Analyzing Multivariate Flow Cytometric Data in Aquatic Sciences

Serge Demers, Junhyong Kim, Pierre Legendre, and Louis Legendre
Institut Maurice-Lamontagne, Pêches et Océans, C.P. 1000, Mont-Joli, Québec, Canada, G5H 3Z4 (S.D.),
Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794-5245 (J.K.),
Département de Sciences Biologiques, Université de Montréal, C.P. 6128, Succursale A, Montréal, Québec, Canada,
H3C 3J7 (P.L.), and GIROQ, Département de Biologie, Université Laval, Sainte-Foy, Québec, Canada,
G1K 7P4 (L.L.)

Received for publication January 4, 1991; accepted August 17, 1991

Flow cytometry has recently been introduced in aquatic ecology. Its unique feature is to measure several optical characteristics simultaneously on a large number of cells. Until now, these data have generally been analyzed in simple ways, e.g., frequency histograms and bivariate scatter diagrams, so that the multivariate potential of the data has not been fully exploited. This paper presents a way of answering ecologically meaningful questions, using the multivariate characteristics of the data. In order to do so, the multivariate data are reduced to a small number of classes by clustering, which reduces the data to a categorical variable. Multivariate pairwise comparisons can then be performed among samples using these new data vectors. The test case presented in the paper forms a time series of observations from which the new method enables us to study on the temporal evolution of cell types.

Key terms: Multivariate analysis, clustering, flow cytometry, aquatic sciences

The advent of flow cytometry in aquatic sciences is a development of great significance for the ecological and physiological study of natural populations of microbial plankton (11, 16). Flow cytometry is a powerful tool for acquiring data on the optical and fluorescence characteristics of particles in the aquatic environment. The flow cytometer was initially developed for biomedical purposes; it was introduced in aquatic sciences at the beginning of the 1980s (16). In contrast to the medical field, where the particles to be analyzed are relatively homogeneous, natural samples of aquatic particles are generally very heterogeneous, i.e., they most often comprise a mixture of different cell populations (e.g., species or other categories). Such cell populations may react differently to environmental variations. In addition, flow-cytometric measurements may be contaminated by inorganic particles and artifacts, which increase the difficulty of interpreting the data. This results in data that are multimodal (multiple peaks corresponding to different populations) and heteroscedastic (heterogeneous variance-covariance structure). In most instances, however, researchers are interested in measuring several variables on the same cell. The important problem is that usually the variables that we measure are not independent of one another and, therefore, we cannot examine results from each variable individually and draw conclusions from them. As an example, imagine a doughnut in two dimensions projected onto the X and Y axes. The shadow of the doughnut on each axis would simply indicate an interval, and we would have no idea that the actual shape was a doughnut. The "hole" in the middle is only realized through coordinated combinations of the X and Y variables. If the number of variables is less then 3, simple plots will help visualize the data and give the investigator some insight into the nature of the data. With higher dimensions, many standard techniques exist for scaling the data in a small number of dimensions while capturing the essence of the variation in the data [e.g., principal component analysis (PCA), factor analysis, metric or nonmetric multidimensional scaling, etc.]. Most of these procedures involve rotating and scaling the data such that the variables become less correlated and more amenable to independent analysis.

The multimodality of the data often reflects the presence of different populations, each clustered around their own centroids (of course, it is also possible for the population to be fundamentally multimodal). In such cases, it is often desirable to separate the populations such that the analysis can be done on each population individually. Standard techniques include cluster anal-

