSS63). For comparison purposes, sets of 20 stations were drawn at random among the 63 stations available (uniform random distribution: subroutine GGUBFS of the IMSL subroutine library) and the TSS20/TSS63 ratio was computed for each of these random selections as well. The random draw of 20 stations was repeated 10000 times, and the results cumulated to obtain a measure of the position of the results from the four proposed methods in the distribution of results one could obtain by random draw of sampling stations. These results are presented in Tab. 5. The lowest TSS20/TSS63 ratio obtained during the 10000 random draws is 0.1797 and the highest value is 0.4865, for an average value of 0.3066. Tab. 5 also shows, for comparison purposes, results that can be obtained from systematically sub-sampling 20 stations among the 63; six systematic sub-sampling designs have been tested.

Variability of the temporal scales

The coefficients of variation computed for the various time scales at each station are shown in Tab. 6, for the

Tab. 5. Comparison of sub-sampling strategies (20 stations among 63).

	Method of selecting 20 sampling stations among 63	TSS20/TSS63[1]	Probability of obtaining, at random, a larger value of TSS20/TSS63[2]
I.	Univariate clustering, consensus	0.4260	0.0247
II.	Multivariate clustering, random choice	0.4301	0.0203
III.	Multivariate clustering, centroids	0.4348	0.0166
IV.	Multivariate clustering, maximum TSS[3]	0.5226	0.0000
Systematic sub-sampling designs[4]		0.2549 to 0.3813	0.7868 to 0.1018

[1] The TSS statistic is the sum of squared deviations of the means of the standardized variables, summed over the 40 data vectors (10 variables × 4 dates). TSS20 = TSS for the 20 selected stations; TSS63 = TSS for all 63 stations. Notice that TSS63 is a constant in the study.

[2] See text. No random selection of stations did produce a TSS20/TSS63 ratio *equal* to the values observed for the various selected solutions.

[3] Method IV, which consists of selecting the group of 20 stations presenting the largest TSS statistic with a single station drawn from each of the groups delineated by clustering (see text), is here compared to random selections of 20 stations among 63 without consideration for pre-established groups.

[4] Based upon six systematic (regularly-spaced) sub-samples of 20 stations.

Tab. 6. Coefficients of variation of eight variables, sampled at various time scales. Time scale symbols as in Tab. 1; symbols for variables as in Tab. 2. Right-hand columns: – (+) is the number of times the given time scale displays the lowest (highest) coefficient of variation, for that station.

Station number	Time scale symbol	Variables								–	+
		Bna[a]	Ma[a]	Chl a	Doc	Fc[a]	NH_4	NO_2	No_3		
	M	31.0	10.5	58.1	13.3	81.0	120.0	142.0	99.0	1	4
	D1	32.6	8.0	70.7	8.0	85.0	75.3	83.8	118.5	0	1
16	D2	31.5	5.0	24.6	6.1	171.0	42.4	28.6	120.7	1	1
	H1	12.0	5.0	65.4	9.4	85.4	54.8	20.0	62.2	4	0
	H2	33.0	5.0	19.4	5.6	([b])	36.7	27.3	129.1	4	2
	M	22.9	21.4	67.0	12.6	140.0	132.0	147.0	118.0	2	3
27	D1	41.5	10.0	74.5	14.2	160.0	82.7	99.6	111.4	1	3
	D2	40.7	7.0	26.0	5.0	386.0	33.1	50.0	143.6	5	2
	M	26.0	8.0	71.4	16.1	63.0	207.0	128.0	107.0	0	1
	D1	19.3	9.0	53.6	16.5	48.0	132.5	145.7	98.0	2	1
35	D2	24.6	10.0	74.6	8.9	110.0	70.9	90.0	173.4	0	2
	H1	16.0	3.7	55.5	16.0	36.4	50.4	225.6	111.8	4	1
	H2	33.0	8.0	81.1	5.6	152.0	69.5	24.0	118.6	2	3

[a] Log_{10} transformed data.
[b] No fecal coliforms present.

variables included in this study (except chlorophyll *b* and *c*).

Discussion

Spatial sampling design

The advantages and inconveniences built into the logic of the four proposed methods are the following. Method I, which is univariate in the sense that the variables are considered one at a time, makes it possible to preserve and take into account each variable's own scale of spatial variability. On the contrary, the three other methods are multivariate from the outset and thus artificially create an average scale of variability, that may well hide some of the variables' fine scale variability.

