ICES Journal of Marine Science, 60: 872–884. 2003
doi:10.1016/S1054–3139(03)00025-0

A method for the identification and characterization of clusters of schools along the transect lines of fisheries-acoustic surveys

Pierre Petitgas

Petitgas, P. 2003. A method for the identification and characterization of clusters of schools along the transect lines of fisheries-acoustic surveys. – ICES Journal of Marine Science, 60: 872–884.

The school-aggregation pattern (schools and clusters of schools) is presumed to play a significant role in determining pelagic fish-stock catchability. However, its analysis has seldom been undertaken because it requires field-behavioural data that is seldom available. Such information can now be obtained by analysing school-based data of fisheries-acoustic surveys. This paper proposes a method for doing so. The method allows for the identification of clusters of schools and the estimation of their parameters along one-dimensional, acoustic-survey transect lines. It is based on a spatial point-process approach that considers schools as point events occurring along the track sailed by a ship. More precisely, it is based on defining a maximum distance between schools in a cluster. This distance is chosen to optimize various criteria and in particular that of homogeneity concerning school location inside the clusters and school number per km. The algorithm is described and applied to a series of acoustic surveys carried out in the Bay of Biscay. The pertinence of the clusters obtained by the algorithm is evaluated by analysing which component of the spatial distribution of the schools corresponds to those clusters. This involves considering all the distances between school events and performing simulations of cluster point processes. The school clusters obtained by the proposed algorithm represent a small-range structure of a few kilometres when a longer-range structure of tens of kilometres was also present in the data.

Keywords: acoustics, aggregation, clusters, fish schools, spatial point process.

Received 18 May 2001; accepted 12 December 2002.

P. Petitgas: IFREMER, Fisheries Ecology laboratory, BP 21105, 44311 Cedex 3, Nantes, France. Correspondence to P. Petitgas; tel: +33 240 374163; fax: +33 240 374075; e-mail: pierre.petitgas@ifremer.fr.

Introduction

As fishermen exploit aggregations of fish, the mesoscale characteristics of fish distribution such as clusters of schools are presumed to impact fish-stock catchability (Paloheimo and Dickie, 1964). Fréon and Misund (1999) present a review of theoretical scenarios relating fish density-dependent, spatial distribution and fishing strategies leading to stock depletion. The analysis of the interactions between the fish-aggregation pattern and other factors influencing stock catchability (e.g. Potier *et al.*, 1997) requires quantitative-behavioural data characterizing the schools and their clustering (e.g. the number of clusters, the number of schools in them, dimension of clusters, etc.) which is seldom available. These data can be obtained from the spatial analysis of school-based data derived from fisheries-acoustic surveys. Geo-referenced, school-based data can be routinely acquired now by post-processing the digitally recorded echograms of acoustic surveys using image-analysis software (ICES, 2000; Reid *et al.*, 2000). The objective of the present paper is to provide a procedure that allows the estimation of school-cluster parameters using the school-based data of fisheries-acoustic surveys.

Two approaches are possible for analysing statistically the spatial structure of schools (ICES, 2000): one is to consider the schools as discrete events of a spatial-point process, the other is to consider the number of schools per-unit-sailed distance (km) as a continuous variable forming a density surface. In the latter case, Geostatistics, Generalised Linear Models, or Additive Models (GLM or GAM) have been used. In the former case, point-process analysis can be used. The interest in considering school occurrence as a point process is that one can work on all distances at all

scales as well as estimate parameters of the school clusters. These are more informative for characterizing clusters than the unique value of the variogram range. This is because the variogram range is not only a function of the dimension of density patches but also of the distance separating them (Guardiola-Albert and Gomez-Hernandez, 2000). Clustering of schools has been reported using various statistical methodologies: geostatistics to acknowledge correlation structure in the number of schools per mile (Marchal and Petitgas, 1993; Petitgas and Lévénez, 1996), GAM to show the trend structure in the number of schools per mile (Beare *et al.*, 2000, 2002) and survival approach to illustrate the skew in the distance to the next neighbouring school (MacLennan and MacKenzie, 1988; Swartzman, 1997; Petitgas and Samb, 1998; Soria *et al.*, 1998). None of these statistical analyses allow the explicit definition of clusters of schools as objects and estimate indices for their parameters. In this paper a point-process approach is used to group schools in clusters of schools and estimate parameters that characterize this scale of spatial organization.

Two ways have been used to identify clusters and estimate their parameters to date. In the fisheries-ecological literature, clusters of schools have been identified based on the distance to the nearest neighbouring school (Swartzman, 1997; Soria *et al.*, 1998). A fixed-threshold distance was chosen based on experience that defined the maximum distance between schools in a cluster. Schools occurring at a distance greater than the threshold from the current school were aggregated to another cluster. Cluster parameters were estimated based on the clusters identified. In the statistical literature, Stoyan (1992) proposed a statistical-inference procedure that uses all the distances between schools and, in particular, those beyond the nearest neighbour. First, the spatial structure in the occurrence of point events was characterized by an experimental curve that plotted the probability that two point events occurred at a given distance from each other. Second, cluster-process models were simulated with various values for their parameters until the fit was acceptable between the observed curve and that corresponding to the simulated model. Values retained for the cluster parameters were those corresponding to the best fit.

In this study, schools were aggregated in clusters based on a simple, clustering algorithm using the distance to the next-school neighbour along the acoustic-transect lines. In contrast to Swartzman (1997) and Soria *et al.* (1998) the threshold distance was not fixed but chosen to optimize several criteria. Then, the pertinence of the clusters obtained in this way was evaluated by analysing which component of the school-spatial distribution corresponded to those clusters. To do this, the approach proposed by Stoyan (1992) was used. The aggregation in school occurrence was characterized by the pair-correlation function, an analogue to the variogram but for point process, which used all distances between schools. Cluster processes were simulated along the survey-transect lines with cluster parameters equal to those inferred by the next-neighbour, clustering algorithm. The pair-correlation function of the simulated processes was then compared with that of the school data. This made it possible to evaluate which part of the school-correlation structure corresponded to the identified clusters.

Materials and methods

School occurrence along acoustic-transect lines

A series of French pelagic acoustic surveys undertaken by IFREMER with RV "Thalassa" in the Bay of Biscay was worked on. Each survey was designed to monitor population abundance of the target species anchovy and sardine. Non-target species also present were sprat, horse-mackerel, and mackerel. Four surveys were selected which were similar in their design (Figure 1), with parallel East–West transects, regularly spaced every 10 nautical miles (nmi), from the Spanish border (43°30′N) to the isle of Ré (46°30′N) and traversing the entire continental shelf. Only daytime surveying was performed as fish aggregations dispersed too greatly at night making the separation between fish and plankton difficult (Massé, 1988). Transects were covered at 10 kn. The acoustic equipment was a hull-mounted, OSSIAN 38 kHz echosounder with a nominal beam angle of 7°. The pulse duration was 1 ms and the ping-repetition rate was 1 s^{-1}. The back-scattered acoustic signal was digitized providing acoustic samples of 10 cm in height and 5 m in length and these formed the echogram. Acoustic samples with a volume backscatter higher than −70 dB were saved. In the laboratory, the surveys were replayed for echo integration by school using a sample threshold of −60 dB (Petitgas *et al.*, 1998). School objects were identified and extracted from the echogram using MOVIES software (Weill *et al.*, 1993). Thresholds for minimum school-object length (1 ping: 5 m), height (two samples: 20 cm) and average density (−55 dB) were defined which allowed for their automated extraction. Extracted school objects with a length smaller than two beam widths at depth were rejected as being too small to be adequately characterized (ICES, 2000; Diner, 2001). Individual-school echotraces were not identified to species because the diagnostic power of school parameters was too imprecise in Biscay (Scalabrin *et al.*, 1996). All schools from all species were considered. For each school, latitude and longitude of the school centre was estimated as the average of latitude and longitude between start and end points of the school echotrace. The occurrence of schools along the survey track was studied as a one-dimensional point process and the school centres were considered as point events along the transect lines. Schools could occur at different depths in the water column. School locations were collapsed vertically and the distance between schools was the horizontal distance along the acoustic-transect lines. We focused on the one-dimensional analysis along the actual track steamed. A point-process analysis in 2D would require all process events to be recorded in 2D as is done, e.g. on a sonar image (ICES, 2000). Here school events are recorded along