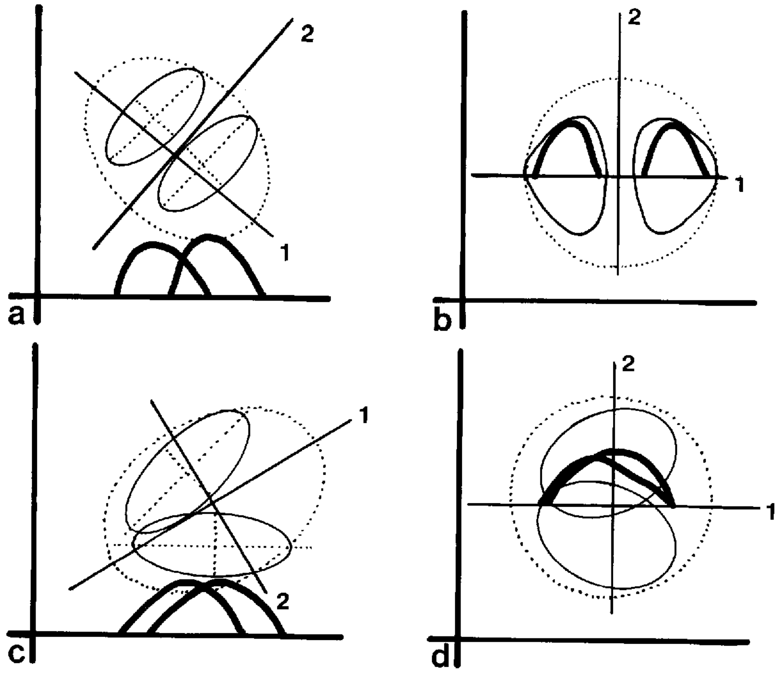


FIG. 1. Schematic diagram of PCA transformation of data showing clusters. The ellipses indicate the variation within each cluster. The large ellipse in dotted line indicates the variability of the total population, i.e., the two clusters together. Minor axes labeled 1 and 2 indicate the PCA axes for the total population. The thick lines indicate the clusters when the populations are projected onto a single axis. a) Two population with homogeneous variance-covariance structure (as indicated by thin dotted lines within each ellipse); b) hypothetical PCA transformation of (a); c) two populations with heterogeneous variance-covariance structure and (d) PCA transformation of (c) and the projection of the populations onto the first PCA axis.

ysis and discriminant analysis. Discriminant analysis tries to separate out populations in such a way that the between-group sums of squares are maximized; cluster analysis, on the other hand, bases its criterion on some distance measure defined on the data space. The reason for the separation of populations is, of course, that in many cases the investigator is interested in variable values associated with single populations; the populations must first be identified before statistical analysis can be performed.

Heteroscedastic data present more complicated problems for standard analyses. A heterogeneous variance-covariance structure usually results from the biological characteristics of the populations. For example, the optical characteristics of the cells may show a range of intraspecific variability that is due to variations in physiological or nutritional states, with the resulting variance differing depending of the species (15). Figure 1 illustrates the types of problems that may be presented by such data. Figure 1a shows two populations (solid line ellipses) where the two variables measured are correlated with one another but the correlation is homogeneous among the two populations. Unless one can identify the two population variance-covariance structures a priori, all the analyses with which we work will use the variance-covariance structure of the total observations, in this case that of the two populations combined. The total variance-covariance structure of the data is depicted as a dotted line ellipse in the figure. A principal component analysis applied to such data would then orient the major variation axis as the two axes labeled 1 and 2 in the figure. Rotation and standardization along the two PCA axes would result in Fig. 1b. Under normal conditions, we would be examining higher dimensional data, which we would project onto PCA axes in order to obtain insight into the original multidimensional structure. This is shown as the thick solid lines in the 4 figures composing Fig. 1. As can be seen with Figs. 1a and 1b, although we could not tell the 2 populations apart by examining either the X-axis or Y-axis projection alone in Fig. 1a, the PCA axis 1 projection in Fig. 1b allows us to easily separate the two populations. Consider now Fig. 1c, where the variance-covariance structures of the 2 populations differ from each other (the small axes within the ellipses have different angles). As can be seen in Fig. 1d, a projection onto the first PCA axis does not easily differentiate the 2 populations. Although the structure of the 2 populations was intuitively obvious in this case, since the example was in 2 dimensions only, this would not be the case when the data are in high dimensions and can only be visualized by using some ordination to lower dimensionality. Flow cytometry also generates large data sets, which considerably limits the applicability of the technique for routine studies. Up to now, flow-cytometric data have generally been analyzed by simple methods only, such as frequency histograms, bivariate scatter diagrams, etc.; the multivariate potential of the data generally remains unexploited. A major development for aquatic studies would be to improve the numerical procedures for analyzing multivariate flow cytometric data. The present paper describes an approach that fully takes into account the multivariate nature of flow-cytometric data and presents an example of results obtained using this approach.

MATERIALS AND METHODS

Numerical Analysis

The numerical procedure described here is based on cluster analysis. The purpose of cluster analysis is to allocate objects to groups or clusters, which are not defined a priori but emerge from the data such that objects in a given cluster tend to be similar to one another and objects in different clusters tend to be dissimilar. This allocation can be based on a single variable measured on the objects, but it is more often based on multivariate data. Because of the characteristics mentioned above in the Introduction, clustering of flow cytometric data presents unusual problems. First, since the number of cells analyzed by flow cytometry may easily exceed 10^4, it is not practical to consider clustering algorithms that require the computation of pairwise distance matrices, e.g., UPGMA clustering (5). Second, multivariate flow cytometry data of multiple populations often show variance-covariance structures (matrices) that present difficulties with some of the commonly used clustering algorithms; i.e., the within population variance-covariance structure may differ widely from one population to another. This is quite natural, since the biological response of different populations to the measured variables may not be the same. Most clustering algorithms assume that the variance-covariance matrices of the different populations are equal or similar. Therefore, application of algorithms that do not account for different variance-covariance structures, e.g., k-means algorithm (1, 8), may result in partitions that seem unnatural and are unable to identify accurately different populations in the data.