Methods I and II contain a random choice element, instead of optimizing an objective criterion by trying in turn all possible solutions that obey the logic of the given method. For method I in particular, optimizing the TSS20/TSS63 criterion by trying in turn all possible orderings of stations, which we have not done, could help better its performance in Tab. 5; this same criterion, applied to method II, would lead to the solution found by method IV. On the contrary, methods III and IV present the advantage of "objectivity", in the sense that the ecologist does not have to intervene once he has decided on the statistical criterion he wishes to optimize. We have checked however that solutions II, III and IV did not lead to possible realisations of method I, some group(s) for some of the variables being left unrepresented in these solutions; this shows that objectivity may not be the utmost criterion one has to consider.

Tab. 4 shows that whatever the method used, the structures best reproduced by the four spatial sampling designs are the longitudinal and transversal gradients. The designs most efficient in reproducing these gradients are II and IV. Patches larger in size than the sampling interval (here, 1 km), as well as holes, are better reproduced by methods I and III. In the absence of a regular structure, the best methods were II and III. These results illustrate the fact that no unique sampling design is always better than all others; depending on the type of structure one has to sample, each of the proposed methods may turn out to be the most appropriate.

Tab. 5 shows clearly that all four methods proposed in this paper are far more efficient than random sampling, or systematic sampling, at selecting 20 stations capable of rendering a large proportion of all 63 stations' variability. Even "the worst" of the four solutions proposed here (only in terms of the criterion of maximum variability, measured by TSS20/TSS63) stands among the top 2.5%, while the best of the systematic design solutions stands much farther (top 10%) among the 10000 reference random solutions.

Solution IV, which stands out as the best one in terms of the objective TSS20/TSS63 criterion, has been designed to systematically select, in each group, the station most different from all other groups; the selected station is then often marginal with reference to its own group, and this may be seen as an undesirable characteristic of the method. For this reason, or because of logistic considerations, one can prefer one of the three other solutions, which, in any case, do not differ much from solution IV in terms of the TSS20/TSS63 criterion.

How can one optimize a sampling program in a water body, when the samples have been taken at various depths? How can one apply the methods proposed in this paper? The problem is not pertinent to the present study, since all samples were taken at the same depth of

0.5 m. It is however worth commenting upon, since it often occurs in aquatic ecosystem studies that environmental variables, measured across depths, display a strong vertical structure. There are at least five ways of extending the methods presented in this paper, not all equally good. Each one is briefly described.

First, one has to decide whether one wishes to analyze the various depth strata separately, or not. Considering the depth strata to be independent from one another may make sense ecologically. It would, for instance, for a sampling conducted in the surface layer, mid-way through the photic zone, mid-way through the aphotic zone, and at the sediment interface; this would be reason enough to consider the strata as independent statistical populations when modelling biotic processes. In the absence of such compelling ecological reasons, one may still wish to know whether the strata represent the same multivariate statistical population, or not; this can easily be tested using the well-known Wilks' lambda statistic, that measures whether there is significant variation among groups of observations in multidimensional space. The result of this test indicates whether it is appropriate to consider the layers as separate problems. In any case, practicing ecologists may find it more useful to visualize and model their water body as a series of separate layers. Coming back to the methods for selecting a smaller number of sampling stations: (1) after carrying out clustering with spatial contiguity constraint (see Methods) for each stratum separately, one may run any one of the four selection methods described in this paper. The major disadvantage of this approach is that it is very unlikely to produce the same set of sampling stations for all strata, so that logistic problems will incur during the actual sampling activity. For this reason, this solution is far from ideal. (2) To overcome this problem and reduce the number of stopping points of the sampling boats, one may modify the station-selection methods and impose that they always select as a group all depths of any one sampling station, instead of selecting individual samples; it can be done in a straightforward manner for selection methods I and IV only. This is our first viable solution.

On the other hand, in a water body which is well-mixed (at most stations, most of the time), ecologists may wish to consider samples from all depths as pertaining to the same statistical population. An easy solution may then be (3) to consider the samples from the various depths as replicates and to average the values obtained, before carrying out the selection analysis. This solution is recommended only in cases where the analysis of the pre-sampling data has clearly established that there is no vertical stratification of the water column; otherwise, one looses pertinent information when averaging values. It may be appropriate, however, to consider each depth-sample as representative of a segment of the water column, and to integrate the readings to get a value of the variable per surface unit, as is common practice in limnology and in oceanography. Another way is (4) to consider the various depths as new samples. There are then more stations to be considered for clustering, while the number of variables at each station is not increased. For clustering with spatial contiguity constraint, the vertical and horizontal spatial adjacency of the sampling stations has to be described to the clustering program, so that the resulting clusters will form homogeneous zones in a volume; one should notice that when doing this, the vertical links describing the spatial adjacency of neighbouring samples in a given water column are treated in exactly the same way by the clustering program as the horizontal links between sampling localities in the same stratum. Any one of the four methods described in this paper can then be applied to choose a smaller number of representative stations in these homogeneous zones. Here again, the sample-selection methods are unlikely to select the same stations at all depths, unless one imposes to the methods that they do so. (5) One last solution is to consider the various depths to be additional sets of variables of the same sampling localities; the number of sampling stations one is choosing from is the same as the number of surface stations, but the number of variables is multiplied by the number of depths considered. One then proceeds with any one of the four methods of selection described in this paper. This is a simple way of solving the problems encountered with methods (3) and (4). It requires of course that there be no missing depth-samples in the data. It has the drawback that a phenomenon occurring at only a few stations of a single depth may pass unnoticed and be forgotten in the following selection of sampling stations.