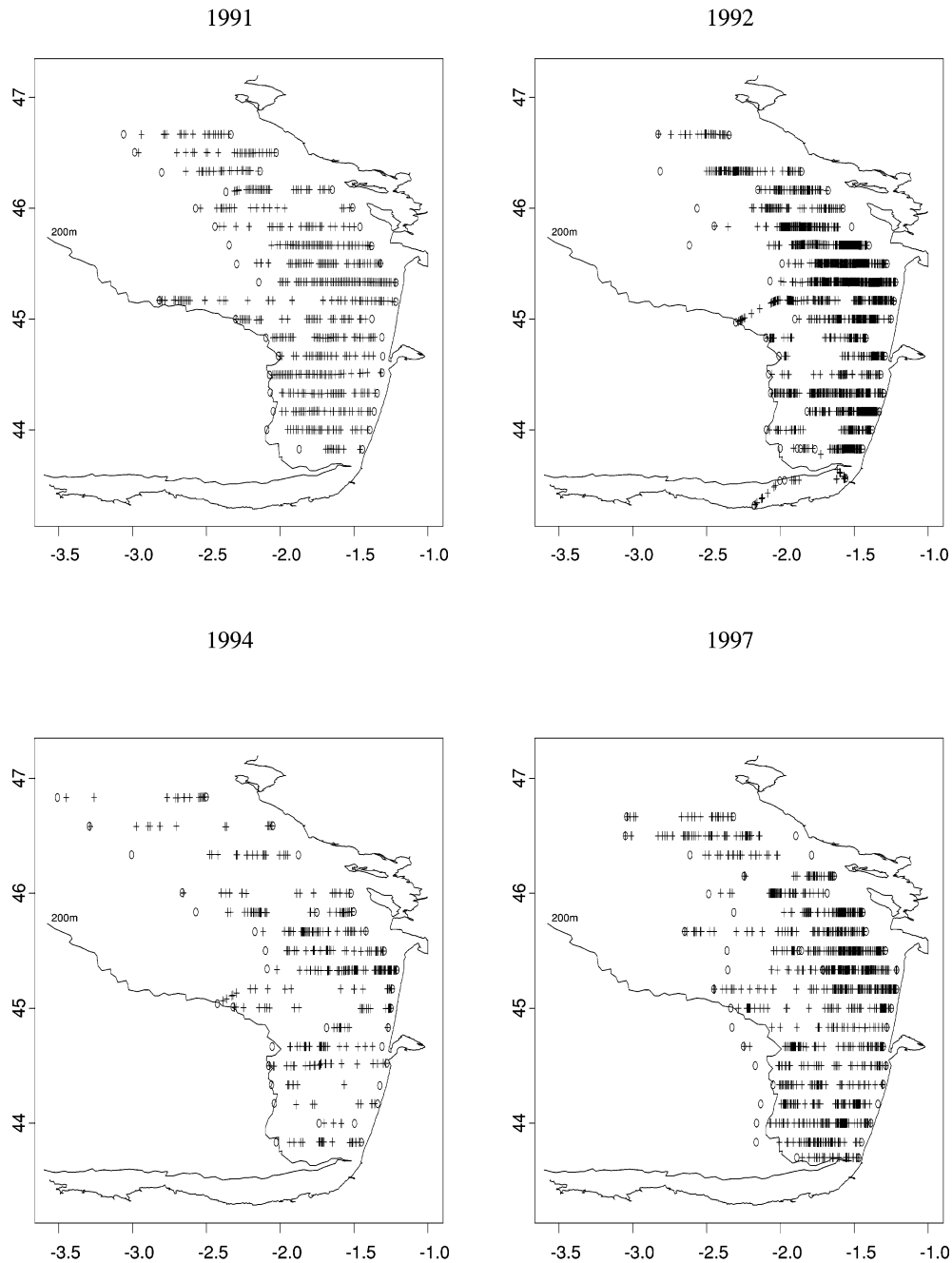

Figure 1. The school occurrence along the one-dimensional transect lines. Spring acoustic surveys of IFREMER carried out by RV "Thalassa" on the French continental shelf in the Bay of Biscay. Each cross indicates a school centre. The empty circles indicate the extremities of the transects. (The figures on the vertical axis show latitude North in degrees and those on the horizontal axis indicate longitude West in 0.5 degree intervals.)

parallel-retransect lines only because of the survey design. Therefore, a two-dimensional reconstruction of the school point process would require some hypothesis on the isotropy or anisotropy of school occurrence across the transects. No such assumptions were made.

Correlation-structure analysis

The correlation structure in the school occurrence along the transect lines was analysed using the pair-correlation function. Consider two infinitesimally small segments of length dx and dy with distance r between their centres.

Let P(r) denote the probability that a segment of length r has one school event in each extremity dx and dy. P(r) writes (Penttinen *et al.*, 1992; Stoyan and Stoyan, 1994):

$$P(r) = \lambda^2 g(r)dxdy.$$

The parameter λ is the intensity of the point process (i.e. the number of schools per km). The function g(r) is called the pair-correlation function. It is the probability-density function of the distance between pairs of point events. When there is no correlation, the pair-correlation function equals 1 for all distances. Interactions between point events (i.e. the correlation structure) will be shown by the departure of g(r) from the unit value. If r_0 denotes the range of the interactions, a cluster process will show values of g(r) higher than 1 for r smaller than r_0 while an inhibition process will show values of g(r) smaller than 1 for r smaller than r_0. Essentially, the pair-correlation function reads like a classical correlation. The pair-correlation function g(r) is the derivative of Ripley's K-function (Stoyan and Stoyan, 1994). It is here preferred to Ripley's K-function for two reasons. First, it interprets in a similar manner to a spatial correlation function and second, it is easier to estimate as no edge effects need to be corrected for at-transect extremities. We used the estimate for g(r) proposed by Stoyan and Stoyan (1994). Along a transect line, the estimate becomes:

$$g(r) = \frac{1}{\lambda^2} \sum_i \sum_{j \neq i} \frac{k_h(r - d_{ij})}{l_{TR} - d_{ij}}$$

where r is the distance at which the function is computed, k_h is a kernel of width h, d_{ij} the distance between school i and school j on the same transect TR, l_{TR} the length of transect TR, and λ the intensity. Because g(r) is a density function, a kernel estimator is useful. A kernel is a weighting function which allows more weight to be given to those distances d_{ij} that are closer to the distance r at which the function g(r) is being estimated. Following Stoyan and Stoyan (1994) the Epaneçnikov kernel (see Appendix) was used. The larger the value of the kernel bandwidth h, the smoother is the estimated curve $\hat{g}(r)$. After different trials the value $h = 0.25$ km was retained as satisfactory and used for all the years being studied. The distance between schools, d_{ij}, was computed as the distance between their centres. The estimate of g(r) was computed by pooling all transects together: the intensity λ was the total number of schools in the survey divided by the total survey length, N/L_{TR}; all the terms $k_h(r - d_{ij})/(l_{TR} - d_{ij})$ were summed with no reference to the transects. The estimate of the square intensity was $\hat{\lambda}^2 = N(N-1)/L_{TR}$. The pair-correlation function was estimated with a distance lag of 1 km.