In the algorithm described below, a variation of the normal mixture algorithm is used (4). The normal mixture model assumes that the observations come from k populations, each with an arbitrary variance-covariance matrix. Using an iterative process, the algorithm assigns probabilities P(j) (j = 1,k) to each observation, where P(j) denotes the probability that the observation belongs to cluster j. As in the k-means algorithms (1, 8), the number of clusters (k) is fixed a priori by the investigator. The iterations of the normal mixture algorithm try to maximize a function that describes the log likelihood of the observed data, given the k different allocations of each particle to the clusters. The algorithm is modified here by positively allocating each observation to one of the k possible clusters in an iterative manner.

In a first step, our algorithm tries to estimate the mean of each of the k clusters. This is achieved in the same way as in Hartigan's leader-algorithm (4). First, k initial seeds are selected as starting points for the k centroids. The initial seeds may be picked randomly from the data set or selected by the investigator. These initial seeds are also the initial centroids for each cluster. Second, Mahalanobis generalized distances (9) are calculated between each observation and each of the k initial centroids:

$$D^2_{ij} = X'_{ij}\, S^{-1}_j\, X_{ij}. \qquad (1)$$

D^2_{ij} is the squared distance between observation i and centroid j, X_{ij} denotes the vector of the multivariate differences between observation i and centroid j, and S^{-1}_j is the inverse of the variance-covariance matrix for cluster j. In the initial step, all S_j's are set as the identity matrix. Observations are assigned to clusters with the smallest D_{ij} value. The values of the k centroids are updated at each step to include the newly assigned observations.

The second step assumes that only the observations that are close to the centroids (within distance c_j) are correctly allocated, and therefore are appropriate to be included in the estimation of the variance-covariance matrices. If such a restriction is not set, each iteration will allocate more and more observations to the cluster possessing the largest number of observations. In this step, a matrix S_j is computed for each of the k clusters. First, for each centroid, Euclidean distances are calculated to each of the other k-1 centroids and the smallest distance is determined. This distance is then multiplied by a constant z, which results in a unique value c_j for that centroid. Initially, constant z is set to 0.25, resulting in c_j's that are 1/4 of the distance to the nearest centroid. The variance-covariance matrix S_j for each cluster j is calculated by computing the variances and covariances of all observations that have been allocated to cluster j (during the previous step) and that are within distance c_j. In subsequent iterations, this value is increased under the assumption that the allocation of observations becomes more reliable for calculating the variance-covariance matrices.

In the third step, distances D_{ij} are computed using the new S_j matrices, and observations are reassigned to the nearest cluster based on these new distances. The first step is repeated, but now using the newly calculated S_j's instead of the old ones. A cost function is then computed as

$$\mathrm{Cost} = \sum_{i=1}^{n} d^2_i. \qquad (2)$$

Here d^2_i is the Mahalonobis distance (9) between observation i and the centroid of the cluster to which it has been assigned. This function is analogous to the log likelihood used in the usual mixture algorithm. Additional iterations of steps 1–3 are performed to improve the clustering results, using minimization of the cost function as the criterion for goodness of fit.

In the example discussed below, the data are a series of measurements taken at regular time intervals. If there are reasons to believe that the parameters of the subpopulations (in terms of the variables measured by the flow cytometer) do not change among samples, but that only the number of cells in each sample varies with time, we are justified to analyze all the samples simultaneously. Observations can be assigned to the k clusters based on the whole of the t samples, after which the observations in each cluster can be reassigned to the original t times. As seen below, such an analysis allows the study of individual population dynamics for mixed community samples. The algorithm presented above and the assignment procedure have been implemented in a user-friendly program written in PASCAL.

Biological Sampling

It is important to note that the flow-cytometric data concerning phytoplankton populations are used here only to illustrate the numerical procedure described above. These data come from an experiment on the feeding behavior of mussels (*Mytilus edulis*) in the presence of the toxic microalga *Alexandrium excavatum*. However, it is not in our intention to interpret the ecological aspects of the results in this paper; the ecological implications of these results will be presented in detail elsewhere.