Temporal sampling design

At stations 16 and 35, where observations have been made for all three time scales, the smallest values for the coefficients of variation are most often found in the hourly series of results (lines H1 and H2 in the penultimate, minus-sign column of Tab. 6). Remember that we are looking for the temporal scale that offers more variability; if sampling was carried out at that scale, one would incur the risk of mistaking local spatial variation for temporal variation, at least for those variables that display small coefficients of variation, because the local variability around single stations is, for these variables, of the order of 10 to 20%.

Still at the hourly scale, some variables display on the contrary a high degree of variability (ex. Fc, NO_2 and NO_3, used to trace continental runoffs in the lagoon). The raw data show the average concentrations of these variables to be small so that they are difficult to measure with precision, given our methods of measurement, except when precipitations bring them in larger amounts in the lagoon. To use the hourly scale of observation, one would have for some of the variables to replicate the samples at each station to tell the temporal variability apart from the local variability, or, for other variables, to increase the volume of the sample to be analyzed. Using nested analysis of variance, Troussellier et al. (1986) have shown for instance what are the minimum volume and the minimum number of samples one has to analyze, at the scale of a single sampling station in the Thau lagoon, in order to obtain a given level of precision for bacterial counts.

Tab. 6 (right-hand column) shows that the highest coefficient of variation values are evenly distributed between the montly and the daily sampling scales. This was confirmed by a principal component analysis of the Tab. 6 data (not reported in more detail here). So, not only the monthly scale (appropriate to display the annual cycle), but also the daily scale is worth sampling to study the dynamics of these variables in the Thau lagoon. It is to be noticed that the daily sampling scale is but little used by ecologists; it may turn out that important components of the variability will be explained by models based on that sampling scale. Contrary to the hourly scale, the variability of most variables is higher at the daily sampling scale than what could be attributed to local within-station variability (Tab. 6). So, for phase 2 of the ECOTHAU program, it was decided to sample first at the daily scale for 15-day sequences and, given the importance of the monthly variation in Tab. 6, to repeat these sequences several times per year. A duration of 15 day was determined as a compromise between (a) the desire to have as many samples as possible available for statistical analysis and, in any case, more samples than variables in each model, and (b) the handling capacity of the labs where the samples will be analyzed.

One last point: studying the scales of temporal variability can help increase the reliability of the spatial sampling design. Indeed, since the temporal variability at the daily scale is large, one must wonder about the representativity of a spatial sampling carried out only once, without replication. Using the daily sampling series from phase 1, we created series of observations made of one, three, five, seven and 15 consecutive days, and compared the means of these series. Considering all variables, the 95% confidence intervals around the mean values show that the means computed from three consecutive days are very often significanty different from single-day observations, but are rarely different from the means computed for five and for seven days. The means of the 15-day series are however generally different from all others. Variations of the means in the 15-day series are accompanied by an increase in variance, which indicates that too long a time series describes a different ecological process; in this case, temporal variability may be important enough to mask the spatial variability. For this reason, and considering the maximum sampling effort that can be tolerated by the field and laboratory teams, we believe that to correctly represent the spatial variability of the variables of interest in the ECOTHAU program, a meaningful spatial sampling design should consist of repeating the spatial sampling of 20 stations during three consecutive days.

Ackowledgements – This is publication No. 5 of the ECOTHAU program. This paper would not have been possible without the work of the 50-and-some research scientists who took part in the ECOTHAU pre-sampling campaigns and correlative laboratory work. The ECOTHAU program is financed by C.N.R.S. (program PIREN), Ministère de l'environnement, and I.F.R.E.M.E.R. This research has also been supported by grants from the France-Québec exchange program, a research fellowship of Ministère des Affaires étrangères (France) to V. Jarry, and by the NSERC grant No. A7738 (Canada) to P. Legendre.