Multicriteria, NND-clustering procedure

Because the transect lines were sailed in a particular direction, the lines were considered oriented. The next-neighbour distance (NND) was computed for each school along the transect lines. Schools were grouped in clusters based on the NND distances. This involved setting a threshold NND beyond which the next school was taken to be in a different cluster (i.e. the maximum distance between schools in a cluster). Schools beyond the threshold NND were considered too far to belong to the same cluster. There was little behavioural knowledge with which to define such threshold distance *a priori*. Historically, this was done using an empirical approach (Swartzman, 1997; Soria *et al.*, 1998). Basically, the researcher defined a distance based on his observations. The procedure proposed here was a more statistical approach to the problem.

For a given-threshold NND, schools were grouped into clusters and the following parameters estimated:

- The number of clusters (Nclus)
- Their length (Lclus)
- The number of schools per unit cluster length (λclus)
- The number of solitary schools (Nsoli)
- The homogeneity of the spatial distribution of schools within a cluster (Homog).

The length of clusters (Lclus) was estimated as the distance between the first and last schools of the cluster. Solitary schools were clusters of zero length and were not considered in the estimation of the average cluster length. The number of schools per-unit-cluster length (λclus) was estimated by the slope of the regression of the number of schools in a cluster on the cluster length. Clusters containing at least two schools were considered in the regression. The internal homogeneity of school occurrence inside the clusters was tested using a Kolmogorov test. Let x denote the distance to the next school in the cluster and τ_c its average. If the schools were Poisson distributed inside the cluster, the empirical distribution function of x would satisfy the exponential curve: $F(x) = 1 - e^{-x/\tau_c}$. The departure from the Poisson curve was checked using the Kolmogorov test, which is based on the value of the maximum difference between the empirical and model values for F(x). The significance level was set at 5%.

A range of threshold NNDs were considered. The threshold-NND retained minimized the following empirical criteria:

- Not too many clusters
- Not too many solitary schools
- Fewer than 5% clusters with non-homogeneous-school occurrence inside
- High R^2 for the regression of numbers of schools-per-cluster on length of clusters.

The range of threshold NNDs was chosen from the NND-cumulative distribution as the threshold lay in the inflexion part of the distribution. This was because the curvature in the distribution curve was related to the repetition rate of the process and thus to the scale of clustering. Consider the distance x between this school and the next as a survival time with mean τ. The homogeneous Poisson process has a

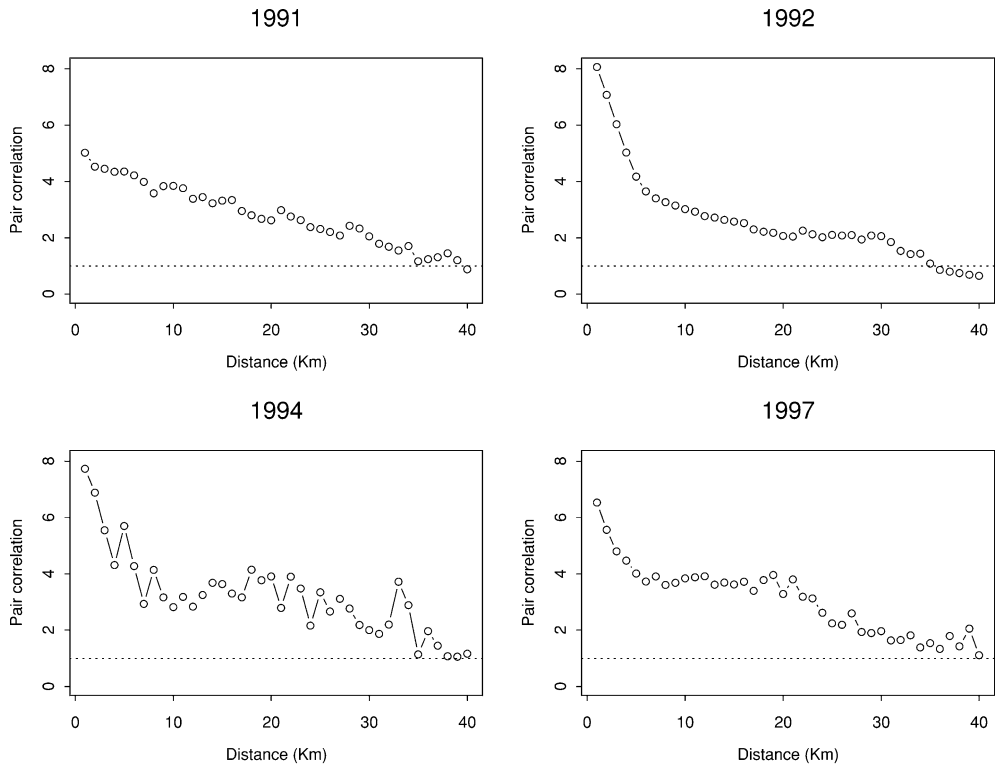

Figure 2. Pair-correlation functions for the four surveys characterizing the spatial structure in the school occurrence along the one-dimensional, survey transect lines.

repetition rate $1/\tau$ and its survival time x follows an exponential distribution $(1 - e^{-x/\tau})$. A process with its repetition rate accelerated by a constant a, a/τ, has its survival time x following a Weibull distribution $(1 - e^{-bx^a}$; with b a scaling factor) that is more skewed than the exponential (McCullagh and Nelder, 1995).

Essentially, the NDD threshold lay in the inflexion part of the NND distribution and was chosen to divide the point process of schools in clusters with similar number of schools-per-km and homogeneous-school distribution inside. The multicriteria, NND-clustering procedure was applied to each survey and cluster parameters estimated for each survey. The reference-conceptual, point-process model was a cluster process, in particular the Matern process (Stoyan and Stoyan, 1994).

Simulations of a cluster process with homogeneous distribution of parent events (Matern process)

A Matern process can be generated in two steps (Stoyan and Stoyan, 1994): first, parent events are positioned following a homogeneous Poisson process with intensity ρ and second, daughter events are uniformly and independently distributed within a distance R from each parent event. The resulting process of daughter events is called a Matern process. The number of daughter events per cluster around a parent point follows a Poisson distribution with parameter μ. A Matern process was simulated along the one-dimensional transect lines of each survey with parameter values for ρ, R and μ equal to those estimated using the multicriteria, NND-clustering procedure: $\hat{\mu} = \overline{L}\text{clus}\ \hat{\lambda}\text{clus}$; $\hat{R} = \overline{L}\text{clus}/2$; $\hat{\rho} = \text{Nclus}/L_{TR}$ with L_{TR} being the total survey length. Ten simulations were performed for each survey. For each simulation, the pair-correlation function was computed. The average for the ten simulations was computed for each survey and compared with that computed on the school data. This revealed that the clustering procedure only accounted for one component in the spatial structure and in particular it did not account for a long-range component observed in the data. Thus a two-stage point process was simulated along the survey-transect lines to account for the two scales in the spatial structuring.