The experimental protocol was as follows: A natural population of marine phytoplankton was incubated for 3 d in a controlled-temperature room (15°C) under continuous light (Optimarc-400; irradiance of 150 μmol of photons m^{-2} s^{-1}). Water was enriched with F/20 medium (3). After this initial incubation, cultured *Alexandrium excavatum* were added to the natural phytoplankton population to obtain a final concentration of about 200 cells/ml. Four liters of this mixture were then fed to 8 mussels (3.0–3.5 cm), for a period of 7 h. A magnetic stirring bar maintained gentle mixing in order to prevent sedimentation of particles. Ten samples were taken during the 7-h period (at 1, 12, 26, 47, 68, 94, 124, 221, 316, and 448 min) to monitor the feeding behavior of the mussels. Sample 0 was taken just before adding the toxic algae. Water samples were analyzed using a Becton-Dickinson FACS Analyzer flow cytometer, equipped with a Coulter-type volume analyzer and a 75-μm square orifice. Measurements on each particle included cell volume, phycoerythrin (FL1), and in vivo fluorescence of chlorophyll *a* (FL2) and 90° light scatter. Cells were excited by a mercury/cadmium arc lamp at a wavelength of 488 nm, and autofluorescence of both chlorophyll *a* emission greater than 665 nm as well as phycoerythrin at 575 nm were detected by photomultiplier tubes. Fluorescence units were normalized relative to 10-μm standard fluorescent beads (Coulter EPICS Division, Hialeah, FL) added to each sample at the beginning of the counts (final concentration of 1000 beads/ml). The samples were analyzed in volumes of 0.5 ml each time. The operating current was set at 0.5 MA, and the gain was on a logarithmic scale. Cell size measurements were calibrated in terms of equivalent spherical diameter, based on standard volume polystyrene beads, and the signal threshold was set on Coulter volume. All data were collected in list mode to allow further analysis. In the present paper, we analyzed only the fluorescent particles.

RESULTS

Figure 2a shows typical results from a Becton Dickinson FACS Analyzer flow cytometer. With the FACS Analyzer, 4 variables can be measured on each particle: Coulter volume, FL1, FL2, and 90° light scatter. Data are normally represented in contour or scatter diagrams (Fig. 2b). Although flow-cytometric measurements are typically multidimensional, it is almost impossible to follow the spatio-temporal changes of each group with such representations. This is because one has to use gating to isolate a given group, chosen on arbitrary criteria based on two variables only (Fig. 2c). Simple statistics are provided for each gate (Table 1). Moreover, each sample must be analyzed separately, which is time consuming. The procedure described in this paper allows the separation of the data in such a way that a multidimensional peak corresponds to a single population, while troughs correspond to boundaries between populations.

Table 2 shows the characteristics of the 5 clusters generated by the above procedure, using the entire time series of 10 samples. The number of cluster has been chosen from the contour graph resulting from the principal component analysis. Each cluster contains cells with similar characteristics, on the basis of the 4 measured variables. Groups 4 and 5 correspond to the toxic algae and fluorescent beads, respectively. In the data set chosen, these additions are independent of the natural population, and so could be used as controls. Microscopic identification showed that group 1 corresponds to *Chaetoceros debilis* (size range 11–13 μm); group 2 corresponds to small cells, such as *Chaetoceros* sp. (5.6–6.4 μm), *Skeletonema costatum* (3.6 μm), and small flagellates (4.8 μm), whereas group 3 corresponds to *Thalassiosira pacifica* (17–20 μm) and *T. conferta* (12.6–13.8 μm).

Table 3 shows changes in the 5 clusters during the course of the experiment. Time 0, which corresponds to the sample before the toxic algae were added, indicates that there were some toxic cells in the natural population; toxic algae are normally present in natural waters at the time of the year the natural sample was collected (August). Decreasing numbers of particles in groups 1–4 (last column) is the result of grazing by blue mussels. Group 5 stays constant during the whole experiment, since the beads were added as controls just before counting the samples on the flow cytometer. Us-