References

Addicott, J. F., Aho, J. M., Antolin, M. F., Padilla D. K., Richardson, J.S. and Soluk, D.A. 1987. Ecological neighborhoods: scaling environmental patterns. – Oikos 49: 340–346.

Amanieu, M., Legendre, P., Troussellier, M. et Frisoni, G.-F. Le programme ÉCOTHAU: théorie écologique et base de la modélisation. – Oceanol. Acta (in press).

Aminot, A. and M. Chaussepied. 1983. Manuel des analyses chimiques en milieu marin. – Centre National pour l'Exploitation des Océans, Paris.

Anderberg, M. R. 1973. Cluster analysis for applications. – Academic Press, New York.

Anscombe, F.J. 1950. Sampling theory of the negative binomial and logarithmic series distributions. – Biometrika 37: 358–382.

Arnold, J. and Anderson, W. W. 1983. Density-regulated selection in a heterogeneous environment. – Am. Nat. 121: 656–668.

Bach, C. E. 1984. Plant spatial pattern and herbivore population dynamics: plant factors affecting the movement patterns of a tropical cucurbit specialist (*Acalymma innubum*). – Ecology 65:175–190.

Bartlett, M. S. 1960. Stochastic population models in ecology and epidemiology. – Methuen, London.

Burkhart, H. E., Stuck, R. D., Leuschner, W. A. and Reynolds, M. R. 1978. Allocating inventory resources for multiple-use planning. – Can. J. For. Res. 8: 100–110.

Cauwet, G. 1984. Automatic determination of dissolved organic carbon in sea water in the sub-parts per million range. – Mar. Chem. 14: 297–306.

Cicéri, M.-F., Marchand, B. et Rimbert, S. 1977. Introduction à l'analyse de l'espace. – Collection de Géographie applicable. Masson, Paris.

Cliff, A. D. and Ord, J. K. 1981. Spatial processes: models and applications. – Pion, London.

Cole, L. C. 1946. A theory for analyzing contagiously distributed populations. – Ecology 27: 329–341.

Conrad, M. 1983. Adaptability. The significance of variability from molecule to ecosystem. – Plenum Press, New York.

Cowie, R. J. and Krebs, J. R. 1979. Optimal foraging in patchy environments – In: Anderson, R. M., Turner B. D and Taylor, L. R. (eds), Population dynamics. Blackwell, Oxford, pp. 183–205.

Darwin, C. 1881. The formation of vegetable mould through the action of worms. – John Murray, London.

David, M. 1977. Geostatistical ore reserve estimation. – Developments in Geomathematics, 2. Elsevier, Amsterdam.

Felsenstein, J. 1976. The theoretical population genetics of variable selection and migration. Ann. Rev. Genet. 10: 253–280.