Simulations of a cluster process with inhomogeneous distribution of parent events

The parent events of the Matern process were positioned according to an inhomogeneous Poisson process (i.e. a Poisson process for which the intensity varies in space). The number of schools was counted in segments of a given length (elementary-sampling, distance unit, ESDUs) along the survey transects. Presence of clusters within the ESDUs was coded 1 for those ESDUs containing at least two

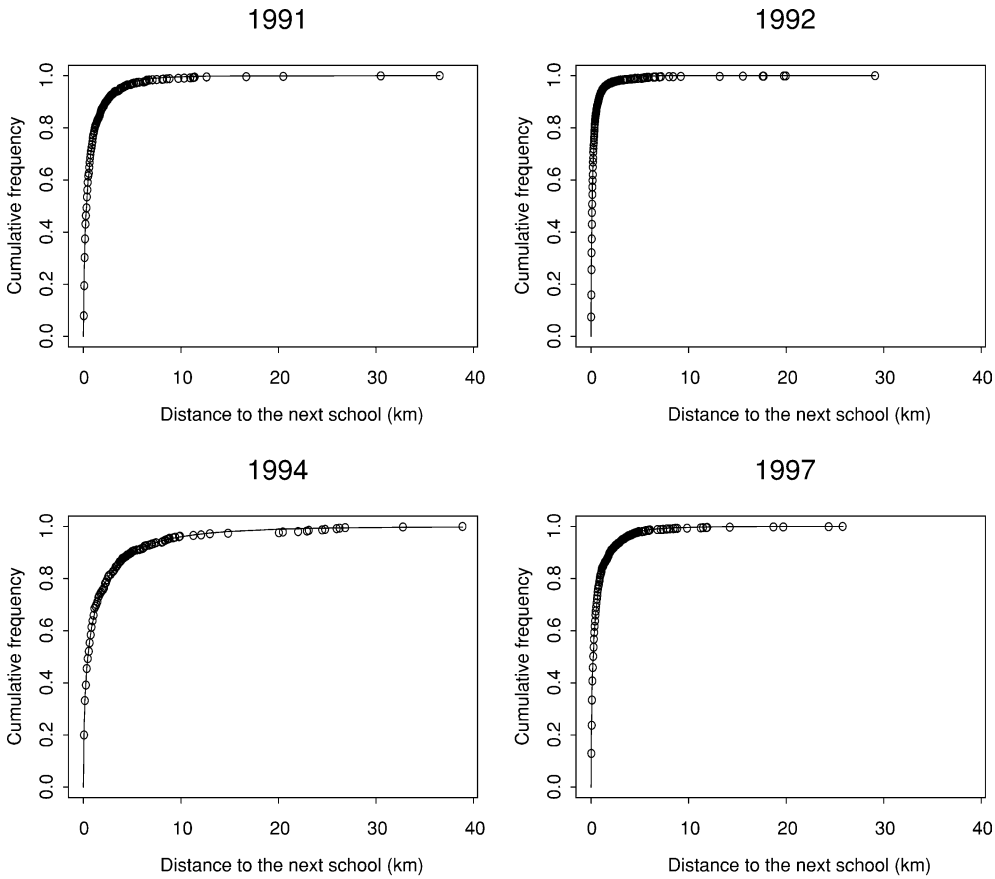

Figure 3. The cumulative distribution of the distance to the next school along the acoustic transect lines (points) together with their fitted Weibull distributions (lines). Model parameters are in Table 1.

schools and 0 otherwise. The probability for a cluster to occur in the ESDUs was estimated using a GAM (Hastie and Tibshirani, 1995). In a GAM, the survey values y_i are considered to come from a random function $Y(y_1, \ldots y_n, \ldots)$ where each sample value y_i is considered to be one outcome from a probability distribution with mean m_i and variance σ^2. The GAM allows for the estimation of the mean m_i as a function of specified, known covariates $X_j (j = 1, \ldots, p)$. The systematic components of the model are the link function L and the linear predictor η: $L(m) = \eta$ and $\eta = \sum_j S_j(X_j)$ where S are smoothers. The random component of the model is the probability distribution assumed for the random variables y_i (i.e. error distribution). The GAM allowed for the estimation of the probability π_i to have a cluster in ESDUs. The link function was the logit, the error distribution was binomial. The smoother used was the locally-weighted, regression-smoother of Cleveland (1979) called "loess" that is found in SPLUS (Inc.) software. The covariate considered was the interaction between latitude and longitude. No other covariate was tried as focus was put on extracting the crude trend in the data. The model fitted can be written as:

$$\mathrm{Log}(\pi_i/(1-\pi_i)) = \mathrm{loess}(\mathrm{lat},\ \mathrm{long},\ 0.25) + \text{binomial error.}$$

The goodness-of-fit of the model was quantified by the deviance, an analogue to the R-square in least-square, linear regression. The GAM provided a trend surface for the probability π_i of cluster occurrence. A threshold value was

Table 1. Modelling the probability distribution, F(x), of the distance to the next school as a survival time, x, using a Weibull distribution, $F(x) = 1 - e^{-bx^a}$. The table gives the values of the Weibull parameters, a and b, estimated by non-linear least square. R^2 is the R-square of the fit.

Years	a	b	R^2
1991	0.060	1.412	0.997
1992	0.623	2.876	0.990
1994	0.497	1.031	0.999
1997	0.551	1.638	0.996

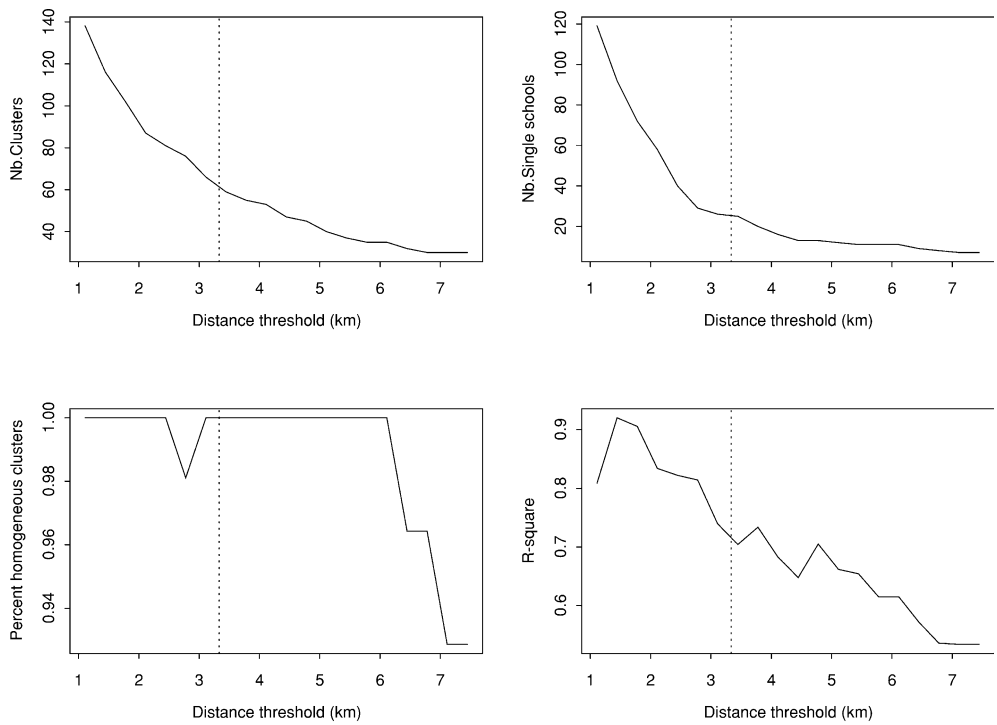

Figure 4. Multicriteria curves for 1991. The threshold distance chosen is indicated by the dotted, vertical line.