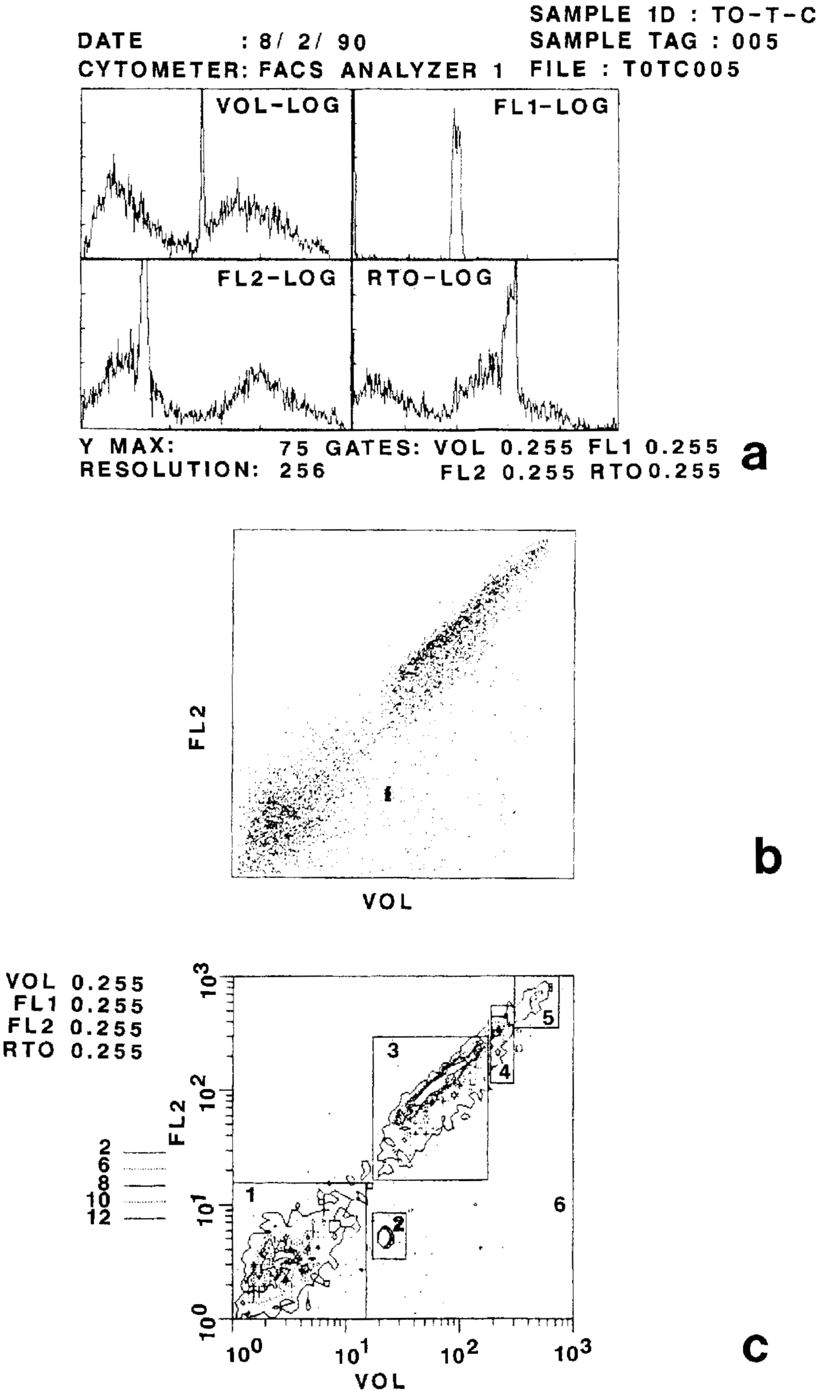


FIG. 2. **a**) Typical results from a Becton Dickinson FACS Analyzer flow cytometer. With the FACS Analyzer, 4 variables can be measured on each particle: Coulter volume (VOL), FL1, FL2, and 90° light scatter (SSC). Flow-cytometric data are normally represented using **b**) a scatter diagram (in this case the x-axis is the volume and the y-axis is the red fluorescence emitted by chlorophyll *a*) or **c**) contour diagrams on which the user can gate the different populations arbitrarily (the same axes as in b).

ing these results, it is possible to estimate the grazing rate by blue mussels for each cluster, and thus to assess whether or not the mussels select their food. Figure 3 shows the temporal evolution of each cluster on the basis of two variables, i.e., size and chlorophyll *a* fluorescence; alternatively, all the variables can be used together to produce a 2-dimensional principal components plot on which the clusters can be shown. Plots of

Table 1
Simple Statistics Given by the Standard Software of the Becton Dickinson FACS Analyzer

Date : 8/2/90	Sample ID : TO-T-C		
Cytometer : FACS Analyzer I	Sample tag : 005		File : TOTCOO5
	Contour statistics		
Parameters : VOL FL2	Gated events : 4240		Total events : 4240

#	X & Y Lower	X & Y Upper	Events	% Gated	% Tot	X & Y Mean	X & Y Mode	Peak
1	1.00 1.00	13.89 15.51	1714	40.42	40.42	3.79 4.05	1.93 3.73	16
2	17.30 3.34	33.40 8.03	559	13.18	13.18	21.84 5.11	21.54 5.18	254
3	17.30 17.30	173.02 268.27	1392	32.83	32.83	74.91 102.41	46.42 80.31	18
4	193.07 124.52	299.36 517.95	190	4.48	4.48	231.99 321.34	193.07 299.36	13
5	299.36 372.76	719.69 1000.00	129	3.04	3.04	424.49 594.28	299.36 464.16	7
6	1.00 1.00	1000.00 1000.00	4240	100.00	100.00	58.23 71.20	21.54 5.18	254

The last row 6, represents the whole population of fluorescent cells, while rows 1–5 represent the 5 gates shown in Fig. 2c.

Table 2
Centroids of Each Cluster in Terms of Cell Size, FL1, FL2, and Light Scatter

Cluster N	Size (μM)	FL1	FL2	Light scatter	Number of observations	%
1	11.57	1.0205	143.2775	109.7040	2903	(11.9)
2	5.46	1.0000	45.0766	35.5094	11149	(45.8)
3	15.55	1.2332	169.7905	143.7354	2940	(12.1)
4	22.15	2.6059	209.4775	175.3376	1695	(7.0)
5	9.84	98.9518	59.6884	154.4194	5652	(23.2)

Total numbers of observations in the whole series are indicated in the second to last column.