Gold, H. J. 1977. Mathematical modelling of biological systems: an introductory guidebook. – Wiley, New York.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. – Biometrika 53: 325–338.
Gower, J. C. 1971. A general coefficient of similarity and some of its properties. – Biometrics 27: 857–871.
Haeckel, E. 1891. Plankton Studien. – Jena Z. Naturw. 25: 232–336.
Hanson, H. C. 1934. A comparison of methods of botanical analysis of the native prairie in western North Dakota. – J. Agric. Res. 49: 815–842.
Hassell, M. P. and May, R. M. 1974. Aggregation in predators and insect parasites and its effect on stability. – J. Anim. Ecol. 43: 567–594.
Hensen, V. 1884. Ueber die Bestimmung der Planktons oder des im Meer triebenden Materials an Pflanzen und Tieren. – Bericht Commission Wiss. Unters. Deutschen Meere 5.
Huffaker, C. B. 1958. Experimental studies on predation: dispersion factors and predator-prey oscillations. – Hilgardia 27: 343–383.
Journel, A. G. and Huijbregts, C. 1978. Mining geostatistics. – Academic Press, London.
Legendre, L. and Demers, S. 1985. Auxiliary energy, ergoclines and aquatic biological production. – Naturaliste can. (Québec) 112: 5–14.
– and Legendre, P. 1984a. Écologie numérique, 2ième éd. Tome 2: La structure des données écologiques. – Masson, Paris et les Presses de l'Université du Québec.
Legendre, P. 1985. The R package for multivariate data analysis. – Dépt. sci. biol., Univ. de Montréal.
– 1987. Constrained clustering. – In: Legendre, P. and Legendre, L. (eds.), Developments in numerical ecology. NATO ASI Series, Vol. G 14. Springer, Berlin, pp. 289–307.
– and Fortin, M.-J. Spatial pattern and ecological analysis. – Vegetatio (in press).
– and Legendre, V. 1984b. Postglacial dispersal of freshwater fishes in the Québec peninsula. – Can. J. Fish. Aquat. Sci. 41: 1781–1802.
– and Troussellier, M. 1988. Aquatic heterotrophic bacteria: modeling in the presence of spatial autocorrelation. – Limnol. Oceanogr. 33: 1055–1067.
– , Troussellier, M. and Baleux, B. 1984. Dynamics of pollution-indicator and heterotrophic bacteria in sewage treatment lagoons. – Appl. Eviron. Microbiol. 48: 586–593.
Levene, H. 1953. Genetic equilibrium when more than one ecological niche is available. – Am. Nat. 87: 331–333.
Levin, L. A. 1984. Life history and dispersal patterns in a dense infaunal polychaete assemblage: community structure and response to disturbance. – Ecology 65: 1185–1200.
Levin, S. A. and Paine, R. T. 1974. Disturbance, patch formation, and community structure. – Proc. Nat. Acad. Sci. U.S.A. 71: 2744–2747.
MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. – In: Le Cam, L. M. and Neyman, J. (eds), Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability, Vol 1. Univ. of California Press, Berkeley, pp. 281–297.
Margalef, R. 1968. Perspectives in ecological theory. – Univ. of Chicago Press, Chicago.
– 1974. Ecología. – Ediciones Omega, Barcelona.
Marsh, C. D. 1897. On the limnetic crustacea of Green Lake. – Trans. Wisc. Acad. Sci. Arts Lett. 11: 179–224.
Matheron, G. 1962. Traité de géostatistique appliquée. Tome 1. – Editions Technip, Paris.
May, R. M. 1974. Stability and complexity in model ecosystems, 2nd ed. – Princeton Univ. Press, Princeton, NJ.
Moberg, E. G. 1918. Variation in the horizontal distribution of plankton in Devils Lake, North Dakota. (Edited by R. T. Young.) – Trans. Am. Microsc. Soc. 37: 239–267.
Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. – Biometrika 37: 17–23.
Neveux, J. and Panouze, M. 1987. Spectrofluorometric determination of chlorophylls and pheophytins. – Arch. Hydrobiol. 109: 567–581.
Pielou, E. C. 1977. Mathematical ecology, 2nd ed. – Wiley, New York.
Ripley, B. D. 1981. Spatial statistics. – Wiley, New York.
– 1987. Spatial point pattern analysis in ecology. – In: Legendre, P. and Legendre, L. (eds), Developments in numerical ecology. NATO ASI Series, Vol. G 14. Springer, Berlin, pp. 407–429.
Sokal, R.R. and Oden, N. L. 1978. Spatial autocorrelation in biology. 1. Methodology. – Biol. J. Linn. Soc. 10: 199–228.
– and Rohlf, F. J. 1981. Taxonomic congruence in the Leptopodomorpha re-examined. – Syst. Zool. 30: 309–325.
Spieth, P.T. 1979. Environmental heterogeneity: a problem of contradictory selection pressures, gene flow, and local polymorphism. – Am. Nat. 113: 246–260.
Stanton, M. L. 1982. Searching in a patchy environment: foodplant selection by *Colias p. eriphyle* butterflies. – Ecology 63: 839–853.
Taylor, L. R. 1984. Assessing and interpreting the spatial distribution of insect populations. – Ann. Rev. Ent. 29: 321–357.
Taylor, P. R. and Littler, M. M. 1982. The roles of compensatory mortality, physical disturbance, and substrate retention in the development and organization of a sand-influenced, rocky-intertidal community. – Ecology 63: 135– 146.
Troussellier, M., Baleux, B. and André, P. 1986. Échantillonnage de variables bactériologiques dans les milieux aquatiques. – In: GERBAM (éd.), Deuxième colloque international de bactériologie marine. Actes de Colloques no 3, CNRS/IFREMER, Brest, pp. 23–33.
Vance, R. R. 1984. The effect of dispersal on population stability in one-species, discrete-space population growth models. – Am. Nat. 123: 230–254.
Watanabe, J. M. 1984. The influence of recruitment, competition, and benthic predation on spatial distribution of three species of kelp forest gastropods (Trochidae: *Tegula*). – Ecology 65: 920–936.
Wcislo, W. T. 1984. Gregarious nesting of a digger wasp as a "selfish herd" response to a parasitic fly (Hymenoptera: Sphecidae; Diptera: Sacrophagidae). – Behav. Ecol. Sociobiol. 15: 157–160.
Williams, G. C. 1975. Sex and evolution. – Princeton Univ. Press, Princeton, NJ.