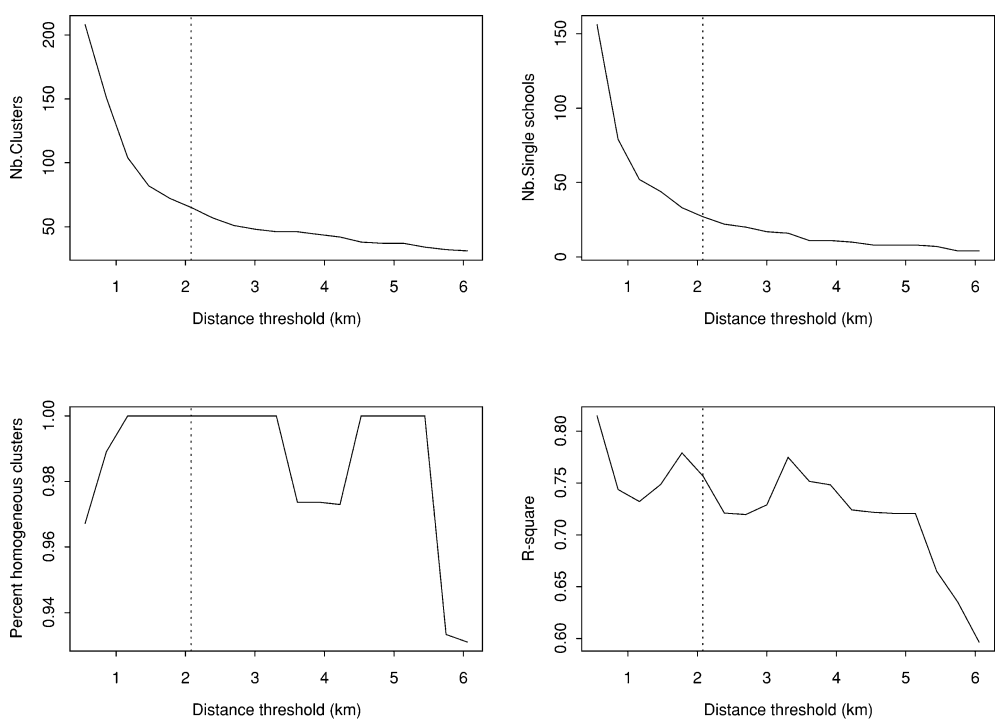

Figure 5. Multicriteria curves for 1992. The threshold distance chosen is indicated by the dotted, vertical line.

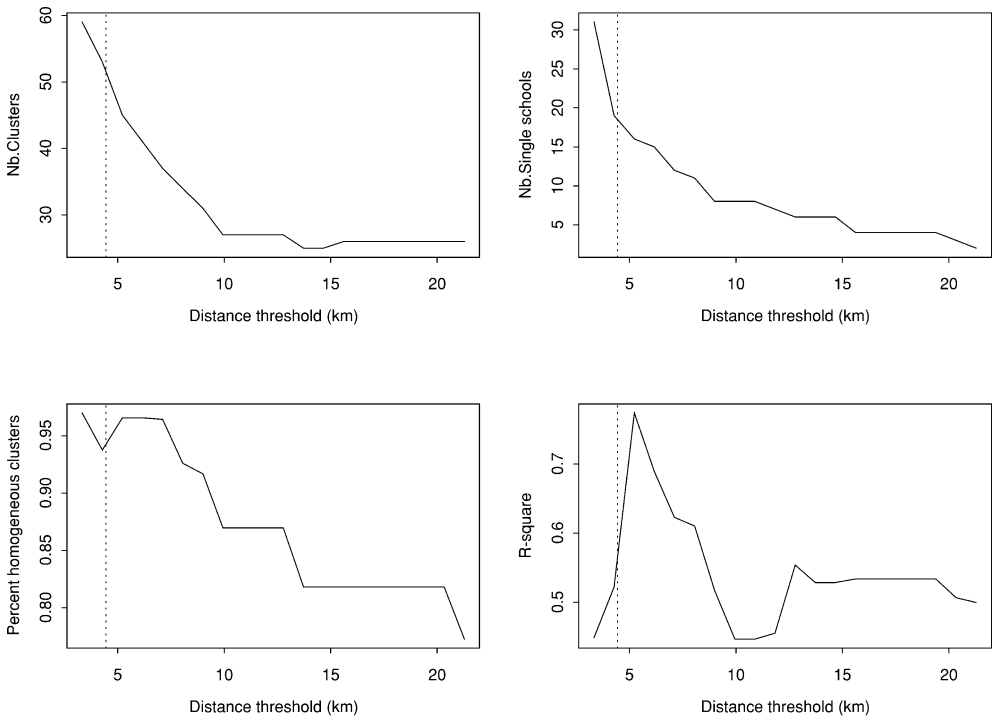

Figure 6. Multicriteria curves for 1994. The threshold distance chosen is indicated by the dotted, vertical line.

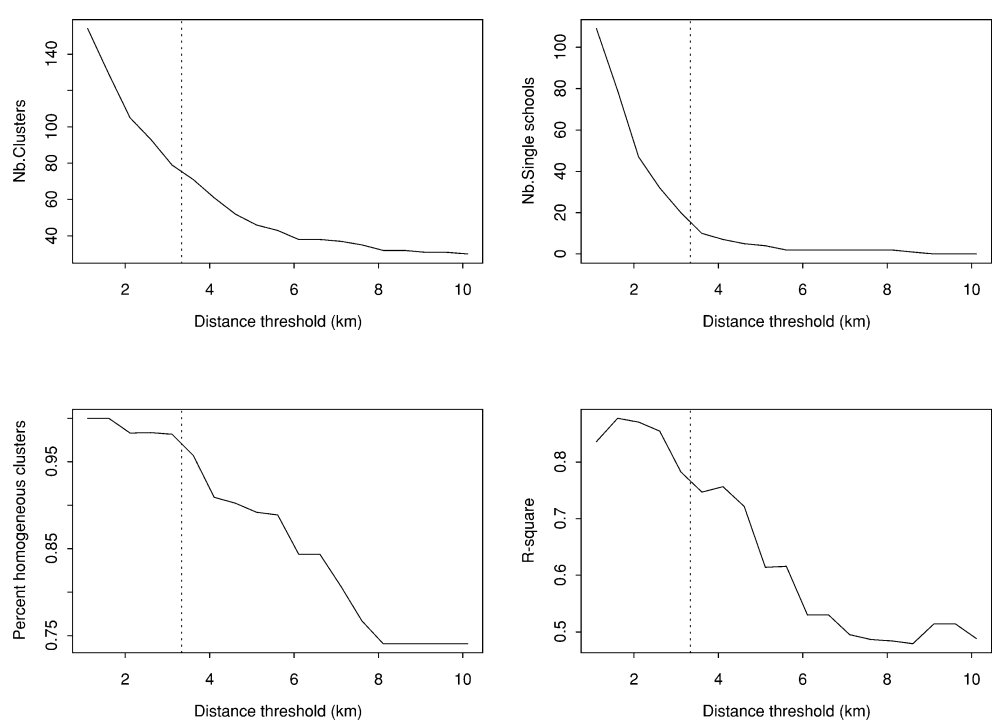

Figure 7. Multicriteria curves for 1997. The threshold distance chosen is indicated by the dotted vertical line.