Table 3
Number of Observations in Each Cluster at Different Times

Time (min)/ cluster	1	2	3	4	5	Total 1–4
0	570	2215	741	275	506	3801
1	615	2312	666	389	545	3982
12	575	1969	554	312	577	3410
26	440	1682	484	308	567	2914
47	443	1756	383	217	550	2799
68	339	1297	298	166	531	2100
94	255	1036	291	169	557	1751
124	187	815	211	121	578	1334
221	30	140	21	3	543	194
316	13	91	22	9	607	135
448	6	51	10	1	597	68

Time 0 is before the toxic algae had been added. Figures are numbers of cells per 0.5 ml.

all pairs of variables may also be produced (Fig. 4). These graphs allow visualization of the criteria used to establish each cluster.

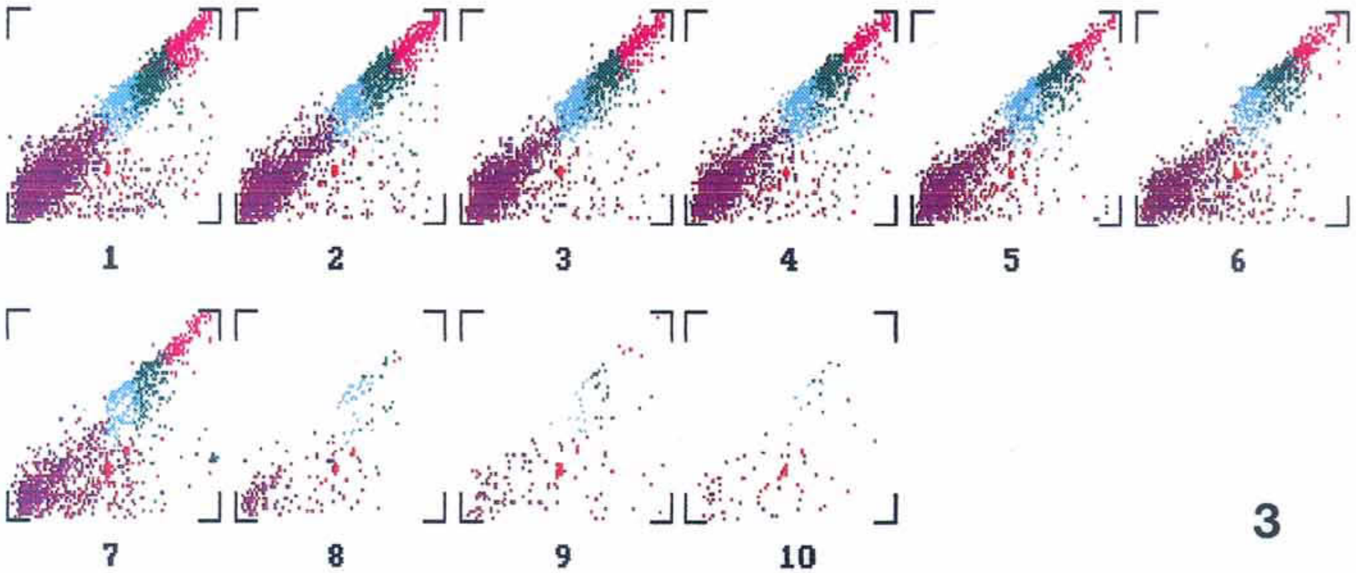
The organization of flow-cytometric data in a table such as Table 3 opens the way to more advanced statistical analyses, such as temporal autocorrelation analysis, time-constrained clustering, ordination analysis of the time series, search for discontinuities, regression analyses, and so on. Likewise, observations obtained in the spatial domain can be analyzed using multivariate spatial statistical methods (7).

DISCUSSION

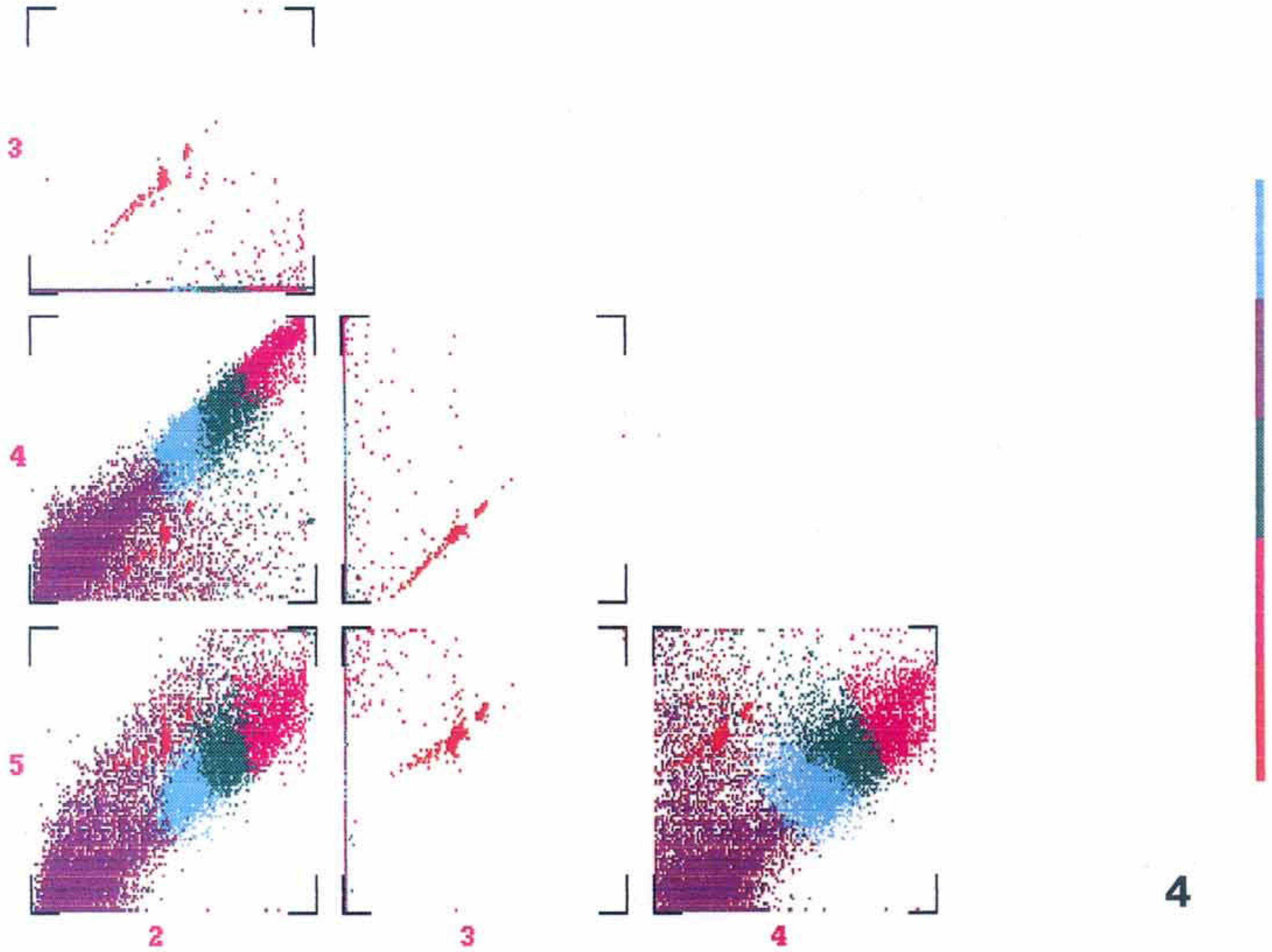
Using flow cytometry to analyze natural populations of aquatic particles has already resulted in significant discoveries (6), such as, for example, the existence of large populations of prochlorophytes in the oceans (2). However, the analysis and interpretation of flow-cytometric data are not a trivial task, and this has limited, so far, the usefulness of flow cytometry in studying oceanographic and limnological problems. This is because it has generally been impossible to relate flow-cytometric data to other oceanographic and limnological variables. As a result, flow cytometry is a powerful tool that has been mostly limited to a descriptive role in oceanography and limnology.

The clustering procedure described above is somewhat similar to the linear adaptive clustering scheme of Rohlf (12), which also computes distances between observations based on different variance-covariance matrices for each cluster. The linear adaptative clustering algorithm adjusts for cluster size by multiplying the determinant of the variance-covariance matrix (a measure of the hypervolume of the cluster) with the distance measure, thereby preventing the allocation of most points to one large cluster. In our algorithm, this is achieved by the use of critical distances.

It is important to note that although the cost function defined above is used to assess the goodness-of-fit measure, it does not have all the desirable statistical



3



4

FIG. 3. Temporal evolution of each cluster on the basis of two variables, i.e., size and chlorophyll *a* fluorescence. The different colors represent clusters (blue is cluster 1, purple is cluster 2, green is cluster 3, red is cluster 4, and orange is cluster 5). The decrease in the number of objects from time 1 to time 10 is the result of feeding by the mussels (x-axis is the volume and y-axis is the fluorescence emitted by chlorophyll *a*).