Table 2. Parameter values inferred by the multicriteria, NND-clustering procedure. Ntot, total number of schools; Lsurv, sum of transect lengths; Av.dtn, average next-neighbour distance; DT, threshold distance; Nsoli, number of solitary schools; Nclus, number of clusters; Lclus, average cluster length; λclus, average number of schools per unit cluster length; R^2, R-square of the linear regression of the number of schools in the clusters on the length of clusters; Homog, percent of clusters with homogeneous distribution of schools.

Years	Ntot	Lsurv	Av.dtn	DT	Nsoli	Nclus	Lclus	λclus	R^2	Homog
1991	1111	1202	0.97	3.33	25	60	9.98	2.36	0.70	1.00
1992	2671	1119	0.37	2.08	27	65	8.84	5.71	0.76	1.00
1994	482	1140	2.07	4.45	19	52	6.53	1.77	0.84	0.95
1997	1487	1341	0.78	3.33	13	79	8.46	2.66	0.77	0.97

then chosen on this probability and one parent event was positioned at random in the ESDUs that had a probability higher than the threshold. The next step was to position the daughter points at random within distance R from the parent points. In a Matern process, the parent events are positioned randomly and uniformly but in this case their position was constrained by the probability surface. For each survey, 10 simulations were performed and the pair-correlation function computed for each simulation. The average of the 10 curves was then estimated and compared with the pair-correlation function computed on the school data.

Results

School occurrence and pair-correlation function along transect lines

For each survey, school occurrence as extracted from the echogram was plotted along the acoustic transect lines (Figure 1). The schools were grouped visually in small clusters which themselves were structured in a regional pattern. In particular, there were many schools in all the years studied in the area in front of the Gironde estuary and this influenced the regional pattern. The pair-correlation function was computed for each survey along the transect lines (Figure 2). The deviation of the curves from the Poisson, horizontal-unit line suggested a clustering of schools. In all years except 1991, the curves showed a similar behaviour: a rapid decrease with a correlation range between 5 and 10 km and then a slower decrease with a correlation range between 35 and 40 km. Such behaviour agreed with that of the variogram of school number per km computed for the same years (Petitgas, 2000). From this it can be inferred that the continuous-density-surface approach or the point-process approach were similar in characterizing school-correlation structure.

Table 3. Parameter values used for simulating Matern processes along the survey lines. λ, school intensity; ρ, cluster intensity; μ, average number of schools per cluster; 2R, maximum cluster length.

Years	λ	ρ	μ	R
1991	1.20	0.050	23	5
1992	2.40	0.058	50	4.5
1994	0.55	0.046	12	3
1997	1.36	0.059	23	4

Multicriteria, NND-clustering procedure

The cumulative distribution of the distance to the next school (NND) along the transects was computed for each survey by pooling all NNDs from all transects (Figure 3). The distributions showed a concave shape modelled by a Weibull distribution (Table 1) fitted by non-linear least squares (routine "nls" in SPLUS, Mathsoft Inc.). For all the surveys, the inflexion part of the curve was between 5 and 10 km and this was in coherence with the small correlation range visible on the pair-correlation functions (Figure 2). The threshold NND was expected to lie in the range 5–10 km. For each survey, several threshold distances were tried in that range, schools were grouped in clusters, and criteria parameters estimated (Figures 4–7). For each survey, one threshold distance was retained and cluster parameters estimated: Nclus, Lclus, λclus, Nsoli, Homog (Table 2). Average cluster lengths (Lclus) were in agreement with the ranges observed on the pair-correlation functions, except for 1991 (see Discussion).

Simulation of a Matern process

The parameters of a Matern process that were considered were parameters μ, R, and ρ estimated by the clustering procedure described above (Table 3). In the simulations, the number of schools in a cluster was a random variable following a Poisson probability distribution with parameter μ. Parameters R and ρ were constants. The school intensity λ (average school number per km) varied in each simulation because it was an outcome from a Poisson distribution with parameter ρ μ. In each year, the pair-correlation function corresponding to the simulated Matern process was estimated. It showed a small-range structure in agreement with that of the school-data, pair-correlation function (Figure 8). However, the slow decrease in the school-data, pair-correlation function was not reproduced by the

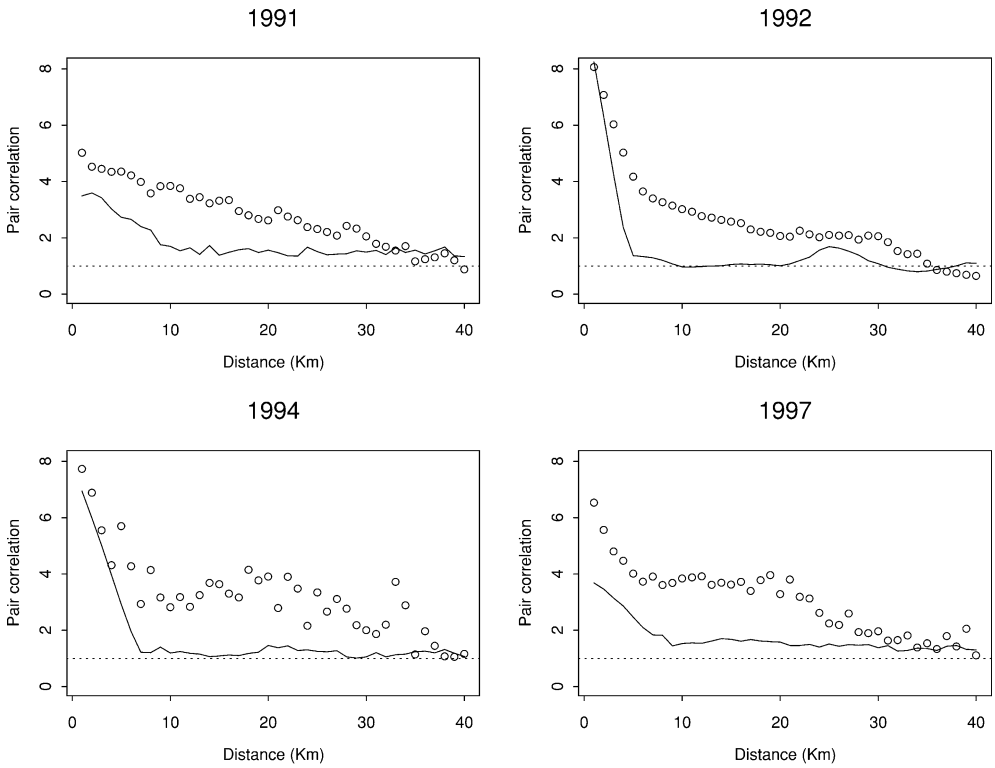

Figure 8. Pair-correlation functions compared between the school data (empty circles) and simulated Matern process (lines).

simulations of the Matern process. The fact that the small-range structure was well reproduced meant that the multi-criteria, NND-clustering procedure identified the small scale correlation range only. In the simulations the clusters were positioned at random along the survey transects. The slow decrease in the experimental curves presumably accounted for the structure made by a non-random positioning of the clusters relative to one another. In order to investigate this assumption further a model was simulated where clusters were positioned around parent points with their locations constrained by a trend surface.

Simulations of a cluster process with inhomogeneous parent events

The survey transect lines were binned in ESDUs of 10 km and the number of schools counted in each. The presence of clusters within the ESDUs was coded 1 for ESDUs containing at least two schools and 0 otherwise. The size of 10 km for the ESDUs was chosen because average cluster lengths were in the range 6–10 km for the different surveys (Table 2). The probability π_i for a cluster to occur in the ESDUs was estimated by fitting a GAM on presence or absence data, using the routine "gam" in Splus (Mathsoft, Inc.). Those ESDUs were then selected which had their π_i greater than a given probability threshold. The threshold was set so as to produce a cluster intensity ρ compatible with that estimated on the data by the multicriteria, NND-clustering procedure. The probability-threshold values for the different years 1991, 1992, 1994, and 1997 were, respectively, 0.95, 0.96, 0.62 and 0.80, and the explained deviances were 12, 18, 8, and 9%. The explained deviances were small, meaning that the trend estimated represented only a small part of the spatial variability. For each year, the pair-correlation function corresponding to the simulations was computed and compared with that of the school data (Figure 9). The small-range structure as well as the slow decrease observed on the school data were both reproduced in the simulations. This confirmed the first interpretation that school clusters were positioned relative to one another according to a regional pattern that is modelled here by a trend surface.

Discussion and conclusion

The distribution of the along-transect NND was well modelled by a Weibull distribution. MacLennan and MacKenzie (1988) and Petitgas and Samb (1998) had already used the Weibull on their data (northern North Sea herring and sardinella off Senegal). The Weibull has two parameters which certainly make it flexible in many cases. However, the model also has the property of characterizing

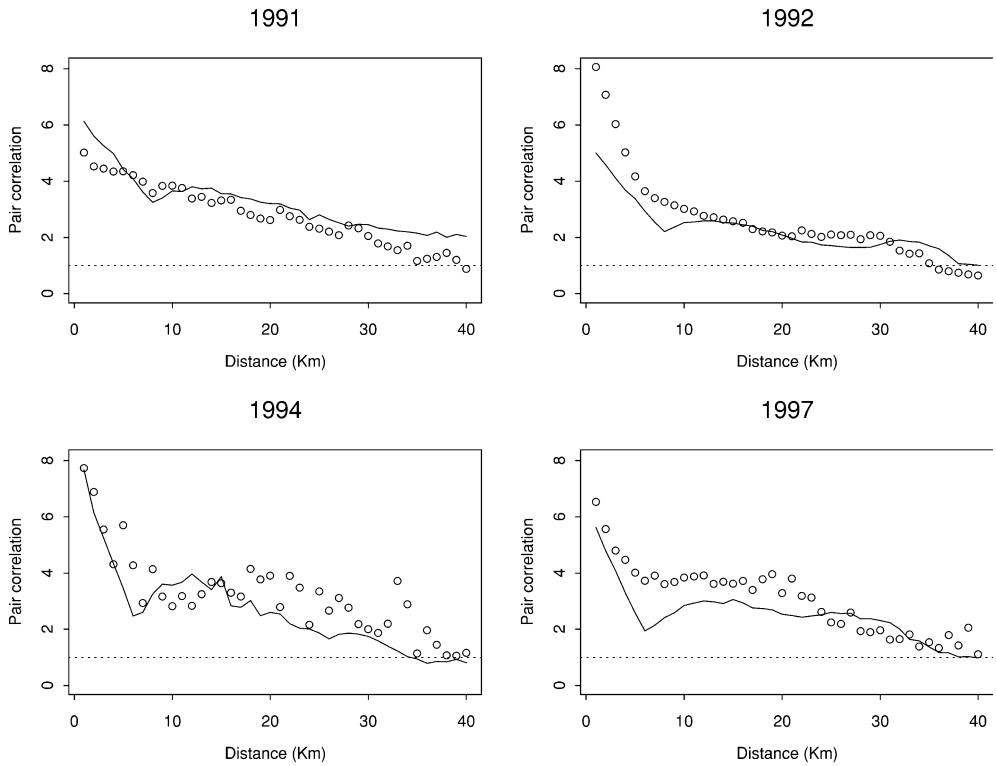

Figure 9. Pair-correlation functions compared between the school data (empty circles) and a two-stage, cluster process (lines). The first stage of the process is an Inhomogeneous-Poisson process giving the locations of parent events. The second stage is a Matern-cluster process around the parent events.

a process with a repetition rate higher than for a Poisson process, meaning that its use is adapted to cluster processes (McCullagh and Nelder, 1995).

In the multicriteria, NND-clustering procedure the threshold distance varied from year to year in accordance with the year to year change in the statistical characteristics of the spatial pattern. Other authors used a fixed threshold (Swartzman, 1997; Soria *et al.*, 1998). A varying threshold will tend to produce clusters that are homogeneous in every year whereas a fixed threshold may produce inhomogeneous clusters for particular years. It seems difficult to say which is better without behavioural knowledge of the schools within clusters. Although variable, the threshold distance did not vary much across years (2–5 km).

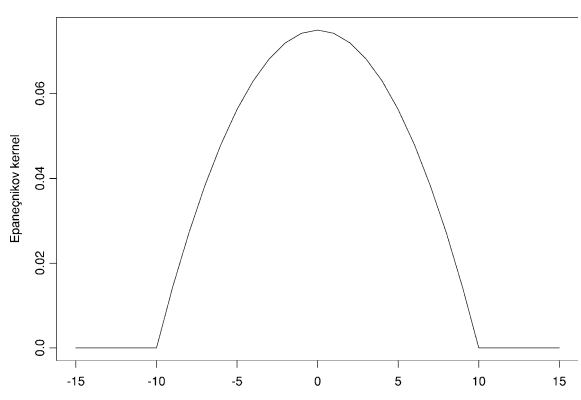

Figure 10. Epaneçnikov kernel with a bandwidth h = 10.

In this paper the estimation of synthetic variables characterizing a fish school aggregation pattern was felt to be a necessary first step to the analysis of how this spatial pattern determined population catchability when interacting with other factors such as e.g. fishing tactics or the environment. The focus of the approach taken was on the methods needed to group schools in clusters and estimate their parameters, school-based data being derived from image analysis of echograms of echosounder surveys. A simple procedure was to group schools in clusters depending on the NND along the track sailed. The clusters estimated by the multicriteria, NND clustering, procedure had internal homogeneity characteristics, i.e. similar number of schools per km and the homogeneous distribution of schools. The clusters estimated in this way corresponded to a small-range spatial structure (5–10 km) leaving a longer-range structure (35–40 km) uncharacterized. The longer-range structure was not identifiable on the NND cumulative distribution

as distances between clusters were too few in number. Therefore, the analysis of the school pattern using only the NND can lead to larger-scale spatial structures that correspond to assemblages of clusters being missed. A practical solution could be to estimate each and every time the pair-correlation function or the variogram of school number per unit distance.

The pair-correlation function allowed for the characterization of the spatial structure over all its spatial scales. Two scales of structuring were found. The clusters identified by the multicriteria, NND-clustering procedure corresponded to the smaller structure (5–10 km). The larger structure was generated by a non-random positioning of the clusters relative to one another. This correspondence was found by undertaking simulations of cluster-point processes and comparing the pair-correlation function of the school data with that of the simulations. The approach was proposed by Stoyan (1992) for estimating cluster parameters and was very appropriate to the situations found in this case where school occurrence was structured at different spatial scales.

For 1991, the pair-correlation function did not identify a small-range structure (Figure 2) whereas the multicriteria, NND-clustering procedure did (Figure 3). Clusters were identifiable visually when plotting the school occurrence along the transect lines (Figure 1). Cluster parameters were not found to be different for that year compared with other years (Table 2). Observed and simulated pair-correlation functions agreed (Figure 9). Thus, clusters of a few kilometres in length were real but their relative positions in this particular year would have made the cluster structure not visible on the pair-correlation function.

Between the smallest spatial structure, the school, and the largest, the population, this study showed two scales of spatial organization: the clusters of schools and the assemblage of clusters (either in larger clusters or along trends). It can be hypothesized that the clusters of schools could be controlled by the dynamical behaviour of schools at the small scale whereas the assemblage of clusters could be organized under environmental factors structured at the regional scale.