FIG. 4. Plots of all pairs of variables. Variable 2 is the volume; variable 3 is the FL1; variable 4 is the FL2 emitted by chlorophyll *a*; and variable 5 is the 90° light scatter. The colors correspond to the same clusters as in Fig. 3.

properties. First, it is not guaranteed that the algorithm will necessarily find the minimum solution. Empirically, the algorithm seems to converge within a few iterations (which is desirable for large data sets), but cluster allocations may not be optimal. One point that alleviates this problem is the fact that, in our program, the clustering results can be checked visually on the computer screen; a trained user may detect grossly inadequate results. Second, the solution found may not be unique, i.e., there may exist other cluster assignments that have the same cost function value. Third, comparing the cost function value for k clusters with that for, say k + 1 clusters is difficult. It must be understood that when the investigator a priori selects k as the number of desired clusters, a particular statistical model is also being selected. This is similar to choosing a model in regression analysis: The investigator may decide to fit a linear model or a quadratic function. In both cases, a R^2 statistic is computed and the investigator must decide whether an increase (or decrease) in the R^2 value is meaningful. At the limit, R^2 can be increased to 1.0 (perfect fit) by allowing as many model parameters as there are data points. In the same way, one can fit clusters perfectly by allowing as many clusters as there are observations.

For reasonably well-defined clusters, as we expect to find for mixed samples of very distinct populations, the above caveats should not be much of a problem. The large number of observations found in flow cytometry are advantageous in this case, since the large ratio of the number of observations to the number of model parameters (the k centroids, and the elements of the k variance-covariance matrices) results in more robustness. Another advantage of flow-cytometric data is that they are relatively low-dimensional (usually less then 10 variables). This allows a fairly accurate ordination of the data through such methods as principal component analysis. The ordination can be used by the investigator to determine whether well-defined clusters are present and, if so, how many such clusters exist and approximately where their centroids lie. The program we have developed allows preliminary visual analysis of the data. If the ordinations do not reveal a clear structure, numerous procedures have been developed for determining the correct number of clusters in a cluster analysis. Milligan and Cooper (10) have studied 30 such indices through a simulation study. They found that the most desirable index for their simulated data was the cubic clustering criterion, which is computed by the SAS package (13, 14). This index is the product of 2 components; the exact formula is given by Milligan and Cooper (10). The results must be taken with some caution, since again they may be data-dependent. As in all multivariate studies, the investigator should first look at the general structure of the data before proceeding with further analyses.

ACKNOWLEDGMENTS

We would like to thank Guylaine Desbiens, Réal Gagnon, and Nathalie Simard for their technical support during the course of this study, and Drs M. Fréchette, G. Salzman, E. Sakshaug, and F. J. Rholf for fruitful discussions. This work is a contribution of the Maurice Lamontagne Institute, Department of Fisheries and Oceans, Canada.

LITERATURE CITED

1. Anderberg MR: Cluster Analysis for Applications. Academic Press, New York, 1973, xiii + 359 p.
2. Chisholm SW, Olson RJ, Zettler ER, Waterbury J, Goericke R, Welschmeyer N: A novel free-living prochlorophyte occurs at high cell concentrations in the oceanic euphotic zone. Nature 334:340–343, 1988.
3. Guillard RRL, Ryther JH: Studies of marine planktonic diatoms. I. *Cyclotella nana* Huestedt and *Detonula confervacea* (Cleve) Gran. Can J Microbiol 8:229–239, 1962.
4. Hartigan JA: Clustering Algorithms. John Wiley & Son, New York, 1977, 351 p.
5. Jain AK, Dubes RC: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ, 1988, xiv + 320 p.
6. Legendre L, Yentsch CM: Overview of flow cytometry and image analysis in biological oceanography and limnology. Cytometry 10:501–510, 1989.
7. Legendre P, Fortin M-J: Spatial pattern and ecological analysis. Vegetatio 80:107–138, 1989.
8. MacQueen J: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Le Cam LM, Neyman J (eds). University of California Press, Berkeley, 1967, pp 281–297.
9. Mahalanobis PC: On the generalized distance in statistics. Proc Natl Inst Sci India 2:49–55, 1936.
10. Milligan GW, Cooper MC: An examination of procedures for determining the number of clusters in a data set. Psychometrika 50:159–179, 1985.
11. Neale PJ, Cullen JJ, Yentsch CM: Bio-optical inferences from chlorophyll *a* fluorescence: What kind of fluorescence is measured by analytical flow cytometry? Limnol Oceanogr 34:1739–1748, 1989.
12. Rohlf FJ: Adaptative hierarchical clustering schemes. Syst Zool 19:58–82, 1970.
13. Sarle WS: Cubic clustering criterion. Tech. Rep. A-108, Cary, NC: SAS Institute Inc., 1985.
14. SAS User's Guide: Statistics. Version 5, 1985.
15. Spinrad RS, Yentsch CM: Observations on the intra- and interspecific single cell optical variability of marine phytoplankton. Appl Opt 26(2):357–362, 1987.
16. Yentsch CM, Horan PK, Muirhead K, Dortch Q, Haugen E, Legendre L, Murphy LS, Perry MJ, Phinney DA, Pomponi SA, Spinrad RW, Wood M, Yentsch CS, Zahuranec BJ: Flow cytometry and cell sorting: A technique for analysis and sorting of aquatic particles. Limnol Oceanogr 28(6):1275–1280, 1983.