In the series of surveys analysed here, 1994 and 1992 were two contrasting years with, respectively, low and high total school numbers. Table 2 shows a relationship between the total number of schools and the school-clustering pattern. When the total number of schools was low (in 1994) the number of solitary schools was high and the number of clusters, their length, and the number of schools in them was low. In contrast, when the total number of schools was high (in 1992), only one cluster parameter responded, i.e. the number of schools in the clusters. These findings agree with the general behaviour of a variety of European pelagic stocks (Petitgas *et al.*, 2001) relating the number of schools and the clustering pattern.

Acknowledgements

This work was supported in part by the European Union program CLUSTER (FAIR CT 96.1799). Particular thanks go to J. Massé, P. Beillois, and N. Schreiber for the provision of the school data sets. The idea of the multi-criteria, NND-clustering procedure was formulated at the London meeting of the CLUSTER program in 1999 via the input of D. Reid and based on the work of Petitgas and Samb (1998).

References

Beare, D., Reid, D., Petitgas, P., Carrera, P., Georgakarakos, S., Haralambous, J., Iglesias, M., Liorzou, B., Massé, J., and Muiño, R. 2000. Spatio-temporal patterns in pelagic fish school abundance and size: a study of pelagic fish aggregation using acoustic surveys from Senegal to Shetland. ICES CM 2000/K: 03.

Beare, D., Reid, D., and Petitgas, P. 2002. Spatio-temporal patterns in herring (*Clupea harengus* L.) school abundance and size in the northwest North Sea: modelling space-time dependencies to allow examination of the impact of local school abundance on school size. ICES Journal of Marine Science, 59: 469–479.

Cleveland, W. 1979. Robust locally-weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74: 829–836.

Diner, N. 2001. Correction on school geometry and density: approach based on acoustic image simulation. Aquatic Living Resources, 14: 211–222.

Fréon, P., and Misund, O. 1999. Dynamics of Pelagic Fish Distribution and Behaviour: Effects on Fisheries and Stock Assessment. Fishing News Books. Blackwell, Oxford. 348 pp.

Guardiola-Albert, C., and Gomez-Hernandez, J. 2000. Average length of objects defined from binary realisations. *In* GeoEnv III—Geostatistics for Environmental Applications, pp. 323–332. Ed. by P. Monestier, D. Allard, and R. Froidevaux. Kluwer Academic Publishers, 537 pp.

Hastie, T., and Tibshirani, R. 1995. Generalised Additive Models. Chapman and Hall, London, 335 pp.

ICES. 2000. Report on Echotrace Classification. Ed. by D. Reid. ICES Cooperative Research Report, No. 238. 107 pp.

MacLennan, D., and MacKenzie, I. 1988. Precision of acoustic fish stock estimates. Canadian Journal of Fisheries and Aquatic Sciences, 45: 605–616.

Marchal, E., and Petitgas, P. 1993. Precision of acoustic fish stock estimates: separating the number of schools from the biomass in the schools. Aquatic Living Resources, 6(3): 211–219.

Massé, J. 1988. Utilisation de l'écho-intégration en recherche halieutique. Rapport IFREMER, DRV-88030-RH/Nantes.

McCullagh, P., and Nelder, J., 1995. Generalised Linear Models, 2nd edition. Chapman and Hall, London. 511 pp.

Paloheimo, J., and Dickie, L.1964. Abundance and fishing success. Rapports et Procès-Verbaux des Réunions du Conseil International pour l'Exploration de la Mer, 155: 152–163.

Penttinen, A., Stoyan, D., and Henttonen, H.1992. Marked point processes in forest science. Forest Science, 38: 806–824.

Petitgas, P. 2000. On the clustering of fish schools at two scales and their relation with meso-scale physical structures. ICES CM 2000/K: 25.

Petitgas, P., and Lévénez, J.-J. 1996. Spatial organization of pelagic fish: echogram structure, spatio-temporal condition, and biomass in Senegalese waters. ICES Journal of Marine Science, 53: 147–154.

Petitgas, P., and Samb, B. 1998. On the clustered occurrence of fish schools along acoustic survey lines and its relation with population abundance. ICES CM 1998/J: 5.

Petitgas, P., Diner, N., Georgakarakos, S., Reid, D., Aukland, R., Massé, J., Scalabrin, C., Iglesias, M., Muino, R., and Carrera, P. 1998. Sensitivity analysis of school parameters to compare schools from different surveys: a review of the standardisation task of the EC FAIR program CLUSTER. ICES CM 1998/J: 3.

Petitgas, P., Reid, D., Carrera, P., Iglesias, M., Georgakarakos, S., Liorzou, B., and Massé, J. 2001. On the relation between schools, clusters of schools, and abundance in pelagic fish. ICES Journal of Marine Science, 58: 1150–1160.

Potier, M., Petitgas, P., and Petit, D. 1997. Interaction between fish and fishing vessels in the Javanese purse seine fishery. Aquatic Living Resources, 10: 149–156.

Reid, D., Scalabrin, C., Petitgas, P., Massé, J., Aukland, R., Carrera, P., and Georgakarakos, S. 2000. Standard protocols for the analysis of school based data from echosounder surveys. Fisheries Research, 47: 125–136.

Scalabrin, C., Diner, N., Weill, A., Hillion, A., and Mouchot, M.-C. 1996. Narrowband acoustic identification of monospecific fish shoals. ICES Journal of Marine Science, 53: 181–188.

Soria, M., Petitgas, P., and Bahri, T. 1998. On the size of fish schools and clusters: a spatial analysis of multibeam sonar images in the Mediterranean Sea. ICES CM 1998/J: 8.

Stoyan, D. 1992. Statistical estimation of model parameters of planar Neyman-Scott Cluster processes. Metrika, 39: 67–74.

Stoyan, D., and Stoyan, H. 1994. Fractals, random shapes and point fields. Wiley, New York. 389 pp.

Swartzman, G. 1997. Analysis of the summer distribution of fish schools in the Pacific Eastern Boundary Current. ICES Journal of Marine Science, 54: 105–116.

Weill, A., Scalabrin, C., and Diner, N. 1993. MOVIES-B: an acoustic detection description software, application to shoal species classification. Aquatic Living Resources, 6: 169–296.

Appendix: Kernel estimator of a probability-density function (after Stoyan and Stoyan, 1994)

We wish to estimate a probability-density function on a continuous range of values x using a finite, discrete number of sample values $x_1, \ldots, x_n$. Let f(x) be the density function to be estimated. Let k(x) be another density function, the kernel function. The kernel is usually a symmetrical function: $k(-x) = k(x)$. The kernel estimator of f(x) is:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k(x - x_i).$$

There are different kernels possible. The Epaneçnikov kernel is often used. It is defined by:

$$e_h(t) = \begin{cases} 0.75(1 - t^2/h^2)/h & (-h \le t \le h) \\ 0 & \text{otherwise} \end{cases}$$

where h is a bandwidth centred on x and $t = x - x_i$. The kernel allows the estimation of f(x) at the unknown value x by weighting those sample values that stand in the bandwidth h centred on x. The sample values outside the bandwidth do not intervene in the estimate of f at x. The estimated f(x) will be rough for small h and will get smoother as h increases. The particular choice of the form of the kernel is less decisive than the bandwidth and it is recommended that preliminary trials are carried out with various bandwidths. The Epaneçnikov kernel with a bandwidth of 10 is plotted in Figure 10 as an example.